SAND2024-15493O

Sandia National Laboratories

URA UNIVERSITIES RESEARCH ASSOCIATION

# AICoM

APRIL 2-3, 2024

## WORKSHOP REPORT

# AI-Enhanced Co-Design for Next Generation Microelectronics

PRINCIPAL ORGANIZER: **Conrad James,** *Sandia National Laboratories*

U.S. DEPARTMENT OF ENERGY NNSA

# Table of Contents

# Executive Summary

The power of artificial intelligence/machine learning (AI/ML) necessitates that the R&D community convenes regularly to discuss research directions, collaboration opportunities, and application impact. For microelectronics, a field that is facing tremendous pressure in terms of Moore's law and energy consumption, it is especially important that researchers discuss ideas for AI/ML to improve microelectronic hardware design, fabrication, and production. The AI-Enhanced Co-Design for Next Generation Microelectronics workshop series was started in 2021 to bring together academic, industry, and government researchers at every wedge of the co-design circle of innovation [1] to share their research and insights on co-design, and to build a community of practice that can leverage new tools, algorithms, experimental techniques, etc. Topics have spanned a wide range from high performance computing, neuromorphic computing, AI algorithm development, electronic design automation, electronic materials, and microfabrication techniques. We have also conducted several forums around workforce development given how crucial the needs are for properly skilled and trained technicians, engineers, and scientists. Given the CHIPS and Science Act legislation and other federal government funding opportunities in microelectronics and AI/ML, we look forward to continued engagement among this community to push the R&D and technology maturation needed to meet the world's growing needs for computing.

## *References*

[1] https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf

# Introduction

Conrad James
Manager, R&D Science and Engineering, Sandia National Laboratories

## Co-Design Remains an Innovative Approach to Meet Our Current Technology Demands

Unrelenting consumer demand for semiconductor products has led researchers around the world to search for new ways to fabricate and operate microelectronic systems with reduced energy consumption. The drastic increase in the prevalence and use of artificial intelligence/machine learning (AI/ML) has also contributed to this pressure as more mechanical and human-operated systems are being filled with microelectronics and automated. The purpose of this workshop over the last four years has been to provide a forum for researchers to present their work in AI/ML and microelectronics, and to encourage the cross-disciplinary co-design research activities needed for the next generation of AI/ML- enhanced algorithms, circuit architectures, and microelectronic systems. These workshops have also led to new collaborations between researchers and has helped to identify opportunities for new tools or methodologies to be developed to further co-design. As part of the 2024 workshop report, we wanted to provide a short history of this series of workshops that began in 2021.

### *Workshop I, 2021*

In our first workshop [1], we discussed the most pressing opportunities and challenges in microelectronics, how modern artificial intelligence (AI) can accelerate microelectronics co-design research, and prospects for collaborative advances spanning multiple knowledge domains and scientific disciplines. As part of an exercise to prepare for the workshop, teams of SNL researchers were tasked with developing examples of co-design research involving microelectronics and AI/ML. During the workshop, two of these exemplar problems were presented to the participants. The first exemplar focused on probabilistic computing and raised the question of what types of computational problems could be solved if we developed new energy-efficient and small-footprint hardware capable of generating large amounts of random numbers. Within four months of the workshop, this exemplar had been turned into a DOE-SC proposal and then funded microelectronics co-design project, COINFLIPS (CO-designed Improved Neural Foundations Leveraging Inherent Physics Stochasticity) [2]. The second exemplar focused on non-volatile memory devices designed for synaptic weight storage in neuromorphic computing hardware. This exemplar later became a component of an Energy Frontier Research Center that was funded in 2022, REMIND (Reconfigurable Electronic Materials Inspired by Nonlinear Neuron Dynamics) [3].

### *Workshop II, 2022*

In the second AICoM workshop in April 2022, we had several speakers provide updates on the use of AI in microelectronics codesign and we focused some of our discussion around the challenges with moving into a post-Moore's Law era. We also began a dialogue on workforce development, an important challenge for interdisciplinary research. Participants discussed the need for new tools and validation approaches to advance co-disciplinary research, and we also discussed the challenges of synchronizing collaboration between disciplines that have different time-constants of innovation (e.g., software changes rapidly whereas hardware is slower). The role of AI in data analysis has changed since its first use to generate phenomenological models from large volumes of data. In the first workshop in 2021, the maturation of AI from a data analysis tool to a tool that could be used to explain general and physically relevant trends in

experimental data was noted. In the 2022 workshop, participants described the potential use of AI/ML to identify the next set of experiments to conduct to generate more data for optimizing physics-informed models and improving their fidelity.

### *Workshop III, 2023*

For the 2023 workshop, we discussed the extent to which the R&D community had embraced co-design. Participants described the academic community as moving relatively quickly into the co-design approach, but that it was still in its infancy. We discussed the impact of the DOE basic research needs (BRNs) workshop reports from the last several years, including the BRNs for microelectronics, extreme heterogeneity, and advanced scientific computing. Since the 2018 BRN report on microelectronics, many new emerging devices for computing have been developed, and participants stressed the need for continued analysis, characterization, and modeling of new devices – not as standalone components but within the computing environment - to determine their usefulness for various applications. This effort will require new tools and validation methods. We also discussed advances in 2.5D/3D integration to produce CMOS+X hardware systems while remaining mindful of the new challenges that integration generates in electrical conduction, heat conduction, photon transmission, electron-photon transduction, and other phenomena that impact computing metrics.

### *Workshop IV, 2024*

For the 2024 workshop, we again invited speakers from a variety of programs that engage in AI/ML and microelectronics R&D, including speakers from Energy Frontier Research Centers, the National Laboratory Microelectronics Codesign program, and the Department of Defense Microelectronics Commons Hubs program. The topics covered next-generation microelectronics, advanced packaging and heterogeneous integration, neural-inspired & emerging computing technologies, trust in AI, and electronic design automation. We also hosted a panel on workforce development for the microelectronics industry, and we noted that the skillsets required for the workforce are varied and span many disciplines, and that all educational levels from vocational certificates to associate degrees to doctorates are needed. The panelists who were comprised of leaders from federal government agencies to two-year community colleges to non-profits, emphasized the importance of coordination in our education system to ensure that K-12 schools are preparing students for all levels of educational attainment, and working hand-in-hand with industry to provide the knowledge and skills needed to succeed. What follows below are the abstracts from speakers at the 2024 workshop.

## *References*

[1] https://www.sandia.gov/ccr/publications/details/ai-enhanced-co-design-for-next-generation-microelectronics-innovating-innov-2021-04-06/

[2] https://coinflipscomputing.org/

[3] https://remind.engr.tamu.edu/

# Session I: Next-generation and emerging microelectronics I

## NorthPole Chip: Neural Inference at the Frontier of Energy, Space, and Time

Filipp Akopyan
Principal Research Scientist | Chip Design Team Lead | International Business Machines (IBM)

### *Key Takeaways*

- NorthPole is a low-precision, massively parallel, densely interconnected, energy-efficient, and spatial computing architecture with a co-optimized, high-utilization programming model.
- NorthPole chip outperforms all prevalent architectures, even those that use more-advanced technology processes.

### *Summary*

Computing, since its inception, has been processor-centric, with memory separated from compute. Inspired by the organic brain and optimized for inorganic silicon, NorthPole is a neural inference architecture that blurs this boundary by eliminating off-chip memory, intertwining compute with memory on-chip, and appearing externally as an active memory chip. NorthPole is a low-precision, massively parallel, densely interconnected, energy-efficient, and spatial computing architecture with a co-optimized, high-utilization programming model. On the ResNet50 benchmark image classification network, relative to a graphics processing unit (GPU) that uses a comparable 12-nanometer technology process, NorthPole achieves a 25 times higher energy metric of frames per second (FPS) per watt, a 5 times higher space metric of FPS per transistor, and a 22 times lower time metric of latency. Similar results are reported for the Yolo-v4 detection network. NorthPole outperforms all prevalent architectures, even those that use more-advanced technology processes.

The NorthPole axiomatic architecture and implementation is shown in Figure 1. Shown in the figure are the following: (**A** and **B**) Core architecture schematic (A) and corresponding core circuit physical layout (B): Compute (red) includes the vector-matrix multiplier (VMM) and vector compute, which is integrated with memory (blue), and is driven by eight control threads (yellow). VMM supports 8-, 4-, and 2-bit precisions (axiom 2). Memories for weights, programs, and neuron activations are shared and unified. The unified memory is placed near the VMM and vector compute units (axiom 4). The physical placement of the VMM, vector units, and control units is intertwined, respectively, with weight buffer, partial sum buffer, and program buffer memories (axiom 4). At each cycle on each core, the VMM multiplies an input activation vector from the unified memory with a weight matrix from the weight buffer and transmits the output to the vector units on adjacent cores and, eventually transmit the output to an activation function unit that, finally, writes the output activations back into the unified memory. The control unit has eight concurrent threads for orchestrating compute and communication (axiom 7). (**C** and **D**) Core NoC schematic (C) and core wire physical layout (D): A 512-wire partial sum NoC (green) transfers values between vector compute units in adjacent cores, enabling spatial computing (axiom 5); a 256-wire activation NoC (orange) transfers

input frames to the cores from the framebuffer, transfers output frames from the cores to the framebuffer, and enables shuffling neuron activations among the unified memory of cores (axiom 5); a 1024-wire model NoC (magenta) carries distributed synaptic weights to the weight buffer (axiom 6); and a 256-wire instruction NoC (teal) delivers distributed instructions to the program buffer (axiom 6). (**E**) Schematic of spatially tiled by two-by-two core array illustrating core-to-core communication using NoCs. (**F**) Die photograph of the manufactured, fully functional NorthPole chip with a 25-mm-by-31.8-mm footprint, illustrating periphery (input and output, queuing, and scheduling; axiom 10), framebuffer memory (for input and output tensor storage; axiom 10), and a 16-by-16 core array (axiom 3) where each core is 1.53-mm-by-165-mm and has 768 kilobytes of memory per core (axiom 4).

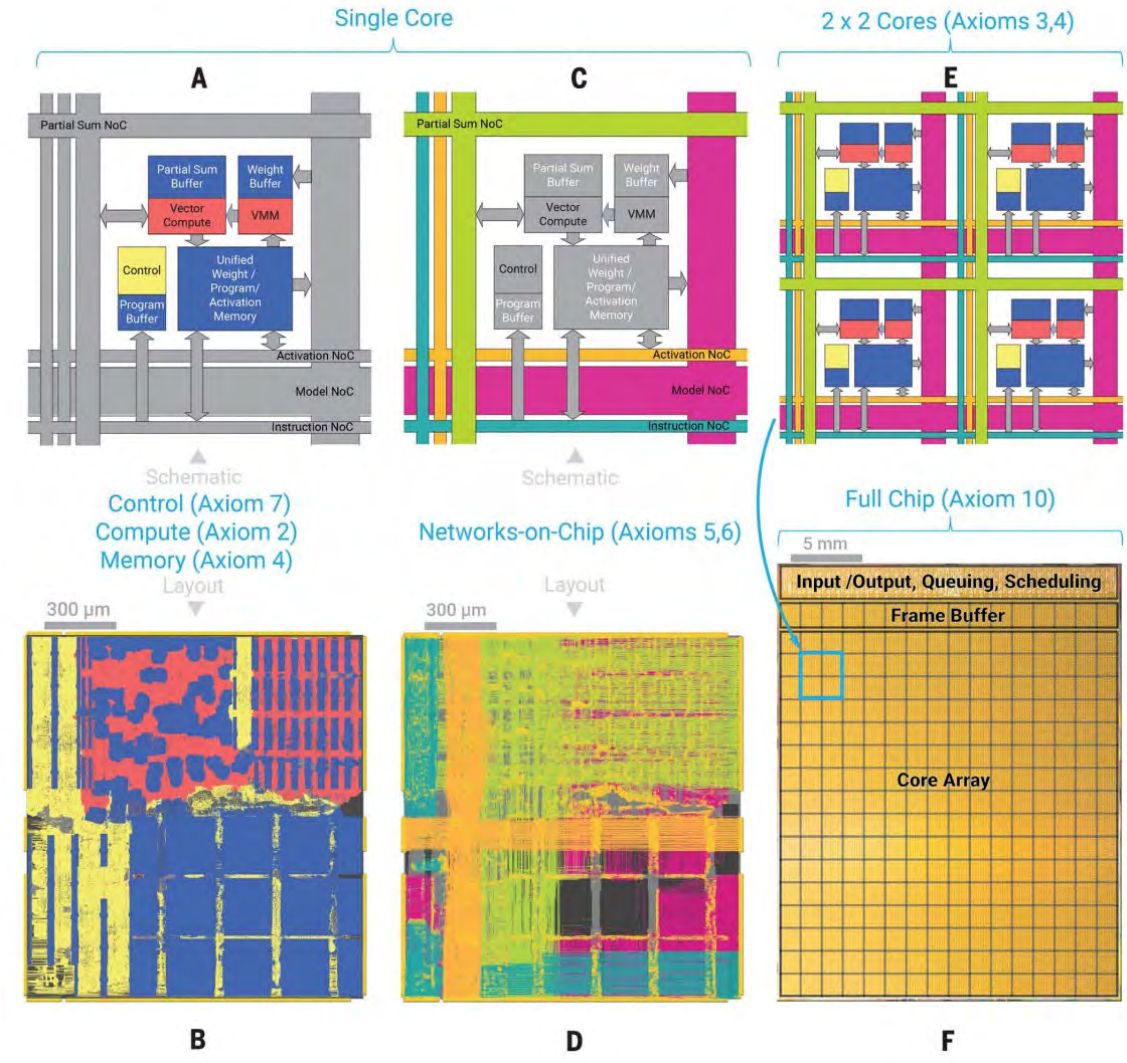



*Figure 1. NorthPole axiomatic architecture and implementation (see text for description.)*

## *Conclusion*

As seen from the inside of the chip, at the level of individual cores, NorthPole appears as memory near compute; as seen from the outside of the chip, at the level of input-output, it appears as an active memory. NorthPole is an architectural innovation at the intersection of brain-inspired computing and semiconductor technology, which defines a frontier that promises to expand. Specifically, the NorthPole trajectory promises to evolve along dimensions of algorithms (natively optimized for NorthPole), systems (scale-out with multiple boards, scale-up with multiple chips per board, and direct sensor integration), modularity (chiplets with varying core array sizes from tiny to large), packaging (heterogeneous chiplet integration), architecture (tileability, three-dimensional integration), silicon scaling, silicon optimization, and, potentially, post-silicon technologies. Interestingly, the EDVAC report began with neurons and synapses, and now the rising importance of neural inference applications has brought the field full circle.

## *References*

[1] D.S. Modha, F. Akopyan, et al. “Neural Inference at the Frontier of Energy, Space, and Time”. Science Journal. 19 Oct 2023. Vol 382, Issue 6668, pp. 329-335.
https://www.science.org/doi/10.1126/science.adh1174

[2] https://modha.org/

[3] https://www.linkedin.com/in/filipp-akopyan-1908901/

## *Acknowledgements*

The authors would like to thank many collaborators: I. Ahmed, T. Amberg, N. Antoine, F. Baez, H. Baier, L. Berge, R. Bhimarao, J. Bjorgaard, M. Boraas, B. Brezzo, M. Butterbaugh, E. J. Campbell, E. Colgan, M. Criscolo, C. di Nolfo, M. Doyle, M. Grassi, C. Guido, D. Hitchings, R. Hoke, K. Holland, S. Hui, T. Jones, S. Kedambadi, P. Kourdis, A. LaPiana, G. LaPiana, S. Lekuch, Z. Liu, R. Loboprabhu, A. Lonkar, D. Ma, G. Maass, S. Machha, V. Mahadevan, P. Mann, M. Mastro, K. O’Connell, B. Parikh, H. Penner, R. Purdy, M. Rajwadkar, L. Rapp, K. Rice, C. Sassano, K. Schoneck, R. Scouller, J. Shaffer, D. Shukla, A. Sigler, C. Smallwood, N. Subbiah, S. Tian, T. Tidwell, D. Turnbull, M. Uman, G. Van Leeuwen, S. Vispute, I. Vo, J. Wang, C. Werner, S. Wundavilli, and C. N. Xiong. We also thank H. Chandran, J. Saxena, J. Lee, A. Prashantha, S. Rao, and the Anora Labs.

# Machine Learning Enabled Co-Design of Microelectronic Systems

Madhavan Swaminathan
Department Head of Electrical Engineering & William E. Leonhard Endowed Chair Director, Center for Heterogenous Integration of Micro Electronic systems (CHIMES) | The Pennsylvania State University

## *Summary*

As Heterogeneous Integration (HI) is beginning to take center stage as a replacement to Monolithic Integration, the role of automated tools that accelerate the design process is becoming increasingly important. The requirement for co-design is now a norm and a necessity, where capturing the interactions between the various multi-physics and signaling domains is becoming very critical. EDA tools have come a long way but are still limited in their ability to support design space exploration, optimization, and robust design with statistical relevance. In this presentation, progress made in Machine Learning specific to HI in the context of advanced packaging will be discussed. Approaches used such as convolutional neural networks, Bayesian learning, reinforcement learning, transfer learning, forward and inverse design applied to a suite of packaging problems related to digital/RF signaling, power delivery networks, beamforming and others will be discussed.

