# A memristor-CMOS hybrid architecture concept for on-line template matching

Alexander Serb$^*$, Christos Papavassiliou$^\dagger$, Themistoklis Prodromakis$^*$
$^*$Electronics and Computer Science Department, University of Southampton, SO17 1BJ, UK.
$^*$Electrical and Electronic Engineering Department, Imperial College, London, SW7 2AZ, UK.

Email: {A.Serb, T.Prodromakis}@soton.ac.uk, {c.papavas}@imperial.ac.uk

***Abstract*—The ability to identify (detect) and categorise (sort) neural spikes in real-time and under highly restrictive power/area budgets is a major enabling technology towards the development of intelligent implantable systems. In this work we propose a memristor-CMOS hybrid architecture concept that relies on a 'template pixel' (texel) circuit combining CMOS and memristive devices to perform on-line spike sorting through template matching. We show through simulation how the texel is capable of comparing an input voltage against a stored (in the memristors) value and converting the degree of matching between input and stored pattern into a current. We further illustrate the fundamental texel design space that includes tuning it to a different preferred input voltage and controlling the sharpness of the tuning. Finally, we estimate that even in an unoptimised technology and design a texel array capable of recognising three different 10-point patterns will consume a very promising maximum of $3.15\,\mu W$ for a footprint of approx. $500\,\mu m^2$.**

***Index Terms*—CMOS, memristors, spike sorting**

## I. Introduction

A key component of the global effort to understand the functioning of the human brain pertains to the development of brain-machine interfaces capable of recording neuronal activity in-vivo; itself a whole research area progressing under its own version of Moore's law [1]. Typically, large-scale neural activity monitoring is achieved through implantable systems that consist of three broad blocks: a) an electrode array [2], b) an Analogue Front-End (AFE) block usually consisting of amplification, filtering and digitisation [3], c) a signal processing block (Back-End) that may either concentrate on performing single-neuron activity detection (spike detection [4] or sorting [5]) or include Local Field Potential (LFP) extraction [6] and d) telemetry for transmitting the extracted data to the external receiver (Fig. 1).

In this work we concentrate on the back-end block and specifically single-neuron activity detection. At the algorithmic level this is currently achieved through a multitude of approaches such as threshold detection [7], non-linear energy operator [8], template matching [9] and others; each offering a different solution on the implementation complexity vs. accuracy trade-off space [10]. What all these methods share in common, however, is the objective of compressing a high data-rate voltage-time series signal arriving from the AFE block into a low data-rate/high-information output signal encoding the timing of neuronal action potentials (spikes) only, while suppressing noise (fig. 1).

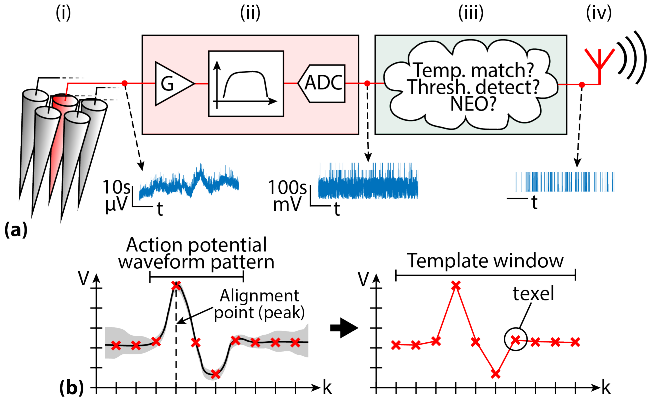

Fig. 1. Processing neural data. (a) Typical example of single-unit neural activity monitoring system channel architecture. Shown are: (i) the electrodes, (ii) the analogue front end, (iii) the signal processing (back-end) block, (iv) the telemetry and antenna module. Below, the progressive transformation of the raw waveform into a low bandwidth raster plot is shown as the signal propagates through the system. (b) Template matching basics: The noisy action potentials identified as originating from the same neuron are aligned (e.g. peak-aligned) and aggregated (gray shade), an average waveform is extracted (black trace - left panel) and the corresponding template is built from its samples (right panel). Notably, the template may not last exactly as long as the stereotyped action potential waveform.

