Dendritic Computation through Exploiting Resistive Memories as both Delays and Weights

Melika Payvand*[1], Simone D'Agostino*[1,2], Filippo Moro*[2], Yigit Demirag[1], Giacomo Indiveri[1], Elisa Vianello[2]

[1]*Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland*

[2]*CEA-Leti, Université Grenoble Alpes, F-38000, Grenoble, France*

Summary. The dendritic branches of biological neurons perform complex spatio-temporal feature detection of the streaming input they receive in the form of spikes. This is done through detecting the spatial and temporal locality of the signals, through using both spatial parameters, weight, and temporal parameters, delays. Inspired by this mechanism, we propose a neuromorphic hardware architecture equipped with multi-scale dendrites, each of which has synapses with tunable weight and delay elements. Weights and delays are both implemented using Resistive Random Access Memories (RRAM). We exploit the variability in the high resistance state of RRAM to implement a distribution of delays for enabling the real-time processing of sensory signals. In particular, we perform RRAM-aware simulations on heartbeat anomaly detection tasks, and show that by incorporating delays directly into the network, the power and memory footprint can be reduced by up to 100x and 10x respectively, compared to the state-of-the-art AI hardware.

In the most commonly used models of artificial neural networks, a neuron's output is a nonlinear transformation of the sum of the product of its input and its weight parameters, known as the point-neuron model. Although for static rate-based information encoding, point-neuron models have enough complexity to perform computation, they are not tailored for detecting the temporal aspects of dynamic input patterns. Neuroscience findings show that the dendritic arbor of a neuron implements non-linear integration and decodes spatio-temporal locality of arriving events, a mechanism known as coincidence detection (CD) [1]. Inspired by this mechanism, we propose a neuromorphic hardware architecture equipped with multi-scale dendrites, each of which has synapses with tunable RRAM-based weight and delay elements (Fig. 1). The delay is implemented using the RRAM-C combination, while the weight is represented by one RRAM device. Training this architecture can find a set of values for RRAMs, such that when a temporal feature is present in the sensory signal, the delayed spikes align, resulting in a spike produced by the neuron, acting as a CD.

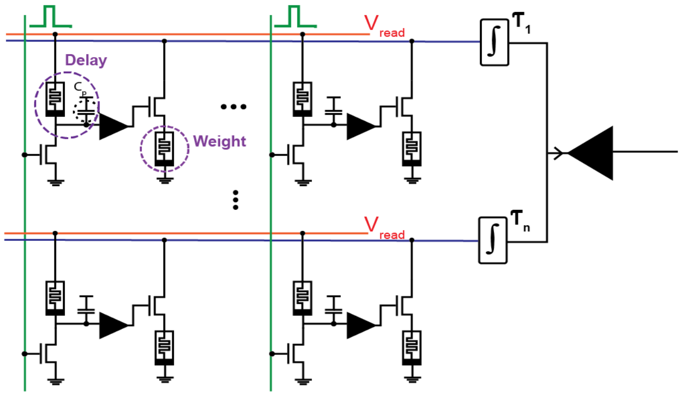

Figure 1. Dendritic architecture using hybrid CMOS and resistive memory technologies.

To enable real-time sensory processing, the delayed elements should be in the range of the time constant of the sensed real-world signals, e.g., in the order of 10s-100s of ms. Thus, to implement such delays on-chip, while reducing the capacitor size, we propose to exploit the High Resistive State (HRS) of RRAMs. Since the conductive filament resulting in resistive switching is very weak in the HRS, controlling the precise value of the resistance of RRAM in the HRS is difficult. This can be seen in the HRS measurements shown in Fig. 2, with large variability in the HRS following a log-normal distribution. The mean of this distribution is a function of the reset voltage with which the device is switched to the HRS. Thus, we propose to use the reset voltage as a knob that determines the order of magnitude of the delays in each compartment. Due to variability, using the same reset voltage to reset the delay devices of each compartment results in samples from the corresponding log-normal distribution. The network objective is then to learn the correct weights corresponding to each delay, such that the neuron performs CD, detecting the temporal features of the signal.

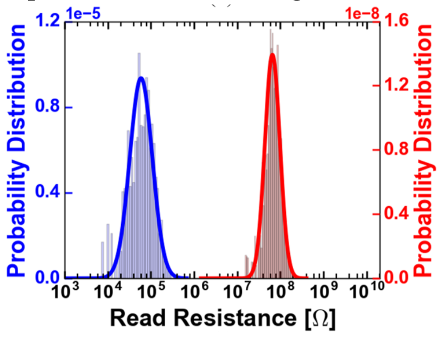

Figure 2. Variability of RRAMs in their HRS follows a wide log-normal distribution. The shift in the distribution is as a result of different reset voltages.

We have benchmarked our architecture on a real-time sensory processing task, namely heartbeat anomaly detection, using MIT-BIH dataset [2]. The dendritic architecture with a single neuron achieves 95% accuracy using 10x less memory footprint compared to the number of resources required in a Spiking Recurrent Neural Network (SRNN) for the same accuracy. Moreover, we have compared the estimated power consumption of our approach against other state-of-the-art methods, illustrated in Table 1. It shows that our approach's power savings is up to 100x, thanks to the passive delay elements, which are keeping the information in time, without having to actively keep the memory.
In conclusion, we have introduced an RRAM-aware dendritic architecture, which is empowered by delays, and as a result can introduce temporal richness to a feedforward network which can classify a sensory processing task with 100x less power consumption and 10x savings in memory footprint compared to the state-of-the art AI hardwares.

Table 1. Power consumption comparison of our architecture compared to other works on ECG anomaly detection task.

Work	This Work	[3]	[4]	[5] SNN
Accuracy	95.94%	99.3%	92.11%	79%
Power	0.47µW	48.6µW	516.1µW	64mW
Energy per class.	88nJ	2.25µJ	258.08µJ	32mJ

References

[1] Paugam-Moisy et al., “Computing with Spiking Neuron Networks”, Handbook of Natural Computing, 2012.

[2] Moody, G.B. and Mark, R.G., 2001. The impact of the MIT-BIH arrhythmia database. *IEEE engineering in medicine and biology magazine*, *20*(3), pp.45-50.

[3] J. Liu et al. “4.5 BioAIP: A Reconfigurable Biomedical AI Processor with Adaptive Learning for Versatile Intelligent Health Monitoring”. IEEE International Solid- State Circuits Conference (ISSCC). Vol. 64. 2021, pp. 62–64. [4] Bauer et al. “Real-Time Ultra-Low Power ECG Anomaly Detection Using an Event-Driven Neuromorphic Processor”. In: IEEE Transactions on Biomedical Circuits and Systems 13.6 (2019), pp. 1575–1582.

[5] Z. Yan, J. Zhou, and W.F. Wong. “Energy efficient ECG classification with spiking neural network”. In: Biomedical Signal Processing and Control 63 (2021).