Breaking the Conversion Wall in Mixed-Signal Systems Using Neuromorphic Data Converters

Loai Danial and Shahar Kvatinsky

Andrew and Erna Viterbi Faculty of Electrical Engineering,
Technion - Israel Institute of Technology, Haifa 3200003, ISRAEL, Email: sloaidan@campus.technion.ac.il

***Abstract*— Data converters are ubiquitous in mixed-signal systems, becoming the computational bottleneck in traditional data acquisition and emerging neuromorphic systems. Unfortunately, conventional Nyquist data converters trade off speed, power, and accuracy. Therefore, they are exhaustively customized for special purpose applications. Furthermore, intrinsic real-time and post-silicon variations dramatically degrade their performance along with the CMOS technology downscaling. Here, we review on our neuromorphic analog-to-digital (ADC) and digital-to-analog (DAC) converters that are trained using the online stochastic gradient descent algorithm to autonomously adapt to different design specifications, including multiple full-scale voltages, number of resolution bits, and sampling frequencies. We demonstrate the feasibility of our converters by simulations and preliminary experiments using memristive technologies. We show collective properties of our converters in application reconfiguration, logarithmic quantization, mismatches calibration, noise tolerance, and power optimization. The proposed data converters achieve a superior figure-of-merit (FoM) of 1 fJ/conv.**

Keywords*— *Analog-to-digital conversion, digital-to-analog conversion, memristors, machine learning, neuromorphic.

I. Introduction

The evolution of data-driven edge devices towards the internet-of-things era has paved the way to emerging interacting and varying applications where data converters play a key role [1]. With the advent of mixed-signal neuromorphic systems, there is an ever-growing demand for accurate, fast, and energy-efficient data converters [2]. Unfortunately, the intrinsic speed-power-accuracy tradeoff in Nyquist's analog-to-digital converters (ADCs) constitutes a computational bottleneck [3]. Furthermore, with the downscaling of technology anticipated by Moore's law, this tradeoff is exacerbated due to alarming deep sub-micron effects [4]. Those effects are poorly handled with particular technology-dependent design techniques that overload data converters with enormous overhead, degrading their performance [3]. Conventional data converters lack design standards and are customized with sophisticated design flow for special purpose applications, from high-speed, to high-resolution, to low-power applications [3]. These methods require exhaustive characterization and massive validation, and are expensive to develop, with a long time-to-market.

This paper reviews a different systematic approach, beyond the conventional Nyquist's data conversion paradigm, inspired by artificial neural networks (ANNs) to design general purpose data converters. We propose that the converted data be used to train the converter to autonomously adapt to the exact specifications of the running application, including multiple full-scale voltages, number of resolution bits, sampling frequencies, quantization scale, and adjust to environmental variations. This approach will reduce the time to market, efficiently scale with newer technologies, drastically reduce its cost, significantly standardize the design flow, and enable a generic architecture [5].

The proposed trainable data converters utilize machine learning (ML) algorithms to train a memristor based ANN architecture [3][5-6]. Memristors are widely adopted in the design of ANNs due to their analog storage properties, energy efficiency, and dimensions [7]. These characteristics allow for synapse-like behavior, where the memristor's conductance implements the synaptic weight. It is trained to obtain high-precision, high-speed, low-power, a simple cost-efficient, and reconfigurable single-channel converter that breaks through the speed-power-accuracy tradeoff [3].

II. The Conversion Wall in Mixed-Signal Systems

In this section, we discuss the limitations of conventional Nyquist's data converters in meeting the demands of emerging data-intensive applications.

