University of Southampton Research Repository

UNIVERSITY OF SOUTHAMPTON

Neural Spike Classification Acceleration with RRAM Technologies

by

Patrick Foster

A report submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering and Physical Sciences
Zepler Institute

November 2023

Academic Thesis: Declaration of Authorship

I, Patrick Foster declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Neural Spike Classification Acceleration with RRAM Technologies

I confirm that:

This work was done wholly or mainly while in candidature for a research degree at this University;

Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

Where I have consulted the published work of others, this is always clearly attributed;

Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

I have acknowledged all main sources of help;

Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Parts of this work have been published as:

Patrick Foster, Jinqi Huang, Alex Serb, Spyros Stathopoulos, Christos Papavassiliou, and Themis Prodromakis. An FPGA-based system for generalised electron devices testing. Scientific Reports, 2022.

P. Foster, J. Huang, A. Serb, T. Prodromakis, and C. Papavassiliou. An FPGA based system for interfacing with crossbar arrays. In 2020 IEEE International Symposium on Circuits and Systems (ISCAS), pages 1–4, 2020.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

UNIVERSITY OF SOUTHAMPTON

<u>ABSTRACT</u>

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
ZEPLER INSTITUTE

by Patrick Foster

Spike classification is an area of critical importance in neurological medicine. The behaviour of neurons is critical in diagnosis of disease, understanding neural structures, and operating prostheses. The probe technology used to gather in situ data from neurons has seen significant advances in the past decade, but the technology required to process this vast amount of data lags behind and this thesis aims to address the data processing aspect of this issue. A novel analogue circuit that conducts most of the sorting in the pre-processing stage is presented, demonstrating the feasibility of such a system using memristive devices. This thesis covers the progress made during this research project; the development of an new instrument for the testing of memristors and memristor related circuits, the design of a new analogue cell for use in template matching systems, and the testing and simulation of this cell both in isolation and in a simple template matching system. The circuit developed demonstrated comparable energy dissipation to current state of the art spike sorting systems, without the need to digitise the signals being processed. This development opens the path to the fabrication of an integrated memristor based spike sorting system suitable for neural signal processing.

Contents

List of Figures

List of Tables

Nomenclature

ADC	Analogue to Digital Converter
ASIC	Application Specific Integrated Circuit
CAM	Content Addressable Memory
DAC	Digital to Analogue Converter
DUT	Device Under Test
EMI	Electro-Magnetic Interference
FeRAM	Ferroelectric Random Access Memory
FIFO	First-In, First-Out
FPGA	Field Programmable Gate Array
FWHM	Full Width at Half Maximum
GPIO	General Purpose Input/Output
IC	Integrated Circuit
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MRAM	Magnetoresistive Random Access Memory
NLTL	Non-Linear Transmission Line
NMOS	N channel MOSFET
PCB	Printed Circuit Board
PCM	Phase Change Memory
PMOS	P channel MOSFET
RRAM/ReRAM	Resistive Random Access Memory
SEPIC	Single Ended Primary Inductor Converter
SMU	Source Measurement Unit
SPI	Serial-Parallel Interface
SPST	Single Pole Single Throw
SRD	Step Recovery Diode
TXL	Template pixel

Acknowledgements

I would like to offer thanks to Alex Serb, who has provided guidance and support throughout this project. Without his advice, this project would never have reached completion.

I would like to offer thanks to Jinqi Huang, who collaborated on the development of the ArC TWO, providing the configuration for the FPGA. The countless hours spent in the laboratory brought my creations to life. Her work is as much an accomplishment as anything found in this document.

I would like to offer thanks to Spyros Stathopoulos, who provided the software used to operate the ArC TWO and conduct every experiment in this project. Jinqi may have brought the system to life, but Spyros made it dance.

I would like to offer thanks to Themis Prodromakis, who brought us all together and provided the academic and administrative support required to make everything happen.

