A mixed-signal spatio-temporal signal classifier for on-sensor spike sorting.

Germain Haessig*, Daniel Garcia Lesta†§, Gregor Lenz‡, Ryad Benosman‡ and Piotr Dudek‡§
*Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland. Email: germain@ini.uzh.ch
†Centro Singular de Investigación en Tecnoloxías Intelixentes, Santiago de Compostela, Spain.
‡Institut de la Vision, Sorbonne Université, Paris, France
§The University of Manchester, Department of Electrical and Electronic Engineering, Manchester, UK

***Abstract*—Neuromorphic systems provide an alternative to conventional computing hardware, promising low-power operation suitable for sensory-processing and edge computing. In this paper, we present a mixed-signal processing system designed to provide on-sensor classification of signals obtained from multi-electrode array neural recordings. The designed circuits implement a real-time spike sorting algorithm, and operate on signals represented by asynchronous event streams. We combine analog circuits computation primitives (temporal surface generation, distance computation, winner-take-all) to implement a spatio-temporal clustering algorithm, classifying signals acquired by neighbouring electrodes. The prototype chip has been submitted for fabrication in a 180nm CMOS technology. The circuits are designed to fit, alongside signal conditioning and conversion circuits, in the area under the recording electrodes (below 80x80um per electrode). Circuit implementation details and simulation results are presented. The expected neural spike recognition rates of 75% in a single-layer network and 88% in a 2-layer network are comparable with a software implementation, while the system is designed to provide a low-power embedded real-time solution. This work provides a foundation towards the design of a large scale neuromorphic processing system, to be embedded in brain-machine interfaces.**

I. Introduction

Real time neural activity decoding is essential for brain-machine interfaces (e.g. for prosthetics) and to enable closed-loop experiments in neuroscience. Prior work [1]–[3] has shown that neural activity in the human brain can be decoded in real-time from a Multi Electrode Array (MEA), after a daily recalibration of the system. These systems usually require a wired connection, as it is a challenge for a wireless system to deal with the amount of recorded data while keeping heat dissipation to a required minimum. Spike sorting is a fundamental pre-processing task that allows to distinguish the activity of one or more neurons that have been recorded with the same electrode (or, as in this work, by a set of adjacent electrodes). This is achieved through the classification of the shapes of the recorded neural activation pulses (spikes). When done near sensor, this has the potential to improve system latency, and significantly reduce data rates, enabling wireless systems. Several hardware approaches to spike sorting have been presented [3]–[7].

Conventional signal processing systems use the Nyquist-rate to sample and quantise signals. Introducing a different data representation can lead to significant gains in performance and power efficiency. It has been shown that an event-based approach can offer advantages in spatio-temporal pattern recognition tasks [8]–[12]. The availability of neuromorphic sensors such as silicon retinas [13], [14] and silicon cochleas [15], [16] will lead to further development of algorithms handling these event-based signal representations. Instead of the classic scheme of periodic sampling, these neuromorphic sensors only transmit data whenever there is a significant change in the signal, leading to a sparse representation and providing high temporal precision at low data bandwidth. An event-based neural recording platform has been recently introduced [17], and will be used in this work as a source of neural recording data. We present a 180nm CMOS implementation of a spike sorting algorithm presented in [18]. Our ultimate goal is to integrate the presented circuitry into the recording electrode array. Section II details the topology and behavior of the proposed system, Section III provides circuit descriptions and Section IV shows simulation results that are compared to the expected behavior of the original algorithm.

II. Concept

A. Towards an event-based representation

The event-based approach asynchronously transmits *events* for a pre-set change in the signal level (see Figure 1) similar to Lebesgue sampling [19]. For a signal coming from a Multi-Electrode Array, we can define an event $ev_i = (t_i, \mathbf{x}_i, p_i)$ as a tuple of a time of appearance t, a spatial position in the array $\mathbf{x}_i$ and a polarity p, indicating the direction of the change. $p = 1$ (ON event) indicates that the signal increased, whereas $p = 0$ (OFF event) indicates that the signal decreased.

