Energy-Efficient Deep Neural Networks with Mixed-Signal Neurons and Dense-Local and Sparse-Global Connectivity

(Invited Paper)

Baibhab Chatterjee and Shreyas Sen
bchatte@purdue.edu,shreyas@purdue.edu
Purdue University
West Lafayette, Indiana, USA

ABSTRACT

Neuromorphic Computing has become tremendously popular due to its ability to solve certain classes of learning tasks better than traditional von-Neumann computers. Data-intensive classification and pattern recognition problems have been of special interest to Neuromorphic Engineers, as these problems present complex use-cases for Deep Neural Networks (DNNs) which are motivated from the architecture of the human brain, and employ densely connected neurons and synapses organized in a hierarchical manner. However, as these systems become larger in order to handle an increasing amount of data and higher dimensionality of features, the designs often become connectivity constrained. To solve this, the computation is divided into multiple cores/islands, called processing engines (PEs). Today, the communication among these PEs are carried out through a power-hungry network-on-chip (NoC), and hence the optimal distribution of these islands along with energy-efficient compute and communication strategies become extremely important in reducing the overall energy of the neuromorphic computer, which is currently orders of magnitude higher than the biological human brain. In this paper, we extensively analyze the choice of the size of the islands based on mixed-signal neurons/synapses for 3-8 bit-resolution within allowable ranges for system-level classification error, determined by the analog non-idealities (noise and mismatch) in the neurons, and propose strategies involving local and global communication for reduction of the system-level energy consumption. AC-coupled mixed-signal neurons are shown to have 10X lower non-idealities than DC-coupled ones, while the choice of number of islands are shown to be a function of the network, constrained by the analog to digital conversion (or vice-versa) power at the interface of the islands. The maximum number of layers in an island is analyzed and a global bus-based sparse connectivity is proposed, which consumes orders of magnitude lower power than the competing powerline communication techniques.

CCS CONCEPTS

• **Computing methodologies → Neural networks**.

ASPDAC '21, January 18–21, 2021, Tokyo, Japan

ACM ISBN 978-1-4503-7999-1/21/01…$15.00
https://doi.org/10.1145/3394885.3431614

KEYWORDS

artificial neural networks, CMOS, low-energy, mixed-signal, neuromorphic computing, local and global connectivity

ACM Reference Format:
Baibhab Chatterjee and Shreyas Sen. 2021. Energy-Efficient Deep Neural Networks with Mixed-Signal Neurons and Dense-Local and Sparse-Global Connectivity : (Invited Paper). In *26th Asia and South Pacific Design Automation Conference (ASPDAC '21), January 18–21, 2021, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394885.3431614

1 INTRODUCTION

The energy efficiency of the biological human brain is orders of magnitude better than today's conventional von-Neumann computers. The human brain (100 billion neurons, 20 W power, ≈ 20 fJ/synaptic operation), when emulated on a super computer requires 500 MW of power [9]. As an alternative computing paradigm, Neuromorphic computing uses artificial neural networks for computation, and has found success in applications involving image/pattern recognition, miniaturized autonomous robots and neural prosthetic. However, in terms of the energy efficiency, the multiply and accumulate (MAC) operation alone in a neuromorphic synapse can consume ≈ 200 fJ for a traditional digital implementation [1]. For analog/mixed-signal implementations, the energy consumption promises to be lower [3, 6, 7, 17, 20], as shown in Fig. 1. However, unlike digital, analog implementations suffer from noise and variability. For a von-Neumann computer, the non-idealities of analog systems reduces the computing precison, which makes their energy-efficiency less attractive. In deed, digital circuits perform better for applications that require high (> 60 dB) signal-to-noise ratio (SNR) for information processing [18]. However, when SNR requirements are relaxed, analog computation can be orders of magnitude better in terms of energy and area efficiency. This is because the analog macros can represent a mathematical function with intrinsic device/circuit dynamics, instead of relying on the logical implementation of the function using digital gates. As an example, an analog multiplier can be implemented with a single transistor, biased for a fixed transconductance (g_m), which produces an output current as a multiplication of the input voltage and the fixed g_m of the device, with almost infinite resolution. On the other hand, to implement the multiply-and-accumulate (MAC) operation using 8b digital multipliers, ≈ 3000 transistors are required [6]. The combined static leakage of such a large number of transistors is usually found to be comparable to, if not greater than the bias current of the analog implementation in scaled technologies. At this point, it is interesting to note that the von-Neumann architecture achieves

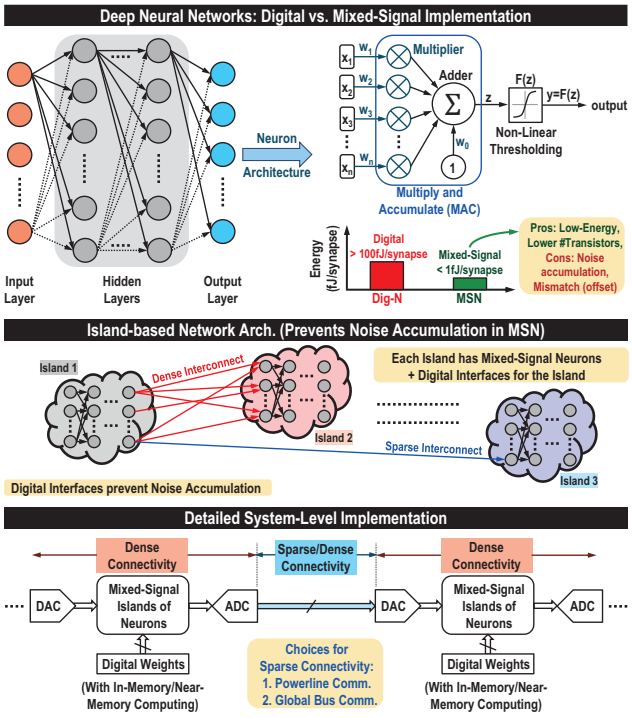

Figure 1: Motivation of using Mixed-Signal Implementation for Deep Neural Networks: Mixed-Signal Neurons (MSN) achieves almost two orders of magnitude better energy efficiency than digital, but requires the input signals to be analog, and suffers from noise accumulation/mismatch as the number of layers increase. The proposed island-based architecture with dense-local and sparse-global connections, with analog-to-digital conversion at the output interface of the islands, eliminates the noise accumulation for MSN.

