Exploiting Inherent Error Resiliency of Deep Neural Networks to Achieve Extreme Energy Efficiency Through Mixed-Signal Neurons

Baibhab Chatterjee, *Student Member, IEEE*, Priyadarshini Panda, *Student Member, IEEE*, Shovan Maity, *Student Member, IEEE*, Ayan Biswas, *Student Member, IEEE*, Kaushik Roy, *Fellow, IEEE*, and Shreyas Sen, *Senior Member, IEEE*

***Abstract*— Neuromorphic computing, inspired by the brain, promises extreme efficiency for certain classes of learning tasks, such as classification and pattern recognition. The performance and power consumption of neuromorphic computing depend heavily on the choice of the neuron architecture. Digital neurons (Dig-N) are conventionally known to be accurate and efficient at high speed while suffering from high leakage currents from a large number of transistors in a large design. On the other hand, analog/mixed-signal neurons (MS-Ns) are prone to noise, variability, and mismatch but can lead to extremely low-power designs. In this paper, we will analyze, compare, and contrast existing neuron architectures with a proposed MS-N in terms of performance, power, and noise, thereby demonstrating the applicability of the proposed MS-N for achieving extreme energy efficiency (femtojoule/multiply and accumulate or less). The proposed MS-N is implemented in 65-nm CMOS technology and exhibits >100× better energy efficiency across all frequencies over two traditional Dig-Ns synthesized in the same technology node. We also demonstrate that the inherent error resiliency of a fully connected or even convolutional neural network can handle the noise as well as the manufacturing nonidealities of the MS-N up to certain degrees. Notably, a system-level implementation on CIFAR-10 data set exhibits a worst case increase in classification error by 2.1% when the integrated noise power in the bandwidth is $\sim 0.1\mu$ V2, along with $\pm 3\sigma$ amount of variation and mismatch introduced in the transistor parameters for the proposed neuron with 8-bit precision.**

***Index Terms*— Artificial neural network (ANN), CMOS, high-speed neuromorphic computing, low energy, mixed signal (MS).**

I. Introduction

THERE has always been a huge gap between the energy efficiencies of the human brain and the von-Neumann model of computing which dominates the consumer market. Software simulations of the brain of a mouse with 2.5 million neurons are 9000 times slower than real-time when running on a personal computer [1]. Moreover, it consumes 400-W power as compared to the paltry 10 mW of a biological mouse brain. To emulate a human brain (100 billion neurons, 20-W power), a supercomputer requires 500 MW [2] of power. Such large differences in energy efficiency, coupled with the rebirth of the deep learning paradigms in the past decade, have forced researchers all across the world to look into alternate models of computation.

Neuromorphic computing, which loosely models the brain and uses artificial neural networks (ANNs) for computation, has found significant success in applications involving image and pattern recognition, miniaturized autonomous robotics [3], and neural prosthesis [4]. However, the performance and energy efficiency of neuromorphic computing depend heavily on the choice of the neuron architecture, operating frequency, resolution, and accuracy required. Digital implementation of a neuron has been the preferred choice for computing in SpiNNekar [5] and TrueNorth [6] projects due to the excellent noise immunity, variability tolerance, and technology scaling of digital designs. While SpiNNekar had no dedicated hardware for its neuron model and consumed 1-W power, IBM's TrueNorth had a dedicated point neuron model for its 1 million neurons (256 synapses each) and consumed only 65 mW. TrueNorth's primary design emphasis was on minimizing active as well as static power for a spiking neural network (SNN) by using an event-driven architecture [7] and having a compact physical design for increased parallelism on a 28-nm process that is well known for power efficiency.

Analog/mixed-signal (MS) computational models can be easily affected by noise, variability, and mismatch, which makes its energy efficiency less attractive. In an interesting study of digital versus analog circuits for computing [8], it was shown that digital circuits perform better for high signal-to-noise ratio (SNR) applications (>60 dB). However, if SNR requirements are relaxed, analog computation could be orders of magnitude more energy and area efficient. This is because analog macros, for example, a multiplier, use only one differential pair of MOSFETs which is sufficient to represent the circuit dynamics using intrinsic device parameters. On the other hand, a digital multiplier computes the same dynamics using ∼1000 transistors, the combined static leakage of

Manuscript received July 8, 2018; revised November 22, 2018; accepted January 7, 2019. Date of publication March 1, 2019; date of current version May 22, 2019. This work was supported by the National Science Foundation Secure and Trustworthy Cyberspace (SaTC), Division of Computer and Network System (CNS) under Grant 1719235. *(Corresponding author: Baibhab Chatterjee.)*

B. Chatterjee, P. Panda, S. Maity, K. Roy, and S. Sen are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: bchatte@purdue.edu; pandap@purdue.edu; maity@purdue.edu; kaushik@purdue.edu; shreyas@purdue.edu).

A. Biswas is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: ayan_biswas@eecs.berkeley.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2019.2896611

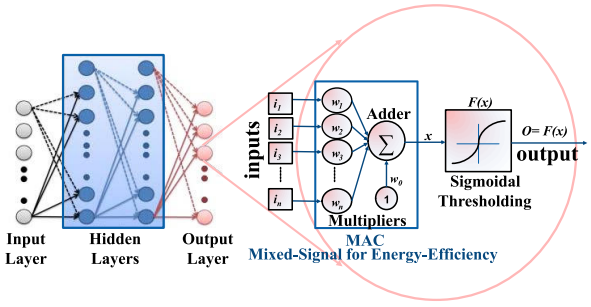

Fig. 1. Neuron model in a feedforward CNN. MAC with thresholding.

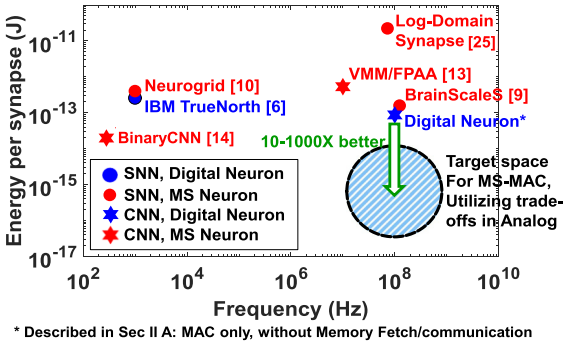

Fig. 2. Energy efficiency per synapse for previously reported related works, a Dig-N (Section II-A) and target MS design.

which could be comparable to the bias current of analog in scaled technologies. It is interesting to note that von-Neumann architectures depend on the high accuracy of a multidigit representation that necessitates a digital implementation. ANN, however, has multiple connections from inputs to output due to its distributed nature, and hence, the noise and variability of analog transistors can be tolerated to some extent due to this inherent error resiliency.

A. Related Work

Currently, several research groups are working on the implementation of large-scale SNNs with analog/MS neuron (MS-N) model. The BrainScaleS project (HICANN chip) [9] at Heidelberg University aims to develop a brainlike system that uses analog computation and digital asynchronous communication and runs 1000–10 000 times faster than real time. The design consists of 200 000 analog neurons with 40 million addressable synapses and consumes about 1 kW at 125-MHz frequency. The Neurogrid project at Stanford [10] also uses a mixed design approach and reduces transistor count further by sharing synapses and dendritic tree circuits [11]. Neurogrid has one million neurons each with 8000 synapses and consumes 3.1 W for real-time brain computations. However, both of these designs are for SNNs that aims to model the spiking neural activities by using a current-switching neuron architecture and requires complex learning models such as spike-timing-dependent plasticity. Convolutional neural networks (CNNs), on the other hand, can employ simple backpropagation algorithms using a multiply-and-accumulate (MAC) model [12] (shown in Fig. 1), which is more suitable than SNNs in pattern recognition applications and in scenarios involving generative-adversarial networks. In [13], a large-signal current-mode MAC implementation is demonstrated. However, in a large-signal implementation, the bandwidth of the design keeps on changing with varying bias currents, and hence, the frequency of the input signals is limited by the minimum large-signal current for any practical application. A small-signal implementation, on the other hand, would be much more attractive in terms of the power, bandwidth, and scalability. Also, a differential voltage-mode architecture would help in reducing the impact of common-mode noise present in the system. Recently, in [14], a 3.8-μJ/inference CNN processor with on-chip memory was presented, which utilizes an MS approach for MAC operations, with XNOR gates as multipliers and switched capacitors serving as additive elements. However, this architecture is only applicable for BinaryNets, where the weights and activations are constrained to be ± 1. For applications with real-valued multiplication, the energy/MAC would increase exponentially. The energy efficiencies of previously reported work are compared with our target MS-N in Fig. 2.

This paper presents three different architectures, leading to a compact, differential MS implementation of a MAC-based neuron in Section III, which operates in small-signal mode with linear power-bandwidth characteristics [unlike digital neurons (Dig-N)] and can help to build an extremely energy-efficient CNN. The key contributions of this paper are listed as follows.