I would like to offer thanks to Christos Papavassiliou, who invited me to begin this project, and provided advice along the way.

I would like to offer thanks to Angela Westley, who coordinated the research group.

I would like to offer thanks to Kate Cooley, who picked up where Angela left off.

Lastly, I would like to offer thanks to my family, whose unconditional support made this possible.

Chapter 1

Introduction

1.1 Motivation

The monitoring of neuron activity is an important element of neurological medicine. By recording and analysing neuron impulses researchers can achieve a greater understanding of the human brain, allowing for better treatments of neurological disorders, and improve neuroprosthetics. The neural interfaces used in this manner can be broadly grouped into two categories: noninvasive interfaces that use electrodes placed externally on the body and invasive interfaces that use surgically implanted probes. The latter of these can provide far more detailed information, but existing systems require extensive pattern matching to sort neural activity from the different neurons in the vicinity of the probe[1]. This is typically done by amplifying the signals recorded by the probe, digitising those signals, and then sending them to an external computer for processing[2]. This processing cannot generally be done in-situ, as the waste heat from the power hungry processors risks damage to surrounding tissue[3]. Such processing often involves the sorting of spikes in the signal to determine the neuron of origin, tying all of the activity of a neuron to a single identity in the post-experiment analysis. As a result of the large bandwidth required to stream data from several hundred probe channels, many current systems fail to take advantage of the high channel count probes that currently exist[4]. If real-time spike sorting could be implemented in implanted hardware[5][6], the effective compression of the data at this stage would ease the reliance on external computation[7]. This would permit far more complex and capable implants, enabling more thorough research and treatment, possibly as far as sophisticated neuroprosthetics[8]. This is not the first attempt to solve this issue and sorting systems have been developed to address this issue[9], but most such attempts still process digital signals. This requires a power hungry analogue to digital converter between the amplifiers and filters and the sorting circuit. It may be possible to substantially reduce the power requirements of the spike sorting task by conducting the sorting and digitisation in the same step. This would

require the development of low power tunable analogue circuits. Adding digital to analogue converters for every tunable parameter would be wholly impractical, so any such circuit would inevitably require the inclusion components which store information in their electrical properties, rather than in conventional digital memory. An example of such a device would be memristors, whose resistive state is a function of the current that has passed through the device. Memristors have been demonstrated in spike sorted applications, using the integrating behaviour of the such devices to compress unique spike shapes into corresponding step changes in resistance[10]. A further expansion on this concept produced an amplifier for neural signals that used memristors to tune a threshold detector[11]. An alternative approach added memristive devices to simple logic gates to provide low-power mixed-signal circuits[12]. This approach yielded flexible analogue-in, digital-out designs that have the potential to displace ADCs, and warrants further investigation. Using memristors to adjust the threshold of a circuit such as a window comparator would allow for an array of such comparators to be used in a spike sorting method such as a template matching system suitable for neural processing applications. This project develops and demonstrates at a proof-of-concept level a memristor tunable circuit for template matching of neural spikes with a competitive power dissipation below the $80\,\mathrm{mW\,cm^{-2}}$[3] threshold for tissue damage.

1.2   Research Objectives

This research project seeks to demonstrate, at a proof-of-concept level, a template matching system suitable for use with neural spike signals. To achieve this, a window comparator with non-volatile tuning elements must be designed. As the operation of large numbers of such circuits is not possible with existing instruments, a new platform for conducting these experiments is also required. This gives rise to the following objectives:

- Develop an instrument to serve as a platform for characterising memristors, memristor crossbar arrays, and testing a wide range of memristor based designs.

- Design an analogue circuit to facilitate template matching tasks in the context of neural signals.

- Simulate the designed analogue circuit, assessing its suitability for further development with a focus on integration.

- Produce a model of the analogue circuit for testing with physical memristor devices.

- Demonstrate the analogue circuit in a template matching task.

