# Neuromorphic Devices & AI Hardware Accelerators

# Secure Digital In-memory Computing for Privacy and Integrity

M. Ashok, S. Maji, X. Zhang, J. Cohn, A. P. Chandrakasan

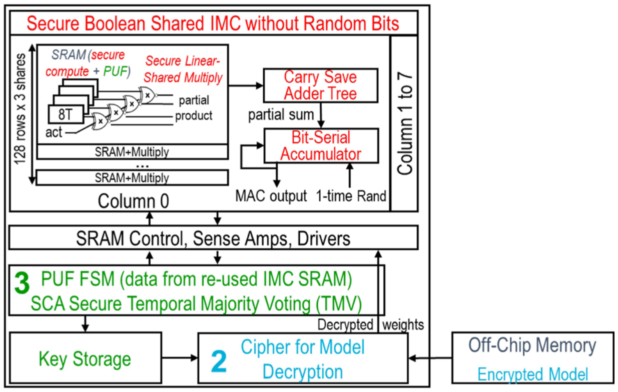
Sponsorship: MIT-IBM Watson AI Lab, NSF Graduate Research Fellowship Program (Grant No. 1745302), MathWorks Engineering Fellowship

Machine learning accelerators are used in a wide variety of applications to achieve high energy efficiency at the expense of generalization. One specific architecture, digital in-memory compute (IMC), reduces data transfer energy by interleaving compute and memory while still targeting high accuracy, even with technology scaling.

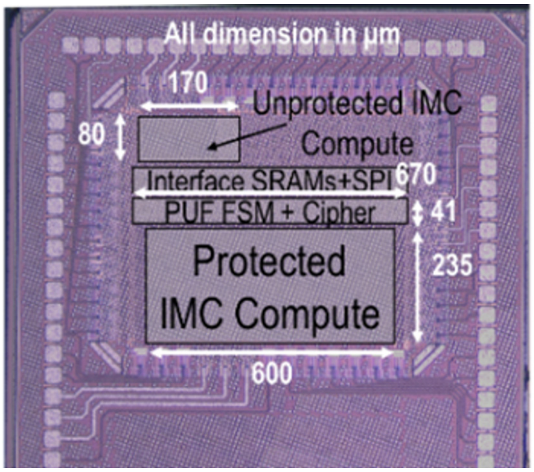
The security of these accelerators is essential to protect both the inputs and model. The input activations often reflect private data collected on a sensor, such as face images. The weights reveal information about private training datasets, which can be used to mount adversarial attacks. Two types of passive attacks can help an attacker gain this information: side channel attacks (SCAs), which correlate the integrated circuit's (IC) power consumption or electromagnetic missions to the data, and bus probing attacks (BPAs), which directly tap the traces between the IC and off-chip memory.

We propose addressing these security threats through an IMC macro with three key protections. The computation is split up into random shares, such that the total power consumption is uncorrelated to the data. By targeting practical attacker scenarios, this architecture is designed to eliminate the need for fresh random bits while minimizing area and energy overheads when possible. In addition, a National Institute of Standards and Technology standard lightweight cipher is included locally so that the model is present only in an encrypted form off-chip. Finally, a feedback-cut physical unclonable function that takes advantage of manufacturing variations in existing IMC memory is designed to generate secret keys on-chip.

With a combination of these contributions, we achieve a secure IMC macro in 14-nm complementary metal-oxide semiconductor (CMOS) technology. The design does not require any random number generators and has no limitations on neural network accuracy, providing a generalized solution for privacy and integrity in a variety of machine learning applications.


Secure Boolean Shared IMC without Random Bits
128 rows x 3 shares
SRAM (secure compute + PUF)
Secure Linear-Shared Multiply
8T
act
partial product
SRAM+Multiply
...
SRAM+Multiply
Column 0
Carry Save Adder Tree
partial sum
Bit-Serial Accumulator
MAC output
1-time Rand
Column 1 to 7
SRAM Control, Sense Amps, Drivers
3 PUF FSM (data from re-used IMC SRAM)
SCA Secure Temporal Majority Voting (TMV)
Decrypted weights
Key Storage
2 Cipher for Model Decryption
Off-Chip Memory
Encrypted Model

▲Figure 1: Proposed secure digital in-memory compute macro architecture with 3 key contributions (Boolean shared compute for SCA security, model decryption on-chip for BPA security, and secret key generated on-chip) highlighted.



▲Figure 2: Die micrograph of proposed protected and baseline unprotected macros in 14-nm CMOS technology. Key blocks and dimensions are noted.

## FURTHER READING:

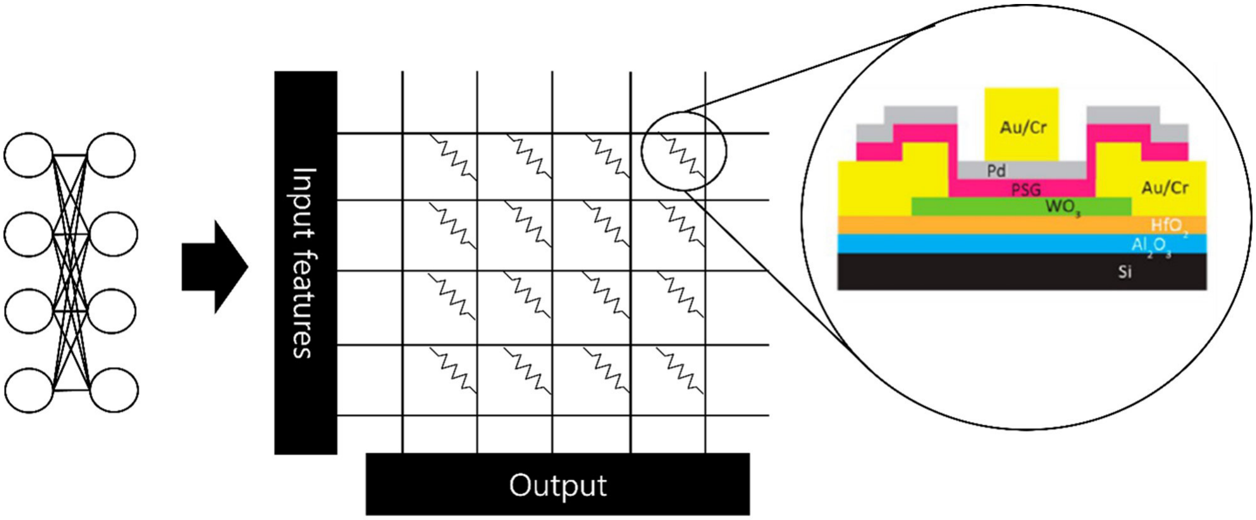
- M. Ashok, S. Maji, X. Zhang, J. Cohn, and A. P. Chandrakasan, "A Secure Digital In-Memory Compute Macro with Protections Against Side-Channel and Bus Probing Attacks," *IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1-2, 2024.

# Analog In-memory Computing with Protonic Synapses

J. Lee, J. A. del Alamo

Sponsorship: IBM Watson

Though demand for computation in neural networks is skyrocketing, conventional computing resources are still constrained by their limited energy efficiency. One of the most promising solutions to this is analog in-memory computing (AIMC). In AIMC, synaptic weights of neural networks are encoded into the conductance of devices which are then configured into crossbar-arrays that perform the matrix-vector multiplication operation. Recently, protonic synapses were demonstrated with suitable properties for AIMC such as linear and symmetric conductance modulation. To study this, we have built a simulation model for protonic synapses and simulated a crossbar-array operation using IBM's AIHWKIT. AIMC with protonic synapses showed 97.1% accuracy in MNIST classification with a linear-4-layer network. This is comparable to a digital chip accuracy of 98%. We also tested the Tiki-taka algorithm that compensates for device non-idealities and identified the relation between the conductance modulation shapes and the performance of algorithm.



