Low-Power, Low-Mismatch, Highly-Dense Array of VLSI Mihalas-Niebur Neurons

Jamal Lottier Molin*, Adebayo Eisape*, Chetan Singh Thakur*,
Vigil Varghese†, Christian Brandli‡ and Ralph Etienne-Cummings*
*Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218
†Centre of Excellence in IC Design (VIRTUS), Nanyang Technological University, Singapore 639798
‡Insightness AG, Zurich, Switzerland 8005
Email: {jmolin1, aeisape1, cthakur2, retienne}@jhu.edu

***Abstract*—We present an array of Mihalas-Niebur neurons with dynamically reconfigurable synapses implemented in 0.5 μm CMOS technology optimized for low-power, low-mismatch, and high-density. This neural array has two modes of operation: one is each cell in the array operates as independent leaky integrate-and-fire neurons, and the second is two cells work together to model the Mihalas-Niebur neuron dynamics. Depending on the mode of operation, this implementation consists of 2040 Mihalas-Niebur neurons or 4080 I&F neurons within a 3mm $\times$ 3mm area. Each I&F neuron cell consumes an area of 1495μm^2 and the neural array dissipates 360pJ of energy per synaptic event measured at 5.0V power supply ($\sim$14pJ at 1.0V estimated from SPICE simulation).**

I. Introduction

The human brain is by far the most computationally complex, efficient, and robust computing system operating under low-power and small-size constraints. It utilizes over 100 billion neurons and 100 trillion synapses in achieving these specifications. Within the field of neuromorphic engineering, we seek to design systems (typically in Very-Large-Scale Integration (VLSI) technology) which mimic the physical characteristics, functionality, and communication scheme of these neurons. There has been much interest in the design of neural networks for object recognition, classification, and similar visual tasks using these same neurally inspired systems. For feasible integration with cutting-edge technology including autonomous cars, drones and brain machine interfaces, it is essential that these neural networks function under such low-power, small-size, and real-time speed constraints.

Current state-of-the-art large-scale neural arrays implemented in VLSI technology include the Neurogrid (Stanford University) [1], TrueNorth (IBM) [2], SpiNNAker (University of Manchester) [3], and BrainScales (University of Heidelberg) [4]. The work presented here is inspired by the integrate-and-fire array transceiver (IFAT). Although there exists various derivations, in its originality, the IFAT is an array of neurons with dynamic, reconfigurable synapses stored in a memory-based look-up-table (LUT) external to the neuron array. It is implemented in mixed-signal VLSI technology and uses an Address-Event Representation (AER) communication protocol. An AER receiver and transmitter at the periphery of the array allows address-events (AE) to be received and transmitted asynchronously in an event-based, time-division multiplexed fashion. Originally, the IFAT was designed using integrate-and-fire neuron models with both probabilistic and conductance-based synapses [5], [6]. More recently, a 65k-neuron array using a two-compartment neuron cell with conductance-based synapses has been designed in Gert Cauwenberghs' lab [7]. Giacomo Indiveri's lab has designed an array of 32 neurons with local, on-chip asynchronous SRAM for storing synaptic weights in $0.35\mu m$ technology [8]. Indiveri's group also designed a reconfigurable on-line learning spiking (ROLLS) neuromorphic processor [9]. The neuron circuits consists of synapses with bi-stable, spike-based plasticity to achieve short-term and long-term learning.

In this work, we describe a novel array of 2048 Mihalas-Niebur (or 4080 integrate-and-fire) neurons designed in 0.5 μm CMOS technology. In previous work [10], our circuit implementation of this neuron model has been shown to produce various spiking behaviors using an adaptive threshold, allowing for a wide-range of applications. Here we introduce a unique design and functionality of the array architecture that allows for lower power dissipation and increase in number of neurons per mm^2 of silicon area in comparison to other neural arrays in comparable feature-size technologies. Furthermore, this approach allows for significantly reduced mismatch between neuron operation. We finally confirm proper operation of the array by using it to perform a visual processing task, demonstrating its applicability to real-world visual systems.