A. Nyquist ADCs/DACs Bottleneck

The Nyquist's ADC is comprised of a sampler that discretely samples the continuous-time signal at a constant rate equal twice the signal bandwidth, and a quantizer that converts the sampled value to the corresponding discrete-time *N*-bit resolution binary-coded form. Data converters are widely used in two computing systems, as shown in Fig. 1:

- *Data acquisition systems* (DAQ): where an ADC can be a dominant interface between the analog data, produced by sensors, and the digital domain, processed by a digital signal processor (DSP), and vice versa for DAC. Such systems are traditionally used in sensory nodes and now can be an embedded part of an ultra-low power edge device for smart sensors applications in internet-of-things network. In conventional DAQ systems, 6% of the power is dissipated on the conversion [8], while the rest is consumed by the memory and processing units at the DSP. Thus, ADCs load digital systems with non-negligible power, latency, and imprecision overheads.
- *Mixed-signal neuromorphic systems*: in attempts to mitigate the von Neumann bottleneck and the memory

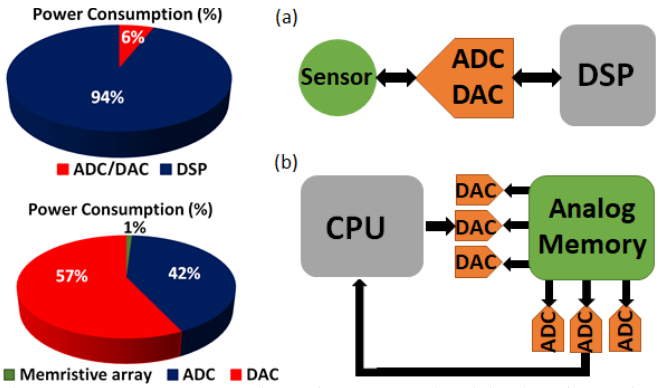

Fig. 1. Data converters are becoming computational bottlenecks as they evolved from: (a) traditional data acquisition systems where they convert analog information from sensors to digital signal processors and vice versa, with 6% power consumption, to (b) emerging neuromorphic systems that use analog in-memory computing for multiply-accumulate operations and digital neural activation functions, with 99% power consumption [1-2].

