Neuromorphic Networks using Nonlinear Mixed-feedback Multi-timescale Bio-mimetic Neurons

Kangni Liu
Department of ECE
University of Pittsburgh
Pittsburgh, USA
connie.liu@pitt.edu

Shahin Hashemkhani
Department of ECE
University of Pittsburgh
Pittsburgh, USA
shh199@pitt.edu

Jonathan Rubin
Department of Mathematics
University of Pittsburgh
Pittsburgh, USA
jonrubin@pitt.edu

Rajkumar kubendran
Department of ECE
University of Pittsburgh
Pittsburgh, USA
rajkumar.ece@pitt.edu

***Abstract*—Biological neurons exhibit rich and complex nonlinear dynamics, which are computationally expensive and power-hungry for hardware implementation. This paper demonstrates the design and development of a hardware-friendly nonlinear neuron model based on an intuitive control theory perspective. The neuron consists of a mixed-feedback system operating at multiple timescales to exhibit a variety of modalities that resemble the biophysical mechanisms found in neurophysiology. The single neuron dynamics emerge from four voltage-controlled current sources and features spiking and bursting output modes that can be controlled using tunable parameters. The bifurcation structures of the neuron, modeled as a 4D dynamical system, illustrate the roles of sources acting on different timescales in shaping the neural dynamics. For the first time, a neural network test chip consisting of 6 nonlinear bio-mimetic neurons and 10 tunable synapses was designed on 180nm CMOS technology. A 4-neuron network with inhibitory synapses of increasing strength was verified to achieve coupled rhythms. The test chip has an area of 0.6mm x 2mm and consumes 0.753mW of total average power.**

Keywords—neuromorphic, nonlinear dynamics, mixed feedback control, bio-inspired, coupled neural networks

I. Introduction

Neuron models used in Artificial Neural Networks (ANNs) are highly simplified in function and operation, to facilitate scaling to large networks [1]. The most popular implementation that is widely adopted in modern AI hardware is the Rectified Linear Unit (ReLU) [2]. The ReLU neuron primarily realizes a linear activation function for all positive inputs, while rectifying negative inputs to zero, thus introducing a simple form of nonlinearity, as shown in Fig. 1.

On the other hand, Neuromorphic Spiking Neural Networks (SNNs), which intend to be bio-inspired, use a Leaky Integrate-and-Fire (LIF) neuron which can generate voltage spikes or 'events' similar to biological neurons. The LIF neuron is implemented using an RC network that linearly integrates an applied external current to produce a voltage ramp [3]. If this ramp output exceeds a threshold voltage, a spike is generated. Thus, the LIF neuron realizes the step activation, which is also a simple form of nonlinearity generating 'all-or-nothing' events as voltage spikes.

Though ReLU and LIF neurons have been extremely successful when used in a predominantly feed-forward network, they are not suitable to demonstrate neuroscientific principles such as multiple modes of operation (spiking, bursting, etc), coupled oscillations with excitatory and inhibitory connections, and rhythm generation (in/anti-phase).

Biological neurons exhibit rich nonlinear dynamics both at the single neuron level and at the network level, which is enabled through neuromodulation, using multiple feedback paths (local and global) operating at multiple timescales [4-5]. Complex neuron models inspired by neurophysiology have been proposed before, like the Hodgkin-Huxley model [6] and the Izhikevich model [7]. These models capture the biophysics of the neuron accurately but are based on non-intuitive computationally expensive differential equations that are hard to implement on hardware with minimal circuit elements.

Neuromodulation of a single neuron can be efficiently implemented using a nonlinear circuit model with mixed feedback paths (positive and negative), operating at different timescales [8]. This paper presents, for the first time to the best of our knowledge, the complete hardware implementation of a neural network consisting of the nonlinear neurons proposed in [8], which only showed SPICE simulations of a single neuron circuit. Section II provides a brief background of the nonlinear neuron model. Section III presents the bifurcation study of the software model. Section IV describes the chip architecture and neural network circuit design with tunable neurons and synapses. Section V discusses measured results of our test chip, and Section VI summarizes our contributions.

