A VLSI field-programmable mixed-signal array to perform neural signal processing and neural modelling in a prosthetic system

Simeon A. Bamford[1,3*], Roni Hogri[2*], Andrea Giovannucci[3], Aryeh H. Taub[2], Ivan Herreros[3], Paul F.M.J. Verschure[3,4], Matti Mintz[2], Paolo Del Giudice[1]

[1]Complex Systems Modelling Group, Istituto Superiore di Sanità, Rome, simeon.bamford@iss.infn.it.

[2]Psychobiology Research Unit, Dept. of Psychology, Tel Aviv University.

[3]Laboratory for Synthetic, Perceptive, Emotive and Cognitive Systems, Universitat Pompeu Fabra, Barcelona.

[4]Catalan Institute of Advanced Studies.

[*]Equal contributors - see Attributions and Acknowledgements for details.

***Abstract*—A VLSI field-programmable mixed-signal array specialised for neural signal processing and neural modelling has been designed. This has been fabricated as a core on a chip prototype intended for use in an implantable closed-loop prosthetic system aimed at rehabilitation of the learning of a discrete motor response. The chosen experimental context is cerebellar classical conditioning of the eye-blink response. The programmable system is based on the intimate mixing of switched capacitor analogue techniques with low speed digital computation; power saving innovations within this framework are presented. The utility of the system is demonstrated by the implementation of a motor classical conditioning model applied to eye-blink conditioning in real time with associated neural signal processing. Paired conditioned and unconditioned stimuli were repeatedly presented to an anaesthetised rat and recordings were taken simultaneously from two precerebellar nuclei. These paired stimuli were detected in real time from this multi-channel data. This resulted in the acquisition of a trigger for a well-timed conditioned eye-blink response, and repetition of unpaired trials constructed from the same data led to the extinction of the conditioned response trigger, compatible with natural cerebellar learning in awake animals.**

I. Introduction

Where brain functions are impaired through brain damage or through degeneration caused by ageing, it may be possible to develop prostheses which could interact with the brain in order to replace this functionality. While existing neural prostheses either provide input to the nervous system (e.g. cochlear prostheses [1], deep-brain stimulators [2] etc.) or take output from it (e.g. motor cortical prostheses [3]), a largely unmet challenge is the creation of devices that take input from the brain and provide output to it, in order to replace or supplement the functionality of a circuit internal to the brain, although software-based prototypes are appearing [4, 5].

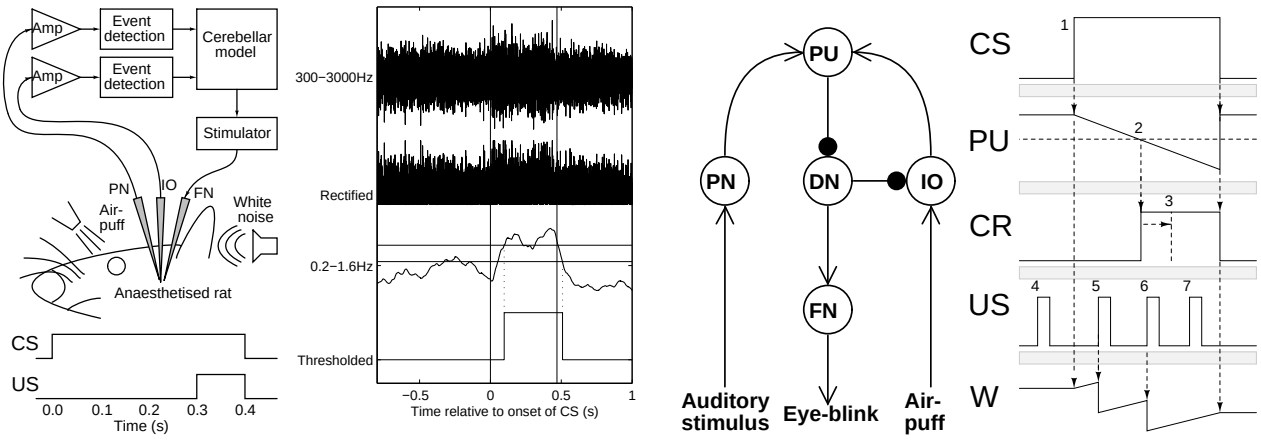
The aim of the European ReNaChip project [6] was to provide a proof of concept for such a closed-loop prosthetic system. The cerebellum was chosen as a target brain area because its well-defined inputs and outputs facilitate physical interventions whilst its relatively simple internal structure have proved fertile grounds for neural modelling from Marr onwards [7]. Eye-blink conditioning was chosen as a well studied target behaviour against which success can be measured. It is intended that the replacement system should be biomimetic, i.e., its architecture and functionality should mimic the characteristics of the area which it replaces according to a neural model of the behaviour of the area. Whilst the system is not specifically intended for clinical application, there has been a focus on practical constraints such as miniaturisation and power constraints for implantability. The project has involved electrode design, neurophysiology, modelling of cerebellar learning, signal processing methods, real-time system integration and chip design. This article focuses on chip design, particularly how a field-programmable mixed-signal array is used to fulfil the computational requirements. Firstly, in sect. II, the target system is described, including: the eye-blink paradigm; electrode placements for recording and stimulation; signal processing methods for real-time extraction of stimulus related events from neural recordings; and the model of cerebellar function which allows on-line learning. Then in sect. III the chip prototype is introduced and its features explained. The key experiment by which the performance of the developed circuitry is demonstrated is the real-time acquisition and extinction of a learnt timed response based on *in vivo* recorded data, for which methods and results are presented in sect. IV and V, respectively.

II. Target prosthetic system

A. Eye-blink conditioning

Eye-blink conditioning is a form of classical conditioning that is commonly investigated with the delay paradigm [8]. An auditory stimulus (conditioned stimulus - CS) and air-puff to the eye (unconditioned stimulus - US) are applied according to the timing scheme in fig. 1a (bottom), in which the CS onset precedes the US onset by an inter-stimulus interval (ISI) of a few hundred ms and the two stimuli then co-terminate. A US alone causes the subject, whether human or rodent, to blink; this is called an unconditioned response (UR). After many repetitions of these paired stimuli, however, the subject learns to blink in response to the CS, prior to the US, at an appropriate time to anticipate the aversive stimulus; this is called a conditioned response (CR). It is known that the cerebellum is necessary for this learning to occur [9]. The target structure for replacement, therefore, is a microcircuit of the cerebellum.

The cerebellum has two inputs and one output, as shown in fig. 1c. Inputs related to all sensory stimuli come from the pontine nucleus (PN) while sensory inputs related to inherently aversive stimuli (US) also come from the inferior olive (IO). Both inputs arrive at the Purkinje cells (PU). Output from PU is inhibitory to the deep cerebellar nuclei (DN). A learnt timed response manifests itself as activation of specific DN cells, from where signals go to premotor nuclei including the red nucleus and on to motor nuclei, such as the facial nucleus (FN) from where, in the case of this paradigm, an eye-blink is elicited.



(a) (b) (c) (d)

Figure 1: (a) System overview: electrodes in PN and IO bring signals to amplifiers; amplified signals are processed leading to detection of CS and US events, respectively (see b); the detected events are inputs to a model of cerebellar learning (see d); the output triggers a stimulator to elicit an eye-blink response via an electrode in FN; in a "paired" trial for CR acquisition, a white noise stimulus (CS) and an air-puff (US) occur in the sequence shown at the bottom, i.e. US starts 0.3 s after CS, and they co-terminate after another 0.1 s; (b) Event detection example for multi-channel data from PN encompassing a CS event: the period of the CS is shown as vertical lines at 0 s and 0.47 s; incoming data from the PN electrode is amplified, band-pass filtered (300-3000 Hz) and the 3 channels are summed together to yield the top trace; this is rectified (second trace), and then band-pass filtered (0.2-1.6 Hz) to yield the third trace; hysteretic thresholds, shown as two horizontal lines, are then applied to yield the detected digital event in the bottom trace; the delay of onset detection of $\approx$ 100 ms is partly explained by neural transduction through the auditory pathway and mainly by the time taken to aggregate information before making a decision; the hysteretic threshold captures the duration of the CS event (after the onset delay) without detecting false alarms from the smaller threshold incursions; magnitudes on y-axis are arbitrary. (c): Simplified cerebellar microcircuit. (d) Cerebellar learning model: CS onset (1) triggers slow reduction of PU activation; when this goes below a threshold (2) it triggers CR, and after a fixed delay (3) IO is inhibited; synaptic weight (W) rises during CS due to LTP; the US events shown here do not refer to the protocol in (a) but rather indicate how the learning model responds to paired stimuli with various timings; a US event prior to CS (4) does not trigger LTD but during CS (5) does cause a fixed amount of LTD (reduction of W); after CR is produced, US events still trigger LTD (6) until the IO inhibition (3) after which they do not (7); during CS, therefore, W reduces in response to US occurring before and slightly after CR, and otherwise increases.

The intended overall system is shown schematically in fig. 1a. Recording electrodes are inserted in PN, where a neural response to the CS can be detected, and in IO, where a response to the US can be detected. The signals from the recording electrodes are amplified and go through various stages of filtration (as detailed in the figure caption and sect. IV-B), resulting in detections of CS and US events. These are input to a model of cerebellar function, whose output may be a timed response to a CS event. This output (the modelled CR) triggers a stimulator which elicits an eye-blink (behavioural CR) through an electrode implanted in FN. The system is therefore meant to bypass and emulate the neural circuitry that implements learning and effects the appropriately timed response. The following sections provide more detail on the aforementioned parts of this system.