# Data-Driven Synthesis and Characterization for Accelerated Electronic Materials Discovery

Christopher Hinkle
Professor, Department of Electrical Engineering | University of Notre Dame

## *Key Takeaways*

- Machine learning-aided analysis of materials characterization data improves accuracy, reduces bias, and enables active learning experiment design.
- Cross-correlations in different data streams are demonstrated as a way to reduce time and resource expenditures in experimental materials and device research.
- These techniques accelerate materials synthesis and optimization and, along with new tools, can enable precision synthesis with real-time monitoring.

## *Summary*

In this presentation, we will discuss our recent work using image segmentation and neural networks to analyze materials characterization data of synthesized electronic materials. We will describe the implementation of domain expertise into the models to enable more accurate and accelerated analysis of diffraction, XPS, ellipsometry, and transport data. We will also talk about cross-correlations in these multi-modal data streams that enables the elimination of some experimental procedures, reducing time and money during materials discovery and/or optimization. Using these data streams to inform active learning experiment design can lead to orders of magnitude acceleration in materials synthesis. Finally, we will discuss the prospects of embedding new characterization streams into deposition tools for real-time monitoring to enable precision synthesis.

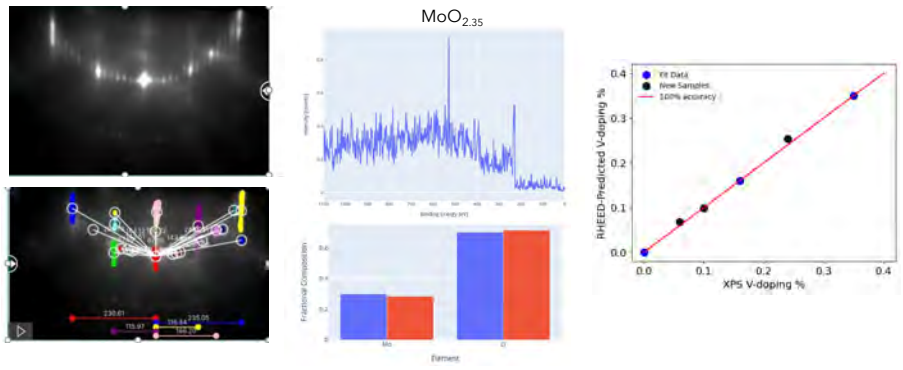



*Figure 1. (Left) Reflection high-energy electron diffraction (RHEED) data along with an example of the segmented and quantified data following deep learning analysis. (center) Accelerated X-ray photoelectron spectroscopy (XPS) quantification using machine learning. (right) Example of cross-correlation in the data, helping to reduce experiment time.*

## *Acknowledgements*

This work was supported in part by the National Science Foundation (NSF) through the Division of Materials Research (DMR) Award No. 2324172. This work was also supported in part by SUPREME, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

# Discussion – Next-Generation and Emerging Microelectronics I

Chris Allemang, Moderator
Sandia National Laboratories

We started this session with a look at computer systems, specifically the IBM NorhPole chip with a talk by Filipp Akopyan. Using the ResNet50 benchmark, the NorthPole chip outperforms all prevalent architectures even with more-advanced nodes demonstrating the benefit of a low-precision, massively parallel, densely interconnected, energy-efficient, and spatial computing architecture with a co-optimized, high-utilization programming model. Next, Madhavan Swaminathan gave a talk on complex microelectronic systems made through heterogenous integration (HI). His talk focused on the role of machine learning in tools for HI to solve packaging problems relating to signaling, power delivery, and beamforming. Finally, Christopher Hinkle shifted the focus from systems to materials and devices. The focus of his talk was on machine learning analysis of materials characterization to efficiently perform material and device research.

Summary of the guided and moderated discussion:

- What is co-design for next generation and emerging microelectronics?
  - Chris Hinkle – from the material and device perspective it is talking to the circuit designers and system architects to understand where value of materials and devices are to appropriately focus research.
  - Filipp Akopyan – hardware and software co-design, they cannot be designed independently and optimization requires co-design.
  - Summary: talk to people you do not normally work with and figure out what is important for things to work together efficiently by deciding where value is and optimization opportunities.
- What are the challenges and opportunities for AI-enhancement for this co-design?
  - Chris Hinkle – training data is extremely limited and non-diverse. Need larger data sets.
    - We even need “bad data” published and to implement active learning strategies.
  - Filipp Akopyan – shares sentiment on lack of data, good data, labeled data. Need to able to predict future so you do not miss out on the useful stuff.
  - Summary: need more data!
- What are the opportunities to improve on biological design?
  - Filipp Akopyan – use CPUs for precise mathematical operations. Use a specific tool for its own job and let the AI accelerators augment CPUs. GPUs are the right architecture for training.
- What are the opportunities for emergent devices to augment silicon?
  - Filipp Akopyan – augmentation comes from architecture but transistors that are energy efficient, switch fast, and are reliable are needed.
  - Madhavan Swaminathan – need a platform for integration and is where HI comes in. Packaging technologies needs to be available when emerging device technology is available.
  - Chris Hinkle – misplaced trust in emerging devices. They are very far away. Interconnect is a place where earlier wins may be available for new materials.

# Session II: Advanced packaging and heterogeneous integration

## AI for Testing AI Hardware: A Virtuous Cycle

Krishnendu Chakrabarty
Fulton Professor of Microelectronics, School of Electrical, Computer and Energy Engineering | Arizona State University

### *Summary*

The ubiquitous application of deep neural networks (DNN) has led to a rise in demand for AI accelerators. For example, the Tensor Processing Unit from Google and its variants are of considerable interest for DNN inferencing using AI accelerators. DNN-specific functional criticality analysis identifies faults that cause measurable and significant deviations from acceptable requirements such as the inferencing accuracy. This talk will show how AI can be used for classifying the criticality of structural faults in the processing elements (PEs) of systolic-array accelerators. The speaker will first present a two-tier machine learning-based method to assess the functional criticality of faults. The problem of minimizing misclassification will be addressed by utilizing generative adversarial networks. While supervised learning techniques can be used to accurately estimate fault criticality, it requires a considerable amount of ground truth for model training. To address this problem, the speaker will present a neural-twin framework for analyzing fault criticality with a negligible amount of ground-truth data. A misclassification-driven training algorithm will be used to sensitize and identify biases that are critical to the functioning of the accelerator for a given application workload. The proposed framework achieves up to 100% accuracy in fault-criticality classification in 16-bit and 32-bit PEs by using the criticality knowledge of only 2% of the total faults in a PE. Time permitting, the speaker will describe a black-box optimization method to generate functional test patterns for AI inferencing accelerators.

# 2.5D/3D Heterogeneous Integration Co-Design for In-Pixel Processing, LLM Acceleration and Bioinformatics

Shimeng Yu
Professor, School of Electrical and Computer Engineering | Georgia Institute of Technology

## *Key Takeaways*

- *Key takeaway 1—2.5D/3D heterogeneous integration enables versatile sensing, compute and memory/storage dies for improved energy efficiency for AI workloads.*
- *Key takeaway 2—Electro-thermal co-simulation is important for thermal management in 2.5D/3D integrated system.*
- *Key takeaway 3—System-technology co-optimization is the key for optimally partitioning the large AI model for smaller chiplets.*
- *Key takeaway 4—In-memory computing is a compelling paradigm for saving the data movement cost in 2.5D/3D integrated system.*

## *Summary*

This presentation will discuss the recent progresses in system-technology co-design of 2.5D/3D heterogeneous integration on silicon interposer for representative applications including the following 1) in-pixel processing in the context of autonomous driving, that adds the frontend feature extraction capability to the CMOS image sensor for reducing the data volume being transferred for the backend accelerator (a vision transformer); 2) large language model (LLM) acceleration, that decomposes the model into a sea of 3D cubes where each cube consists of versatile accelerator chiplets including analog compute-in-memory, digital compute-in-memory, tensor processing unit, etc.; 3) Genome sequencing based on large volume of the bioinformatic dataset, where a 3D NAND Flash is modified as associative content-addressable memory for pattern matching and is proposed to be bonded with a logic die for hyperdimensional computing.

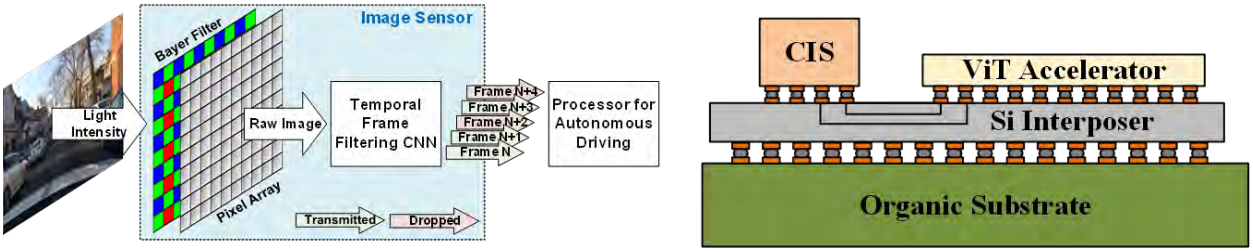



*Fig. 1 In-pixel processing*

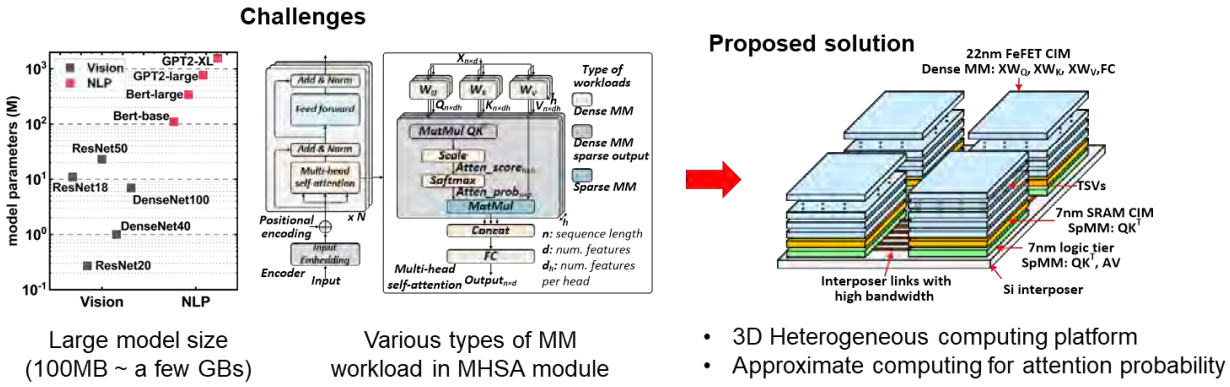



*Fig. 2 LLM acceleration*

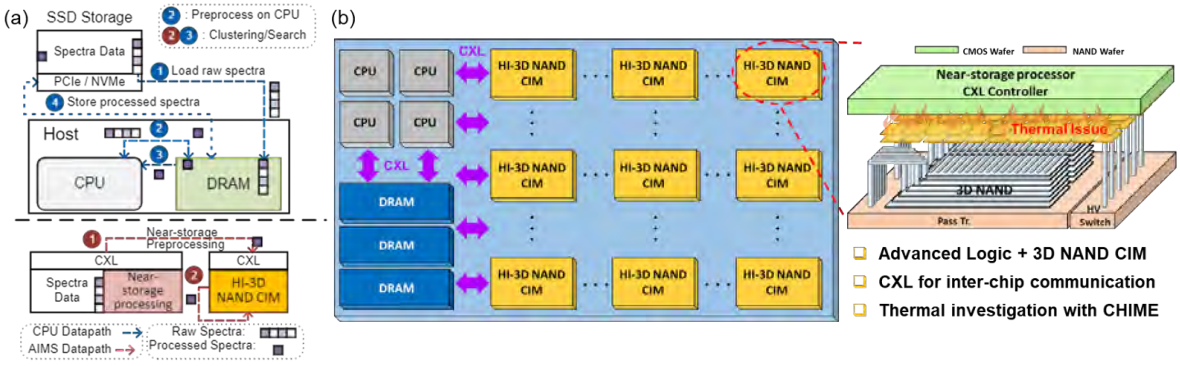



*Fig. 3 Genome sequencing*

## *Conclusion*

This presentation shows the potentials of large-scale 2.5D/3D system integration.

## *References*

[1] J. Sharda, W. Li, Q. Wu, S. Chang, S. Yu, "Temporal frame filtering for autonomous driving using 3D-stacked global shutter CIS with IWO buffer memory and near-pixel compute," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 70, no. 5, pp. 2074-2084, 2023.

[2] J. Sharda, M. Manley, A. Kaul, W. Li, M. S. Bakir, S. Yu, "Design and thermal analysis of 2.5D and 3D integrated system of a CMOS image sensor and a sparsity-aware accelerator for autonomous driving," IEEE Journal of the Electron Devices Society, accepted.

[3] Y. Luo, S. Yu, "H3D-Transformer: A heterogeneous 3D (H3D) computing platform for transformer model acceleration on edge devices", ACM Transactions on Design Automation of Electronic Systems (TODAES), accepted.

[4] P.-K. Hsu, V. Garg, S. Yu, "A heterogeneous platform for 3D NAND-based in-memory hyperdimensional computing engine for genome sequencing applications," IEEE Transactions on Circuits and Systems I: Regular Papers, accepted.

[5] P.-K. Hsu, W. Xu, T. Rosing, S. Yu, "An in-storage processing architecture with 3D NAND heterogeneous integration for spectra open modification search," ACM/IEEE International Symposium on Memory Systems (MEMSYS) 2023.

## *Acknowledgements*

This research is supported in part by DARPA IP2 program, PRISM and CHIMES, two of SRC/DARPA JUMP 2.0 centers.

# Co-Design Challenges in Modern Heterogeneously Integrated Microsystems

Christopher D. Nordquist
RF Microsystems Department, Microsystems Science Engineering & Applications (MESA) Center
Sandia National Laboratories

## *Key Takeaways*

- Heterogeneous integration entails the incorporation of diverse technologies from different sources, with each technology having a different simulation and modeling heritage.
- Co-design amongst the different technologies in heterogeneous microsystems requires accurate modeling and simulation of interactions across multiple domains including semiconductor physics, electrical signals, RF & signal integrity, mechanical stress & strain, materials, thermal transport & management, and system engineering.
- There are currently no commercial or custom solutions that provide insight across all of the domains and scales required to enable first-pass success for future complex HI microsystems, presenting rich opportunities for AI-enabled co-design.

## *Summary*

Successful realization of future microsystems via Heterogeneous Integration (HI) requires co-design across multiple technologies, physics, scales, and domains. HI, defined as *"…the integration of separately manufactured components into a higher-level assembly that, in the aggregate, provides enhanced functionality and improved operating characteristics*" [1], has been established in high-volume commercial products to enable design disaggregation and technology mixing for improving yield, lowering cost, and increasing performance [2,3]. The investments and coordination required for these commercial successes are supported by incredibly large production volumes and coordination across technologies, vendors, and processes. In order to maintain access to state-of-the-art microelectronics, the research and development (R&D) community must leverage these developments and supply chain.

For government and research, HI provides a potential route towards intimate affordable integration for creating prototype microsystems [4,5]. However, typical research and development (R&D) activities often have low overall volumes, limited budgets, and short timescales. These constraints limit the level of investment and technology customization allowed, encouraging the leveraging of external technologies through multi-project wafers and existing integrated circuits where possible. HI provides an opportunity to merge these different technologies with emerging research technologies to achieve form factors and performance beyond that obtainable with traditional packaging and printed circuit board (PCB) approaches.

As an example of an HI prototype, Figure 1 shows a Sandia optical receiver that includes five different die of three different technologies on a glass interposer. The die are attached to the glass interposer with a

combination of gold thermocompression bonding and solder balls. All of the die were fabricated at Sandia, but the process has also been demonstrated using externally fabricated die. This prototype achieves order-of-magnitude miniaturization relative to a PCB-based implementation of the same function. This prototype illustrates the need for co-design: most of the integration design and verification was performed manually because the three different die technologies and the interposer were designed by four different teams using four different design procedures.

This example illustrates how multiple sources and technologies in future HI microsystems present several co-design challenges beyond those of working on a single chip or board-level integration. These challenges are associated with the integration density, variety of technology options, intimate interfaces, and long fabrication times for HI microsystems. The following bullet list provides a brief summary of these challenges.

- Chips from different sources are modeled and designed in different design tools that are not compatible. Even if the tools are compatible, users and integrators do not have full design visibility into commercial integrated circuit models. Additionally, even tools from a single company are often not compatible across different domains. This limits the ability to co-simulate multiple chips from different vendors to produce a full digital twin of an electronic system.
- The intimate integration of multiple chips and technologies exacerbates multi-physics interactions. Modern microsystems require understanding and modeling amongst semiconductor physics, electrical signals, RF & high-speed signal integrity, power distribution, mechanical stress & strain, materials, and thermal transport & management. In previous electronic systems these different domains were handled separately using specific tools, but the intimate nature of modern microsystems requires capturing interactions amongst different physics on simulation timescales.
- HI enables the rapid incorporation of new device technologies, but often the models of these new technologies are incomplete or primitive. Co-design must accommodate these immature technologies into microsystems including more mature and established technologies.
- Different technology teams have different cultures and nomenclatures that result in different expectations and assumptions. Proper co-design must uncover and accommodate these differences to allow for effective interactions amongst technologies.
- System engineering uses “interfaces amongst black boxes” approaches that rely on requirements flowdown and interface definitions but do incorporate comprehensive outcomes and results from engineering and multi-physics simulations or measured data. This abstraction of interfaces provides for standardization at the cost of efficiency and modeling accuracy.
- Close placement of integrated circuits with different functions has the potential to cause near-field interference or crosstalk between different chips and functions in an HI microsystem. There are currently no mechanisms to simulate and capture this type of near-field interference and its impact on the behavior of microsystems.

