Neuromorphic Devices & AI Hardware Accelerators

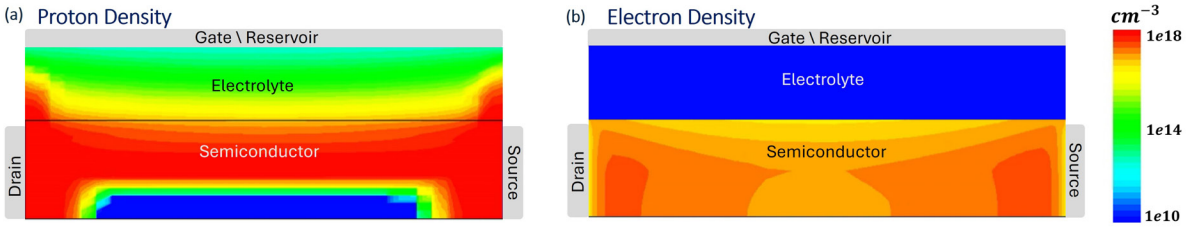
Electronic-protonic Conduction in Electrochemical RAM

S. Bitton, J. A. del Alamo
Sponsorship: MIT-IBM Watson AI Lab, Fulbright Fellowship, Intel International Science and Engineering Fair Fellowship, Zuckerman STEM Leadership Program, Schmidt Israeli Women's Postdoctoral Fellowship

Electrochemical random-access memory (ECRAM) is a promising candidate for analog in-memory neural network accelerators. Unlike traditional digital memory, which stores binary states (0 or 1), ECRAM can store a continuous range of values by modulating the conductivity of a semiconductor channel. This enables efficient analog computation directly within memory, reducing data movement between memory and processor, lowering energy consumption, and enhancing performance for artificial intelligence workloads.

The non-volatile analog states in ECRAM are achieved through proton (i.e., hydrogen cations) intercalation from a hydrogen reservoir, through an electrolyte, into a semiconductor channel. As protons accumulate in the semiconductor channel, they attract electrons, which increases the channel's conductivity. While this basic mechanism is known, the underlying electronic–protonic interactions that govern device behavior are not yet fully understood. These interactions are key to unlocking the full potential of ECRAM for neuromorphic computing.

In this work, we investigate these mechanisms using a combination of advanced two-dimensional (2D) device simulations and targeted experiments. Our simulation framework provides a theoretical foundation that complements the experimental results, offering insights into proton dynamics and their effect on conductivity. This integrated approach helps identify the key factors that influence device performance and guides strategies for further optimization.



▲ Figure 1: Simulated 2D ECRAM structure based on a WO_3 channel showing the spatial distribution of (a) proton density and (b) electron density after 50 potentiation pulses.

FURTHER READING:

- M. Onen, N. Emond, B. Wang, D. Zhang, F. M. Ross, J. Li, B. Yildiz, and J. A. del Alamo, "Nanosecond Protonic Programmable Resistors For Analog Deep Learning," *Science*, vol. 377, pp. 539-43, 2022.
- M. Onen, J. Li, B. Yildiz, and J. A. Del Alamo, "Dynamics of PSG-Based Nanosecond Protonic Programmable Resistors for Analog Deep Learning," *2022 International Electron Devices Meeting (IEDM)*, vol. 2, no. 6, pp. 1-4, 2022.

Impact of Annealing on Ferroelectric Properties of HZO for Non-volatile Memory

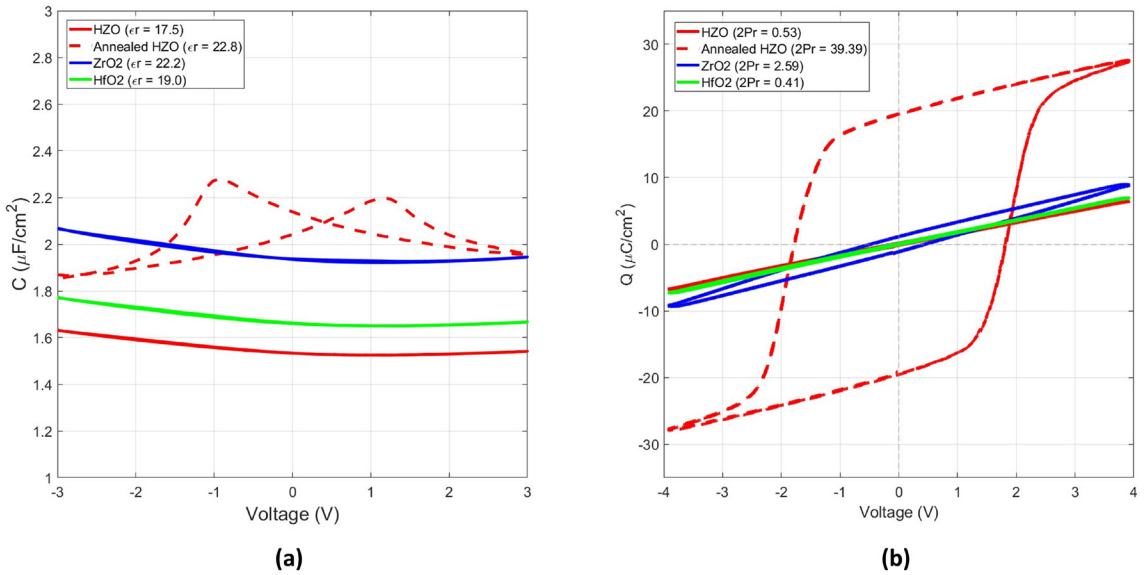
H. Choi, J. C.-C. Huang, Y. Shao, T. E. Espedal, D. A. Antoniadis, J. A. del Alamo
Sponsorship: SRC

Ferroelectric (FE) materials, particularly hafnium oxide (HfO_2)-based thin films, have attracted significant interest for complementary metal-oxide-semiconductor-compatible non-volatile memory technologies. In this study, we examine how thermal annealing affects the FE properties of HZO thin films compared to unannealed films (HZO, pure HfO_2, and ZrO_2). We fabricated metal-insulator-metal structures using plasma-enhanced atomic layer deposition for the insulator and sputtering for the tungsten electrodes.

Capacitance-voltage (C-V) and charge-voltage (Q-V) measurements at 100 kHz reveal clear differences between annealed and unannealed films. Annealed HZO samples show distinct butterfly-shaped C-V loops (Figure 1a), indicating robust FE behavior. The strong hysteresis in the Q-V characteristics (Figure 1b) demonstrates a remnant polarization (2Pr) of approximately 39.4 μC/cm², where each polarization state represents a distinct memory state suitable for non-volatile memory.

Unannealed HZO, HfO_2, and ZrO_2 films exhibit relatively featureless C-V characteristics but subtle yet measurable hysteresis in their Q-V characteristics, suggesting previously over-looked polarization. Although significantly smaller than annealed HZO, the observed remnant polarization (below 3 μC/cm²) challenges the assumption that these materials lack ferroelectricity without annealing.

These findings highlight that annealing greatly enhances polarization stability and ferroelectricity in HZO thin films, guiding the development of efficient, high-performance FE memory devices. Additionally, unannealed HfO_2 and ZrO_2 films may also be useful for applications requiring minor polarization effects. Future research should explore other activation methods to optimize FE polarization properties.



▲ Figure 1: (a) C-V loops and (b) Q-V loops of metal/ferro/metal with different FE materials.

FURTHER READING:

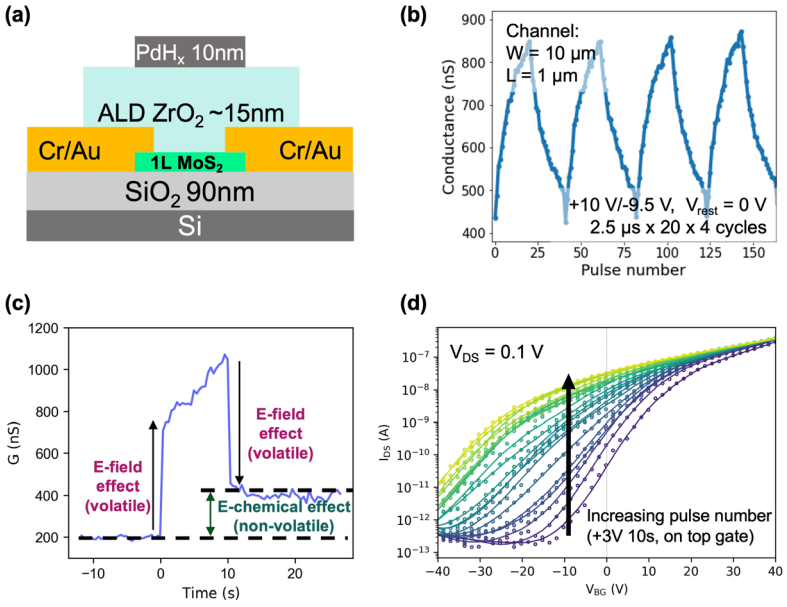
- T. Kim, J. A. del Alamo, and D. A. Antoniadis, "Switching Dynamics in Metal–Ferroelectric HfZrO2–Metal Structures," *IEEE Transactions on Electron Devices*, vol. 69, no. 7, pp. 4016-4021, July 2022. DOI: 10.1109/TED.2022.3175444
- T. Kim, J. A. del Alamo, and D. A. Antoniadis, "Dynamics of $HfZrO_2$ Ferroelectric Structures: Experiments and Models," *2020 IEEE International Electron Devices Meeting (IEDM)*, vol. 21, no. 4, pp. 1-4, 2020. DOI: 10.1109/IEDM13553.2020.9372013

Atomistic Simulations on Ion Incorporation in 2D Channel Materials for Fast Conductivity Modulation in Electrochemical Random-access Memory Devices

V. Fotopoulos, M. Siebenhofer, M. Huang, L. Xu, B. Yildiz
Sponsorship: SRC

Electrochemical random-access memory (ECRAM) has emerged as a novel type of pro-grammable resistor for crossbar arrays—a promising architecture for implementing energy-efficient artificial neural networks. ECRAMs consist of three key functional layers: an ion reservoir, a solid electrolyte, and a channel (Figure 1(a)). Through voltage-driven intercalation of mobile ions (e.g., H^+), the electronic conductivity of the channel can be finely modulated, enabling precise control over the resistance state of the device. Mixed ionic and electronic conducting oxides, e.g., WO_3, have been investigated as channel materials. However, their bulk nature necessitates three-dimensional ion redistribution, leading to undesirably long conductivity settling times. Two-dimensional (2D) materials, including monolayers of transition metal dichalcogenides (TMDs), such as MoS_2, offer a promising alternative, yet the mechanisms of interfacial ion transport and their impact on the conductivity of the channel remain underexplored.