1) *Power-bandwidth-scalable MS-N for CNN:* This paper focuses on the extreme power and energy efficiency of an MS-N (over Dig-N) that can be achieved by judiciously utilizing various tradeoffs (energy versus frequency, energy versus linearity, and energy versus total integrated noise/variations) and strategies in analog design. To the best of our knowledge, this is the first work that extensively analyzes various tradeoffs in an MS-N for energy efficiency.
2) *Proposed DFE-inspired MS-N with bandwidth extension that can architecturally support both binary and nonbinary multiplications:* The design of a small-signal MS-N with resistive feedback is presented in Section III, which helps to achieve much better energy efficiency (ratio of power to bandwidth) by extending the bandwidth. The MS-N architecture is inspired by decision feedback equalizers (DFEs) employed in wireline communication [15] and can support nonbinary multiplications, unlike previous implementations of MS-N [14]. The bandwidth extension technique also allows the W/L ratio of the input transistors to increase, thereby reducing the effects of mismatch at no additional power cost. This results in a subfemtojoule MAC operation that is $\sim$100$\times$ improvement over the state of the art. Lower energies at the neuron level directly result in lower computation power/ops at the network level for a fixed architecture and bit precision. An analysis of the communication/interfacing and memory fetch power is out of the scope of this paper.
3) *Proof-of-concept simulations in system level:* A system-level analysis of the inherent error resiliency

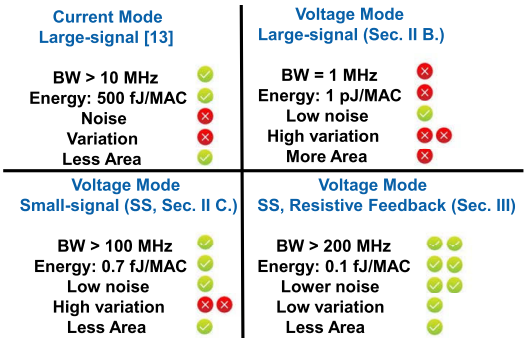

Fig. 3. Comparison of current-mode MS MAC [13] with various voltage-mode MS MACs presented in this paper.

of neural networks is illustrated, justifying how a low/medium-complexity CNN/ANN for Internet of Things/healthcare applications [16]–[22] can tolerate the effects of noise and mismatch in the MS-N to a considerable degree.

In essence, the proposed MS-N achieves order(s) of magnitude better energy efficiency as compared to digital and traditional analog architectures but can suffer from noise and variations. A method to reduce those nonidealities has been proposed, and the classification errors resulting from analog computation have been quantified in a system-level application. Such MS-N architectures are most useful for smaller networks but could be made part of larger networks where multiple local tiles/islands of analog computing units are connected with digital interfaces. In this paper, we primarily focus on applications involving low- to medium-complexity networks as shown in Section VI.

The rest of this paper is organized as follows. Section II describes three different implementations of a neuron—a fully Dig-N, an MS-N operating in the large-signal mode, and an MS-N operating in the small-signal mode. Section III depicts the proposed MS-N with increased bandwidth and reduced offset. The specific advantages and disadvantages of these neurons are shown in Fig. 3 and will be discussed in detail in Sections II and III. Section IV presents the comparison between Dig-N and the small-signal MS-N, while Section V presents a detailed discussion on the tradeoffs and theoretical limits of the MS-N shown in Sections II and III. A system-level application of the MS-N is presented in Section VI, wherein the error resiliency of a CNN and a fully connected network is demonstrated separately. Section VII compares our design with other state-of-the-art neuron architectures. We conclude the work in Section VIII by summarizing our major contributions.

II. Neuron Architectures: Dig-N and MS-N

It is well-established that analog design is superior to digital in terms of power and area for applications that require <8-bit precisions [8]. It is also indicated in [23] that >8-bit fixed point precision is redundant for most ANN applications. In this paper, our target application is a classification problem for digit/image recognition, using the MNIST data set [24] for handwritten digits and the CIFAR-10 data set [25] for

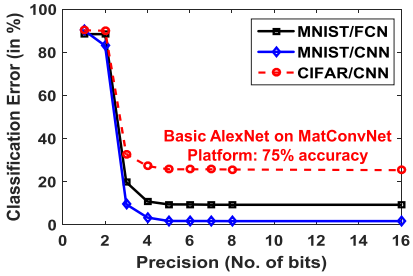

Fig. 4. System-level precision analysis for target applications.

images. The CNN and fully connected neural network (FCN) architectures are used at the system level, as will be shown in Section VI. Fig. 4 shows the classification error for the applications and different network architectures as a function of the fixed-point bit precision. It can be observed that the classification error does not increase significantly from the baseline if the precision is reduced from 16 to 6 bits. The error rate starts to increase significantly at 3-bit precision. 2-bit precision results in a classification error of >80%. Based on these numbers, we have limited ourselves with digital and MS-Ns having a precision in the range of 3–8 bits. The CNN architecture is assumed to be LeNet for MNIST and basic AlexNet for CIFAR-10, both of which are highly susceptible to nonidealities in the neuron (noise and mismatch in MS-N) due to the lack of regularization techniques such as dropout and data augmentation. This assumption leads to the high baseline error rate as shown in Fig. 4.

A. Digital Neuron

The basic functionality of a MAC-based neuron is to evaluate the weighted sum of input signals followed by a thresholding function for activation. Hence, based on the nomenclature used in Fig. 1, we can write the output of a neuron as $o = F(\Sigma x_i w_i)$ for $i = 1, 2, 3, \ldots, n$ where n is the total number of synapses, F is the activation function, which can be hard-limiting (e.g., step function) or soft-limiting (e.g., log/tan-sigmoid or rectified linear function). w_i is the weight corresponding to the ith multiplier having input voltage x_i. The number of bits (N) in w_i or x_i defines the precision of the MAC architecture. An 8-bit MAC needs an 8×8 bit multiplier and 17-bit adder, while a 3-bit MAC needs a 3×3 bit multiplier and 7-bit adder.

We have synthesized an 8×8 bit Wallace tree (WT) multiplier with a 17-bit ripple-carry adder in 65-nm CMOS technology, along with comparators for activation logic. The 3-bit version of the same design uses a 3×3 bit WT multiplier and a 7-bit ripple carry adder. A carry look-ahead adder or a carry-save adder does not result in significant speed advantage at such low precisions at the expense of more hardware. Although WT multipliers are fast, they consume higher power than most other multiplier architectures. For this reason, we have also synthesized 8×8 bit and 3×3 bit array multipliers (AMs) that consume less power. The number of different cells and transistors in an N-bit digital MAC (both WT and AM) is given in Table I. Unlike analog implementations that rely on intrinsic device dynamics, digital logic

TABLE I
NUMBER OF CELLS AND TRANSISTORS IN N-bit DIGITAL MAC

N	Number of AND gates	Number of (HA, FA)[a]	Total Number of transistors
3	9	(2,8)-WT[c], (3,3)-AM[d]	426-WT, 234-AM
4	16	(3,14)-WT, (4,8)-AM	738-WT, 504-AM
5	25	(4,22)-WT, (5,15)-AM	1146-WT, 870-AM
6	36	(9,30)-WT, (6,24)-AM	1638-WT, 1332-AM
7	49	(12,42)-WT, (7,35)-AM	2274-WT, 1890-AM
8	64	(14,56)-WT, (8,48)-AM	2988-WT, 2544-AM

[a]HA: half-adder, FA: full-adder
[c]WT: Wallace-tree based MAC, [d]AM: Array-Multiplier based MAC

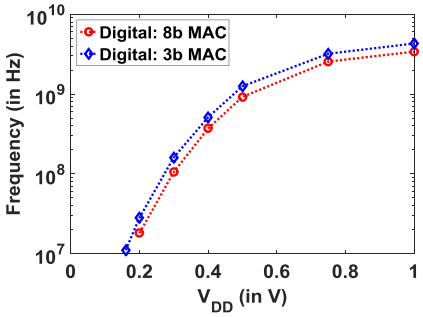

Fig. 5. Bandwidth of the Dig-N (WT) with different supply voltages.

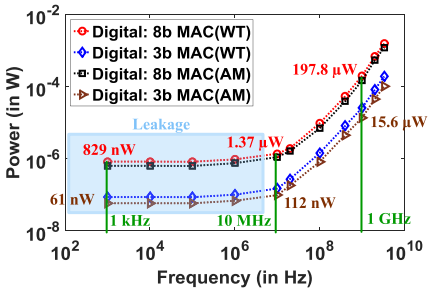

Fig. 6. Power consumption of the Dig-Ns across frequencies with DVFS.

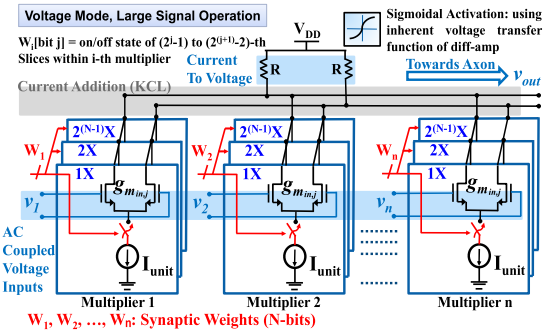

Fig. 7. MAC-based MS-N operating in large-signal mode, with total number of slices in kth multiplier $= 1 + 2 + \cdots + 2^{(N-1)} = 2^N - 1$.

computes the circuit dynamics algorithmically, leading to a transistor count as high as 2500 for the 8-bit case.

1) Bandwidth: Bandwidth of the Dig-N is dependent on the supply voltage (V_{DD}), as shown in Fig. 5. The BW is slightly higher for the 3-bit Dig-N as the critical-path delays are less for a low-complexity design. The scaling of bandwidth with supply voltage allows dynamic voltage and frequency scaling (DVFS) over different frequencies.

2) Power Versus Performance: The power in digital circuits usually consists of two components: 1) dynamic power and 2) static leakage power. Dynamic power in the design is given in the following equation:

$$P_{\mathrm{Dig}} = \sum_{i=1}^{N_{\mathrm{Dig}}} \alpha_i C_i V_{\mathrm{DD,Dig}}^2 f \tag{1}$$

where α_i is the activity factor of the ith node, N_{Dig} is the total number of nodes, C_i is the capacitance at a switching node, $V_{\mathrm{DD,Dig}}$ is the supply voltage, and f is the operating frequency.

The static leakage current is due to: 1) subthreshold conduction; 2) reverse-biased p-n-junction conduction; and 3) gate-induced drain leakage, out of which subthreshold conduction is the dominant factor [26]. Fig. 6 shows total power versus frequency for the 8-bit and 3-bit Dig-Ns designed in 65-nm CMOS process with DVFS. The dynamic power dominates for frequencies >10 MHz. However, at lower frequencies, power consumption is dominated by leakage, which increases proportionally to the number of transistors (8-bit WT has ~13 times more leakage than that of a 3-bit AM Dig-N, which corresponds to the ratio of transistors present in corresponding designs). The minimum energy consumption of the 8-bit AM Dig-N can be calculated as 87 fJ at 10 MHz (137 fJ for WT, 8-bit).

