A 1.52 pJ/Spike Reconfigurable Multimodal Integrate-and-Fire Neuron Array Transceiver

Rajkumar Kubendran
University of California San Diego
La Jolla, CA, USA
rchinnak@ucsd.edu

Weier Wan
Stanford University
Stanford, CA, USA
weierwan@stanford.edu

Siddharth Joshi
University of Notre Dame
Notre Dame, IN, USA
sjoshi2@nd.edu

H.-S. Philip Wong
Stanford University
Stanford, CA, USA
hspwong@stanford.edu

Gert Cauwenberghs
University of California San Diego
La Jolla, CA, USA
gert@ucsd.edu

ABSTRACT

An ultra-low power integrate-and-fire neuron array transceiver with a multi-modal neuron architecture is presented. The design features an array of 16×16 charge-mode mixed-signal neurons that can be configured to implement a variety of activation functions, including step, sigmoid and Rectified Linear Unit (ReLU), through reconfiguration of clocking waveforms through partial reset in charge accumulation and additive stochastic noise by Linear Feedback Shift Register (LFSR) coupling. The neuron outputs spike-based sparse synchronous events, which are either binary (event/no event) or ternary (positive/negative/no events). The reconfigurable energy-efficient design makes this architecture suitable for deep learning and neuromorphic applications like Restricted Boltzmann Machines, Convolutional Neural Networks and general event-driven computing. The 1.796 mm^2 chip fabricated in 130nm CMOS technology consumes 140.6 μW from a 1.8V supply at 92.5 MSpikes/s achieving an energy efficiency Figure-of-Merit (FoM) of 1.52 pJ/Spike. A CNN architecture implemented on the chip using sigmoid and ReLU activation achieves MNIST prediction accuracy of 94.8% and 96.9%.

KEYWORDS

Deep learning, neuromorphic computing, Restricted Boltzmann Machines (RBM), Convolutional Neural Networks (CNN), Integrate and Fire (I&F), Rectified Linear Unit (ReLU).

ACM Reference Format:
Rajkumar Kubendran, Weier Wan, Siddharth Joshi, H.-S. Philip Wong, and Gert Cauwenberghs. 2020. A 1.52 pJ/Spike Reconfigurable Multimodal Integrate-and-Fire Neuron Array Transceiver. In *International Conference on Neuromorphic Systems 2020 (ICONS 2020), July 28–30, 2020, Oak Ridge, TN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3407197.3407209

ICONS 2020, July 28–30, 2020, Oak Ridge, TN, USA

ACM ISBN 978-1-4503-8851-1/20/07.
https://doi.org/10.1145/3407197.3407209

1 INTRODUCTION

Deep learning methods have accelerated the adoption of artificial neural networks (ANN) in many applications, spanning and impacting every field that requires automation. This has resulted in a tremendous demand for hardware accelerators, primarily graphical processing units (GPU) and recently field programmable gate array (FPGA) based systems and custom hardware. There has been an escalated interest in both industry and academia, in developing robust, energy efficient and re-configurable hardware.

On the other hand, neuromorphic systems have gained significant prominence recently, as it holds promise to extreme energy efficiency and low latency, which is suitable for edge computing and Internet-of-Things (IoT) applications [3]. Inspired by biophysical principles, the neuron design can use noise and mismatch variations to its advantage, while generating and processing spike-based events that can be efficiently implemented using switched-capacitor circuits and techniques for analog implementations [5, 6, 8] and advanced technology nodes for digital implementations [1, 2, 7].

This paper proposes an architecture consisting of 256 neurons and supporting peripheral drivers and circuits, that can be programmed and configured to implement a variety of features tailored for different ANNs. A voltage sensing stochastic integrate-and-fire (I&F) analog neuron forms the core component, which can be reused for correlated double sampling (CDS), deterministic or stochastic voltage integration, and binary or ternary level output from comparison of integrated input with a threshold window. The neuron design supports a variety of activation functions, that can cater to deep learning and neuromorphic applications. At only a fraction of the energy efficiency compared to the state-of-the-art implementations, this architecture can be easily adopted and scaled for large networks [9].