In this work, we employed density functional theory to investigate hydrogen (H) incorporation at interfaces between a solid electrolyte (SiO_2) and a monolayer (MoS_2) channel (Figure 1(b)). We explored a range of SiO_2 surface chemistries—including surfaces with unsaturated Si and O dangling bonds and reconstructed surfaces with fully saturated bonds—and showed that surface termination determines the most stable H incorporation sites. For defect-free MoS_2, H is stable on MoS_2 only when the underlying SiO_2 surface is fully saturated (Figure 1(c)(i)), increasing the channel's conductivity through n-type doping (Figure 1(d)(i)). In contrast, when unsaturated O and/or Si dangling bonds are present (Figure 1(c)(ii)), H preferentially binds to the electrolyte surface, remaining electronically decoupled from MoS_2 (Figure 1(d)(ii)). Additionally, introducing a sulfur vacancy (V_s) in MoS_2 alters this behavior: across all surfaces, H stabilizes inside the vacancy, leading to n-type doping. These findings highlight how interfacial structure and defect engineering can enhance ionic modulation of conductivity in 2D ECRAMs.



▲ Figure 1: (a) ECRAM device. (b) MoS_2/SiO_2 interface. (c) (i) At fully saturated interfaces, H is stable on MoS_2. (ii) In interfaces with dangling bonds, H is stable on SiO_2. (d) (i) Density of states (DOS) of MoS_2 in saturated interface. (ii) DOS of MoS_2 in interface with dangling bonds.

FURTHER READING:

- A. A. Talin, J. Meyer, J. Li, M. Huang, M. Schwacke, H. W. Chung, L. Xu, E. J. Fuller, Y. Li, and B. Yildiz, "Electrochemical Random-Access Memory: Progress, Perspectives, and Opportunities," *Chemical Reviews*, vol. 125, no. 4, pp. 1962-2008, 2025.
- M. Schwacke, P. Žguns, J. A. del Alamo, J. Li, and B. Yildiz, "Electrochemical Ionic Synapses with Mg2+ as the Working Ion," *Advanced Electronic Materials*, vol. 10, no. 5, p. 2300577, 2024.
- M. Huang, M. Schwacke, M. Onen, J. A. del Alamo, J. Li, and B. Yildiz, "Electrochemical Ionic Synapses: Progress and Perspectives," *Advanced Materials*, vol. 35, no. 37, p. 2205169, 2023.

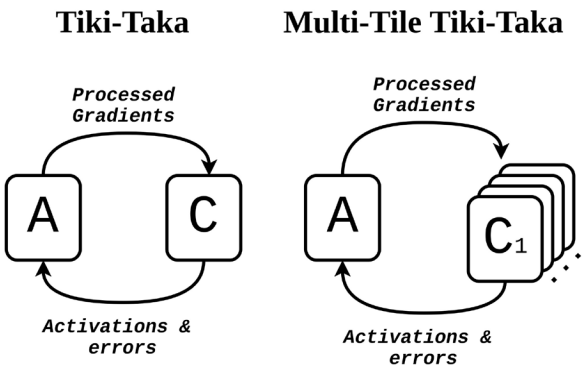
Accelerated Analog In-memory Computing for Neural Network Training

I. J. Gallo, J. A. del Alamo
Sponsorship: MIT-IBM Watson AI Lab

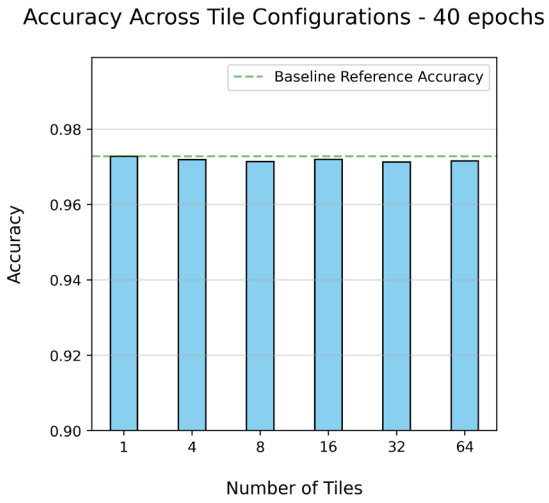
Neural network training demands enormous computational resources, leading to high energy consumption and long processing times. Analog in-memory computing offers a promising solution by performing matrix-vector multiplications directly within memory arrays, leveraging the physical properties of analog devices such as protonic synapses. However, a fundamental limitation of analog crossbar arrays is their inability to process multiple inputs simultaneously.

We propose a multi-tile parallel processing architecture that accelerates existing algorithms such as Tiki-Taka by introducing parallelism at the algorithmic level, as shown in Figure 1. Our approach distributes computation across multiple crossbar arrays, enabling the parallel processing of inputs while maintaining high accuracy. Using IBM's AIHWKIT simulation frame-work, we demonstrate that this parallel architecture achieves comparable accuracy to conventional sequential implementations while significantly reducing training time.

As shown in Figure 2, our parallel implementation maintains over 97% accuracy on the Modified National Institute of Standards and Technology (MNIST) dataset while achieving substantial speedup compared to standard Tiki-Taka implementations. By combining parallelization with analog in-memory computing, our approach delivers both dramatic improvements in energy efficiency over conventional digital methods and significantly accelerated training times.



▲ Figure 1: Architecture diagram of the pro-posed multi-tile parallel processing approach for accelerated Tiki-Taka.



▲ Figure 2: Comparison of classification accuracy in MNIST between standard and parallel implementations, demonstrating maintained performance despite parallelization.

FURTHER READING:

- M. Onen, T. Gokmen, T. K. Todorov, T. Nowicki, J. A. del Alamo, J. Rozen, W. Haensch, and S. Kim, "Neural Network Training with Asymmetric Crosspoint Elements," *Frontiers in Artificial Intelligence*, vol. 5, Article 891624, 09 May 2022. DOI: 10.3389/frai.2022.891624

Predicting Energy Materials Properties with Artificial Intelligence

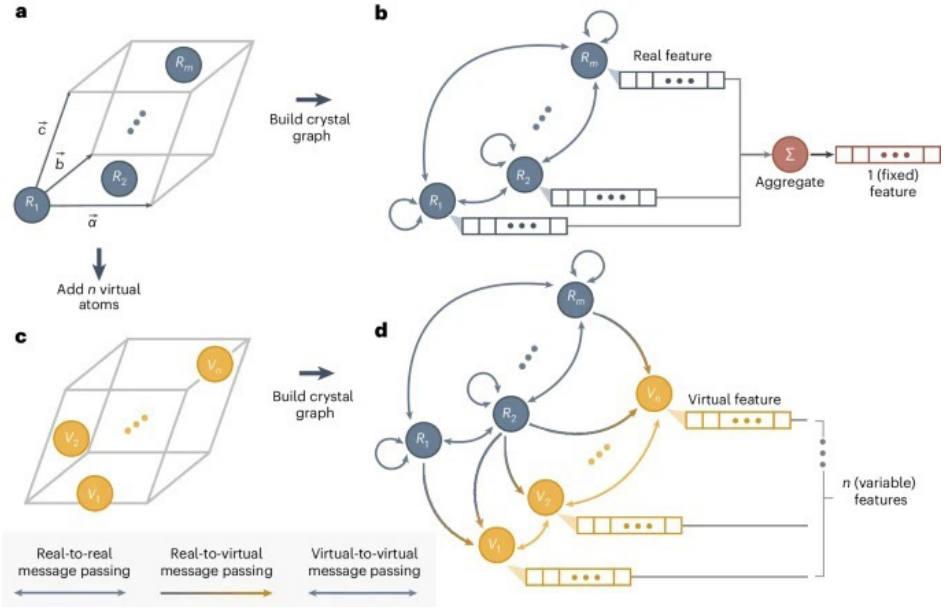
R. Okabe, A. Chotrattanapituk, M. Li
Sponsorship: NSF, Department of Energy

Accurate prediction of materials properties is essential for the discovery and design of next-generation energy materials. While first-principles calculations provide reliable insights, they are often computationally intensive, particularly for complex properties such as optical response and phonon spectra. Recent advances in machine learning offer promising alternatives, but key challenges remain in embedding atomic structures and handling variable-length outputs across materials.

To address these challenges, we introduce two complementary artificial intelligence (AI) frameworks tailored for property prediction: GNNOpt and the virtual node graph neural network (VGNN). GNNOpt is an equivariant graph neural network designed to predict optical properties—including absorption, refractive index, and reflectance—directly from crystal structures. It leverages universal atomic embeddings and the Kramers-Krönig relations to deliver accurate predictions on a dataset of only 944 materials, demonstrating strong agreement with first-principles calculations. Applications include screening photovoltaic candidates by spectroscopic efficiency and discovering topological quantum materials such as SiOs with high quantum weight.

In parallel, VGNN addresses the challenge of predicting phonon-related properties, which often have materials-dependent dimensionality. By incorporating virtual nodes into the crystal graph, this approach enables efficient and accurate prediction of Γ-point phonon spectra and full phonon dispersion relations. With significantly reduced computational cost and high accuracy, VGNN has generated large-scale databases, including over 146,000 Γ-phonon entries and phonon band structures for zeolites.

Together, these AI models demonstrate how tailored GNNs can achieve scalable, flexible, and high-fidelity prediction of energy-relevant properties, accelerating the discovery and design of functional materials for energy conversion, transport, and storage applications.



▲ Figure 1: A GNN processes a crystalline material with m atoms per unit cell, where each atom is a real node. After message passing, local features are aggregated into a fixed-size output. To enable flexible outputs, n virtual nodes are added to the crystal graph, allowing representations of variable length not limited to real-node aggregation.