3) Noise: The dominant source of noise in digital circuits is quantization noise. Thermal noise-induced bit-flipping is practically rare due to high noise margin. Assuming a uniform distribution of error, the quantization noise voltage (in volts) of the Dig-N can be expressed as shown in the following equation:

$$N_Q = \frac{(V_{\mathrm{high}} - V_{\mathrm{low}})}{\sqrt{12(2^N - 1)}} \tag{2}$$

where N is the number of bits (precision), and $(V_{\mathrm{high}} - V_{\mathrm{low}}) \approx V_{\mathrm{DD}}$. As N is reduced, N_Q increases exponentially. For N-bit precision, the SNR in presence of N_Q can be calculated as SNR $= 6.02N + 1.76$ (in decibel). This results in ~50-dB SNR for 8-bit precision, and ~20-dB SNR for 3-bit precision.

B. Mixed-Signal Neuron (Large-Signal Mode)

The Dig-N is not energy efficient at frequencies <10 MHz where it suffers from static leakage power. MS-N with analog computation can potentially have far better energy efficiency, as they can be designed with only a few transistors unlike Dig-Ns. Fig. 7 shows an N-bit, differential-amplifier-based subthreshold MS-N architecture with n synaptic weights. The N-bit weights are coming from a digital memory while the MAC operation is performed in an analog fashion, hence the name MS-N. Bit j ($j = 0, 1, 2, \ldots, N-1$) of the ith weight activates switches at the tail current sources for each of the 2^j slices (from slice number $(2^j - 1)$ to slice number $(2^{(j+1-2)})$ in the ith multiplier (for all $i = 1, 2, 3, \ldots, n$). The tail current source for each slice has a value of I_{unit} when on. The overall circuit performs a vector MAC operation as follows: the alternating current through the jth slice in a single multiplier (let us take the multiplier with the input v_1)

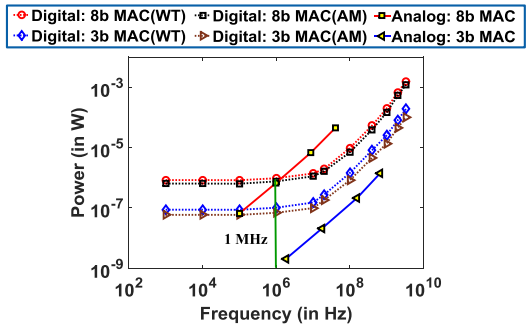

Fig. 8. Power consumption of the large-signal MS-MAC versus frequency.

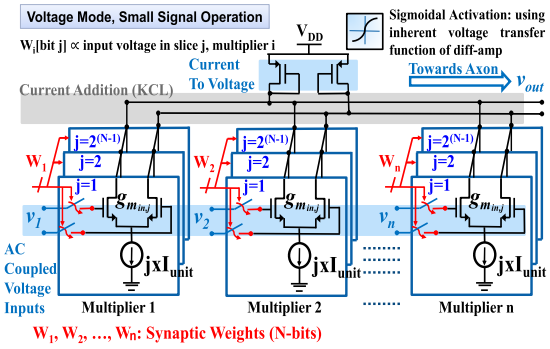

Fig. 9. MAC-based MS-N operating in small-signal mode, with a total number of slices in the kth multiplier $= N$.

is equal to $g_{m_{in,j}} v_1$ ($g_{m_{in,j}}$: transconductance of the input transistors of the jth slice). With the weight w_1 controlling the number of slices that will be on in that multiplier, the total current through the multiplier is $g_{m_{\text{in},j}} v_1 w_1$. Combining all the multipliers, the total current through the neuron becomes $g_{m_{\text{in}}} \sum_{k=1}^{n} w_k v_k$ (assuming every $g_{m_{\text{in},j}} = g_{m_{\text{in}}}$ in a regular, repeated structure). The voltage output with load resistors R would be $V_{\text{out}} = F(g_{m_{\text{in}}} R \sum_{k=1}^{n} w_k v_k)$, where F denotes the voltage transfer function (sigmoidal activation) for a differential amplifier. Designing the scaling factor $g_{m_{\text{in}}} R$ to be equal to 1, $V_{\text{out}} = \sum_{k=1}^{n} w_k v_k$ in the nonsaturated region, which is a vector MAC operation between the input voltages v_k and weights w_k. Interestingly, the total bias current through the load also keeps on changing with the weight, which means the effective bandwidth of the system corresponds to the minimum weight while the maximum power consumption of the system corresponds to the maximum weight. Moreover, changing the large-signal current with the weight necessitates a resistive load to ensure linearity of multiplication (a PMOS load will be nonlinear). This leads to significant area penalty for an on-chip implementation.

Fig. 8 illustrates the power consumption in the MS-N with respect to frequency and compares it to the power consumption of the Dig-Ns. The 8-bit analog MAC has a constant energy consumption of ≈0.9 pJ across all frequencies and has better power efficiency than digital MACs at frequencies <1 MHz. The 3-bit analog MAC has better power efficiency than the 3-bit digital MACs at all frequencies.

C. Mixed-Signal Neuron (Small-Signal Mode)

As illustrated in Fig. 9, to achieve a better power-bandwidth scalability, the weights can be used to activate switches at the gate of the input subthreshold transistors while a fixed current $I_{\text{bias}} = \sum_{j=1}^{2^x} (N-1) j I_{\text{unit}} = (2^N - 1) I_{\text{unit}}$ flows through the ith multiplier (for all $i = 1, 2, 3, \ldots, n$), enabling a small-signal mode of operation. When a switch is off, the corresponding input is connected to ground to avoid floating nodes. A PMOS load can now be employed as the bias current through the neuron is fixed. Since the gain of the neuron needs to be ≤1 (depending on the weight) to avoid saturating subsequent stages, high-impedance PMOS loads now allow a smaller effective transconductance which leads to better energy efficiency. Also, the number of slices in the ith multiplier is now reduced to N (from $(2^N - 1)$ in the large-signal case) while the current source of the jth slice now carries a current of $j \times I_{\text{unit}}$. This reduces the effective capacitance at the output of the neuron, thus increasing its bandwidth. Output of the MS-N is modeled in the following equation, following a similar derivation as shown in Section II-B:

$$V_{\text{out}} = F\left(A \times \sum_{k=1}^{n} w_k v_k\right) \tag{3}$$

where F is the voltage transfer function of a differential amplifier, which acts as a sigmoidal activation function in the context of a neuron, and w_k and v_k are the weights and the ac-coupled input voltage of the kth multiplier, respectively. A is the small-signal voltage gain of the unit slice of the kth multiplier where $A \approx g_{m_{\text{in1}}}/g_{m_p}$, $g_{m_{\text{in1}}}$ is the input transconductance of the unit slice of the multiplier, g_{m_p} is the transconductance of the PMOS load, and $(g_{m_{\text{in1}}} \sum_{k=1}^{n} w_k v_k)$ represent the Kirchoff's current law (KCL) addition of currents at the output nodes. The detailed expression for total gain A_v for the kth multiplier is obtained through a small-signal analysis [27] and is shown in the following equation:

$$A_v = -\frac{C_{\text{coupling}}}{C_{\text{coupling}} + N \times C_{gg}} \times \frac{\sum_{j=1}^{N} g_{m_{\text{in},j}}}{g_{m_p} + g_{ds_p} + \sum_{j=1}^{N} g_{ds_{\text{in},j}}} \tag{4}$$

where C_{coupling} is the ac coupling capacitance for each multiplier (100 fF—not shown in Fig. 9), C_{gg} is the effective gate-to-ground capacitance at the input of a multiplier, and $g_{m_{\text{in},j}}$ and $g_{ds_{\text{in},j}}$ are the transconductance and output conductance, respectively, for the input NMOS transistors in the jth slice of a multiplier. Similarly, g_{m_p} and g_{ds_p} are the transconductance and output conductance, respectively, for the PMOS load. The g_m quantities are in the same range of g_{ds}, and hence, g_{ds} cannot be ignored. Any source resistance/resistance at gate is ignored as it will be very small and will create a nondominant pole at a very high frequency.

1) Bandwidth: The linearity of bandwidth with respect to power is substantiated from (5) which shows the bandwidth at the output node of an N-bit synapse in terms of the subthreshold current [27]

$$\text{BW} = \frac{g_{m_p}}{2\pi \times C_{\text{eff}}} = \frac{(2^N - 1) I_{\text{unit}}}{4\pi \times \eta V_{\text{T}} \times C_{\text{eff}}} \tag{5}$$

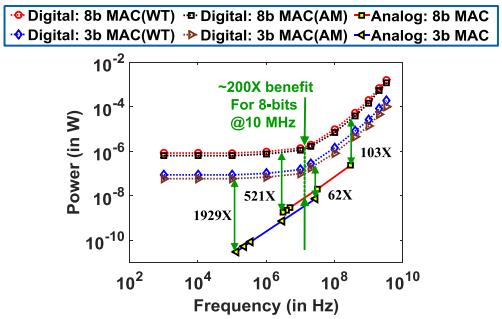

Fig. 10. Power consumption of the small-signal MS-MAC versus frequency.

where $g_{m_p} = (I_p/\eta V_T)$ is the transconductance of the PMOS load, I_p is the large-signal current through the PMOS load which is equal to half of the total current through a multiplier, C_{eff} is the effective output capacitance, and I_{unit} is the unit current of a multiplier. All input transistors are designed to carry subthreshold current.