The paper is organized as follows. Section 2 presents an overview of the chip architecture and its component blocks. The neuron design stages, including sampling, integrator, comparator, latch and readout, are described in detail in this section. In Section 3, the different modes of operation of the neuron are elaborated, illustrating how the same design can be configured to implement various activation functions. The measurement results of the chip are presented in Section 4, and Section 5 concludes the paper with a summary of the design features and possible applications.

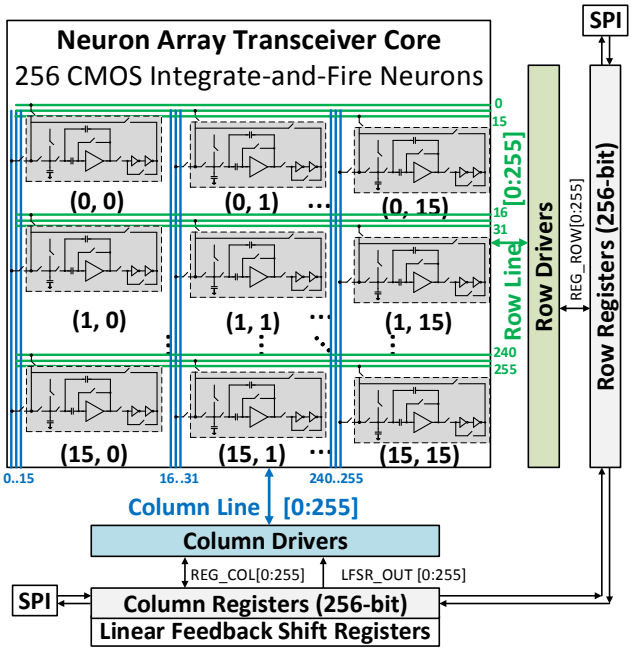

Figure 1: Neuron array transceiver block diagram. 16×16 I&F neuron core with peripheral drivers, biasing, LFSR and SPI.

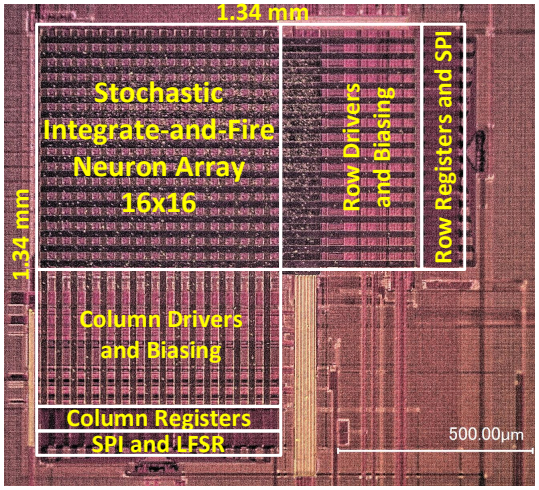

Figure 2: Neural array transceiver chip micrograph.

2 SYSTEM-ON-CHIP ARCHITECTURE AND IMPLEMENTATION

Fig. 1 shows the block diagram of the neuron array transceiver architecture composed of 16×16 I&F neurons with shared rows and columns. The peripheral circuitry consists of row and column drivers and biasing circuits. Linear feedback shift register (LFSR) chains add stochasticity to input samples in order to shape the neuron activation function. The 256-bit row and column registers along the periphery provide I/O communication through a serial peripheral interface (SPI). Configuring the row/column-select switches connects the I&F neuron to its respective row and column. The input pulses can be applied at the rows (or columns), and sampled by the neuron through the column-select (or row-select) switches. The neuron outputs are written into column (or row) registers through the column-select (or row-select) switches.