II. Background

The proposed nonlinear neuron model is made up of an excitable membrane that can be modeled as a highly parallel circuit, including a passive RC network in parallel with voltage-gated conductance channels inspired by the Na^+/K^+ ion channels commonly found in neurophysiology [5]. These conductance channels provide positive and negative feedback in addition to the applied external stimulus and passive leakage currents, thereby forming a mixed-feedback system. Moreover, different conductance elements in the model are tuned to operate at distinct timescales to achieve modular control of the excitability properties of the entire neuron.

The software model of a single neuron has 4 conductance elements, forming 2 positive and 2 negative feedback loops and operates at 3 different timescales – fast, slow, and

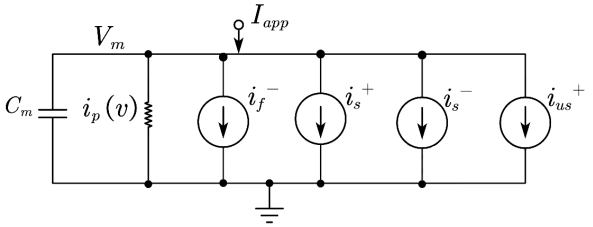

Fig. 1. Neuron circuit model. Adopted from [8]. Passive RC network in parallel with four conductance elements operating at three timescales. The nonlinear feedback currents are fast negative, slow positive, slow negative, and ultra-slow positive. I_{app} provides external input stimulus.

2023 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-6654-5109-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCAS46773.2023.10182201

ultraslow, as shown in Fig. 1. The fast negative conductance (i_f^-) operates in microsecond (μs) range, the slow positive (i_s^+) and slow negative (i_s^-) conductances operate in millisecond (ms) and ultra-slow positive conductance (i_{us}^+) operates in 100s of ms range. The four voltage-controlled conductance channels act as dependent current sources, I_x where x = f/s/us for fast, slow and ultra-slow timescales respectively. Each conductance element incorporates a hyperbolic tangent (tanh) transfer function, which combined together enables a complex nonlinear activation function for the neuron. The nonlinearity is particularly important at the fast and slow timescales, which can be seen as 'N-shaped' IV curve for the corresponding conductance channel. The nonlinear differential equations that describe the proposed neuron model are summarized as,

$$C_m \frac{dV_m}{dt} = I_{app} - I_p(V_m) - \sum I_x^{\pm} \quad (1)$$

$$\tau_x \frac{dV_x}{dt} = V_m - V_x \quad (2)$$

In equation (1), C_m is the membrane capacitance, I_{app} is the externally applied current, conductance channel currents $I_x^{\pm}$ are nonlinear functions of the channel voltages, V_x, given by $I_x^{\pm} = \pm F_x(V_x) = \pm \alpha_x^{\pm} tanh(V_x - \delta_x^{\pm})$ and I_p is the passive current linearly dependent on V_m, for example, $I_p = \kappa V_m$. In equation (2), τ_x is the time delay and V_x is the delayed output voltage of the conductance channel following V_m.

III. BIFURCATION STUDY OF THE SOFTWARE MODEL

The model proposed in [8] was simulated in software as a dynamical system for studying the transient behavior, phase portrait trajectories and bifurcation structures present in the model. The values for all parameters (α_x, δ_x, etc) were chosen carefully by tuning the fast, slow, and ultraslow I-V curves, based on the methods used in [8]. All quantities – voltages, currents and time are dimensionless. Neuron membrane voltage (V_m) and fast conductance voltage (V_f) are considered equal in the presented study, to reduce the 4D nonlinear dynamical system (V_m, V_f, V_s, V_{us}) to a 3D system aiding better visualization of the results. Bifurcation analysis was done using XPPAUT based on the methods described in [9].