B. Event detection

The signals from the electrodes are treated as multi-unit; i.e., the aim is to detect energy related to a population of spikes rather than to identify spikes from particular neurons; an increase in energy is observed in response to the stimuli, which is typically sustained in the case of PN [10] and phasic in the case of IO. The signal is amplified (gain $\approx$10000$\times$) and filtered in the frequency band associated with spikes (typically 300-3000 Hz), resulting in traces of magnitude $\approx$0.1V RMS. For the multi-channel electrode in the PN, the signals are summed together according to a weighting calculated offline, based on the quality of event detection that can be obtained from each channel separately. Then signals are rectified and band-pass filtered to yield a measure monotonically related to the energy over a small window of time (the energy envelope), and a threshold is applied to yield onsets and offsets of detected events. The high cut-off frequency of the band-pass filter is a compromise between the need to detect events immediately to act on them in real-time, and the need to aggregate more information over a longer period to make better detections. The low cut-off frequency is not critical but removes long-term drifts in the background energy in traces, as can be observed in acute experiments with anaesthetised animals. For PN, where detections may last a few hundred ms, the band is on the order of 0.1-1 Hz (CS detection should at least occur prior to the minimum ISI that can be learnt, which might be$\approx$150 ms [11]), whereas for IO, where the phasic response may be as short as 25 ms, the band is $\approx$1-10 Hz. The thresholding of the PN trace is hysteretic, so that given the typical pattern of response with a large phasic component followed by a smaller sustained component, the offset time

can be detected without lowering the threshold, which would increase false positive detections. Fig. 1b shows an example of this procedure (which is common for a range of biosignals [12, 13]).

C. Cerebellar model

The learning model of the system presented here is based on a biologically constrained model of the cerebellum and its role in classical conditioning [14, 15, 16]. Fig. 1d presents a simplified scheme. The time course of the CR depends on the total effective excitatory drive onto the PU cells that is adjusted through the interplay of long-term potentiation (LTP) caused by the CS and long-term depression (LTD) caused by the US in the presence of the CS. Learning through LTD causes the CS derived input to a specific PU to diminish. As a result this PU will start to pause in its response to a CS. Due to the absence of PU activity the DN is released from inhibition and a CR is triggered. LTD caused by coincident CS and US events incrementally reduces the input to the PU and brings forward in time the moment at which a CR is triggered. Over many pairings, the timing of the eye-blink will precede that of the US and will be considered a CR. The correct timing of the response is stabilised through a negative feedback from DN to IO which, once activated to deliver a CR, also blocks US signals from being conveyed to the cerebellum, thus preventing further LTD. The feedback delay of this loop is tens of ms [17], which serves to match peripheral delays in the production of an eye-blink. In the continued absence of paired trials, LTP caused by the CS alone will ultimately extinguish a previously learnt timed response.

The real-time features of this model have been previously assessed using robotic experiments and key features of this model have already been implemented in an aVLSI form [16]. Further validation has been obtained by interfacing it directly to the brain [4]. The model is interpreted in this work as high-level, not concerned with details at the level of spiking transmission or molecular mechanisms of plasticity, and not necessarily indicative of the behaviour of individual PU cells but rather as an aggregate behaviour. Nevertheless the model contains some elements common to neuromorphic electronic design, such as decaying time courses (as in the activation of PU cells during the CS), events triggered by threshold crossings (as in the CR event caused by the reduction of PU to DN inhibition below a certain level), the need for the storage of a value representing (in this case aggregate) synaptic weight and integration of plasticity events on that value, based on relative timing of events (as in the application of LTP based on a CS event and the application of LTD based on the arrival of a US event during a CS event).

III. Chip design

A. Prototype chip

A chip prototype has been designed and fabricated to implement the cerebellar microcircuit replacement prosthesis described in sect. II. The design respects many of the constraints of implantation, although the current prototype does not offer a standalone solution. The chip (fig. 2a) contains three cores, (1) a voltage bias generator; (2) low noise neural amplifiers; (3) a field-programmable mixed-signal array (FPMA). The FPMA core is capable of implementing event detection (sect. II-B) and the cerebellar model (sect. II-C) and is the focus of this article. Other cores are not used in this work; any voltage biases necessary are supplied externally, and the amplifier core (which would include the first stage of filtration in fig. 1b) is by-passed, with pre-amplified and pre-filtered signals brought to the inputs of the programmable array. Note that a complete solution would also contain a core for generating stimulation pulses, whereas this prototype can be used to trigger an external stimulator.

This section introduces and describes the programmable core. A typical field-programmable gate array (FPGA) contains an array of digital logic primitives which are surrounded by a matrix of programmable interconnect such that primitives can be wired together by setting digital switches; thus arbitrary digital computers can be constructed. Such devices are commonly used especially in prototyping systems. The field-programmable analogue array (FPAA) concept is similar except with analogue computational primitives. Various authors have attempted to use diverse primitives in FPAAs, including transistors [18] current-mode circuits [19], switched capacitors (SC) [20], and higher level compound blocks [21, 22]. The many different possible requirements of analogue circuits suggest a spectrum of different design choices from the choice of primitives upwards and dictate against the generality achievable with FPGAs, limiting application of a given FPAA architecture to a given application domain. It will be argued in sect. VI-B that, with certain design choices, neural signal processing and neural modelling is a promising domain for this technology.

The core that has been created is a field-programmable mixed-signal array (FPMA), but not in the usual sense of an FPGA and an FPAA core on the same chip with a layer of analogue-to-digital and digital-to-analogue converters (ADCs and DACs) separating their domains [23, p. 71] [24]. Rather digital and analogue signals are mixed "intimately", sharing the same routing resources, and a key novelty is the method of controlling currents to allow this mixing (sect. III-E). The general approach taken is to work with discrete-time voltage-mode signals by means of SC circuits; this is a common choice for academic and commercial designs alike [20, 25, 26, 27]. The SC technique emulates resistances by switching the terminals of capacitors; this standard technique will not be explained here.

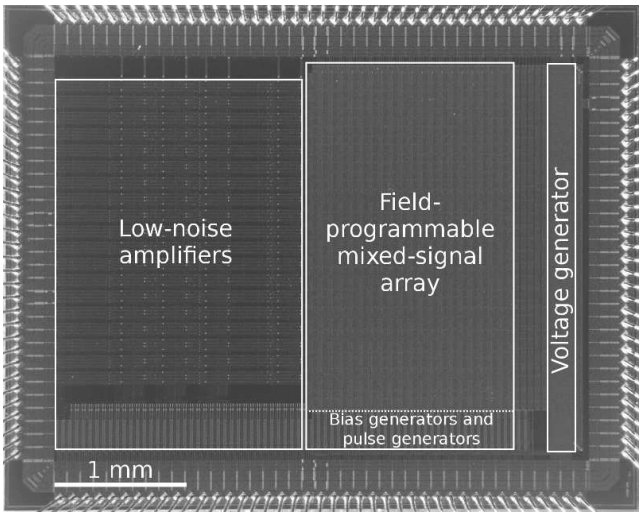
The primitives (hereafter "components") are of 4 types: pulse generator (PGN), configurable switched capacitor (CSC), operational transconductance amplifier (AMP) and configurable logic block (CLB); schematics are shown in fig. 2b. They are laid out in an island-style topology [28], with relatively permissive routing which is not optimised for low path impedance. Configuration of components and routing is by the row-parallel programming of SRAM cells distributed throughout the chip. There are 500 components of the various types;

this is therefore a fine-grained design, (whereas most commercial designs have offered a small number of components [27, 26, 29]), and the intention is to operate with many small, low-quality components, using a combination of calibration and pooling of components to deliver accuracy where it is required. For details of the core architecture see fig 2c.

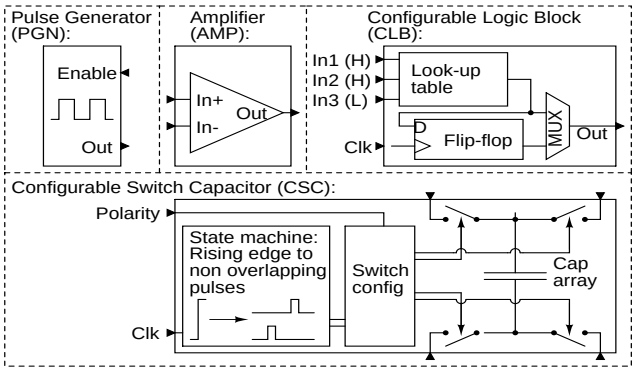
Limitation of power consumption is a major concern for implantable hardware and a prominent reason for working with analogue circuitry. In the following 4 sections, key aspects of this design are described that limit power consumption and otherwise make it fit for the domain of neural signal processing and neural modelling. These aspects are: switched capacitor optimisation (sect. III-B); current control (sect. III-C); leakage limitation (sect. III-D); and the mixing of analogue and digital signals (sect. III-E). Then sect. III-F, shows how rectification is performed, as an example computation which utilises all components and which is part of the signal processing chain of sect. II-B.