At the implementation level on-chip spike detectors and sorters have to: a) operate using minimal area and power budgets, b) achieve maximum data compression to ease the power budget of the telemetry block and c) avoid placing excessively area/power expensive signal preconditioning requirements on the AFE block. Whilst architectures utilising fully digital [11] and mainly analogue [5] techniques seeking to address these requirements have already been proposed, more recently we have demonstrated a memristor-based methodology for spike detection and rudimentary sorting [12]. Memristors are electronic components that change their resistive state when appropriate voltages are applied across their terminals and can be practically implemented using a variety of technologies [13]. This fundamental property allows memristors to act as thresholded integrators, an ability that the proposed memristor-based system exploits in order to perform threshold detection of action potentials in a single component.

In this work we build on [12] by presenting a memristor-CMOS hybrid circuit for performing non-rudimentary on-line spike sorting via template matching. The CMOS component performs the template matching in the analogue domain whilst the memristors are used to store the template library in an area-efficient way (leveraging their back-end-of-line integrability). The rest of the paper is organised as follows: In section II the operating principles and architecture of the proposed circuit are shown. Section III shows simulation results detailing the limits of configurability afforded to the template matching system by the memristive components whilst section IV discusses some of the important practical operating considerations and concludes the paper.

## II. Concept and operational principles

Template matching-based spike sorting systems rely on the basic principle that action potentials emitted by the same neuron will be recorded as stereotyped waveforms as shown in fig. 1(a). Repeatedly time-sampling these stereotyped waveforms allows the generation of templates that a discrete-time system such as our proposed one could then recognise and discriminate from other templates. We shall refer to each template sample as a 'template pixel' ('texel').

### *A. Overall system architecture*

The cornerstone of the proposed architecture is a standardised, programmable texel circuit that assembles into arrays capable of scanning conditioned input signals for the presence of spikes on-line, as shown in Fig. 2. The system operates as follows: Each channel receives a preconditioned neural data signal $f(t)$ from the AFE which then feeds into a comparator (U1). Once $f(t)$ exceeds 'spike detection' threshold $TH_1$ a spike is considered to be occurring (a technique also used in [4]). That triggers the Finite State Machine (FSM) which draws a fixed number of incoming analogue input signal samples and distributes them to a bank of Sample & Hold (SH) circuits; effectively an analogue register. Every tap in the register then feeds an entire column of texels, but only texels that receive an input sufficiently close to their preferred, programmed input will respond by outputting a current on to the output line of the template they belong to. The different spike templates to be recognised are stored in the different texel rows of the array. The current entering the output line of each template is integrated on a capacitor (C1) for a suitable amount of time and if an adjustable 'spike recognition' threshold $TH_2$ is exceeded a spike belonging to the corresponding template is registered. This small-bandwidth/high-information flag signal can then be transmitted outside the body. At the end of the entire procedure C1 is reset. The system operates in discrete time as concerted by a clock enforcing a suitable sampling rate (typ. $7-28\,kHz$ [10]).

### *B. Block description*

*1) Input comparator, FSM and SH:* The comparator can be implemented using either a standard low power clocked latch design (Fig. 3(a)) or a continuous time operational amplifier. In the former case power will be saved, but in the latter the expected reduction in sampling jitter may improve performance accuracy. The FSM can be implemented either as a counter or as a linear one-hot register, in both cases feeding a multiplexer. The multiplexer, in turn, routes a succession of input signal samples to the SH bank. Once triggered the FSM cannot be re-triggered until the SH bank is full. Finally, the SH circuit is implementable using the switch capacitor circuit topology carrying out correlated double sampling in CMOS imagers [14] (Fig. 3(b)). In the present system, however, the input is single-ended. The optimisation of the circuitry supporting the texel array lies outside the scope of this paper.