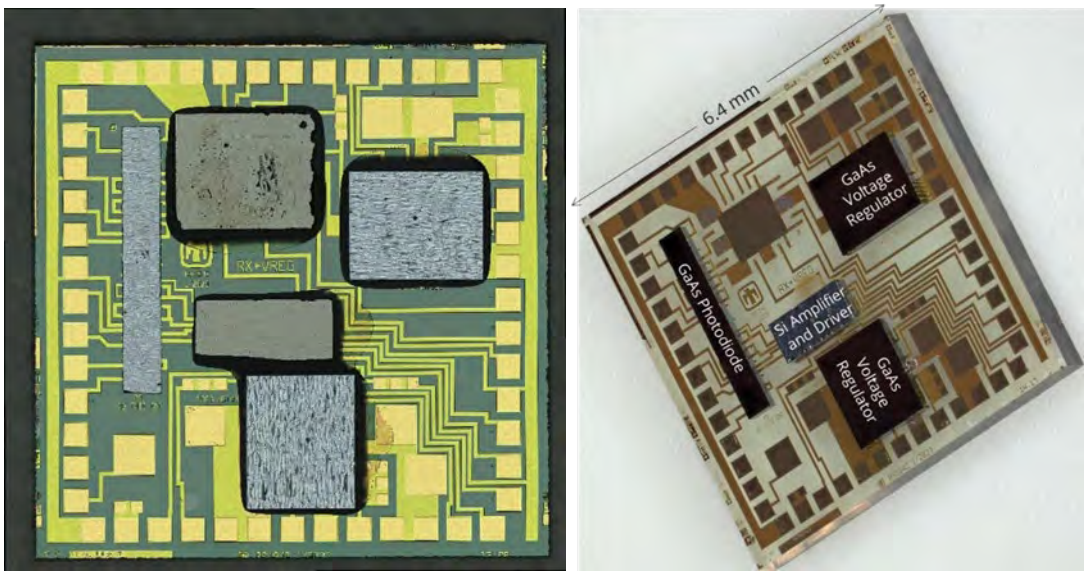



*Figure 1: Images of a Sandia HI optical receiver prototype that includes Si CMOS, GaAs HBT, and optoelectronic components assembled onto a glass interposer.*

## *Conclusion*

Heterogeneous integration presents the potential for order-of-magnitude improvements in size and performance for future microsystems, but requires co-design that crosses between the various technologies, interfaces, and physics to reach its true promise. This broad design space provides rich future opportunities in the area of AI-enabled co-design for enhancing design fidelity, reducing cycle time, and characterizing heterogeneously integrated microsystems.

## *References*

[1] https://eps.ieee.org/technology/heterogeneous-integration-roadmap.html

[2] S. S. Iyer, "Heterogeneous Integration for Performance and Scaling," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, no. 7, pp. 973-982, July 2016, doi: 10.1109/TCPMT.2015.2511626.

[3] J. Lau, "Recent Advances and Trends in Chiplet Design and Heterogeneous Integration Packaging," J. Electronic Packaging, vol. 146, no. 1, 010801 (31 pages), March 2024, doi: 10.1115/1.4062529.

[4] D. S. Green, C. L. Dohrman, J. Demmin, Y. Zheng and T. -H. Chang, "A Revolution on the Horizon from DARPA: Heterogeneous Integration for Revolutionary Microwave\/Millimeter-Wave Circuits at DARPA:

Progress and Future Directions," in *IEEE Microwave Magazine*, vol. 18, no. 2, pp. 44-59, March-April 2017, doi: 10.1109/MMM.2016.2635811.

[5] C. D. Nordquist, et al., "Heterogeneous Integration of Silicon Electronics and Compound Semiconductor Optoelectronics for Miniature RF Photonic Transceivers," *ECS Transactions*, vol. 98, no. 6, pp. 93-106, Sept 2020.

## *Acknowledgements*

Heterogeneous Integration requires the participation of a broad range of teams and individuals. In this case, Sandia's MESA Microfabrication (Matt Jordan, Mieko Hirabayashi, Patrick Finnegan) and packaging (Ray Haltli, Tipp Jennings) teams, Sandia's HBT (Rafmag Cabrera and Collin Burt) and RFIC (Travis Forbes, Justine Saugen, Stefan Lepkowski) teams, and Sandia's optoelectronic (Michael Wood, Erik Skogen) teams have provided key technical contributions while MESA management (Gideon Robertson, Gary Patrizi, Drew Hollowell, Ken Dean, Keith Tracey) have provided support and vision. Additional individual contributors have been too numerous to list here.

*Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.*

*This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.*

# Discussion - Advanced Packaging and Heterogeneous Integration

Cale Crowder, Moderator
Sandia National Laboratories

## Session Summary

Heterogeneous integration (HI) offers the opportunity to innovate by connecting multiple technologies together. However, there are several challenges that are preventing HI from gaining additional traction. Krishnendu Chakrabarty talked about challenges in predicting critical faults in AI accelerator hardware and about how AI digital twins can be used accelerate the development of AI models for predicting critical faults. Shimeng Yu spoke about applications of 3D HI and about the importance of modeling thermal effects and electrical parasitics when building 3D HI systems. Chris Nordquist spoke about the challenges that HI practitioners face, including the need for cheap, quick, low-volume manufacturing; difficulties in running multi-physics simulations; the lack of standardization across manufacturers; and the large HI design space. The priorities for progress in HI that AI can potentially accelerate include multi-physics (particularly electrothermomechanical) modeling and the co-design of sub-systems for power distribution and thermal management.

Main points:

- Challenges for HI
  - Testing, error-proofing, defect screening
  - Implementation of HI when vendors lack standardization
  - Fast turnaround, cheap, and at low volume
  - Multi-domain physics
    - Theory doesn't always work at the domain interfaces
  - Power delivery, especially for 3D
    - Need power conversion next to load to prevent IR drop
  - Trust & insurance
    - How do we know that we can trust chiplets?
  - Partitioning problem solving between academia, industry, and government
- Biggest priorities
  - Thermal modeling (because everyone wants more power)
  - Power distribution
  - Development of standards to facilitate HI
- Opportunities for AI in HI
  - Big-data approaches for identifying outlier system behavior, even when it passes tests
  - Accelerating and improving multi-physics simulations
  - AI digital twins for microelectronics, including for fault prediction
  - Multi-physics (including electrical, thermal, optical, and mechanical) modeling

- Navigating the HI co-design space
  - Types of semiconductors
  - Types of devices (e.g., electrical vs. optical)
  - Commercial off the shelf vs. custom designs
  - Types of solder joints

# Session III: Neural-inspired and other emerging microelectronics and computing

## Efficient Implementation of High-Order Ising Machines

Dmitri Strukov
Professor, Electrical and Computer Engineering | UC Santa Barbara

### *Key Takeaways*

- Key takeaway 1— We developed a hardware approach for massively parallel in-memory computing of high-degree polynomial function gradients. Such operations are essential in many biologically and physics-inspired computing concepts such as Ising machines, Boltzmann machines, and Hopfield neural networks.
- Key takeaway 2— The developed in-memory gradient computing idea is extended and generalized into a unified framework for implementing high-order Ising machines. We discuss how this framework enables orders of magnitude advantages in density and performance when applying Ising machines to solving optimization problems.
- Key takeaway 3—We developed field-programmable architectures for Ising Machine solvers that take advantage of the inherent sparsity of the hard (constraint satisfaction) optimization problems. Our approach could lead to further >10x improvements in density or speed compared to baseline implementations.

### *Summary*

Ising Machines and related approaches such as Hopfield Neural Networks and Boltzmann Machines are promising computing concepts for solving optimization problems. High-order Ising machines are expected to be most useful since many practical optimization problems result in higher-degree polynomial objective functions. However, efficient hardware implementations of such solvers are lacking, partly due to challenges in computing partial derivatives of high-degree functions of the underlying optimization problems in an efficient and reconfigurable manner. For example, a typical approach is to convert a high-order problem to an equivalent quadratic one and then solve it with a conventional second-order Ising machine. However, such an approach comes with high complexity (i.e., many auxiliary variables) and latency (slower convergence due to spurious minima) overheads.

This talk focuses on very recent work by my research group and our DARPA QuICC program collaborators in addressing these challenges [1-3]. It consists of three connected parts. First, I will discuss a novel approach for massively parallel gradient calculations of high-degree polynomials, which is conducive to efficient mixed-signal in-memory computing circuit implementations and whose area complexity scales linearly with the number of variables and terms in the function and, most importantly, independent of its degree (Fig. 1a). To validate this approach, we experimentally demonstrated solving a small-scale third-order Boolean satisfiability problem based on integrated metal-oxide memristor crossbar circuits, one of

the most prospective in-memory computing device technologies, with a competitive heuristics algorithm (Fig. 1b,c). Second, I will discuss how the developed in-memory gradient computing idea can be extended and generalized into a unified framework to implement high-order Ising machines efficiently. Two major types - monomial and derivative-based Ising machines were developed, and their optimal matching to problem encoding was investigated. Simulation results for larger-scale, constrained satisfaction problems showed orders of magnitude improvements in the density and related advantages in speed and energy efficiency compared to the state-of-the-art. Third, I will discuss in-memory computing architecture for Ising machines inspired by island-type field programmable gate arrays (Fig. 2a,b). The architecture takes advantage of high sparsity in coupling matrixes of many practical (hard) optimization problems. By adapting open-source design automation tools, we co-optimized problem embedding and architecture parameters and modeled architecture physical performance. The developed architecture shows potential to improve density further (by at least 10x) and speed up operation compared to the baseline approaches (Fig. 2c). Finally, I will discuss how our work could enable even higher-performance systems after co-designing architectures and algorithms.

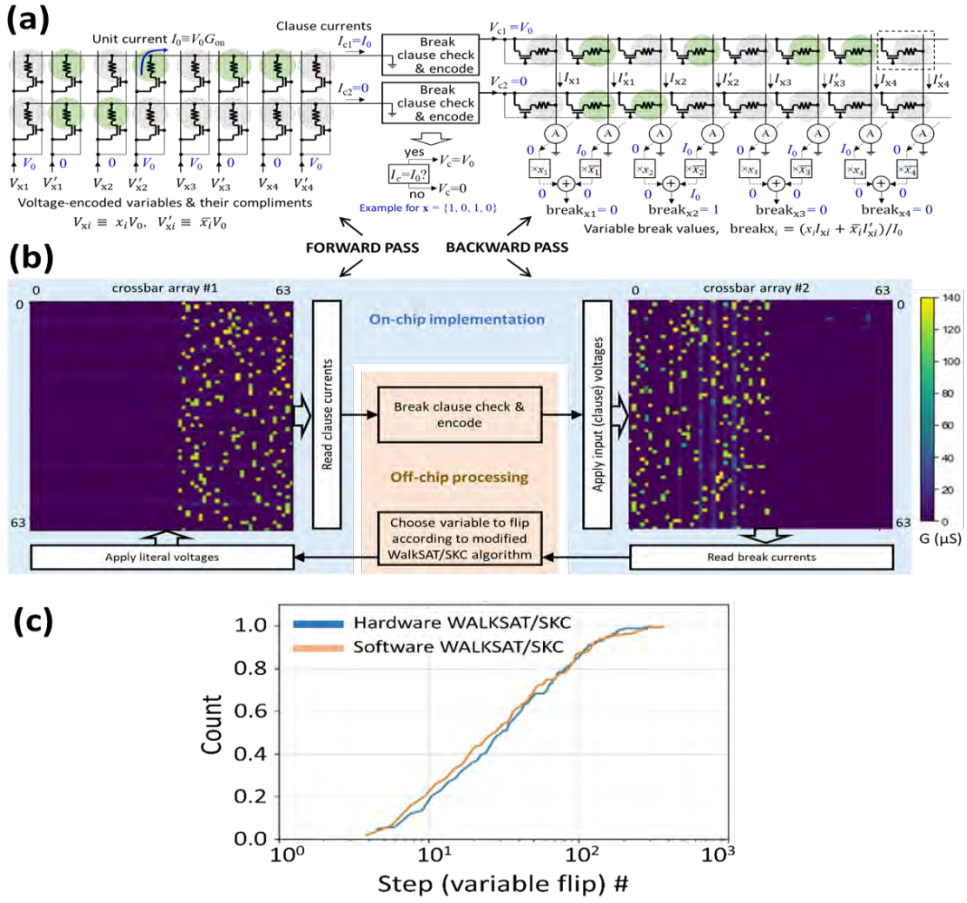



*Figure 1. (a) Equivalent circuit for 1T1R crossbar array implementation. For clarity, only portion of a crossbar array circuits are shown. (b) Experimental setup based on HPE's SuperT system comprising of two 64×64 1T1R memristor crossbars integrated with CMOS sensing, driving, and programming circuitry (c) Runtime length distribution for solving 14-variable 64-clause 3-SAT with WalkSAT/SKC heuristics.*

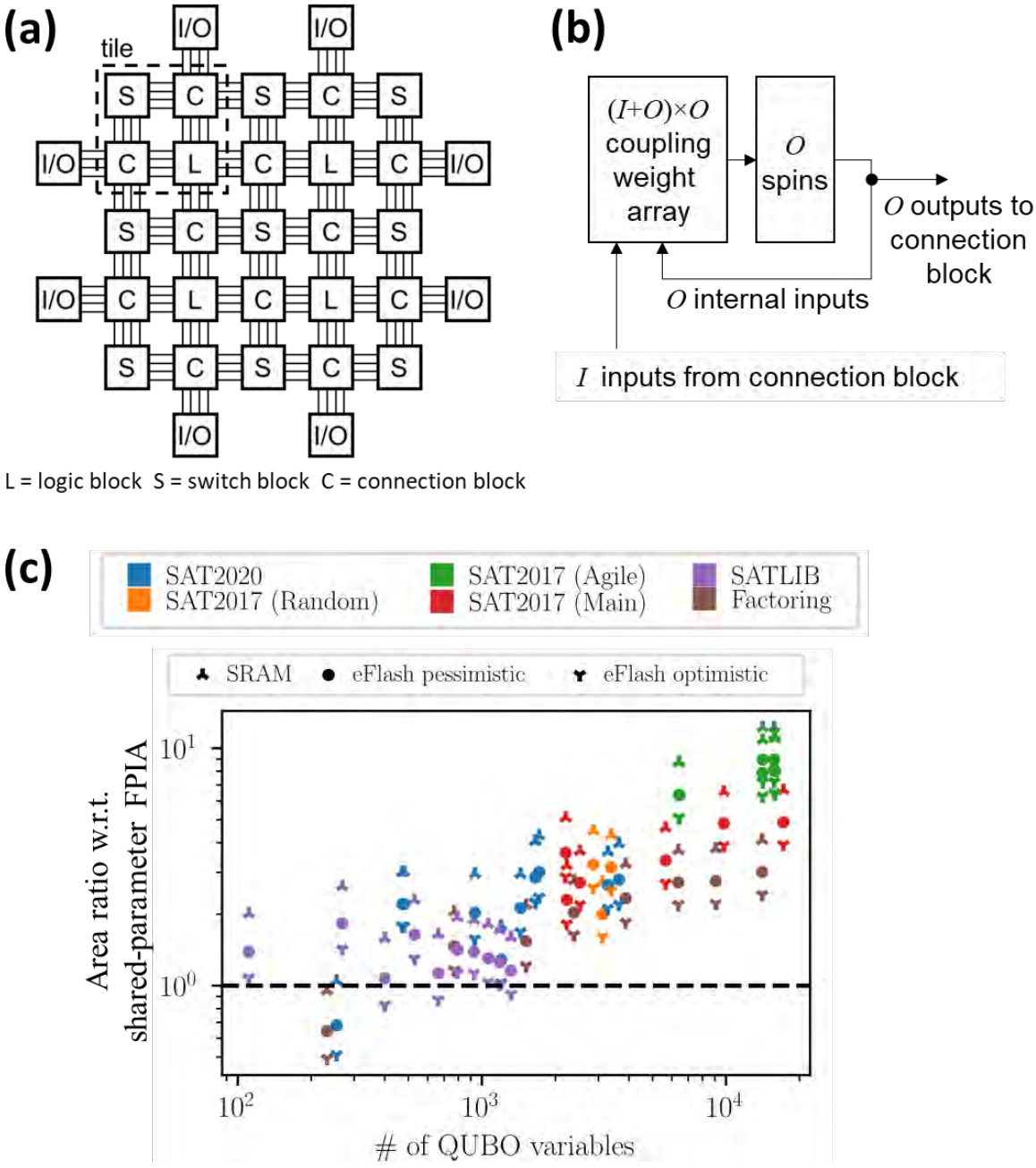



*Figure 2. (a) Top level architecture and (b) logic block for quadratic field programmable Ising array (FPIA). (c) Modeling of density advantage over baseline based on ISAAC architecture for common benchmark problems.*

## *References*

[1] M. Hizani et al. “Memristor-based hardware and algorithms for higher-order Hopfield optimization solver outperforming quadratic Ising machines”, arXiv:2311.01171v1, Nov. 2023

[2] T. Bhattacharya et al. “Computing high-degree polynomial gradients in memory”, arXiv:2401.16204, Jan 2024.