B. Switched capacitor optimisation

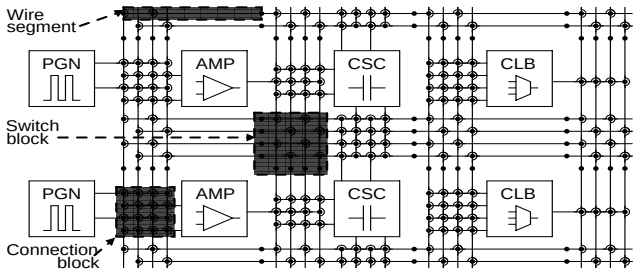
The choice of SC circuitry allows great flexibility but is not ideal for power consumption, since repetitive charging and discharging of clock nodes can pass significant current with respect to the charging and discharging of the voltage-mode signal nodes that they act on. Nevertheless there is much that can be done to limit power consumption. Firstly, CSCs are clocked by a single signal and each contain a state machine for locally generating a pair of non-overlapping pulses in response to a rising edge (fig. 2b). This halves the power used in charging and discharging clock nodes compared to transmitting the two non-overlapping clocks on separate wires. Secondly, clocks are not global but rather generated by PGNs and routed only to where they are needed. The CSCs take their clock signals from the programmable matrix, also allowing them to pass single packets of charge in response to irregular events generated elsewhere within the array; this has possible uses in neuromorphic modelling, a novelty which sets this design apart from other SC FPAAs, but which is not exploited in this article. The aforementioned state machine is insensitive to the slew rate of the clock, thus reducing the requirement for the strength of the driver of the clock signal, which needs to source and sink current only just fast enough to charge and discharge the clock node once per cycle. (The state machine is based on the slew-rate insensitive D-type flip-flop of [30]). This can reduce the effect of clock noise in the system, since clock nodes typically slew much more slowly than in digital systems, meaning that driven nodes onto which these signals are coupled may have much smaller transients as a result. Thirdly, PGNs can be enabled by routed digital signals, thus processes that are active with only a short duty cycle (there are many within the cerebellar model, see sect. IV-D) may consume much less power than if they were continuously clocked.



(a) The chip uses a 0.35 μm process (Austria Microsystems), and has dimensions 4.8×3.8 mm. Cores are indicated on the photo.



(b) Simplified schematics of the 4 components types. The CLB (sect. III-E) is as in [30] with an additional high-starved input ("H"). The CSC (sect. III-B) has 2 switched inputs to each side of its capacitor and can work in lossless as well lossy modes [31], or as an analogue switch or a static capacitor; the capacitance is an array programmable by SRAM [as in 20, fig. 6], in the range ≈50 fF - 1.6 pF. "Polarity" explained in sect. III-F. For PGN see sect. III-B; For AMP see sect. III-D.



(c) Components are laid out in an island-style topology [28]. Each routing bus has 8 wires (only 4 are shown for clarity), wire segments span one row or column of the component array and the switch blocks connect each wire terminal to 5 others (relatively permissive [28, sect. 5.1.3]). Switches (shown as dots) are transmission gates (T-gates; i.e. an NMOS and a PMOS in parallel) each gated by a dedicated SRAM cell. Configuration of components and routing is by row-parallel programming of SRAM distributed throughout the chip; the chip is laced with 337 word lines and 410 bit-line pairs driven by standard programming circuitry in the periphery; these define a grid of possible SRAM locations of which ≈60,000 are occupied. Each 8-transistor SRAM/T-gate cell uses $\approx 54\mu m^2$. The core has dimensions 1.6×2.9 mm; this area is dedicated to: routing switches - 58%; components - 34%; and bias generators (sect. III-B) and decoupling capacitors - 8%. There are 500 components: 40 PGNs, 120 CSCs, 180 AMPs and 160 CLBs.

Figure 2: Chip design overview

C. Current control

The signals involved in the initial stages of the chain of filters must pass signals of up to 3 kHz, implying a Nyquist rate of 6 kHz and a clock frequency for CSCs significantly higher (the core has been designed for frequencies up to $\approx$ 100 kHz). Later stages in the process have high cut-off frequencies on the order of just 1 Hz, and the cerebellar model of sect. II-C needs a slowly ramping signal representing PU activation (trace 2 of fig. 1d) which decreases over a period of order 1 s, for which clocked processes of order 10-100 Hz may be sufficient. There is therefore a range of greater than 3 orders of magnitude of different frequencies of operation and it should be possible to set the currents associated with these various processes appropriately so as not to waste power. The core is divided into 10 bands of components, each of which has associated bias currents which can be set to bias the AMPs, the CSCs' state machines, and the CLBs (sect. III-E). It is intended that different circuitry operating at different speeds be placed within these bands, so that only those components with a high speed requirement are run at high power. The 24-bit programmable current generators of [32] have been reworked for SRAM programming. The currents are used both to bias components and to drive oscillators in the PGNs. The current of each generator can be individually altered over several orders of magnitude from a master current of 2 μA down to < 1 pA, producing oscillator frequencies from $\approx$ 100 kHz down to << 1 Hz. Taking the aforementioned slowly ramping PU-activation signal as an example, this was constructed as a SC integration [31], with a PGN driving a CSC, an AMP for active operation and a CLB (sect. III-E) controlling the activation of the ramping. The PGN was biased at 330 pA, giving a frequency of $\approx$100 Hz, which (for chosen capacitor ratios) set the speed of the ramping. (Other less critical biases were set in a similar range: 3 nA for the AMP and 250 pA for the CLB and CSC).

D. Leakage limitation

Since some signals, e.g. the level of PU activation, are intended to vary with a time constant of order 1 s or below, the leakage of charge through switches to such nodes becomes a cause for concern. Leakage is reduced in a mode suggested by [33]. The chip has two pairs of power rails, an inner and an outer pair. The outer pair, *vdd* and *gnd*, are separated by a standard 3.3 V, whereas the inner pair are offset by programmable voltages from the outer pair, e.g. to 3.1 V and 0.2 V respectively. All inputs to the programmable interconnect are powered by the inner power rails and are thus constrained to remain between them, whereas the SRAM cells which control the T-gate switches are powered by the outer rails. This means that if a node required to carry a stable voltage is separated from other nodes carrying unknown voltages by a switched off T-gate, V_{gs} is guaranteed to be a maximum of -0.2 V (for the NMOS), thus limiting the currents through the transistors to the fA range. A suitable choice of the offset voltage at each power rail can reduce the currents through the transistors until they are comparable to the reverse diode leakage current,

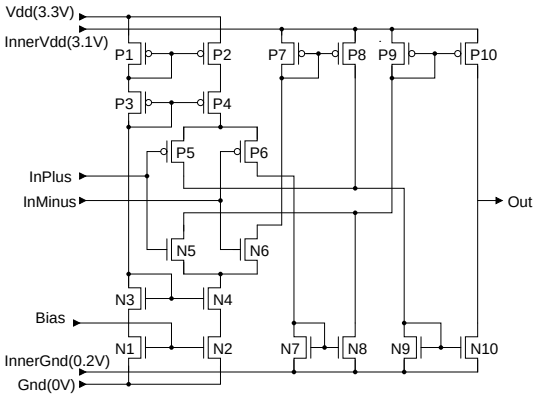


Figure 3: Amplifier topology. The input stage (P/N1-6) is cascoded (P/N3-4) to reduce offset, with the bias (N1-2) mirrored with pMOSFETs (P1-2) for rail-to-rail operation. The input stage operates between the outer power rails for maximal input range whereas the intermediate mirrors (P/N7-8) and output stage (P/N9-10) operate between the inner power rails as is the limitation for all signals which pass through the interconnect network.

which ultimately limits the stability of a node. The use of inner and outer power rails to reduce leakage has been demonstrated in a different context in [34, sect. 3B]. Measurements on this chip show that a typical net consisting of 30 routing wire segments and only parasitic capacitance can achieve a leak as low as 35 mV/s, a 200-fold reduction compared to when no offset is used. Thus this technique can reduce leakage by orders of magnitude and allow voltages stored on capacitors to remain almost stable over time scales relevant for neural modelling. For this, a proportion of the voltage range available for analogue computation has been sacrificed. The transistor-level design of the AMP component is given in fig. 3 as an example of how the dual power rails are utilised. It is a single-ended output amplifier based on a standard rail-to-rail topology but is altered so that its output stage is limited to the inner power rails whereas its input stage operates between the outer power rails, optimising linearity over the input range.

E. Intimate mixing of analogue and digital signals with asymmetric logic

Digital logic is used to supplement analogue computations where required. For example, in the model described in sect. II-C, the direction of synaptic plasticity depends on the timed convergence of direct and modulatory inputs on synapses from CS and US signals respectively; such a decision can be implemented with a logical AND gate. Digital circuitry also allows the building of stable binary-valued memories of arbitrary precision, e.g. to store the weight value in the model. The CLB component allows these possibilities. In search of a simple flexible design, the CLBs have been placed in the same matrix of programmable interconnect as the other components (fig. 2c), such that any component can act as an input to any other, e.g. an AMP implementing a threshold can act as an input to a CLB.

A standard approach to power reduction in digital logic is to increase the slew rate of signals so as to reduce "crowbar

current". This is the current which flows through a logic gate, e.g. an inverter, when its input is not saturated at one of the power rails. In a system where analogue signals may be used as digital inputs, slew rates may be arbitrarily slow, and thus a different solution is required. The CLB design (fig. 2b) has been described in [30]. To summarise, this uses starved logic gates to limit crowbar current. As AMPs and the state machines of CSCs can be biased to define their speed of operation, the maximum currents that flow through the digital gates of the CLBs are likewise programmable, also defining their intended speed of operation. The logic gates are starved asymmetrically, and this asymmetry allows useful circuits such as a D-type flip-flop which is insensitive to the slew rate of its clock, and a CLB configuration which checks the digital saturation of an input, as used in this experiment, see sect. III-F.

