A 3.43TOPS/W 48.9pJ/Pixel 50.1nJ/Classification 512 Analog Neuron Sparse Coding Neural Network with On-Chip Learning and Classification in 40nm CMOS

Fred N. Buhler[1], Peter Brown[1], Jiabo Li[1,2], Thomas Chen[1], Zhengya Zhang[1] and Michael P. Flynn[1]
[1]University of Michigan, Ann Arbor, MI, [2]Intel, Hillsboro

Abstract

A digital-analog hybrid neural network exploits efficient analog computation and digital intra-network communication for feature extraction and classification. Taking advantage of the inherently low SNR requirements of the Locally Competitive Algorithm (LCA), the internally-analog neuron is 3x smaller and 7.5x more energy efficient than an equivalent digital design. This work demonstrates large-scale integration of 512 analog neurons using a traditional scalable digital workflow to achieve a best-of-class power efficiency of 3.43TOPS/W for object classification. At 48.9pJ/pixel and 50.1nJ/classification, the prototype 512-neuron IC achieves 2x efficiency over the digital design while maintaining reliable classification results over PVT.

Introduction

The current revolution in machine learning is based on digital feed forward CNNs comprising many layers. These are notoriously power hungry. In contrast, the Locally Competitive Algorithm accomplishes similar tasks with fewer layers [1]. LCA uses lateral inhibitory connections in addition to feed forward connections in a spiking neural network, as shown in Figure 1. LCA relies on a biologically-inspired analog Leaky Integrate-and-Fire (LIF) neuron model, suggesting an efficient analog implementation.

We introduce an analog neuron that is 3x smaller and consumes 7.5x less power than the comparable digital neuron, taking advantage of the sparse spiking behavior and low SNR requirement of LCA. Our internally-analog neuron presents fully digital I/O to facilitate efficient all-digital inter-neuron communication via address events (AE) and enables compatibility with a digital CAD flow. The prototype 512 neuron IC achieves 48.9pJ/Pixel and 50.1nJ/Classification, 2x more efficient than the state-of-the-art digital implementation [2], while maintaining reliable classification over a wide temperature (0C to 100C) and over 0.8 to 0.95V supply range for 20 measured devices.

Analog/Digital Neuron Comparison

At the moderate accuracy required by LCA, a fundamental analysis of energy per clock cycle clearly shows that an analog neuron is potentially 1,000x more energy efficient. Fig 2 compares the energy versus resolution in bits for equivalent digital and analog implementations of an LIF neuron (assuming a conservative sparsity of 1 firing each 10 cycles). In this comparison, the digital neuron power consumption is for neurons synthesized in 40nm CMOS with various bit-widths. The accuracy of the analog LIF neuron is determined by the kT/C noise of the integration capacitor and comparator noise. The energy consumption of the analog LIF neuron is dominated by the CV^2 energy loss at higher resolutions. Comparator noise, a function of load capacitance, sets comparator energy. Since the integration capacitor energy is dissipated only when an LIF neuron fires, the analog neuron takes significant advantage of LCA's sparse firing behavior.

Analog Neuron

The compact analog LIF neuron (Fig. 3) operates in two phases: excitation and inhibition. During excitation, the neuron accumulates the weighted stimuli from the input pixels and stores the sum as an excitation term. During inhibition, the neuron evolves according to the dynamics of LCA, accumulating the excitation term while subtracting both a leakage term and any weighted inhibition terms triggered by incoming spikes from other neurons. When the neuron potential reaches a threshold, it outputs a spike to inhibit other neurons.

Integration in the LIF neuron is implemented as a current multiplying DAC integrating onto a capacitor C_{int} (Fig. 3). We reuse the same MDAC in both excitation and inhibition to save area. During excitation, the digital inputs to the MDAC are pixel values and excitatory weights, and their product is accumulated onto C_{int}. At the end of excitation, the accumulated value is stored in an analog memory implemented as a parallel plate MoM capacitor saving 35μm^2 compared to a 10b digital register. During inhibition, the digital inputs to the MDAC are the spike AEs and inhibitory weights. The analog memory, with a transconductor, forms a current copier that integrates the stored excitation current onto C_{int}. Also integrated onto C_{int} is a programmable leakage current and any spike AE triggered inhibition (negative) current. A comparator detects when the integrated value exceeds a threshold (V_{ref}), then resets C_{int} and generates a binary spike AE to the other neurons. The single-bit spiking removes the need for power hungry ADCs and simplifies inter-neuron digital communication.