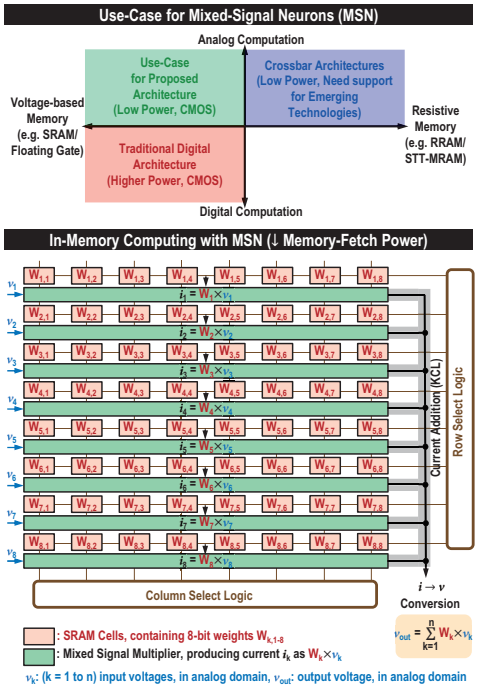

Figure 2: Use-case for the Proposed Architecture in the (memory, computation) space: The proposed system is most suitable when we have voltage-based memory (e.g. SRAM/Floating Gate based memory), with in-memory mixed-signal computation for low-power operation.

high accuracy through multi-bit representation which necessitates a digital implementation. However, today's deep neural networks (DNN) contain multiple connections from its inputs to output due to its distributed multi-path nature and hence the noise and variability of analog transistors can be tolerated to some extent due to this inherent error-resiliency [6].

As the size of the DNN increases to handle a large amount of data (also to find distinguishable features in higher dimensions), today's implementations often divide the computation into multiple cores/islands/processing engines (PEs). The connectivity among these PEs could be dense or sparse, depending on the architectural definitions, physical proximity of the cores and the statistics of the data traffic. Digital designers usually implement the PEs in a modular way, based on the dataflow structure, and hence the number of stages/layers in the PE is determined by the algorithm. However, for analog/mixed-signal implementations, the cumulative effect of noise and mismatch within the island should also be considered for determining the number of consecutive layers in the island. As shown in Fig. 1, the analog/mixed-signal islands would contain digital interfaces (ADC and DAC: analog-to-digital conversion at the output and digital-to-analog-conversion at the input) to eliminate the effect of noise accumulation/transfer to the next island. In this paper, we analyze the allowable size of the islands for mixed-signal implementations. As shown in Fig. 2, the proposed architecture is most suitable for Analog, in-memory (or near-memory) computation with voltage-based memories that can be implemented in a cost-efficient CMOS technology.

The PEs are typically connected together with a power and area-hungry Network-on-Chip (NoC) in digital implementations, which is not efficient for sparse connectivity. Inspired by our earlier work [5, 8, 21] on energy-efficient communication, we also propose the use of a global bus for sparse connectivity which offers energy-efficient inter-island communication, and compare it with the competing power-line communication (PLC) technology [2].

1.1 Related Work

Analog/mixed-signal neuromorphic systems are becoming increasingly popular among researchers for their extreme energy efficiencies, along with their compatibility with CMOS. The BrainScales project [20] helped developing a high-speed (1000-10,000 times faster than the human brain) system that uses analog computation aided by digital asynchronous communication. The design consists ≈ 200k analog neurons with 40M synapses, while consuming ≈ 1kW at 125 MHz frequency (156 fJ per synapse). Neurogrid [3] uses a mixed-signal design approach, and reduces transistor

count by sharing synapses and dendritic trees, as shown in [4]. Neurogrid consists ≈ 1M neurons, each with 8k synapses and consumes 3.1W for real-time computations (390fJ per synapse). Both BrainScales and Neurogrid are aimed towards spiking neural networks (SNNs) which models instantaneous neural activities with a current-switching spiking neuron architecture. This requires complex learning models such as spike timing dependent plasticity (STDP). On the other hand, convolutional neural networks (CNN) employ simple back-propagation-based learning algorithms which utilize a multiply-and-accumulate (MAC) model [15] for the neuron. CNNs are extensively used today for image/pattern recognition applications such as surveillance and weather predictions. Our earlier work on voltage-mode, small-signal mixed-signal neuron (MSN) [6, 7] demonstrated an attractive power-bandwidth trade-off through dominant pole compensation, achieving energy efficiencies < 1fJ for the synaptic MAC operation.

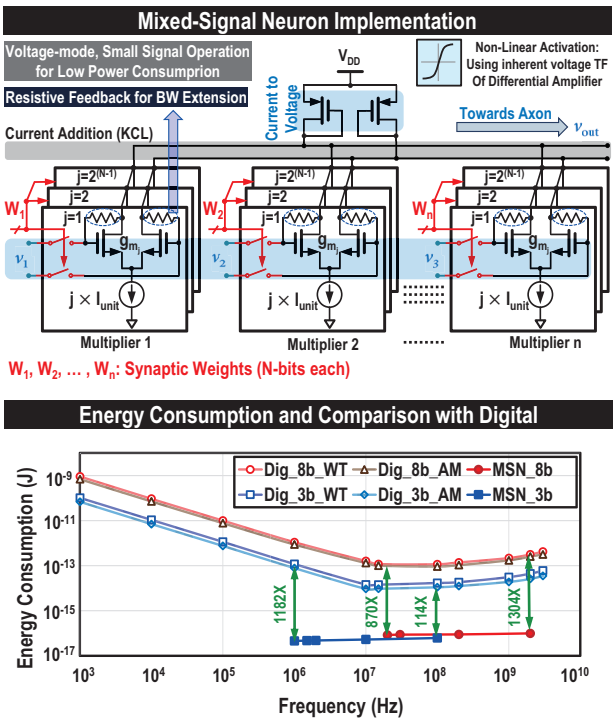

Figure 3: Our previous work [6] on low-power Mixed-Signal Neuron (MSN) implementation, utilizing a DFE-based structure with weight-independent static currents and bandwidth extension through dominant pole compensation. Achieved energy benefits are >100X across all frequencies, compared to digital implementations (WT: Wallace Tree multiplier, AM: Array Multiplier).