FURTHER READING

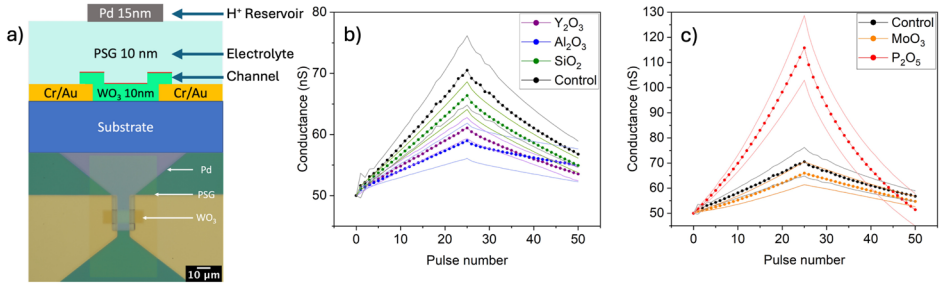
- R. Okabe, A. Chotrattanapituk, A. Boonkird, N. Andrejevic, X. Fu, T. S. Jaakkola, Q. Song, T. Nguyen, N. C. Drucker, S. Mu, B. Liao, Y. Cheng, and M. Li, "Virtual Node Graph Neural Network for Full Phonon Prediction," *Nature Computational Science*, vol. 4, p. 522, 2024.
- N. T. Hung, R. Okabe, A. Chotrattanapituk, and M. Li, "Universal Ensemble-Embedding Graph Neural Network for Direct Prediction of Optical Spectra from Crystal Structures," *Advanced Materials*, vol. 36, p. 2409175, 2024.

Oxide Interface Coatings in Proton-based Electrochemical Ionic Synapse Devices

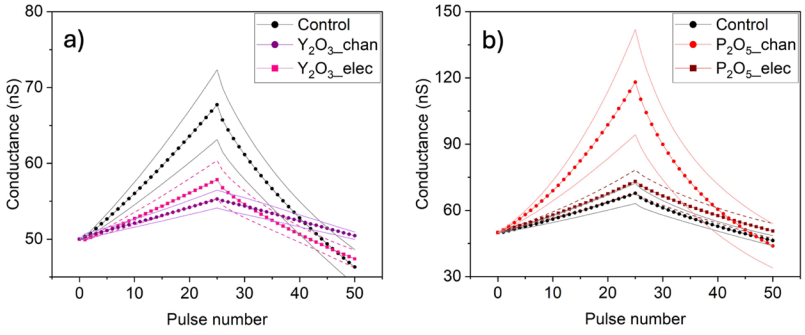
J. Meyer, B. Yildiz
Sponsorship: Department of Defense National Department of Science and Engineer-ing Graduate Fellowship (2022), Department of Energy EFRC Hydrogen in Energy and Information Sciences

Electrochemical ionic synapses (EIS) that modulate the electronic conductivity of a channel by ion intercalation are promising devices for use in energy-efficient neuromorphic computing hardware. Using protons as the working ions enhances the energy-efficiency and programming speed, but proton transfer through the electrolyte and interfaces of the device faces kinetic limitations. This presents a challenge for achieving nanosecond programming at 1 V or less.

Here, thin binary oxide coatings are deposited at the electrolyte-channel (phosphosilicate glass (PSG)-WO_3) interface to modify the interface chemistry and EIS device operation. Figure 1 shows tunable conductance change of the channel in response to voltage gating, depending on the interface oxide. The P_2O_5 interface coating markedly enhances the conductance change obtained under cycling, relative to both the other oxide-coated and unmodified (control) EIS devices. The position of the oxide coating within the device structure, in the middle of the PSG electrolyte versus at the electrolyte-channel interface, also impacts the magnitude of the effect on conductance change (Figure 2). The P_2O_5 and Y_2O_3 coatings increase and decrease the conductance change, respectively, to a greater degree when deposited at the interface. These results inform interface design for EIS devices with greater conductance change per voltage pulse, which can enable lower voltages as the programming speed increases towards the nanosecond scale.



▲ Figure 1: a) EIS schematic and image. (b-c) Cycling of oxide-coated EIS with 10-nm PSG with ±3 V, 100-ms pulses, 25 positive and 25 negative voltage pulses, showing mean and standard deviation for each device chemistry.



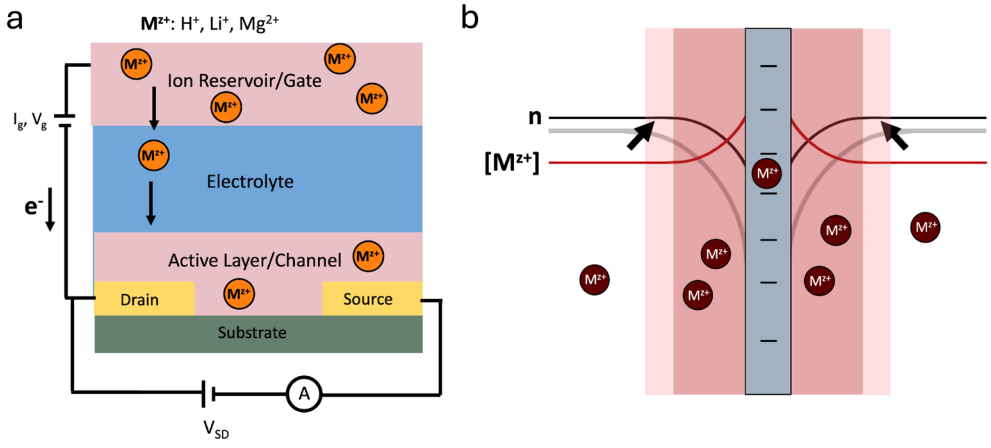
▲ Figure 2: Cycling of a) Y_2O_3-coated and b) P_2O_5-coated EIS with 20-nm PSG with ±4 V, 1 s pulses. Position of oxide coating is labeled for deposition in middle of electrolyte (_elec) versus at electrolyte-channel interface (_chan).

Understanding the Role of Space Charge Resistances in ECRAM

M. Schwacke, M. Siebenhofer, T. Defferriere, H. Tuller, B. Yildiz
Sponsorship: IBM, MIT School of Engineering Mathworks Fellowship

With the rapid rise in the prevalence of artificial intelligence, the energy consumed by training neural networks is also skyrocketing. Moreover, the energy consumed each year by computing is exponentially increasing and rapidly approaching the world's total energy production, making finding more energy efficient methods of computing imperative. Electrochemical random-access memory (ECRAM) is a promising technology for brain-inspired computing and for energy efficient training of neural networks. ECRAM devices act as programmable resistors, where the resistance of a channel material is programmed by electrochemically controlled intercalation of small cations into or out of the channel. A schematic of an ECRAM device appears in Figure 1a. It is generally thought that resistance modulation occurs due to changes in the bulk carrier concentration of the channel, as electrons must accompany cation intercalation to maintain charge neutrality. However, polycrystalline thin films, which are generally used as ECRAM channels, can have many sources of resistance beyond the bulk, including resistances arising from electron depletion in space charge regions at grain boundaries, contacts, and the electrolyte/channel or channel/substrate interfaces.

We use electrochemical impedance spectroscopy to characterize the contributions of bulk, grain boundary, and contact resistances to the total resistance of sputtered, polycrystalline WO_3 thin films, a common ECRAM channel material, before and after various levels of Mg^{2+} intercalation. We find that space charge resistances actually dominate the total resistance of the films and are also modulated by ion intercalation. To understand the mechanism by which space charge resistances are modulated, we develop an electrostatics model of space charge regions. The modeling results suggest that several mechanisms exist by which small concentrations of mobile cations could dramatically reduce the degree of electron depletion in space charge regions, including by cation accumulation in space charge regions and cation insertion directly into grain boundary or interfacial cores, as Figure 1b shows. This work has important implications for understanding the operating mechanisms of ECRAM. It also opens new avenues for informed device design, including by controlling grain size and interfacial chemistries.



▲ Figure 1: Schematic depictions of (a) an ECRAM device and (b) space charge regions adjacent to grain boundaries.

FURTHER READING

- A. A. Talin, J. Meyer, J. Li, M. Huang, M. Schwacke, H. W. Chung, L. Xu, E. J. Fuller, Y. Li, and B. Yildiz, "Electrochemical Random Access Memory: Progress, Perspectives, and Opportunities," *Chemical Reviews*, vol. 125, pp. 1962-2008, Feb. 2025.

Quantifying and Deconvoluting the Variability in Protonic Electrochemical Random-access Memories

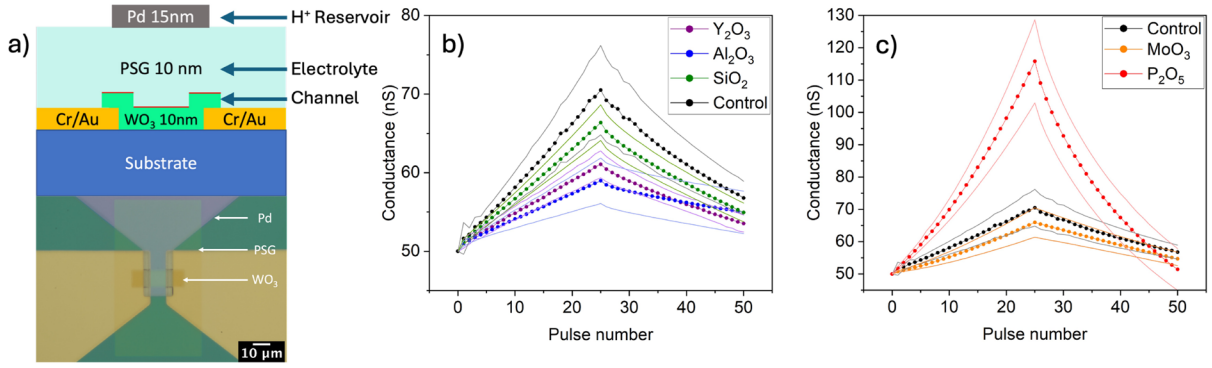
L. Xu, M. Huang, B. Yildiz
Sponsorship: SRC

Electrochemical random-access memory (ECRAM) is a novel programmable resistor candidate powering hardware neural networks (HNNs) based on cross-bar arrays, targeting fast and energy-efficient artificial intelligence (AI) training. Low device variability is crucial to high training accuracy in HNNs. ECRAMs are expected to have low variability compared to other candidates such as resistive random-access memory and phase-change memory, enabled by the deterministic dynamic doping of the channel. Some possible sources, such as the material micro-structure and non-ideal fabrication artifacts, can introduce variations in ECRAMs.

