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

Baibhab Chatterjee, Shreyas Sen
Purdue University

Speaker Bio

Dr. Shreyas Sen is an Associate Professor in ECE, Purdue University and received his Ph.D. degree in ECE, Georgia Tech. Dr. Sen has over 5 years of industry research experience in Intel Labs, Qualcomm and Rambus. His current research interests span mixed-signal circuits/systems and electromagnetics for the Internet of Things (IoT), Biomedical, and Security. Dr. Sen is a recipient of the NSF CAREER Award 2020, AFOSR Young Investigator Award 2016, NSF CISE CRII Award 2017, Google Faculty Research Award 2017, Intel Labs Quality Award for industrywide impact on USB-C type, Intel Ph.D. Fellowship 2010, IEEE Microwave Fellowship 2008 and seven best paper awards including IEEE CICC 2019 and IEEE HOST 2017, 2018, and 2019. He has co-authored 2 book chapters, over 135 journal and conference papers, and has 14 patents granted/pending.
Outline

• Motivation: Low-power Mixed-Signal Neurons and Connectivity Requirements
• Design of Mixed-Signal Neurons for Ultra-Low Power
• Trade-offs: Noise and Mismatch in Mixed-Signal Designs
• Design considerations for Islands with Mixed-Signal Neurons
• Dense Local and Sparse Global Communication
• Global-bus based communication vs Powerline Communication
• Conclusions
Motivation

Low-power Mixed-Signal Neurons and Connectivity Requirements
Neuromorphic Computing: Current Trends

Energy Bottlenecks

- The bottlenecks:
  - Energy for computation (MAC)
  - Energy for Communication
  - Memory fetch energy (for weights)

Deep Neural Networks: Digital vs. Mixed-Signal Implementation

Motivation

Deep neural networks (DNNs) are widely used in various applications, such as image classification, natural language processing, and speech recognition. However, the energy consumption of these networks is a major concern, especially for mobile and embedded systems. Digital implementations consume more energy compared to mixed-signal implementations, but they are more reliable and easier to design.

Trade-offs

Digital implementations are generally more energy-efficient, but they require more transistors. Mixed-signal implementations, on the other hand, consume less energy but are susceptible to noise and mismatch issues.

Island Considerations

ADC and DAC

Global Communication

Conclusions

Input Layer

Hidden Layers

Output Layer

Neuron Architecture

Energy (fJ/synapse)

Pros: Low-Energy, Lower #Transistors,

Cons: Noise accumulation, Mismatch (offset)

Pros: Low-Energy, Lower #Transistors,

Cons: Noise accumulation, Mismatch (offset)
Deep Neural Networks: Digital vs. Mixed-Signal Implementation

Motivation

Why?
- Utilizes Inherent Physical Properties/Dynamics of Devices/Analog Circuits
- Requires lower number of transistors
- Has low Leakage Power
- However, Analog accumulates Noise and Mismatch

Analog is better

Energy efficiency (pJ/computation)

SNR Requirement for Application (dB)

Deep Neural Networks: Digital vs. Mixed-Signal Implementation

Prevent Noise Accumulation in Analog/Mixed-Signal Computation:

- Each Island has Mixed-Signal Neurons + Digital Interfaces for the Island
- The digital Interfaces prevent Noise Accumulation
- Question: How many layers in an Island? How to implement the Sparse Connectivity?
Deep Neural Networks: Digital vs. Mixed-Signal Implementation

Prevent Noise Accumulation in Analog/Mixed-Signal Computation:

- Each Island has Mixed-Signal Neurons + Digital Interfaces for the Island
- The digital Interfaces prevent Noise Accumulation
- Question: How many layers in an Island? How to implement the Sparse Connectivity?
Use Case for Mixed-Signal Neuromorphic Computation

Motivation

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions

Use-Case for Proposed Architecture (Low Power, CMOS)

Traditional Digital Architecture (Higher Power, CMOS)

Crossbar Architectures (Low Power, Need support for Emerging Technologies)

Resistive Memory (e.g. RRAM/STT-MRAM)

Voltage-based Memory (e.g. SRAM/Floating Gate)

Digital Computation

Analog Computation
Use Case for Mixed-Signal Neuromorphic Computation

Motivation

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions

In-Memory/Near-Memory Computation

- SRAM Cells, containing 8-bit weights $W_{k,1-8}$
- Mixed Signal Multiplier, using transconductance ($g_m$) of transistors

$v_k$: (k = 1 to n) input voltages, in analog domain, $v_{out}$: output voltage, in analog domain
Design of Mixed Signal Neurons

- Digital: computed dynamics – High energy consumption (>100fJ/bit), low noise/variation
- Analog: intrinsic dynamics – Ultra-low energy (<1fJ/bit), more noisy, more variation
Digital MAC and Associated Challenges

**Motivation**

**Trade-offs**

**Island Considerations**

**ADC and DAC**

**Global Communication**

**Conclusions**

---

**Takeaway:**

- Higher number of transistors
- Leakage Power dominates @ low freq.