2) Power Versus Performance: The total current is close to 255 times of the unit current for 8-bit MS-N and 7 times of the unit current for 3-bit MS-N. Fig. 10 demonstrates the power consumption in the MS-N with respect to frequency. The linearity of frequency versus power consumption is established from this figure as well. The energy consumption is constant (0.8 fJ for the 8-bit MS-N at all frequencies) and is $>60\times$ better than the AM-based Dig-N.

3) Noise: Unlike Digital-MAC, MS-MAC does not have a noise margin; hence, the accuracy of the network will be affected by the noise in the circuit. The main components of noise in the analog MAC are as follows.

a) MOSFET thermal noise: This is the dominant analog noise source. Considering the input and load transistors are in subthreshold saturation for each multiplier, we calculate the open circuit mean-square noise power per unit bandwidth. The total thermal noise current power per unit bandwidth for the subthreshold input transistors connected to each polarity of the differential input in a multiplier is given in the following equation [28]:

$$\overline{i_{n,\text{in}}^2} = 2qI_n \tag{6}$$

where q is an electronic charge and I_n is the total bias current through the relevant input transistors. Interestingly, I_n is half of the total current through each multiplier, and hence, (6) can be written as the following equation for an N-bit synapse:

$$\overline{i_{n,\text{in}}^2} = 2q \times \frac{1}{2}\sum_{i=1}^{N} 2^{(i-1)} \times I_{\text{unit}} = q \times (2^N - 1)I_{\text{unit}}. \tag{7}$$

The channel noise for the PMOS load in subthreshold is given in the following equation:

$$\overline{i_{n,p}^2} = 2qI_p = 2q \times \frac{1}{2}(2^N - 1)I_{\text{unit}} = q \times (2^N - 1)I_{\text{unit}}. \tag{8}$$

Hence, the total thermal noise power at the output (in V^2/Hz) can be obtained as given in the following equation:

$$\overline{v_n^2} = [q \times (2^N - 1)I_{\text{unit}} + q \times (2^N - 1)I_{\text{unit}}] \times \left(\frac{1}{g_{m_p}}\right)^2$$
$$= [4qI_p] \times \left(\frac{\eta V_T}{I_p}\right)^2 = \frac{4q\eta^2 V_T^2}{I_p} = \frac{8q\eta^2 V_T^2}{(2^N - 1)I_{\text{unit}}}. \tag{9}$$

Therefore, $\overline{v_n^2}$ is inversely proportional with bias current. Equations (5) g and (9) also imply that the integrated thermal noise power over the bandwidth will be constant, as given in the following equation:

$$\overline{v_{n,\text{integrated}}^2} = \frac{4q\eta \text{V}_\text{T}}{2\pi \times C_{\text{eff}}}. \tag{10}$$

This means that to reduce the integrated thermal noise, we need to use larger W and L for the transistors which increases C_{eff}. However, this also reduces bandwidth. We will show a method to overcome this tradeoff in Section III.

b) Flicker noise: This is a ubiquitous noise present in all electronic systems, which is most significant at low frequencies. This noise power is empirically given in the following equation:

$$\overline{v_{n,f}^2} = \frac{K}{f^\alpha} \tag{11}$$

where K is an empirical parameter that is dependent on device type, dimensions, and technology node, and α is an exponent that is usually close to unity [27].

c) Switch noise: Since we have introduced switches in the signal path in our design, any switching activity will give rise to transient noise. However, it must be noted that the weights will be set during training, and hence, there would be no noise from the switches during the testing phase.

d) Quantization noise: A binary code is used to activate the binary weights that connect the inputs to the desired differential pairs. Thus, effectively, a digital-to-analog conversion (DAC) operation is being carried out, which gives rise to quantization noise, given by (2). This is found to be the overall dominant noise because of the low precisions in our application.

In the system-level applications, however, the integrated thermal and flicker noise is of more importance, as quantization noise affects bit precision analysis for the weights (Fig. 4) and creates the same baseline error for both Dig-N and MS-N. Fig. 11 exhibits the noise power (in V^2) integrated over the signal bandwidth, as a function of the total power consumed. As expected, this is relatively constant since bandwidth and noise floor (in V^2/Hz) both scales linearly in equal and opposite amounts with bias current.

4) Effect of Mismatch/DC Offset: There is negligible systematic offset because of the symmetry of the design. However, there will be a random offset because of mismatches in threshold voltages and dimensions during fabrication. These mismatches can be within-die (local) or die-to-die (global) and create an offset at the output nodes of each multiplier. Since the individual multipliers are ac coupled (with 100-fF coupling capacitance), this offset will not propagate to

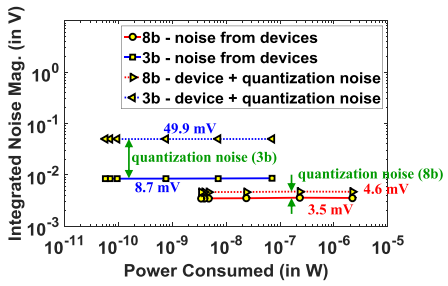

Fig. 11. Integrated device and quantization noise versus power for the MS-MAC.

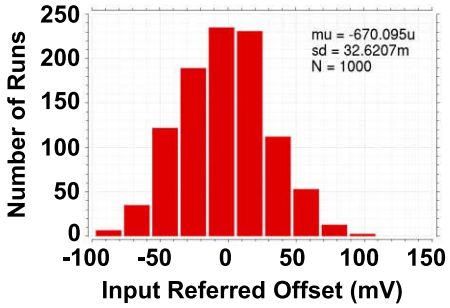

Fig. 12. Mismatch analysis. Input referred offset for the 8-bit MS-MAC.

subsequent stages. However, the input bias points of the two legs of a branch can be different due to the offset, causing variations in gain and swing in the two branches.

The primary contributor to overall mismatch is often the mismatch in threshold voltages ($V_{\rm TH}$) of the otherwise symmetric transistors. The standard deviation for $V_{\rm TH}$ is given in the following equation [31]:

$$\sigma_{V_{\rm TH}} = \frac{A_{V_{\rm TH}}}{\sqrt{WL}} + S_{V_{\rm TH}} \times D_x \tag{12}$$

where $A_{V_{\rm TH}}$ is the $V_{\rm TH}$-mismatch parameter, W and L are the device dimensions, D_x is the distance between the centers of the devices, and $S_{V_{\rm TH}}$ is the distance proportionality constant ($\approx$0 for common centroid layout). Mismatches in sizing, mobility, gate–oxide capacitance, and body bias parameter also contribute to the overall offset. With the standard values of the parameters for a 65-nm process and the values of W and L, $3\sigma_{V_{\rm TH}}$ is calculated to be around 60 mV for each differential leg in our design. Fig. 12 shows the results of the Monte Carlo analysis that indicates a 3σ output offset variation of 98 mV, which is alarmingly high considering the voltage swing to be a few hundred millivolts.

III. Proposed Mixed Signal Neuron: Reduced Mismatch/Offset and Increased Bandwidth

To solve the issue of offset, we propose an MS-N with resistive feedback, as shown in Figs. 13 and 14. The feedback resistance tries to keep the input and output dc points at the same voltage. At the same time, the resistance forms a low-pass filter with the parasitic C_{gd} in the feedback path. This creates a zero in the feed-forward transfer function which, when superimposed on the dominant pole, increases the bandwidth of the circuit. The increase in bandwidth enables us to increase W and L of the MOSFETs (keeping the ratio same) that reduces the input offset and the effect of mismatch/noise.

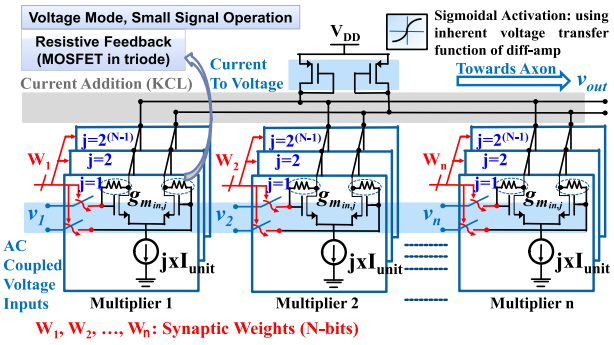

Fig. 13. Proposed small-signal MS-N with resistive feedback (resistance is implemented with an NMOS in triode region). Total number of slices in the kth multiplier $= N$.

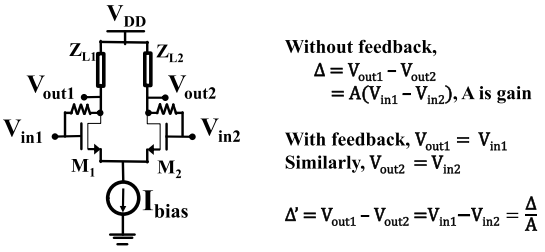

Fig. 14. Offset compensation using resistive feedback. All voltages are dc.

These benefits are obtained at no extra power cost and a minimal area cost.

Since on-chip poly resistors consume a significant area, the resistance in the feedback path is implemented using an NMOS in triode. The nonlinearity introduced due to this does not affect the final results since the input common mode range of the neuron is observed to be >150 mV across process corners.