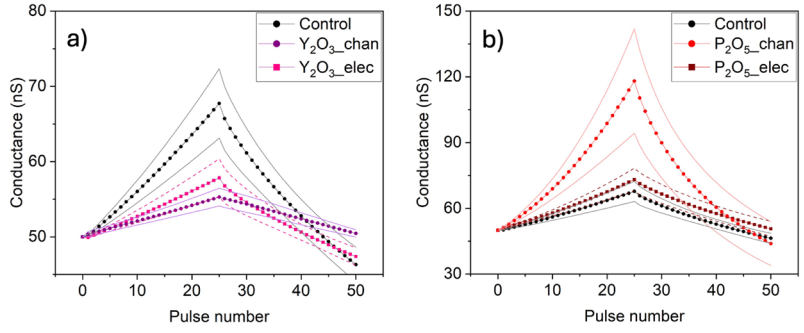
This work systematically quantified the variability of complementary metal-oxide-semiconductor-compatible protonic ECRAMs (PdH_x/HfO_2 or YSZ/WO_3 structure, Figure 1). By examining conductance modulation range and symmetry with over 1000 conductance states, we observed low variations in low-conductance regime (Figure 2). Device-to-device variation showed dependence neither on channel ordering (crystalline/amorphous), nor on channel sizes ranging from 102 μm² to 1502 nm².

Meanwhile, source/drain contact with the channel was investigated as a possible variation source. Correlation between the contact resistance contribution and device modulation behavior was observed. We explored contact improvement via hydrogen plasma treatment or using Ti transition layer and plan further evaluation to assess its impact on variability.

These findings confirm that ECRAM meets variability targets and demonstrates strong potential for downscaling, indicating it can be a promising candidate for programmable resistors. The results also emphasize the importance of microstructure and contact resistance control for consistent, low-variability operation.



▲ Figure 1: ECRAM device structure. (a) Cross section schematic of device structure. (b) Optical image of a device with 10^2-μm2 channel size.



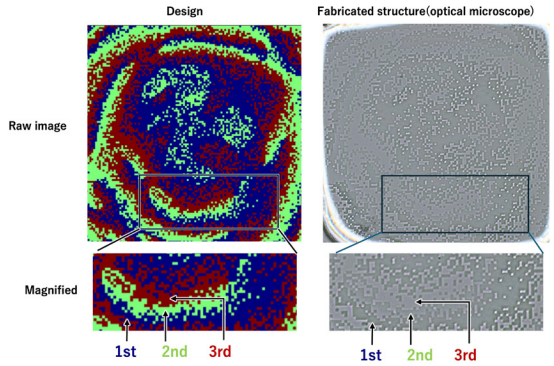
◀ Figure 2: Device variability. (a) ECRAM con-ductance modulation with 3V/-3.2V writing voltage. (b) Device-to-device variation of EC-RAM device with amorphous and crystalline channel, showing 15% variation.

Toward Fast, Nanoscale, and Accurate Fabrication of Diffractive Optical Neural Networks

D. Zhang, H. Kusaka, T. Nambara, Y. Kunai, G. Barbastathis
Sponsorship: Fujikura Ltd.

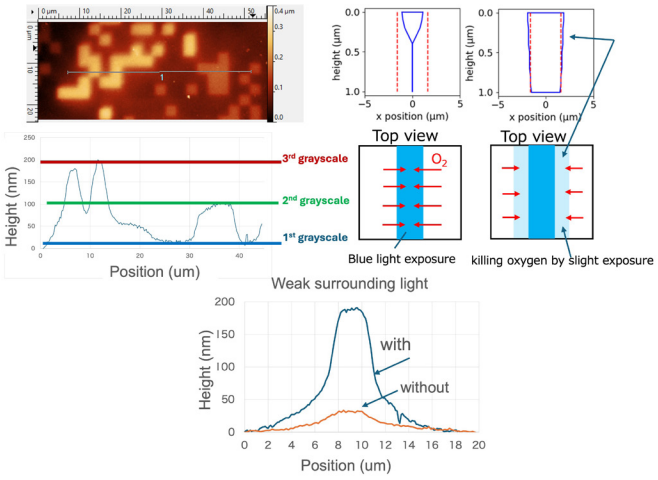
The fusion of machine learning and optics has driven advances in all-optical computing, with diffractive deep neural networks (D2NNs) emerging as a powerful architecture. By using deep learning to design diffractive layers, D2NNs can perform tasks like image classification at the speed of light, offering high parallelism and energy efficiency. However, D2NNs are typically fabricated at the macro scale, which limits their suitability for highly integrated devices. While some have been made at the micron scale using slow, point-by-point methods, these approaches remain unsuitable for large-scale deployment. Moreover, accurately fabricating these structures at wavelength-level resolution remains a major challenge. Achieving fast, scalable, and precise fabrication is critical for practical implementation.

We systematically address these fabrication challenges. For speed, we used a digital micromirror device as a dynamic mask to fabricate entire layer patterns at once, enabling layer-by-layer lithography rather than traditional point-by-point writing (Figure 1). To achieve the nanoscale, we demonstrated the ability to fabricate 2.5-μm pillar features using blue light only (Figure 2). We will also adopt two-color lithography to exceed the diffraction limit, achieving sub-wavelength (λ/2) resolution and enabling precise phase control. In terms of accurate fabrication, we discovered that non-local oxygen inhibition coupled with varying light intensity along the vertical axis due to diffraction can significantly degrade the height profiles of printed pillars. This effect results in inconsistent polymerization across depths, resulting in shape distortion and even fabrication failure. To explain this phenomenon, we modeled the underlying photopolymerization dynamics, capturing both chemical reactions and oxygen diffusion. The model successfully describes the observed behavior and provides a predictive framework. Based on this, we proposed an illumination strategy that introduces weak surrounding light to deplete oxygen, improving quality of printed pillars and enabling consistency across printed layers (Figure 2).



◀ Figure 1: Rapidly fabricated 3-layer structures observed under optical microscope.

▼ Figure 2: (Top left) AFM height profiles of 3-layer structures. (Top right) Simulation of coupling between non-local oxygen inhibition and diffracted light intensity profiles. Introducing weak surrounding light to pre-deplete oxygen effectively resolves this issue and enhances fabrication quality. (Bottom) Experimental validation of simulation results.

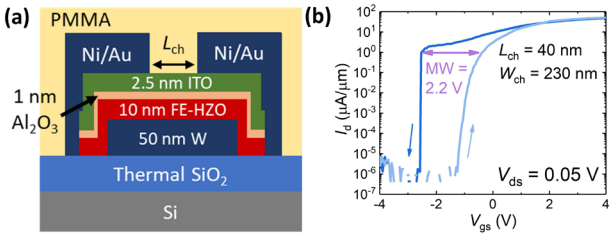


Discrete Domain Switching in Scaled Amorphous Metal-oxide Channel Ferroelectric FETs

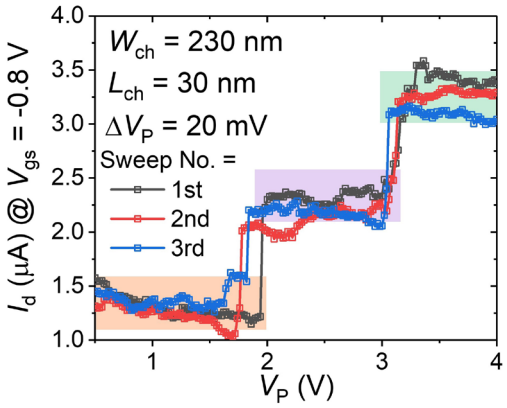
Y. Shao, E. Rafie Borujeny, J. Navarro Fidalgo, J. C.-C. Huang, T. E. Espedal, D. A. Antoniadis, J. A. del Alamo
Sponsorship: Intel Corporation, SRC

Understanding domain structures and domain switching mechanisms in ferroelectric (FE) $Hf_{0.5}Zr_{0.5}O_2$ (HZO) is crucial for its applications in non-volatile memory and analog hardware. Probing the actual size of the FE domains and mapping their individual polarization switching is challenging, but such information is highly valuable for HZO-based FE device design.

In this work, we integrate FE-HZO in complementary metal-oxide-semiconductor (CMOS)-compatible FE-FETs based on an amorphous oxide-semiconductor (AOS) channel (Figure 1a). Extensive electrical characterizations, including large-signal polarization-voltage, small-signal capacitance-voltage, and direct-current current-voltage characteristics, have been carried out on multiple device structures. A large memory window (MW) of 2.2 V @ 1 μA/μm is achieved with a scaled channel length of 30 nm (Figure 1b). Gate voltage pulses with in-creasing amplitudes are applied. After each pulse, channel current is read at a constant gate voltage. Discrete domain switching is observed in narrow devices with reproducible multi-level erasing/programming operations (Figure 2), whereas gradual switching is apparent in wider ones (Figure 2). Moreover, we show that discrete polarization switching acts as a sensitive probe to study intriguing physics in AOS-channel FE-FETs, such as FE fatigue. We observe that FE domain pinning, with domains stuck in the up-polarization state, leads to MW closure and negative threshold voltage shift. Based on a channel length scaling study, we estimate the av-erage FE domain size in our FE-HZO film to be ~40 nm. This work shows the rich physics and countless engineering opportunities in AOS-based FE devices.



▲Figure 1: (a) Schematic of CMOS-compatible FE-FET. (b) Hysteretic transfer characteristics of a highly scaled FE-FET.



▲Figure 2: Drain current obtained at a constant gate voltage as a function of applied positive gate pulse amplitude of a narrow device showing three discrete states.

FURTHER READING

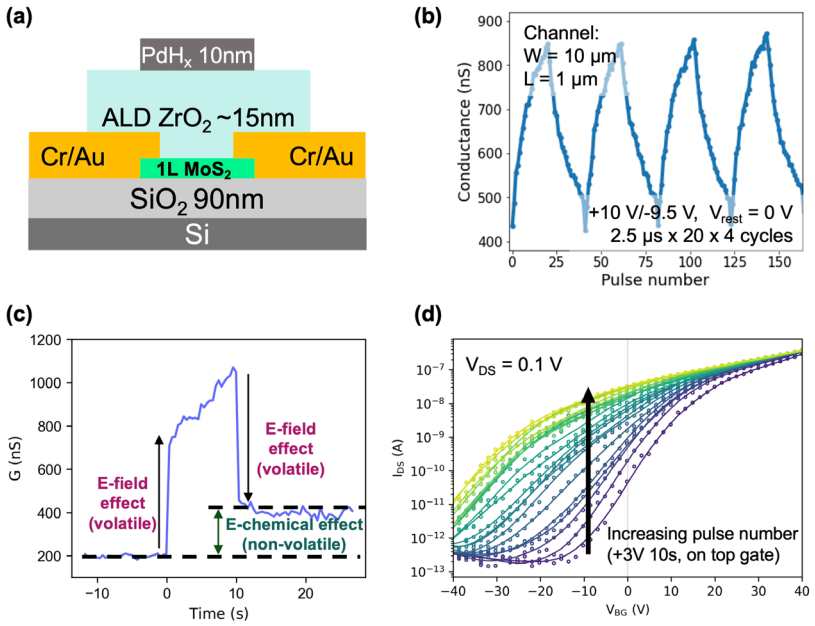
- Y. Shao, E. Rafie Borujeny, J. Navarro Fidalgo, J. C.-C. Huang, T. E. Espedal, D. A. Antoniadis and J. A. del Alamo, "Discrete Domain Switching in Scaled Oxide-Channel Ferroelectric FETs," *presented at 82nd Device Research Conference*, 2024.
- Y. Shao, E. Rafie Borujeny, J. Navarro Fidalgo, J. C.-C. Huang, T. E. Espedal, D. A. Antoniadis and J. A. del Alamo, "Discrete Ferroelectric Polarization Switching in Nanoscale Oxide-channel Ferroelectric Field-effect Transistors," *Nano Letts.*, vol. 25, no. 8, pp. 3173-3179, Feb. 2025.