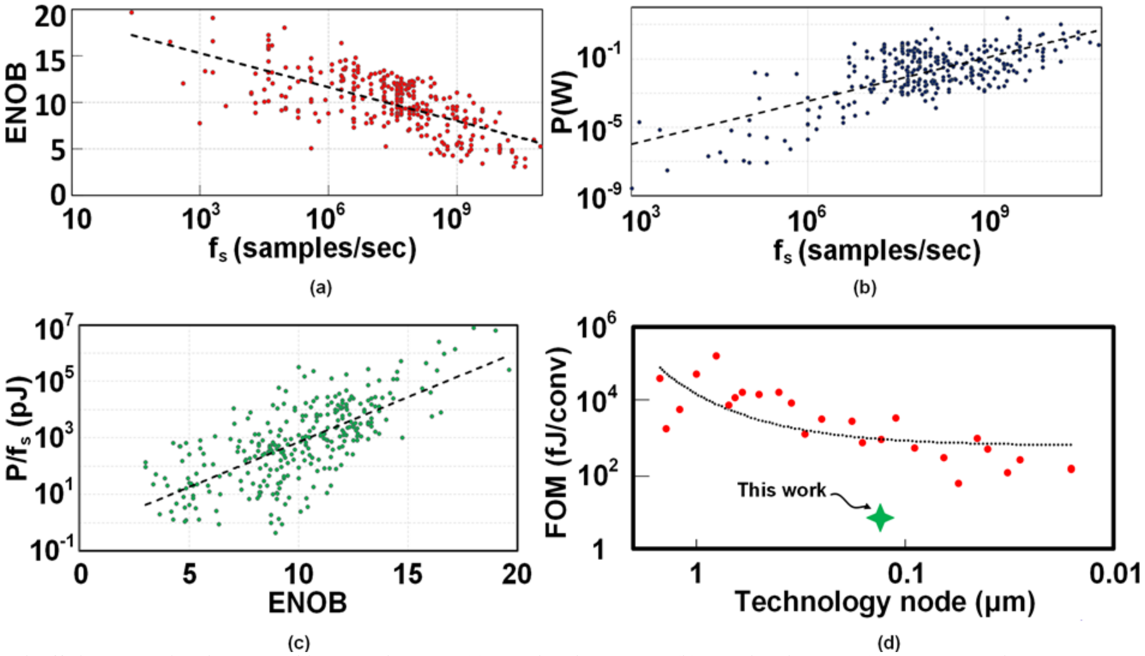

Fig. 2. Tradeoffs in conventional ADC architectures between (a) speed and accuracy, (b) speed and power, (c) accuracy and energy, as reported in [9]. Since the power-accuracy tradeoff depends on the limitations of the underlying architectures, the energy-accuracy is independent of the architecture and shows the tradeoff accordingly. (d) Average figure-of-merit (FOM) evolution versus technology node scale-down of the different ADC architectures and specifications reported in the ADC survey [9]. Overall, the FOM improves with the technology scale-down. However, the asymptotic slowdown in the last decade is shown by the trendline. The green star shows the achieved FOM of this work.

wall in energy-constrained digital processing systems, neuromorphic computing brings processing elements closer to the memory by providing analog pre-processing capabilities on the data collected from sensors. This requires data converters to couple neuromorphic circuits to digital components. The conversions, however, overwhelm a large portion (~99%) of the total power consumption and chip area of neuromorphic accelerators in end-point devices [2]. Thus, beyond von Neumann architectures push the computational bottleneck from memory towards data converters [1-2], constituting a conversion wall.

B. The Speed-Power-Accuracy Tradeoff

The quality of a system is considered ideal when it achieves high speed and accuracy with a low power drain. In practice, however, the resolution decreases as the conversion rate increases, and greater power consumption is required to achieve the same resolution. Device mismatch is the dominant factor affecting system accuracy [3]. Larger devices are necessary to improve accuracy, but capacitive loading of the circuit nodes increases as a result and greater power is required to attain a certain speed. Moreover, four loss mechanisms limit the ADC resolution: quantization noise, jitter, comparator ambiguity, and thermal noise.

Quantization noise is the only error in an ideal ADC. Jitter is a sample-to-sample variation of the instant in time at which sampling occurred. Additionally, the conversion speed is limited by the ability of the comparator to make assertive decisions regarding the relative amplitude of the input voltage [4]. This limitation is called comparator ambiguity and it is related to the speed of the device, f_T. As a result of these limitations, approximately one bit of resolution is lost each time the sampling rate doubles [3]. Whereas distortions and mismatches can be somehow compensated for, thermal white noise cannot. Lowering the noise floor by a factor of two in purely thermal-noise limited circuits would quadruple the power consumption [3]. The speed-power-accuracy tradeoff is illustrated in Fig. 2(a-c); it is based on data that we have processed from Stanford's ADC survey [9]. It has resulted in a wide range of ADC architectures optimized for special purpose applications, from high speed, to high resolution, to low power applications, as shown in Fig. 3(b).

C. Moore's Law of Data Converters

When comparing ADCs with different specifications, a global numerical quantity known as a figure of merit (FOM) is fairly used to characterize the performance of each ADC relative to its alternatives. One of the widely used FOMs is

$$FOM = \frac{P}{2^{ENOB} \cdot f_s}\left[\frac{J}{conv}\right], \tag{1}$$

and relates the ADC power dissipation during conversion, *P*, to its performance in terms of sampling frequency, f_s, and effective number of resolution bits *(ENOB)*. Lower *FOM* values will result in better ADC performance, and it best captures the speed-power-accuracy tradeoff [3]. The *FOM* evolution also best describes Moore's law of ADCs. Technology scaling improves sampling frequencies, because f_T allows for faster operation [3]. However, it is limited by the comparator ambiguity. In the same context, the impact of technology scaling on power dissipation optimization is also limited by the supply voltages, and by leakage currents that lead to an increase in the power consumption required to maintain ENOB [4]. Furthermore, manufacturing variations and device mismatches in advanced technologies limit the resolution of converters [4]. Thus, the speed-power-accuracy tradeoff is becoming dramatically more severe with technology downscaling, pushing future data converters out of the application band of interest [3]. The FOM evolution is shown in Fig. 2(d) based on [9]. The figure shows an overall improvement in the FOM over the technology nodes. However, the improvement has slowed down significantly and performance has recently saturated [3].