Highly energy-efficient integrated circuits implement the neural array, each I&F neuron comprising a single high-gain operational trans-conductance amplifier (OTA), switches and output latch to reconfigure the amplifier feedback loop and enable multiple modes of operation. This integrating amplifier doubles as a comparator for digital output generation through global control over the switching timing waveforms. Correlated double sampling (CDS) provides periodic offset cancellation, mitigating systematic variations in the circuit and also establishes a DC operating point for the capacitively coupled OTA. A detailed schematic and timing waveform for the 16×16 I&F neuron circuit are given in [9], where the neurons were configured to implement step and sigmoid activation with binary output. This work extends the reconfigurability of the neuron, to implement ReLU and logistic sigmoid activation with binary or ternary output. This work also presents a system level evaluation of implementing a CNN, with ReLU and sigmoid activation functions.

The I&F neuron array transceiver was designed and fabricated in 130-nm CMOS technology. Fig. 2 shows the chip micrograph. Active area of the chip is approximately 1.796 mm^2 including input SPI and peripheral circuitry, with the row/column drivers, registers and LFSR dominating the area. Individual I&F neuron has an area of 1200μm^2. The chip operates at 1.8V supply for both digital and analog blocks. Measurements show each neuron operation with 63 nW static power drawn from a 1.8 V supply [9]. The total power (static+dynamic) consumption of the chip (256 I&F neurons, biasing and peripherals) is 140.6 μW for data throughput of 92.5 MSpikes/s. The chip was configured through a Xilinx FPGA development board to program bias voltages and currents, send commands to the chip, and receive the data from the chip. The output data from the chip were sent through USB to a PC where data analysis and post processing was performed to characterize the neuron activation functions and conduct the classifier experiments. Measured results from these experiments are presented next.

3 NEURON ACTIVATION MODES

Fig. 3 show measured waveforms illustrating the wide range of activation functions for the neuron available by configuring global control variables of the neuron array transceiver.

3.1 Step activation function: binary or ternary states

Step activation is realized with a tunable threshold voltage V_{TH} in the comparison that sets the stepping threshold for partial reset of the neural integration variable when the neuron flips from "-1" (negative) to "0" (no event) state, or from "0" (no event) to "+1" (positive event) state. If only one threshold voltage is applied for comparison, the neuron output is binary. If two different threshold voltages are applied for comparison in two phases, the neuron output is ternary. Since the CDS phase eliminates any mismatch or offset in the neurons, the 256 neurons in the chip show minimum variation in the activation function, as can be seen in Fig. 3 for binary or ternary levels.

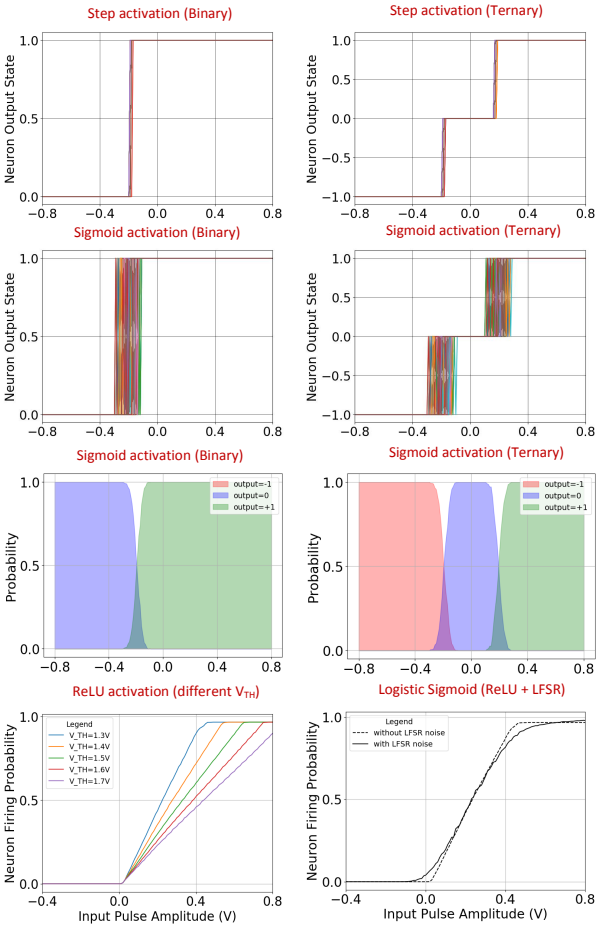

Figure 3: Implementation of step, sigmoid and rectified linear-unit (ReLU) activation functions. Dual thresholds are ±0.2V for both step and sigmoid functions. ReLU activations are shown for different threshold voltage, V_{TH}. Sigmoid activations are generated by coupling LFSR noise to the neuron, for fixed threshold voltage, V_{TH} = 1.3V.