Electrochemical Random-access Memory with Monolayer MoS_2 Channels for Fast Conductivity Modulation and Dynamically Tunable Transistors

M. Huang, L. Xu, X. Zheng, J. Kong, B. Yildiz
Sponsorship: SRC

Electrochemical random-access memories (ECRAMs) are promising three-terminal programmable resistors for powering deep neural network hardware accelerators when arranged in crossbar arrays. They are three-terminal devices with an ion reservoir layer, a solid-state electrolyte layer, and a channel layer. The electronic conductivity of the channel can be modulated by electrochemical ion intercalation with good linearity, symmetry, and low variability.

Conventional ECRAMs using bulk channels could suffer from undesirable relaxation transients due to ion diffusion through their finite thickness. In this work, we investigate ECRAMs with a 2H-MoS_2 monolayer as the channel (Figure 1a). Our findings show that the supply of protons to the 2H-MoS_2 channel enables reversible non-volatile conductance modulation with microsecond voltage pulses (Figure 1b). The response of the device to the applied gate voltage exhibits both a non-volatile electrochemical effect and a volatile electric field effect (Figure 1c). When combined with the Si substrate as an electronic back-gate, a transistor structure forms at the back side, where the source-drain current is tunable via the back-gate voltages. Applying top-gate pulses induces a large threshold voltage shift, suggesting that the hydrogen supplied to the MoS_2 and its surrounding interfaces increases the n-type doping of the channel (Figure 1d) and allows for a large range (105) of non-volatile conductance modulation at a constant back gate voltage. Our findings offer a pathway to develop high-speed programmable resistors and dynamically tunable transistors with 2D channels for hardware neural networks and other in-memory computation architectures.



▲ Figure 1: (a) Schematic of ECRAM device with monolayer MoS_2 channel. (b) Channel conductance modulation with microsecond voltage pulses. (c) Volatile and non-volatile conductance modulation effect from applied gate voltage. (d) Source-drain current as function of back-gate voltage after increasing number of electrochemical gate voltage pulses applied to top, showing shift of threshold voltage towards lower voltage with hydrogen supplied to channel.

FURTHER READING

- A. A. Talin, J. Meyer, J. Li, M. Huang, M. Schwacke, H. W. Chung, L. Xu, E. J. Fuller, Y. Li, and B. Yildiz, "Electrochemical Random-Access Memory: Progress, Perspectives, and Opportunities," *Chemical Reviews*, vol. 125, no. 4, pp. 1962-2008, 2025.

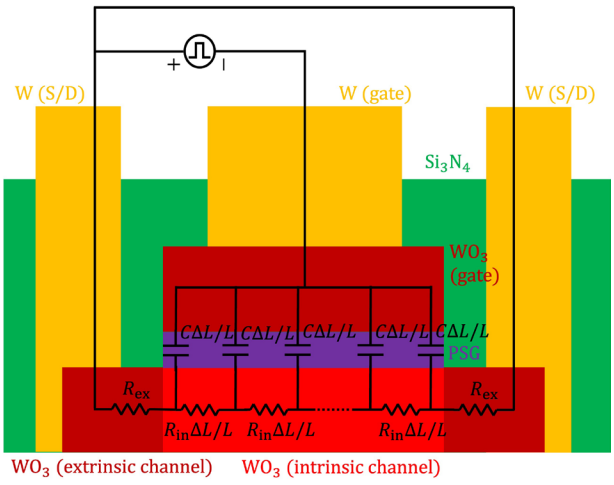
Dynamic Modeling of WO_3-PSG Protonic Devices for Analog Computing

D. Shen, J. A. del Alamo
Sponsorship: MIT-IBM Watson AI Lab

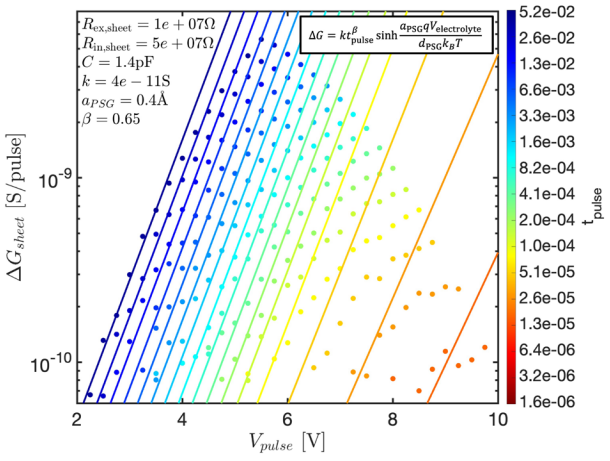
By leveraging local information processing through intrinsic physical properties of devices, analog computing presents a promising approach to overcome computational bottlenecks faced by traditional digital deep learning systems. One prominent strategy involves electrochemical ionic interactions in programmable resistors, where device resistance is adjusted by ionic exchange through an electrolyte. Previous research in our group has demonstrated proton-based non-volatile programmable resistors featuring a tungsten oxide (WO_3) channel, phosphorous-doped silicon dioxide (PSG) electrolyte, and palladium (Pd) gate reservoir. However, enhancing performance and making fabrication complementary metal-oxide-semiconductor (CMOS)-compatible remain crucial for application in future accelerators.

In this study, we optimized the device structure into a symmetric WO_3-PSG-WO_3 stack, making device fabrication fully compatible with standard CMOS fabrication processes. Device programmability is achieved by applying voltage pulses between the gate and source/drain terminals, thereby precisely controlling channel conductance.

To evaluate programming efficiency, we systematically characterized channel conductance responses to various pulse voltages and durations. A distributive resistor-capacitor (RC) model was developed to interpret experimental data and extract critical device parameters by accurately simulating voltage distribution across the electrolyte. This model successfully matched experimental results except under conditions of particularly high voltages or long pulse durations. These discrepancies might highlight the phenomenon of diffusion saturation. When the pulse width is much shorter than the diffusion time of protons inside the WO3, proton flow from the reservoir to the channel will become supply-limited. Diffusion saturation significantly restricts programming speed because conductance changes become proportional to the square root of pulse width. Consequently, improving ion diffusion rates within WO_3 emerges as a critical factor for further enhancing device performance.



▲ Figure 1: WO_3-PSG-WO_3 protonic device structure and schematic for distributive RC model. We use in-situ protonated WO_3 as both channel and gate reservoir, Si_3N_4 as encapsulation, and W as source/drain/gate contacts. The extrinsic channel is not gated and is heavily doped, while the intrinsic channel is gated and lightly doped.



▲ Figure 2: Channel conductance response to different pulse voltages and durations in a 5 μm-by-5 μm device with pulse setup and direction as in Figure 1. Colors indicate pulse durations. Dots show experimental data; lines show simulation results. With the fitting formula on the right, we extract the device parameters on the left.

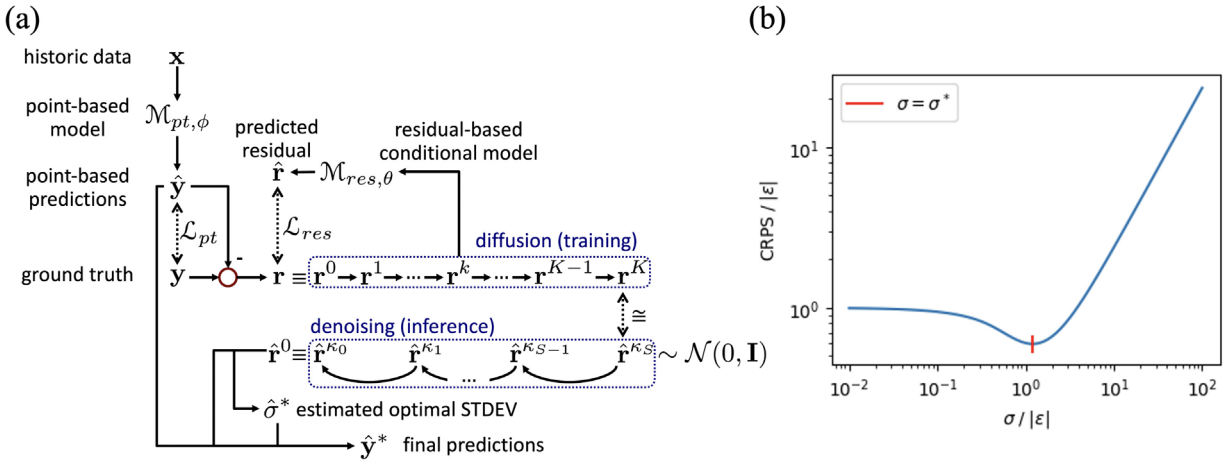
FURTHER READING

- M. Onen, N. Emond, B. Wang, D. Zhang, F. M. Ross, J. Li, B. Yildiz, and J. A. del Al-amo, "Nanosecond Protonic Programmable Resistors for Analog Deep Learning," *Science*, vol. 377, no. 6605, pp. 539-543, Jul. 2022.