The gain of this multiplier structure can be obtained through a small-signal analysis [27] and is given in the following equation:

$$A_v = -\frac{C_{\rm coupling}}{C_{\rm coupling} + N \times C_{gg}} \times \frac{\sum_{j=1}^{N} g_{m_{{\rm in},j}} - \frac{N}{R}}{g_{m_p} + g_{{\rm ds}_p} + \sum_{j=1}^{N} g_{{\rm ds}_{{\rm in},j}} + \frac{N}{R}} \tag{13}$$

where R is the feedback resistance (an NMOS in triode). Again, g_m quantities are in the same range of $g_{\rm ds}$, and hence, $g_{\rm ds}$ terms cannot be ignored. The dominant pole is given by

$$\omega_{p,{\rm dom}} = \frac{g_{m_p} + \frac{N}{R}}{\sum_{j=1}^{N} C_{gd_j}} \tag{14}$$

while the following equation models the zero:

$$\omega_z = \frac{\sum_{j=1}^{N} g_{m_{{\rm in},j}} - \frac{N}{R}}{\sum_{j=1}^{N} C_{gd_j}}. \tag{15}$$

Solving for $\omega_{p,{\rm dom}} = \omega_z$, we can find R, as given by the following equation:

$$R = \frac{2N}{\left|\sum_{j=1}^{N} g_{m_{{\rm in},j}} - g_{m_p}\right|}. \tag{16}$$

TABLE II

PROCESS AND TEMPERATURE VARIATION IN PROPOSED 8-bit MS-N[1] FOR SUPPLY VOLTAGE = 0.75 V

Specification	TT Process @(-25°C, 27°C, 100°C)	FF Process @(-25°C, 27°C, 100°C)	SS Process @(-25°C, 27°C, 100°C)	FNSP Process @(-25°C, 27°C, 100°C)	SNFP Process @(-25°C, 27°C, 100°C)
Power (nW)	(16.6, 20.8, 35.7)	(19.7, 24.49, 42.8)	(14.9, 19.32, 31.9)	(19.6, 23.44, 40.3)	(17.9, 20.27, 35.8)
BW (MHz)	(330.6, 292.2, 201.1)	(417.1, 374, 275.8)	(274.7, 247.4, 173.2)	(443.2, 410.3, 306.4)	(246.3, 231.7, 152.2)
Gain (dB)	(-0.441, -0.457, -0.448)	(-0.355, -0.361, -0.374)	(-0.682, -0.696, -0.691)	(-0.308, -0.315, -0.328)	(-1.31, -1.39, -1.73)
Int. Noise Power (V^2)	(5.9e-8, 6.3e-8, 7.5e-8)	(5.1e-8, 5.4e-8, 6.5e-8)	(6.3e-8, 6.8e-8, 8.1e-8)	(5.9e-8, 6.4e-8, 7.3e-8)	(6.1e-8, 6.6e-8,7.7e-8)
Energy Eff. (Power/BW)	(50, 71, 178) aJ	(47, 65, 155) aJ	(54, 78, 184) aJ	(44, 57, 132) aJ	(73, 87, 235) aJ

[1]: Results are with I_{unit} = 100 pA (mirroring current) and Supply Voltage = 0.75V

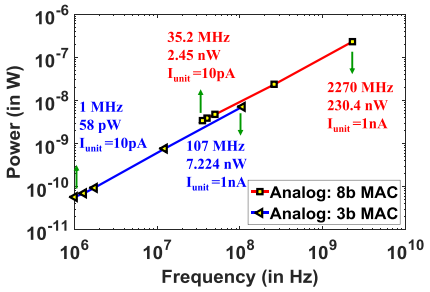

Fig. 15. Power consumption of the proposed MS-MAC versus frequencies.

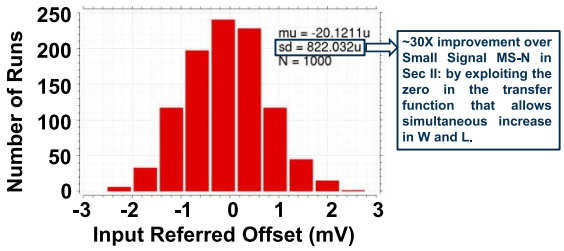

Fig. 16. Mismatch analysis. Input referred offset for proposed MS-MAC.

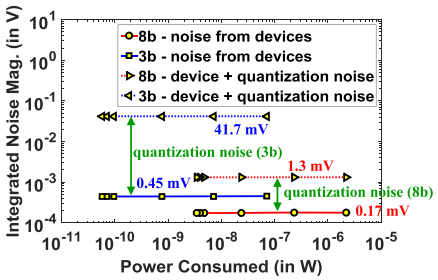

Fig. 17. Integrated noise versus power for the proposed MS-MAC.

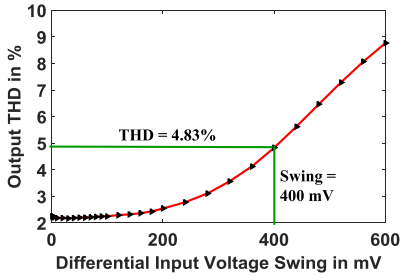

Fig. 18. THD for the 8-bit proposed MS-MAC, with I_{unit} = 100 pA.

This resistance (~100 kΩ) is implemented using an NMOS in the triode region, and the absolute value has a small effect (~5% reduction) on the bandwidth of the circuit when the resistance varies by ±10% around its nominal value. Fig. 13 shows how offset compensation can be achieved using the resistive feedback structure. The output offset with the feedback is reduced by a factor of A as compared to the offset without the feedback. In our design, $A = 1$ that results in a residual offset same as the original offset, but the introduction of the zero due to the resistive feedback enables bandwidth extension which allows larger devices (hence, smaller offset). Fig. 15 exhibits the frequency (bandwidth) versus power characteristics of the proposed MS-N. The energy efficiency is ~100 aJ for the 8-bit design, which is much lower than the Dig-N (87 fJ, the best case for AM-based MAC) and the other designs of the MS-Ns presented in Section II (0.8 fJ, the best case). Fig. 16 exhibits a 3σ output offset variation of 2.5 mV from Monte Carlo simulations of the 8-bit design, while 0 shows that the integrated output noise power <0.1 μV^2 over the bandwidth (integrated noise voltage is ~0.17 mV). Thus, the overall worst case effect of noise and mismatch can be considered to be within 3 mV, from Figs. 16 and 17. Fig. 18 presents the output total harmonic distortion (THD) as a function of the voltage swing when $I_{\text{unit}} = 100$ pA. With a differential input swing of 400 mV, THD is <5%. Hence, the 3-mV error due to mismatch and noise is within 1% of the output swing.

A. Effect of Variability

Apart from noise and mismatch, the MS-N also suffers from process, voltage, and temperature (PVT) variations. Table II lists the important specifications of the MS-N across different process corners and temperatures. For room temperatures and below, the energy efficiency is close to 100 aJ/MAC (also valid for supply voltages in the range 0.7–1 V), while the worst case efficiency at 100° is 235 aJ/MAC at 0.75-V supply (384 aJ at 1 V supply), which is still subfemtojoule/MAC. Since the gain is always close to 1 and the integrated noise is almost constant across process corners, the only limitation posed by the process corners is the bandwidth when NMOS transistors are slow. However, even the worst case bandwidth results in an energy efficiency that is much better than existing architectures, as will be seen in Section VII.

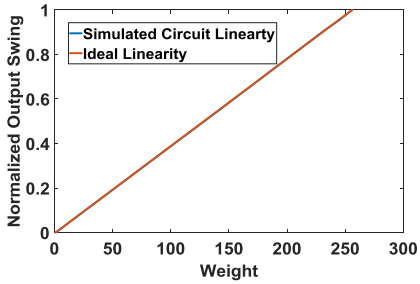

Fig. 19. Linearity for the 8-bit MS-N with 400-mV differential swing and 100-pA unit current (in presence of $\pm 3\sigma$ mismatch in device dimensions).

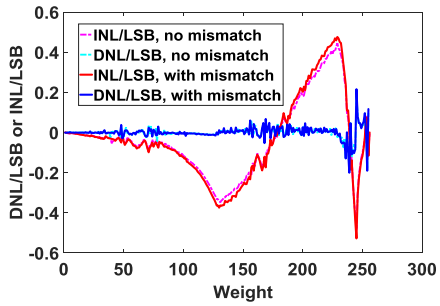

Fig. 20. Simulated INL and DNL of the MS-N (conditions same as Fig. 19).

B. Stability

The bandwidth is extended by making $\omega_{p,\text{dom}} \approx \omega_z$. The feedback resistor can lead to oscillations if the phase margin is not enough. However, the system is always stable because the gain of the system is ≤ 1, and hence, there is no gain crossover frequency and the concept of phase margin does not apply.

C. DNL and INL

The multiplication of the input signal with the weights is effectively a DAC operation. The slices in each multiplier are designed with binary weighted bias currents but with same sized input transistors, and hence, the overdrives are different for the input transistors in each slice, which leads to an effective transconductance that increases in a slightly nonlinear manner with the weight. Hence, this architecture achieves high bandwidth and low power at the cost of nonlinearities in the DAC operation. However, with large input swing (~400-mV differential) and a small unit current (~100 pA), the differential nonlinearity (DNL) and integral nonlinearity (INL) can be kept within ± 0.5 LSB, even in the presence of $\pm 3\sigma$ amount of mismatch as shown in Figs. 19 and 20. This implies that there is no missing code during the MS-N operation.

IV. Comparison: Dig-N Versus Proposed MS-N

The power and energy consumption of the proposed MS-N at different frequencies is shown in Figs. 21 and 22, respectively. The proposed MS-N is two to three orders more energy-efficient as compared to the Dig-N over all frequencies. At frequencies <1 MHz, the all-digital implementation suffers from static leakage currents, which is quite high due to a large number of transistors. Energy consumption at frequencies >500 MHz is also high for the Dig-N because of DVFS.

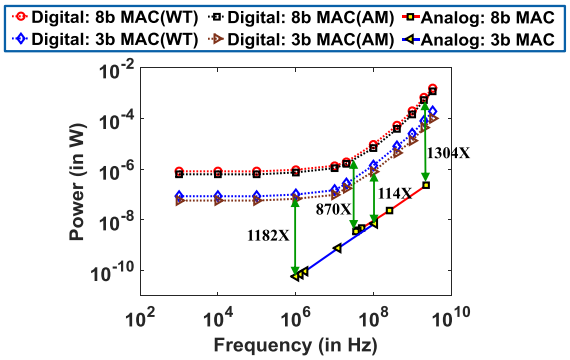

Fig. 21. Power consumption of the proposed MS-MAC against Dig-MAC.