3.2 Sigmoid activation function: binary or ternary states

The sigmoidal graded activation mode is similar to the step activation, but in addition to the input pulses, LFSR pulses are applied to the integrating amplifier input to effect additive noise. Uncorrelated pseudo-random noise is generated through two counter propagating LFSRs whose outputs are modulated and applied to the neurons via column lines. The accumulation of multiple noise pulses smoothens the step transition of the comparator to a sigmoidal function as measurements show in Fig. 3 (second and third rows). Similar to the step activation mode, the sigmoidal graded activation mode can also generate binary/ternary output based on the comparison phase threshold voltages.

3.3 ReLU activation function

Rectified linear-unit (ReLU) activation implements a one-sided hinge function. It is based on rate coding of spikes, unlike the above step and sigmoid modes which are based on individual spike

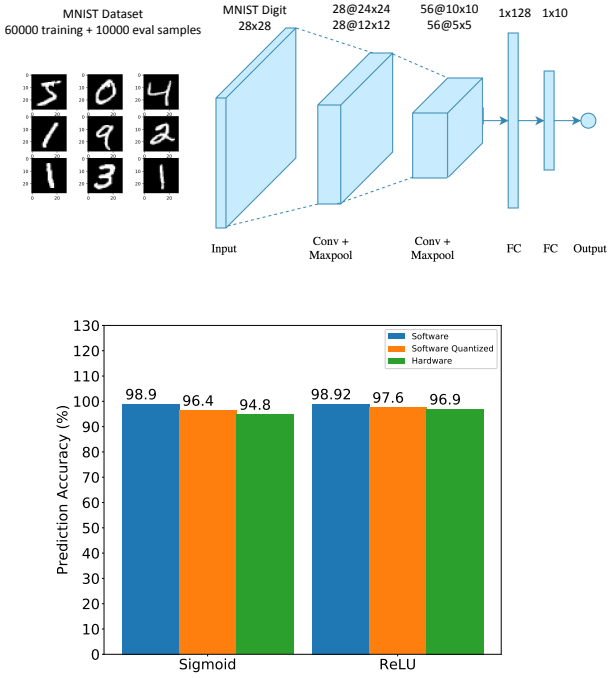

Figure 4: Implemented CNN architecture (top) and evaluation results on MNIST in software vs hardware (bottom).

timing. The ReLU activation is generated using the partial reset mechanism, where the digital output of the neuron is sampled and fed back to the integrating amplifier input to increment or decrement the neural state variable by a fixed amount. As such, the I&F neuron in ReLU mode implements a delta-sigma modulator, with an output mean rate linear in the input, with rectification for inputs below zero. Rectification for inputs above the upper rail is avoided for sufficiently high rail voltage in the upper threshold of the comparator.

Fig. 3 (bottom left) shows the implementation of the ReLU activation for different partial reset threshold voltages. The transfer curve is the average of the 256 neuron output firing probabilities. The number of pulses sent as input to the neuron is fixed to 30. When the input amplitude is negative, the neuron does not fire at all, hence the output probability is zero. When the input pulse amplitude increases above zero, the output firing probability of the neuron increases linearly until it reaches the partial reset threshold voltage where the probability of neuron firing plateaus close to 1. Fig. 3 (bottom right) shows that sigmoid activation function can also be generated by adding LFSR noise to the ReLU neuron.