Fig. 3 presents the MSN architecture shown in [6]. The neuron consists of n multipliers (synapses) of N-bits each. Each slice j in multiplier i produces an AC current equal to $g_{m_j} \times v_i$ where g_{m_j} is the effective transconductance of the input stages in slice j (proportional to the constant current $j \times I_{unit}$ in each slice), and v_i is the input ac voltage to multiplier i. The weight W_i in a multiplier controls which slices contribute to the overall ac current by turning on (or off) the switches for the input v_i. The overall ac current addition happens inherently through KCL at the output node, which is converted to an equivalent voltage (v_{out}) through the PMOS load impedances. v_{out} can then be expressed as shown in eq. 1.

$$v_{out} = F\left(A_v \times \sum_{k=1}^{n} W_k v_k\right) \tag{1}$$

where F denotes the inherent non-linear activation of the differential amplifier-based structure, and A_v is the overall small-signal voltage gain of a multiplier which needs to be made equal to 1 for a CNN. The use of weight-controlled switches at the input allows a fixed static current, thereby allowing non-linear PMOS loads (instead of area-inefficient linear resistors which would have been required had the weights switched the current sources). The use of irregular slices (value of the current source increases in a binary fashion, while the sizing of the input transistor remains constant) improves the energy efficiency by a factor of $\frac{2^{N-1}}{N}$. The Resistive feedback in each slice compensates for the dominant pole, which extends the bandwidth of operation. Alternatively, the dominant-pole compensation allows for the use of larger devices at the same bandwidth, which reduces the effects of noise and mismatch. Further details can be found in [6].

1.2 Our Contribution

In this paper, we present the following:

(1) **Analysis of Non-idealities with increasing synapse area, for DC-coupled/AC-coupled architectures (Section 3)**: We show that for iso-area consumption, AC-coupled MSNs offer >10X better non-ideality percentage (effect of noise and mismatch) than DC-coupled MSNs for allowable non-linearity ranges as defined by system -level simulations.

(2) **Analysis of maximum allowable stages in an island/PE (Section 4.1)**: Given an area and allowable non-ideality ranges from system-level simulations, we show that the maximum number of allowable layers in an island is limited by the resolution required (for e.g., 4 layers for 8b resolution).

(3) **Analysis of the power overhead due to ADC/DACs in terms of the system power (Section 4.2)**: We show that the number of islands in the DNN, before which the ADC/DAC power starts consuming >10% of the system power, is a function of the network size (number of synapses). We also show that for practical networks, island-based communication with MSN offers better energy-efficiency than digital.

(4) **Analysis of global bus-based sparse communication as an alternative to NoC-based or PLC-based architectures (Section 5)**: We analyze the global bus-based sparse connectivity in comparison with PLC-based connectivity and show the energy benefits (which is more than 3 orders of magnitude better than PLC-based architecture).