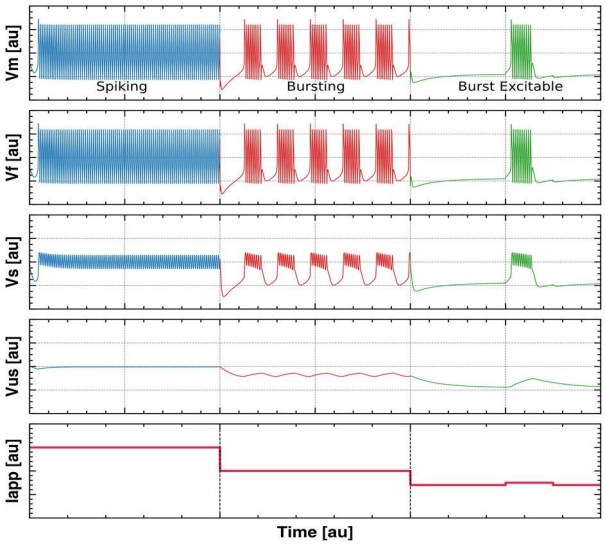

Fig. 2. Transient behavior of the neuron model for different modes of operation – spiking (blue), bursting (red), and burst excitable (green).

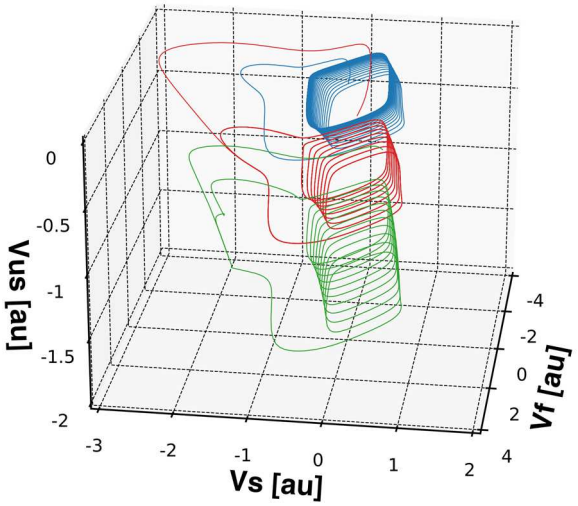

Fig. 3. Phase portrait of the neuron model as a 3D dynamical system with fast conductance voltage (V_f) directly following the membrane voltage (V_m). The trajectories of V_m are for different modes of operation – spiking (blue), bursting (red), and burst excitable (green).

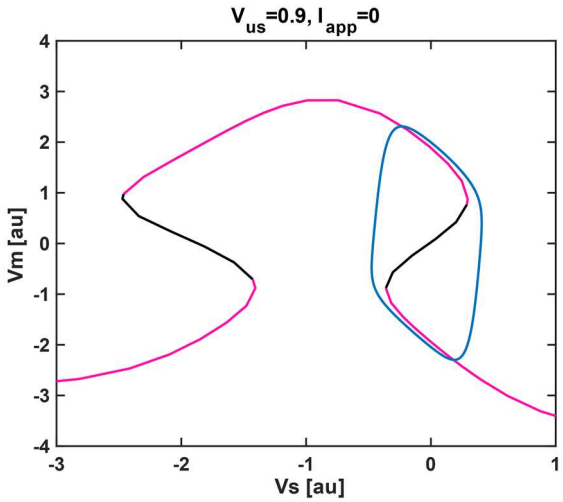

Fig. 4. Bifurcation structure for the fast (V_m, V_f) subsystem with V_s treated as a bifurcation parameter. **a.u.** stands for arbitrary units.

Fig. 2 shows the transient response of the neuron voltage terms, V_m, V_f, V_s, and V_{us}, demonstrating different modes of operation – spiking (blue), bursting (red), and burst excitable (green), simply by tuning the external stimulus current, Iapp. In addition, the plot emphasizes that the slow component is necessary to oscillate while the ultra-slow component is responsible for bursting behavior. Fig. 3 illustrates the trajectories of the system dynamics in 3D highlighting the transition from spiking to bursting to burst excitable behavior.