Outputs of the CLBs are all current-starved in one direction, such that digital signals are allowed in the programmable matrix which transition upwards quickly but downwards more slowly (according to how they are biased). More generally, signals in the matrix are driven by currents which can vary over many orders of magnitude or which are driven only by switched capacitors and therefore undriven between pulses. This introduces several possibilities for signals with large and/or fast swings to couple capacitively to other signals which may be sensitive to noise. Capacitive coupling mainly occurs in the routing matrix and is especially problematic when two signals run alongside each other on parallel wires for long distances. Sect. IV-B gives an example in which a filter design was selected specifically to avoid such a problem. It is also possible for sensitive signals to be protected by the routing algorithm, for example by being flanked by grounded wires, though with an additional resource cost.

F. Full-wave rectification

Rectification, as required in the chain of signal processing leading to event detection, is given as an example of how the components described above can be used together to perform computation. Fig. 4a shows a rectifier circuit, which uses the same principle as [35]. It is based on the active low pass filter circuit shown in fig. 4a (inset), which is mapped into the components previously described. CSC1-2 act as R1-2 respectively and CSC3 acts as C1 (for clarity, the diagram shows only the ports of components which are used). *InputOffset* is a voltage bias at the level around which the input signal *In* is centred. PGN provides the regular pulse stream which drives CSC1-2. Using the same clock for both components simplifies the setting of the gain and cut-off frequency of the filter to a matter of adjusting the ratios of capacitance in CSC1-3. Each CSC shown here may be composed of more than one physical component wired in parallel in order to achieve the desired capacitance. AMP1 determines whether *In* is above *InputOffset*. AMP2 applies further positive feedback to sharpen the previous decision. CSC1-2 act in lossless mode [31], with their ground set to a voltage bias *OutputOffset*. This bias is set in a calibration phase to a level which compensates for any systematic offsets due to mismatch, to deliver an output centred around the desired voltage (as will be described in sect. IV-D). CSC2 acts as a transresistance, whereas the output of AMP2 is used as the "polarity" of CSC1 (a specialisation of the CSC component, which is controlled by the input labelled "P", such that ϕx/y are ϕ1/2 or vice versa), so that CSC1 either acts directly as a transresistance when *In* is below *InputOffse*t, giving negative or inverting gain, or otherwise acts as a negative transresistance, giving positive gain, effectively rectifying the input. The CLB is programmed with the XNOR function to act as a logic level detector [as in 30] on the polarity, disabling the pulse generator when *In* is close to *InputOffse*t to prevent an intermediate polarity input to CSC1 causing an improper switch sequence. An example output from the chip is shown in fig. 4b (the PGN operated at $\approx$50 kHz and the filter was programmed and calibrated for a gain of $\approx 2.2\times$); additional phase shift can be seen, as well as clipping at the bottom towards the threshold due to the clock disablement; however, performance is more than adequate for its subsequent use in energy detection.

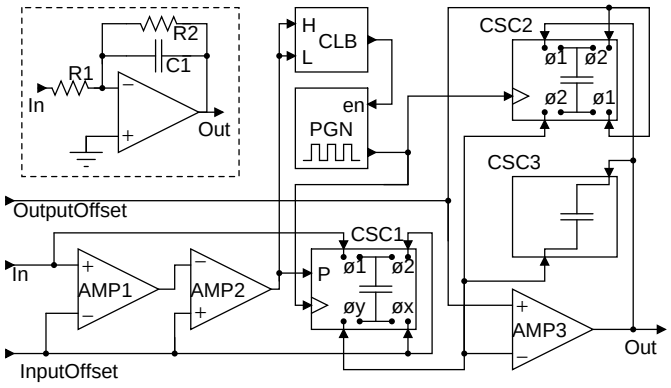
IV. Methods

A. Electrophysiology

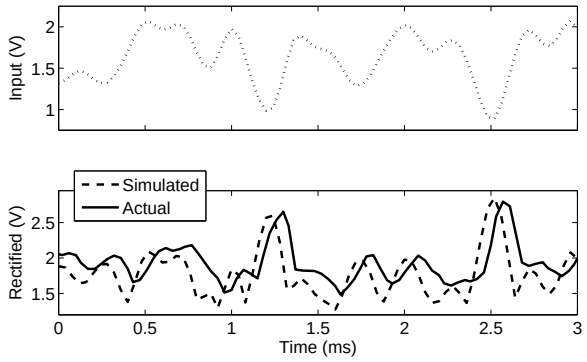
The data was selected from a batch of 6 electrophysiology sessions. In each session an anaesthetised rat had a 3-twisted platinum wire (California fine wire) electrode inserted into the PN to detect the CS and a 5 MΩ tungsten needle electrode (A-M Systems, USA) or a stainless steel entomological pin #000, insulated except for ~0.15 mm tip, into the IO to detect the US. These electrodes were connected to a standard amplification system (MCP-plus, Alpha-Omega, Israel) which applied 10000$\times$ gain and Butterworth filters: 2-pole high-pass at 300 Hz; 4-pole low-pass at 3000 Hz. The 4 signals were then digitised at 14286 Hz per channel with a standard sampling system (Power1401, CED, UK). The CS was a white-noise stimulus of 67-70 dB for 470 ms delivered through a hollow ear-bar of a stereotaxic head holder to the right ear. The US was an air-puff of 1.5 bars at source for 100 ms delivered through a nozzle about 2 cm from the right eye. The ISI was 370 ms, such that CS and US co-terminated. 60 paired CS-US trials were delivered, with an inter-trial interval (not including CS duration) of 8 s. The rat was sacrificed and electrode locations were confirmed with histology. All procedures were approved by the Tel Aviv University Animal Care and Use Committee (P-05-004).

B. Simulation of event detection and parameter setting for model

The signal processing was conceived as a chain of first-order filters, where the first in the chain was rectifying as in sect. III-F, with cut-offs for IO at 3000 Hz LP (rectifying); 30 Hz LP; 6.4 Hz LP; 1 Hz HP. The 30 Hz step was added in order to avoid extreme capacitor ratios in the step down to 6.4 Hz. For PN, the final three cut-off frequencies were instead: 10 Hz



(a)



(b)

Figure 4: Rectification: (a) inset top-left: First order active low-pass filter design. (a) rectifier circuit using all 4 component types - see main text for explanation; CLB ports are "H" (high-starved input) and "L" (low-starved); CSC port "P" controls switching polarity. (b) performance example: The chip was routed to implement the design in (a). It was then calibrated for gain and output offset (see sect. IV-D). Input data (dotted line) was white noise shaped by filters which match the filters offered by the MCP amplifier (see sect. IV-A) with 250 mV RMS magnitude. This was reconstructed by a DAC at 25 kHz and streamed in to the chip. A short section of the output of the filter is shown (full line) against a simulation of the output of an ideal full-wave rectifier and 3000 Hz first order low pass filter.

LP; 1.6 Hz LP and 0.2 Hz HP. It has been noted in sect. II-B that the precise filter frequencies are not critical but are based on heuristics. For PN, an additional 3000 Hz LP filter was included at the beginning of the chain which had one input for each channel and performed weighted summation. Gain was introduced at each filter stage. In the first one or two stages for IO and PN respectively, sufficient gain was introduced to bring the signal to 500 mV RMS. Then gain was 4 and 3 for the two low-pass stages (these values were selected to keep the signal utilising the available voltage range). An active high-pass filter has only parasitic capacitance on its virtual ground and this node can therefore suffer from capacitively coupled clock noise in the programmable interconnect. Thus a passive high-pass filter was used for the final stage (the gain was therefore unity). These signal processing chains were applied in software to each digitised trace separately using IIR filters. Following the final stage, a threshold was applied, where an iterative search yielded the threshold which to the nearest 1 mV maximised bespoke quality measures. In the case of IO, the quality measure was based on the background frequency of US detections being as close as possible to 1 Hz (a level which empirically works well over diverse recordings). In the case of PN, the quality measure rewarded CS detections which started in a time window up to 100 ms after the CS onset and lasted for the correct duration, and punished deviations from this ideal. (Details of the quality measures and further insights on these methods will be published separately). For the PN, which has multiple recording points, the quality measure was used to provide weights for the summation of the channels in the first filter stage, such that channels which individually provided better information about the stimulus contributed more.

The model was parametrised with: a threshold for the production of a CR when PU activation reached a proportion of 0.2 of its full baseline value (an arbitrary choice); a rate of (linear) reduction of PU activation such that it passed from maximum to minimum in 1 second; a delay from the CR onset to inhibition of IO of 80 ms (higher than the 20-30 ms observable in biology [17] in order to accentuate the observable effect in this experiment); and LTP and LTD rates which were set so that an acquisition of a well-timed response would ideally be achieved after 60 paired CS-US trials (a physiologically realistic number of trials would be ≈ 500 for rat but corresponding to rabbit and much fewer in humans) and extinguished after the same number (fluctuations in detection performance would cause deviations from these ideals however).