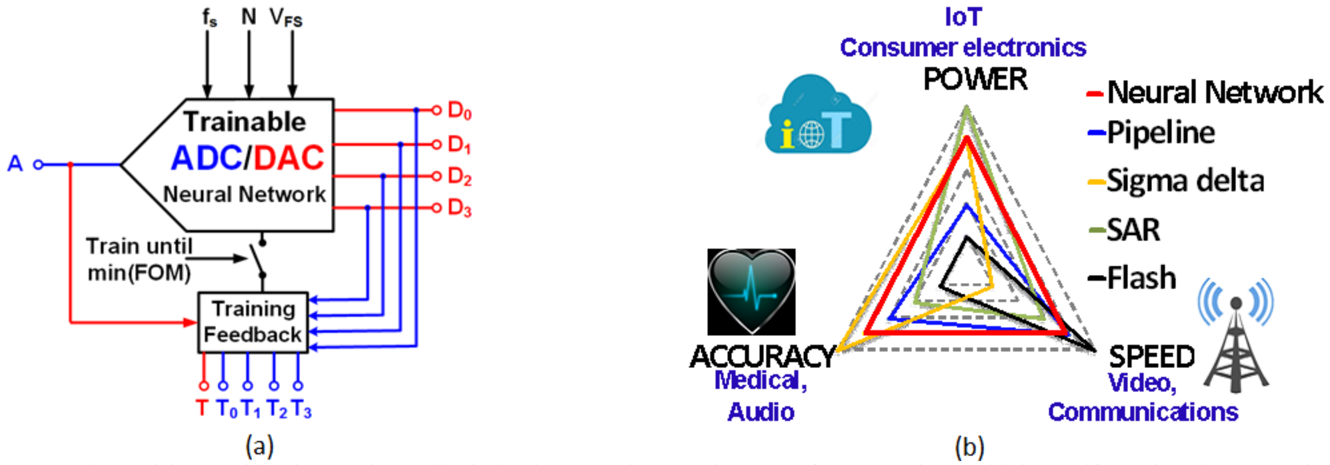

Fig. 3. Scheme of the trainable 4-bit ADC/DAC (blue/red path) ANN. The network receives $f_s, V_{FS}, N,$ and is provided a specific teaching dataset $\{T\}$ for online training. Training continues until the converter achieves the optimal FOM. (b) Spider diagram of ADC architectures (colored), tradeoff, and associated applications (blue text), along with our general purpose, ANN architecture (yellow), which breaks through the tradeoff as shown by the balanced (red) line.

III. Neuromorphic Data Converters

Data conversion could be seen as a pattern classification task and thus could be easily solved using ANNs and trained by ML algorithms. We propose trainable data conversion architectures for general purpose applications [5], as shown in Fig. 3(a). A set of parameters is determined to meet the requirements of the running application. First, f_s is determined, followed by the number of resolution bits N, followed by the full-scale voltage V_{FS}, which specifies the converter amplitude dynamic range $(V_{FS}/2^N)$. Then, the converter is trained by a supervised ML algorithm, called stochastic gradient descent (SGD). The training is done online to optimize the ENOB and power dissipation, where the correct digital labels corresponding to the analog input are supplied and compared to the actual digital outputs, and vice versa for DAC. This procedure is equivalent to a dynamic *FOM* optimization. The technique is not exclusive to application reconfiguration but can also be used for device mismatch self-calibration, power optimization, and noise tolerance with generic methodology. In this section, we present the neuromorphic conversion paradigm including: architectures, fundamentals, theory, and training algorithms.