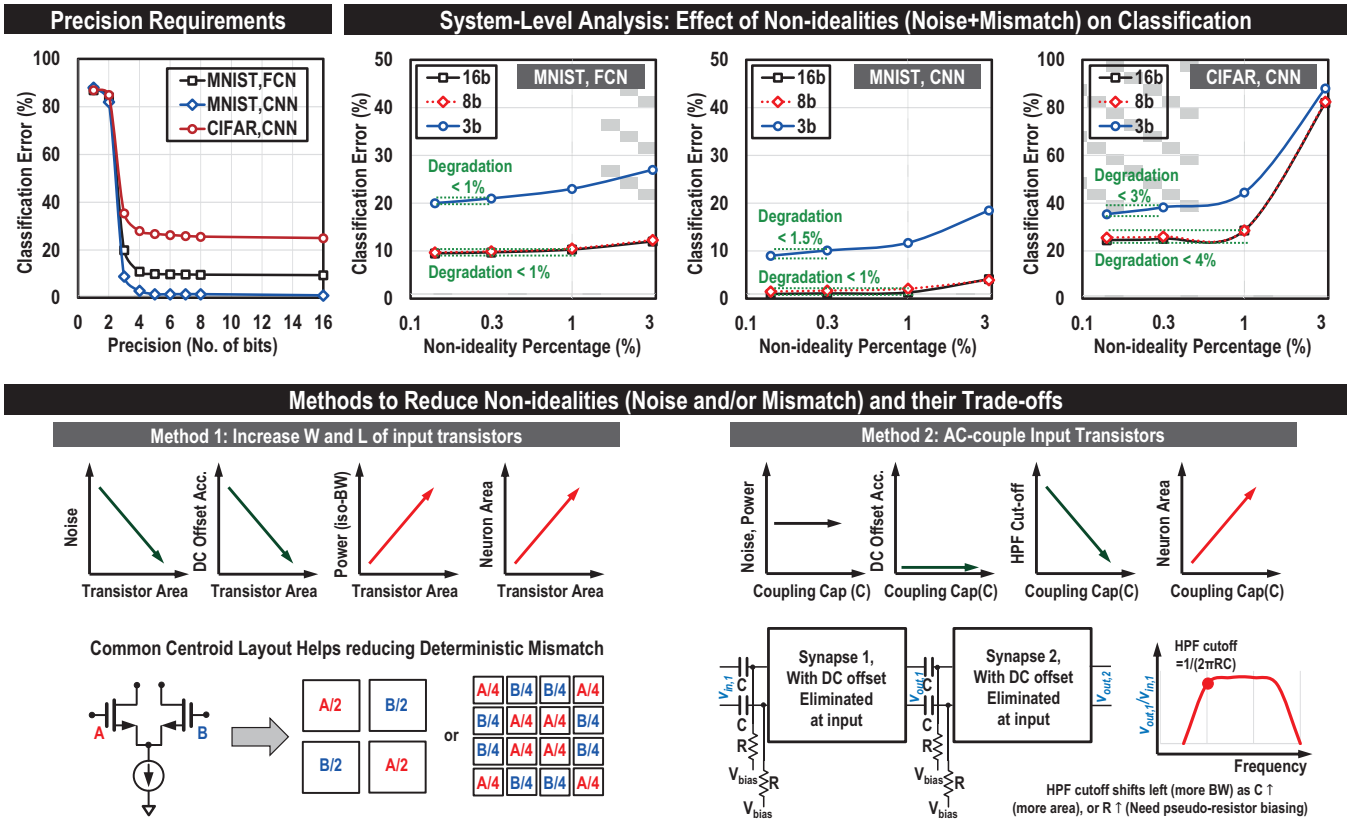

Figure 4: System-level Precision analysis for target applications and System level simulation results for 3-bit, 8-bit and 16-bit MS-N: MNIST [16] with FCN, MNIST with CNN and CIFAR-10 [13] with CNN are considered. Details can be found in [6].

2 SYSTEM-LEVEL REQUIREMENTS AND ANALYSIS

It was shown in [18] that analog design is better than digital in terms of power and area efficiency for applications that can tolerate < 8b precision. Google has also recently indicated that > 8b fixed point precision is redundant for most neural network applications [19]. As a proof-of-concept, we have shown in [6] that for MSN-based networks, the accumulated non-idealities in the form of noise and/or mismatch slightly increases the classification error for different digit/image recognition applications, which is shown in Fig. 4. The analysis is performed on the MNIST dataset [16] for handwritten digits and the CIFAR-10 dataset [13] for images. The network architectures used are CNN (LeNet [15] for MNIST, AlexNet [14] for CIFAR-10) and a 784×100×50×10 fully connected network (FCN) for MNIST. The baseline classification error (Fig. 4, top-left) with digital neurons show that the classification error saturates for >8b resolution, while 3-8b resolution shows reasonably acceptable errors for each application.

To represent the non-idealities present in MSN (noise and mismatch), eq. 1 is modified as shown in eq. 2,

$$v_{out} = F\left(A_v \times \sum_{k=1}^{n} w_k (v_k + \sqrt{A} + \Delta_k)\right) \tag{2}$$

which includes the input referred noise voltage ($\sqrt{A}$, thereby denoting an input referred noise power of A) and input referred DC offset (Δ_k) due to mismatch for the k-th multiplier. Details of the analysis can be found in [6].

Eq. 2 can be re-written as eq. 3.

$$v_{out} = F\left(A_v \times \sum_{k=1}^{n} w_k \left(1 + \frac{\sqrt{A} + \Delta_k}{v_k}\right) v_k\right) \tag{3}$$

The quantity $100 \times \left(1 + \frac{\sqrt{A}+\Delta_k}{v_k}\right)$ is called 'Voltage non-ideality percentage' (V_{NIP}), and will be used as a measure of non-ideality from now on. Maximum v_k is assumed to be 400mV as in [6].

From Fig. 4, it can be observed that for 16b/8b resolution with V_{NIP} <1% the degradation in classification error is <1% for MNIST (FCN/CNN) and <4% for CIFAR (CNN). Similarly, for 3b resolution with V_{NIP} <0.3%, the degradation in classification error is <1.5% for MNIST (FCN/CNN) and <3% for CIFAR (CNN). This analysis indicates that as long as V_{NIP} <1% for 8b resolution (0.3% for 3b), the neural network can tolerate the effects of noise and mismatch.

Fig. 5 shows the noise and mismatch present in one multiplier stage of the MSN of Fig. 3. As explained in [6] and in Fig. 4, larger W and L for the input transistors help in reducing both noise and mismatch (DC offset), at the cost of area. Interestingly, as shown in Fig. 5, the absolute value of the noise voltage is ≈5X smaller than

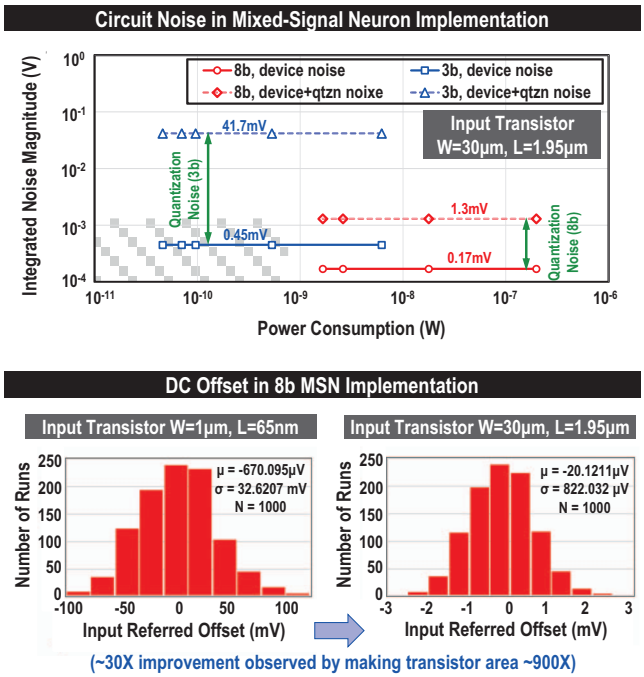

Figure 5: Noise and Input-Referred Mismatch present in one Multiplier of the MSN. The device noise (input referred noise voltage, integrated over the bandwidth) for the 8b neuron is only 0.17mV, while the standard deviation for the input referred DC offset is 0.82mV, indicating that DC offset/mismatch is more detrimental for MSN.

the absolute value of the DC offset (8b scenario), which means that it is possible to reduce V_{NIP} through a combination of AC coupling (removes DC offset with an input coupling capacitor) and device upsizing (reduces noise).