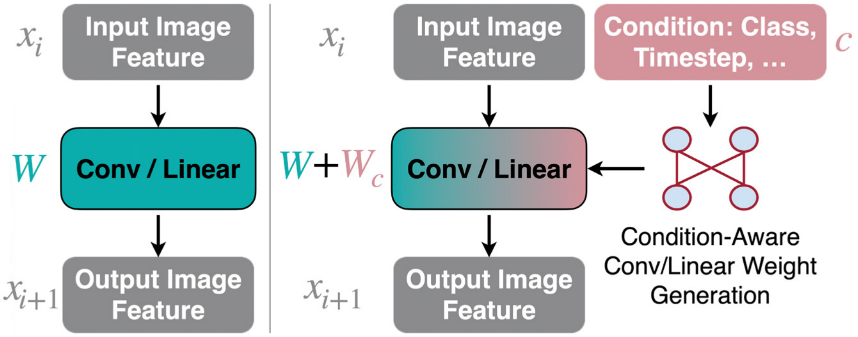
▲ Figure 1: Schematic of cross-bar array with protonic synapses.

# Condition-aware Neural Network for Controlled Image Generation

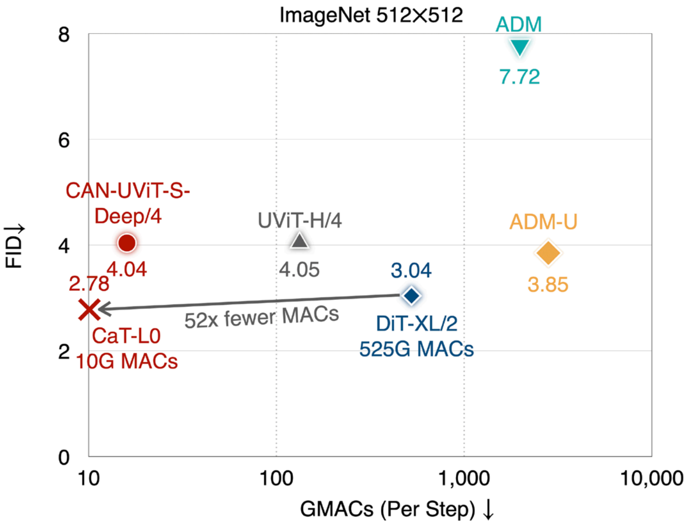
H. Cai, M. Li, Z. Zhang, Q. Zhang, M.-Y. Liu, S. Han

Sponsorship: MIT-IBM Watson AI Lab, Amazon, MIT Science Hub, NSF

We present Condition-Aware Neural Network (CAN), a new method for adding con-trol to image-generative models. In parallel to prior conditional control methods, CAN controls the image-generation process by dynamically manipulating the weight of the neural network. This is achieved by introducing a condition-aware weight generation module that generates conditional weight for convolution/linear layers based on the input condition. We test CAN on class-conditional image gen-eration on ImageNet and text-to-image generation on COCO. CAN consistently de-livers significant improvements for diffusion transformer models, including DiT and UViT. In particular, CAN combined with EfficientViT (CaT) achieves 2.78 FID on ImageNet 512x512, surpassing DiT-XL/2 while requiring 52x fewer MACs per sampling step.



▲Figure 1: Illustration of CAN.



◀Figure 2: Comparing CAN models and prior image-generative models on ImageNet 512X512.

## FURTHER READING

- H. Cai, et al. "Condition-Aware Neural Network for Controlled Image Generation," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 2024.
- H. Cai, et al. "EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction," *Proc. IEEE/CVF International Conference on Computer Vision*, p. 2023.

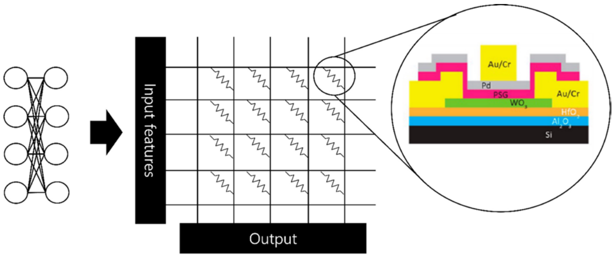
# IBM AIHWKIT Simulations of Analog Training Accelerators Based on Protonic Synapses

J. Lee, J. A. del Alamo

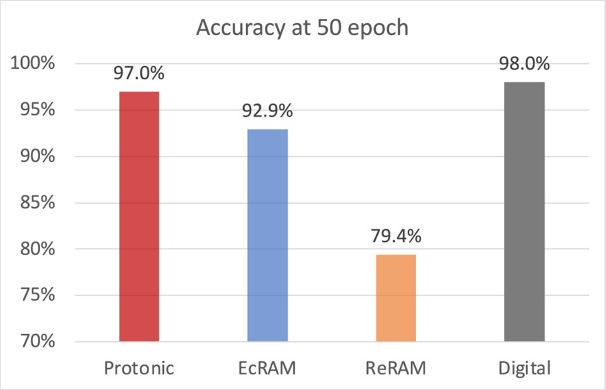
Sponsorship: IBM, Ericsson

Though demand for computation in neural networks is skyrocketing, conventional computing resources are constrained by their limited energy efficiency. One of the most promising solutions to this is analog in-memory computing (AIMC). In AIMC, synaptic weights of neural networks are encoded into the conductance of devices that are then configured into crossbar-arrays that perform the matrix-vector multiplication operation. Recently, our group demonstrated protonic synapses with suitable properties for AIMC such as linear and symmetric conductance modulation. To study the potential of this new device technology for AIMC, we are carrying out simulations of crossbar-array operation using IBM's AIHWKIT.

With AIHWKIT, users can simulate AIMC accelerators and observe how modifying the device properties directly affects the performance of neural networks. Additionally, the tool enables the evaluation of performance in various types of neural networks, including convolutional neural networks, LSTM, Transformer, and others, which are currently under extensive research. Further, this system can also investigate the performance of different algorithms, such as stochastic gradient descent (SGD), Tikitaka, etc. In our studies, we compare MIT's protonic synapses against other analog synaptic devices such as IBM's ReRAM and ECRAM technologies. We demonstrate an accuracy of our protonic devices of 97.0% in Modified National Institute of Standards and Technology (MNIST) hand-written digit image classification using SGD, which is much higher than with ReRAM of 79.4%. Protonic device accuracy approaches that of digital circuits, which typically reach 98%. This enhancement in performance is due to the more symmetric and linear modulation of weights in protonic synapses. Our research is investigating the effect of device non-idealities on accuracy and testing several analog computing algorithms designed to mitigate their degrading impact.



▲Figure 1: Schematic of cross-bar array for neural network with protonic synapses; input features are loaded in analog voltages from left side, and output emerges from bottom. Each crosspoint has a protonic synapse.



▲Figure 2: Comparison of accuracy of 4-layer analog neural networks based on different device technologies after training with MNIST model. For reference, result of digital network is also shown.