B. Original algorithm

Using the precise timing of the event-based representation of the input signal, we can introduce the spatio-temporal context S^i of the i-th event ev_i, representing the past activity on a given surrounding, centered around the incoming event. This context is based on the work presented in [11] and [10], and built by sampling event traces generated through exponential decays:

$$S^i_j = \exp\left(-\frac{t_i - t_j}{\tau}\right) \; for \;\; |\mathbf{x}_i - \mathbf{x}_j| \leq r \qquad (1)$$

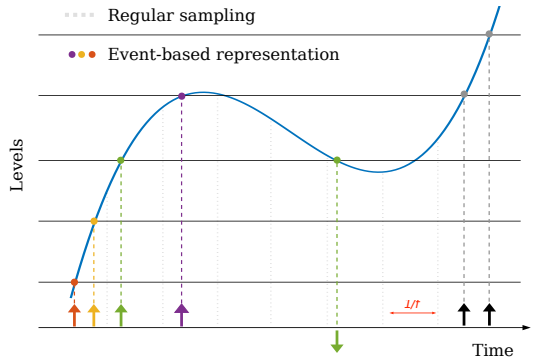

Fig. 1: Event generation. Each time the signal changes by a sufficient amount w.r.t. to the signal at the time of the last event (arrow), a new event is generated. The direction of this change is called polarity and is represented in the orientation of the arrow (up/down). For comparison, the grey dotted lines represent a standard sampling rate at fixed frequency.

where t_j is the timestamp of the last event (ev_j) at the given $\mathbf{x}_j$ position, r the surrounding size (here, $r = 1$, meaning a radius of 1 in a hexagonal grid, *i.e.* 6 neighbours), and τ the time constant of the decaying unit.

Afterwards, this spatio-temporal context is compared to learned templates, in order to find the closest one. A classification unit can be used in order to assign the corresponding class to the input spiking pattern. The processing scheme is illustrated in Figure 2.

C. Constraints

The recording array (32×32 electrodes) has dimensions of $3.4 \times 2.8mm$. Each single electrode occupies an area of $96 \times 79\mu m^2$ [17]. As we aim to implement our circuitry directly below the electrode, we have to keep the are of our circuitry below $0.007mm^2$. The power consumption should be as low as possible, not only to allow for efficient wireless systems but also to keep tissue damage caused by heat dissipation to a minimum. The recording array consumes $145\mu W$ [17]). Given the characteristics of biological signals, the decay time constant for the event trace should be tunable from a few μs up to ms.

III. Circuit description

Our chip needs three main computational blocks: an exponential decay unit to form the event traces composing the spatio-temporal contexts; a comparison unit to compute a distance between the presented context and learned templates and a unit that selects the template that provides the closest match. The circuits implementing these blocks are presented in this section, followed by the overall architecture of our chip.

A. Exponential decay unit for the event trace

The time constant of the decaying unit has to range from μs to ms. A straightforward implementation of a resistor-capacitor (RC) circuit can implement exponential voltage decay, but in order to provide a tuneable time constant and a small circuit area, we implement this with a switched capacitor circuit, controlled by an external clock source. Figure 3a shows the basic circuit. Making the two switches (P_1 and P_2) for this switched capacitance large enough, the intrinsic drain/source capacitances are sufficiently large to avoid the need of an external capacitance (see C_{ds} in Figure 3a). The non-linearity is not a problem, as it simply modifies the overall distance function of the comparison unit (see Section IV-B). The incoming spike V_{spk} resets the capacitance C to V_1 (via N_1), ensuring a quick discharge ($\sim 40ns$), that is negligible given the considered time range. Then, the capacitance charges towards V_2, and the voltage V_{trace} is fed to the next module, for comparison to the template value. Here, the capacitance C has a size of $10 \times 10\mu m$, for a value of $200fF$, while C_{ds} is approximately $24\times$ smaller. Variations of the control clock frequency lead to change in the time constant as shown in Table I.