[3] G. Hutchinson et al. “FPIA: Field-programmable Ising arrays with in-memory computing”, arXiv: 2401.16202, Jan 2024.

## *Acknowledgements*

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under the Air Force Research Laboratory (AFRL) Agreement No. FA8650-23-3-7313. The views, opinions, and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. The author would like to acknowledge the contribution of the DARPA QuICC project collaborators.

# Evolutionary Optimization for Neuromorphic Co-Design

Catherine Schuman
Department of Electrical Engineering and Computer Science | University of Tennessee

## *Key Takeaways*

- Evolutionary optimization can help us to optimize device characteristics and tailor the performance of the device for a particular application
- Evolutionary optimization can be used to optimize within the characteristics and constraints of devices, dealing with issues such as limited device precision or device noise

## *Summary*

Effectively leveraging the characteristics and capabilities of neuromorphic computing systems continues to be a challenge in the field, especially as new hardware, devices, and materials are being developed and applied to neuromorphic implementations. There are several key reasons why evolutionary optimization is useful in designing spiking neural networks for neuromorphic deployment.

First, evolutionary optimization can be used in both device co-design and device understanding. In previous work, we have shown that evolutionary optimization can be used to understand the impact of different memristive materials in neuromorphic applications; in particular, we have trained spiking neural networks for a memristive neuromorphic implementation for a simple classification task, and we have compared how different materials impact energy usage (Figure 1) [1]. We have also used evolutionary optimization approaches to understand the impact of neuron design features such as leak, axonal delay, and refractory periods for superconducting optoelectronic-based neurons [2] and for fully digital neuromorphic systems [3]. We have also previously investigated the impact of synaptic features such as different forms of spike-timing dependent plasticity on application performance [1,4]. Finally, we have used evolutionary optimization to design neuromorphic networks implemented with a variety of emerging device types, including biomimetic devices [5].

A second key advantage of evolutionary optimization is that it can be used to perform multi-objective optimization. We have previously used evolutionary optimization to simultaneously maximize accuracy on an application while also minimizing network size and/or minimizing energy usage [6]. We have also used multi-objective evolutionary optimization to optimize for resilience to noise or failures in the synaptic devices [7,8].

Other advantages of using evolutionary optimization for co-design include that they can be used on a wide variety of different application types, including classification, control, and anomaly detection, they can be used with hardware in the loop to customize to the characteristics of a specific hardware implementation, and they can be used to optimize around other types of noise and constraints of the hardware implementation.

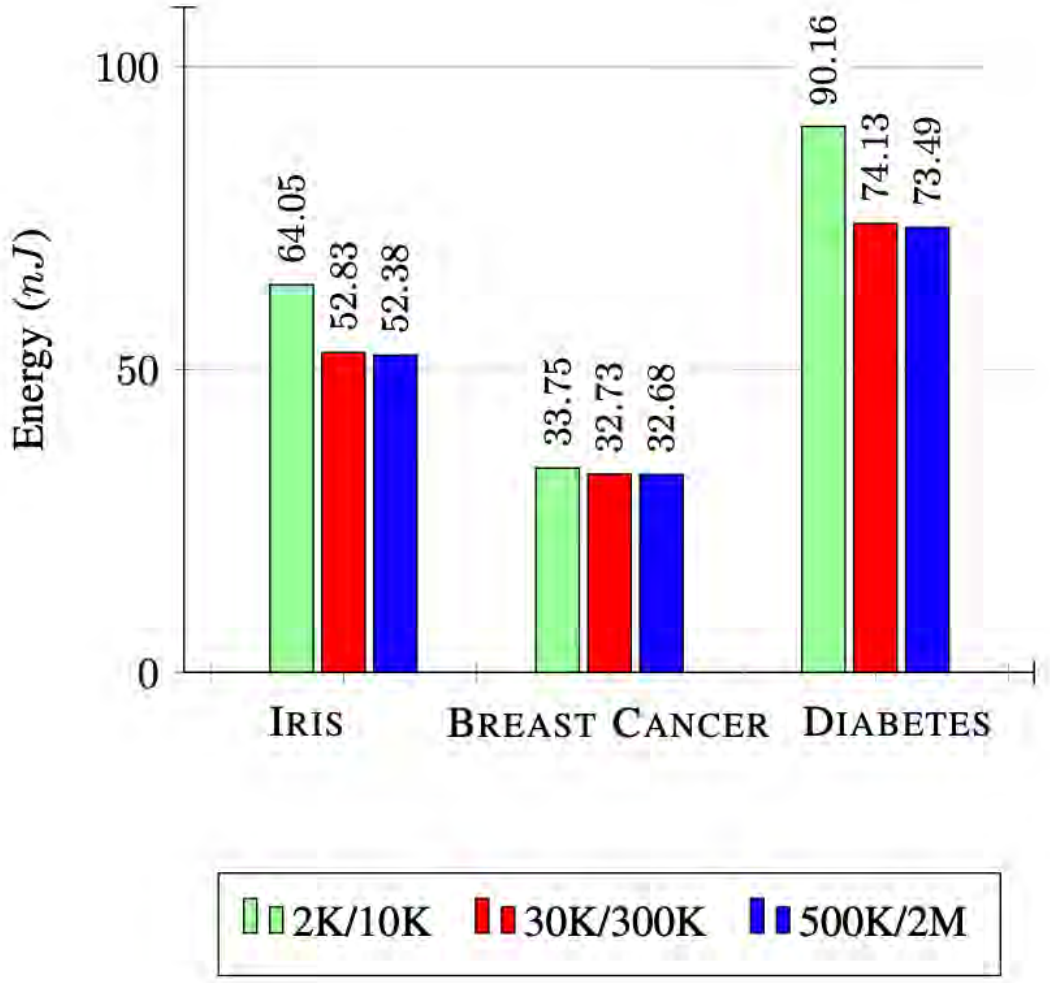



*Figure 1: Low Resistance State/High Resistance State for different metal oxide device types [1].*

## *Conclusion*

There are many key advantages for leveraging evolutionary optimization for co-design in neuromorphic systems. In future work, we would like to leverage hierarchical evolution to simultaneously optimize at multiple levels of the compute stack. For example, we would like to optimize at the device, circuit, and architecture level simultaneously, fully customizing the entire implementation to a given application.

## *References*

[1] Chakma, G., Adnan, M. M., Wyer, A. R., Weiss, R., Schuman, C. D., & Rose, G. S. (2017). Memristive mixed-signal neuromorphic systems: Energy-efficient learning at the circuit-level. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *8*(1), 125-136.

[2] Buckley, Sonia, Adam N. McCaughan, Jeff Chiles, Richard P. Mirin, Sae Woo Nam, Jeffrey M. Shainline, Grant Bruer, James S. Plank, and Catherine D. Schuman. "Design of superconducting optoelectronic networks for neuromorphic computing." In *2018 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1-7. IEEE, 2018.

[3] Schuman, Catherine, Hritom Das, James S. Plank, Ahmedullah Aziz, and Garrett Rose. "Evaluating Neuron Models through Application-Hardware Co-Design." In proceedings of the 2023 Asilomar Conference. 2023.

[4] Ameli, Shelah, Adam Foshie, Drew Friend, James Plank, Garrett Rose, and Catherine Schuman. "Algorithm and Application Impacts of Programmable Plasticity in Spiking Neuromorphic Hardware." In *Proceedings of the 2023 International Conference on Neuromorphic Systems*, pp. 1-6. 2023.

[5] Hasan, Md Sakib, Catherine D. Schuman, Joseph S. Najem, Ryan Weiss, Nicholas D. Skuda, Alex Belianinov, C. Patrick Collier, Stephen A. Sarles, and Garrett S. Rose. "Biomimetic, soft-material synapse for neuromorphic computing: from device to network." In *2018 IEEE 13th Dallas Circuits and Systems Conference (DCAS)*, pp. 1-6. IEEE, 2018.

[6] Schuman, Catherine D., J. Parker Mitchell, Robert M. Patton, Thomas E. Potok, and James S. Plank. "Evolutionary optimization for neuromorphic systems." In *Proceedings of the 2020 Annual Neuro-Inspired Computational Elements Workshop*, pp. 1-9. 2020.

[7] Dimovska, Mihaela, Travis Johnston, Catherine D. Schuman, J. Parker Mitchell, and Thomas E. Potok."Multi-objective optimization for size and resilience of spiking neural networks." In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0433-0439. IEEE, 2019.

[8] Schuman, Catherine D., J. Parker Mitchell, J. Travis Johnston, Maryam Parsa, Bill Kay, Prasanna Date, and Robert M. Patton. "Resilience and robustness of spiking neural networks for neuromorphic systems." In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-10. IEEE, 2020.

## *Acknowledgements*

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under award number DE-SC0022566.

# High Entropy Oxides for Next Generation Microelectronics

Elliot J. Fuller
Sandia National Laboratories

## *Key Takeaways*

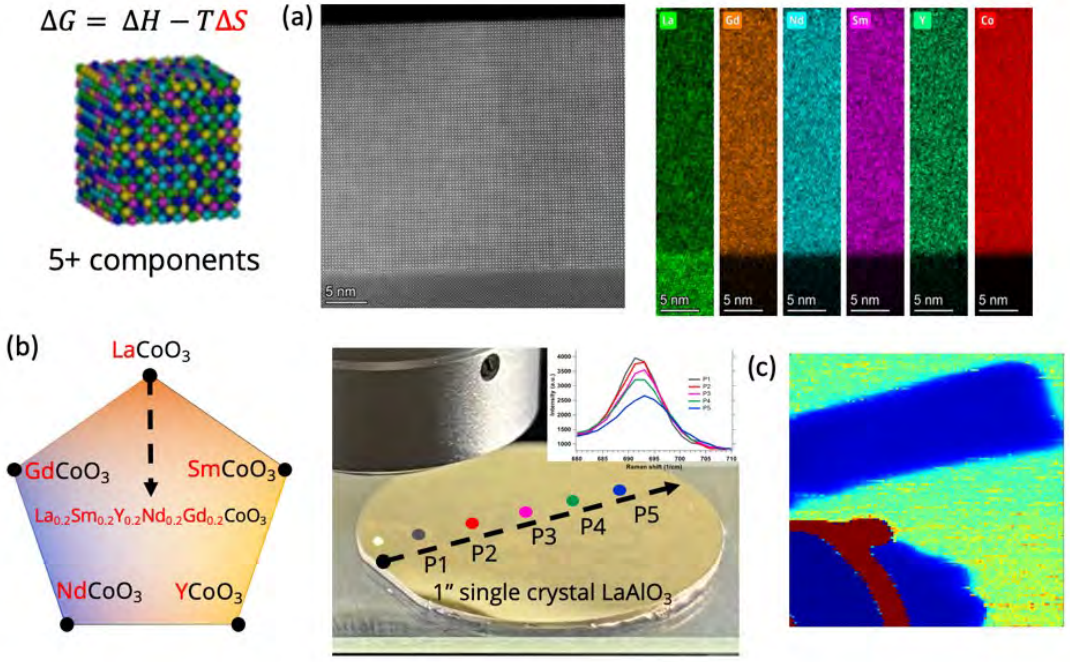
- High entropy materials provide an enormous compositional space with which to design material functionality for microelectronics applications.
- The rational design of high entropy oxides is both a major scientific challenge and an engineering opportunity.

## *Summary*

High entropy compounds are an emerging class of materials with widely tunable electrical and thermal properties that show outstanding resilience to radiation environments[1]. High entropy oxides (HEO) have potential for disruptive performance breakthroughs in microelectronics and related technologies with broad impact. For example, HEOs exhibit tunable electronic transitions for memory devices[2], the highest demonstrated energy density for ceramic capacitors for power applications[3], and widely tunable thermal conductivity for barrier coatings[4]. However, despite breakthroughs, the role of configurational entropy and disorder in HEOs remains unclear, with many compounds exhibiting emergent properties that cannot yet be predicted from first principles.

Here, I will discuss our efforts to establish 1) the role of configurational entropy in thermal and electronic properties through hypothesis driven experiments and to 2) the design of high-performance electronic devices with entropy oxides for next-generation compute paradigms. As a platform to investigate thermal and electronic properties, we have focused our efforts on perovskite systems $ABO_3$ (where A = rare earth and B = transition metal). Using combinatorial pulsed laser deposition on single crystal substrates we can generate compositional gradients to explore various material parameter spaces. This approach enables an enormous range of compositions with just a few starting materials and allows the growth of low-defect epitaxial films free of grain boundaries that obscure underlying physics.

With regards to electronic properties, for oxides competing interactions of spin, orbital and charge over closely spaced energy scales lead to emergent behavior, including metal insulator transitions and new electronic phases. Such behaviors are exquisitely sensitive to small compositional variations and disorder and therefore I will describe instances where high-entropy effects have led to surprising and counterintuitive results, such as electronic carrier-type crossover with ‘valence neutral’ disorder. Additionally, HEOs have emergent and widely tunable thermal properties. I will describe how in a perovskite system, weak A-site disorder can be utilized to increase thermal conductivity, whereas strong B-site disorder can be engineered to lower it. Finally, HEO present a unique and unexplored opportunity for devices that undergo electro-thermal phase transitions because tunable thermal properties can dictate the dynamics of such phase transformations in close feedback with electronic effects.




## *Conclusion*

Functional electronics based upon high entropy oxides will require significant experimental and modeling efforts across enormous compositional spaces but will yield rich rewards for device applications.

## *References*

[1] C. Oses, C. Toher, and S. Curtarolo, "High-entropy ceramics," *Nature Reviews Materials,* vol. 5, no. 4, pp. 295-309, 2020, doi: 10.1038/s41578-019-0170-8.

[2] M. Ahn *et al.*, "Memristors Based on (Zr, Hf, Nb, Ta, Mo, W) High-Entropy Oxides," *Advanced Electronic Materials,* vol. 7, no. 5, 2021, doi: 10.1002/aelm.202001258.

[3] B. Yang *et al.*, "High-entropy enhanced capacitive energy storage," *Nature Materials,* vol. 21, no. 9, pp. 1074-1080, 2022, doi: 10.1038/s41563-022-01274-6.

[4] J. L. Braun *et al.*, "Charge-Induced Disorder Controls the Thermal Conductivity of Entropy-Stabilized Oxides," *Advanced Materials,* vol. 30, no. 51, p. e1805004, Dec 2018, doi: 10.1002/adma.201805004.

## *Acknowledgements*

ReMIND DOE Energy Frontier Research Center, Sandia National Laboratories LDRD

# Discussion – Neural-Inspired and Other Emerging Microelectronics and Computing

Brad Aimone, Moderator
Sandia National Laboratories

The Neural Inspired Technology session had three talks spanning a range of topics. The first talk was by Dmitri Strukov, whose talk focused on how physical hardware can efficiently implement high-order Ising machines, which in turn can be used to address a growing number of combinatorial optimization problems. He showed how near-term technologies, such as eFLASH, can be used for Ising simulations, and highlighted how longer-term options such as resistive memories (ReRAM) can achieve higher densities. He finally showed how higher order approaches where algorithm constraints can be mapped into the Ising formulation can be used to solve more complex problems.

Katie Schuman gave a talk on her group's neuromorphic software stack, including the EONS tool that uses genetic algorithms to identify compact spiking neuron circuits for tasks, ranging from radiation detection to engine control for fuel efficiency. One emerging conclusion from these studies is that different features of more advanced spiking neurons (leak, synaptic delay, etc) offer real advantages in practice.

Finally, Elliott Fuller described how materials exploration can be used to identify new high entropy oxide combinations that can make devices more suitable for performance conditions. The search space for these materials is combinatorially enormous, and classic screening strategies do not scale. Through strategies such as thin films which vary compound mixtures across a surface, device properties can be examined for many combinations so that underlying relationships can be identified that can be used to iteratively improve on device characteristics. Importantly, this accelerated mapping can be used to guide development of not only synapses but other types of devices as well.

## *Discussion*

- We are using AI to go "deep" in materials and circuits, but how do we pull these things together?
  - It isn't feasible to search all materials space inside each EONS optimization; can we start addressing that problem now before we go too deep?
  - EF - We need technology bridges that connect across stack.
  - CS - We need *people* who have visibility across stack.
    - Even some understanding across all levels of stack can be critical in reducing the complexity here.
    - How do we train people who have that broad view?
- We say AI should be good at exploring, but we need that human intuition to tell us when to zoom in on something new. How does this work?
  - EF - We will always need people who can look at the data and identify what is interesting. The exception to the rule is where the interesting science is
  - CS – People have to be in the loop.

- How much “new AI” is needed?
    - CS – Device and materials research is informing new directions for how to use AI
    - EF – Tools today are not as developed as you would like for use in materials research. Tools ask for data to be formatted and organized in way that experiments simply do not provide well.
    - BA – Most modern AI methods assume that everything is perfect. But reality of AI is outside of a few areas (computer vision), all AI is a few blocks from the wild west. Figuring out mapping from science to AI is as hard as developing the AI itself. It isn’t just install PyTorch and hit go.
    - EF – A big problem is that different scientific questions are at different levels of complexity. It isn’t reasonable to ask for automation and AI to solve things like correlated transitions at this, but some tasks will be a lot more straightforward to automate.