The model was simulated based on the detections from the previous stage, to confirm that acquisition and extinction of the learning of a well-timed learnt response was possible in principle based on applying these methods to the available data. To do so, the traces recorded in the 60-trial experiment were repeated twice, allowing there to be a phase of acquisition in which the weight value should decrease, followed by a phase of stability in which the weight value should be maintained in the same region by the negative feedback (in a control systems sense) effected by the (feedforward) inhibition from DN to IO. Thereafter the traces were repeated twice more but with the IO recording shifted forward in time by ITI/2, such that the increase in US-related events did not occur during the CS thus simulating unpaired trials; this allowed another 120 trials in which conditions for extinction were simulated and the weight value should increase to its maximum value and stay close to it thereafter.

Of recordings from the 6 electrophysiology sessions, some had S/N ratios from one or both of the nuclei too low for the described learning to recognisably occur (this will be quantified in a separate publication); the best simultaneous recording from both nuclei was selected for the experiment reported here. Having established that the learning was possible in principle, the same inputs were sent to the chip, yielding the results in sect. V.

C. Chip test environment

The chip was placed on a bespoke PCB providing connections to DACs and ADCs and an integration board (XEM3010, Opal Kelly, USA) hosting an FPGA (Xilinx Spartan 3). The FPGA was used to programme the chip, manage the ADCs and DACs and stream data between the chip and a PC. The chip was designed to be packaged with a minimal pin-out of 56 pins in an 8×8 mm QFP package for implantation; however for testing it has a full pin-out of 144 pins. Of these, 58 are general purpose I/O ports to the FPMA core. Bespoke software for programming of the chip (placing, routing and calibrating) and monitoring of its operation was developed using Matlab (Mathworks, USA). Programming SRAM, for example, involves generation of data words encoding the switch matrix settings generated by a routing algorithm. These words are transmitted via USB to the FPGA, which then effects a serial programming protocol. Programming each of the 337 rows took 2 ms.

D. Chip programming: Place, Route and Calibrate

Various types of sub-circuit were defined, e.g. an active low-pass filter type, with rectifying as a sub-type, as demonstrated in sect. III-F. The event detection chain and cerebellar model were decomposed into sub-circuits and described using a bespoke description in Matlab code. Other sub-circuits included a delay (for example for timing the delay between CR onset and IO inhibition), a linear ramp (for example for describing the behaviour of PU activation following a CS onset), a hysteretic threshold, etc. The delay and linear ramp are two examples of circuits which are event triggered and activate a PGN to drive their process only when required, so as not to waste power on unused clock cycles. Placement of components to form the necessary sub-circuits was performed deterministically based on heuristics from the user; in constructing filters, for example, trade-offs between clock rates and capacitance ratios were calculated from coded heuristics, as well as their relative placement to minimise necessary routing. Routing was then performed using a bespoke algorithm and the chip was programmed. The design used in this experiment employed 43% of CLBs, 89% of CSCs, 21% of AMPs, 38% of PGNs, and 39% of routing wires.

Each stage of processing introduces deviations from ideal performance due to mismatch, for example in amplifier offsets. To compensate for this, calibration routines were devised for each sub-circuit. For example, for active first-order filters, calibration consisted of streaming in a short section of recorded data, recording the filter output, comparing the output to that of the same filter in software and adjusting capacitor ratios and voltage biases to adjust gain and offset respectively. When initially laid out, an excess of programmable capacitance was made available beyond what was needed in the ideal case, to allow the capacitance ratios to be altered to allow for the effects of mismatch and parasitic capacitance from routing wires and switches. The calibration process was iterated until the residual error fell below thresholds chosen by the user, in this case <50 mV offset and <5% difference in gain. A calibration routine could also be devised for cut-off frequency but this has not been implemented. Pulse generator frequency, the basis of filter cut-off frequency and other behaviours, was however calibrated on a component-by-component basis.

To avoid accumulation of offset differences from one stage to the next, input was always to the first filter in the chain and comparison was always with the accumulated effect of all the software filters up to that point in the chain. In case a desired gain could not be programmed because the required capacitance were greater than that allowed for in the placement of CSC components, then the extra gain would automatically be introduced by the calibration in the following stage, correcting the overall behaviour of the signal processing chain (the gain of the final HPF is, however, uncorrectably less than unity due to parasitic capacitance on the output node forming a capacitive divider to ground, but this simply results in altered thresholds for detection).

Having calibrated performance of individual parts of the system, the overall performance was optimised by streaming in the entire sequence of recorded data and optimising: (a) the thresholds for event detection, using iterative search in software as described in sect. IV-B but based on traces recorded from the chip; and (b) the frequencies for plasticity processes, so as to match as well as possible the desired rates of acquisition and extinction of the learnt response.

As mentioned in sect. III-E, the model requires a stable analogue value representing a combined synaptic weight. Given that LTP and LTD need to be finely balanced against each other and that single plasticity events must have sufficiently small effect that learning acts only over many trials, ≈12 bits of analogue depth is required, an accuracy which is difficult to achieve with multi-valued stabilisation mechanisms [36]; however for the timing of eye-blink responses accurate to ≈10 ms, ≈7 bits of accuracy is required, which is more easily achievable. The design therefore used a 12-bit digital incrementer and decrementer (Inc-Dec) circuit, and the 7 most significant bits were converted to an analogue voltage by a DAC circuit. The Inc-Dec circuit was constructed from CLBs and was clocked by the outputs of two PGNs, which were enabled only during plasticity events. A binary-weighted design was used for the DAC, with CSCs emulating resistances. The CSCs were all clocked at the same low rate (since weight changes only slowly) and a calibration phase fine-tuned the capacitance values to maximise the linearity of the conversion given mismatch. This is a case where accuracy can be traded off against resources; the more CSCs used, the better the linearity that can be achieved, see the discussion on accuracy in sect. VI-B1.

V. Results

A. Real-time learning

As described above in sect. IV-B, data from 240 trials (4 repetitions of 60 trials with the first 120 having paired CS-US

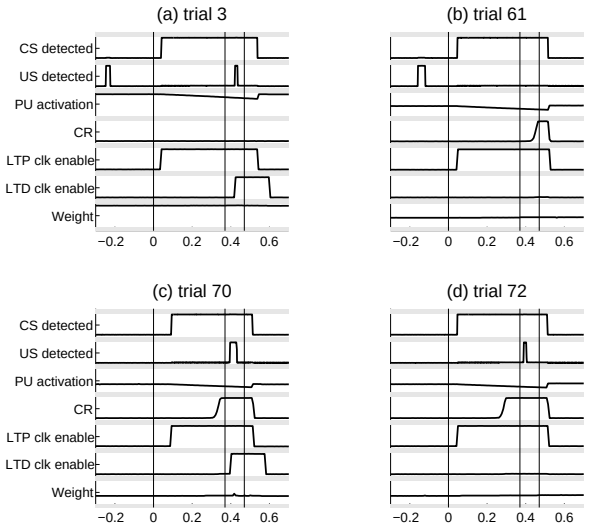


Figure 5: Results. Behaviour of chip signals during 4 example trials. Each graph shows the traces of 7 signals buffered out from the programmable core. Within each signal, the vertical scale shows the voltage range from 0 up to 3.3V. The number of the trial is given in the heading and time on the x-axis is relative to the CS onset, with three vertical lines marking, from left to right: CS onset, US onset and the offset of both stimuli.

events and the rest effectively having CS alone) was streamed to the chip, once it had been programmed to perform event detection and the cerebellar model, and had been calibrated accordingly. Fig. 5 shows the results of selected trials from the experiment. Note the diversity of the signals involved: *CS-detected, US-detected, LTP-clk-enable* and *LTD-clk-enable* are all low-starved digital outputs from CLBs, with biases ranging from 20-500 nA; *CR* is the output of an AMP thresholding *PU-activation* biased at 30 nA (smooth upwards slews can be seen); and *PU-activation* and *Weight* are analogue traces, with *Weight* being the output of a DAC sub-circuit buffered by an AMP, and *PU-activation* being the output of a linear ramp sub-circuit (driven by a CSC). In an early trial, (a), CS and US were both detected, leading to a period of LTP which lasted for the duration of the detected CS, and LTD was applied for a fixed period after the detection of the US. The net effect on the weight was negative, though almost imperceptible in the graph. Note that a detection of IO activity prior to the CS did not cause LTD. During the detected CS, *PU-activation* gradually declined from its baseline level, although not enough to cause a CR. In (b), after *Weight*, and thus the baseline for *PU-activation*, had decreased somewhat, *PU-activation* crossed its threshold causing an output in *CR* around time 0.45, too late to anticipate the aversive stimulus. In (c), with the weight slightly lower, the *CR* event occurred prior to the air-puff, and in (d) the CR happened early enough that the US detection did not lead to LTD, because its action was blocked by the modelled effect of DN to IO inhibition.