*2) The texel:* The basic architecture of the texel is shown in Fig. 4 and consists of two stages. The first stage is an inverter where a memristor-based potential divider has been introduced between its transistors. By changing the resistive states of the memristors the switch point of the inverter can be shifted (see Fig. 4(b,c)). The second stage is another inverter, this one fed through a mirror supply. The output of the texel

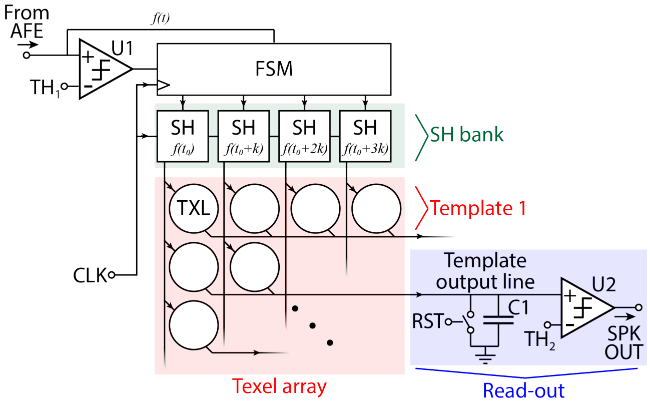

Fig. 2. Block diagram level of proposed system architecture (1 channel). The input signal $f(t)$ is fed into a comparator and an FSM. Once the 'spike detection' threshold $TH_1$ has been exceeded the FSM is activated and distributes a series of input signal samples to a bank of sample & hold circuits. These samples feed the texel array where they cause each texel to respond by pumping current onto its corresponding template output line. The closer the match between input and stored template, the higher the current. Individual templates are stored in texel rows. The aggregate template output current is integrated on C1 and if C1 charges above the 'spike recognition' threshold $TH_2$ the respective template emits a flag.

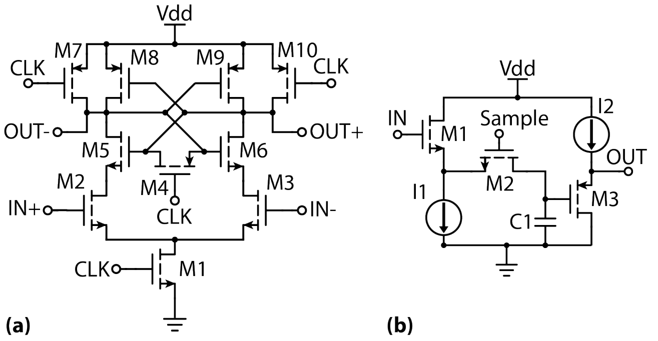

Fig. 3. Supporting circuitry modules. (a) Low power clocked comparator. (b) Sample and hold circuit. I1 and I2 can be gated to save power. I1 should be operational only when a sample is being loaded and I2 when the inputs to the entire texel array are ready and an answer needs to be computed.

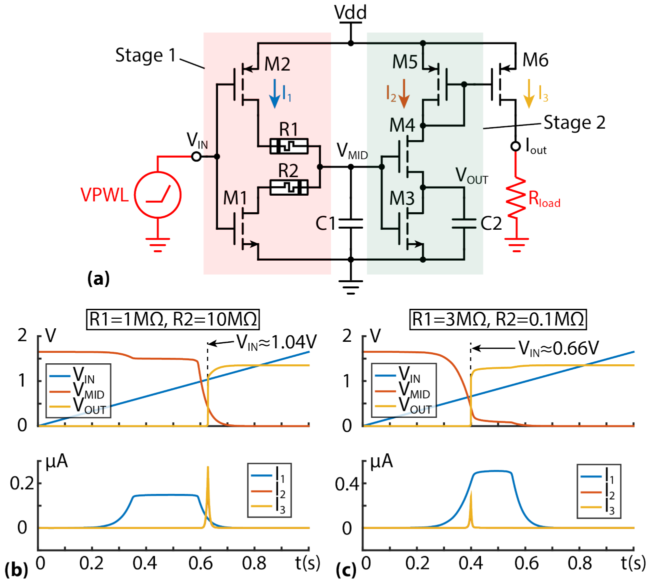

Fig. 4. Texel architecture and basic operation. (a) Circuit diagram with elements belonging to the test bench used for (b,c) coloured red. $R_{load} = 1\,k\Omega$. The memristive devices are marked as R1, R2. (b,c) Examples of texel operation for (b): $R1 = 1\,M\Omega$, $R2 = 10\,M\Omega$ and (c) $R1 = 3\,M\Omega$, $R2 = 0.1\,M\Omega$. The input voltage is swept from GND to VDD over $1\,s$ while current through the three branches (upper panels) and voltage at key nodes (lower panels) is monitored. The control of $V_{pk}$ through the values of R1, R2 is evident, as is the 'plateau' region where the memristor divider dominates the overall resistance of stage 1.