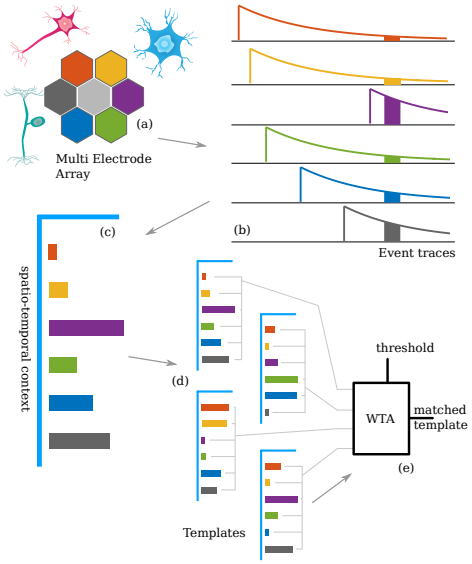

Fig. 2: Functionality of the proposed chip. (a) The hexagonal array records signals from neurons, generating events on multiple channels. (b) Each time an event is triggered, an event trace is generated. When the central event occurs (light gray electrode), the value of the traces on neighbouring channels is sampled and memorised, forming a spatio-temporal context (c). For the sake of understanding, only ON events are represented here. (d) This context is then compared to 4 stored templates. (e) The currents resulting from these comparisons are passed to a Winner-Takes-All block, in order to determine which template is the closest to the current context.

Frequency	τ
5 kHz	4.3 ms
500 kHz	76 μs
f	24/f

TABLE I: Exponential decay time constant τ for the event trace, versus clock frequency, for the switched-capacitor circuitry. Extracted from post layout simulations.

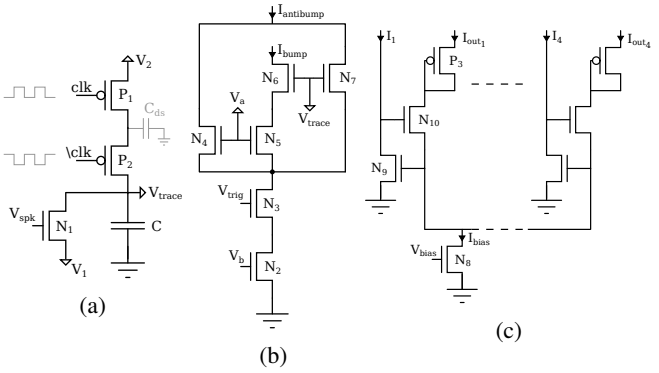

Fig. 3: Basic cells schematics: (a) Event trace circuitry; (b) Comparison circuitry [20]; (c) Basic WTA circuit [21], two inputs are shown. See text for details.

B. Comparison Unit

When the central event occurs, the comparison of each event trace value to an external template value is triggered. The difference between the template value V_a and the voltage V_{trace} is obtained by a bump-antibump circuit [20], shown in Figure 4, outputting the bump-current I_{bump} defined as:

$$I_{bump} = \frac{I_b}{1 + \frac{4}{S}\cosh^2\frac{\kappa\Delta V}{2}} \tag{2}$$

where I_b is the bias current controlled by the V_b, S is the ratio between transistor sizes of (N_4, N_7) and (N_5, N_6), κ the transconductance of the transistors and $\Delta V = V_a - V_{trace}$ the voltage difference.

C. Template matching

The currents representing the differences between event traces and template values obtained as described above, are summed over the entire template (6 currents summed in a node). This provides the overall template matching current for each of the templates. A set of 4 templates is considered sufficient to achieve desired recognition rates in our application (according to [18]). For each neighboring pixel (6 in the hexagonal case), we then need 4 comparison units, giving 32 comparison units in total. Each one of these units needs an analog reference value, which is provided externally in the current design.

D. Winner take all

The summed currents of comparison circuits represent distances between the current spatio-temporal context and the templates. The selection of the best matching template is done by a Winner-Take-All (WTA) circuit. A simple yet efficient implementation of a current-mode WTA circuit was brought forward by Lazzaro et al. [21], using only two transistors per input channel. This design was chosen to minimize the delay up to which the closest matching current is picked, as well as space efficacy. However this also makes the circuit more susceptible to mismatch. As can be seen in Figure 3c, all 4 input cells are connected to a global node that provides a bias current (N_8). This bias current is controlled by an externally provided voltage V_{bias}. The output of the comparison units for the 4 templates (I_1 to I_4) are fed into the WTA circuitry, in order to select the one with the highest current. The WTA will set the corresponding output I_{out1} to I_{out4} to $\approx I_{bias}$, all other outputs will be suppressed (unless the multiple inputs are very closely matched). The output currents are binarised, to provide an indicator of the winning template (in a multi-layer version of the classifier, this would generate an event on the template channel). We also include a circuit to compare the winning current magnitude with a threshold current, to provide a confidence bit indicating that the winner is valid. We do this in order to reject winners that are themselves poor matches.