The neuron (Fig. 4) is designed for speed, compact area and low power while maintaining sufficient linearity. Since capacitors C_C and C_{int} are implemented as parallel-plate MOM capacitors and placed above the active circuitry, they occupy no die area. The shared MDAC comprises positive and negative DACs in parallel for operation up to 250MHz. A resistor-degenerated common-source amplifier forms a compact linear transconductor. Instead of charge sharing the C_C and C_{int} capacitors at the end of excitation, both capacitors are shorted during the entire excitation period, reducing the voltage swing on C_{int} by 2x and significantly improving linearity. The entire neuron occupies 320μm^2.

Implementation and Measurement Results

The prototype IC contains two 256-neuron cores, which can operate independently as two 256-neuron networks, jointly as a 512-neuron network, or hierarchically as a two-layer network (Fig. 1). Each core consists of four spiking LCA networks with 64 neurons arranged in a bus-ring topology: eight neurons are connected by a grid to form a single group, and eight groups are connected in series via a systolic ring. This topology allows spikes from any neuron in the network to propagate to every other neuron within eight clock cycles. The 64 neurons within each network share 16Kb SRAM containing the 4096 4-bit weights for the lateral inhibition between neurons.

The dual-core prototype IC is fabricated in 40nm CMOS with a total active area of 1.31mm^2 (Fig. 5). For testing, the

cores are tasked to process two MNIST images in parallel. The cores run at 250MHz with a 0.9V supply and together perform 1.7M classifications per second with a throughput of 1.78G pixels per second. Each core consumes 43.5mW, resulting in an energy efficiency of 50.1nJ per classification, or 48.9pJ per input pixel, while achieving a classification accuracy of 88% on the MNIST handwritten digit dataset. 20 fabricated dies perform reliably over a voltage range of 0.8 to 0.95V and a temperature range of 0 to 100°C (Fig. 5). Fig. 6 compares our prototype analog LCA LIF neural network with the state of the art. This work demonstrates large-scale integration of 512 analog neurons using a traditional scalable digital workflow to achieve a best-of-class power efficiency of 3.43TOPS/W for object classification. Our analog neuron is 3x smaller and 7.5x more energy efficient than a comparable digital neuron, resulting in a 2x more efficient network.

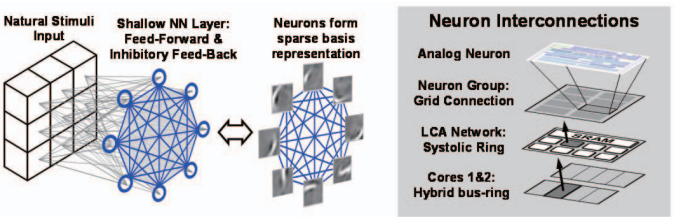

Fig. 1: LCA based neural network.

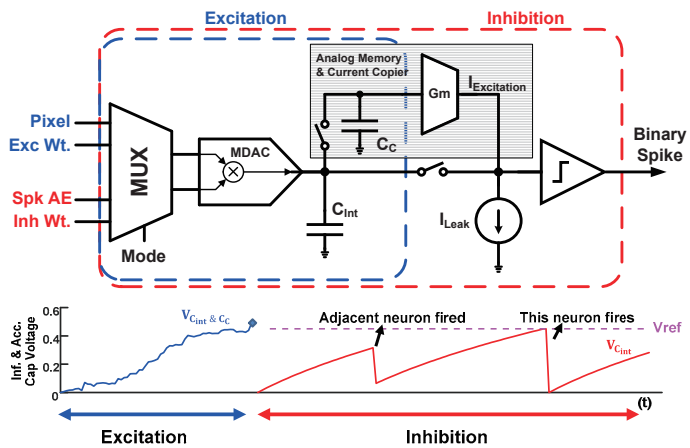

Fig. 3: (Top) Fundamental components of a LIF neuron.
(Bottom) Example waveform of Excitation and Inhibition phases.