A. Neuromorphic ADC

The ANN architecture shown in Fig. 4(a) implements a trainable four-bit ADC using programmable synaptic weights [3] ($W_{i,j}$ is the conductance between a pre-synaptic neuron j and a post-synaptic neuron i), where V_{in} is the analog input and $D_3D_2D_1D_0$ is the corresponding digital form ($i=3$ is the MSB), and each bit (neuron product) has either zero or full-scale voltage, and V_{ref} is a reference voltage equal to the LSB. This network operates successively; first, the MSB (D_3) is determined, and is forward propagated to the LSBs (D_2, D_1, D_0) to be next recursively determined. Thus, this network could be considered as a recurrent single-layer ANN. The synapses are realized using NMOS, PMOS, capacitor and memristor, with the transistor gates connected to a common enable input e, as shown in Fig. 4(b) [7]. The neurons comprise of an inverting op-amp and capacitor for integration and a comparator for decision making. Synaptic weights are tuned to minimize the mean square error (MSE) by using the SGD. The training continues until the error falls to $E_{threshold}$, a predefined constant that defines the learning accuracy.

B. Neuromorphic DAC

The single-layer ANN architecture [6], shown in Fig. 4(c) converts the four-bit digital input code ($B_3B_2B_1B_0$) to an analog output (A) where it is trained as a binary-weighted DAC. The four memristive synaptic weights W_i collectively integrate the inputs through the single op-amp neuron to produce the output. This output is compared to the analog teaching labels using a pulse width modulation (PWM) feedback, which regulates the value of the weights according to the SGD algorithm. Similarly, the feedback is disconnected after the training is complete (when $E < E_{threshold}$).

C. Neuromorphic Pipelined ADC

Four-bit resolution is insufficient for practical applications [4-6], while direct scaling of the proposed architectures is challenging due to the quadratic increase in number of synaptic weights, large area, high power consumption, longer training time, limited sampling frequency, and physical challenges of memristors. Aiming to handle these challenges, we propose a scalable and modular neural network ADC architecture (*e.g.* eight bits) based on a pipeline of four-bit converters [10], as shown in Fig. 4(d). The pipelined ADC preserves the inherent collective advantages of neuromorphic ADCs, while approaching higher resolution and throughput. The sub-ADC coarsely quantizes the sampled input V_{in} to the digital code $D_7D_6D_5D_4$. The output of the sub-ADC is converted back to an analog signal A by the DAC. Next, this output is subtracted from the held input to produce a residue Q. This residue is sent to the next stage of the pipeline, where it is stored in D-flipflop registers and quantize $D_3D_2D_1D_0$. The stages of this architecture are trained simultaneously, by controlling the switches S_i, according to the SGD algorithm.

IV. Generalizing the Concept

Our proposed data converters are simulated in SPICE using memristors fitted by the VTEAM model [11] to a Pt/HfO$_x$/Hf/TiN RRAM [12], achieving a FOM value of 1fJ/conv. Furthermore, we experimentally demonstrate the feasibility of our converters using the mature technology of two-terminal floating-gate memristive devices fabricated in 180nm CMOS process and precisely operated at the sub-threshold mode to achieve 65 resistive levels [13-14]. Besides the pipeline architecture, we are currently investigating