## FURTHER READING

- M. Onen et al., "Nanosecond Protonic Programmable Resistors for Analog Deep Learning," *Science*, vol. 377, no. 6605, pp. 539–543, Jul. 2022. doi: 10.1126/science.abp8064
- M. Onen et al., "Neural Network Training with Asymmetric Crosspoint Elements," *Front. Artif. Intell.*, vol. 5, May, 2022. doi: 10.3389/frai.2022.891624

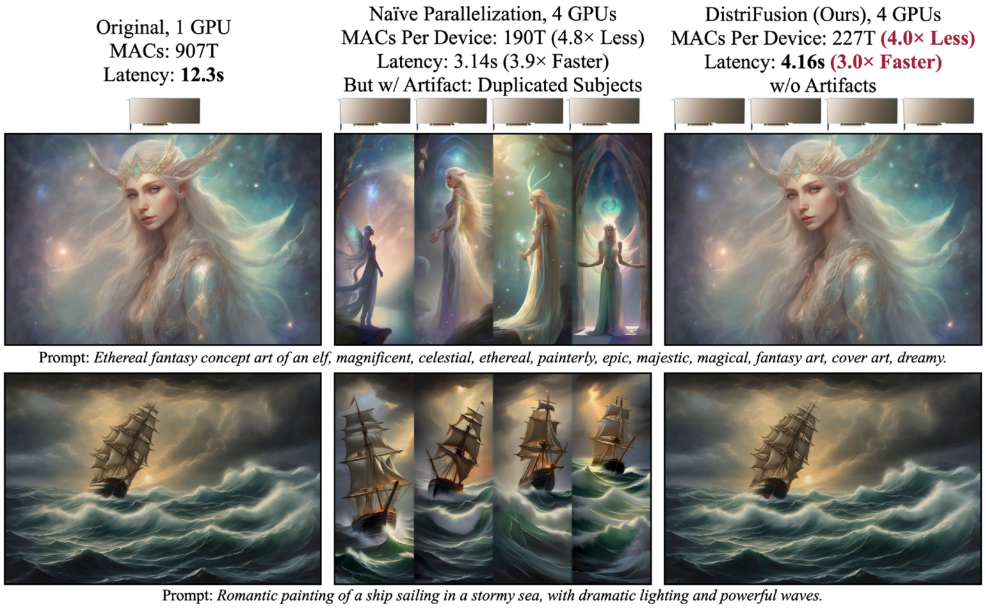
# DistriFusion: Distributed Parallel Inference for High-resolution Diffusion Models

M. Li, T. Cai, J. Cao, Q. Zhang, H. Cai, J. Bai, Y. Jia, M.-Y. Liu, K. Li, S. Han

Sponsorship: MIT-IBM Watson AI Lab, Amazon, MIT Science Hub, NSF

Diffusion models have achieved great success in synthesizing high-quality images. However, generating high-resolution images with diffusion models is still challenging due to the enormous computational costs, resulting in a prohibitive latency for interactive applications. In this paper, we propose DistriFusion to tackle this problem by leveraging parallelism across multiple graphic processing units (GPUs). Our method splits the model input into multiple patches and assigns each patch to a GPU. However, naively implementing such an algorithm breaks the interaction between patches and loses fidelity while incorporating such an interaction will incur tremendous communication overhead.

To overcome this dilemma, we observe the high similarity between the input from adjacent diffusion steps and propose displaced patch parallelism, which takes advantage of the sequential nature of the diffusion process by reusing the pre-computed feature maps from the previous timestep to provide context for the current step. Therefore, our method supports asynchronous communication, which can be pipelined by computation. Extensive experiments show that our method can be applied to recent Stable Diffusion XL with no quality degradation and achieve up to a 6.1× speedup on eight A100 GPUs compared to one.


Original, 1 GPU
MACs: 907T
Latency: **12.3s**
Naïve Parallelization, 4 GPUs
MACs Per Device: 190T (4.8× Less)
Latency: 3.14s (3.9× Faster)
But w/ Artifact: Duplicated Subjects
DistriFusion (Ours), 4 GPUs
MACs Per Device: 227T **(4.0× Less)**
Latency: **4.16s (3.0× Faster)**
w/o Artifacts

Prompt: *Ethereal fantasy concept art of an elf, magnificent, celestial, ethereal, painterly, epic, majestic, magical, fantasy art, cover art, dreamy.*

Prompt: *Romantic painting of a ship sailing in a stormy sea, with dramatic lighting and powerful waves.*

▲ Figure 1: We introduce DistriFusion, a training-free algorithm to harness multiple GPUs to accelerate diffusion model inference without sacrificing image quality. Naive Patch suffers from the fragmentation issue due to the lack of patch interaction. Our DistriFusion removes artifacts and avoids the communication overhead by reusing the features from the previous steps. Setting: SDXL with 50-step Euler sampler, 1280×1920 resolution. Latency is measured on A100s.

## FURTHER READING

- M. Li, J. Lin, C. Meng, S. Ermon, S. Han, and J.-Y. Zhu, "Efficient Spatially Sparse Inference for Conditional Gans and Diffusion Models," *NeurIPS*, 2022.
- M. Li, J. Lin, Y. Ding, Z. Liu, J.-Y. Zhu, and S. Han, "Gan Compression: Efficient Architectures for Interactive Conditional Gans," *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5284-5294, 2020.

# A Massively Parallel In-memory-computing Architecture Using Stochastic Computing

Q. Wang, B. C. McGoldrick, M. A. Baldo, L. Liu

Sponsorship: SRC, Mathworks

The potential of in-memory computing (IMC) to circumvent the inefficiencies of traditional computing architectures has been widely recognized in the field of machine learning. Despite this potential, current IMC models are hampered by the difficulty of embedding complex arithmetic operations on-site, and the lack of capability for large-scale yet versatile parallel computation, which significantly diminish the applicability of IMC in scenarios demanding high computational throughput and intricate mathematical processing. Confronting these limitations, the research introduces a novel IMC architecture that utilizes stochastic MTJs to streamline in-memory operations. Beyond mere energy and space efficiency, this architecture innovates with a hierarchical approach that emulates the structure of biological neural networks, offering a solution to the scalability challenge. Within this hierarchy, local computations are performed by densely connected neuron-like nodes, enabling rapid data processing and communication for simpler tasks. Meanwhile, more complex and distant interactions are managed via a Network on Chip (NoC), facilitating efficient communication across the larger neural network. This dual-level organization allows for both intensive local computation and broader inter-neuron communication, effectively mirroring the parallelism and connectivity of the brain, and unlocking new potentials for IMC in supporting sophisticated and large-scale parallel machine learning workloads.

# 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

Tensor algebra describes a class of applications that are increasingly being used in fields such as machine learning, data science, graph analytics, scientific simulations, and engineering modeling. Although many of these applications operate on tensor data that has very high sparsity (i.e., many zeros), exploiting this sparsity to save both computation and memory is challenging. Prior sparse tensor algebra accelerators have explored splitting tensors into tiles to increase exploitable data reuse and improve throughput, but typically allocates tile size in a given buffer for the least sparse tile and thus limits utilization of available memory resources when sparsity varies between different regions of a tensor.

This paper proposes a speculative tensor tiling approach, called overbooking, to improve buffer utilization by taking advantage of the distribution of nonzero elements in sparse tensors to construct larger tiles with greater data reuse. We propose a low-overhead hardware mechanism, Tailors, that can tolerate data overflow by design while ensuring reasonable data reuse and introduce a statistical tiling approach, Swiftiles, that ensures high buffer utilization by picking a tile size so that tiles usually fit within the buffer's capacity, but can potentially overflow, i.e., it overbooks the buffers. Across a suite of 22 sparse tensor algebra workloads, we show that our proposed overbooking strategy introduces an average speedup of 52.7x and 2.3x and an average energy reduction of 22.5x and 2.5x over an existing sparse tensor algebra accelerator without and with optimized tiling, respectively.

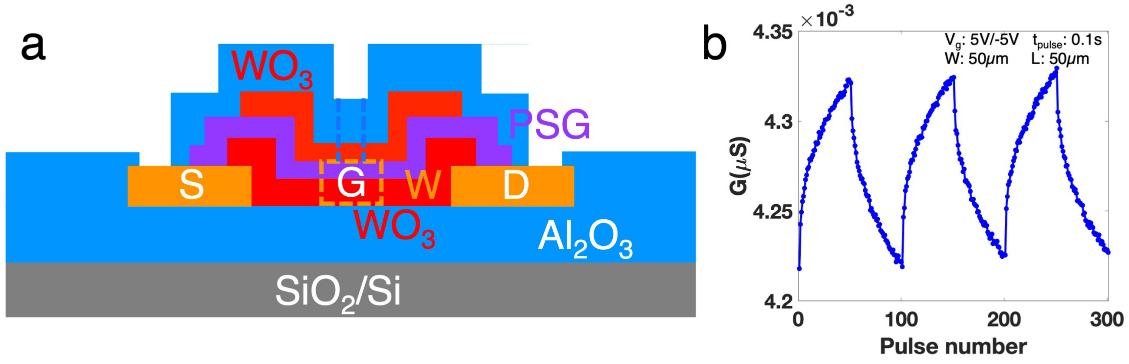
# Symmetric PSG/WO3 Protonic Synapses for Analog Deep Learning

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 with specific devices for matrix multiplication calculations and outer product updates. Among them, electrochemical RAMs modulate channel resistance by ionic exchange between the channel and a gate reservoir via an electrolyte.