Fig. 4 shows the bifurcation structure for the fast (V_m, V_f) subsystem with V_s treated as a bifurcation parameter. Pink traces are stable critical points, while black traces denote unstable critical points. The blue trace is the projection of the oscillating trajectory in this fast-slow system, corresponding to a spike. Fig. 5 shows the bifurcation structure for the (V_m, V_f, V_s) system with V_{us} as a bifurcation parameter. Color codes are similar to Fig. 4 except that now purple traces denote maximum and minimum V_m values along unstable periodic orbits and green traces denote maximum and minimum V_m values along stable periodic orbits The trajectory observed (blue) is now a burst, not a relaxation oscillation. The bifurcation study of the system distinguishes the roles of

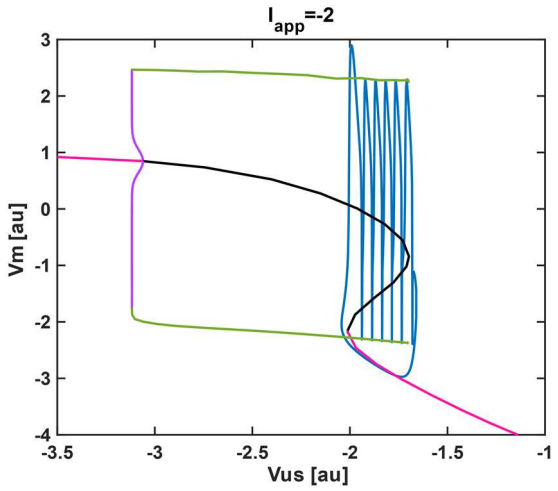

Fig. 5. Bifurcation structure for the (V_m, V_f, V_s) system with V_{us} taken as the bifurcation parameter.

different voltage terms in producing the solutions observed in the transient response simulations and phase portrait.

IV. HARDWARE IMPLEMENTATION

The main contribution of this work is to take the software model proposed in [8], validated with SPICE simulations in 0.35µm process and implement a translated model in TSMC 180nm CMOS process technology node. In addition to implementing single-neuron circuits, we have designed a test chip with multiple neurons forming multiple networks interconnected with tunable synapses in order to test and validate network-level behavior.

A. Chip Architecture

Fig. 6 provides the detailed architecture block diagram of the test chip. The chip consists of 4 neural networks, each consisting of 6 nonlinear neurons and 10 synapses implemented using operational amplifiers (op-amps). Within each neural network in the chip, 5 out of 6 neurons are fully interconnected to each other with synapses with one neuron isolated for single neuron testing, as shown in Fig. 6(a). The circuit model of the individual neuron is shown in Fig. 6(b), as described in the background section. The membrane voltage outputs of the six neurons in each of the 4 networks are sent through a 4:1 multiplexer (MUX). Two input digital select bits S0 and S1 select one of the 4 neural networks as the active network for testing. The output of 4:1 MUX is sent to voltage buffers that can drive large capacitive loads and enable accurate probing of the internal voltages. All neurons in the chip share common input bias voltages and bias currents, while the external stimulus current is set by two different voltages, v_{tune1} for the neurons N1, N2, N3 and v_{tune2} for the remaining 3 neurons N4, N5, and N6 of the 6-neuron network.

B. Chip Design

Fig. 6(b) shows the circuit schematic diagram of the single nonlinear neuron model and Fig. 7(c) shows the circuit topology of the individual conductance channel. In Fig. 6(b), the 20pF membrane capacitance of the neuron is realized using the M1 MOSCAP (Metal-Oxide-Semiconductor) since it has higher capacitance density and hence occupies a smaller area, compared to a MIMCAP (Metal-Insulator-Metal) available in the same technology. The external stimulus current, Iapp is provided using the two PMOS transistors M2 and M3 that is biased using another op-amp. The analog input voltage, v_{tune}, is used to tune the Iapp current. The equations for the dependent nonlinear current sources Ix, delay elements τ_x, and passive current Ip are given by equations (3) – (5).