circuit is not the voltage, but the current draw of the second stage inverter, reflected in $I_{out}$. This is maximised when the voltage at node 'MID' $V_{MID}$ is such that both M3 and M4 are simultaneously maximally open, which will occur at some value $V_{pk}$ determined by the sizing of M3 and M4. The end result of this topology is that by shifting the switch point of stage 1 the texel input voltage causing $V_{MID} = V_{pk}$ and consequently $I_{out}$ to reach its maximum can be controlled.

The operation of the texel can be understood by examining two transfer characteristic ($I_{out} = f(V_{IN})$) examples illustrated in Fig. 4(b,c), each for a different configuration of memristor values. In one case, $I_{out}$ spikes at $V_{IN} \approx 0.66\,V$ whilst in the other case the spike occurs at $\approx 1.04\,V$ yet the fundamental operation mechanism is common. M1, M2, R1 and R2 form a four-component potential divider where for each value of $V_{IN}$ every component shows a fixed static resistance ($R_{stat}(V_{IN}) \equiv R(V_{IN}) = \frac{V(V_{IN})}{I(V_{IN})}$). Sweeping $V_{IN}$ reveals its effect on the balance between these resistances: in both cases the current in stage 1 follows a table-hill form. At the 'table-top' region of the curve the dominant R is the series combination of the two memristors (R1, R2) as evidenced by the fact that $V_{MID}$ is largely constant and determined roughly as $V_{div} \propto \frac{R2}{R1+R2}$. At the 'foothill' regions the gate-source voltage $V_{GS}$ of either M1 or M2 has begun to sink below threshold ($V_{th}$) letting it dominate the divider.

So long as $V_{div} \neq V_{pk}$, it is at one of the foothills that $V_{MID}$ will become equal to $V_{pk}$ and $I_{out}$ will maximise. The sum of $R1 + R2$ and its relation to the $V_{IN}$-dependent static resistances of M1,2 defines the range of input voltages for which the memristors dominate the divider and therefore $V_{MID}$ cannot reach $V_{pk}$. Meanwhile the ratio $\frac{R2}{R1+R2}$ determines $V_{div}$ and thus indirectly the shape of the foothill regions and consequently the point where $V_{MID} = V_{pk}$.

The texel can also be operated at the table-top region. Keeping the ratio between R1 and R2 such that $V_{div} \approx V_{pk}$ will allow $I_{out}$ to maximise for a range of $V_{IN}$ voltages determined by the sum $R1 + R2$. This can be useful when the entire length of the pattern need not be defined. Any texel may be tuned to a broad range of input signals (always on) by setting $V_{div} \approx V_{pk}$ and maximising $R1 + R2$.

Finally, the second stage detects whether the output voltage of the first stage is close to its preferred $V_{pk}$ and outputs a corresponding current in response. In both example cases $I_{out}$ is maximised at $V_{MID} \approx 0.42\,V$ (max. $I_{out} \approx 250\,nA$).

*3) Read-out circuit:* The read-out subsystem can use a clocked latch-based comparator similar to the input comparator. The reset signal that discharges the texel output integrating capacitor can be globally shared. The option of integrating texel output on a capacitor rather than measuring the potential across a load resistor was chosen for power considerations.

## III. Simulation Results

In order to better understand the capabilities of the proposed architecture the texel circuit was simulated over a broad range of resistive state values for R1 and R2. Performance was then assessed using three key metrics: a) The voltage at which $I_{out}$ peaks, b) the breadth of the voltage range for which $I_{out} \geq \frac{I_{out}}{2}$ and c) the maximum overall texel steady-state power dissipation at its preferred $V_{IN}$. Results are summarised in Fig. 5. For these simulations TSMC $0.35\,\mu m$ technology devices have been used, C1 and C2 are set to $1\,pF$ (small but controllably achievable in integrated implementations), and $VDD = 1.65\,V$.