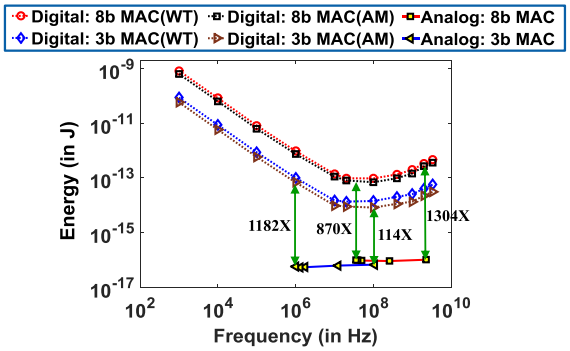

Fig. 22. Energy consumption of the proposed MS-MAC against Dig-MAC.

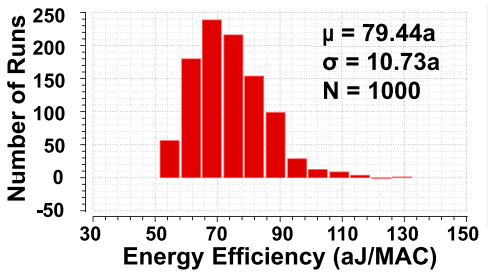

Fig. 23. Monte Carlo simulation showing the statistical distribution of energy efficiency (with 27 °C temperature and 0.75-V supply).

In contrast, power for the analog MAC in the MS-N scales linearly with frequency and is more energy-efficient both at low and high frequencies. For the 8-bit design, the MS-N is ~870× better than Dig-N in terms of energy efficiency at near-threshold point (~10 MHz), which is a further 4× improvement over the performance of the MS-N presented in Section II-C and in [33]. Fig. 23 shows the statistical distribution of energy efficiency with ± 3-σ models for process and mismatch provided by the foundry at 27 °C temperature and 0.75-V supply, corroborating <100 aJ/MAC energy efficiency for the above-mentioned conditions.

V. Theoretical Limits and Tradeoffs for MS-N

To understand the performance benefits of the MS-N over Dig-N, we consider the fundamental design of the 8-bit MS-N presented in Section II-C. The power consumption for an MS-multiplier (P_{MS}) in that case is calculated as given in the

following equation:

$$P_{\mathrm{MS}} = V_{\mathrm{DD,Ana}} I = V_{\mathrm{DD,Ana}} \times (2\pi \times \eta V_T \times C_{\mathrm{eff}} \times \mathrm{BW}) \tag{17}$$

where $V_{\mathrm{DD,Ana}}$ is the supply voltage for the analog MAC, $I = (2^N - 1) I_{\mathrm{unit}} = (2\pi \times \eta V_T \times C_{\mathrm{eff}} \times \mathrm{BW})$ is the total current in the multiplier, as found from (5). On the other hand, the dynamic power consumption of a Dig-N (P_{Dig}) was shown in (1), which leads us to the following equation, which shows the ratio ($P_{\mathrm{Dig}}/P_{\mathrm{MS}}$)

$$\frac{P_{\mathrm{Dig}}}{P_{\mathrm{MS}}} = \frac{\sum_{i=1}^{N_{\mathrm{Dig}}} \alpha_i C_i V_{\mathrm{DD,Dig}}^2 f}{V_{\mathrm{DD,Ana}} \times (2\pi \times \eta V_T \times C_{\mathrm{eff}}) \times \mathrm{BW}.} \tag{18}$$

For isofrequencies ($f = \mathrm{BW}$) and considering $V_{\mathrm{DD,Ana}} = 1$ V, $V_{\mathrm{DD,Dig}} = 0.4$ V (which are the values of analog and digital supplies in the simulations), $\eta = 1.3$, and $V_T = 0.026$ V, we get the following equation:

$$\frac{P_{\mathrm{Dig}}}{P_{\mathrm{MS}}} = \frac{V_{\mathrm{DD,Dig}}^2}{V_{\mathrm{DD,Ana}} \times (2\pi \times \eta V_T)} \times \frac{\sum_{i=1}^{N_{\mathrm{Dig}}} \alpha_i C_i}{C_{\mathrm{eff}}} = 0.75 \times \frac{\sum_{i=1}^{N_{\mathrm{Dig}}} \alpha_i C_i}{C_{\mathrm{eff}}}. \tag{19}$$

Depending on whether the effective output capacitance of the MS-multiplier (C_{eff}) is dominated by its own intrinsic capacitance (from the slices), or by the number of multipliers connected at the output node, or by the number of fanouts from the neuron, the following scenarios may arise.

A. Case 1: Intrinsic Slice Capacitance Dominates

From the simulation results shown in Figs. 6 and 10, it is clear that the worst case energy-benefit ($P_{\mathrm{Dig}}/P_{\mathrm{MS}}$) for the MS-N would be observed near 10-MHz frequency where the energy consumption of Dig-N is at its lowest. At that frequency, $\sum_{i=1}^{N_{\mathrm{Dig}}} \alpha_i C_i$ for the Dig-N can be approximated as $(((1370 - 829) \times 10^{-9})/(0.4 \times 0.4 \times 10^7)) \approx 338$ fF. Hence

$$\frac{P_{\mathrm{Dig}}}{P_{\mathrm{MS}}} = \frac{0.75 \times 338}{C_{\mathrm{eff}}} = \frac{253.5}{C_{\mathrm{eff}}}. \tag{20}$$

If all the slices in the MS-N were designed with unit current, then the unit slice would have required to be repeated 2^j-times for the jth bit of the weight for a multiplier (as shown in Section II-B, $j = N - 1$ is the MSB, and $j = 0$ is the LSB). This would have meant $C_{\mathrm{eff}} = (2^N - 1) C_{\mathrm{unit}} = 255 \times C_{\mathrm{unit}}$, where C_{unit} is the unit capacitance from each slice, connected to the output node. As a result, ($P_{\mathrm{Dig}}/P_{\mathrm{MS}}$) would have been ($253.5/255 \times C_{\mathrm{unit}}$), which is close to a mere factor of 2.5 assuming $C_{\mathrm{unit}} = 0.4$ fF which is the unit node capacitance in both our Dig-N and MS-N implementations. Hence, to get the energy benefits of MS-N, irregular slices had to be adopted as shown in Fig. 9, where the tail current sources are binary-weighted, but each bit of the weight is connected to only one slice. This implies that $C_{\mathrm{eff}} = N C_{\mathrm{unit}} = 8 \times C_{\mathrm{unit}}$ and hence ($P_{\mathrm{Dig}}/P_{\mathrm{MS}}$) = ($253.5/8 \times 0.4$) ≈ 79. Of course, this energy benefit comes with the issue of nonlinearity as discussed in Section III.

The analysis shown earlier only considers the dynamic power of the Dig-N. The total energy benefit considering dynamic + leakage current for each node in the Dig-N in this scenario is given in the following equation (again taking the numbers from Fig. 6):

$$\frac{P_{\mathrm{Dig}}}{P_{\mathrm{MS}}} = \frac{1370 \times 10^{-9}}{V_{\mathrm{DD,Ana}} \times (2\pi \times \eta V_T) \times C_{\mathrm{eff}} \times 10^7} = \frac{630.6}{C_{\mathrm{eff}}} \tag{21}$$

which results in ($P_{\mathrm{Dig}}/P_{\mathrm{MS}}$) ≈ 197 when $C_{\mathrm{eff}} = 8 \times C_{\mathrm{unit}}$ and $C_{\mathrm{unit}} = 0.4$ fF. These results correspond to the graph shown in Fig. 10 where the energy benefit for the MS-multiplier is about 200× that of the digital implementation near 10 MHz. Theoretically, we can arrive at the same result from (19) by noting that ($\sum_{i=1}^{N_{\mathrm{Dig}}} \alpha_i C_i / C_{\mathrm{eff}}$) is essentially the ratio of number of transistors in Dig-N (N_{Dig}) to the number of input transistors in MS-N (N_{Ana}), multiplied with the effective α of the Dig-N. Hence, ($P_{\mathrm{Dig}}/P_{\mathrm{MS}}$) can be expressed as follows:

$$\frac{P_{\mathrm{Dig}}}{P_{\mathrm{MS}}} = 0.75 \times \alpha \times \frac{N_{\mathrm{Dig}}}{N_{\mathrm{Ana}}} \times L \tag{22}$$

where L is the additional contribution from leakage ($L \approx 2.5$ near the threshold frequency of 10 MHz where dynamic power is almost 67% of the leakage power). Writing $N_{\mathrm{Ana}} = (N_{\mathrm{Ana,regular}}/((2^N - 1)/N))$ (N is the number of bits), we get the following equation:

$$\frac{P_{\mathrm{Dig}}}{P_{\mathrm{MS}}} = 0.75 \times \alpha \times \frac{N_{\mathrm{Dig}}}{N_{\mathrm{Ana,regular}}} \times \frac{2^N - 1}{N} \times L$$

(Activity Factor (Dig-N); Ratio of Transistors; Benefit from Irregular slices; Leakage (Dig-N))

$$= 0.75 \times 0.3 \times \frac{2805}{255} \times \frac{2^8 - 1}{8} \times 2.5 = 197 \tag{23}$$

where $N_{\mathrm{Ana,regular}}$ is the number of input transistors in MS-N with regular slices, and $((2^N - 1)/N)$ denotes the benefit gained from irregular slices (as the number of transistors reduce from $(2^N - 1)$ to N). In the calculation, we have assumed $N_{\mathrm{Dig}} \approx 2805$ (in Table I) and $\alpha \approx 0.3$ ($\sum_{i=1}^{N_{\mathrm{Dig}}} \alpha_i C_i = 338$ fF, hence, α can be calculated and verified from Synopsys reports).