E. Implementation

All the above described blocks are assembled in our core, as shown in Figure 6. The aim of this first prototype is to validate the principle and quantify the effects of noise and variability. We replicate the core block many times, multiplexing different intermediate signals to the bonding pads, in order to be able to carry out comprehensive measurements. The core comprises 64 blocks with digital outputs (4 binary WTA outputs + 1 valid bit, 64 blocks with an analog output (current output of the distance to each template), 64 blocks with access to the decaying unit voltage, and 1 full digital block. All the blocks except the last one are multiplexed to the pads via three different 64:1 multiplexers. The system uses 4 templates of 6 values each. Each one of these is an analog reference value. For now, to limit the chip's complexity (at the cost of an increased number of pads), these 32 analog values are fed from an external source. In the final implementation we will use SRAM memory and 32 DACs on-chip such as the one presented in [22]. An alternative that we are exploring in this project is to use non-volatile analog memory devices (memristors) integrated with the CMOS process [23]. Memristors could be then also used to provide programmable decay in a modified event-trace unit circuit.

IV. Simulations

A. Benchmark for performance evaluation

In order to quantify the performances of our implementation, we used artificially generated Multi-Electro-Array (MEA) recordings [24]. These generated signals where filtered accordingly to the amplification stages of the recording unit [17] and then converted to events, as shown in Figure 5.

B. Event trace and distance

Figure 4 presents the simulation of the realized event trace, with its associated template value, and the the output of the distance circuit. The exponential time constant is set to $\tau = 1$ ms, and the template value is 1.1 V, which corresponds to a peak response at around 1.2 ms. The implemented distance function is much sharper than traditional L_1 and L_2 norms used in the original algorithm [18], but is proven to perform well in our target application (see next section).

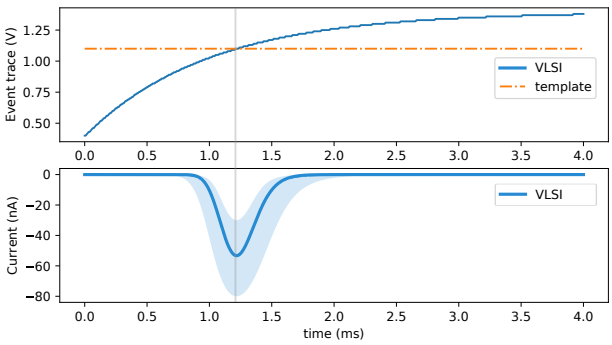

Fig. 4: Circuit simulation results: (top) event trace generated by the decay unit; (bottom) distance between the trace and the fixed template value. Blue line is the simulated distance, the blue shadowed area shows the impact of the fabrication mismatch (Monte Carlo simulation, for the bump circuit only. The light blue area represents the min/max values for a set of 100 samples) on the obtained distance. Average trace differs less than 0.1% from Equation 2.

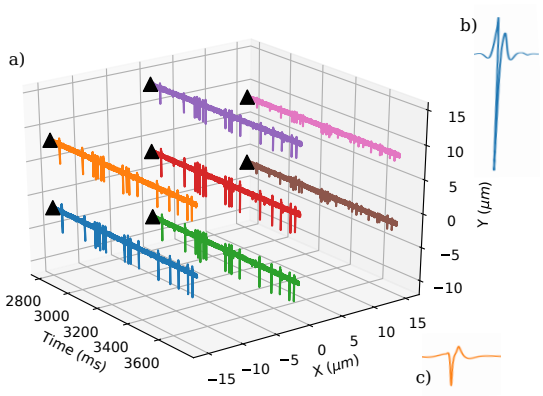

Fig. 5: a) Artificially generated recordings [24] for the probe topology. Here, an hexagonal probe is used. We only show 7 electrodes, in the same configuration as for our chip [25] [26]. Ground truth is available for classification performance estimation. b) and c) 2 different spike shapes extracted from the dataset.

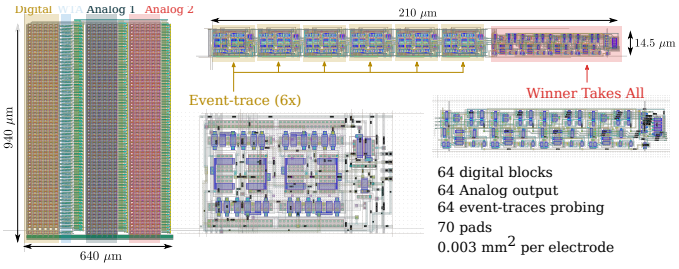

Fig. 6: Chip layout. The core is composed of 3×64 basic blocks, with each block comprising 6 decaying units, 4 template matching circuits and a WTA circuit. The core occupies a space of $640 \times 940 \mu m$, each individual block being $14.5 \times 210 \mu m$.