Fig. 6 shows overall results for the experiment. Fig. 6(a) shows trial-by-trial detection performance for the two nuclei superimposed, as well as the CR events produced. Most CS events were detected shortly after their onset, and in addition there was a low rate of false alarms. Those correctly detected stayed active for an average of 0.46 s. US onsets were detected during the air-puff with a frequency ≈3 times the background rate. The noise inherent to the system is evidenced by the fact that the pattern of detections of CS and US events was similar but not identical from one block of 60 trials to the next, although the inputs were identical. Nevertheless the modelled neural system achieved the acquisition and extinction of a well-timed response to the CS; fig. 6(b) zooms graph (a) in the region of the acquisition of a well-timed response; the first well-timed CR (excluding one produced due to a false detection at trial 59) occurred at trial 69, and from then until trial 120, 88% of CS events caused a well-timed response. Thereafter the last well-timed CR occurred at trial 125 and from trial 132 onwards there were no more responses, i.e. extinction of the CR. Fig. 6(c) summarises the acquisition and extinction of well-timed responses. Fig. 6(d) shows the evolution of the weight during the experiment. There was a period until trial ≈70 in which it descended, after which it remained buffeted around the same level. Then from trial ≈120 onwards the weight ascended until it reached its maximum level, to which it thereafter stayed close. For comparison with fig. 6(d), (e) shows the evolution of the weight variable during the software simulation of the experiment. Although differences are visible, the broad behaviour is the same.

B. Adapted model

A demonstration of the utility of the programmable system is provided by an alternative experiment. Electrical stimulation of the FN to elicit an eye-blink can introduce large artefacts into the recordings from PN and IO which, unless cancellation techniques were applied, would result in detections of CS and US events for the duration of stimulation, corrupting the action of the model. To avoid this without developing artefact cancellation, an alternative form of the model was implemented on the chip, as in [4], in which there was no delayed inhibition of LTD based on the production of a CR, but rather, both forms of plasticity, LTP and LTD, were inhibited for the duration of a CR (this departs from the biomimetic roots of the model for the sake of practicality). In this case, when CRs are well-timed, US events will be blocked by this mechanism and the weight should stabilise in any case. Results are not graphed due to space, but the weight variable followed a similar trajectory, with the first genuine well-timed CR at trial 69, but with the difference that without the delay to regulate the timing of the response there were more late responses, with only 52% well-timed CRs between trials 69 and 120.

C. Power consumption

Power consumption is presented as measurements of current (at room temperature; with Keithley 6487 picoammeter, Keithley Instruments Inc., USA) into outer *vdd* (thus dropping 3.3 V through the outer power rails) or into inner *vdd*

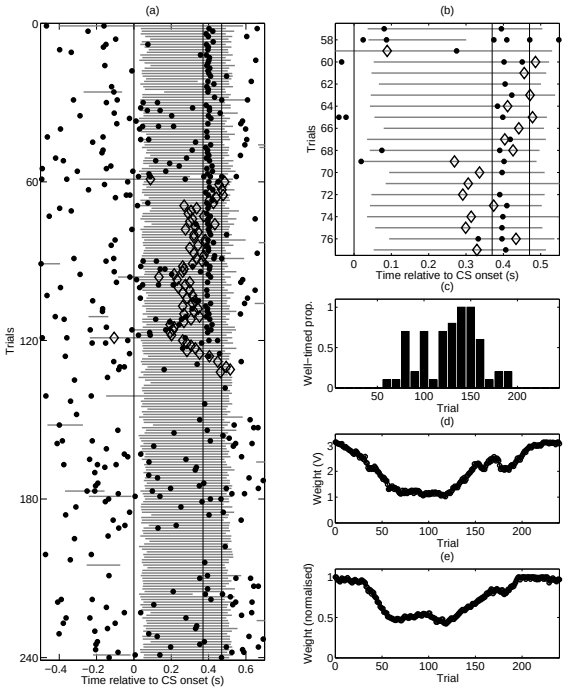


Figure 6: Results. (a): stimulus detections and produced CRs by trial, aligned to CS onsets. Vertical lines mark the CS onsets at 0 s, US onsets at 0.37 s and the joint offsets at 0.47 s (after trial 120 though, the US was artificially displaced outside the CS period); grey lines are CS detections (the output of the hysteretic threshold as indicated in fig. 1b); filled circles are detected US onsets; diamonds are produced CRs. (b): as (a) but zoomed around trials 60-75. (c): proportion of trials evoking well-timed responses, by block of ten trials. (d): weight sampled 1s after the CS onset of each trial. (e): weight sampled 1s after the CS onset of each trial for the software simulation of the experiment.

(dropping 2.9 V through the inner power rails). With the chip powered but all components disabled, 0.9 nA passed through inner *vdd*; this has ≈8000 entry points to the matrix and the components through back-biased transistors, implying ≈110 fA per transistor and demonstrating very low leakage. Outer *vdd* current through the FPMA should be comparably low but interference from other cores on the prototype precludes accurate measurement. The bias generators of sect. III-B leak 2 μA internally, in order to generate currents which may be orders of magnitude lower, and since the core uses 60 of these elements which cannot be enabled separately in this prototype, the quiescent current would be no less than 120 μA. In fact, due to further biases in other cores and to additional buffering, the current through outer *vdd* when they were switched on was 420 μA. Thus biasing overheads are unnecessarily high. During the main experiment (sect. V-A), current increased by 94 μA (outer and inner *vdd* contributed similar currents to this total and are hereafter combined for simplicity). This was dominated by 26 amplifiers in constructed filters and the DAC circuit, which were biased at full strength. 6.4 μA of this was due to switched capacitor operation i.e. to the state machines within CSCs which create non-overlapping clocks, the PGNs (of which 16 were used), and the routing capacitance leading from these to the CSCs; therefore, switched capacitor machinery had a significant but not dominant power cost. Most of the current consumption was due to the fastest processes, i.e. the rectifying filters and the initial summation of inputs from PN electrodes, which operated at ≈50 kHz. A separate experiment was performed in which only the PU activation part of the model was implemented (i.e. 3rd trace in fig. 5). The bias currents used are stated in sect. III-C; this caused <20 nA total additional current during operation.

VI. Discussion

A. Progress and limitations

The work presented here represents a key step in the progress towards an autonomous implantable device which could rehabilitate the function of a circuit internal to the brain. It demonstrates that a device designed specifically for neural rehabilitation has operated in real-time on recorded neurophysiological data to perform the computations necessary for biomimetic replication of the functionality of an internal brain circuit. In this section limitations of the system, both existing and projected, are duly noted.

The target prosthetic system requires a phase of supervised learning in order to optimise performance, e.g. the threshold searches described in sect. IV-B, for which, when applied to the chip, some external programming is required. It is unclear to what extent recalibration would be necessary in a chronic system but it is likely that some degree of reprogramming during operation would be necessary.

The core has not yet operated in a closed loop with a brain, although a software based-system performing similar functions has done so [4]. The data worked with is from anaesthetised animals; operation on data from behaving animals and from chronically implanted electrodes has not yet been demonstrated. The anaesthetised preparation allows a demonstration of rehabilitation without introducing a lesion or acting on aged animals, (where impaired performance might be expected [37]) because under anaesthesia no natural eye-blinks are evident and the rat cannot learn an eye-blink response. The anaesthesia introduces differences from normal neural functioning although previous findings suggest that these differences are minor. It is likely that the microcircuit model approach used here could be applied to other cerebellar learning functions (e.g. vestibulo-occular reflex conditioning) with little effort; however it is not clear to what extent this approach could be parallelised to provide more generalised intervention in the case of a damaged or degraded neural system.

Stimulator circuitry has not been included on the chip prototype. Problems inherent to electrical stimulation include stimulus artefacts in recordings (see sect. V-B), a large current requirement (order 100 μA) for which high voltages are necessary, and chronic problems in the electrode-tissue interface. For these reasons, it is the opinion of these authors that efforts

may be better spent in investigating promising alternatives such as optogenetic stimulation.

B. Application of FPMA to neural signal processing and neural modelling

The field-programmable approach has been useful even within this project, as it allowed chip prototyping to proceed whilst alternative forms of event detection and neural circuit modelling were being investigated without having to first decide on an optimal strategy. This is evidenced for example by the change of cerebellar model in sect. V-B, a trivial change for the programmable core which may have been impossible with a hardwired ASIC implementation [16].

As noted in sect. III the wide range of requirements of analogue circuitry dictate against FPAAs achieving the kind of generality possible with FPGAs in the digital domain, and limit application of a given FPAA architecture to a given application domain. Here, the domain of neural signal processing and neural modelling is proposed as a promising candidate. There have been several discussions regarding design choices in FPAAs, [e.g. 38]. In this section, the design choices that have been made and explained above, particularly the fine-grained topology and discrete-time SC design, are assumed, and issues of noise, speed, power and parallelisation are discussed in reference to neural signal processing and neural modelling.

1) Noise and accuracy: In analogue design there are various sources of inaccuracies including mismatch, noise, large-signal non-linearities and thermal effects. If more area can be dedicated to devices, then in general, inaccuracies due to both mismatch and noise can be reduced due to averaging. Where precision is required it is common to build in the ability to calibrate circuits in some way prior to use, so that the unwanted effects of mismatch can be removed. Calibration procedures for filter sub-circuits were described in sect. IV-D, based on the programmability of capacitance within CSCs. FPAAs offer a more powerful promise for calibration, which has not been demonstrated here. FPAA structure naturally allows access to all components for characterisation of their properties; components could therefore be selected in the placement phase based on their individual properties [an example of this approach from a slightly different domain is 39]. The more fine-grained the design is, the greater flexibility would be available; utilising this approach would be a non-trivial undertaking however, since it would increase the complexity of placing and routing requirements.