The simulations reveal a number of interesting trends: First, $V_{pk}$ is tunable within a range of $\approx 0.55 - 1.05\,V$, which is broadly similar to the range defined by the power supply minus the thresholds of the p- and n-MOS devices $[V_{th,n}, VDD - V_{th,p}]$ (approx. $[0.50, 0.94]\,V$ in this technology). Second, memristor resistive states within $0.1 - 5\,M\Omega$ suffice for covering a large part of that range. Third, the full-width half maximum (FWHM) plot of $V_{pk}$ indicates that to some extent it might be possible to cover most of the $V_{pk}$ range at a controllable degree of sensitivity. The broader the FWHM the blunter the 'tuning curve' of the texel (the broadness of the $I_{out}(V_{IN})$ curve - Fig. 4(b,c)). Finally, the power dissipated at $V_{IN} = V_{pk}$ increases, as expected, when both memristors drop in resistive state, but remains at $< 1.5\,\mu W$ for the majority of the design range. It must be stressed that the power dissipation when the texel is shown a non-preferred input will be generally much lower, as can be inferred from Fig. 4(b,c).

In terms of overall power and area requirements (texel array only) let us investigate a 3-template, 10-texel/template system (30 texels). Power: we consider the case where a 10-point input is being presented to the array and matches one of the

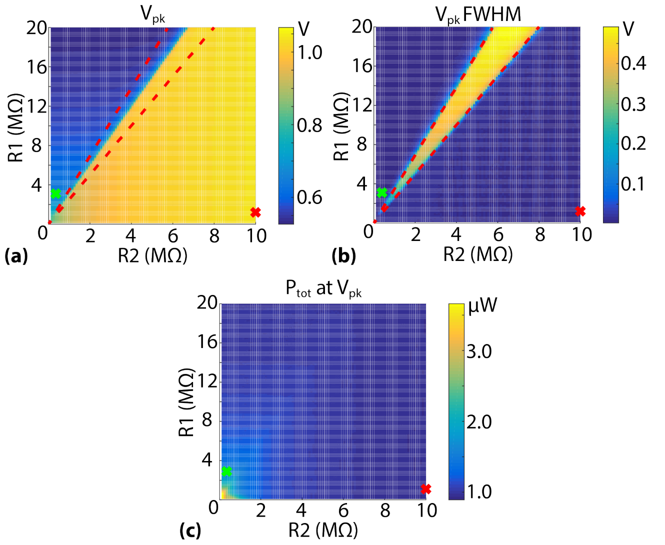

Fig. 5. Key texel performance indicators vs. memristor resistive states. (a) Input voltage at which $I_{out}$ reaches maximum $V_{pk}$. (b) Full-width half maximum (FWHM) of $V_{pk}$. (c) Total texel power dissipation $P_{tot}$ when $V_{IN} = V_{pk}$. Red and yellow crosses indicate the configuration of the simulations in Fig. 4(b) and (c) respectively. Red dashed lines in (a,b) roughly delimit regions where FWHM is broader than the majority of the design space.

stored templates. Whilst the texel array is computing the match the matching row will consume approx. $10\,pts \cdot 1.5\,\mu W = 15\,\mu W$ whilst the non-matching rows will consume approx. $2\,patterns \cdot 10\,pts \cdot 0.5\,\mu A \cdot 1.65\,V = 16.5\,\mu W$ for a total of $31.5 \mu W$ (based on the 'table-top' current in Fig. 4(c)). If the system operates at $12\,kHz$ this comparison can be performed at most at $f_{sample}/(pts/template) = 1.2\,kHz$. If we further assume that a texel assessment can be completed within $\frac{1}{f_{sample}} \approx 83\,\mu s$, then the maximum channel power dissipation for an input signal consisting of a constant stream of back-to-back matchable spikes drops to $3.15\,\mu W$. Area: transistors M4,5,6 occupying a $W \cdot L$ of $120 \times 1\,\mu m^2$ each, comprise the majority of the total nominal transistor $W \cdot L$ of $415.5\,\mu m^2$ footprint ($500\,\mu m^2$ incl. 20% overhead). Note: These calculations are intended to illustrate rough expected power/area overheads only. The technology, transistors sizings and power supply voltage are not optimised and the currents flowing through the system are conservative (e.g. reasonable template-wide match estimate obtainable with $I_{out} < 250\,nA$/texel).