RDIT: Residual-based Diffusion Implicit Models for Probabilistic Time Series Forecasting

C.-Y. Lai, D. S. Boning

We propose RDIT (Residual-based Diffusion Implicit modeling for probabilistic Time series forecasting), a novel framework designed to address key limitations in recent probabilistic time series forecasting (PTSF) methods. While traditional TSF models have increasingly adopted deep learning architectures, many of these rely on assumptions—such as linear mappings and channel independence—that are ill-suited for accurately modeling uncertainty. Additionally, standard PTSF approaches often conflate point estimation and noise modeling, resulting in reduced flexibility and suboptimal uncertainty quantification. To overcome these issues, RDIT separates the forecasting process into two stages: a plug-and-play model performs point-based prediction (estimating the conditional mean or median), while a conditional diffusion model captures the distribution of the residuals. This modular design not only improves adaptability across different domains but also allows the diffusion model to focus purely on learning noise characteristics. For fast inference, RDIT employs denoising diffusion implicit models (DDIM), significantly reducing sampling time compared to traditional diffusion models. Furthermore, we derive a theoretical formulation for optimizing the continuous ranked probability score (CRPS), a common metric in probabilistic forecasting, under the assumption of Gaussian-distributed errors. Based on this, we introduce an error-aware expansion mechanism that adjusts the learned distribution to better match the evaluation metric. Experimental results across 8 datasets and 6 forecasting horizons show that RDIT consistently outperforms state-of-the-art TSF and PTSF models in terms of accuracy, uncertainty calibration, and generation speed. Our work provides a practical and theoretically grounded approach to modeling uncertainty in time series forecasting, with potential applications in risk-sensitive domains such as finance, healthcare, environmental monitoring, and industrial process optimization.

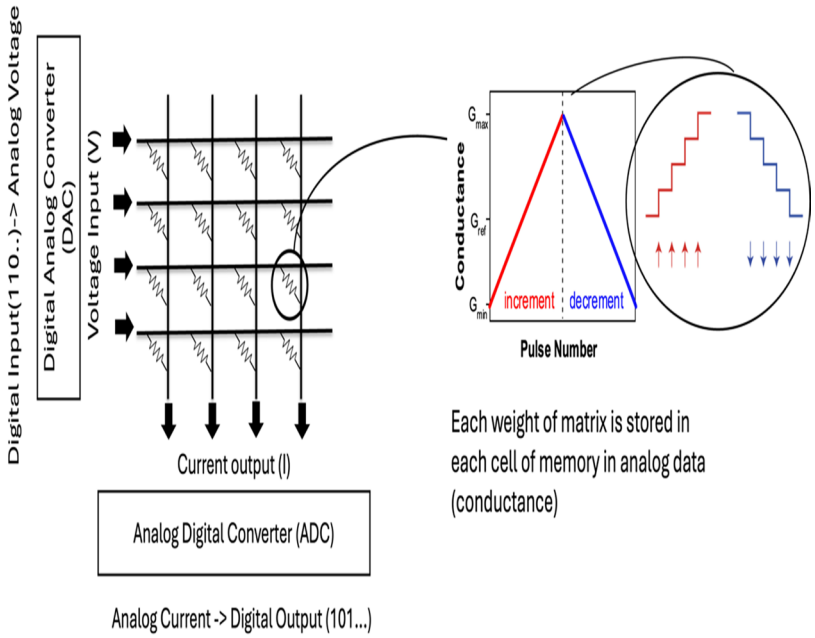


▲ Figure 1: (a) General scheme of this work. The point-based model is used to predict the point-based predictions; the residual-based conditional model is used to predict residuals from the k-th diffusion step residual while conditioning on the input and point-based prediction. (b) Normalized CRPS plotted against normalized standard deviation when the predictions are drawn from a normal distribution with mean and standard deviation, with the ground truth being y.

Design Considerations of Analog Accelerators for Machine Learning Applications

J. Lee, J. A. del Alamo
Sponsorship: IBM, Ericsson

Recently, there has been tremendous progress in machine learning, leading to a dramatic increase in its applications, such as image classification and natural language processing. As a result, there has been an explosion in demand for Graphics Processing Units and various accelerators that perform the computation required for machine learning training and inference. The widespread use of currently dominant digital accelerators requires a massive amount of energy, which is becoming a significant global issue. In response, analog computing using devices such as protonic synapses and ReRAM has been proposed as an alternative that can significantly enhance energy efficiency. Analog devices still face issues such as nonlinearity, asymmetry, and noise. Moreover, their performance heavily depends on components like Analog-to-Digital Converters (ADCs) and Digital-to-Analog Converters (DACs). In this work, we investigate how non-idealities degrade the performance of analog computing. We also evaluate different analog algorithms that mitigate performance degradation in several tasks such as Convolutional and Recursive Neural Networks with IBM AIHWKIT.



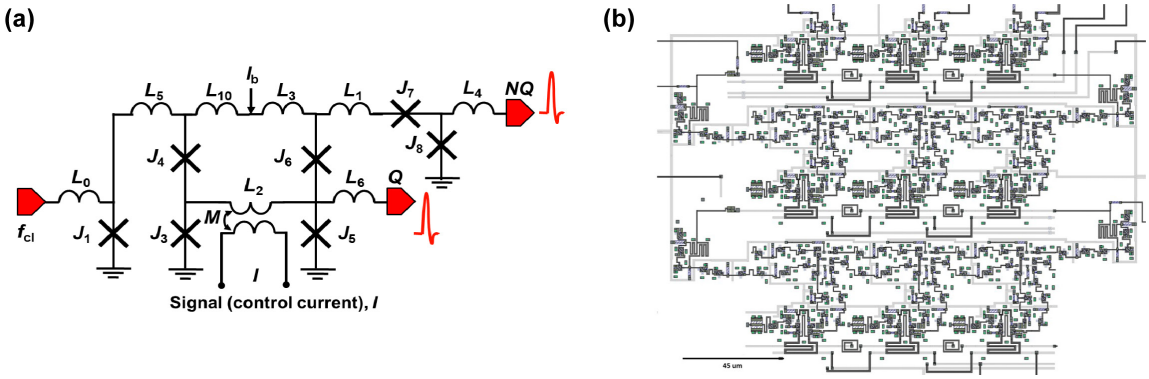
▲ Figure 1: Schematics of analog computing with analog crossbar array consisted by non-volatile memory cells.

Development of a Neuromorphic Network using BioSFQ Circuits

E. B. Golden, A. Qu, V. K. Semenov, K. K. Berggren, S. K. Tolpygo
Sponsorship: Air Force Contract No. FA8702-15-D-0001

Superconducting single-flux quantum (SFQ) circuits are promising candidates for neuromorphic hardware accelerators. They are extraordinarily fast and energy-efficient and use asynchronous pulse-rate data encoding, much like biological neurons. BioSFQ is an SFQ-based family that uses these neuromorphic features of SFQ circuits to process mixed analog/digital logic. BioSFQ circuits are also programmable, enabling mixed modes of operation and resilience to fabrication variation and flux trapping.

In this work, we design, fabricate, and measure a 3x3 network of bioSFQ comparators, the fundamental building block of bioSFQ circuits. We also demonstrate novel techniques for network calibration, integrating on-chip memory, and image processing using this 3x3 network.



▲ Figure 1: (a) Schematic of a bioSFQ comparator and (b) layout of the 3x3 comparator network.

CiMLoop: A Flexible, Accurate, and Fast Compute-In-Memory Modeling Tool

T. Andrulis, J. S. Emer, V. Sze
Sponsorship: Ericsson, TSMC, the MIT AI Hardware Program, MIT Quest, Samsung Semiconductor Fellowship, Siebel Scholars Fellowship

Compute-In-Memory (CiM) is a promising solution to accelerate Deep Neural Networks (DNNs) as it can avoid energy-intensive DNN weight movement and use memory arrays to perform low-energy, high-density computations. These benefits have inspired research across the CiM stack, but CiM research often focuses on only one level of the stack (i.e., devices, circuits, architecture, workload, or mapping) or only one design point (e.g., one fabricated chip). There is a need for a full-stack modeling tool to evaluate design decisions in the context of full systems (e.g., see how a circuit impacts system energy) and to perform rapid early -stage exploration of the CiM co-design space.

To address this need, we propose CiMLoop: an open-source tool to model diverse CiM systems and explore decisions across the CiM stack. CiMLoop introduces (1) a flexible specification that lets users describe, model, and map workloads to both circuits and architecture, (2) an accurate energy model that captures the interaction between DNN operand values, hardware data representations, and analog/digital values propagated by circuits, and (3) a fast statistical model that can explore the design space orders-of-magnitude more quickly than other high-accuracy models. Using CiMLoop, researchers can evaluate design choices at different levels of the CiM stack, co-design across all levels, fairly compare different implementations, and rapidly explore the design space.

Ultra-low Power Superconducting Electronics for Deep Learning Accelerator Architectures: Evaluating Energy Efficiency and Scalability

L. C. Blackburn, E. Golden, T. Andrulis, V. Sze, J. S. Emer, N. Gershenfeld, K. K. Berggren
Sponsorship: MIT Lincoln Laboratory, the MIT AI Hardware Program

Since the invention of the Josephson junction in the 1960s, superconducting electronics have shown promise for high-speed and energy-efficient computing. Since 2013, the Adiabatic Quantum Flux Parametron (AQFP) device has gained popularity for its ultra-low energy dissipation. AQFP inverters dissipate 10-^{21}J per switching event, 100 less than other superconductor logic, and 10^6X less energy than modern-day CMOS transistors or 10^3X when including the cryogenic cooling cost. As Moore's law ends and energy efficiency emerges as a limit on today's computing systems, superconducting AQFP logic is a promising technology to address these energy challenges.

Although individual AQFP device performance is impressive, superconducting electronics have failed to replace CMOS systems in the past in part due to the high cost of cryogenic low-noise testing environments and the limitations of superconductor memory scaling. To realize the promise of superconducting electronics, there is a need to architect full systems that can leverage the benefits of the unique superconductor physics (e.g., low-energy logic, low-energy interconnects on zero-resistance wires) while addressing the challenges (e.g., using low-noise cryogenic environments commoditized by the quantum computing industry, constructing a memory hierarchy that addresses the lack of a scalable, high-density superconducting memory).

In this work, we extend Timeloop/Accelergy accelerator modeling tools to support superconducting accelerators. This framework explores the design space of deep learning accelerator architectures with a toolbox of superconducting circuits from various logic families. We present results demonstrating the tradeoffs between superconductor vs. CMOS accelerators while running a range of deep learning workloads.

Single-Shot Matrix-Matrix Multiplication Optical Processor for Deep Learning

C. Luan, D. Englund, R. Hamerly
Sponsorship: NTT Research, TSMC, DARPA NaPSAC

The computational demands of modern AI have sparked interest in optical neural networks (ONNs), which offer the potential benefits of increased capacity and lower power consumption. Notable progress includes the demonstration of optical matrix-vector multiplication (MVM) processors based on cascaded Mach-Zehnder interferometer arrays using coherent light as the data carriers and thermo-optic phase shifters as weighting. Broadcast-and-weight MVM optical processors using different wavelengths as data carriers and tunable add-drop micro-ring resonators as weighting elements have also been demonstrated. Recent advancements in delocalized photonic deep learning also shows the advantages of using optical fan-out and analog time integrator based optical MVM processors on the Internet's edge. So far, limited by the low parallelism, most existing systems operate vector-vector multiplication (VVM) or MVM with O(N) or O(N^2) scaling in system throughput. To fully unlock the potential of optical computing, a parallel matrix-matrix multiplication (MMM) processor will allow better throughput and efficiency scaling than ordinary MVMs, but its realization is challenging due to the 3D data structure and high parallelism requirements.