4 SYSTEM-LEVEL PERFORMANCE

Fig. 4 shows the prediction scores of a CNN implementing MNIST classification [4] in software vs hardware. A quantization aware network, with 2 convolutional and max-pooling layers followed by a fully connected hidden layer and a final output layer for classification, was trained using Keras and Tensorflow with ReLU and sigmoid activation functions, to obtain the weights. The activations

Table 1: Neuron Architecture and Performance Comparison

Parameter	ROLLS Processor [8]	Braindrop [5]	IFAT [6]	This Work
Technology	180nm CMOS	28nm CMOS FDSOI	90nm CMOS	130nm CMOS
Supply Voltage	1.8 V	1 V	1.2 V	1.8 V
Neuron Count	256	4096	65536	256
Activation Function	Step, Sigmoid	Step	Step	Step, Sigmoid, ReLU
Output Levels	Binary	Binary	Binary	Binary, Ternary
Power Consumption	4 mW	NA	1.572 mW	140.6 μW
FoM	NA	381 fJ/Syn.Op.	22 pJ/Spike.	1.52 pJ/Spike.
Active Area	51.44 mm^2	0.65 mm^2	16 mm^2	1.796 mm^2

at each convolution and dense layer was implemented on hardware using the proposed reconfigurable neuron.

The performance metrics of the neuron array transceiver are summarized in Table 1 in comparison with other spike-based transceiver architectures reported in the literature. Since each architecture has different number of neurons, that have different complexity of implementation and operating at different supply voltages, power consumption varies significantly. However, energy efficiency provides an effective Figure-of-Merit (FoM) for comparison of these architectures, given by, FoM = E_{op} / N_{op}, where E_{op} is the energy consumed for synaptic operations N_{op}, performed. The presented architecture achieves 1.52 pJ/Spike, where each pre-synaptic input event constitutes one operation. While this level of energy efficiency compares favorable with most other analog implementations, e.g. [6, 8], Braindrop [5] offers superior FoM owing to using a substantially smaller technology node 28nm in an advanced fully depleted silicon-on-insulator (FDSOI) process. We project similar substantial energy savings when porting the current design to comparable deep-submicron technology nodes that benefit from reverse body biasing (RBB) to provide significant energy savings from leakage.

5 CONCLUSION

An ultra-low power neuron array transceiver with a multi-modal integrate-and-fire neuron architecture was presented. A variety of activation functions were realized by using additional pseudo-random noise or partial reset mechanism that makes this chip versatile to cater to different ANN architectures. The highly reconfigurable and energy efficient design makes this transceiver suitable for deep learning applications such as RBMs, CNNs using neurons with ReLU or sigmoidal activation. Neurons with step or sigmoidal activation, with binary or ternary output levels, can be used for both rate coding or spike timing based neuromorphic applications.

REFERENCES

[1] F. Akopyan et al. 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.

[2] M. Davies et al. 2018. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro* 38, 1 (2018), 82–99.

[3] G. Indiveri et al. 2011. Neuromorphic silicon neuron circuits. *Frontiers in Neuroscience* 5, 73 (2011).

[4] Y. LeCun et al. 1999. The MNIST Database of Handwritten Digits. (1999). http://yann.lecun.com/exdb/mnist/

[5] A. Neckar et al. 2019. Braindrop: A Mixed-Signal Neuromorphic Architecture With a Dynamical Systems-Based Programming Model. In *Proceedings of the IEEE*, Vol. 107. 144–164.

[6] J. Park, S. Ha, T. Yu, E. Neftci, and G. Cauwenberghs. 2014. A 65k-neuron 73-Mevents/s 22-pJ/event asynchronous micro-pipelined integrate-and-fire array transceiver. In *Proceedings of the IEEE Biomedical Circuits and Systems Conference*. 675–678.

[7] J. Pei et al. 2019. Towards artificial general intelligence with hybrid tianjic chip architecture. *Nature* 572 (2019), 106–111.

[8] N. Qiao et al. 2015. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. *Frontiers in Neuroscience* 9 (2015), 141.

[9] W. Wan et al. 2020. A 74 TMACS/W CMOS-ReRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models. *IEEE International Solid State Circuits Conference, ISSCC* (2020).