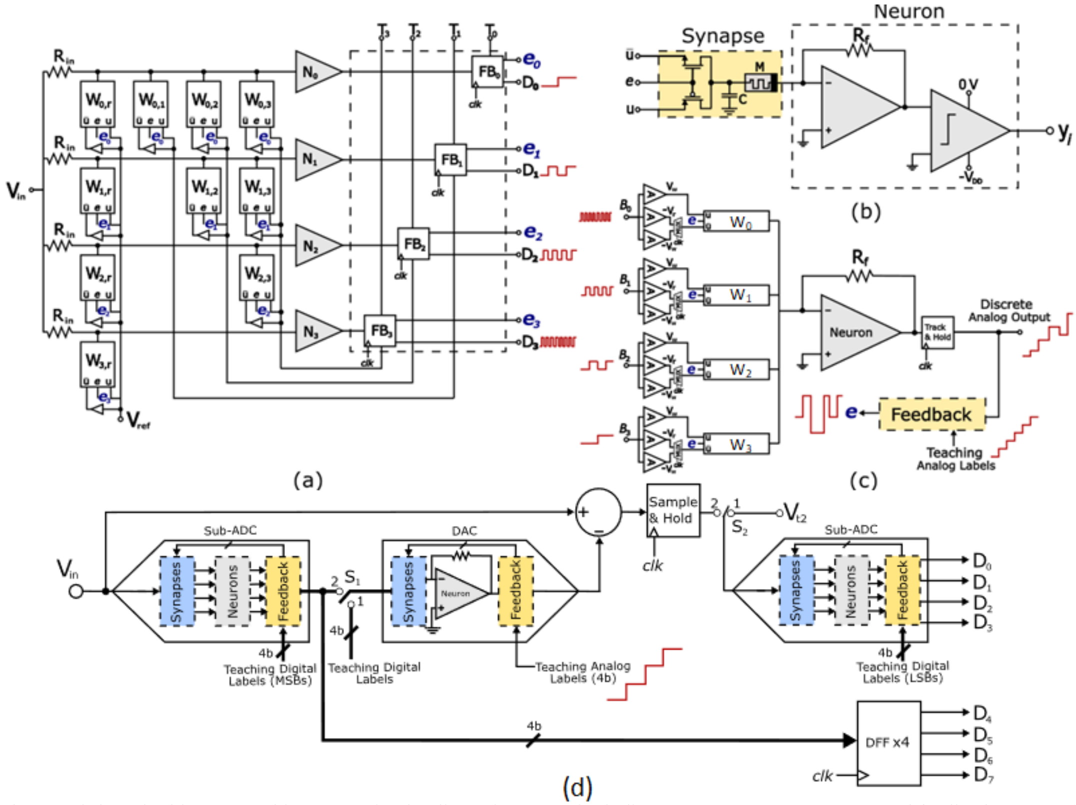

Fig. 4. (a) Neural network-based 4-bit ADC architecture trained online using SGD, including synapses $W_{i,j}$, neurons N_i, and feedback FB_i [3], (b) A memristive synapse connected to an artificial neuron implemented as an inverting opAmp and a comparator. (c) ANN-based 4-bit DAC architecture, including synapse W_i, a neuron implemented as an opAmp, and a PWM-based feedback circuit [6] for the SGD. (d) Proposed architecture of a two-stage pipelined ADC trained online using SGD. The first stage consists of four-bit single-layer neural network sub-ADC and DAC. The second stage consists of another four-bit ANN ADC. Both stages operate simultaneosuly to increase the conversion throughput and their intermediate results are temporarily stored in D-flipflop registers.

mixed-signal techniques to scale up our concept (sigma-delta modulation [15] and logarithmic quantization [16]). Moreover, deep neural networks (DNNs) with offline training have been utilized using RRAM [17]. However, the degradation in data retention limits the accuracy and sacrifices power, area, and latency by increasing the number of network layers. The continuous unsupervised learning by spiking neural networks (SNNs) using high-endurance RRAM can be employed to obtain adaptive data converters.

In the broader scope of using various network paradigms and memristive technologies, we generalize our concept of neuromorphic data converters to include a family of converters from/to different domains: analog, digital, spike, time, and frequency. Finally, we believe that such family of data converters will be embedded in neuromorphic accelerators (DNNs/SNNs) to break the conversion wall.