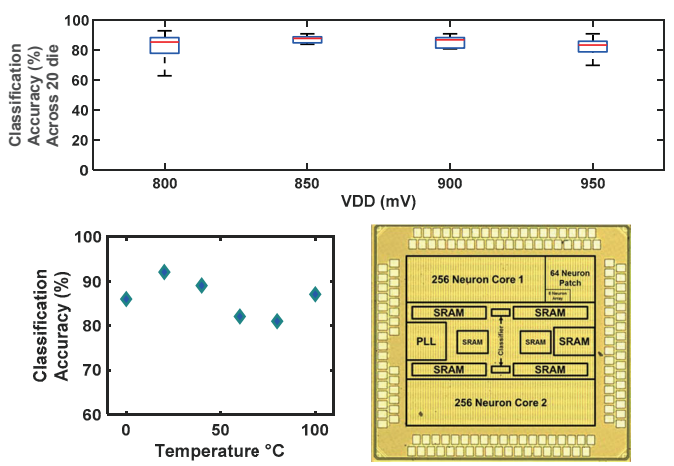

Fig. 5: (Top) Measured classification for 20 devices at different supply voltages.
(Bottom Left) Measured classification accuracy versus temperature.
(Bottom Right) Die photograph, 40nm CMOS, active area 1.31mm^2.

Acknowledgments

This work was supported in part by DARPA and Intel.

References

[1] J. K. Kim, et al., *VLSI Circuits*, 2015.
[2] B. V. Benjamin, et al., *IEEE Proc.*, May 2014.
[3] J. Sim, et al., *ISSCC Dig. Tech. Papers*, 2016.
[4] Y.H. Chen, et al., *ISSCC Dig. Tech. Papers*, 2016.
[5] S. Park, et al., *ISSCC Dig. Tech. Papers*, 2015.
[6] B. Moons, et al., *ISSCC Dig. Tech. Papers*, 2016.

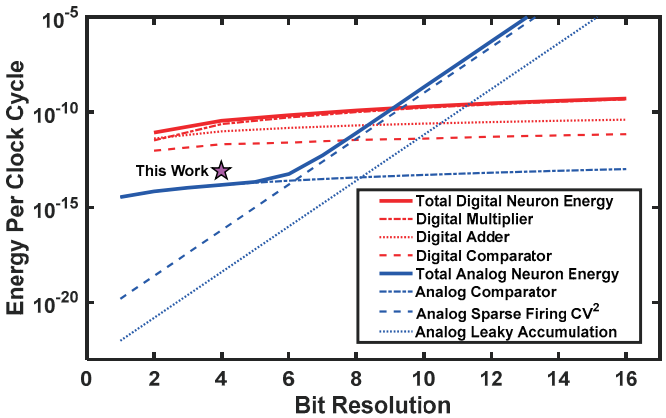

Fig. 2: Energy versus bit resolution for equivalent digital and analog implementations of an LIF neuron.

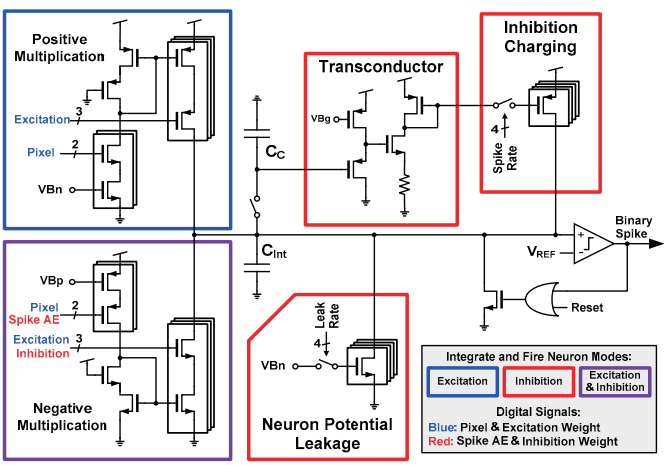

Fig. 4: Simplified transistor implementation of LIF neuron.

Metric	This Work	[1]	[3]	[4]	[5]	[6]
Application	Spiking LCA	Spiking LCA	CNN	CNN	DNN	CNN
Process (nm)	40	65	65	65	65	40
Supply (V)	0.9	1	1.2	1	1.2	1.1
Clock (MHz)	250	635	125	200	200	102
MPixel/sec	1778	2540	—	1.788	1475	2.42
Power (mW)	87	268	45	278	213	76
pJ/pixel	48.93	105.51	—	155481	144.41	31405
GOPS/W	3430	—	1420	—	1930	2600

Fig. 6: Comparison Table.