3 ANALYSIS: DC-COUPLING/ AC-COUPLING OF MSN STAGES

As shown in Fig. 4 (bottom-right), AC-coupling subsequent stages of MSN results in a high-pass filtering (HPF) action, with the cutoff frequency of the HPF being $\frac{1}{2\pi RC}$, where R denotes the biasing resistance and C represents the AC-coupling capacitance. Assuming R=20MΩ or higher (which requires a pseudo-resistor-based MOSFET implementation [11] for the biasing resistances), the cutoff frequency of the HPF is shown in Fig. 6 w.r.t. the capacitance value. We calculate that a capacitance of 80fF is required for an HPF cutoff of 100kHz, which is suitable for operating frequency >1MHz. The area of the 80fF MIMCap (Metal-Insulator-Metal Capacitor) in a standard 65nm technology is about 100μm^2.

Fig. 7 shows the variation in the V_{NIP} as a function of the multiplier area. For the DC-coupled MSN, the (W/L) of the input transistors are varied from (1μm/65nm) to (30μm/1.95μm) and the V_{NIP} is simulated. For the AC-coupled MSN, 80fF coupling capacitors were placed at the inputs (total 160fF for differential inputs), and then the (W/L) of the input transistors are varied from (1μm/65nm) to (30μm/1.95μm). For V_{NIP} <1% (as obtained as a requirement from the system level simulations) with 8b resolution, total multiplier area is only 330μm^2 for the AC-coupled case and about 1050μm^2 for the DC coupled case. For 3b MSN, the requirement of V_{NIP} <0.3% is only fulfilled with AC-coupled architecture, with multiplier area ≈400μm^2. For a nominal multiplier area of 400μm^2 or above, AC-coupled architecture achieves ≈10X better V_{NIP} than DC-coupled architecture. Also, the digital implementation of the Wallace-Tree (WT) multiplier consumes ≈700μm^2 area, obtained using Synopsys EDA tools, implying that ***the MSN can achieve acceptable system-level performance with ≈100X lower power (Fig. 3) and ≈1.75X lower area (Fig. 7) than digital implementations.***

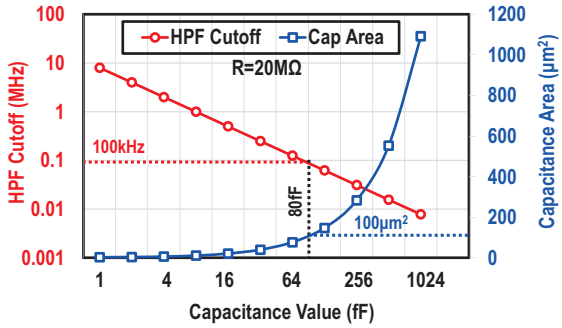

Figure 6: Analysis of the required capacitance for AC coupling (Assumption: biasing resistance = 20MΩ): The HPF cutoff frequency and the capacitance area in a 65nm technology are shown in the two y-axes.

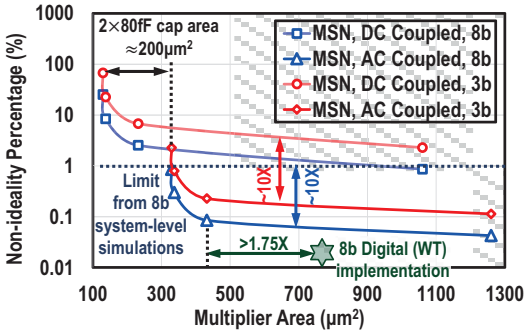

Figure 7: Synapse (one multiplier) Area vs Non-ideality Percentage for DC coupled and AC coupled Mixed-Signal Neurons, and their comparison with WT-based digital multiplier. The non-ideality of the digital multiplier is effectively assumed to be zero.

4 ANALYSIS: NUMBER OF STAGES IN ISLAND

4.1 Limits arising from Resolution

For the AC-coupled scenario, the accumulated non-ideality (which is only the noise voltage, obtained from the linear addition of noise

power for subsequent stages) is plotted in Fig. 8 as a function of the number of stages in the island. For v_k=400mV and 8b resolution (i.e. $\frac{400}{2^8-1}$mV = 1.5686mV resolution), the allowable accumulated one-sided non-ideality is only 1.5686mV/2, which corresponds to a V_{NIP} limitation of 0.196%. Assuming a multiplier area of ≈400μm^2, and taking the non-ideality of the individual stage accordingly, we have obtained Fig. 8. For 8b resolution, the allowable limit for consecutive MSN stages is 4. This increases to ≈ 850 for 3b resolution since the allowable accumulated V_{NIP} increases exponentially with lowered resolution, while V_{NIP} itself increases in proportion to the square-root of the number of stages.

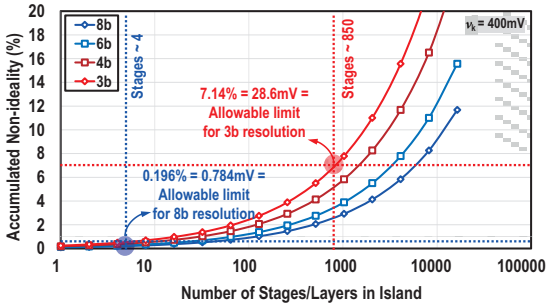

Figure 8: Analysis on the number of maximum consecutive MSN stages/layers for which accumulated non-ideality in MSN surpasses the resolution limit. The resolution limit comes from the maximum signal swing, divided by the number of resolution-steps.