Regarding noise, FPAAs trade area of devices against flexibility of design, by using area for configuration circuitry and for resources which may not be used in a given application. Additionally, connecting analogue circuits with switches for flexibility can add noise compared to a monolithic design. Making components larger to give them better noise performance reduces the number of components that can be placed in a given area, and thus reduces flexibility and ultimate complexity of circuits. Thus in general, domains requiring high accuracy are a poor fit to the FPAA concept. In neural signal processing, the quality of signals from electrodes is limited by noise from electrode impedance and early amplification stages; depending on the application, the inherent signal-to-noise ratio of the neural signal may also be a limiting factor. Thus as long as a signal processing system introduces noise at lower levels than those existing in the amplified input, its contribution should be irrelevant. The approach taken here was to use a fine-grained design with many relatively small, low quality components, and the approach proved to be sufficient for the required processing.

Regarding non-linearities, for certain operations such as energy detection and thresholding, linearity of operations need not be precise and monotonicity is a sufficient constraint. Thus the processing of neural signals, once amplified, may have a lower fidelity requirement than in other domains, e.g. audio processing.

Regarding thermal effects, the ultimate application of low-power implanted devices may offer the possibility of disregarding performance change with temperature since temperature is well-regulated inside the body.

2) Speed: To achieve programmability, components which would be directly connected by a wire in a monolithic design are instead connected through a matrix of switches. Each of the switches adds resistance and capacitance to the signal path. Adding impedance to signal paths is made irrelevant by the SC approach, providing only that settling times are long enough during phases in which switches are closed. Given that neural signal processing is concerned with frequencies <10 kHz, even allowing SC circuits to run $10\times$ faster for anti-aliasing, the requirement for the clock rates is <100 kHz, which is not a difficult constraint. This relaxes requirements on the design of switchable interconnect with respect to other application domains, i.e. switches can be small and numerous.

3) Power: Implantable devices have tight power budgets. Excessive heat dissipation can cause tissue damage and beyond this, the lower the power consumption, the smaller can the implanted batteries be and the longer the times between recharging. Notwithstanding recent improvements in the performance of low-power digital processors (e.g. the ARM Cortex) one of the promises of analogue computations is to reduce the power consumption, compared to an equivalent digital implementation of the same computation. The increased capacitance on signals due to programmable interconnect increases the power consumption, and therefore an ultimate design for a low power device would likely be a monolithic circuit. The use of SC circuitry itself is not an ideal choice for power consumption due to the necessity of charging and discharging clock nodes from rail to rail. Certain approaches used here help to limit these losses: non-overlapping clocks produced locally requiring the delivery of only one pulse stream; pulse generators disabled when not required; and pulse streams routed only where required. The results of sect. V-C show that the contribution of SC circuitry to power consumption was not dominant. Nevertheless the prototype presented here is just at the beginning of what could be

achieved regarding power limitation. A priority in a future prototype will be the reduction of the large overhead of the bias generator unit, perhaps by moving to individual biases for each component created by floating-gate transistors. Such a move would also ease placement constraints caused by the banding of components, and would avoid power losses from the inability to individually optimise the current usage of components in the same band. Another design revision may be the elimination of PGN components and the use of the other components to construct oscillators where required. This could reduce routing and hence capacitance for clock signals as well as simplifying the overall design.

4) Parallelisation of hardware: As noted above, it is unclear to what extent the prosthetic intervention presented here could be parallelised, for example allowing the pairing of more neutral and aversive stimuli in order to provide more generalised cerebellar functionality. In general however, neural processing gets its power from its massive parallelism. To achieve parallel processing, in a digital system it is typical, though not essential, to time-multiplex a single or small number of processing cores, whereas in analogue design it is typical to parallelise hardware for computations which must operate simultaneously. Typical digital processing therefore approaches speed constraints whereas typical analogue processing approaches area constraints. A fine-grained design which offers many small, low quality components is a better match to the demands of neural modelling since more resources allow the construction of more circuits in parallel. Parallelism has not been demonstrated here, with only two parallel signal processing chains and the summation of three input channels implemented. In the present prototype, notwithstanding its fine-grained design, area constraints are severe, with 500 components of various types being sufficient but not over-abundant for the task at hand. Nevertheless since the ultimate limits of VLSI scaling are not known, it is too early to conclude that a field-programmable approach would not provide dense enough circuitry for continued use in implantable devices. The majority of the area of the prototype core is occupied by minimum-sized devices and it is hoped to investigate the scaling potential of such a design.

In conclusion, it is argued here that fine-grained FPAA designs applied to neural signal processing and neural modelling may not suffer from some of the drawbacks which limit their applicability to other domains. Meanwhile this approach may offer benefits of rapid prototyping and quicker time to market especially for low-power implantable prosthetic applications.

VII. Conclusion

An FPMA specialised for neural signal processing and neural modelling has been designed and fabricated as a core on a chip prototype intended for use in an implantable closed-loop prosthetic system aimed at rehabilitation of a function internal to the brain. Novelties in the design of the FPMA include: the intimate mixing of SC analogue techniques with current-starved digital computation and power saving innovations within this framework; and the adaptation of components for use within a switch-leakage-resistant framework employing inner- and outer- power rails. The utility of the system has been demonstrated by the implementation of classical conditioning of an eye-blink reflex, resulting in the acquisition of well-timed responses to paired conditioned and unconditioned stimuli, which have been detected in real-time from multi-channel data recorded simultaneously from two sub-cerebellar nuclei, and the extinction of those responses given unpaired trials constructed from the same data.

Attributions and Acknowledgements

Simeon Bamford designed the chip; Roni Hogri and Aryeh Taub performed the electrophysiology; Simeon Bamford and Andrea Giovannucci developed the event detection methods; Andrea Giovannucci and Ivan Herreros developed and tested different versions of the model; Paolo Del Giudice, Matti Mintz and Paul Verschure provided scientific and practical oversight. This work was funded by the ReNaChip EC project grant agreement no. 216809. The authors would like to thank: Massimiliano Giulioni, who designed some parts of the chip not reported in this paper; Robert Prückl, who developed a parallel software-based system in which alternative forms of event detection and modelling were tried; Ari Magal for helpful advice which contributed to chip design; Tobi Delbrück, who contributed the bias generator design; and Angela Silmon for organisational oversight.

References

[1] W. House, "Cochlear implants," *Ann Otol Rhinol Laryngol*, vol. 85 suppl 27(3Pt2), pp. 1–93, 1976.

[2] R. Kumar, A. Lozano, Y. Kim, W. Hutchison, E. Sime, E. Halket, and A. Lang, "Double-blind evaluation of subthalamic nucleus deep brain stimulation in advanced Parkinson's disease," *Neurology*, vol. 51, pp. 850–855, 1998.

[3] D. Taylor, S. Tillery, and A. Schwartz, "Direct cortical control of 3D neuroprosthetic devices," *Science*, vol. 296, pp. 1829–32, 2002.

[4] R. Prueckl, A. Taub, R. Hogri, A. Magal, I. Herreros, S. Bamford, R. Ofek Almog, Y. Shacham, P. Verschure, M. Mintz, J. Scharinger, A. Silmon, and C. Guger, "Behavioral rehabilitation of the eye closure reflex in senescent rats using a real-time biosignal acquisition system," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2011.

[5] T. Berger, R. Hampson, D. Song, A. Goonawardena, V. Marmarelis, and S. Deadwyler, "A cortical neural prosthesis for restoring and enhancing memory," *Journal of Neural Engineering*, vol. 8, 2011.

[6] http://www.renachip.org/.

[7] D. Marr, "A theory of cerebellar cortex," *J. Physiol.*, vol. 202, pp. 437–470, 1969.

[8] D. Woodruff-Pak, M. Papka, and R. Ivry, "Cerebellar involvement in eyeblink classical conditioning in humans," *Neuropsychology*, vol. 10, pp. 443–458, 1996.

[9] V. Bracha, M. Webster, N. Winters, K. Irwin, and J. Bloedel, "Effects of muscimol inactivation of the cerebellar interposed-dentate nuclear complex on the performance of the nictitating membrane response in the rabbit," *Experimental Brain Research*, vol. 79, pp. 453–468, 1994.

[10] A. Taub and M. Mintz, "Amygdala conditioning modulates sensory input to the cerebellum," *Neurobiology of Learning and Memory*, vol. 94, pp. 521–529, 2010.

[11] M. Smith, S. Coleman, and I. Gormezano, "Classical conditioning of the rabbit's nictitating membrane response at backward, simultaneous, and forward CS-US intervals," *J. Comp. Physiol. Psychol.*, vol. 69, pp. 226–231., 1969.

[12] E. Rodriguez-Villegas, P. Corbishley, C. Lujan-Martinez, and T. Sanchez-Rodriguez, "An ultra-low-power precision rectifer for biomedical sensors interfacing," *Sensors and Actuators A: Physical*, vol. 153, pp. 222–229, 2009.

[13] E. Stark and M. Abeles, "Predicting movement from multiunit activity," *Journal of Neuroscience*, vol. 27, pp. 8387–8394, 2007.

[14] P. Verschure and M. Mintz, "A real-time model of the cerebellar circuitry underlying classical conditioning: A combined simulation and robotics study," *Neurocomputing*, vol. 38-40, pp. 1019–1024, 2001.

[15] C. Hofstotter, M. Mintz, and P. Verschure, "The cerebellum in action: a simulation and robotics study," *European Journal of Neuroscience*, vol. 16, pp. 1361–1376, 2002.

[16] C. Hofstötter, M. Gil, K. Eng, G. Indiveri, M. Mintz, J. Kramer, and P. Verschure, "The cerebellum chip: an analog VLSI implementation of a cerebellar model of classical conditioning," in *Advances in Neural Information Processing Systems* (L. Saul, Y. Weiss, and L. Bottou, eds.), vol. 17, pp. 577–584, 2005.