## IV. Discussion and conclusions

The proposed architecture features its own set of design considerations. First, the AFE block is affected through its input signal range requirements (usable range of $\approx 0.5\,V$), the most notable feature being the difference between input range and power supply. This may be potentially addressed using lower threshold transistors in suitable CMOS technologies, in which case the supply voltage may be able to drop without loss of performance. Another important consideration is noise. This is mitigated by the capacitor in the SH circuit (Fig. 3(b)) and by the integrator-based read-out approach (Fig. 2). Next, memristors are non-linear I-V elements which may render control over the precise distribution of voltage in the first stage of each texel challenging. This can be mitigated by using a 1/chip (or few/chip), normally-off programmer that programs the texel array one row at a time; accessing the memristors in each texel individually and manipulating them until the pattern current is maximised at the correct input (currently under development).

Simultaneously, the architecture inherently allows control over the tuning sharpness for each template through adjustment of $TH_2$ in Fig. 2, shows great promise in terms of down-scaling both in area (6 transistors/texel + back-end elements) and power and obviates the need for an analogue-to-digital converter (ADC) anywhere in the system.

In conclusion we have presented a concept architecture for a memristor-CMOS hybrid on-line template matcher with a view towards integrated implementation. We discussed the basic operating principles, gave simulation results based on TSMC 0.35 micron technology illustrating the reconfigurability of the texel circuit underlying the architecture and performed back-of-the-envelope power and area overhead calculations. With further optimisation this technology offers a potentially disruptive solution to the problem of brain recording.

## References

[1] I. H. Stevenson and K. P. Kording, "How advances in neural recording affect data analysis." *Nature neuroscience*, no. 2, pp. 139–142, feb.

[2] "The utah intracortical electrode array: A recording structure for potential brain-computer interfaces," *Electroencephalography and Clinical Neurophysiology*, vol. 102, no. 3, pp. 228 – 239, 1997.

[3] D. Y. Barsakcioglu *et al.*, "An Analogue Front-End Model for Developing Neural Spike Sorting Systems," *IEEE Transactions on Biomedical Circuits and Systems*, no. 2, pp. 216–227, apr.

[4] A. Rodriguez-Perez *et al.*, "A low-power programmable neural spike detection channel with embedded calibration and data compression," in *IEEE Transactions on Biomedical Circuits and Systems*, no. 2, apr, pp. 87–100.

[5] S. E. Paraskevopoulou and T. G. Constandinou, "A sub-1W neural spike-peak detection and spike-count rate encoding circuit," in *2011 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, nov, pp. 29–32.

[6] A. Berényi *et al.*, "Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals." *Journal of neurophysiology*, no. 5, pp. 1132–49, mar.

[7] R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering," *Neural Computation*, no. 8, pp. 1661–1687, aug.

[8] S. Gibson *et al.*, "An efficiency comparison of analog and digital spike detection," in *2009 4th International IEEE/EMBS Conference on Neural Engineering*. IEEE, apr, pp. 423–428.

[9] M. S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," jul.

[10] J. Navajas *et al.*, "Minimum requirements for accurate and efficient real-time on-chip spike sorting," *Journal of Neuroscience Methods*, vol. 230, pp. 51–64, 2014.

[11] M. Chae *et al.*, "A 128-channel 6mW wireless neural recording IC with on-the-fly spike sorting and UWB Tansmitter," in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*. IEEE, feb, pp. 146–603.

[12] I. Gupta *et al.*, "Real-time encoding and compression of neuronal spikes by metal-oxide memristors," *Nature Communications*, pp. 1–16, sep.

[13] R. Waser and M. Aono, "Nanoionics-based resistive switching memories," *Nature materials*, vol. 6, no. 11, pp. 833–840, 2007.

[14] S. Yoshizaki *et al.*, "Octagonal CMOs image sensor with strobed RGB LED illumination for wireless capsule endoscopy," in *Proceedings - IEEE International Symposium on Circuits and Systems*, 2014, pp. 1857–1860.