1.3   Structure

This thesis is structured as follows:

Chapter 2 provides an overview of the subject of this research, covering the topic at hand and related subjects, such as spike detection and memristors.

Chapter 3 describes the development and testing of the instrumentation intended to act as the platform for this project.

Chapter 4 discusses the design of several circuits as candidates for further development.

Chapter 5 presents the selected circuit and experiments conducted with it, including integrated circuit simulations and tests with physical model.

Chapter 6 concludes the thesis and discusses the direction of future research

1.4   Contributions

The work described in Chapter 3 was presented at ISCAS 2020[13], and a more thorough paper written for Scientific Reports[14]. These papers present the instrument and analyse its performance in a variety of tasks.. In addition to this, the instrument is in a sufficiently complete state that units have already been sold, and several researchers are already using it in their projects. In some cases, specialised daughterboards were designed for their applications (appendix E).

A paper on the TXL circuit discussed in Chapter 5 as an isolated cell is in progress[15]. This paper will present the Split TXL circuit in isolation, showing simulations of the circuit in integrated circuit development tools, along with test of physical models using memristors. A paper on a split TXL based template matching channel is planned. This paper will present a demonstrator of the Split TXL based template matching channel, showing tests of a physical model on synthetic and recorded signals.

Chapter 2

Processing of Neuronal Spike Signals

2.1 Neural Signals
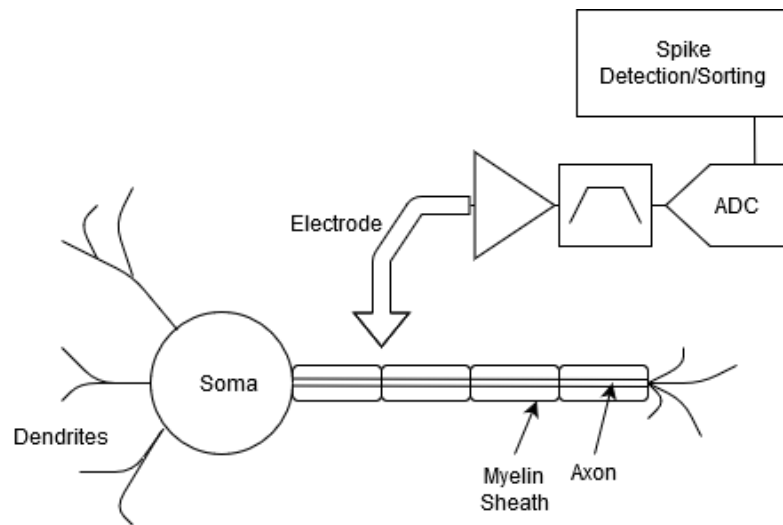
Neural tissue is comprised of neurons, which are specialised cells with a dense array of dendrites branching from the soma and long axon insulated with a myelin sheath (Fig. 2.1). The ends of the dendrites are connected to the axons of other neurons in a structure called a synapse. When triggered by an electrochemical pulse called an action potential, the axon terminal of a neuron releases chemicals called neurotransmitters. These chemicals bind to receptors on the dendrite, which in turn open ion channels. The influx of ions changes the local membrane potential, which triggers voltage sensitive ion channels in the surrounding cell membrane. If enough receptors are stimulated, the