$$I_x^{\pm} = \pm i_x^{\pm} \tanh\left(V_x - V_{\delta_x}^{\pm}\right) \text{ and } i_x^{\pm} = i_0 e^{\frac{V_{bx}}{nV_T}} \quad (3)$$

$$\tau_x = \frac{C_{T_x}}{G_{T_x}} \text{ and } G_{T_x} = \frac{I_{T_x}}{nV_T} = \frac{i_0 e^{\frac{V_{T_x}}{nV_T}}}{nV_T} \quad (4)$$

$$I_p(V_m) = G_m\left(V_m - V_{ref}\right) \text{ where } G_m = \frac{I_{bias}}{nV_T} = \frac{i_0 e^{\frac{V_b}{nV_T}}}{nV_T} \quad (5)$$

The passive resistance is realized using an op-amp transconductance, $G_m=1/R_p$. Equations for G_m and I_p, shown in equation (5), can be tuned using the analog bias voltages V_b and V_{ref} respectively. The current sources Ix are implemented using two op-amps and 1 MOSCAP (2pF) each, as shown in Fig. 6. The first op-amp acts as a voltage-follower with a delay that can be tuned using the bias voltage, V_{Tx}. The second op-amp provides the tanh transfer function as described by equation (3). The slope and offset of the tanh function can be tuned by bias voltages V_{bx} and $V_{\delta x}$ respectively. Positive and

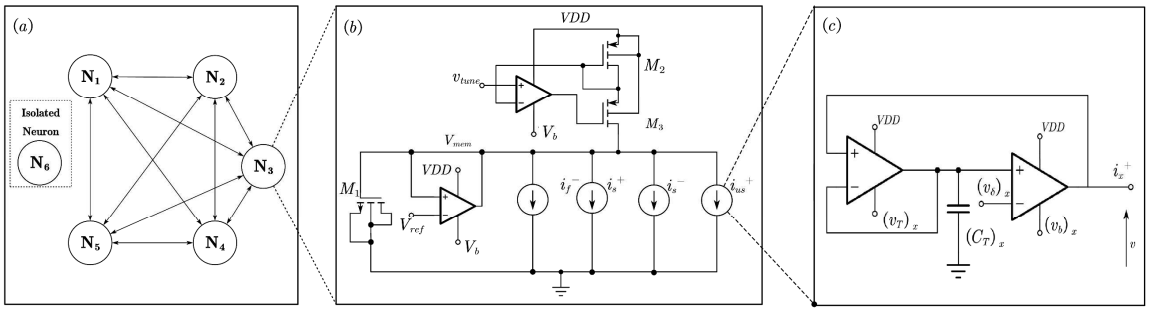

Fig. 6. Chip Architecture. (a) 6 neurons with 10 synapses formed 1 network. 5 neurons are connected with 10 opamp-based synapses. One neuron is isolated from the network for unit testing. (b) Neuron circuit schematic. Passive RC network in parallel with four conductance elements operating at three timescales. The nonlinear feedback currents are fast negative, slow positive, slow negative and ultra-slow positive. The amplifier on top with two PMOS is a current tuning structure to provide the external current, I_{app}. Another op-amp acts as tunable passive resistance. A MOSCAP with 20µm × 20µm NMOS provides 20pF membrane capacitance. (c) Circuit schematic of negative conductance channel. Adopted from [8]. First op-amp acts as a first-order tunable delay stage. Second op-amp provides e nonlinear tanh activation and decides the feedback type of the element. For the positive conductance element, input terminals of the second op-amp are swapped.

negative feedback of the dependent current source is obtained by flipping the positive and negative terminal of the second op-amp. All op-amps operate in the sub-threshold region of operation thereby enabling ultra-low-power consumption. The synapses connecting the neurons are implemented as a trans-linear device using the same op-amp used for the neuron.