# Session IV: Next-generation and emerging microelectronics II

## Tackling Unpredictability in Emerging Memory Devices: the Bayesian Approach

Damien Querlioz
Université Paris-Saclay, CNRS, Centre de Nanosciences et de Nanotechnologies

### *Key Takeaways*

- Bayesian methods can mitigate the variability and unpredictability of resistive memory devices, enhancing AI accelerators by quantifying prediction uncertainty.
- This study shows three methodologies—Bayesian machines, BNNs, and Metropolis-Hastings MCMC technique—that leverage resistive memory's randomness for efficient and explainable AI applications.
- Demonstrations of these approaches reveal their practical viability and effectiveness in real-world applications, such as arrhythmia detection and cancerous tissue identification, with notable energy efficiency and accuracy.

### *Summary*

Resistive memory devices offer outstanding potential for near- and in-memory artificial intelligence (AI) accelerators due to their nonvolatility, high density, and multilevel capacity. However, these devices also present significant challenges due to their high variability and unpredictability, essentially rendering them functionally similar to random variables. This feature presents an intriguing parallel with machine learning, where Bayesian methodologies are expressly designed to work with random variables, delivering the added advantage of being able to quantify prediction uncertainty—a hurdle that conventional artificial intelligence techniques often struggle to overcome. In this study, we propose that these Bayesian methods could serve as an effective strategy to harness the potential of resistive memory devices, without the accompanying drawbacks. Our exploration introduces three distinct approaches that increasingly integrate the random nature of resistive memories. Each methodology is substantiated through practical implementation using circuits fabricated in a hybrid CMOS/hafnium-oxide memristor process.

We first introduce Bayesian machines, which leverage near-memory computing for energy-efficient Bayesian reasoning, enabling explainable decision-making under uncertain conditions by using all available data and knowledge. These machines digitally encode parameters in resistive memory, minimizing energy costs associated with data movement. Two designs were realized using a 130-nm hybrid CMOS/HfOx memristor process, featuring stochastic computing [1] and integer logarithmic computation [2] for efficient processing, both demonstrating remarkable energy efficiency.

Then, we show the realization of Bayesian neural networks (BNNs), which treat synapses as random variables to model uncertainty in predictions and utilize resistive memories to simulate the random behavior of synapses [3]. This approach, validated through the programming of 50 in-memory neural networks for arrhythmia detection, allows the networks to quantify prediction certainty by leveraging the

inherent randomness of memristors. The performance of these BNNs remains stable over time, incorporating resistive memory fluctuations into their uncertainty modeling.

Finally, we introduce the Metropolis-Hastings Markov Chain Monte Carlo (MCMC) technique, tailored to exploit the random nature of memristors for learning, enabling synaptic weights to be sampled in a way that mimics natural Gaussian distributions [4]. This method was applied to train experimentally a 16k array of resistive memories for cancerous tissue identification, achieving a 98% accuracy rate, comparable to traditional software-based methods, thereby illustrating the potential of memristors in advanced Bayesian learning applications.

## *References*

[1] K.-E. Harabi et al, "A memristor-based Bayesian machine", *Nature Electronics* 6, 52, 2023

[2] C. Turck et al, "Energy-Efficient Bayesian Inference Using Near-Memory Computation with Memristors", DATE 2023.

[3] D. Bonnet et al, (2023). “Bringing uncertainty quantification to the extreme-edge with memristor-based Bayesian neural networks”, *Nature Communications* 14, 7530 (2023).

[4] T. Dalgaty et al, "In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling", *Nature Electronics*. 4, 151, 2021

## *Acknowledgements*

This work was supported by ERC grants NANOINFER (715872) and DIVERSE (101043854) and a France 2030 government grant (ANR-22-PEEL-0010).

# Ising-CIM: Advancing Ising Accelerators for Combinatorial Optimization Problems with Compute-in-Memory Design Approach

Jaydeep P. Kulkarni
The University of Texas at Austin

## *Key Takeaways*

- *Key takeaway 1— Combinatorial optimization problems (COP) find applications in multiple domains such as logistics, resource management, healthcare, etc. Finding the optimum solution with the brute force search method is computationally expensive.*
- *Key takeaway 2—Multiple COPs can be solved efficiently by mapping into a generic Ising model framework, which emulates the spin dynamics in ferromagnets to reach the minimum energy state (optimal solution)*
- *Key takeaway 3 – There is a need for energy-efficient and cost-effective Ising accelerators. One promising approach is to perform Ising computation within the existing memory arrays, minimizing the data movement energy overheads; Silicon prototype (Ising-CIM) results showed >1000x improvement in a COP solution time over CPU.*

## *Summary*

Combinatorial optimization problems (COP) find many real-world social and industrial data-intensive computing applications. Examples include optimization of mRNA sequences for COVID-19 vaccines, semiconductor supply chains, and financial index tracking, to name a few. Such COPs are predominantly NP-hard, and performing an exhaustive brute-force search becomes untenable as the COP size increases. An efficient way to solve COPs is to let nature perform exhaustive searches in the physical world using the Ising model, which can map many COPs. The Ising model describes spin dynamics in a ferromagnet, wherein spins are naturally oriented to achieve the lowest ensemble energy state of the Ising model, representing the optimal COP solution.

Previous Ising model-based hardware designs have been realized by emulating spin dynamics using quantum bits, lasers, or coupled oscillators. However, these methods incur high cooling power costs for the quantum approach, large-size components for the optical approach, and high-power consumption for coupled oscillators, making them difficult to scale to large-size COPs. Another approach to designing Ising hardware is to abstract spin dynamics with a mathematical model (Hamiltonian), and update spins iteratively to minimize the Hamiltonian. Spin update computation involves numerous memory read and write operations. Consequently, iterative Ising model computing with conventional von Neumann architectures results in frequent off-chip memory accesses, causing performance and energy overheads. To reduce overhead, compute-near-memory (but not within memory) designs, which perform massively parallel computations near the memory bitcells, have been proposed earlier. Prior compute-near-memory-based Ising accelerators have reported up to 26000x speedup and ~1000x lower energy than CPUs. These custom designs implement multiple instances of digital arithmetic logic adjacent to unit memory segments.

This incurs large area overhead, increasing the hardware cost and drastically degrading the memory bit density by a factor of 3-10x. Despite impressive performance gains, these overheads have limited the adoption of compute-near-memory based Ising designs in current SoCs. Therefore, there is a need for research on energy-efficient and cost-effective hardware designs to advance COP accelerator development.

We propose Ising-CIM [1], an Ising accelerator design based on computation within prevalent embedded memories (as shown in Fig. 1). In contrast to prior near-memory digital arithmetic computing, this compute-in-memory (CIM) approach performs Hamiltonian computations in the analog domain within a memory array with minimal circuit changes. It prudently maps Hamiltonian computations onto the available memory wordline and bitline circuitry, which has remained a critical technical challenge. The key attribute of Ising-CIM is that all Hamiltonian calculations can be performed massively parallelly as analog bitline computations with minimal changes to peripheral array circuits. Although analog, CIM performs a one-bit spin update during a Hamiltonian iteration, with a local write-back mechanism using available sense amplifiers in a memory array. It acts as a one-bit comparator and eliminates the need for complex data converters. The Silicon prototype, implemented in a 65-nm CMOS process, demonstrates the Ising-CIM concept and achieves a 1091× speedup in annealing time compared to the CPU. This research can facilitate the rapid adoption of CIM-based Ising accelerators in modern SoCs at a low cost.

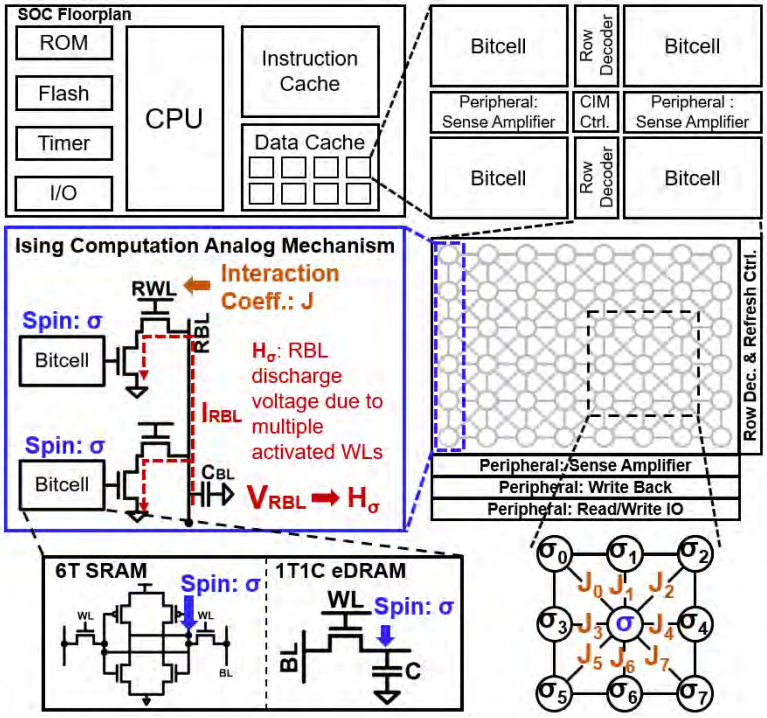



*Figure. 1 Ising-CIM Big picture: SoC floorplan (top), Hamiltonian computation with a memory array (SRAM or eDRAM) (bottom left), and King's graph mapping onto a memory bank (bottom)*

## *Conclusion*

Compute-in-memory-based Ising accelerator is a promising approach to performing massive parallel spin Hamiltonian computations and finding an optimal solution for combinatorial optimization problems. Future research may target scaling up the Ising-CIM approach to support full-spin connectivity and multi-bit interaction coefficients and applying similar CIM techniques for various optimization problems.

## *References*

[1] S. Xie, S. Raman, C. Ni, M. Wang, M. Yang, and J. P. Kulkarni, "Ising-CIM: A Reconfigurable and Scalable Compute within Memory Analog Ising Accelerator for Solving Combinatorial Optimization Problems"' IEEE Journal of Solid-State Circuits (JSSC), vol. 57, no. 11, pp. 3453-3465, Nov. 2022

## *Acknowledgments*

This research is supported by the NSF Career Award and the TSMC University shuttle program.

# The Future of Hardware Technologies for Computing

Subhasish Mitra
Professor of Computer Science | Stanford University

## *Summary*

The computation demands of 21st-century abundant-data workloads, such as AI/machine learning, far exceed the capabilities of today's computing systems. For example, a Dream AI Chip would ideally co-locate all memory and compute on a single chip, quickly accessible at low energy. Such Dream Chips aren't realizable today. Computing systems instead use large off-chip memory and spend enormous time and energy shuttling data back and forth. This memory wall gets worse with growing problem sizes, especially as conventional transistor miniaturization gets increasingly difficult.

The next leap in computing requires transformative NanoSystems by exploiting the unique characteristics of nanotechnologies and abundant-data workloads. We create new chip architectures through ultra-dense 3D integration of logic and memory – the N3XT 3D approach. Multiple N3XT 3D chips are integrated through a continuum of chip stacking/interposer/wafer-level integration — the N3XT 3D MOSAIC. To scale with growing problem sizes, new Illusion systems orchestrate workload execution on N3XT 3D MOSAIC creating an illusion of a Dream Chip with near-Dream energy and throughput.

Several hardware prototypes, built in commercial and research fabrication facilities, demonstrate the effectiveness of our approach. We target 1,000X system-level energy-delay-product benefits, especially for abundant-data workloads. We also address new ways of ensuring robust system operation despite the growing challenges of design bugs, manufacturing defects, reliability failures, and security attacks.

# Discussion – Next-Generation and Emerging Microelectronics II

Conrad James, Moderator
Sandia National Laboratories

The second session on next-generation and emerging microelectronics began with a talk from Damien Querlioz from Université Paris Saclay. His talk focused on the use of Bayesian models implemented in hardware to solve confidence-based problems. One of the challenges of using Bayesian models is that they are computationally expensive. He discussed the use of memristive device technologies to reduce the power consumption of such hardware, the use of 2T/2R architectures to reduce errors, and the use of a technique referred to as a logarithmic machine approach wherein multiplication operations are converted into additions to reduce energy consumption. The second speaker of the session was Jaydeep Kulkarni from the University of Texas at Austin. His talk described the use of Ising model based accelerators for solving combinatorial optimization problems such as Maxcut and the traveling salesman. He discussed several academic and commercial implementations of Ising models and then focused on a CMOS coupled oscillator based system fabricated in 65nm technology. Our third speaker of the session was Subhasish Mitra from Stanford University. He discussed the use of 3D microsystems as a way to circumvent limitations in memory and miniaturization, specifically the use of ultra-dense 3D integration where the vertical connections are on the order of 100nm wide and spaced.

- What does the application space look like for alternative computing methods such as quantum and neuromorphic computing?
    - The von Neumann vs non- von Neumann architecture distinction can be a distraction because the memory “wall” (too much data to move) is still not addressed
    - Applications need to achieve large improvements with a new technology in order to gain interest.
- How do the investments of domestic and foreign governments impact R&D in emerging microelectronics? What changes to the fab ecosystem need to occur to support innovation?
    - Foundries are being leveraged to support the development of emerging device and architecture technologies;
    - Fab engineers, architecture experts, and algorithm developers need to communicate more amongst each other to facilitate a deeper understanding across disciplines.
    - The access to fabs also needs to be improved to lower the barrier for entry for new researchers and smaller institutions.
    - EDA tools need to be improved in regard to reliability and security. 3D systems will present more of a challenge for design tools.

# Session V: Trust in AI for microelectronics design and future of ML in compiler construction

## Why Should AI-Enhanced Design of Microelectronics Care About Adversarial Machine Learning?

Evercita Eugenio
Sandia National Laboratories

### *Key Takeaways*

- Key takeaway 1 – Machine learning introduces an additional set of algorithmic vulnerabilities that allows those models to be subverted even when hardware, software, and network environments are perfect.
- Key takeaway 2 – Developing and using a machine learning hygiene checklist is important and writing down adversary models is a good practice to have.

### *Summary*

Machine learning (ML) advancements have grown rapidly in the last few years. However, machine learning is heavily dependent on data, as machine learning models are trained on vast datasets that are then used to provide insight or draw conclusions. This reliance on data means that machine learning can be subverted by an adversary who is able to manipulate even just a small portion of the data. This type of subversion is an example of an algorithmic vulnerability in machine learning that exists even when the networking, hardware and software environments are perfect.

The following is one possible taxonomy for adversarial machine learning (AML). There are many possible taxonomies, and this is just a starting point to consider.

- Subvert: Adjust the training data to undermine the model.
  - e.g., label poisoning, “bad nets”
- Evade: Adjust the test data to avoid correct classification.
  - e.g., adversarial test samples
- Reveal: Extract sensitive information from the machine learning model.
  - e.g., membership interference, model stealing.
- Apply: Use machine learning in adversarial ways.
  - e.g., deep fakes, toxic chemical discovery
- Other: Many new and creative edge cases are constantly emerging.

Based on these different categories, one can begin to understand the different vulnerabilities that exist in machine learning. Thus, this leads to some things to think about in relation to adversarial machine learning. The first is that as machine learning is being developed and used, it is important to have a machine learning hygiene checklist [15], [16]. Second, ML security should in many ways be treated like cyber security, where one does end-to-end analysis, risk assessments, considers the supply chain, etc.

Furthermore, when working with ML it is important to write down the adversary model so that one has a clear picture of what things could or could not be done to exploit the vulnerabilities of the ML. If sharing or acquiring ML models, insisting on training data and white box access to supplied machine learning systems would be extremely beneficial. Then, once one has those, one should carefully inspect those systems, even when tools are scare for this. Additionally, if you are the one sharing your ML, then minimizing what you share is important (i.e., exposing no more model information than necessary).

Finally, it is important to think carefully about releasing anything more than data categorization and be cautious about providing explainability tools because they may provide information you did not intend to release.

## *References*

[1] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey, 2021. URL https://arxiv.org/abs/2108.00401.

[2] Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freem an, Fabio Pierazzi, and Kevin A. Roundy. ”Real attackers don’t compute gradients”: Bridging the gap between adversarial ml research and practice, 2022. URL https://arxiv.org/abs/2212.14315.

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In 35th International Conference on Machine Learning, 2018.

[4] Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In 2022 IEEE Symposium on Security and Privacy (SP). IEEE, may 2022. doi: 10.1109/sp46214.2022.9833572. URL https://doi.org/10.1109%2Fsp46214.2022.9833572.

[5] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles, 2021. URL https://arxiv.org/abs/2112.03570.

[6] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023. URL https://arxiv.org/abs/2301.13188.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321–357, 2002. URL http://adsabs.harvard.edu/abs/2011arXiv1106.1813B.

[8] Nitesh V. Chawla, Lawrence O. Hall, Kevin W. Bowyer, and W. Philip Kegelmeyer. Learning ensembles from bites: A scalable and accurate approach. Journal of Machine Learning Research, 5:421–451, 2004.

[9] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pages 1322–1333, 2015.