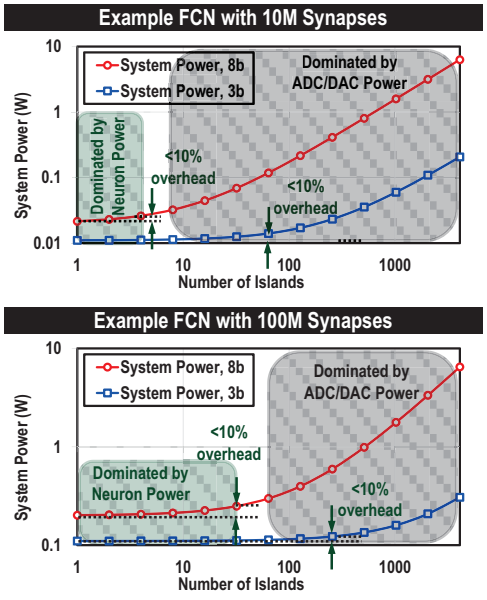

Figure 9: Overall System Power Consumption with two example FCNs, containing 10M and 100M synapses respectively. With the smaller network (10M synapses in this case), for 8b resolution, the system power becomes almost entirely dominated by the ADC/DAC power as the number of islands increase beyond 4. With a realistic DNN (100M synapses), ADC/DAC power is <10% of the total system's power up to 32 islands.

4.2 ADC/DAC Power and System Power

As the number of islands increase in larger and larger DNNs, the power consumption in the interfacing circuitry (ADC and DAC) increases. This is shown in Fig. 9, with two example DNNs with 10M and 100M synapses. With 8b resolution for the smaller network (10M synapses), the ADC/DAC power becomes >10% of the overall system power as the number of islands increase beyond 4. For this analysis, the ADC/DACs are assumed to have an energy efficiency of 30fJ/conversion step [10]. With a realistic DNN (100M synapses), the power consumption in ADC/DAC is <10% of the total system's power up to 32 islands. Since MSN is already > 100X better in terms of energy efficiency than digital neurons, this additional power overhead is insignificant. At this juncture, it is important to note that the initial characterization networks used in [6] are fairly small (<100k synapses for the 784×100×50×10 FCN, ≈1M synapses for LeNet and ≈60M synapses for AlexNet). However, the biggest network, AlexNet contains 8 layers, which is more than the allowable Number of contiguous stages/layers in an island as shown in Fig. 8. As a result, the system-level degradation in CIFAR-10 classification with AlexNet CNN is >1%. With 2 islands, this degradation is expected to return to ≈1%.

5 CHOICES FOR SPARSE INTERCONNECTS

For longer distances, the inter-island connectivity is expected to be sparse through architecture design. Today's PE-based DNNs achieve inter-island connectivity through NoCs, which are inefficient in terms of area an power. In this section, we compare two techniques for low-power, long-distance sparse connectivity.

5.1 Powerline Communication (PLC)

Our earlier work [2] introduced the powerline communication approach for enhanced connectivity in neuromorphic systems, and analyzed a hybrid PLC-NoC-based memristive architecture for high throughput and improved energy efficiency. In PLC, small and sparse data is injected onto the power line through a transmitting buffer (Tx), while the data is recovered at the receiver (Rx) side with help of a level shifter, followed by a low-noise amplifier (LNA), optional variable gain amplifiers (VGAs) and a sampler. The amplifier chain is necessary because of the small amount of signal injected by the Tx and the channel loss in the powerline (≈ 10-20dB for the longest path in a 1960μm × 1960μm power grid, depending on the metal trace [2]). The average energy benefit over several benchmarks was observed to be ≈39% at comparable latency, when compared with NoC-based architectures.

5.2 Global Bus-based Communication (GBC)

Noting that PLC requires driving the power grid with a high-power Tx and also needs a highly sensitive Rx because of the small amplitudes of signal received, we propose the use of a dedicated global bus for the sparse connectivity. Fig. 10 compares PLC with the