Distance	Recognition rate	
	1 layer	2 layers
$\mathcal{L}_1$	60%	68%
$\mathcal{L}_2$	73%	82%
Bhattacharyya [27]	78%	89%
This work	75%	88%

TABLE II: Recognition rate for different distance metrics, and our implemented model. We can notice that the bump distance performs almost as good as the Bhattacharyya [27], which is significantly more complex to implement on a chip due to more complex computation (square root, logarithm).

C. Classification rate

The data set was split into one training set containing 1370 spikes and a testing set containing 930 spikes. Training and testing is done using the distance behavior as extracted from post-layout simulations, including variability. Classification scores are given in Table II. We can expect this chip to behave almost as well as the method introduced in the original paper, with computational time being highly reduced. This is especially true for hierarchical structures that cause algorithmic complexity to significantly increase due to the additional number of templates [11]. In our chip, the *computation time* grows linearly with the number of layers. Our simulations achieve a score of 75% on a single layer and 88% on a 2-layer architecture, with computation time of $80/160ns$ respectively. These scores are average over 100 samples of a Monte-Carlo design simulation including device mismatch. The computation time corresponds to the propagation time in the digital circuitry and is substance to further optimization. For the original algorithm, computation time is around $5\mu s$ for a single layer architecture, using Python code executed on a Core i7-8700K @ 3.7 GHz computer.

V. Conclusion

The presented system is a first step towards a fully neuromorphic signal processing pipeline for neural decoding applications. It implements the essential primitive computational blocks that will be embedded below each pixel of our recording array. Due to their event output, these computational blocks can be chained to form a hierarchical processing pipeline in more complex processing scenarios than considered in this paper. All the results were obtained via simulations (including post-layout and Monte Carlo), further work will be to fully characterize the fabricated chip, specifically to quantify the parameter variations due to fabrication mismatch and analyze the impact of those on the classification result. Our aim is to design a fully functional integrated Micro-Electrode Array system, with on-chip spike sorting. We are also working on embedding memristive memories for parameter configuration. We anticipate this work to push the boundaries of state-of-the-art low-power embeddable brain-machine interfaces.

Acknowledgment

This work was partially supported by the Swiss National Science Foundation Sinergia project #CRSII5-18O316 and UK EPSRC Grant EP/R024642/1 (FORTE).

REFERENCES

[1] B Wodlinger, JE Downey, EC Tyler-Kabara, AB Schwartz, ML Boninger, and JL Collinger. Ten-dimensional anthropomorphic arm control in a human brain- machine interface: difficulties, solutions, and limitations. *Journal of neural engineering*, 12(1):016011, 2014.

[2] Masoud Rezaei, Esmaeel Maghsoudloo, Mohamad Sawan, and Benoit Gosselin. A 110-nw in-channel sigma-delta converter for large-scale neural recording implants. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5741–5744. IEEE, 2016.

[3] David A Borton, Ming Yin, Juan Aceros, and Arto Nurmikko. An implantable wireless neural interface for recording cortical circuit dynamics in moving primates. *Journal of neural engineering*, 10(2):026010, 2013.

[4] Thilo Werner, Elisa Vianello, Olivier Bichler, Daniele Garbin, Daniel Cattaert, Blaise Yvert, Barbara De Salvo, and Luca Perniola. Spiking neural networks based on oxram synapses for real-time unsupervised spike sorting. *Frontiers in Neuroscience*, 10:474, 2016.

[5] Sivylla E Paraskevopoulou, Deren Y Barsakcioglu, Mohammed R Saberi, Amir Eftekhar, and Timothy G Constandinou. Feature extraction using first and second derivative extrema (fsde) for real-time and hardware-efficient spike sorting. *Journal of neuroscience methods*, 215(1):29–37, 2013.

[6] Ian Williams, Song Luan, Andrew Jackson, and Timothy G Constandinou. Live demonstration: A scalable 32-channel neural recording and real-time fpga based spike sorting system. In *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–5. IEEE, 2015.