change in membrane potential triggers a critical mass of voltage sensitive ion channels, and the resulting wave of membrane potential and ion transport travels down the axon as an action potential. The movement of ions in and out of the intracellular medium causes a current that can be measured. This is done by inserting small electrodes[16] in to the area of interest and amplifying[17] the resulting output to a usable level. Amplification is necessary because the electrical signals involved in firing of a neuron are very low, typically below $100\,\mu V_{pk-pk}$[18]. These spikes typically last around $2\,ms$[19] and occur at varying frequencies in different parts of the nervous system. The cerebral cortex is an area of high interest, as it is responsible for motor control and perception; it typically has spike rates of $30-80\,Hz$[20], although spike frequencies of up to $450\,Hz$[21] have been observed in human cortical samples. In addition to the spikes, electrodes can also detect electric fields caused by bulk current flow into or out of neurons as they spike and recover. These local field potentials typically have maximum frequencies at or below the firing rate of nearby neurons[22], and can be detected a significant distance away from the neurons contributing to the measured signal[23]. They are usually filtered out in applications focusing on the action potentials, as they are a much less local phenomenon and could otherwise interfere with detection and measurement of the spikes. The information collected from these spike signals can be used to control robotic prostheses, or map neurons for installation of neuroprosthetics, such as cochlear implants[24]. Before this can be done, the action potentials in neurons of interest must first be distinguished from the measurement noise and the action potentials of surrounding neurons.

2.2 Spike Detection and Sorting

The raw neural signals are usually recorded by the implant, and the task of picking the voltage spike of an action potential out from the surrounding noise is left to an external computer[25]. The computer will apply sorting methods to cluster recorded events, which can then be associated with a specific neuron. This approach requires that large quantities of data be streamed from the implant: an array of 100 probes[26] operating at sample rate of $20\,kHz$ with a 12-bit resolution requires at least $24\,Mbit\,s^{-1}$ of bandwidth. This can be reduced by using in situ methods of spike detection, such as voltage thresholds or nonlinear energy operators[27] to limit the transmitted data to only the spike events. However, any given neuron might be firing at up to 200 times per second, so energy savings are not guaranteed. Preferably, an in situ method of classifying neural signals would be used, but most systems still use many microwatts of power per channel[28].

2.2.1   Software Sorting

The software approach to spike sorting typically takes one of two forms: template matching or principal component analysis.

Template matching compares the recorded signals with an example spike. This comparison can be done by calculating the distance between the waveform and the template[29], but it can also be done by calculating the cross correlation[30]. Cross correlation or convolution can be very effective at detecting coincident spikes from different neurons.

Principal component analysis[31] considers samples captured from the waveform as a multidimensional array and processes this array with matrix transformations to find new variables in which the spikes of different neurons can be grouped. This is a more recent approach than template matching, as it requires significant computing power for any dataset that isn't trivially small.

Modern implementations of template matching and principal component analysis can achieve very high accuracy[32] with a wide range of waveforms, but their reliance on external computing makes them unsuited for achieving the goals of this project and as such they will not be considered going forwards.

2.2.2   Hardware Sorting

Amplitude alone can be used to detect spikes, but in a situation where two neurons are producing differently shaped spikes of similar amplitude the accuracy of this method suffers. A simple way around this is to use a time amplitude discriminator[33]. This is a circuit that returns a match if the input waveform falls within an amplitude window at a set time after the spike has passed an amplitude threshold.

A more sophisticated approach to in-hardware spike sorting is template matching, where a fixed number of samples are collected from the input waveform when an event is detected. The resulting values are then compared to a known pattern and the number of matching values used to determine if the waveform matches the template. This approach has two advantages: the circuit can sleep to reduce power consumption when not sampling and the same set of samples can be used as input to several sets of matching hardware where each is tuned to detect a different pattern. Since checking a sample for a match requires some sort of window comparator and many samples must be compared, the obvious path to improve on this approach is to reduce the power consumption and complexity of the window comparator. Simple window detection circuits, known as analogue content addressable memory, are a subject of current research[12][34], although the technology is not yet mature.

Another approach might be to implement one of the many spike sorting algorithms into a digital application specific integrated circuit (ASIC). As modern digital circuitry can operate at very low supply voltages, the power requirements of the sorting element of such a system can be very low. A system of this type was reported to have a dissipation of $64\,\text{nW}$[9], setting an ambitions standard to beat, although this is quoted for a signal that has already been digitised.