B. Case 2: Load Capacitance Is Also Significant

The load capacitance consists the fanouts from the neuron (same for both the Dig-N and MS-N) and the number of multipliers (n) connected to the output node (considered only for MS-N). Since the fanout can be taken care of by inserting properly sized buffers for both Dig-N and MS-N, we only consider the effect of n in this analysis and write the ratio ($P_{\mathrm{Dig}}/P_{\mathrm{MS}}$) as shown in the following equation:

$$\frac{P_{\mathrm{Dig}}}{P_{\mathrm{MS}}} = \frac{630.6}{C_{\mathrm{eff}}} = \frac{630.6}{n \times N C_{\mathrm{unit}}} = 0.75 \times \alpha \times \frac{N_{\mathrm{Dig}}}{n \times N_{\mathrm{Ana,regular}}} \times \frac{N - 1}{N} \times L. \tag{24}$$

Thus, the worst case energy benefit reduces by a factor of 10 when $n = 10$. This means that the MS-N is largely

TABLE III
LEARNING ANN/CNN ARCHITECTURES AND PERFORMANCE EVALUATION (NEURON-ONLY POWER CONSUMPTION WITHOUT PIPELINING/RESOURCE SHARING)

Application	Learning Architecture	Power (Dig-N: AM) @100MHz	**Energy (Dig-N: AM) @100MHz**	Power (MS-N) @100MHz	**Energy (MS-N) @100MHz**
MNIST_FCN: 84060 MACs	784×100×50×10 (2 Hidden Layers)	582mW (8b), 69mW(3b)	**5.82nJ (8b), 690pJ (3b)**	841μW (8b), 315μW (3b)	**8.41pJ (8b), 3.15pJ (3b)**
MNIST_CNN: 2.3M MACs	LeNet [12]	15.92W (8b), 1.89W(3b)	**159.2nJ (8b), 18.9nJ (3b)**	23mW (8b), 8.6mW(3b)	**230pJ (8b), 86pJ (3b)**
CIFAR_CNN: 46.2M MACs	AlexNet [29] for 32x32 images	319.9W (8b), 37.9W(3b)	**3.2μJ (8b), 379nJ (3b)**	462mW (8b), 173mW (3b)	**4.62nJ (8b), 1.73nJ (3b)**

beneficial for a CNN where the number of connections to a neuron is limited. However, cascode topologies are shown to be useful when a number of input/outputs are large in an analog design [34] and such structures can be employed for implementing a fully connected network using the proposed MS-N.

In summary, this analysis shows that the fundamental energy benefit for the MS-N is a direct effect of: 1) lower number of transistors due to the ability to represent complex functions with intrinsic dynamics and 2) irregular slice structure, which trades off with the linearity of the neuron. The leakage (at low frequency) and DVFS (at high frequency) in Dig-N further improves this energy benefit at frequencies other than the near-threshold point (10 MHz) as the power consumption for MS-N is linear with frequency. The other significant tradeoffs are between total integrated noise [which reduces by increasing device dimensions according to (10)] versus bandwidth (which degrades as per (5) when device dimensions are increased). The dc offset due to device mismatch can also be reduced at the cost of reduced bandwidth. The proposed design in Section III alleviates these tradeoffs by extending the bandwidth of the circuit using pole-zero compensation, which allows for increased device dimensions to reduce the effects of noise and mismatch.

VI. EXPLOITING NEUROMORPHIC ERROR RESILIENCY AT THE SYSTEM LEVEL FOR LENET AND ALEXNET

To analyze the energy benefits and performance of the proposed MS-N, a cohesive circuit-algorithmic framework is developed that uses two well-known benchmark image recognition data sets, MNIST [24] and CIFAR-10 [25]. MNIST is a standard data set of handwritten digits that contains 60 000 training and 10 000 test patterns of 28×28 pixel-sized greyscale images of the digits 0–9. CIFAR10 is a more complex data set that consists of 60 000 colored images belonging to ten classes. Each image has 32×32 pixels. The first 50 000 images were used for training and the last 10 000 images were used for testing.

Table III shows the different deep learning and fully connected implementations used to evaluate the data sets. It is to be noted that the learning architectures employed are the standard networks that have shown reasonable accuracy on the various benchmarks for low-to-medium-complexity applications [12], [29], as sensor nodes targeted toward edge analytics do not require deeper networks like ResNet or GoogLeNet to be implemented on the edge device. Each of the architectures shown in Table III was implemented using the widely used MatConvNet [30] platform, a deep learning toolbox used for training and evaluating the performance of the benchmark applications. While training, 16-bit precision was used to get a reasonable accuracy for the baseline network. However, for most ANNs, the bit precision can be scaled down to 8 bits without incurring any accuracy degradation as shown in Fig. 4 (error versus bit precision figure). Scaling below 8-bits may cause accuracy loss, a part of which can be restored by retraining the network. Therefore, for lower bit precision (starting from 6 bits, down to 3 bits), incremental retraining was performed with bit width restrictions in place on the weights and neuron outputs to reclaim a significant portion of the accuracy ceded by scaling. Bit width scaling (and retraining) helps to get an optimized CMOS digital framework for our precision-constrained MS multipliers. It also helps in obtaining an optimized digital baseline framework for fair energy/performance comparison with our MS-N. Hence, the software baseline implementation was aggressively optimized for performance.

The trained baseline network with appropriate bit restrictions on the learned weights is then evaluated on the testing set of the benchmark to obtain the performance or accuracy. The analog noise and mismatch models obtained from circuit simulations are incorporated in the software during the evaluation phase. The total integrated noise power was calculated in the range 1 kHz–1 GHz (ac coupling capacitors filter out the low-frequency noise) using the bsim4 noise model in Cadence Virtuoso tool flow, and with a standard 65-nm technology model for transistors (obtained from foundry). The mismatch/input-referred-offset was simulated using Monte Carlo analysis (in Cadence Virtuoso) with $\pm 3-\sigma$ models for process and mismatch. Since both dc offset and analog noise come from the multiplier units that perform the multiplication of the weight values with the corresponding input, they are included within a modified weight value. The output at a particular neuron without noise and mismatch is given by (3), while the same output is calculated using the following equation in presence of noise and mismatch:

$$V_{\text{out}} = F\left[A_v \times \sum_{k=1}^{n} w_k(v_k + \sqrt{A} + \Delta_k)\right] = F\left[A_v \times \sum_{k=1}^{n} w_k\left(1 + \frac{\sqrt{A} + \Delta_k}{v_k}\right)v_k\right] \quad (25)$$

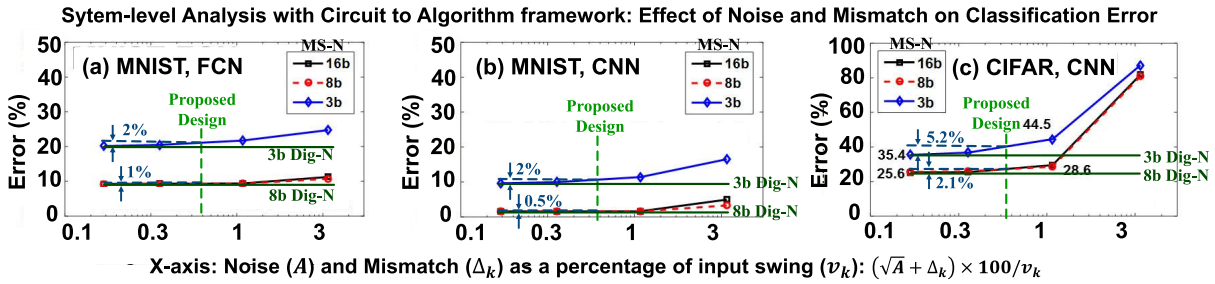

Fig. 24. System level simulation results for 3-, 8-, and 16-bit MS-N. (a) MNIST [24] with FCN. (b) MNIST with CNN. (c) CIFAR-10 [25] with CNN.

TABLE IV
COMPARISON RESULTS WITH STATE-OF-THE-ART NEURON ARCHITECTURES

	SpiNNekar[1] [5]	TrueNorth [6]	Neurogrid [10]	BrainScaleS [9]	VMM [13]	ISSCC 2018 MS-N [14]	Traditional MS-MAC (Sec II A.)[3]	Proposed Work (Section III)[4]
Meas./Sim.	Measured	Measured	Measured	Measured	Measured	Measured	Simulation	Analysis/Simulation
Neuron Type	Digital	Digital	Mixed-Signal	Mixed-Signal	Mixed-Signal	Mixed-Signal	Mixed-Signal	Mixed-Signal
SNN/CNN	SNN	SNN	SNN	SNN	CNN	CNN	CNN	CNN
Architecture	ARM-core	Point neuron	Current mode	Current mode	Current mode	BinaryNet+SC[2]	Current mode	Small signal
MAC based	No	No	No	No	Yes	Yes	Yes	Yes
CMOS Tech.	130 nm	28 nm	180 nm	180 nm	500 nm	28 nm	65 nm	65 nm
Power/synapse	-	254 nW	390 pW	19.5 µW	5.310 µW	4.62 pW	0.9 µW	10 nW
Frequency	200 MHz	1 kHz	1 kHz	125 MHz	10 MHz	237 FPS	1 MHz	**100 MHz**
Energy/synapse	-	254 fJ	390 fJ	156 fJ	531 fJ	19.5 fJ	900 fJ	**100 aJ**

[1]: Neurons are simulated on ARM-core (no physical implementation), [2]: Switched-Capacitor (SC) addition with XNOR multiplication,
[3,4]: The ANN architectures are not available on hardware

where A is the integrated noise power (V^2) within the bandwidth at a given bias current, while Δ_k is the dc offset in the kth synapse. If the swing of v_k is high, the effect of noise and mismatch is minimal. Fig. 24 presents the classification error for (a) MNIST_FCN, (b) MNIST_CNN, and (c) CIFAR_CNN as a function of the nonideality percentage (NIP), which we define as NIP $= (\sqrt{A} + \Delta_k/v_k) \times 100$. We observe that the worst case increase in error from the baseline (8-bit and 3-bit Dig-Ns) is only 5.2% in 3 bit and 2.1% in the 8-bit proposed MS-N (both for CIFAR-10, with AlexNet CNN) with a differential input swing of 400 mV and NIP of 0.75% $\approx$ 3 mV which is found from circuit-level simulations.