[7] Nur Ahmadi, Matthew L Cavuto, Peilong Feng, Lieuwe B Leene, Michal Maslik, Federico Mazza, Oscar Savolainen, Katarzyna M Szostak, Christos-Savvas Bouganis, Jinendra Ekanayake, et al. Towards a distributed, chronically-implantable neural interface. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 719–724. IEEE, 2019.

[8] Thusitha N Chandrapala and Bertram E Shi. The generative adaptive subspace self-organizing map. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 3790–3797. IEEE, 2014.

[9] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. *arXiv preprint arXiv:1803.07913*, 2018.

[10] Gregory K Cohen, Garrick Orchard, Sio-Hoi Leng, Jonathan Tapson, Ryad B Benosman, and André Van Schaik. Skimming digits: neuromorphic classification of spike-encoded images. *Frontiers in neuroscience*, 10:184, 2016.

[11] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.

[12] Saeed Afshar, Libin George, Jonathan Tapson, André van Schaik, and Tara J Hamilton. Racing to learn: statistical inference and learning in a single spiking neuron with adaptive kernels. *Frontiers in neuroscience*, 8:377, 2014.

[13] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *Solid-State Circuits, IEEE Journal of*, 46(1):259–275, 2011.

[14] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240$\times$ 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[15] Vincent Chan, Shih-Chii Liu, and Andr van Schaik. Aer ear: A matched silicon cochlea pair with address event representation interface. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(1):48–59, 2007.

[16] Shih-Chii Liu, André Van Schaik, Bradley A Mincti, and Tobi Delbruck. Event-based 64-channel binaural silicon cochlea with q enhancement mechanisms. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 2027–2030. IEEE, 2010.

[17] Federico Corradi and Giacomo Indiveri. A neuromorphic event-based neural recording system for smart brain-machine-interfaces. *IEEE transactions on biomedical circuits and systems*, 9(5):699–709, 2015.

[18] Germain Haessig, Kevin Gehere, and Ryad Benosman. Spikes decoding spikes : A neuromorphic event-driven framework for real-time unsupervised spike sorting. *in Press*, 2019.

[19] Karl Johan Astrom and Bo M Bernhardsson. Comparison of riemann and lebesgue sampling for first order stochastic systems. In *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, volume 2, pages 2011–2016. IEEE, 2002.

[20] Tobi Delbruck. 'bump'circuits for computing similarity and dissimilarity of analog voltages. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 1, pages 475–479. IEEE, 1991.

[21] John Lazzaro, Sylvie Ryckebusch, Misha Anne Mahowald, and Caver A Mead. Winner-take-all networks of o (n) complexity. In *Advances in neural information processing systems*, pages 703–711, 1989.

[22] Tobi Delbruck, Raphael Berner, Patrick Lichtsteiner, and Carlos Dualibe. 32-bit configurable bias current generator with sub-off-current capability. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 1647–1650. IEEE, 2010.

[23] Spyros Stathopoulos, Ali Khiat, Maria Trapatseli, Simone Cortese, Alexantrou Serb, Ilia Valov, and Themis Prodromakis. Multibit memory operation of metal-oxide bi-layer memristors. *Scientific reports*, 7(1):1–7, 2017.

[24] Alessio P Buccino and Gaute T Einevoll. Mearec: a fast and customizable testbench simulator for ground-truth extracellular spiking activity. *bioRxiv*, page 691642, 2019.

[25] Joana P Neto, Gonçalo Lopes, João Frazão, Joana Nogueira, Pedro Lacerda, Pedro Baião, Arno Aarts, Alexandru Andrei, Silke Musa, Elvira Fortunato, et al. Validating silicon polytrodes with paired juxtacellular recordings: method and dataset. *Journal of neurophysiology*, 116(2):892–903, 2016.

[26] Richárd Fiáth, Bogdan Cristian Raducanu, Silke Musa, Alexandru Andrei, Carolina Mora Lopez, Chris van Hoof, Patrick Ruther, Arno Aarts, Domonkos Horváth, and István Ulbert. A silicon-based neural probe with densely-packed low-impedance titanium nitride microelectrodes for ultrahigh-resolution in vivo recordings. *Biosensors and Bioelectronics*, 106:86–92, 2018.

[27] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.