Neuromorphic sorting methods could be a better fit, as the crossbar structures common in two-terminal memory technologies lend themselves well not only to the matrix multiplications that are central to neuromorphic applications, but also to deconvolution tasks[35], and in memory computing[36]. Samples are collected in a similar manner as with the earlier example of template matching, but the samples are used to bias the wordlines of a crossbar array instead. The current in a given bitline can then be used to determine a match for the template set by the devices on that line. Cutting edge designs already show considerable improvement, cutting down both the power consumption and substrate area of the sorting system[37]. In 2020, a research group demonstrated a neural signal processing system based on a neural network implemented as a memristor crossbar array[38], for the purpose of detecting epilepsy-related neural signals. They report that their work used $1/400^{th}$ the power of a contemporary CMOS ASIC system, without sacrifice in accuracy.

2.3   Memory

In all of the hardware sorting methods discussed the measured signal is being compared against information stored in the circuit, whether that be in a conventional databank, the configuration components of a template, or the weighting of a neural network. Thus the limits of spike sorting tools are defined by the memory technologies available. Higher density of suitable memory would allow for an expansion of spike sorting systems, highlighting the importance of selecting the right technology (Tab. 2.1).

	FLASH	PCM	MRAM	FeRAM	ReRAM	Molecular
Data type	Multi level[39]	Multi level[40]	Binary[41]	Binary[42]	Analogue[43]	Multi level[44]
Area	$50 \times 50\,\text{nm}$[45]	$35 \times 35\,\text{nm}$[46]	$0.26 \times 0.54\,\mu\text{m}$[47]	$80 \times 34\,\text{nm}$[42]	$30 \times 30\,\text{nm}$[48]	$0.01 \times 2\,\mu\text{m}$[44]
Write Speed	$0.4 - 1.5\,\text{ms}$[39]	$100\,\text{ns}$[49]	$1\,\text{ns}$[47]	$1\,\mu\text{s}$[50]	$5\,\text{ns}$[51]	$1\,\text{ms}$[52]
Retention	10^4h[53]	10^5h[49]	10^5h[54]	10^5h[55]	10^5h[56]	$600\,\text{h}$[44]
Endurance	10^6[57]	10^{11}[49]	10^{11}[47]	10^8[50]	10^8[51]	10^{12}[58]

TABLE 2.1: A table comparing non volatile memory technologies.

The most common non volatile memory in use is flash, which operates by trapping charge on an electrically isolated gate between the gate and channel of an otherwise relatively conventional planar MOSFET. Developed from EEPROM in the 1980s[59], flash is by far the most mature memory technology covered in this chapter, with high densities of stacked triple level cells (cells capable of enough discreet levels for 3 bits

of information) being common in consumer products. While well established, flash has significant limitations. The mechanism for placing charge onto the floating gate requires high voltage supply, typically 12 V. Not only does this result in high write energies, subjecting the cell to such high voltages also causes the cell to fail with enough write-erase cycles. Flash ICs are typically rated for no more than $100,000$ cycles, although the actual endurance is often higher[57].

Phase change memory (PCM) is a newer technology, comprised of a layer of chalcogenide glass between metal contacts. Passing current through the cell heats the glass, with different heating profiles allowing for the cell to set in amorphous or crystalline state, changing the resistance[60]. Most memory of this type only stores a binary state, but two level cells have been demonstrated[40]. Phase change memory offers vastly superior endurance to flash, with many millions of cycles being possible. As there are no trapped charges to escape, it also displays better retention, at least when ambient temperatures are at a reasonable level[49]. The downsides of this technology are that high current densities are required to achieve the necessary heating, and that the chalcogenide glass is not commonly found in integrated circuit fabrication processes, complicating its implementation.