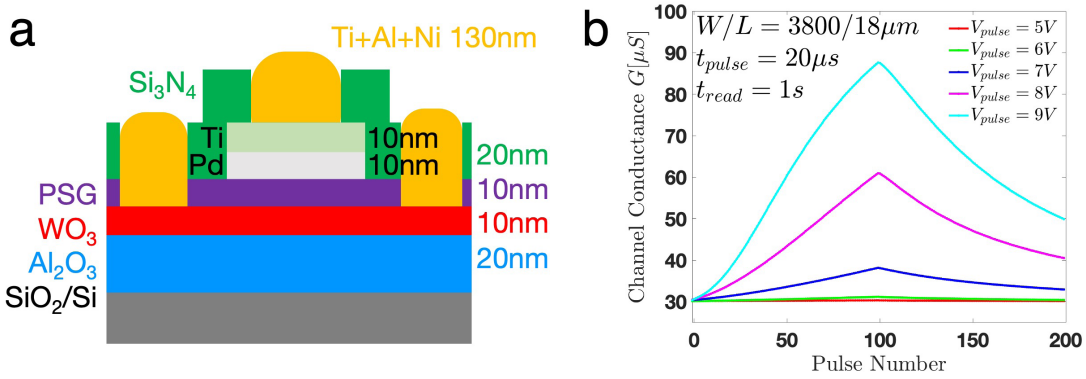
In this work, we propose and experimentally demonstrate a 3D grating-based ONN architecture using time-wavelength-spatial domain data flows with parallel operations in 16 space-degrees of freedom to improve the output capacity and energy efficiency. We experimentally demonstrate a parallel matrix-matrix multiplication processor using 4x4 input and output fiber arrays with 16 channel frequency comb lines of 7 different wavelengths, 32 broadband $LiNbO_3$ intensity modulators for weight matrix and input matrix encoding, a blazed reflective grating for low loss beam routing, and 16 analog time integrators for signal accumulation and network scaling, yielding a total operation-throughput of 64 MACs/shot with a high bit precision of 8-bits.

200 mm Wafer Diameter Process of Pd/PSG/WO3 Protonic Synapses for Analog Deep Learning

D. Shen, J. A. del Alamo
Sponsorship: MIT-IBM Watson AI Lab

To solve the overcoming computational bottlenecks for deep learning, analog deep learning accelerators process information locally using special-purpose devices for matrix multiplication calculations and outer product updates. Among them, Electrochemical Random-Access Memories modulate channel resistance by ionic exchange between a semiconductor channel and a gate reservoir via an electrolyte. This design aims to enable neural network training with enhanced energy efficiency, non-volatility, and low latency.

Our research focuses on proton-based ionic synapses featuring in-situ hydrogenated H_xWO_3 (as channel) and PdH_y (as gate reservoir), and phosphosilicate glass (PSG) (as electrolyte). In a close collaboration with IBM Research, we have fabricated devices on a 200-mm wafer from IBM Research using a CMOS back-end-of-line compatible process. We have successfully demonstrated linear and symmetrical channel conductance modulation under different voltage pulses across gate and channel.



▲ Figure 1: (a) Schematic of Pd/PSG/WO_3 protonic synapse fabricated in a joint MIT-IBM Research process on 200-mm wafers. (b) Conductance modulation under voltage pulses from 5 V to 9 V with a width of 20 μs that are fired every 1 s.

Tailor Swiftiles: Accelerating Sparse Tensor Algebra by Overbooking Buffer Capacity

Z. Y. Xue, Y. N. Wu, J. S. Emer, V. Sze
Sponsorship: MIT AI Hardware Program, Mathworks Fellowship, NSERC PGS-D

Many applications operate on tensor data that has high sparsity (i.e., many zeros) with large variations sparsity between regions of a tensor. Prior sparse tensor algebra accelerators partition the tensor into equal shape tiles that all fit in a buffer, limiting utilization of buffer resources for more sparse tiles. Our key insight is that we can overbook the buffer by allocating tiles that occasionally exceed the capacity of the buffer. We propose to combine a low-overhead data orchestration mechanism, Tailors, with a statistical tiling approach, Swiftiles, in order to support tiles that overbook the buffer and improve utilization of buffer resources and thus improve on-chip data reuse. Across a suite of 22 sparse tensor algebra workloads, we show that Tailors and Swiftiles introduce an average speedup 2.3x over an existing sparse tensor algebra accelerator with optimized tiling.

Stable and Accurate Nano-Resistor for Reliable Fixed AI Inference Tasks

G. Lee, M. Song, K. Kwon, J. Kim
Sponsorship: Samsung Semiconductor Fellowship (3965323)

As artificial intelligence (AI) technology continues to advance in everyday applications, there is a rising demand for seamless, private, and sophisticated AI functionalities. On-device AI solutions, embedded within commercial mobile devices, eliminate the need for external server communication, thereby enhancing response times and data privacy. However, existing on-device AI technologies are limited by substantial power consumption and insufficient computational capacity to support advanced generative AI models. Memristor-based analog AI accelerators have emerged as a potential solution to the von Neumann bottleneck, a major limitation in achieving greater speed and energy efficiency in AI computing. Despite their promise, memristors are hindered by issues with conductance state stability and the complexity of required programming algorithms and circuitry, which constrains their widespread adoption in industry. In this study, we introduce an ultra-reliable nano-resistor array that enables robust analog AI inference for specific tasks, minimizing the dependence on complex circuitry. Conductance states are fixed and geometrically defined through a single micro-nano patterning process, removing the need for stochastic programming and reducing the complexity of programming circuits typically required in memristor-based accelerators. We achieved 6.8-bit programming accuracy and stable 8-bit conductance levels. Additionally, experimental results from multiply-accumulate (MAC) operations show the feasibility of achieving 8.2-bit accuracy in a passive 28x28 array with simple circuit-level compensation. This nano-resistor array offers a reliable and precise platform for AI computing, tailored for daily AI tasks while reducing peripheral circuitry.

DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads

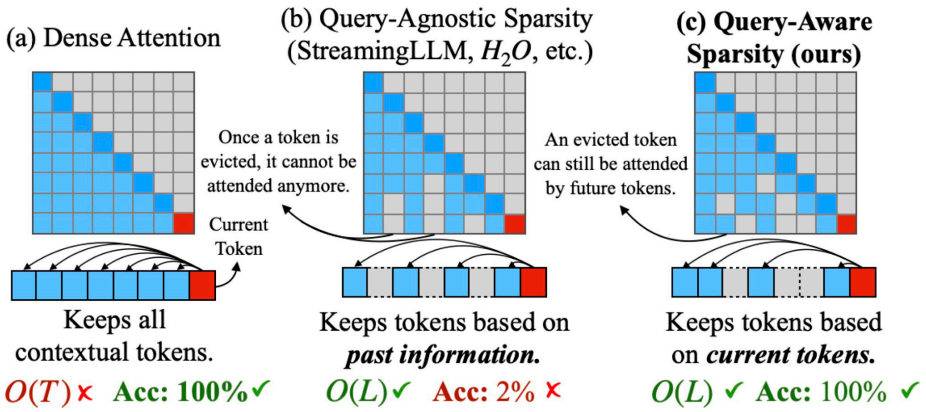
G. Xiao, J. Tang, J. Zuo, J. Guo, S. Yang, H. Tang, Y. Fu, S. Han
Sponsorship: National Science Foundation, MIT-IBM Watson AI Lab, MIT AI Hardware Program, Amazon, Samsung, Hyundai

Deploying long-context large language models (LLMs) is essential but poses significant computational and memory challenges. Caching all Key and Value (KV) states across all attention heads consumes substantial memory. Existing KV cache pruning methods either damage the long-context capabilities of LLMs or offer only limited efficiency improvements. In this paper, we identify that only a fraction of attention heads, a.k.a, Retrieval Heads, are critical for processing long contexts and require full attention across all tokens. In contrast, all other heads, which primarily focus on recent tokens and attention sinks--referred to as Streaming Heads--do not require full attention. Based on this insight, we introduce DuoAttention, a framework that only applies a full KV cache to retrieval heads while using a light-weight, constant-length KV cache for streaming heads, which reduces both LLM's decoding and pre-filling memory and latency without compromising its long-context abilities. DuoAttention uses a lightweight, optimization-based algorithm with synthetic data to identify retrieval heads accurately. Our method significantly reduces long-context inference memory by up to 2.55x for MHA and 1.67x for GQA models while speeding up decoding by up to 2.18x and 1.50x and accelerating pre-filling by up to 1.73x and 1.63x for MHA and GQA models, respectively, with minimal accuracy loss compared to full attention. Notably, combined with quantization, DuoAttention enables Llama-3-8B decoding with 3.3 million context length on a single A100 GPU.

Quest: Query-Aware Sparsity for Efficient Long-Context LLM Inference

J. Tang, Y. Zhao, K. Zhu, G. Xiao, B. Kasikci, S. Han
Sponsorship: NSF, MIT-IBM Watson AI Lab, MIT AI Hardware Program, Amazon, Samsung, Hyundai

As the demand for long-context large language models (LLMs) increases, models with context windows of up to 1M tokens are becoming prevalent. However, long-context LLM inference is challenging since the inference speed decreases significantly as the sequence length grows. This slowdown is primarily caused by loading a large KV cache during attention. Previous works have shown that a small portion of critical tokens will dominate the attention outcomes. However, we observe the criticality of a token highly depends on the query. To this end, we propose Quest, a query-aware KV cache selection algorithm. Quest keeps track of the minimal and maximal Key values in KV cache pages and estimates the criticality of a given page using Query. By only loading the Top-K critical KV cache pages, Quest significantly speeds up attention without sacrificing accuracy. We show that Quest can achieve up to 2.23x attention speedup, which reduces inference latency by 7.03x with negligible accuracy loss.



▲ Figure 1: Quest Algorithm Overview.