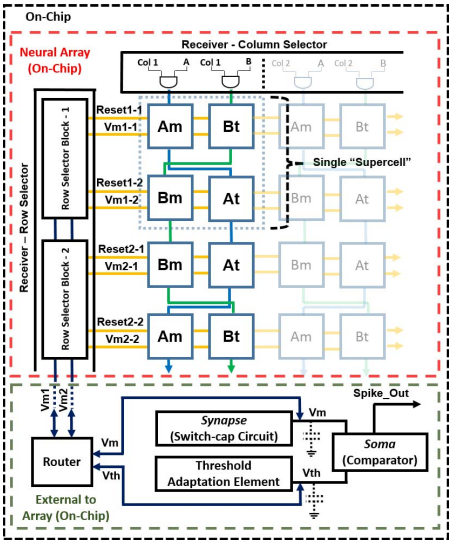

Fig. 1: Block diagram of complete neural array design.

978-1-4673-6853-7/17/$31.00 ©2017 IEEE

II. Neural Array Design

The full neuron array chip block diagram can be seen in Fig.1. It was implemented with means to maximize the neuron array density, minimize power consumption, and reduce mismatch due to process variation. This is achieved by utilizing a single membrane synapse (switch-capacitor circuit) and soma (comparator) shared by all neurons in the array. The connections between neurons is reconfigurable via an off-chip look-up table (LUT). Presynaptic events are sent first through the LUT where the destination addresses and synaptic strengths are stored. Post-synaptic events are then sent to the chip. These events are sent as address-events (AE) along a shared address bus decoded by the row decoder and column decoder on-chip. The incoming address corresponds to a single neuron in the array.

The neuron array is made up of supercells, each containing four cells labeled, Am, At, Bm, and Bt. Each supercell contains two Mihalas-Niebur (M-N) neurons, one using Am and At cells, and the second using Bm and Bt cells. Each of these M-N neurons can also operate as two independent leaky integrate-and-fire (I&F) neurons resulting in a total of four leaky I&F neurons (Am, At, Bm, and Bt). Incoming AE select the supercell in the array and also consists of two additional bits for selecting one of the two M-N neuron (A or B) within the supercell, or one of the four cells when operating as I&F neurons. Finally, the voltage across the storage capacitance for both the membrane cell and threshold cell is buffered to the processor via the router ($Vm1$-X and $Vm2$-X, where X is the row selected). The router is used for selecting which voltage (from the membrane cell or threshold cell) is buffered to the processor as the membrane voltage and/or threshold voltage, depending on the mode selected (M-N mode or I&F mode). This router is necessary for allowing the voltage from the threshold (At or Bt) cell to be used as the membrane voltage when in I&F mode. After the selected neuron cell(s) buffer its stored voltage to the external capacitances C_m and C_t, the synaptic event is applied and the new voltage is buffered back to the same selected cells that received the event. The synapse and threshold adaptation elements execute the neuron dynamics as events are received. If the membrane voltage exceeds the threshold voltage, there is a single comparator (soma) that outputs a logic high (event).

An output arbiter/transmitter is not necessary in our design considering that a neuron only fires when it receives an event. The single output signal always corresponds to the neuron that receives the incoming event. Having a single comparator not only reduces power consumption but also reduces the required number of pads for digital output. In this design we compromise speed (for low-power and low-area) due to the time necessary to read and write to and from the neuron. However, we are still capable of achieving a maximum input event rate of $\sim$1 MHz for proper operation.

III. Circuit Implementation of Mihalas-Niebur Neuron Model

Each cell pair (Am/At and Bm/Bt) in this neural array models the Mihalas-Niebur neuron dynamics [11]. The M-N neuron uses linear differential equations and parameters with biological facsimiles. It consists of an adaptive threshold and was shown to be capable of modeling all of the biologically-relevant neuron behaviors. It uses differential equations modeling the internal currents, membrane voltage, and adaptive threshold voltage dynamics. Update rules are applied for each time the membrane voltage exceeds the adaptive threshold voltage [11]. For circuit implementation of this M-N model, we make a few modifications. The first is omitting internal induced-spike currents, and the second is setting the reset voltage equal to the resting potential. We make these modifications at the expense of the generality of the model. However, this modified M-N model is still capable of implementing 9 of the biologically-relevant spiking behaviors [10]. The modified differential equations for CMOS implementation are as follows:

$$V'm(t) = \frac{g_l^m}{C_m}(V_r - V_m(t)) \tag{1}$$

$$\theta'(t) = \frac{g_l^t}{C_t}(\theta_r - \theta(t)) \tag{2}$$

$$V_m(t+1) = V_m(t) + \frac{C_s^m}{C_m}(E_m - V_m(t)) \tag{3}$$

$$\theta(t+1) = \theta(t) + \frac{C_s^t}{C_t}(V_m(t) - V_r) \tag{4}$$

and,

$$g_l^{m,t} = \frac{1}{r_l^{m,t}} = f_l^{m,t} C_l \tag{5}$$

This modification allows the use of two neuron cells to work together to model a single M-N neuron as discussed in Section II. Eq. (3) and (4) model the change in membrane potential (V_m) and threshold potential (θ) at each time step as the neuron receives an input. C_s^m and C_s^t are the switch-capacitor capacitance depicting the synapse conductance or threshold adaptation conductance, respectively. Cm and Ct are the storage capacitance for the membrane and threshold cells, respectively. E_m is the synaptic driving potential. Eq. (1) and (2) model the leakage dynamics, independent of synaptic connections. $g_l^{m,t}$ are the leakage conductances for the membrane and threshold and are dependent on the clock frequency, $f_l^{m,t}$. The update rules for this modified M-N neuron model are as follows:

$$V_m(t) \leftarrow V_r \tag{6}$$

$$\theta(t) \leftarrow \begin{cases} \theta(t), & \text{if } \theta(t) > V_m(t) \\ \theta_r, & \text{if } \theta(t) \leq V_m(t) \end{cases} \tag{7}$$

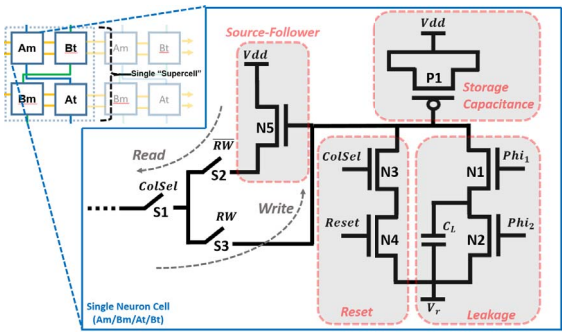

Fig. 2: Neuron cell schematic.

A. Neuron Cell Circuit

The PMOS transistor, $P1$, is the storage capacitance ($\sim$440fF), C_m or C_t, (depending on whether the cell is being used to model the membrane or threshold dynamics) implemented as a MOS capacitor with its source and drain tied to Vdd. Transistors $N1$ and $N2$ model the leakage (Eq. (1) and (2)) via a switch-capacitor circuit with Phi_1 and Phi_2 pulses at a rate of $f_l^{m,t}$ (also $C_L << C_m$). Transistors $N3$ and $N4$ allow for resetting the neuron when selected ($ColSel = 1$). Transistor $N5$ forms a source-follower when coupled with a globally-shared variable resistance located in the processor of the neural array. It is implemented as an NMOS transistor with a voltage bias (V_b). In read mode ($RW = 0$), switch $S2$ is closed such that the voltage across the storage capacitance is buffered to an equivalent capacitance coupled to the synapse and/or threshold adaptation element. In write mode ($RW = 1$), switch $S3$ is closed such that the new voltage from the synapse/threshold elements (after an event is received) is buffered to the storage capacitance.

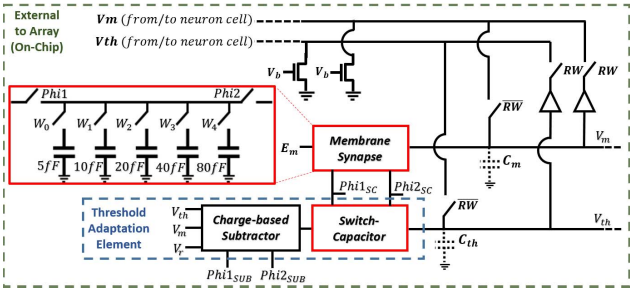

Fig. 3: Single, shared, on-chip synapse and threshold adaptation schematic external to the neural array.