[17] P. Svensson, F. Bengtsson, and G. Hesslow, "Cerebellar inhibition of inferior olivary transmission in the decerebrate ferret," *Experimental Brain Research*, vol. 168, pp. 241–253, 2006.

[18] J. Langeheine, J. Becker, S. Folling, K. Meier, and J. Schemmel, "A CMOS FPTA chip for intrinsic hardware evolution of analog electronic circuits," in *Evolvable Hardware, Proceedings of the NASA/DoD Workshop on*, pp. 172–175, 2001.

[19] O. Fares and M. Abuelmaatti, "Configurable analogue building blocks for field-programmable analogue arrays," *International Journal of Electronics*, vol. 95, no. 10, pp. 1009–1028, 2008.

[20] E. Lee and W. Hui, "A novel switched-capacitor based field-programmable analog array architecture," *Analog Integrated Circuits and Signal Processing*, vol. 17, pp. 35–50, 1998.

[21] H. Klein, "The EPAC architecture: An expert cell approach to field programmable analog devices," *Analog Integrated Circuits and Signal Processing*, vol. 17, pp. 91–103, 1998.

[22] A. Basu, S. Brink, C. Schlottmann, S. Ramakrishnan, C. Petre, S. Koziol, F. Baskaya, C. Twigg, and P. Hasler, "A floating-gate-based field-programmable analog array," *IEEE Journal of Solid-State Circuits*, vol. 45, pp. 1781–1794, 2010.

[23] Y. Sun and R. Lala, *Wireless communication circuits and systems*, ch. Field-programmable and reconfigurable analogue and mixed-signal arrays, pp. 53–76. IET, 2004.

[24] SIDSA, *FIPSOC Mixed Signal System-on-Chip*, 2000.

[25] H. Kutuk and S. Kang, "A switched capacitor approach to field-programmable analog array (FPAA) design," *Analog Integrated Circuits and Signal Processing*, vol. 17, pp. 51–65, 1998.

[26] Motorola, *MPAA020 Field-programmable analog array datasheet*. Motorola, 1997.

[27] Anadigm, *AN120E04I datasheet configurable FPAA*. Anadigm, 2003.

[28] I. Kuon, R. Tessier, and J. Rose, "FPGA architecture: Survey and challenges," *Foundations and Trends in Electronic Design*, vol. 2, no. 2, pp. 135–253, 2007.

[29] Zetex, *TRAC-S2 datasheet*. Zetex, 2000.

[30] S. Bamford and M. Giulioni, "Intimate mixing of analogue and digital signals in a field-programmable mixed-signal array with lopsided logic," in *IEEE Biomedical Circuits and Systems Conference (BIOCAS)*, pp. 234–237, 2010.

[31] K. Martin, "Improved circuits for the realization of switched-capacitor filters," *Circuits and Systems, IEEE Transactions on*, vol. 27, no. 4, pp. 237–244, 1980.

[32] T. Delbrück and P. Lichtsteiner, "Fully programmable bias current generator with 24-bit resolution per bias," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2849–2852, 2006.

[33] B. Linares-Barranco and T. Serrano-Gotarredona, "On the design and characterization of femtoampere current-mode circuits," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 8, pp. 1353–1363, 2003.

[34] S. Bamford, A. Murray, and D. Willshaw, "Spike-timing-dependent plasticity with weight dependence evoked from physical constraints," *IEEE Transactions on Biomedical Circuits and Systems*, 2012 in press.

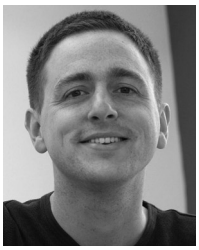
[35] Y. Tanaka and H. Iseki, "Switched capacitor rectifier circuit," 1987.

[36] E. Vittoz, H. Oguey, M. Maher, O. Nys, E. Dijkstra, and M. Chevroulet, *Introduction to VLSI-Design of Neural Networks*, ch. Analog storage of adjustable synaptic weights, pp. 47–63. Kluwer Academic Publ., 1991.

[37] J. Coffin and D. Woodruff-Pak, "Delay classical conditioning in young and older rabbits: Initial acquisition and retention at 12 and 18 months," *Behavioral Neuroscience*, vol. 107, pp. 63–71, 1993.

[38] A. Blaszkowski, W. Ciazynski, M. Garlicki, and J. Kulisz, "A comparative analysis of continuous-and discrete-time field programmable analogue arrays," in *Programmable devices and systems 2003 (PDS 2003): a proceedings volume from the 6th IFAC Workshop*, pp. 183–188, 2003.

[39] E. Neftci and G. Indiveri, "A device mismatch compensation method for VLSI neural networks," in *IEEE Biomedical Circuits and Systems Conference (BIOCAS)*, 2010.



Simeon A. Bamford received a BA in Artificial Intelligence in 1995 from the School of Cognitive and Computing Sciences at the University of Sussex. After an entrepreneurial career he returned to study and in 2009 received a PhD from the Neuroinformatics Doctoral Training Centre at the University of Edinburgh. For this project he was employed by SPECS lab, Universitat Pompeu Fabra, and seconded to the Complex Systems Modelling Group at Istituto Superiore di Sanità, where he now researches neural and neuromorphic engineering, in pursuit of these treasures: the elegant circuit, the ideal application, and insight into the mind.

Roni Hogri received an M.A. in psychology ('08) from Tel Aviv University (TAU). He has been a research student in Matti Mintz's lab at TAU's Psychobiology Research Unit since 2006, and during this period he has studied cerebellar function and learning. His PhD studies focus on the inferior olive's responses to noxious somatosensory stimuli, their modulation and behavioral consequences. Within the ReNaChip project, he has been a member of the Neurophysiology and Behavior team, focused on establishing paradigms for interfacing neural signals with synthetic devices. He has been funded by the EC's FP7 and the Israel Science Foundation's Converging Technologies grants, and by the Dan David Prize Scholarship and the Michael Myslobodsky Fellowship.

Andrea Giovannucci received his BSc degree in electronic engineering from the Milan Politecnico, Italy in 2002; in 2006 and 2008 he received his MSc and PhD degrees in computer science from the Spanish National Research Council (IIIA-CSIC), where he investigated auction mechanisms and optimization. From 2008 to 2010 he was a postdoctoral fellow at SPECS, Pompeu Fabra University, Barcelona, where he focused on computational neuroscience and neuroprosthetics. Since 2010 He has been a postdoctoral research associate in the molecular biology department of Princeton University (Wang Lab), where he is specializing in novel experimental neuroscience methodologies.

Aryeh H. Taub received an MA in Psychobiology ('07) from Tel-Aviv University. He did his PhD in Tel-Aviv University, studying cerebro-cerebellar interaction and biocompatibility of chronically implanted microelectrode ('07-11). During that time, he developed a biocompatible protein-coated electrode for chronic implantation. During his PhD studies, he was also group leader at the Physiology Group for the ReNaChip project The project's goal was studying the feasibility for behavioral rehabilitation by brain-computer interface. Since 2011, he has been a postdoctoral fellow at the lab of Prof. Ilan Lampl at the Neurobiology Department at the Weizmann Institute.

Ivan Herreros (Madrid, 1975) is a Computer Engineer both from the Universitat Politecnica de Catalunya (Barcelona) and Istituto Politecnico di Torino (Turin, Italy). Since 2007 he has been a PhD candidate at the Universitat Pompeu Fabra in Barcelona. There, he's part of the SPECS Lab, where he pursues thesis work in the field of computational neuroscience, designing computational models of the cerebellum to be used in the context of associative learning, such as the one implemented in the neuroprosthetic solution developed for the ReNaChip project.

Paul Verschure received his MA and PhD in psychology and has worked at leading institutes including: the Neurosciences Institute, the Salk Institute, the University of Amsterdam, University of Zurich and the Swiss Federal Institute of Technology-ETH, where he was a founding member of the Institute of Neuroinformatics. Paul has published widely in leading scientific journals including Nature, Science, Neuron, PNAS, TINS, PLoS Biol, etc and conferences in a range of disciplines from microelectronics, theoretical physics and neuroscience to psychology, art, robotics and neurology. Paul has presented over 25 installations, exhibitions and performances based on his research.

Matti Mintz received his PhD in Psychobiology from Tel Aviv University. In 1983 he joined the Department of Psychology of Tel Aviv University and established the laboratory of emotional and motor learning. Current research is directed toward decomposition of the cerebellar microcircuit implied in the classical conditioning of discrete motor responses. Testing this circuit culminated in this multicenter project aimed at replacing the cerebellar microcircuit by a synthetic analog. Other research areas aim at revealing the role of emotion in hippocampal processing of spatial information.

Paolo Del Giudice is a senior researcher at the Italian National Institute of Health. Physicist by training, he works on the theory and electronic implementation of neural models. On the subject he published over 30 journal papers and 30 conference papers and book chapters. He is responsible for a unit of the Italian National Institute for Nuclear Research; adjunct professor of neural networks at the Physics Department of Rome University Sapienza; member of the editorial board of Advances in Artificial Neural Systems, Associate Editor of Frontiers in Neuroengineering and Frontiers in Neuromorphic Engineering; co-organizer of six workshops/schools on neural networks; recipient of several EU and bilateral Italy-US grants.