[10] Maoguo Gong, Yu Xie, Ke Pan, Kalyuan Fend, and Alex Kai Qin. A survey on differentally private machine learning. IEEE Computational Intelligence Magazine, 15(2): 49-64, 2020

[11] Sanghyun Hong, Yiitcan Kaya, Ionu-Vlad Modoranu, and Tudor Dumitra. A panda? no, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference, 2020. URL https://arxiv.org/abs/2010.02432.

[12] Philip Kegelmeyer, Timothy M. Shead, Jonathan Crussell, Katie Rodhouse, Dave Robin-son, Curtis Johnson, Dave Zage, Warren Davis, Jeremy Wendt, Justin "J.D." Doak, Tiawna Cayton, Richard Colbaugh, Kristin Glass, Brian Jones, and Jeff Shelburg. Counter adversarial data analytics. Technical report, Sandia National Laboratories, 2015.

[13] W. Philip Kegelmeyer, Jr., Joe M. Pruneda, Philip D. Bourland, Argye Hillis, Mark W. Riggs, and Michael Nipper. Computer-aided mammographic screening for spiculated lesions. Radiology, 191(2):331–337, May 1994.

[14] Yisroel Mirsky, Tom Mahler, Ilan Shelef, and Yuval Elovici. CT-GAN: Malicious tampering of 3D medical imagery using deep learning, 2019. URL https://arxiv.org/abs/1901.03597.

[15] Bruce Nagy. Level of rigor for artificial intelligence development. Technical Report AD1173626, Naval Air Warfare Center Weapons Division, April 2022. Version 1.

[16] National Cyber Security Centre. Principles for the security of machine learning. Technical report, General Communications Headquarters, United Kindom, August 2022. Version 1.

[17] Florian Tramer, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. arXiv:2204.00032, 2022. URL https://arxiv.org/abs/2204.00032.

[18] Simon Weckert. Google maps hacks, performance & installation, 2020. URL https://www.simonweckert.com/googlemapshacks.html.

[19] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors, 2019. URL https://arxiv.org/abs/1910.14667.

[20] Irad Zehavi and Adi Shamir. Facial misrecognition systems: Simple weight manipulations force DNNs to err only on specific persons, 2023. URL https://arxiv.org/abs/2301.03118.

# Making Compilers More Evolvable with Learned Cost Models

Charith Mendis,
University of Illinois Urbana-Champaign

## *Summary*

Compilers use cost models either implicitly or explicitly to make optimization and code generation decisions. Traditionally, these cost models are constructed manually as analytical models or are embedded into the compiler code directly. However, with the diversity and evolving nature of the workloads and hardware, manually maintaining such cost models is both cumbersome and error prone. Such cost models easily become stale leading to poor optimization decisions. In this talk, I will show how we can automatically generate such cost models from data as opposed to manually writing them. Not only are these cost models significantly more accurate, but they are easier to develop and to maintain. I will demonstrate how to build data-driven cost models for different hardware (e.g., commodity hardware and domain-specific accelerators) as well as for different workloads (e.g., dense and sparse) that enable better code optimizations. As a testament to the success of this approach, Google has already deployed a learned cost model targeting TPUs, that not only enables better code generation, but also allows organic evolvability and adjustability with automatic periodic retraining. Such evolvability is not possible in traditional analytical cost models.

# Session VI: Microelectronics Workforce Panel

## The New Workforce Skills in Leveraging the Power of AI in Manufacturing

Valinda Scarbro Kennedy
HBCU SkillsBuild Program Manager | IBM

### *Key Takeaways*

- Cross pollution of workforce skills leveraging AI
- Developing sustainable semi-conductor ecosystem
- Competitiveness of US Market

### *Summary*

As we look at the semiconductor history in the US, we have an audience who has experience in many scenarios with the manufacturing industry in their community. Education on how the market is different today and how that will bring different long-term opportunities is critical across the US to have the impact needed for the US Semiconductor opportunity to deliver the financial impact, chip availability and scalability needed. Globally we are all more dependent on semi-conductor chip technology, the time to capture the US community attention is now. Enabling the US industry and education reach with internships, apprenticeships along with advanced master's and PhD degrees is a key foundational step that has to be redeveloped today.

Advancements of various technologies today, including artificial intelligence, provide a new opportunity for communities who previously may not have participated in this space to be uniquely positioned for success today. The investment in education in the classroom, leveraging digital assets and global relationships provides a platform previously unavailable.

### *Conclusion*

The semiconductor market is volatile as the need is present, but market forces are creating an unstable environment [1]. Developing a long-term road map based on an interconnected ecosystem will deliver the most stable global economic winners. Built on an established education, production and projected needs means the opportunity both short term and long term can deliver a model that meets the market need with a stable execution approach especially with the foundation models now provided by artificial intelligence.

### *References*

[1] Gartner Forecasts Worldwide Semiconductor Revenue to Grow 17% in 2024 Press Release 12/5/23, 8:23 Stamford, Conn.

*Acknowledgements*

Thank you to the entire IBM Team constantly adapting and changing to meet the needs of our global ecosystem.

# A National Strategy to Solve America's Tech Talent Shortage

*Creating Pathways to Advanced Degrees and Tech Careers Through Skills-Based Learning and Registered Apprenticeships*

Michael Russo
President and CEO | National Institute for Innovation and Technology (NIIT)

## *Key Takeaways*

- **Strategic Development of the U.S. Talent Pipeline:** The nation's need to build a resilient and skilled talent pipeline in strategic industry sectors is paramount. This is particularly the case for the semiconductor manufacturing and nanotechnology-related industries, which are critical for the nation's national security and global competitiveness.
- **Elevation of Skills-Based Learning and Registered Apprenticeships:** At the heart of the strategy the Institute is executing on is a commitment to skills-based education and Registered Apprenticeship (RA) models. These frameworks are instrumental in bridging the gap between academic preparation and the practical demands of the industry, ensuring that learners acquire relevant, foundational, and actionable skills.
- **Inclusive and Collaborative Engagement:** The strategy the Institute is executing on underscores the value of engaging a broad spectrum of stakeholders, including young students, returning service members, and adults aspiring to pivot to high-quality careers. This inclusive approach is designed to diversify and expand the talent pool, ensuring a robust and continuous supply of skilled professionals in semiconductor manufacturing.

## *Summary*

In this presentation, attendees can anticipate an in-depth exploration of the U.S.' strategic response to a pressing talent shortfall in the semiconductor and advanced manufacturing sectors. As the nation stands at a critical crossroads, aiming to recapture its global leadership in these industries that are critical to national security and global competitiveness, a concerted effort to solve the talent scarcity is underway. Spearheaded by the National Institute for Innovation & Technology as a U.S. Department of Labor Intermediary, the proposed strategy is a spectrum of initiatives designed to foster skills-based learning and Registered Apprenticeships (RAs), ultimately forging accessible pathways to high-value careers within the industry.

This comprehensive strategy introduces tools and infrastructure accessible to states and regions, optimized for efficiency and cost-effectiveness. This approach not only bolsters regional talent pipelines and encourages place-based economic growth but also streamlines the strategic and developmental process for regional entities, allowing them to focus resources on regional deployment.
Key facets of the strategy include:

1. **Development of a Versatile Talent Pipeline:** Creating a pipeline of individuals endowed with foundational skills that are transferable across various roles within the semiconductor and

advanced manufacturing sectors is essential. This strategic move guarantees a workforce capable of adapting to the ever-changing industrial landscape.

2. Skills-Based Learning in Public Education: Embedding skills-based STEM-related learning within the public education system ensures that students from a young age are prepared with the practical skills and knowledge essential for succeeding in high-tech industries.

3. Dynamic Tools for Job Connectivity: The strategy also includes the development of dynamic tools aimed at aligning individual skills with industry needs, like the Institute's National Talent Hub, facilitating seamless transitions for job seekers from a variety of fields.

4. 'Learn and Earn' Opportunities: Emphasizing the importance of 'learn and earn' models, the strategy seeks to open new opportunities for a broad demographic, particularly those previously disengaged from more traditional educational pathways, enabling them to pursue high-skill careers.

5. Expansion of the Talent Pool: A critical component is the significant expansion of the participant pipeline to embrace a diverse population. This includes enhancing the existing public education systems (from K-12 to post-secondary), reintegrating returning service members and their families, and providing adults with avenues to pivot into advanced careers in the industry at any stage.

The strategy's embodiment, the National Talent Pipeline Development Initiative (TPDI), is a testament to the nation's commitment to attract, engage, and create career pathways for a diverse workforce. By connecting industry, education, and job seekers, TPDI is vital in cultivating a talent pipeline that is central to national security and global competitiveness.

Moreover, the strategy is forward-looking, addressing the entire supply chain talent pipeline from its construction phase to operational maturity. It is designed to not only support existing ecosystems but also to anticipate and prepare for the emergence of new ones, ensuring the industry is primed with a ready pool of skilled professionals.

In summary, this presentation will articulate how the United States is fortifying its semiconductor and nanotechnology-related workforce through a strategic, inclusive, and comprehensive approach to talent development. By underscoring the significance of skills-based learning, RAs, and collaborative engagement with a broad spectrum of the population, the initiative aims to ensure that educational outcomes are in tandem with the needs of industry. These novel approaches are necessary for cementing the nation's stature as a global vanguard in technological innovation and manufacturing excellence, fostering economic growth, and safeguarding national security against the backdrop of an accelerating global technological race.


*Figure 1. National Talent Pipeline Development Initiative Strategy & Ecosystem*

## *Conclusion*

To transform these strategies into tangible results, the following next steps are proposed:

1. **Establishment of Public-Private Partnerships:** Forging strong collaborations between government entities, educational institutions, and industry leaders to ensure that the initiatives have the support, funding, and guidance necessary to succeed.

2. **Implementation of Skills-Based Curriculum:** Working closely with educational policymakers to integrate skills-based learning into the core curriculum across K-12 and post-secondary education, preparing the next generation of workers for the demands of the industry.

3. **Expansion of Registered Apprenticeship Programs (RAPs):** Scale up RAPs to provide more individuals with on-the-job training, complemented by related technical instruction, thus creating a ready-to-deploy workforce.

4. **Inclusive Outreach Programs:** Develop targeted outreach programs aimed at diverse groups, including returning service members, their families, and adults seeking new career paths, to ensure that all potential talent pools are tapped into.

5. **Continued Expansion of the Institute's National Talent Hub:** Building on the centralized platform the Institute has created where industry needs, educational resources, and job seekers can converge, providing real-time data and analysis to align workforce development with market demands.

6. **Investment in Infrastructure:** Allocating resources towards the development of the necessary infrastructure to support the growth of new and existing ecosystems, anticipating the needs of the industry before they arise.

7. **Monitoring and Evaluation:** Implementing a robust system of monitoring and evaluation to track the progress of the initiatives, allowing for adjustments and improvements based on performance data and feedback from stakeholders.

8. **National Campaign for Awareness:** Initiating continued national campaigns to raise awareness about the importance of the semiconductor and advanced manufacturing sectors, and the rewarding career opportunities they offer, to inspire a new generation of innovators and builders.

By following these steps, the U.S. can ensure the successful execution of its national strategy for talent development. The goal is not just to fill jobs but to catalyze a fundamental shift in how the nation develops its workforce, ultimately ensuring that it remains at the forefront of technological innovation and manufacturing excellence. This strategic effort represents a commitment to the economic prosperity and national security of the U.S., paving the way for a future where the tech talent pipeline is as dynamic and innovative as the industries it serves.

## *References*

[1] https://www.niit.org/the-national-talent-hub-niit

[2] https://www.niit.org/apprenticeship-gains

[3] https://www.niit.org/sam-tap-program

[4] https://www.niit.org/vetconnect-veterans-fellowship-program

[5] https://www.niit.org/real-robotics-for-kids-program

[6] https://www.niit.org/csbl

## *Acknowledgements*

We extend our gratitude to the USDOL for its unwavering support and commitment to workforce development. The collaboration with the agency has been pivotal in developing RAPs that are key to our strategic talent development.

Furthermore, we acknowledge the White House's Advanced Manufacturing Sprint, which has provided crucial momentum to our efforts. This initiative underscores the federal government's dedication to reinforcing America's manufacturing capabilities and has significantly contributed to the impetus behind our national strategy.

The synergy between these esteemed bodies and our initiative is a testament to the power of collaborative effort. Together, we are forging a path towards a brighter future for the American workforce and ensuring that our nation continues to lead in innovation and manufacturing on the global stage.

# Creating Opportunities Everywhere: STEM Investments in Education and Workforce Development

James L. Moore III
Assistant Director for the Directorate for STEM Education | U.S. National Science Foundation

## *Key Takeaways*

- *Key takeaway 1— NSF's investments in semiconductors and microelectronics will help the Nation keep pace with a changing, global innovation landscape.*
- *Key takeaway 2— NSF is committed to educating and training a semiconductors and microelectronics workforce and advancing opportunities for equitable STEM education.*
- *Key takeaway 3— NSF is leveraging its partnerships to accelerate at speed and scale the translation of research and workforce development.*

## *Summary*

For decades, the U.S. National Science Foundation (NSF) has made significant contributions to the field of semiconductors and microelectronics, ranging from supporting education and research, funding scholarships and fellowships, and investing in the Nation's workforce development efforts. And, more recently, the Foundation has begun to accelerate its efforts to leverage partnerships with industry, including Micron and Intel, to expand microelectronics education and workforce development. In this presentation, James L. Moore III, assistant director for STEM Education at NSF, will both highlight and discuss the Foundation's efforts in developing the education, technology, workforce, and infrastructure necessary to strengthen the Nation's semiconductors and microelectronics field.


*Figure 1. Advanced Technological Education (ATE) program is NSF's flagship Community College program focused on high skilled technical education, essential for the semiconductor workforce; impacts over 40K students and 90K faculty.*

# Higher Education at the Speed of Life

Molly Dodge
Senior Vice President for Workforce and Careers | Ivy Tech Community College

## *Key Takeaways*

- ***Ivy Tech is Indiana's Community College:*** *Ivy Tech offers full-service campuses and manufacturing labs across Indiana, and serves various sectors, including K-12, veterans, and Department of Correction to offer degrees, diplomas, and certification training.*
- ***Industry and Community Partnerships:*** *Working closely with Indiana Economic Development Corporation and their industry partners in semiconductor and EV battery manufacturing, these groups have evaluated and validated Ivy Tech's curriculum to ensure that students have the skills and competencies that are aligned to employer needs.*
- ***Evolving education to meet national needs:*** *Ivy Tech is developing new programs that train students in the skills currently needed to advance national microelectronic and semiconductor industry. Training in these areas is offered via stackable credentials, allowing flexibility to progress from a short-term certificate to certifications and degrees.*

### *Summary*

***Ivy Tech is Indiana's Community College***: Ivy Tech offers 19 full-service campuses, 9 smart manufacturing labs and multiple advanced manufacturing labs across Indiana. It's the largest singly accredited Community College in the nation with 70 programs aligned to Indiana's economy and the largest dual credit institution serving 90,000 high school students across the U.S. It also serves K-12 career and technical education programs, tailors course work to Indiana veterans, and is contracted with the Department of Correction to offer the high school equivalency diploma and industry certification training.

**Industry and Community Partnerships:** In partnership with the Indiana Economic Development Corporation and industry partners, **industry certifications** have been facilitated by numerous vendor relationships across all sectors, ensuring that students have the skills and competencies that are validated by a third party to align to employer needs. By participating in curriculum design, industries are more likely to hire Ivy Tech students and may sometimes invest in employees training by sponsoring continued education at Ivy Tech. Ivy Tech also partners with Purdue University where students can transfer to study for a smart manufacturing industrial informatics bachelor's degree.


*Figure 1. Ivy Tech's STEM programs and curriculum evolves to meet Indiana's manufacturing needs.*

Evolving education programs to meet national needs

- Stackable credentials are a flexible pathway to not only associate degrees, but also long-term or technical certificates that stack into those associate degrees and are often designed by working with employers who identify emerging skills and competencies needed and map them to courses that compose a particular certificate or associate degree. These also serve as on and off ramps for students in which employers will hire students with short term certificates and then pay to send them back to Ivy Tech to continue their education, leveraging their tuition reimbursement programs. Education attained via stackable degrees certificates require approximately 21 credit hours, technical certificates ~ 31 hours, and associate degrees ~70 credit hours.

- In 2022, Indiana's manufacturing industry and institutions evolved into Industry 4.0. As industry shifted so did Ivy Tech which became one of the first community colleges to launch the Smart Manufacturing and Digital Integration degree, which introduces students to the industrial Internet of Things, industrial robots with vision capabilities, collaborative robots, data analytics, a high level of automation, cyber physical systems, and discovery on how to integrate new technologies within an automated system. The program is designed to provide a lifetime career opportunity for current and future students and ensure Indiana's economic competitiveness with the latest trends in industry. Third party certifications are embedded within the Smart Automation Certification Alliance or SACA certifications, offered as standalone non-credit training.

- New certificates to support Semiconductor Manufacturing & the Technician Roles SK Hynix recently announced a $4 billion investment in Indiana to create a semiconductor chips and innovation facility, a credit to Purdue University as an asset in this sector. In anticipation, Ivy Tech has aligned the technician role to the smart manufacturing digital integration degree, with electives such as art, industrial technology, design technology and manufacturing production and operations, from which employers can create interdisciplinary degrees that align to specific roles and can choose courses from each of these programs. In this way Ivy Tech is creating the technician workforce for data analytics, cyber security. Courses can also stack into the smart manufacturing degree or be offered as a standalone technical certificate for those needing upscaling to enter the semiconductor industry.