This study focuses on proton-based ionic synapses featuring in-situ protonated $H_xWO_{3-x}$ as both channel and gate reservoir, and phosphorus-doped silicon dioxide (PSG) as electrolyte. This design aims to enable neural network training with enhanced energy efficiency, non-volatility, and low latency. We have experimentally implemented this design in a CMOS and back-end-of-line compatible process. Our protonic synapse with a vertically symmetric structure enables non-volatile and repeatable channel conductance modulation under voltage pulses across gate and channel, thus showing promising applications in deep learning accelerators.



▲ Figure 1: (a) Schematic of symmetric WO$_3$-based protonic synapse with tungsten contacts, PSG electrolyte and Al$_2$O$_3$ encapsulation, fabricated in a lift-off-free and in-situ protonation process. (b) Conductance modulation under voltage pulses shows repeatable control.

# Analysis of Memristor Device Requirements and Update Thresholding Strategy for Precision Programming of Neuromorphic Memristor Arrays

G. Lee, J. Kim

Neuromorphic computing, inspired by neural systems, presents advantages such as minimized data transmission, reduced power consumption, and parallel computation, spanning applications in artificial intelligence (AI), scientific computing, and security. Neuromorphic computing, specifically with a memristor crossbar array, shows promise as an AI inference accelerator, demonstrated through on-chip deep neural network inferences comparable to software-based methods. Despite the prevalent one-transistor one-resistor (1T1R) system ensuring precise memristor programming and overcoming sneak path issues, it compromises power, speed, and design simplicity. For these reasons, the simplest one-resistor (1R) system that can accurately program memristors and minimize the sneak path problem would be the ultimate solution. However, precise conductance adjustments at cross-points remain challenging due to asymmetry, poor selectivity, sneak paths, and stochasticity.

This study introduces the simplest automatic programming algorithm, employing iterative conductance reads and updates. It analyzes memristor requirements for high programming accuracy, establishing relationships with array size and parameters like interconnect-to-memristor conductance ratio. Selectivity is identified as enhancing convergence speed. Notably, update asymmetry hampers programming, which is addressed by an update event thresholding algorithm, significantly improving accuracy. This analysis and programming strategy hold the potential to aid the 1R memristor array system in overcoming programming challenges, realizing an ideal AI inference accelerator.

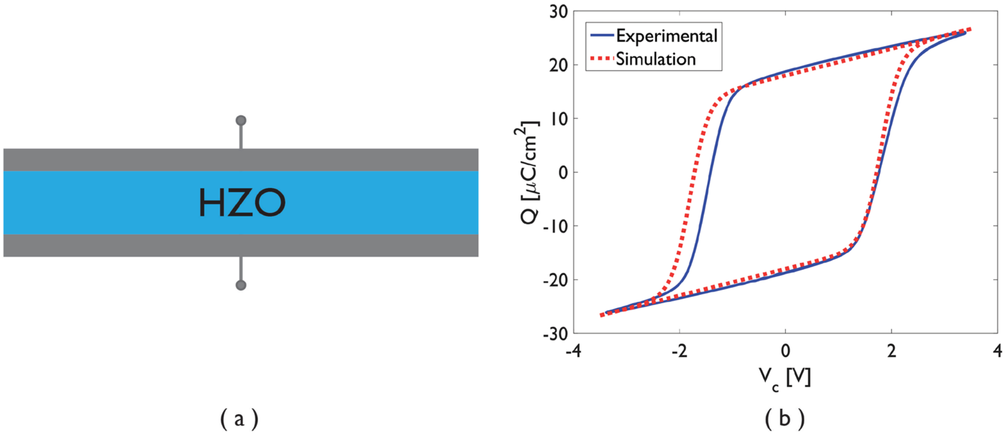
# Large-signal TCAD Simulation of Polarization Switching in Ferroelectric Devices

J. Navarro, Y. Shao, J. A. del Alamo

Sponsorship: Technical University of Madrid (UPM), Semiconductor Research Corporation

Ferroelectric materials enjoy spontaneous and non-volatile charge polarization with an orientation that can be switched by the application of a voltage. Recently, the discovery of ferroelectricity in HZO ($Hf_{0.5}Zr_{0.5}o_2$) thin films, compatible with current CMOS fabrication processes, have made ferroelectric field-effect transistors (FE-FETs) attractive as low power, nonvolatile memory devices for neuromorphic computing. However, the physics behind HZO polarization switching, especially in FE-FETs, are not widely understood.

In this work, we carry out TCAD simulations on FE capacitors and FE-FETs under large-signal operation. We have employed the Preisach model to reproduce the experimental polarization-voltage hysteresis loop, including transient behavior for low-frequency signals. Currently we are studying the nonvolatile memory behavior in FE-FETs. These results allow us to study how current physical models differ from experimental data, providing insights on the essential FE physics.



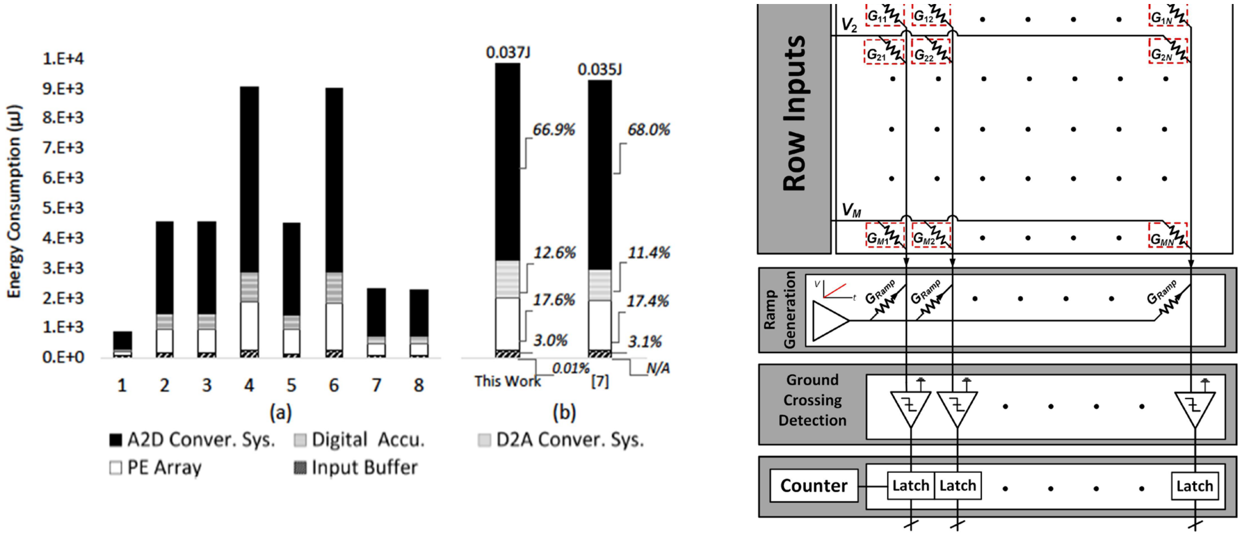
▲Figure 1: (a) FE Capacitor structure (MFM) and (b) experimental and simulated polarization-voltage (P-V) hysteresis loops

# Interface Circuits for Analog In-Memory Computing

M. A. G. Elsheikh, H.-S. Lee

Sponsorship: MIT/MTL Samsung Semiconductor Research Fund

Machine learning (ML) has found its way into our daily lives in applications including image processing, voice recognition and healthcare. The energy and speed bottlenecks in ML stem from data movement back and forth between the memory and the processing element. Analog compute in-memory (CIM) presents an alternative where the processing is done locally within the memory, and the analog nature of it allows the parallelization of a significant number of multiply-and-accumulate operations. However, current read-out circuits of CIM circuits are energy and area hungry leading a slow-down and energy inefficiency in the overall accelerator performance. In this work, we proposed a new design for interface circuits using single-slope analog to digital converters that exploit the typical statistics of neural network outputs to optimize speed and efficiency. This paves the way for the incorporation of CIM accelerator in low-power applications such as health monitoring and mobile applications.