V. Measurement Result

Fig. 7 shows the micrograph of the prototype chip fabricated on 180nm CMOS technology. The supply voltage of the chip is 3.3V. Area of each neuron is 105µm x 65µm and consumes an average power of 31.4µW and 23.1µW during spiking and bursting mode. The prototype chip occupies an area of 0.6mm x 2mm and consumes 753µW power.

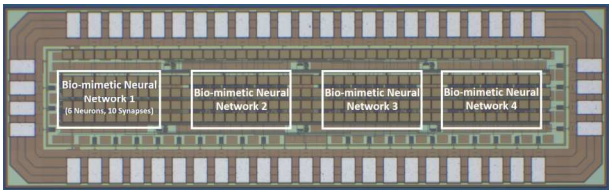

Fig. 7. Prototype chip die micrograph, implemented on TSMC 180nm technology. The total chip area is 0.6mm x 2mm

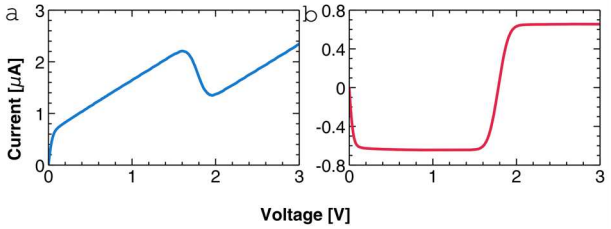

Fig. 8. IV curve of conductance element. (a) N-shape IV curve of conductance element in parallel with a passive resistive component. (b) Hyperbolic tangent (tanh) activation curve of conductance element.

TABLE I. ARCHITECTURE AND PERFORMANCE SUMMARY

Parameters	AdExp IF [10]	Braindrop [11]	NeuRRAM [12]	This work
Technology	22nm CMOS	28nm FDSOI	180nm CMOS	180nm CMOS
Supply Voltage (V)	0.8	1	1.8	3.3
Activation Function	Step and Sigmoid	Step	Step, Sigmoid, ReLU	Nonlinear Temporally Dynamic
Power/neuron µW	15.62	NA	0.55	31.4 (spike) 23.1 (burst)
Average Power (Total)	4mW	NA	140.6µW	753µW
Energy per spike	990fJ	380fJ	13.5fJ	179nJ(spike) 330nJ(burst)

Each neuron requires several bias voltages and currents as described in the previous sections. Op-amp design with MOS transistors in the sub-threshold region was successfully implemented to provide the tanh transfer function. Using 2 op-amps, the 'N shaped' IV curve of the fast timescale conductance channel was generated, as shown in Fig. 8. This confirms the nonlinear operating region of the dependent current source i_f.. Four conductance channels were connected together to simulate a single neuron. Tuning the gain of the slow conductance channels, the neuron was able to produce spiking and bursting rhythms at output membrane voltage, as shown in Fig. 9, which also shows the phase transition between spiking and bursting modes by controlling the external current, I_{app}. Fig. 10 demonstrates a 4-neuron network with inhibitory synapses to show coupled busting with increasing strength of synapses. Table I summarizes the results of our work compared to the state-of-the-art. The power and area per neuron of this work is comparable to

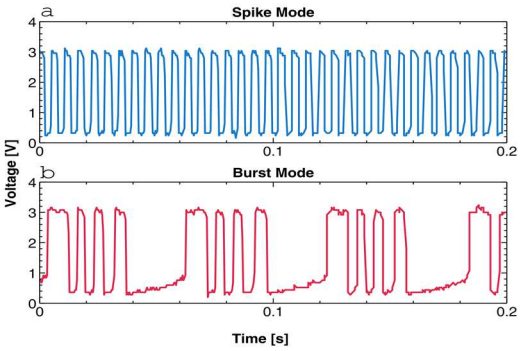

Fig. 9. Phase transition between spiking and bursting. (a) Neuron membrane voltage in spiking mode. Apply Iapp = 300nA. (b) Neuron