Acknowledgements
*Abstract composed from presentation recording*

# Discussion – Microelectronics Workforce Panel

Alan Mitchell, Moderator
Sandia National Laboratories

## *Key Takeaways from Panel Discussion*

Following presentations, panelists discussed the most significant obstacles for advancing the microelectronics workforce development today, and shared the following perspectives on what was needed to remove the obstacles.

Student gateways to the microelectronics industry

- **Paid micro internships** are an emerging trend, allowing students to work on smaller projects for 10,15 or 20 hours at entrepreneurial companies, and are beneficial in helping students understand different applications of AI, data science etc., connecting them with companies, and the industrial work environment. This is a great way to incentivize and upskill and generates a greater understanding of how the world and employment needs are changing.
- Many in the US population do not have good gateways to even realize that microelectronics is a good career path nor have the support to get students into that pathway, **career coaching** is critical. An interesting topic of discussion was noted that there are few jobs and that industry hiring is cyclical, but that is in contrast to the CHIPS & Science Act which forecasts workforce development in the tens of thousands of needed jobs to fill. Time will tell how much the CHIPS Act accomplished to create those jobs.

Preparation & Connection

- It's a challenge for students to be interested in something that they've never been exposed to. An **awareness campaign** and an opportunity for as early as middle school students to job shadow, participate in plant tours, and engage their families is needed. One of NSF's grand challenges is to increase student exposure to microelectronics applications and research to foster their interest and enthusiasm for the field, and to do so through strong mentoring. Achievement is highly correlated with the relationship that a student has with his or her teacher, as well as the number of contact hours spent with teachers. Thus, mentoring is an essential aspect
- Across the country, not all students have access to a quality STEM education and resources, and **standardization of public-school STEM resources are needed** (teachers, curricular opportunities, facilities etc.). For example, all AP classes are not created equally, so that upon matriculation, students may not have been exposed to the same rigor of other high schools. Serving rural communities to ensure that their education and technology needs are met is also of importance.

Skill building

- For students to have viable chances of success in these careers, greater access to quality STEM education is needed for much of the US. In particular, the **K-12 system will need to adapt to equip students with new skills that are foundational and transferable lational** to support artificial intelligence, data science, and cybersecurity and to gain admission to jobs. Basic upskilling for the high school teachers and curriculum will be needed.

- As seen with Ivy Tech Community College, non-degree, **stackable credentials, and certificate programs** are mechanisms that allow upskilling which is vital both for those already possessing some level of postsecondary education as well as for entry into further post-secondary education and microelectronics workforce opportunities.

- **Partnerships** with organizations in the semiconductor industry across the country is a path to successful workforce development as it involves **sharing, learning, and distributing the capabilities that already exist**. This is more effective as opposed to attempting to redevelop capabilities to grow the skills base. For example, in Indiana, the workforce talent pipeline encourages participation not only from higher education and employers but also community and faith-based partners which help to broaden the reach of semiconductor ecosystems.

# Session VII: Electronic design automation and ML

## AI for Chip Design/EDA: Everything, Everywhere, and All at Once (?)

David Z. Pan
Professor, Silicon Laboratories Endowed Chair in Electrical Engineering | The University of Texas at Austin

### *Key Takeaways*

- AI for chip design and EDA touches everything and everywhere, in all kinds of design objectives, all key design steps, and digital / analog designs.
- AI for EDA is still not yet all at once, e.g., generative AI from design specification to layout, in a correct-by-construction manner.
- More research is needed in particular for AI for Analog and RF IC designs.

### *Summary*

AI for chip design and EDA has received tremendous interests from both academia and industry in recent years. It touches *everything* that chip designers care about, from PPA to cost/yield, turn-around-time, security, and so on. It is *everywhere*, in all levels of design abstractions, testing, verification, DFM, mask synthesis / ILT, and some aspects of analog/mixed-signal/RF designs as well. It has also been used to tweak the overall design flow and hyper-parameter tuning, but not yet *all at once,* e.g., generative AI from design specification to layout, in a correct-by-construction manner. In this talk, I will cover some recent advancement/breakthroughs in AI for chip design/EDA and share my perspectives.

AI/ML can serve as hammers, bridges, and optimizers for prediction, generation, acceleration, and optimization. For acceleration, DREAMPlace was proposed to solve the state-of-the-art nonlinear placement problem through a **novel analogy** by casting it into a neural network training problem, and then leverage deep learning hardware and software to offer 30x speedup [1]. The DREAMPlace family has been extended to handle congestion, timing, macro-placement, and FPGA placement (DREAMPlaceFPGA). Generative AI was shown to help provide routing guidance and direct lithography printed-image prediction (LithoGAN [3]). For analog design automation, an open-sourced layout design automation tool, MAGICAL has been developed to generate layout directly from a circuit netlist [2] (Fig. 1). We are also working on analog sizing [5] and topology generation [6] leveraging ML, hoping to build an end-to-end analog "RTL to GDSII" framework. We also plan to extend our ML-assisted design automation framework to advanced packages [4] and RF/mmWave ICs.

To summarize, AI for chip design/EDA is everything, everywhere, in PPACT, digital, analog, …, with all kinds of AI/ML algorithms, to perform acceleration, prediction, generation, architecture exploration. However, there are still lots to be done: DATA, model generalization, transferability, bias, interpretability/explainability, optimality… And it is still far away from **"all at once"** (the ultimate dream). As quoted in the Communications of ACM Jan. 2023 "Member News" interview, "My dream is to have a silicon compiler which can let people design chips as easily as they can write software".

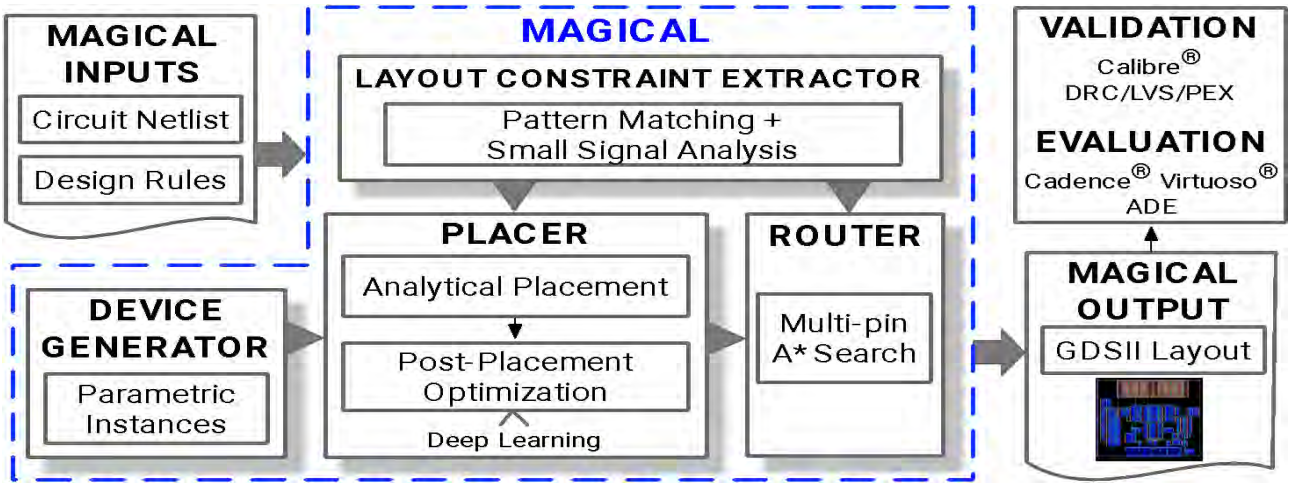



*Fig. 1 The MAGICAL analog layout system [2].*

## *References*

[1] Yibo Lin, Zixuan Jiang, Jiaqi Gu, Wuxi Li, Shounak Dhar, Haoxing Ren, Brucek Khailany, and David Z. Pan, "DREAMPlace: Deep Learning Toolkit-Enabled GPU Acceleration for Modern VLSI Placement," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Vol. 40, no. 4, pp. 748-761, April 2021. (**Donald O. Pederson Best Paper Award**)

[2]Keren Zhu, Hao Chen, Mingjie Liu, and David Z. Pan, "Tutorial and Perspectives on MAGICAL: A Silicon-Proven Open-Source Analog IC Layout System," *IEEE Transactions on Circuits and Systems II: Express Briefs (TCAS-II)*, Date of Publication: May 2022. (**Invited**)

[3] Wei Ye, Mohamed Baker Alawieh, Yibo Lin, and David Z. Pan, "LithoGAN: End-to-End Lithography Modeling with Generative Adversarial Networks," *ACM/IEEE Design Automation Conference (DAC)*, Las Vegas, NV, Jun. 2-6, 2019. **(Best Paper Award Candidate)**

[4] Hyunsu Chae, Keren Zhu, Bhyrav Mutnury, Douglas Wallace, Douglas Winterberg,Daniel de Araujo, Jay Reddy, Adam Klivans, and David Z. Pan, "ISOP+: Machine Learning-Assisted Inverse Stack-Up Optimization for Advanced Package Design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 43, no. 1, pp. 2-15, Jan. 2024.

[5] Ahmet F. Budak, Prateek Bhansali, Bo Liu, Nan Sun, David Z. Pan, and Chandramouli V. Kashyap, "DNN-Opt: An RL Inspired Optimization for Analog Circuit Sizing using Deep Neural Networks," *ACM/IEEE Design Automation Conference (DAC),* San Francisco, CA, Dec. 5-9, 2021. (one of the five Best Paper Candidates, out of 900+ submissions)

[6] Souradip Poddar, Ahmet Budak, Linran Zhao, Chen-Hao Hsu, Supriyo Maji, Keren Zhu, Yaoyao Jia, and David Z. Pan, "A Data-Driven Analog Circuit Synthesizer with Automatic Topology Selection and Sizing," *IEEE/ACM Design, Automation and Test in Europe (DATE)*, Valencia, Spain, March 25-27, 2024.

## *Acknowledgements*

This work is supported in part by NSF, DARPA, SRC, AMD, Cadence, Google, Intel, Nvidia, Samsung, Synopsys.

# AI: Trailblazing the Next Generation of Semiconductor Innovation

Vishal Khandelwal
Synopsys Inc.

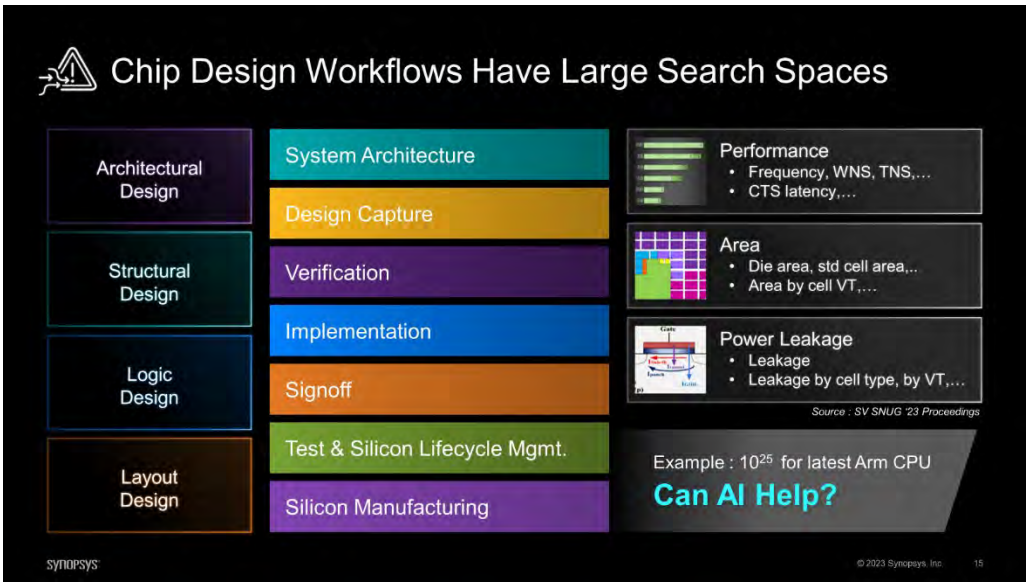
## *Key Takeaways*

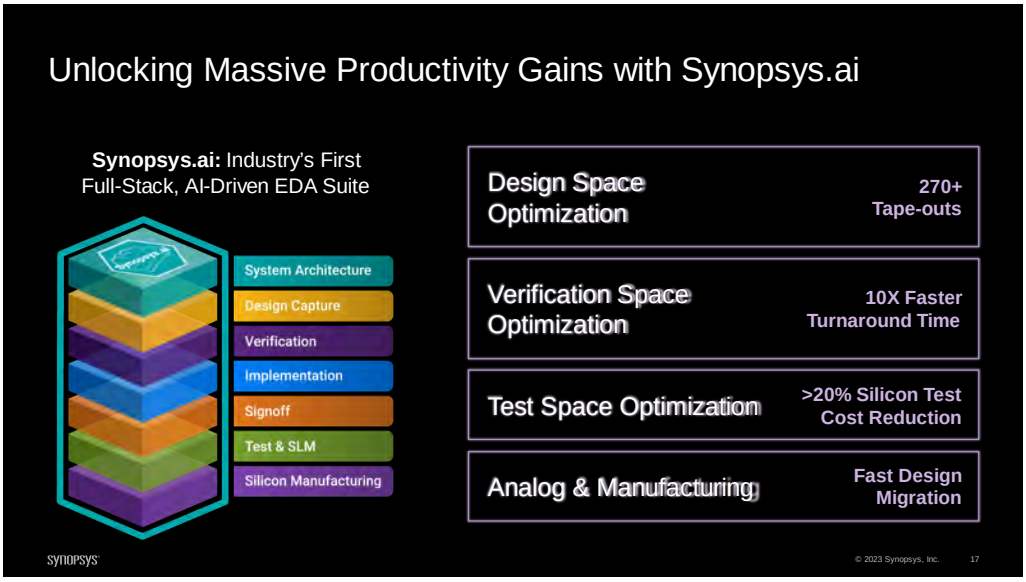
- AI growth is fueling exponential complexity increase in digital SOCs, paving the way for multi-die systems to bridge together chip design complexity and a decelerating Moore's Law. There is tremendous potential (and necessity) in leveraging AI to innovate and transform chip design.
- Hardware design innovation requires AI optimized solutions for the entire tool stack, driving architecture, verification, implementation, signoff and manufacturing.
- Significant productivity, quality-of-results and time-to-market boost from AI in EDA tools, in times where talent shortage is widespread in the semiconductor industry.

## *Summary*

With AI-based applications coming pervasively mainstream through ChatGPT, digital healthcare, and almost everything else we do with our devices on a daily basis, digital chip design is seeing an exponential push towards complexity, performance/power and time-to market. As semiconductor industry pivots towards multi-die system, built on a mix of advanced and mainstream process nodes (balance cost vs. performance), the EDA tool flows are becoming mission critical. The advent of transformer models has led to an explosion of compute, where hardware design is now dealing with "processing", "memory" and "communication bandwidth" walls, to meet the demands of a hardware-led 4$^{th}$ industrial revolution that is driving everything towards being intelligent, connected, monitored and data-driven.

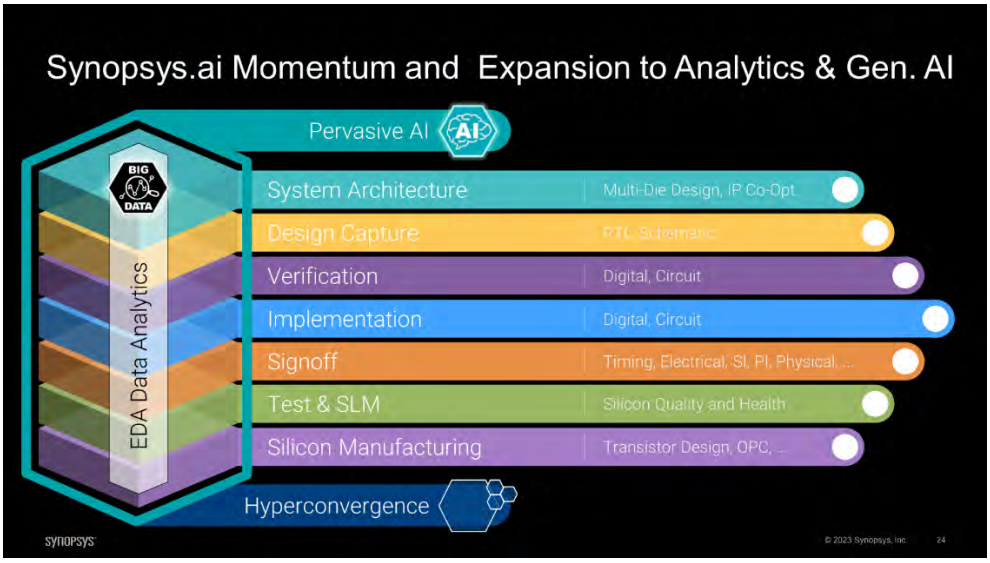
In this talk we go into the details of how hardware design and EDA tools are evolving to deliver the next generation of performance, power, and productivity (PP&P) boost. Even with exponential growth in design productivity in the past decade, overall design cycle is not meeting its intended targets. The technical complexity of advanced node design with new dominant effects like thermal and IR, are leading to highly complex tool/methodology flows.