▲ Figure 1: (a)Energy breakdown in analog CIM systems (b) proposed single-slope analog to digital converter readout circuits for analog CIM

# Sub-Nanometer Interface Modifications for Conductance Modulation in Proton-Based ECRAM Devices

J. Meyer, M. Huang, B. Yildiz

Sponsorship: DoD NDSEG Fellowship (2022 Fellow), DOE EFRC Hydrogen in Energy and Information Sciences (HEISs)

Three-terminal electrochemical random-access memory (ECRAM) devices have gained interest for use as resistive elements in energy-efficient neuromorphic computing architectures. ECRAM devices display promising non-volatility, reversibility, and symmetric switching operations through solid-state ion intercalation of a channel material, typically $WO_3$. However, these devices remain limited in energy-efficiency and programming speed, displaying operating voltages above 1 V and programming on the order of microseconds. Advances in thin high ionic conductivity electrolytes have decreased the operating voltage required for switching, but as electrolyte thickness approaches a few nanometers, other optimizations are required.

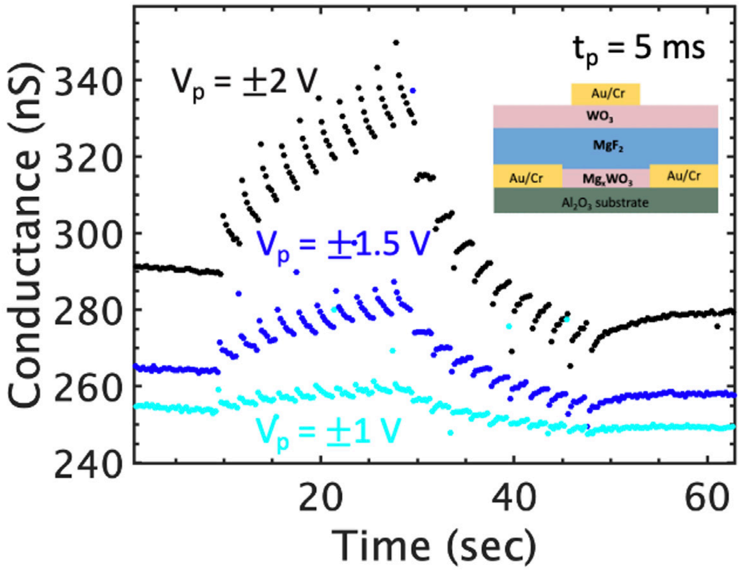
The structure of ECRAM devices is similar to solid-state lithium-ion batteries, for which electrode-electrolyte interfaces can be significant impedances to ion transport. For proton-based ECRAM, much less is known about proton transport across solid-state interfaces. This work chemically modifies the electrolyte-channel interface in a proton-based ECRAM device with sputtered metals and metal oxides. Differences in conductance modulation across interface-modified devices are presented. Surface-modified $WO_3$ films are examined with X-ray photoelectron spectroscopy. An electrochemical impedance spectroscopy setup is also used to separate the impedance contributions of electrolyte and interface in the devices. In doing so, we demonstrate some tunability of ECRAM device operation with sub-nanometer interface modifications. These results suggest modifying the electrolyte-channel interface in ECRAM devices may be another route to achieve faster neuromorphic computing with lower operating voltages.

# Magnesium Fluoride as a Thin Film Electrolyte in Mg-based Electrochemical Ionic Synapse (EIS) Devices

M. L. Schwacke, J. A. del Alamo, J. Li, B. Yildiz

Sponsorship: Semiconductor Research Corporation (Task ID: 3010-001), MIT Quest

Dynamic doping by electrochemical ion intercalation is a promising mechanism for modulating electronic conductivity, allowing for energy-efficient, brain-inspired computing hardware. Programmable resistors which operate by this mechanism are called electrochemical ionic synapses (EIS). Use of $Mg^{2+}$ as the working ion in EIS allows for both low energy consumption (compared to O-EIS) and good retention in ambient air (compared to H-EIS). However, while Mg-EIS devices based on organic electrolytes have provided a promising proof-of-concept; inorganic, thin film Mg electrolytes are needed for compatibility with CMOS processing. Here we report development of $MgF_2$ as a thin film electrolyte for Mg-EIS. We investigate how deposition technique (RF sputtering, electron beam evaporation) affects the $MgF_2$ film properties and by extension device characteristics. We show that Mg-EIS with $MgF_2$ electrolyte films are promising for CMOS-compatible EIS with long-term retention and low energy consumption.



▲Figure 1: Channel conductance of Mg-EIS device with $MgF_2$ electrolyte (schematic shown in inset) in response to 10 positive followed by 10 negative voltage pulses.

# Polaronic Carrier Mobility in Channel Materials for CMOS-compatible ECRAM

P. Žguns, B. Yildiz

Sponsorship: SRC JUMP 2.0 SUPREME Center

Neuromorphic computing hardware based on electrochemical random access memory (ECRAM) enables low-energy computing and nanosecond timescale operation [1–3]. In these devices, the electrochemical intercalation of hydrogen modulates the electronic conductivity of the channel material, where high sensitivity of conductivity to hydrogen insertion is needed for high energy efficiency. Therefore, to select promising channel materials, the impact of hydrogen on carrier mobility needs to be understood. Here, we computationally study hydrogenated, CMOS-compatible channel materials $H_xWO_3$, $H_xV_2O_5$, and $H_xMoO_3$. We apply the DFT+*U* method to describe the electronic structure and polaron localization in the host oxide lattice. We identify the ground state configurations of the proton–polaron pairs and probe microscopic factors that control polaron mobility, viz. the polaron migration barriers in pristine lattice and proton–polaron association energy (i.e., the energy required to break the pair). The polaron migration barriers are comparable to the polaron–proton association energies (which are on the order of 0.1 eV–0.2 eV), highlighting that the attraction between polarons and protons plays an important role in polaron mobility. These computational findings will help us to identify the most suitable CMOS-compatible oxide channel material for low-power ECRAM devices for neuromorphic computing applications.

## FURTHER READING

- X. Yao et al., "Protonic Solid-State Electrochemical Synapse for Physical Neural Networks," *Nat. Commun.* 11, 3134 (2020)
- M. Onen et al., "CMOS-Compatible Protonic Programmable Resistor Based on Phosphosilicate Glass Electrolyte for Analog Deep Learning," *Nano Lett.* 21, 14, 6111–6116 (2021)
- M. Onen et al., "Nanosecond Protonic Programmable Resistors for Analog Deep Learning," *Science* 377, 6605, 539-543 (2022)

# Analog Complex-valued Neural Networks for Quadratic Energy Savings

M. G. Bacvanski, S. K. Vadlamani, D. R. Englund
Sponsorship: NSF RAISE-TAQS (1936314), NSF C-Accel (2040695), NTT Research Inc.