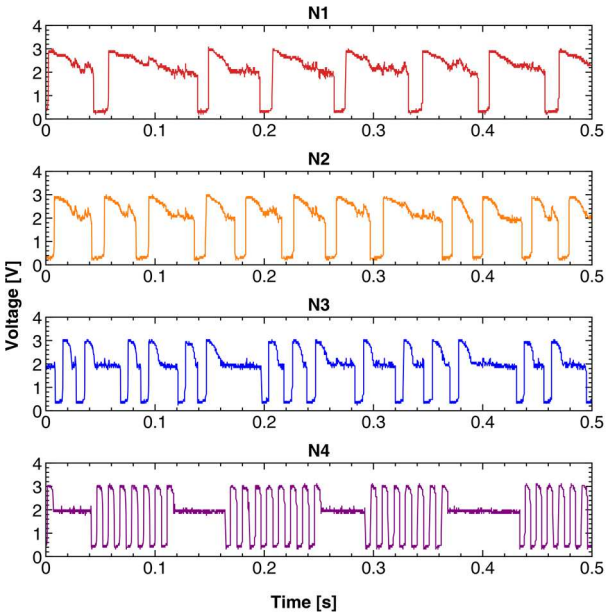

Fig. 10. Four neurons in bursting mode interconnected with inhibitory synapses form a network. N1 receives the highest inhibition from the other three neurons and N4 receives the minimum inhibition.

digital and mixed-signal implementations of state-of-the-art.

VI. Conclusion

This paper presented the hardware implementation of a nonlinear neuron model designed from a control theory perspective. The neuron is modeled as a mixed-feedback system operating at multiple timescales that can be tuned to produce spiking and bursting behavior. Bifurcation study of the neuron model as a 4D dynamical system describes the expected phase trajectories of the membrane voltage. A test chip consisting of 24 neurons and 40 synapses was designed and implemented on 180nm CMOS process. The neurons can exhibit rhythmic patterns found in tightly coupled neural networks in biology, such as the Central Pattern Generator (CPG), allowing neuromodulatory control on hardware at nodal and network levels. Recreating the CPG, crucial for movement in animals, can enable robust sensorimotor control for robotic locomotion. It is possible to scale the size and power of the neuron further by porting to a deep-submicron technology node, thereby enabling large-scale neuromorphic neural networks with nonlinear dynamics inspired by biology.

REFERENCES

[1] Buchanan, Bruce G. "A (very) brief history of artificial intelligence." Ai Magazine 26.4, 53-53, 2005.

[2] Szandała, Tomasz. "Review and comparison of commonly used activation functions for deep neural networks." Bio-inspired neurocomputing. Springer, Singapore, 203-224, 2021.

[3] R. Kubendran et al. "A 1.52 pJ/Spike Reconfigurable Multimodal Integrate-and-Fire Neuron Array Transceiver." International Conference on Neuromorphic Systems, 2020.

[4] Izhikevich, Eugene M. Dynamical systems in neuroscience. MIT press, 2007.

[5] Marder, Eve. "Neuromodulation of neuronal circuits: back to the future." Neuron 76.1, 1-11, 2012.

[6] Hodgkin, Alan L., and Andrew F. Huxley. "A quantitative description of membrane current and its application to conduction and excitation in nerve." The Journal of physiology 117.4, 500, 1952.

[7] Izhikevich, Eugene M. "Simple model of spiking neurons." IEEE Transactions on neural networks 14.6, 1569-1572, 2003.

[8] Ribar, Luka, and Rodolphe Sepulchre. "Neuromodulation of neuromorphic circuits." IEEE Transactions on Circuits and Systems I: Regular Papers 66.8, 3028-3040, 2019.

[9] Ermentrout, B., “Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students”, Society for Industrial and Applied Mathematics (SIAM), 2002.

[10] A. Rubino, M. Payvand and G. Indiveri, "Ultra-Low Power Silicon Neuron Circuit for Extreme-Edge Neuromorphic Intelligence," 2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Genoa, Italy, 2019.

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

[12] Wan, W., Kubendran, R., Schaefer, C. et al., “A compute-in-memory chip based on resistive random-access memory.” Nature 608, 504–512, 2022.