V. Conclusion

This paper reviews the neuromorphic data conversion paradigm which employs memristors and can be trained using learning algorithms to break through the speed-power-accuracy tradeoff in conventional Nyquist's data converters and the conversion wall in mixed-signal systems. We first presented our four-bit ADC and DAC. Based on that, we presented a large-scale pipelined ADC. Finally, we discussed the recent approaches in this field and extended our concept to include a family of data converters that could be used as an interface to any type of neuromorphic system.

Acknowledgements

This research was partially supported by the Israeli PBC Fellowship, by the Viterbi Graduate Fellowship, and by the POLAK fund for applied research at the Technion.

References

[1] A. Fayyazi *et al.*, “An Ultra Low-Power Memristive Neuromorphic Circuit for IoT Smart Sensors,” *IoT-J*, Vol. 5, No. 2, pp. 1011–1022, Apr. 2018.

[2] X. Liu *et al.*, “RENO: A High-Efficient Reconfigurable Neuromorphic Computing Accelerator Design,” pp. 1–6, *DAC*, Jun. 2015.

[3] L. Danial *et al.*, “Breaking Through the Speed-Power-Accuracy Tradeoff in ADCs Using a Memristive Neuromorphic Architecture,” *TETCI*, Vol. 2, No. 5, pp. 396-409, Oct. 2018.

[4] Y. Chiu *et al.*, “Scaling of Analog-to-Digital Converters into Ultra-Deep-Dubmicron CMOS,” *CICC,* pp. 375-382, Sep. 2005.

[5] L. Danial, and S. Kvatinsky, “Real-Time Trainable Data Converters for General Purpose Applications”, pp. 34-36, *NanoArch*, July 2018.

[6] L. Danial *et al.*, “DIDACTIC: A Data-Intelligent Digital-to-Analog Converter with a Trainable Integrated Circuit using Memristors,” *JETCAS*, Vol. 8, No. 1, pp. 146-158, March 2018.

[7] D. Soudry *et al.*, “Memristor-Based Multilayer Neural Networks With Online Gradient Descent Training,” *TNNLS*, Vol. 26, No. 10, pp. 2408-2421, Oct. 2015.

[8] V. Konstantakos *et al.,* “Energy consumption estimation in embedded systems,” *TIM*, vol. 57, no. 4, pp. 797–804, Apr. 2008.

[9] B. Murmann, “ADC Performance Survey 1997-2019,” [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html

[10] L. Danial, K. Sharma, and S. Kvatinsky, “A Pipelined Memristive Neural Network Analog-to-Digital Converter,” pp. 1-5, *ISCAS*, May 2020.

[11] S. Kvatinsky *et al.*, “VTEAM: A General Model for Voltage-Controlled Memristors,” *TCASII*, Vol. 62, No. 8, pp. 786-790, Aug. 2015.

[12] J. Sandrini *et al.*, “Effect of Metal Buffer Layer and Thermal Annealing on HfOx-based ReRAMs,” *ICSEE*, pp. 1-5, Nov. 2016.

[13] L. Danial *et al.,* **“**Two-Terminal Floating-Gate Transistors with a Low-Power Memristive Operation Mode for Analogue Neuromorphic Computing,**”** *Nature Electronics,* Vol. 2, pp. 596-605, December 2019**.**

[14] L. Danial *et al.*, “Modeling a Floating-Gate Memristive Device for Computer Aided Design of Neuromorphic Computing,” pp. 472-477, *DATE,* March 2020.

[15] L. Danial *et al.*, “Delta-Sigma Modulation Neurons for High-Precision Training of Memristive Synapses in DNNs,” pp. 1-5, *ISCAS,* May 2019.

[16] L. Danial *et al.*, “Logarithmic Neural Network Data Converters using Memristors for Biomedical Applications,” pp. 1-4, *BioCAS*, Oct. 2019.

[17] W. Cao *et al.*, “NeuADC: Neural Network-Inspired RRAM-Based Synthesizable Analog-to-Digital Conversion with Reconfigurable Quantization Support,” pp. 1456-1461, *DATE*, March 2019.