---

Motivation

MSN

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions

Mixed-Signal MAC

Voltage-mode, Small Signal Operation for Low Power Consumption

Resistive Feedback for BW Extension

Current Addition (KCL)

Multiplier 1

Multiplier 2

Multiplier n

W₁, W₂, … , Wₙ: Synaptic Weights (N-bits each)

Mixed-Signal MAC

**Motivation**

**Trade-offs**

**Island Considerations**

**ADC and DAC**

**Global Communication**

**Conclusions**


\[ i \propto g_{m_1}v_1 \]

\( W_1, W_2, \ldots, W_n \): Synaptic Weights (N-bits each)
Mixed-Signal MAC

W₁, W₂, …, Wₙ: Synaptic Weights (N-bits each)

Mixed-Signal MAC

Motivation

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions


Current Addition (KCL)

\[ i \propto g_m \sum W_k v_k \]

\[ j \times I_{\text{unit}} \]

\[ j = 2^{(N-1)} \]

W_1, W_2, ..., W_n: Synaptic Weights (N-bits each)
Mixed-Signal MAC

Motivation

MSN

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions

Mixed-Signal MAC

Motivation

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions

Mixed-Signal MAC

Voltage-mode, Small Signal Operation for Low Power Consumption

Resistive Feedback for BW Extension

Current Addition (KCL)

Multiplier 1

\[ j \times I_{\text{unit}} \]

\[ W_1, W_2, \ldots, W_n: \text{Synaptic Weights (N-bits each)} \]

Motivation
Trade-offs
Island Considerations
ADC and DAC
Global Communication
Conclusions

Mixed-Signal MAC

Resistive Feedback for BW Extension

Extends bandwidth at iso-power
Or
Reduces power at iso-bandwidth
Or
Allows higher W and L of transistors at iso-bandwidth and iso-power (helps reducing noise and mismatch)

Theoretical Proof of Energy-Efficiency for Mixed-Signal (Please see Ref.):

Trade-offs for Mixed-Signal Design

- Noise
- Mismatch/Input Offset
- Bit resolution requirements and system-level analysis
Noise in Mixed-Signal Neuron

Circuit Noise in Mixed-Signal Neuron Implementation

Mismatch/Input Offset in Mixed-Signal Neuron

DC Offset in 8b MSN Implementation

Input Transistor W=1μm, L=65nm

- Input Referred Offset (mV)
  - Number of Runs: ~30X improvement observed by making transistor area ~900X

Input Transistor W=30μm, L=1.95μm

- Input Referred Offset (mV)

Motivation

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions

Mismatch/Input Offset in Mixed-Signal Neuron

Motivation

Trade-offs

Island Considerations

ADC and DAC

Global Communication

Conclusions


DC Offset in 8b MSN Implementation

Input Transistor W=1μm, L=65nm

\[ \mu = -670.095\mu V \quad \sigma = 32.6207\, mV \quad N = 1000 \]

Input Transistor W=30μm, L=1.95μm

\[ \mu = -20.1211\mu V \quad \sigma = 822.032\, \mu V \quad N = 1000 \]

Introduce the Concept of Voltage Non-ideality Percentage (\(v_{\text{NIP}}\))

\[ v_{\text{out}} = F(\Sigma W_k \left(1 + \frac{\sqrt{A+\Delta_k}}{v_k}\right) v_k) \]

(with noise and mismatch)

\[ v_{\text{NIP}} = \left(1 + \frac{\sqrt{A+\Delta_k}}{v_k}\right) \times 100\% \]

(~30X improvement observed by making transistor area ~900X)
System-Level Classification Error due to Noise/Mismatch

Baseline Precision Requirements (MNIST and CIFAR-10) with Digital Neurons

![Graph showing classification error (%) vs. precision (No. of bits) for MNIST, FCN, MNIST, CNN, and CIFAR, CNN.

Learning ANN Architectures for Evaluation

<table>
<thead>
<tr>
<th>Application</th>
<th>Learning Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST_FCN</td>
<td>784×100×50×10 (2Hidden Layers)</td>
</tr>
<tr>
<td>MNISTCNN</td>
<td>LeNet</td>
</tr>
<tr>
<td>CIFAR_CNN</td>
<td>AlexNet</td>
</tr>
</tbody>
</table>

System-Level Classification Error due to Noise/Mismatch

System-Level Analysis: Effect of Non-idealities (Noise+Mismatch) on Classification

Limit for 8b operation: $v_{NIP} < 1\%$, Limit for 3b operation: $v_{NIP} < 0.3\%$

Methods to Reduce Noise/Mismatch

Method 1: Increase W and L of input transistors

Method 2: AC-couple Input Transistors

Common Centroid Layout Helps reducing Deterministic Mismatch

Synapse 1, With DC offset Eliminated at input

Synapse 2, With DC offset Eliminated at input

HPF cutoff shifts left (more BW) as C (more area), or R (Need pseudo-resistor biasing)
Island Design Considerations

- Reduce Noise and Mismatch: Upsizing or AC-Coupling?
- How many continuous stages/layers are allowed in an Island?
- Total System Power and ADC/DAC Power
Analysis of Capacitance Area for AC Coupling

- Values correspond to a standard commercially available 65nm Process
- 100kHz HPF cut-off ensures that > 1MHz operation is feasible
- Each input needs to have 80fF capacitance (consumes 100μm² Area)
Single-Stage $v_{NIP}$ with DC and AC coupling

- Limit on non-ideality is imposed from System Level Simulations (Slide 27)
- AC Coupling consumes ~10X lower area than DC coupling
- AC coupled Mixed-signal Neuron consumes >1.75X lower Area than a Wallace-Tree based digital implementation (also >100X lower power, as shown in slide 21)
How many Stages in an Island?