VII. PERFORMANCE COMPARISON AND DISCUSSION

Table IV shows the comparison of the performance of the proposed design with the existing neuron architectures. The proposed design achieves the best energy efficiency and can work at high frequencies, which makes it suitable for neuromorphic computing applications. Although the power consumption per synapse is lower for Neurogrid, we must note that Neurogrid runs at a much lower frequency. The bias currents in the proposed design can be reduced for low-frequency applications and have a better power per synapse value. Since the power in MS-N scales linearly with frequency, this will not degrade the energy efficiency.

It must be noted that the proposed work is based on simulation and analysis, while the other works presented in Table IV have measured data that consider the energy and latency of communication, memory fetch, data management, and streaming, which often proves to be a worse energy bottleneck than computation. However, in-memory [35] or near-memory [36] architectures help in reducing the memory-fetch power. As shown in [11], several power-reduction strategies such as event-driven computing, overlapping dendritic trees, island formation, hierarchical axonal structures, power-gating, multiplexed signaling, and coordinated processing can be employed to reduce the communication energy, further exploiting the improved energy efficiency of the proposed neuron at a system level. To account for the increased loading at the output nodes in a fully functional ANN, a cascode topology as presented in [34] can be adopted as well. The proposed neuron model can, thus, be utilized to improve the power versus frequency performance for the architectures demonstrated in the references shown in Table IV, with similar hardware for communication, memory fetch, data management, and streaming, and motivates the need for future research in this direction.

VIII. CONCLUSION

We have presented a MAC-based MS-N architecture that can achieve extreme energy efficiency by employing a small-signal voltage mode multiplication using a differential amplifier with resistive feedback. Compared to a traditional Dig-N, the proposed MS-N is $\sim$ 1000$\times$ more energy efficient at both low frequencies (<1 MHz) and very high frequencies (>500 MHz), and > 100$\times$ more energy-efficient across frequencies in the range of 1–500 MHz without significantly affecting the classification error rate for digit/image recognition applications. The proposed implementation promises to achieve better energy efficiency than prior analog/MS designs (20–40$\times$ better than digital implementations) as well as memristor-based designs (3–4$\times$ better than prior MS designs) [37]. Digital implementations can be duty cycled (using power gating) to reduce the power consumption.

However, duty cycling does not reduce the on-time energy consumption. Moreover, for applications with low-frequency input signal, duty-cycled digital implementations require input and output FIFOs and suffer from the tradeoff between FIFO size and frequency of turn-on/turn-off of the computation unit. The sources of the energy benefit in the MS-MAC for such scenarios are thoroughly analyzed and the significant tradeoffs are identified, which are: 1) energy versus linearity; 2) energy versus total integrated noise; and 3) energy versus variability. A bandwidth extension technique is proposed, which helps in alleviating the tradeoffs and leads to a 0.1 fJ/MAC implementation of the 8-bit MS-N, which provides enough headroom for mimicking a biological neuron (20 fJ/MAC [11]) at a system level. As a future extension of this paper, the memory fetch and communication energies of FCN and CNNs will be analyzed and implemented with near-memory computation, which promises to achieve an energy-efficient MS neural network.

REFERENCES

[1] C. Eliasmith *et al.*, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202–1205, Nov. 2012.

[2] H. Markram, "The human brain project," *Sci. Amer.*, vol. 306, no. 6, pp. 50–55, May 2012.

[3] T. S. Clawson, S. Ferrari, S. B. Fuller, and R. J. Wood, "Spiking neural network (SNN) control of a flapping insect-scale robot," in *Proc. IEEE Conf. Decis. Control*, Dec. 2016, pp. 3381–3388.

[4] J. Dethier, P. Nuyujukian, S. I. Ryu, K. V. Shenoy, and K. Boahen, "Design and validation of a real-time spiking-neural-network decoder for brain–machine interfaces," *J. Neural Eng.*, vol. 10, no. 3, 2013, Art. no. 036008.

[5] E. Painkras *et al.*, "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, Aug. 2013.

[6] F. Akopyan *et al.*, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[7] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[8] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Comput.*, vol. 10, no. 7, pp. 1601–1638, 1998.

[9] J. Schemmel, D. Briiderle, A. Griibl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proc. IEEE ISCAS*, May/Jun. 2010, pp. 1947–1950.

[10] B. V. Benjamin *et al.*, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, May 2014.

[11] K. Boahen, "A neuromorph's prospectus," *IEEE Comput. Sci. Eng.*, vol. 19, no. 2, pp. 14–28, Mar./Apr. 2017.

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[13] R. Chawla, A. Bandyopadhyay, V. Srinivasan, and P. Hasler, "A 531 nW/MHz, 128×32 current-mode programmable analog vector-matrix multiplier with over two decades of linearity," in *Proc. IEEE Custom Integr. Circuits Conf.*, Oct. 2004, pp. 651–654.

[14] D. Bankman *et al.*, "An always-On 3.8 μ J/86% CIFAR-10 mixed-signal binary CNN Processor with all memory on chip in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, Jan. 2019.

[15] B. Kim, Y. Liu, T. O. Dickson, J. F. Bulzacchelli, and D. J. Friedman, "A 10-Gb/s compact low-power serial I/O with DFE-IIR equalization in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 12, pp. 3526–3538, Dec. 2009.

[16] N. Cao, S. B. Nasir, S. Sen, and A. Raychowdhury, "Self-optimizing IoT wireless video sensor node with *in-situ* data analytics and context-driven energy-aware real-time adaptation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 9, pp. 2470–2480, Sep. 2017.

[17] N. Cao, S. B. Nasir, S. Sen, and A. Raychowdhury, "In-sensor analytics and energy-aware self-optimization in a wireless sensor node," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2017, pp. 200–203.

[18] S. Sen, D. Banerjee, M. Verhelst, and A. Chatterjee, "A power-scalable channel-adaptive wireless receiver based on built-in orthogonally tunable LNA," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 5, pp. 946–957, May 2012.

[19] D. Banerjee, B. Muldrey, X. Wang, S. Sen, and A. Chatterjee, "Self-learning RF receiver systems: Process aware real-time adaptation to channel conditions for low power operation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 1, pp. 195–207, Jan. 2017.

[20] S. Sen, V. Natarajan, S. Devarakond, and A. Chatterjee, "Process-variation tolerant channel-adaptive virtually zero-margin low-power wireless receiver systems," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 12, pp. 1764–1777, Dec. 2014.

[21] S. Sen, "Invited–Context-aware energy-efficient communication for IoT sensor nodes," in *Proc. 53rd Annu. Design Automat. Conf. (DAC)*, Jun. 2016, Art. no. 67.

[22] B. Chatterjee, D. Das, S. Maity, and S. Sen, "RF-PUF: Enhancing IoT security through authentication of wireless nodes using *in-situ* machine learning," *IEEE Internet Things J.*, to be published.

[23] K. Sato *et al. An in-depth look at Google's first Tensor processing Unit (TPU)*. Accessed: Jul. 3, 2017. [Online]. Available: https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

[24] Y. LeCun *et al. The MNIST Database*. Accessed: Jun. 10, 2017. [Online]. Available:http://yann.lecun.com/exdb/mnist/

[25] A. Krizhevsky *et al. The CIFAR-10 Dataset*. Accessed: Jun. 10, 2017. [Online]. Available: https://www.cs.toronto.edu/ kriz/cifar.html

[26] W. Nebel and J. Mermet, *Low Power Design in Deep Submicron Electronics*. Amsterdam, The Netherlands: Springer, 1997.

[27] B. Razavi, *Design of Analog CMOS Integrated Circuits*. New York, NY, USA: McGraw-Hill, 2000.

[28] R. Sarpeshkar, T. Delbruck, and C. A. Mead, "White noise in MOS transistors and resistors," *IEEE Circuits Devices Mag.*, vol. 9, no. 6, pp. 23–29, Nov. 1993.

[29] A. Krizhevsky, I. Sutskever, and G. F. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] *MatConvNet: CNNs for MATLAB*. Accessed: Jun. 19, 2017. [Online]. Available: http://www.vlfeat.org/matconvnet/

[31] W. M. C. Sansen, *Analog Design Essentials*. Amsterdam, The Netherlands: Springer, 2006, pp. 421–455.

[32] J. Park, S. Ha, T. Yu, E. Neftci, and G. Cauwenberghs, "A 65k-neuron 73-mevents/s 22-pJ/event asynchronous micro-pipelined integrate-and-fire array transceiver," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2014, pp. 675–678.

[33] B. Chatterjee, P. Panda, S. Maity, K. Roy, and S. Sen, "An energy-efficient mixed-signal neuron for inherently error-resilient neuromorphic systems," in *Proc. IEEE Int. Conf. Rebooting Comput. (ICRC)*, Nov. 2017, pp. 1–2.

[34] C. Thakkar and E. Alon, "Design of multi-Gb/s multi-coefficient mixed-signal equalizers," M.S. thesis, Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, 2014.

[35] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan. (2017). "Computing in memory with spin-transfer torque magnetic RAM." [Online]. Available: https://arxiv.org/abs/1703.02118

[36] D. Patterson *et al.*, "Intelligent RAM (IRAM): Chips that remember and compute," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 1997, pp. 224–225.

[37] I. E. Ebong and P. Mazumder, "CMOS and memristor-based neural network design for position detection," *Proc. IEEE*, vol. 100, no. 6, pp. 2050–2060, Jun. 2012.