Neural networks (NNs) that are composed of complex weights and operate over complex inputs have demonstrated excellent performance in domains like medical data and signal processing because of their natural ability to manipulate phase and amplitude. However, implementing these complex-valued NNs on conventional digital GPU hardware entails performing several times more multiplication and addition operations than real-valued NNs of the same size. In this work, we propose an AI accelerator that uses standard telecommunications hardware to efficiently implement complex-valued NNs by combining standard modulation schemes with homodyne detection. Through extensive numerical experiments, we demonstrate that these telecom-based analog NNs perform identically to traditional real-valued neural networks on several standard real-valued datasets, while consuming quadratically lower amounts of energy.

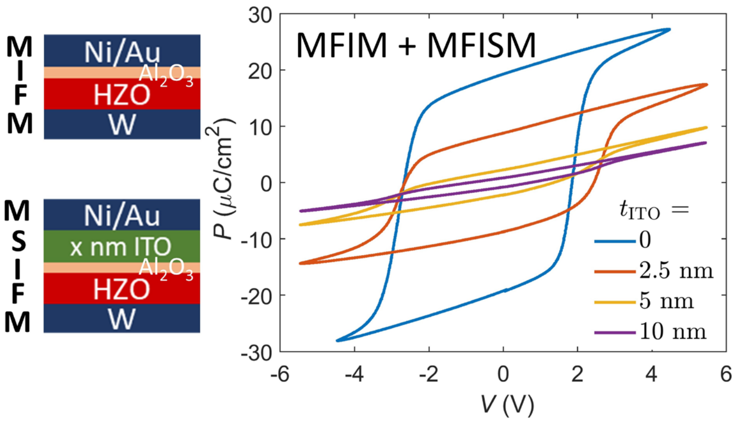
# CMOS-compatible Ferroelectric Memory for Analog Neural Network Accelerators

Y. Shao, E. R. Borujeny, T. Kim, T. E. Espedal, J. C.-C. Huang, J. Navarro, D. A. Antoniadis, J. A. del Alamo

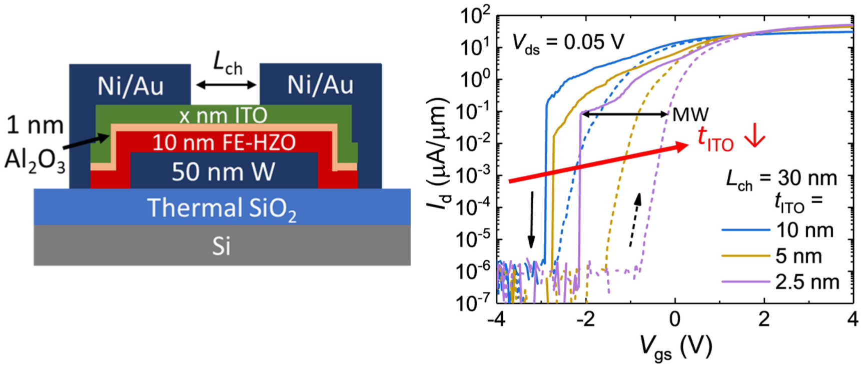
Sponsorship: Intel Corporation, SRC

Artificial intelligence has irreversibly changed the way information is stored and processed. However, the huge energy consumption and enormous computation time required for training modern deep learning models highlights the urgent need for energy- and time-efficient hard-ware uniquely designed to implement AI algorithms. Recently, analog computing has been proposed as an alternative to the digital counterpart. In an analog neural network accelerator, core computations are carried out in the analog domain, exploiting unique device properties and fundamental physical laws.

In this work, we examine the potential of complementary metal-oxide semiconductor- (CMOS) compatible ferroelectric field-effect transistors (FE-FETs) with a metal-oxide channel as the core device for analog accelerators. Extensive electrical characterizations, including large-signal polarization-voltage (P-V), small-signal capacitance-voltage, and direct-current current-voltage characteristics have been carried out on multiple device structures. We have found that the thickness scaling of the channel material, i.e., indium-tin-oxide (ITO) in our study, plays a key role in the enhancement of FE polarization switching (Figure 1). To scale down ITO thickness for better device performance, we studied thin-film transistors with ITO channel thickness down to 2.5 nm and obtained a field-effect mobility of 34 cm²/V·s in 2.5-nm-thick ITO films. In addition, we designed and fabricated FE-FETs with highly scaled channel length and source/drain contact length in a back-gate configuration. Thanks to the scaled ITO thickness, we observed a large memory window of ~2 V (Figure 2), which is among the best reported in the literature. On the other hand, the speed of FE polarization switching in the transistors/synapses, which determines the operation frequency of the analog crossbar array, is of great importance. In our fabricated devices, preliminary results show that sub-microsecond FE polarization switching could be obtained. These results pave the way to the realization of CMOS-compatible ferroelectric programmable resistors for analog accelerator arrays.



◀ Figure 1: P-V loops of metal/FE-HZO/insulator/metal and metal/FE-HZO/insulator/semiconductor/metal structures with different ITO thicknesses.



▶ Figure 2: FE-FET structure (left) and subthreshold characteristics of devices with different ITO thicknesses (right). Memory window is defined at 100 nA/μm.

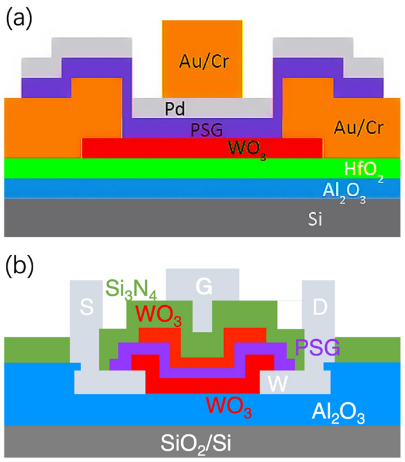
# Device Stack Optimization for Protonic Programmable Resistors

D. Shen, J. A. del Alamo

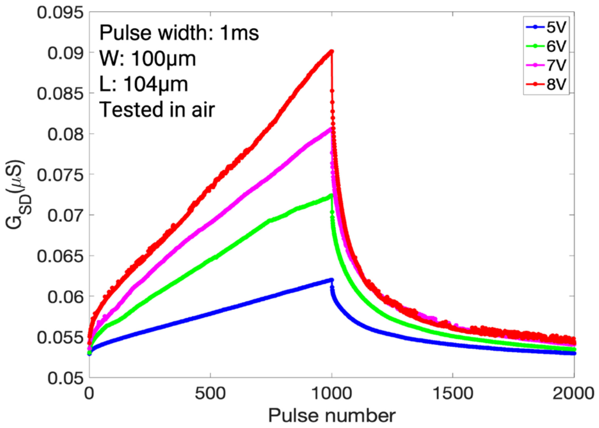
Sponsorship: MIT-IBM Watson AI Lab

Analog computing offers a potential solution for overcoming computational bot-tlenecks in traditional digital systems utilized for deep learning. The fundamental con-cept of analog deep learning accelerators involves processing information locally by leveraging the physical properties of devices, rather than conventional Boolean arith-metic—specifically, using Ohm's and Kirchhoff's laws for matrix inner product calcula-tions and threshold-based updating for the outer product. Among various physical principles, electrochemical ion-intercalation makes possible a three-terminal device with a channel resistance that is modulated by ionic exchange between the channel and a gate reservoir via an electrolyte. This study focuses on such ionic programmable resistors featuring $WO_3$ as a channel and protons as the working ions, aiming to pro-vide information processing with increased energy savings, non-volatility, and low la-tency. Our group's previous work, with a device structure shown in Figure 1a, has demonstrated Si-compatible nanoscale devices that are 1,000 times smaller than bio-logical neurons, enabling channel conductance modulation over a 20x range with na-nosecond operation at room temperature.