B. Synapse and Threshold Adaptation Circuits

The schematic for modeling the neuron dynamics can be seen in Fig.3. When a neuron receives an event, $RW = 0$, and the neuron's cell is selected and its stored membrane voltage is buffered to the capacitance C_m. In the same manner, if in M-N mode, the threshold voltage is buffered to C_t. The $Phi1_{SC}$ and $Phi2_{SC}$ pulses are then applied (off-chip), adding (excitatory event) or removing (inhibitory event) charge to C_m via the synapse using a switch-capacitor circuit. A second, identical switch-capacitor circuit is used for implementing the threshold adaptation dynamics. As a neuron receives events, the same $Phi1_{SC}$ and $Phi2_{SC}$ pulses are applied to the threshold adaptation switch-capacitor circuit which adds or removes charge to C_t. The new voltage is then buffered ($RW = 1$) back to the neuron cells for storing the new membrane voltage (as well as the threshold voltage if in M-N mode). When using each neuron independently as leaky I&F neurons, the threshold adaptive element is bypassed and an externally applied fixed threshold voltage is used. A charge-based subtractor is used in the threshold adaptation circuit for computing $V_{th} + (V_m - V_r)$ in modeling Eq.4. This subtraction output is the driving potential for the threshold switch-capacitor circuit. An externally applied voltage, E_m, is the synaptic driving potential for the membrane synapse and is used for modeling Eq.3. Finally, the comparator outputs an event when the membrane voltage exceeds the threshold voltage. An external reset signal for both the neuron cell modeling the membrane voltage and cell modeling the threshold voltage is activated for the selected neuron (via $Reset1$-X and $Reset2$-X) when a spike is outputted.

IV. Results

A. Neuron Area

A single neuron cell in this array has dimensions of $41.7\mu m \times 35.84\mu m$. A comparison to other state-of-the-art neural array chips can be seen in Table II. It consumes only 62.3% of the area consumed by a single neuron cell in [6], also designed in a $0.5\mu m$ process. While we achieve 668.9 I&F neurons/mm^2, [6] achieves only 416.7 neurons/mm^2 and [12] achieves only 387.1 neurons/mm^2. We seek to further increase the number of neurons/mm^2 by optimizing the layout of the neuron cell and implementing in smaller feature-size technology.

B. Mismatch

For analyzing mismatch (due to process variations) across the neuron array, we observed the output event to input event ratio for a fixed synaptic weight and input event rate of 1 MHz for each neuron in the array. With this fixed synaptic weight, the 2040 M-N neurons have a mean output to input event ratio of 0.0208 $\pm$1.22e-5. In the second mode of operation, the 4080 I&F neurons have a mean output to input event ratio of 0.0222 $\pm$5.57e-5. For fair comparison, we compare to the results from a similar experiment performed in the $0.5\mu m$ conductance-based IFAT in [6] (See Table I). Our design shows significantly less deviation. Small amounts of mismatch can be taken advantage of in applications that require stochasticity. However, for those spike-based applications that do not benefit from mismatch, in this neural array, it is more controlled. This again is a result of utilizing a single, shared synapse, comparator, and threshold adaptive element for all neurons in the array. The mismatch between neurons is only due to the devices within the neuron cell itself (i.e. membrane capacitance, source-follower transistor).

TABLE I: Array Characterization: Output Events / Input Event

Neural Array	Mean Ratio (μ)	SD (σ)	# Neurons
This Work[a]	0.0222	$\pm$5.57e-5	4080
Vogelstein et al. [6]	0.0210	$\pm$1.70e-3	2400

[a]Characterization using I&F neurons

C. Power Consumption

Another design goal was to minimize power consumption. At an input event rate of 1 MHz, we measure an average power consumption of $360\mu W$ at 5.0V power supply. A better representation of the power consumption is energy per incoming event. From these measurements, this chip consumes 360pJ of energy per synaptic event. A comparison to other state-of-the-art neural array chips can be seen in Table II. Compared to those chips designed in 500nm technology [6] and 800nm [12], we see a significant reduction in energy per synaptic event. Due to complications in the circuit board, we were unable to measure low-voltage operation. However, from simulations we have validated proper operation at 1.0V (at slower speeds). Assuming dynamic energy scales with V^2 (capacitance remains the same), we estimate $\sim$14.4pJ of energy per synaptic event at 1.0V. These results are promising and can be further optimized in smaller feature-size technology.

TABLE II: Neural Array Chip Comparison

Neural Array	Process	Vdd Supply	Neuron Design	Neuron Area	Energy/Event
This Work	500nm	5.0V [1.0V[a]]	Analog	$1495\mu m^2$	360pJ [14.4pJ[b]]
Vogelstein et al. [6]	500nm	5.0V	Analog	$2400\mu m^2$	645pJ
Indiveri et al. [12]	800nm	3.3V[c]	Analog	$2583\mu m^2$	900pJ[c]
Neurogrid [1]	180nm	1.8V	Analog	$1800\mu m^2$	31.2pJ
TrueNorth [2]	45nm SOI	0.85V	Digital	$3325\mu m^2$	45pJ
SpiNNaker [3]	130nm	1.2V	Digital	N/A	43nJ
BrainScales [4]	180nm	N/R[d]	Analog	$1500\mu m^2$	N/R
Park et al. [7]	90nm	1.2V	Analog	$140\mu m^2$	22pJ
Qiao et al. [9]	180nm	N/R	Analog	$918\mu m^2$	4mW[e]