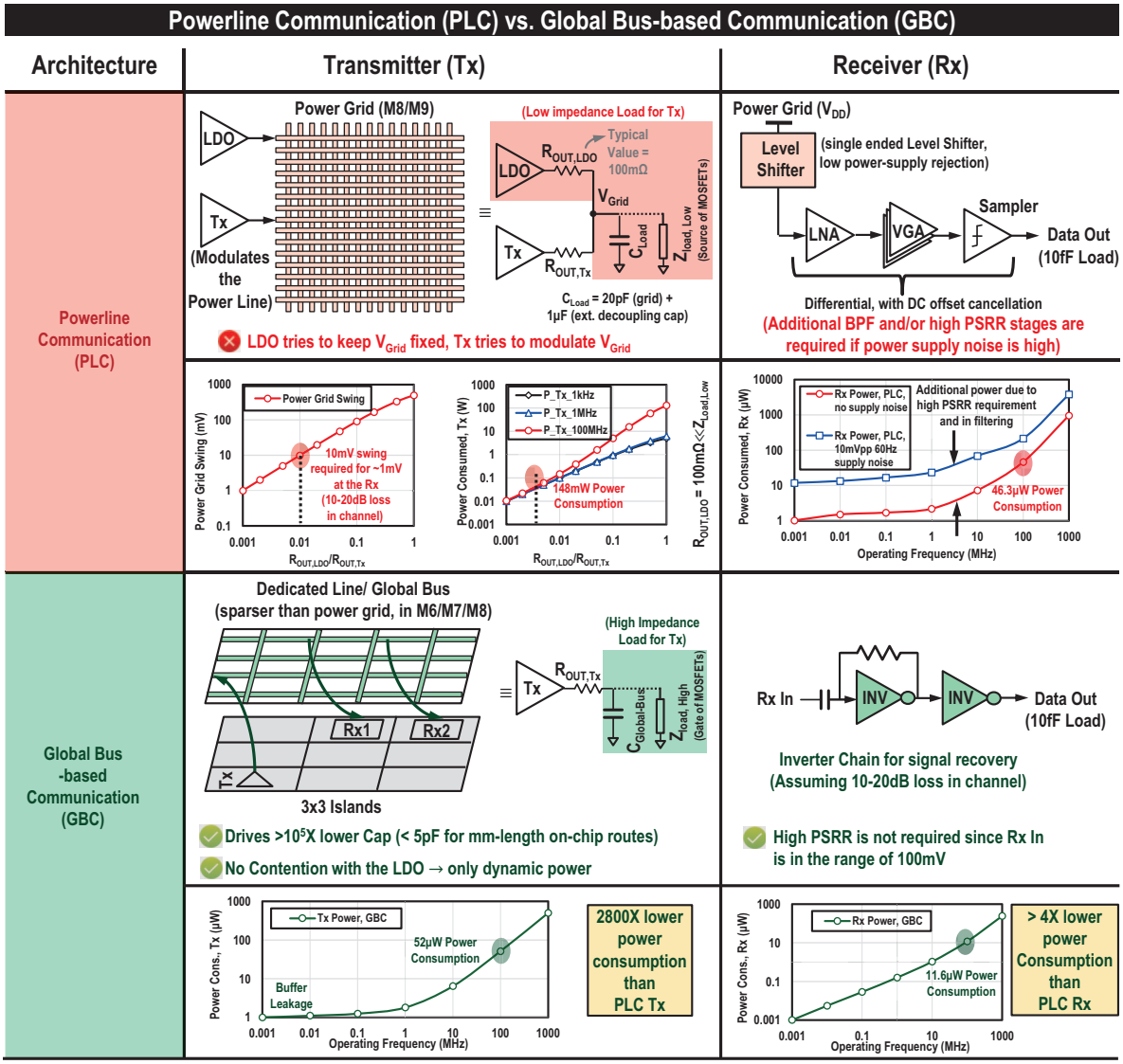

Figure 10: Powerline Communication (PLC) and Global Bus-based communication (GBC) as two techniques for achieving sparse global connectivity among islands. The PLC Tx needs to drive the power grid (a direct contention with the LDOs in the system), thereby consuming a high amount of power (100mW or more), while only imparting a small-amplitude signal (10mV) on the powerline. The PLC Rx, on the other hand, needs to recover the small received signal. In presence of power-supply noise, the PLC Rx becomes increasingly power hungry due to high PSRR and filtering requirements. The GBC Tx and Rx overcomes this problem by having a dedicated global bus between the Tx and the Rx. Since there is no contention with the LDO for the GBC Tx (also, the capacitance to be driven is much lower than PLC Tx), the GBC Tx achieves ≈2800X lower power at 100MHz, while transmitting rail-to-rail signals. The GBC Rx receives higher amplitude signals, and hence can operate at lower power (≈4X lower than PLC Rx, even in the case when PLC Rx does not suffer from power supply noise).

global bus-based Communication (GBC). Please note that for any System on a Chip (SoC) or embedded system, a low-dropout regulator (LDO) is employed which drives the power grid and tries to keep the voltage constant in the grid. Hence, the PLC Tx would always need to burn additional power to overcome the driving capabilities of the LDO (a typical output resistance of a commercial LDO is $R_{OUT,LDO} \approx 100mΩ$ [12]). The output swing for the PLC Tx, based on the ratio of resistances ($R_{OUT,LDO}/R_{OUT,Tx}$) is plotted in Fig. 10, showing that signal swings >10mV can be achieved for $R_{OUT,LDO}/R_{OUT,Tx}$ >0.01 (with 1V supply). The 10mV swing is assumed at the Tx side so that it is distinguishable from noise, and provides enough signal (>1mV) at the Rx side with 10-20dB channel loss [2]. For $R_{OUT,LDO}/R_{OUT,Tx}$ >0.01, the power consumption at the PLC Tx > 148mW for 100MHz operating frequency. For higher

frequencies, the dynamic power consumption becomes dominant, as the power grid is often connected to 100s of nF, or even μF external capacitances for decoupling. A power analysis at the Rx side, using a standard 65nm process, shows 46.3μW power consumption for 100MHz operating frequency, in absence of any power supply noise (with only the level shifter, LNA and sampler). When a 10mVpp, 60Hz noise in the powerline is considered, the power consumption increases due to the requirements of high power supply rejection ratio (PSRR) and filtering.

For GBC, a dedicated global bus is utilized for sparse connectivity. The capacitance of this bus is in the range of a few pF in the worst case (when the bus travels several mm within the chip), as extracted from a standard 65nm process, which results in a dynamic power consumption of 52μW at 100MHz. Since there is no contention with the R_{OUT} of the LDO in this case, the signal swing at the Tx is rail-to-rail (1V), while simultaneously achieving a power benefit of $\approx$ 2800X as compared with the PLC Tx.

Since the voltage transmitted by the GBC Tx is 1V, and the channel loss, in the worst case, is still only 10-20dB, the received signal for GBC Tx is expected to exceed 100mV, providing very good SNR, and hence a simple inverter chain (or a simple resistive feedback LNA followed by an inverter) will be able to recover the signal. The power consumption for such a chain is found to be only 11.6μW at 100MHz, thereby providing a 4X additional power benefit at the Rx side (even when the PLC Rx is considered to have no supply noise). The combined power benefit for the GBC Tx+Rx exceeds 10,000X, when compared with PLC.

6 CONCLUSION

Mixed-signal neuromorphic computing promises almost two orders of magnitude better energy efficiencies than digital implementations, at the cost of additional classification error arising from the analog non-idealities (noise and mismatch). This paper extensively analyzes the usability of mixed-signal neurons (MSN) for deep neural networks, with an island-based architecture with digital interfaces for preventing the effects of accumulated non-idealities. AC coupling is shown to be a more area-efficient method than device upsizing with DC coupling, for reducing the effects of noise and mismatch. The resolution of the application is shown to limit the maximum allowable accumulated non-ideality in an MSN-based island, from which the maximum number of stages/layers in the island can be found out. Multiple islands are shown to incur negligible power overhead (<10%, arising from the ADCs and DACs at the island interface) when the number of synapses contributing to the system energy consumption is more than a few million. Finally, global-bus based communication is shown to be more than three orders of magnitude more power efficient than powerline communication for implementing sparse global connectivity.