SORBET: Secure Off-chip Memory Interface for Deep Neural Network Accelerators

K. Lee, G. Das, D. Han, A. P. Chandrakasan
Sponsorship: Samsung Electronics

As deep neural networks (DNNs) are deployed in high-stakes applications, ensuring their confidentiality and integrity becomes crucial. Trusted execution environments (TEEs) offer a potential solution by cryptographically encrypting and authenticating all data traffic to and from DNN accelerators without relying on off-chip hardware or system software to provide security. However, hardware memory encryption and authentication for DNN accelerators is challenging due to the large memory footprints of DNNs and the impact of cryptographic operations on the data access pattern of the accelerators. To address these challenges, we present SORBET, a secure off-chip memory interface for DNN accelerators. SORBET efficiently manages the altered data access patterns resulting from cryptographic authentication and leverages lightweight cryptography to minimize overhead. Also, we designed our DNN accelerator to support fused-layer processing, a technique that reduces the overall off-chip data traffic, to alleviate pressure on the cryptographic engine. Our implementation of a secure DNN accelerator equipped with SORBET supports memory encryption and authentication with only 1-22% performance overhead, 5.6-7.9% of the chip area, and 18.4% energy overhead. These results are verified with an ASIC implementation using TSMC 28nm technology. Overall, we show that memory security of TEEs can be practically achieved for resource-constrained DNN accelerators.

LoopTree: Exploring the Fused-layer Dataflow Accelerator Design Space

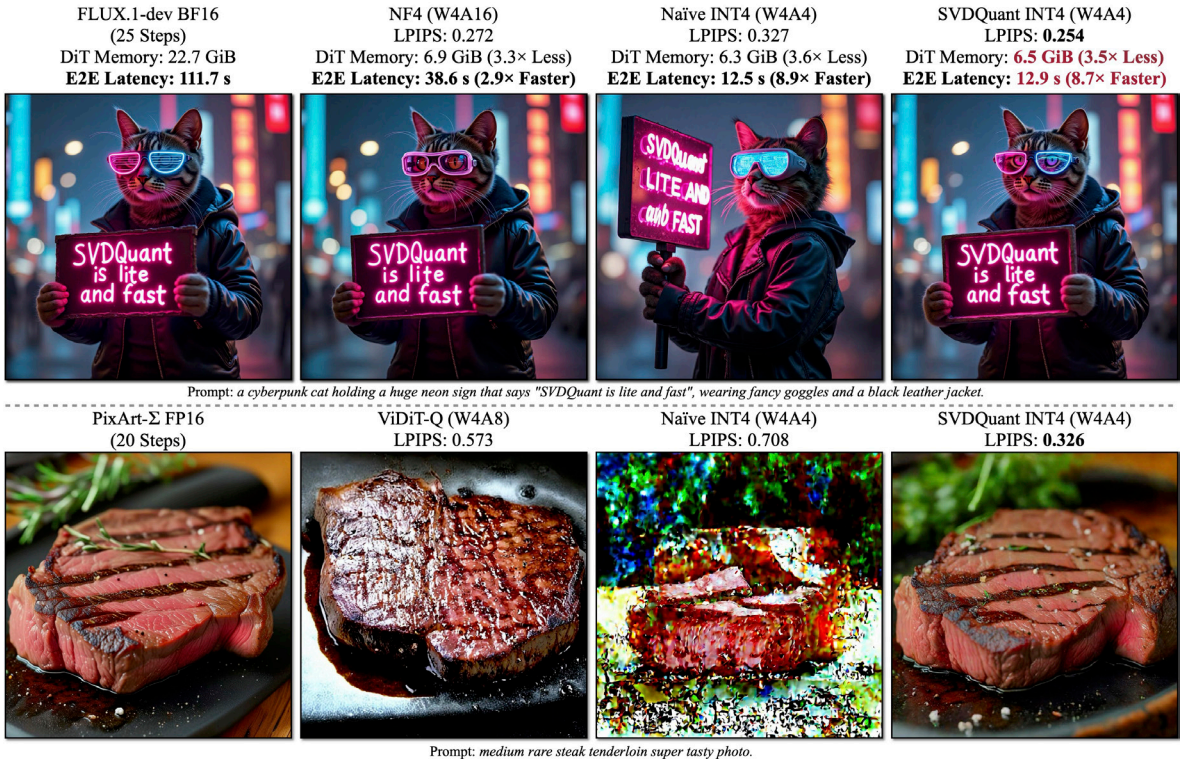
M. Gilbert, Y. N. Wu, V. Sze, J. S. Emer
Sponsorship: MIT AI Hardware Program

Deep neural network (DNN) accelerators often process DNNs one layer at a time, keeping intermediate data in off-chip DRAM. However, DRAM data transfers consume more energy than on-chip transfers and may increase latency due to limited DRAM bandwidth. Recent work has proposed fused-layer accelerators, which do not transfer intermediate data to/from DRAM but must recompute or retain data on-chip. This retention-recomputation trade-off results from the order of operations (dataflow) and the data tiles retained on-chip (partitioning). However, prior work has only explored a subset of this design space. We propose (1) an expanded design space, and (2) a model, LoopTree, to evaluate the latency and energy consumption of accelerators in this design space. We validate LoopTree against prior architectures (worst-case 4% error). Finally, we show how exploring this larger space results in more efficient designs (e.g., up to 10× buffer capacity reduction to achieve the same off-chip transfers).

SVDQuant: Absorbing Outliers by Low-Rank Components for 4-Bit Diffusion Models

M. Li, Y. Lin, Z. Zhang, T. Cai, X. Li, J. Guo, E. Xie, C. Meng, J.-Y. Zhu, S. Han
Sponsorship: NSF, MIT-IBM Watson AI Lab, MIT AI Hardware Program, Amazon, Samsung, Hyundai

Diffusion models have been proven highly effective at generating high-quality images. However, as these models grow larger, they require significantly more memory and suffer from higher latency, posing substantial challenges for deployment. In this work, we aim to accelerate diffusion models by quantizing their weights and activations to 4 bits. At such an aggressive level, both weights and activations are highly sensitive, where conventional post-training quantization methods for large language models like smoothing become insufficient. To overcome this limitation, we propose SVDQuant, a new 4-bit quantization paradigm. Different from smoothing which redistributes outliers between weights and activations, our approach absorbs these outliers using a low-rank branch. We first consolidate the outliers by shifting them from activations to weights, then employ a high-precision low-rank branch to take in the weight outliers with Singular Value Decomposition (SVD). This process eases the quantization on both sides. However, naïvely running the low-rank branch independently incurs significant overhead due to extra data movement of activations, negating the quantization speedup. To address this, we co-design an inference engine Nunchaku that fuses the kernels of the low-rank branch into those of the low-bit branch to cut off redundant memory access. It can also seamlessly support off-the-shelf low-rank adapters (LoRAs) without the need for re-quantization. Extensive experiments on SDXL, PixArt-∑, and FLUX.1 validate the effectiveness of SVDQuant in preserving image quality. We reduce the memory usage for the 12B FLUX.1 models by 3.5×, achieving 3.0× speedup over the 4-bit weight-only quantized baseline on the 16GB laptop 4090 GPU, paving the way for more interactive applications on PCs. Our quantization library and inference engine are open-sourced.



▲ Figure 1: SVDQuant is a post-training quantization technique for 4-bit weights and activations that well maintains visual fidelity. On 12B FLUX.1-dev, it achieves 3.6× memory reduction compared to the BF16 model. By eliminating CPU offloading, it offers 8.7× speedup over the 16-bit model when on a 16GB laptop 4090 GPU, 3× faster than the NF4 W4A16 baseline. On PixArt-Σ, it demonstrates significantly superior visual quality over other W4A4 or even W4A8 baselines. "E2E" means the end-to-end latency including the text encoder and VAE decoder.

LongVILA: Scaling Long-Context Visual Language Models for Long Videos

Q. Hu, H. Tang, S. Yang, S. Han

Long-context capability is critical for multi-modal foundation models, especially for long video understanding. We introduce LongVILA, a full-stack solution for long-context visual-language models by co-designing the algorithm and system. For model training, we upgrade existing VLMs to support long video understanding by incorporating two additional stages, i.e., long context extension and long video supervised fine-tuning. However, training on long video is computationally and memory intensive. We introduce the long-context Multi-Modal Sequence Parallelism (MM-SP) system that efficiently parallelizes long video training and inference, enabling 2M context length training on 256 GPUs without any gradient checkpointing. LongVILA efficiently extends the number of video frames of VILA from 8 to 2048, MM-SP is 2.1x - 5.7x faster than ring style sequence parallelism and 1.1x - 1.4x faster than Megatron with a hybrid context and tensor parallelism.

QServe: W4A8KV4 Quantization and System Co-design for Efficient LLM Serving

Y. Lin, H. Tang, S. Yang, Z. Zhang, G. Xiao, C. Gan, S. Han
Sponsorship: NSF, MIT-IBM Watson AI Lab, MIT AI Hardware Program, Amazon, Samsung, Hyundai

Quantization can accelerate large language model (LLM) inference. Going beyond INT8 quantization, the research community is actively exploring even lower precision, such as INT4. Nonetheless, state-of-the-art INT4 quantization techniques only accelerate low-batch, edge LLM inference, failing to deliver performance gains in large-batch, cloud-based LLM serving. We uncover a critical issue: existing INT4 quantization methods suffer from significant runtime overhead (20-90%) when dequantizing either weights or partial sums on GPUs. To address this challenge, we introduce QoQ, a W4A8KV4 quantization algorithm with 4-bit weight, 8-bit activation, and 4-bit KV cache. QoQ stands for quattuor-octo-quattuor, which represents 4-8-4 in Latin. QoQ is implemented by the QServe inference library that achieves measured speedup. The key insight driving QServe is that the efficiency of LLM serving on GPUs is critically influenced by operations on low-throughput CUDA cores. Building upon this insight, in QoQ algorithm, we introduce progressive quantization that can allow low dequantization overhead in W4A8 GEMM. Additionally, we develop SmoothAttention to effectively mitigate the accuracy degradation incurred by 4-bit KV quantization. In the QServe system, we perform compute-aware weight reordering and take advantage of register-level parallelism to reduce dequantization latency. We also make fused attention memory-bound, harnessing the performance gain brought by KV4 quantization. As a result, QServe improves the maximum achievable serving throughput of Llama-3-8B by 1.2x on A100, 1.4x on L40S; and Qwen1.5-72B by 2.4x on A100, 3.5x on L40S, compared to TensorRT-LLM. Remarkably, QServe on L40S GPU can achieve even higher throughput than TensorRT-LLM on A100. Thus, QServe effectively reduces the dollar cost of LLM serving by 3x.