[a]SPICE Simulation showed proper operation at 1.0V, but slower speeds
[b]Using simulated operation at 1.0V Vdd and assuming energy scales with V^2
[c]Reported Vdd and power consumption from version of neural array designed in 350nm technology
[d]Not Reported
[e]Reported average power consumption for a typical experiment

V. Event-based Image Filtering Task

To demonstrate an application using all 4080 I&F neurons, we use the IFAT to perform an image processing application. Given an input image we generate probabilistic AE streams (on PC) such that each address corresponds to a pixel in the image (64×60). The output event-rate of each pixel has a mean firing rate that is proportional to the pixel intensity. These AEs go through a LUT implemented in memory on an FPGA, holding the associated destination addresses and synaptic weights. We first show results from a one-to-one connection between the pixel and its corresponding neuron. Secondly, we show results from a simultaneous warping (by 45°) and spreading (low-pass/blurring filter) operation in which each incoming address event is projected not only to its warped destination, but also to its neighboring neurons at that location with a synaptic weight (stored in the LUT) based on the kernel: [0.2, 0.2, 0.2, 0.2, 0.2]. The output event rates are decoded into pixel intensities and the results are shown (See Fig. 4). These results confirm proper operation of each neuron in the array and ability to use the complete array for performing event-based visual tasks.

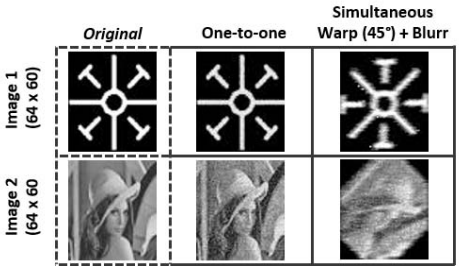

Fig. 4: Visual processing task via chip (one-to-one and warp + blur).

VI. Conclusion

We have demonstrated a $3mm \times 3mm$ neural array of Mihalas-Niebur neurons implemented in 500nm CMOS technology. Its novel design enables minimal power consumption, neuron area, and mismatch by trading-off speed. However, this chip is still capable of surpassing biological real-time speeds. Speed can be increased linearly with the number of shared synapse, threshold adaptation element, and soma circuits per neural array. In continuing work we will design this architecture in a 130nm CMOS technology or smaller to achieve higher neuron density and minimal power consumption.

VII. Acknowledgments

We acknowledge the support from Draper Laboratory.

References

[1] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 699–716, 2014.

[2] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. S. Modha, "A digital neurosynaptic core using embedded crossbar memory with 45pj per spike in 45nm," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*. IEEE, 2011, pp. 1–4.

[3] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[4] 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 *Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on*. IEEE, 2010, pp. 1947–1950.

[5] D. H. Goldberg, G. Cauwenberghs, and A. G. Andreou, "Probabilistic synaptic weighting in a reconfigurable network of vlsi integrate-and-fire neurons," *Neural Networks*, vol. 14, no. 6, pp. 781–793, 2001.

[6] R. J. Vogelstein, U. Mallik, J. T. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses," *IEEE transactions on neural networks*, vol. 18, no. 1, pp. 253–265, 2007.

[7] 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 *Biomedical Circuits and Systems Conference (BioCAS), 2014 IEEE*. IEEE, 2014, pp. 675–678.

[8] S. Moradi and G. Indiveri, "A vlsi network of spiking neurons with an asynchronous static random access memory," in *Biomedical Circuits and Systems Conference (BioCAS), 2011 IEEE*. IEEE, 2011, pp. 277–280.

[9] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses," *Frontiers in neuroscience*, vol. 9, p. 141, 2015.

[10] V. Varghese, J. L. Molin, C. Brandli, S. Chen, and R. E. Cummings, "Dynamically reconfigurable silicon array of generalized integrate-and-fire neurons," in *Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE*. IEEE, 2015, pp. 1–4.

[11] Ş. Mihalaş and E. Niebur, "A generalized linear integrate-and-fire neural model produces diverse spiking behaviors," *Neural computation*, vol. 21, no. 3, pp. 704–718, 2009.

[12] G. Indiveri, E. Chicca, and R. Douglas, "A vlsi array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE transactions on neural networks*, vol. 17, no. 1, pp. 211–221, 2006.