7 ACKNOWLEDGMENTS

This work is supported in part through the NSF Career Award (1944602), and in part through the NSF CRII Award (CNS 1657455).

REFERENCES

[1] Filipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, Brian Taba, Michael Beakes, Bernard Brezzo, Jente B. Kuang, Rajit Manohar, William P. Risk, Bryan Jackson, and Dharmendra S. Modha. 2015. TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34, 10 (2015), 1537–1557. https://doi.org/10.1109/TCAD.2015.2474396

[2] Ayush Ankit, Minsuk Koo, Shreyas Sen, and Kaushik Roy. 2019. Powerline Communication for Enhanced Connectivity in Neuromorphic Systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27, 8 (2019), 1897–1906. https://doi.org/10.1109/TVLSI.2019.2907096

[3] Ben Varkey Benjamin, Peiran Gao, Emmett McQuinn, Swadesh Choudhary, Anand R. Chandrasekaran, Jean-Marie Bussat, Rodrigo Alvarez-Icaza, John V. Arthur, Paul A. Merolla, and Kwabena Boahen. 2014. Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations. *Proc. IEEE* 102, 5 (2014), 699–716. https://doi.org/10.1109/JPROC.2014.2313565

[4] Kwabena Boahen. 2017. A Neuromorph's Prospectus. *Computing in Science Engineering* 19, 2 (2017), 14–28. https://doi.org/10.1109/MCSE.2017.33

[5] Baibhab Chatterjee, Ningyuan Cao, Arijit Raychowdhury, and Shreyas Sen. 2019. Context-Aware Intelligence in Resource-Constrained IoT Nodes: Opportunities and Challenges. *IEEE Design Test* 36, 2 (2019), 7–40. https://doi.org/10.1109/MDAT.2019.2899334

[6] Baibhab Chatterjee, Priyadarshini Panda, Shovan Maity, Ayan Biswas, Kaushik Roy, and Shreyas Sen. 2019. Exploiting Inherent Error Resiliency of Deep Neural Networks to Achieve Extreme Energy Efficiency Through Mixed-Signal Neurons. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27, 6 (2019), 1365–1377. https://doi.org/10.1109/TVLSI.2019.2896611

[7] Baibhab Chatterjee, Priyadarshini Panda, Shovan Maity, Kaushik Roy, and Shreyas Sen. 2017. An Energy-Efficient Mixed-Signal Neuron for Inherently Error-Resilient Neuromorphic Systems. In *2017 IEEE International Conference on Rebooting Computing (ICRC)*. 1–2. https://doi.org/10.1109/ICRC.2017.8123656

[8] Baibhab Chatterjee, Dong-Hyun Seo, Shramana Chakraborty, Shitij Avlani, Xiaofan Jiang, Heng Zhang, Mustafa Abdallah, Nithin Raghunathan, Charilaos Mousoulis, Ali Shakouri, Saurabh Bagchi, Dimitrios Peroulis, and Shreyas Sen. 2020. Context-Aware Collaborative Intelligence with Spatio-Temporal In-Sensor-Analytics for Efficient Communication in a Large-Area IoT Testbed. *IEEE Internet of Things Journal* (2020), 1–1. https://doi.org/10.1109/JIOT.2020.3036087

[9] Chris Eliasmith, Terrence C. Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, and Daniel Rasmussen. 2012. A Large-Scale Model of the Functioning Brain. *Science* 338, 6111 (2012), 1202–1205. https://doi.org/10.1126/science.1225266

[10] Pieter Harpe, Cui Zhou, Xiaoyan Wang, Guido Dolmans, and Harmke de Groot. 2010. A 30fJ/conversion-step 8b 0-to-10MS/s asynchronous SAR ADC in 90nm CMOS. In *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*. 388–389. https://doi.org/10.1109/ISSCC.2010.5433967

[11] Reid R. Harrison and Cameron. Charles. 2003. A low-power low-noise CMOS amplifier for neural recording applications. *IEEE Journal of Solid-State Circuits* 38, 6 (2003), 958–965. https://doi.org/10.1109/JSSC.2003.811979

[12] Texas Instruments. 2019. *TPS7A52: 2-A, low-VIN (1.1-V), low-noise, high-accuracy, ultra-low-dropout (LDO) voltage regulator*. Retrieved November 2, 2020 from https://www.ti.com/product/TPS7A52

[13] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. *The CIFAR-10 dataset*. Retrieved November 2, 2020 from https://www.cs.toronto.edu/~kriz/cifar.html

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[15] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based Learning Applied to Document Recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[16] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. 1998. *The MNIST Database of Handwritten Digits*. Retrieved November 2, 2020 from http://yann.lecun.com/exdb/mnist/

[17] Ning Qiao and Giacomo Indiveri. 2016. Scaling mixed-signal neuromorphic processors to 28 nm FD-SOI technologies. In *2016 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 552–555. https://doi.org/10.1109/BioCAS.2016.7833854

[18] Rahul Sarpeshkar. 1998. Analog Versus Digital: Extrapolating from Electronics to Neurobiology. *Neural Computation* 10, 7 (1998), 1601–1638. https://doi.org/10.1162/089976698300017052

[19] Kaz Sato, Cliff Young, and David Patterson. 2017. *An in-depth look at Google's first Tensor Processing Unit (TPU)*. Retrieved November 2, 2020 from https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

[20] Johannes Schemmel, Daniel Brüderle, Andreas Grübl, Matthias Hock, Karlheinz Meier, and Sebastian Millner. 2010. A Wafer-scale Neuromorphic Hardware System for Large-Scale Neural Modeling. In *2010 IEEE International Symposium on Circuits and Systems (ISCAS)*. 1947–1950. https://doi.org/10.1109/ISCAS.2010.5536970

[21] Shreyas Sen. 2016. Invited: Context-aware Energy-efficient Communication for IoT Sensor Nodes. In *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. 1–6. https://doi.org/10.1145/2897937.2905005