In this work, we have optimized the device stack in four directions and used nano-porous phosphosilicate glass (PSG) to demonstrate a symmetric $WO_3$-PSG-$WO_3$ structure in a complementary metal-oxide semiconductor- (CMOS) compatible and lift-off-free process, with the help of a circular transfer length method, which efficiently ex-amines the resistance properties of $WO_3$. We have explored: (a) device protonation as part of the fabrication process, (b) encapsulation preventing proton depletion during device fabrication and operation, (c) contact metal optimization to replace gold with a CMOS-compatible material, and (d) a PSG evaluation vehicle to optimize device per-formance. Putting it all together, we have designed and fabricated a symmetric device, shown in Figure 1b, which enables non-volatile and repeatable channel conductance modulation under voltage pulses across gate and channel with results shown in Fig-ure 2, thus showing promising application in deep learning accelerators.



▲ Figure 1: $WO_3$ protonic device structures: (a) original device structure with $WO_3$ as channel, PSG as electrolyte, and Pd as hydrogen reservoir and controlling gate. (b) Optimized symmetric device structure with W as contact metal, in-situ protonated $WO_3$ as both channel and gate reservoir, and $Si_3N_4$ as encapsulation.



▲ Figure 2: Conductance modulation curve for different voltage pulses of a symmetric device. The device shows linear conductance increase with positive pulses on gate. With encapsulation and in-situ protonation, the device can repeatably modulate its conductance under air.

## FURTHER READING

- O. Murat, 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.

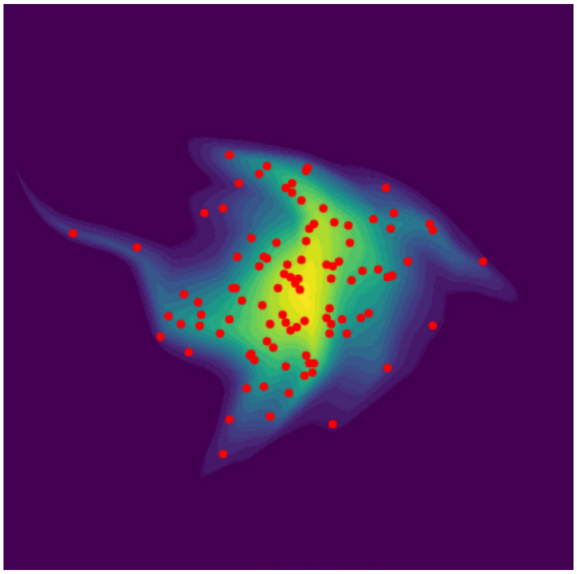
# The Use of Cross-validation in Semi-supervised Anomaly Detection

F.-K, Sun, D. S. Boning

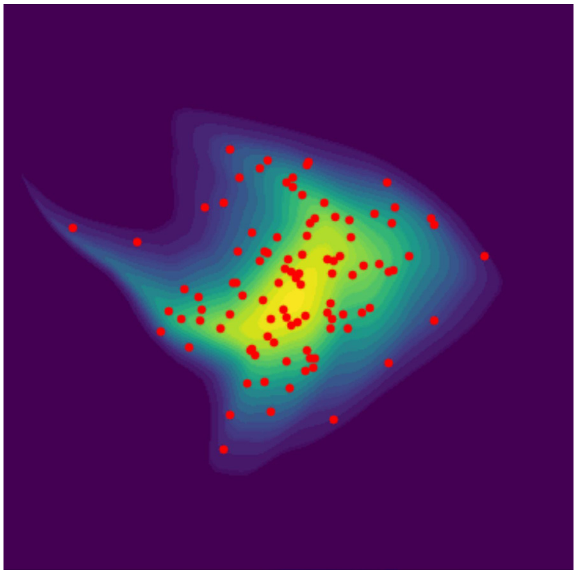
Sponsorship: Analog Devices

Unplanned industrial downtime can result in significant financial losses and operational disruptions. To address these issues, companies are investing in sensors designed to detect anomalies and reduce downtime. However, anomaly detection is challenging because anomalies are rare. Anomaly detection algorithms are typically developed assuming that only known good data are used for training and anomalies will only occur during inference. These assumptions can lead to overfitting, where an algorithm performs well on the training set but poorly during actual use.

In this work, we propose using a validation set to help regularize and reduce overfitting. With a random training-validation split, ideally, the losses on both sets should be similarly distributed. However, due to the expressiveness of neural networks, we often see the training set loss continuously decrease while the validation set loss decreases only moderately or even increases, leading to overfitting. We suggest matching the loss distributions from the training and validation sets using the Kolmogorov–Smirnov (KS) statistic. Since the KS statistic is originally non-differentiable, we designed a sigmoid-smoothed version to enable gradient descent. Additionally, we propose randomly splitting the training and validation sets each epoch, allowing the model to train on all data samples. Our techniques result in smoother decision boundaries and better generalization across various datasets, as demonstrated in the following results.



▲ Figure 1: An example showing the decision boundary without using a validation set.



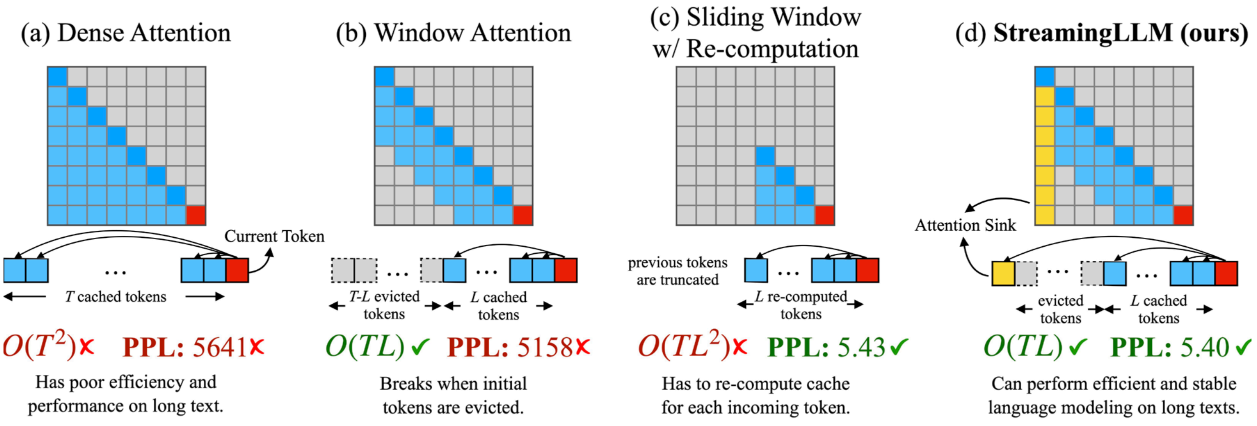
▲ Figure 2: The same example with a validation set incorporated, resulting in a much smoother decision boundary and reduced overfitting.

# Efficient Streaming Language Models with Attention Sinks

G. Xiao, Y. Tian, B. Chen, S. Han, M. Lewis

Sponsorship: MIT-IBM Watson AI Lab, Amazon, MIT Science Hub, NSF

Deploying large language models (LLMs) in streaming applications such as multi-round dialogue, where long interactions are expected, is urgently needed but poses two major challenges. Firstly, during the decoding stage, caching previous tokens' key and value states (KV) consumes extensive memory. Secondly, popular LLMs cannot generalize to longer texts than the training sequence length. Window attention, where only the most recent KVs are cached, is a natural approach, but we show that it fails when the text length surpasses the cache size. We observe an interesting phenomenon, namely an attention sink, that keeping the KV of initial tokens will largely recover the performance of window attention. In this paper, we first demonstrate that the emergence of an attention sink is due to the strong attention scores towards initial tokens as a "sink" even if they are not semantically important. Based on the above analysis, we introduce StreamingLLM, an efficient framework that enables LLMs trained with a finite length attention window to generalize to infinite sequence lengths without any fine-tuning. We show that StreamingLLM can enable Llama-2, MPT, Falcon, and Pythia to perform stable and efficient language modeling with up to 4 million tokens and more. In addition, we discover that adding a placeholder token as a dedicated attention sink during pre-training can further improve streaming deployment. In streaming settings, StreamingLLM outperforms the sliding window re-computation baseline by up to 22.2x speedup.