This is where at Synopsys, we have taken a full tool-stack approach to build pervasive AI into our tools, boosting multiple aspects of PP&P across our customers. We will share examples of applications ranging from design implementation, verification, test, and analog/manufacturing to showcase the potential of this

technology and how it is reshaping chip design workflows. Many of these technologies are taking us closer to the no-human-in-the-loop goal for chip design. With a serious talent shortage and the increase is solution complexity beyond human expertise, AI-augmented solutions for chip design is the way forward.




## *Conclusion*

Hardware design innovation is at the heart of the pervasive AI boom around us. The complexity of technology, shortage of talent and push for product differentiation necessitates the use of AI across the design tool flows, leading to gains in productivity, performance/watt, and time-to-market. We see pervasive AI across the EDA tool stack as a key enabler for semiconductor industry.

## *References*

[1] Yi-Chen Lu, Wei-Ting Chan, Deyuan Guo, Sudipto Kundu, Vishal Khandelwal, and Sung Kyu Lim, “RL-CCD: Concurrent Clock and Data Optimization using Attention-Based Self-Supervised Reinforcement Learning”, In 2023 60th ACM/IEEE Design Automation Conference (DAC)

## *Acknowledgements*

Many thanks to Synopsys colleagues for various technical collaborations.

# ChipNeMo: Domain-adapted LLMs for Chip Design

Haoxing (Mark) Ren
NVIDIA

## *Key Takeaways*

- LLM can provide chip designers with both knowledge assistance and coding assistance.
- Domain adapted LLM models perform better than general purpose LLM models of similar size.
- Retrieval augmented generation (RAG) is very effective in improving model performance. Domain adapted models together with RAG give the best overall performance.

## *Summary*

ChipNeMo aims to explore the applications of large language models (LLMs) for industrial chip design. Instead of directly deploying off-the-shelf commercial or open-source LLMs, we instead adopt the following domain adaptation techniques: domain-adaptive tokenization, domain-adaptive continued pretraining, model alignment with domain-specific instructions, and domain-adapted retrieval models. We evaluate these methods on three selected LLM applications for chip design: an engineering assistant chatbot, EDA script generation, and bug summarization and analysis. Our evaluations demonstrate that domain-adaptive pretraining of language models, can lead to superior performance in domain related downstream tasks compared to their base LLaMA2 counterparts, without degradations in generic capabilities. In particular, our largest model, ChipNeMo-70B, outperforms the highly capable GPT-4 on two of our use cases, namely engineering assistant chatbot and EDA scripts generation, while exhibiting competitive performance on bug summarization and analysis. These results underscore the potential of domain-specific customization for enhancing the effectiveness of large language models in specialized applications.

## *References*

[1] Mingjie Liu, et al, “ChipNeMo: Domain-adapted LLMs for Chip Design”, https://arxiv.org/abs/2311.00176

# Discussion - Electronic Design Automation and ML

Ben Feinberg, Moderator
Sandia National Laboratories

## *Session Summary*

AI and ML methods have enormous potential to significantly accelerate complex tasks in electronic design automation (EDA). Each of the speakers, emphasize different aspects of ML for EDA and the different classes or problems which can be improved through ML models. David Pan from the University of Texas at Austin spoke about the use of ML tools for traditional EDA tasks such as placement and routing, where the problem is NP-Hard. ML models carefully trained models can function as heuristic optimizers and can significantly out-perform traditional heuristics. Professor Pan also noted the use of ML models as surrogates to replace costly simulation such as for lithographic hotspot detection and power, performance, and area (PPA) prediction. Vishal Khandelwal from Synopsys discussed how ML models like the ones described by Professor Pan are currently being used and shipped as products by Synopsys. Dr. Khandelwal also emphasized the virtuous cycle of ML models since hardware is powering the current AI/ML revolution. Improvements in EDA, both in designed productivity and result quality improve hardware which enables even better models. Finally, Mark Ren from Nvidia discussed Nvidia's use of large language models (LLMs) to improve developer productivity for chip design. Dr. Ren described how Nvidia has trained an LLM specifically for a variety of chip design Q&A applications including answering questions about internal tools and documentation and generating code examples for those internal tools. Dr. Ren also noted that a relatively small number of Nvidia-developed prompts were needed in the training set to achieve good results.

## *Key Points:*

- AI/ML models are already very successful in a wide-range of EDA and chip-design relevant problems.
  - ML models as optimizers for tasks such as placement, routing, and cell layouts for library generation.
  - Surrogate models for thermals, lithography, and PPA.
  - Code generation for HDLs and test-benches.
  - Generally difficult to determine a priori which tasks are well-suited for ML models vs not, although data availability is a major bottleneck.
- Data availability is a significant challenge.
  - Low-level design information is deeply proprietary but large amounts of data are needed to effectively train models.
  - When working with commercial data from multiple customers, significant care must be taken to avoid data-leakage between customers. Individual customer-specific models are likely safer than trying to avoid data-leakage.
  - ML-generation of examples may be able to help fill data gaps, although likely not as a useful as human-developed designs.
  - For companies with significant chip design experience, internal data only may be sufficient to achieve high-quality results from ML models.

# Conclusions

The 2024 Workshop on AI-Enhanced Co-Design for Next-Generation Microelectronics was the fourth and culminating workshop in the series organized by Sandia National Laboratories. Progress has been rapid, and drivers have evolved throughout the time spanning the four workshops, 2021-2024, including the passage of the CHIPS & Science Act in 2022 as well as other federal funding opportunities.

These workshops have been aimed at bringing together experts from universities, industry, and national labs to survey the state-of-the-art and to discuss future needs, directions, and priorities for R&D, and to consider the important educational and workforce development efforts needed to build the next generation of technicians, engineers, and scientists.

The key focus of ongoing work and of future directions has been the need to overcome the limitations currently slowing Moore's Law and to develop greater energy and operational efficiency of microelectronics for computing, sensing at the edge, and other applications, as has been the enabling use of AI/ML and co-design.

# APPENDIX A: Workshop Agenda

**Day 1: AI-Enhanced Co-Design for Next Generation Microelectronics**

| | |
|---|---|
| 9:00-9:15 | *Welcome and workshop introduction* – Conrad James, Sandia National Laboratories |

**Next-generation and emerging microelectronics I**

| | |
|---|---|
| 9:15-9:55 | *NorthPole Chip: Neural Inference at the Frontier of Energy, Space, and Time* – Filipp Akopyan, IBM |
| 9:55-10:35 | *Machine Learning Enabled Co-Design of Microelectronic Systems* – Madhavan Swaminathan, Penn State University |
| 10:35-10:45 | Break |
| 10:45-11:25 | *Data-Driven Synthesis and Characterization for Accelerated Electronic Materials Discovery* – Christopher Hinkle, Notre Dame |
| 11:25-11:40 | Guided and moderated discussion with the session's speakers and audience – Chris Allemang, Sandia National Laboratories, moderator |
| 11:40-12:10 | Lunch |

Advanced packaging and heterogeneous integration

| | |
|---|---|
| 12:10-12:50 | *AI for Testing AI Hardware: A Virtuous Cycle* – Krishnendu Chakrabarty, Arizona State University |
| 12:50-1:30 | *2.5D/3D Heterogeneous Integration Co-Design for In-Pixel Processing, LLM Acceleration and Bioinformatics* – Shimeng Yu, Georgia Tech |
| 1:30-1:40 | Break |
| 1:40-2:20 | *Co-Design Challenges in Modern Heterogeneously Integrated Microsystems* – Chris Nordquist, Sandia National Laboratories |
| 2:20-2:35 | Guided and moderated discussion with the session's speakers and audience – Cale Crowder, Sandia National Laboratories, moderator |
| 2:35-2:45 | Break |

Neural-inspired and other emerging microelectronics and computing

| | |
|---|---|
| 2:45-3:25 | *Efficient Implementation of High-Order Ising Machines* – Dima Strukov, University of California Santa Barbara |
| 3:25-4:05 | *Evolutionary Optimization for Neuromorphic Co-Design* – Katie Schuman, University of Tennessee and Oak Ridge National Laboratory |
| 4:05-4:15 | Break |
| 4:15-4:55 | *High Entropy Oxides for Next Generation Microelectronics* – Elliot Fuller, Sandia National Laboratories |
| 4:55-5:10 | Guided and moderated discussion with the session's speakers and audience – Brad Aimone, Sandia National Laboratories, moderator |
| 5:10 | Day 1 wrap up and Day 2 outlook – Conrad James, Sandia National Laboratories |

Day 2: AI-Enhanced Co-Design for Next-Generation Microelectronics

| | |
|---|---|
| 9:00-9:10 | Welcome, impressions from Day 1 and outlook for Day 2 – Conrad James, Sandia National Laboratories |

Next-generation and emerging microelectronics II

| | |
|---|---|
| 9:10-9:50 | *Tackling Unpredictability in Emerging Memory Devices: The Bayesian Approach* – Damien Querlioz, Université Paris Saclay |
| 9:50-10:30 | *Ising-CIM: Advancing Ising Accelerators for Combinatorial Optimization Problems with Compute-in-Memory Design Approach* – Jaydeep Kulkarni, University of Texas at Austin |
| 10:30-10:45 | Break |
| 10:45-11:25 | *The Future of Hardware Technologies for Computing* – Subhasish Mitra, Stanford |
| 11:25-11:40 | Guided and moderated discussion with the session's speakers and audience – Suma Cardwell, Sandia National Laboratories, moderator |
| 11:40-12:10 | Lunch |

Trust in AI for microelectronics design and future of ML in compiler construction

12:10-12:40 *Why Should AI-Enhanced Design of Microelectronics Care About Adversarial Machine Learning* – Evercita Eugenio, Sandia National Laboratories

12:40-1:20 *Making Compilers More Evolvable with Learned Cost Models* – Charith Mendis, University of Illinois Urbana-Champaign

1:20-1:30 Break

Microelectronics Workforce Panel

1:30-1:40 *The New Workforce Skills in Leveraging the Power of AI in Manufacturing* – Valinda Kennedy, IBM

1:40-1:45 Q&A

1:45-1:55 *A National Strategy to Solve America's Tech Talent Shortage Creating Pathways to Advanced Degrees and Tech Careers Through Skills-Based Learning and Registered Apprenticeships* – Mike Russo, National Institute for Innovation and Technology

1:55-2:00 Q&A

2:00-2:10 *Creating Opportunities Everywhere: STEM Investments in Education and Workforce Development* – James L. Moore III, National Science Foundation

2:10-2:15 Q&A

2:15-2:25 *Empowering Tomorrow's Workforce: The Role of AI in Smart Manufacturing* – Molly Dodge, Ivy Tech Community College

2:25-2:30 Q&A

2:30-2:50 Guided and moderated discussion with the session's speakers and audience – Alan Mitchell, Sandia National Laboratories, moderator

2:50-3:00 Break

Electronic design automation and ML

3:00-3:40 *AI for Chip Design/EDA: Everything, Everywhere, and All at Once (?)* – David Pan, University of Texas at Austin

3:40-4:20 *AI: Trailblazing the Next Generation of Semiconductor Innovation* – Vishal Khandelwal, Synopsys

4:20-4:30 Break

| | |
|---|---|
| 4:30-5:10 | *ChipNeMo: Domain-Adapted LLMs for Chip Design* – Mark Ren, NVIDIA |
| 5:10-5:25 | Guided and moderated discussion with the session's speakers and audience – Ben Feinberg, Sandia National Laboratories, moderator |
| 5:25 | Workshop wrap up and outlook – Conrad James, Sandia National Laboratories |

# APPENDIX B: Attendees

| NAME<br>AFFILIATION | | NAME<br>AFFILIATION |
|---|---|---|
| Akopyan, Filipp<br>IBM | | Diaz, Jacquelin<br>Sandia National Laboratories |
| Addamane, Sadhvikas<br>Sandia National Laboratories | | Dodd, Mark<br>Sandia National Laboratories |
| Agarwal, Sapan<br>Sandia National Laboratories | | Dodge, Molly<br>Ivy Tech Community College |
| Aimone, Brad<br>Sandia National Laboratories | | Eugenio, Evercita<br>Sandia National Laboratories |
| Allemang, Chris<br>Sandia National Laboratories | | Feinberg, Ben<br>Sandia National Laboratories |
| Bentz, Brian<br>Sandia National Laboratories | | Fischer, Art<br>Sandia National Laboratories |
| Bogart, Greg<br>Sandia National Laboratories | | Flores Mendoza, Irma T.<br>Sandia National Laboratories |
| Bojja, Ram<br>Sandia National Laboratories | | Frederick, Stephanie<br>National Institute for Innovation and Technology |
| Borna, Amir<br>Sandia National Laboratories | | Fritzemeier, Ron<br>Intel |
| Briggs, Ron<br>Sandia National Laboratories | | Fuller, Elliot<br>Sandia National Laboratories |
| Cardwell, Suma<br>Sandia National Laboratories | | Gnecco, Austin<br>Sandia National Laboratories |
| Chakrabarty, Krishnendu<br>Arizona State University | | Hinkle, Christopher<br>University of Notre Dame |
| Chance, Frances<br>Sandia National Laboratories | | Hoekstra, Rob<br>Sandia National Laboratories |
| Chapman, William<br>Sandia National Laboratories | | Holder, Aaron<br>DOE – Office of Science |
| Chen, Elton<br>Sandia National Laboratories | | Huq, Ashfia<br>Sandia National Laboratories |
| Chung, Helen<br>Sandia National Laboratories | | Incorvia, Jean Anne<br>University of Texas at Austin |
| Crowder, Cale<br>Sandia National Laboratories | | Jacobs-Gedrim, Robin<br>Sandia National Laboratories |
| Dasu, Aravind<br>Intel | | James, Conrad<br>Sandia National Laboratories |
| DeAntonio, Michael<br>Sandia National Laboratories | | Jarmon, Chris<br>Sandia National Laboratories |

| NAME<br>AFFIIATION | | NAME<br>AFFILIATION |
|---|---|---|
| Jimenez, Nick<br>Sandia National Laboratories | | Pan, David<br>University of Texas at Austin |
| Joshi, Siddarth<br>University of Notre Dame | | Pan, Wei<br>Sandia National Laboratories |
| Kaplar, Bob<br>Sandia National Laboratories | | Park, Minseong<br>University of Virginia |
| Kennedy, Valinda<br>IBM | | Parpillon, Benjamin<br>Fermilab |
| Khandelwal, Vishal<br>Synopsys | | Perry II, Derrick<br>Sandia National Laboratories |
| Kim, Seung Ju<br>University of Southern California | | Pino, Robinson<br>DOE – Office of Science- ASCR |
| Kulkarni, Jaydeep<br>University of Texas at Austin | | Querlioz, Damien<br>Université Paris – Saclay |
| Kumar, Prabhat<br>Lawrence Berkeley National Laboratory | | Quintana, Jasent<br>Sandia National Laboratories |
| Madireddy, Sandeep<br>Argonne National Laboratory | | Ren, Haoxing (Mark)<br>NVIDIA |
| Mailhiot, Christian<br>Sandia National Laboratory | | Russo, Mike<br>National Institute for Innovation and Technology |
| Marsiske, Helmut<br>DOE – Office of Science- HEP | | Sawant, Saurabh<br>Lawrence Berkeley National Laboratory |
| Meinelt, Zach<br>Sandia National Laboratories | | Schneider, Michael<br>NIST |
| Mendis, Charith<br>University of Illinois Urbana-Champaign | | Schuman, Catherine<br>University of Tennessee |
| Miner, Nadine<br>Sandia National Laboratory | | Seagrave, Dorian<br>Sandia National Laboratories |
| Mitchell, Alan<br>Sandia National Laboratories | | Severa, William<br>Sandia National Laboratories |
| Mitra, Subhasish<br>Stanford University | | Sharps, Paul<br>Sandia National Laboratories |
| Nguyen, Richard<br>Sandia National Laboratories | | Sickafoose, Shane<br>Sandia National Laboratories |
| Nordquist, Chris<br>Sandia National Laboratories | | Smith, Darby<br>Sandia National Laboratories |
| Nowlin, Nathan<br>Sandia National Laboratories | | Song, Wenhao<br>University of Southern California |

| NAME<br>AFFILIATION | | NAME<br>AFFILIATION |
|---|---|---|
| Strukov, Dmitri<br>University of California, Santa Barbara | | Wixom, Ryan<br>Sandia National Laboratories |
| Swaminathan, Madhavan<br>Penn State | | Wondra, Nicholas<br>Sandia National Laboratories |
| Taylor, Brady<br>Sandia National Laboratories | | Xiao, Patrick<br>Sandia National Laboratories |
| Tsao, Jeff<br>Sandia National Laboratories | | Young, Aaron<br>Oak Ridge National Laboratory |
| Tumeo, Antonino<br>Pacific Northwest National Laboratory | | Young, Andrew<br>Sandia National Laboratories |
| Vetter, Jeff<br>Oak Ridge National Laboratory | | Yu, Shimeng<br>Georgia Institute of Technology |
| Vineyard, Craig<br>Sandia National Laboratories | | Zhang, Xin<br>IBM |
| Wakeland, Anna<br>Sandia National Laboratories | | Zhao, Ruoyu<br>University of Southern California |