- Motivation
- MSN Trade-offs
- Island Considerations
- ADC and DAC
- Global Communication
- Conclusions

**Accumulated Non-ideality (%)**

<table>
<thead>
<tr>
<th>Number of Stages/Layers in Island</th>
<th>1</th>
<th>10</th>
<th>100</th>
<th>1000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>8b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 7.14% = 28.6mV = Allowable limit for 3b resolution
- 0.196% = 0.784mV = Allowable limit for 8b resolution

- Non-ideality is added in power, and then $v_{\text{NIP}}$ is calculated

- For 3b, step size = $\frac{vk}{2^{3}-1}$ = 57.14mV
- For 8b, step size = $\frac{vk}{2^{8}-1}$ = 1.568mV

- Allowable error/non-ideality limit = $0.5 \times$ step size

- Note that in the System-level Analysis (Slide 21), For CIFAR-10 classification with AlexNet, the degradation is more than the FCN and LeNet architectures. This is because AlexNet has 8 Conv/Maxpool layers which is more than the allowable number of continuous stages for 8b neurons.
- ADC Efficiency is assumed to be 20fJ/conversion step (state-of-the-art)
- ADC+DAC power is <10% of the System Power if Number of Islands are <4 (10M synapses), and <32 (100M synapses). Practical Networks such as AlexNet/ResNet contain >>10M synapses
Sparse Global Connection among Islands

- Dense local and Sparse Global Connection
- Powerline Communication vs. Global Bus based Communication for Sparse Connectivity
Dense Local and Sparse Global Connection

• Each Island has Mixed-Signal Neurons + Digital Interfaces for the Island
• The digital Interfaces prevent Noise Accumulation
• Sparse Interconnects: Using Powerline? Or Through a Global Bus?
Powerline Communication (PLC): Transmitter Design

- Motivation
- Trade-offs
- Island Considerations
- ADC and DAC
- Global Communication
- Conclusions

Power Grid (M8/M9)

(Low impedance Load for Tx)

Typical Value = 100mΩ

C_Load = 20pF (grid) + 1μF (ext. decoupling cap)

LDO tries to keep V_{Grid} fixed, Tx tries to modulate V_{Grid}

- Assumption: R_{OUT,LDO} ≈ 100mΩ << Z_{load,Low}
- 148mW power consumption, even at low frequencies, for driving low impedances with a swing of around 10 mV
**Powerline Communication (PLC): Receiver Design**

- Complex Receiver, with High PSRR Requirement

**Power Grid (\(V_{DD}\))**

- Level Shifter
  - (single ended Level Shifter, low power-supply rejection)

- LNA → VGA → Sampler
  - Data Out (10fF Load)
  - Differential, with DC offset cancellation
  - (Additional BPF and/or high PSRR stages are required if power supply noise is high)

**Graph**

- Rx Power, PLC, no supply noise
- Rx Power, PLC, 10mVpp 60Hz supply noise
- Additional power due to high PSRR requirement and in filtering
- 46.3\(\mu\)W Power Consumption
Global-Bus based Communication (GBC): Transmitter Design

Dedicated Line/Global Bus (sparser than power grid, in M6/M7/M8)

- Drives >10^5X lower Cap (< 5pF for mm-length on-chip routes)
- No Contention with the LDO → only dynamic power

- Drives High Impedance in Voltage-Mode
- 2800X Lower Power Consumption than PLC Tx
Global-Bus based Communication (GBC): Receiver Design

- Inverter Chain for signal recovery (Assuming 10-20dB loss in channel)
- High PSRR is not required since Rx In is in the range of 100mV

- Simple Receiver, with relaxed PSRR requirements
- >4X Lower Power Consumption than PLC Rx
Conclusions
Conclusions

- Mixed-Signal Neurons offer >100X Energy benefits than Digital Neurons.
- Noise and Offset affect Classification Accuracy in a Mixed-Signal system, and needs to be reduced.
- AC coupling offers 10X lower area than DC coupling for same amount of analog non-idealities.
- AC coupled Mixed-Signal Neurons are >1.75X Area-Efficient than Digital Neurons.
- Island based architecture with digital interface prevents noise and offset accumulation.
- Number of continuous stages in an island should be fairly low (~4 for 8b implementation).
- Number of total islands could be fairly high (up to 32 for 8b implementation with ~100M synapses) before the ADC+DAC power starts dominating the system power.
- Global-bus, if adopted for sparse communication among islands, offers ~2800X better power efficiency in the transmitter and ~4X better power efficiency in the receiver, as compared to PLC.
THANK YOU