▲ Figure 1: StreamingLLM efficiently handles long texts by keeping initial tokens for stable attention computation, combined with recent tokens, offering consistent performance without excessive computational overhead compared to existing methods with limitations on text length or efficiency.

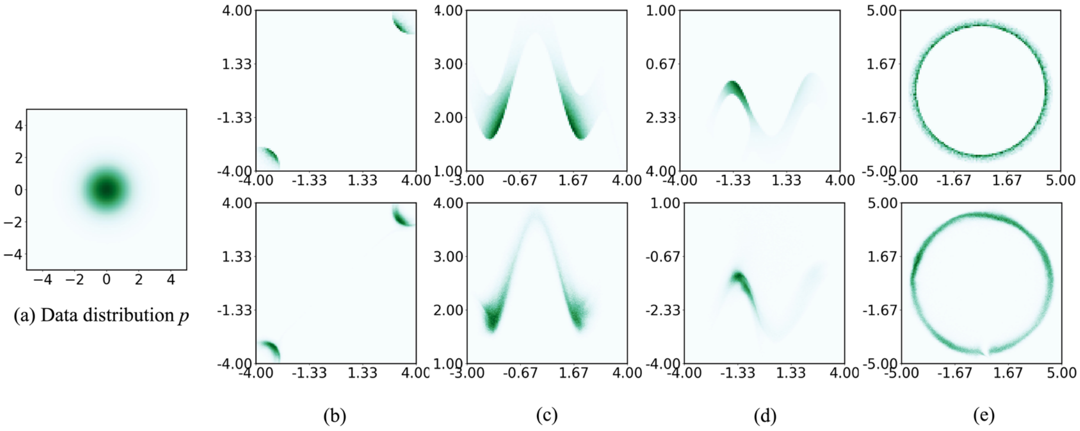
## FURTHER READING

- G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models," *Proc. 40th International Conference on Machine Learning Proc. Machine Learning Research* 202:38087-38099, 2023. Available from https://proceedings.mlr.press/v202/xiao23c.html.
- I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," 2020. *ArXiv.* /abs/2004.05150.

# Rare Event Probability Learning by Normalizing Flows

Z. Gao, L. Daniel, D. S. Boning

A rare event is defined by a low probability of occurrence. In semiconductor manufacturing, accurate estimation of such small probabilities (e.g., the rare failure rate of a process, a device, or a circuit system) is of utmost importance. Conventional Monte Carlo methods are inefficient, demanding an exorbitant number of samples to achieve reliable estimates. Inspired by the exact sampling capabilities of normalizing flows, we propose normalizing flow assisted importance sampling, termed NOFIS. NOFIS first learns a sequence of proposal distributions associated with predefined nested subset events by minimizing KL divergence losses. Next, it estimates the rare event probability by utilizing importance sampling in conjunction with the last proposal. The efficacy of our NOFIS method is substantiated through qualitative visualizations, affirming the optimality of the learned proposal distribution, as well as a series of quantitative experiments, which highlight NOFIS's superiority over baseline approaches.



(a) Data distribution $p$

(b)

(c)

(d)

(e)

▲Figure 1: (a) The heatmap represents the data generating distribution. (b)-(e) The top row displays the theoretically optimal proposal distribution, while the bottom row illustrates the learned proposal distribution learned by our Algorithm.

# On-the-Fly Learning for DNN Monocular Depth Estimation

S. Sudhakar, Z.-S. Fu, S. Karaman, V. Sze

Sponsorship: National Science Foundation, Real-Time Machine Learning program grant no. 1937501, the MIT-Accenture Fellowship, the Huang Phillips Fellowship, gift from Intel

Monocular depth estimation with deep neural networks (DNNs) is critical for resource-constrained robots to avoid using power-hungry depth sensors. Since DNNs are known to perform poorly when the deployment environment is different from its training environment, we design a computation-efficient framework where the robot can adapt a monocular depth estimation DNN to a new environment on-the-fly. This form of training is compute and storage-limited, online (sequential video inputs), and self-supervised, making it challenging to accomplish with conventional training practices. As training on every image is computation-intensive, we propose using a novel acquisition function to select a subset of images for the DNN to train on. This function for training data selection is based on a unique combination of uncertainty, diversity, and self-supervised loss quality. With reducing the frequency of training to only 2.5% of the time instead of at every timestep as is conventionally done, leveraging our acquisition function achieves 5.6% higher accuracy on the explored dataset and 3.6% higher accuracy on the unexplored dataset compared to a fixed-periodic image selection strategy using the same number of training images. In addition, we require a buffer capacity of only 9 images, making it more suitable for storage-limited platforms compared to existing works that require two orders of magnitude more capacity. Finally, we deploy MC-dropout on the final layer of the DNN to efficiently estimate the uncertainty term in the acquisition function, which reduces the computational cost of the acquisition function by up to 9.1× compared to the traditional ensemble method.

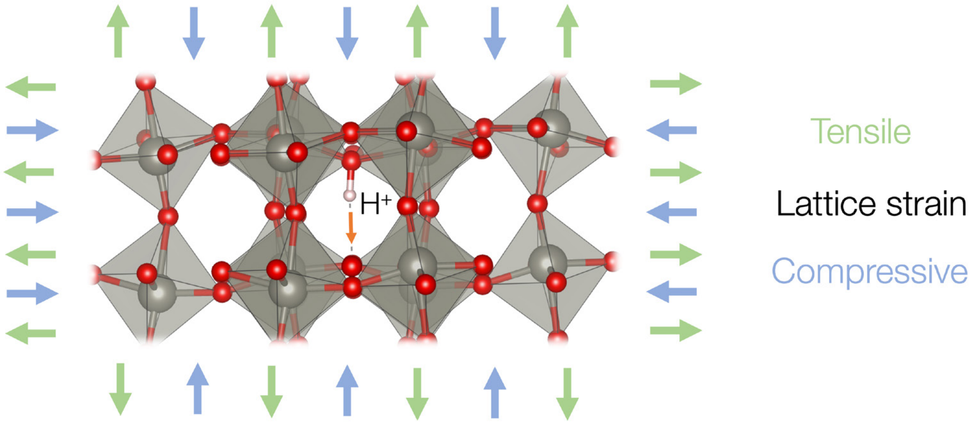
# Tungsten Bronzes forNeuromorphoc Computing: Can Lattice Strain Enhance Proton Migration in $WO_3$?

M. Siebenhofer, P. Zguns, B. Yildiz

Sponsorship: Max Kade Fellowship 2022 for M. Siebenhofer

Tungsten bronzes are insertion compounds with the general formula $MxWO_3$, where M (e.g. H, Li or Mg) is intercalated into $WO_3$. These compounds have a variety of applications, ranging from superconducting or electrochromic materials to catalysts and analog programmable resistors for artificial neural networks. For its application in neuromorphic computing, where the conductivity of $WO_3$ is modulated by ion intercalation, fast ion diffusion is key to obtain homogeneous ion distribution and fast switching between distinct resistive states. However, exact migration pathways in $WO_3$ are not well understood and novel strategies to enhance proton movement are desired.

In this contribution, we report the results of ab-initio investigations on proton migration processes in $WO_3$. We identify favourable binding sites for protons and migration pathways through the lattice. Using the nudged-elastic-band method, we investigate energetic barriers for proton hopping in $WO_3$ and evaluate their modulation in structures under tensile and compressive strain. If successful, targeted introduction of lattice strain (e.g. by doping) may be essential to improve both neuromorphic and electrochromic devices.



▲ Figure 1: Tensile and compressive lattice strain present a potential tool to modulate proton migration in $WO_3$.