Magnetoresistive memory (MRAM) uses a pair of ferromagnetic layers with a thin insulating layer between. When the magnetic fields in the ferromagnetic layers are aligned, the probability of an electron tunnelling across the insulator layer increases, effectively reducing the resistance of the cell[61]. Writing to MRAM is typically done using a grid of wires, running across the entire array along the horizontal and vertical cell rows[62]. Selecting the wires passing across a cell, and passing a current through them, forms a small magnetic field in the cell, altering the state of the ferromagnetic layers. While relatively simple, this method also subjects horizontally or vertically aligned cells to a smaller magnetic field, limiting the strength of the writing field. This is relevant because the retention of the cell is a function of the writing field strength. An alternative method of writing to MRAM is passing spin-aligned electrons through the cell. When electrons pass into a layer that forces them to change their spin, they transfer some of their angular momentum to the layer, altering its field[41]. This approach avoids the issues of the more conventional approach, allowing smaller cells, but must still make compromises between power, speed, and retention. MRAM endurance is limited by the breakdown of the thin tunnelling layer, but under typical operating conditions endurance can exceed many millions of cycles[63]. Both high temperatures and strong magnetic fields can compromise the retention of MRAM, but outside of such situations the retention is excellent. While the binary nature of MRAM makes it unsuited to the applications under consideration in this thesis, its characteristics make it a popular choice for non volatile RAM research.

Ferroelectric (FeRAM) memory uses a layer of ferroelectric material between two plates to store information. The cell is written by applying a voltage across it, causing dipoles

within the ferroelectric material to align with the applied field[64]. To read the cell, a voltage is applied across it, much the same as when writing. If the read pulse is of the same orientation as the write pulse nothing will happen, but if it's the opposite direction, applying a voltage causes a small pulse of current as the dipoles change orientation. Unlike all the other technologies discussed in this review, this destroys the data stored in the cell. Because FeRAM uses ferroelectric material rather than ferromagnetic, resistance to magnetic fields is superior to MRAM. The endurance is substantially worse though, as the previous writes to a cell can cause it to develop a preferential polarisation. An alternative to FeRAM is FeFETs[42], which use the ferroelectric material in place of the gate insulator of a MOSFET. FeFETs operate in a very similar manner to flash cells, although they store their state as a dipole moment rather than static charge.

Molecular memory operates in a similar manner to FeFETs, but uses a chemical reaction instead of a ferroelectric layer to store the charge that controls the channel of the FET. This is the least explored of the memory technologies in this review but despite this, triple level cells have already been demonstrated[44] using In_2O_3 nanowires coated with Fe^{2+} -terpyradine.
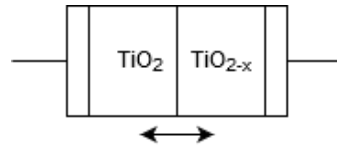
FIGURE 2.2: Diagram of a TiO_{2-x} memristor.

Memristors[65] (RRAM or ReRAM) are two terminal electrical devices with a non-linear current/voltage relationship that is dependant on the charge that has passed through the device. The memristor was first identified in theory in 1971[66] as a fourth fundamental circuit element, described by the equation $M = \frac{d\phi}{dq}$. A physical approximation of such a device was not identified as such until 2008[67], although the exact classification of this device is disputed[68]. This ReRAM device was constructed from a bi-layer of TiO_2 and TiO_{2-x} (Fig. 2.2), and then exposed to high voltage to form conductive defects[69] such as metal filaments in a process called electroforming. The oxygen vacancies in the TiO_{2-x} layer act as p-type dopants, and drift in the applied electric field. This changes the thickness of the conductive and insulating layers, with the conductive layer acting on the filament in a similar manner to the wiper on a rheostat. Memristors based on other materials have been demonstrated[70], some of which remove the requirement to electroform the pristine devices[71]. The low power and high speed of switching, the high level of CMOS compatibility, and granularity of control, make them an attractive choice for low power memory and signal processing tasks. This type of memory is already finding applications in fields relevant to this project, such as a neural signal amplifier using memristors as integrating elements[11].

2.3.1 Crossbar Arrays
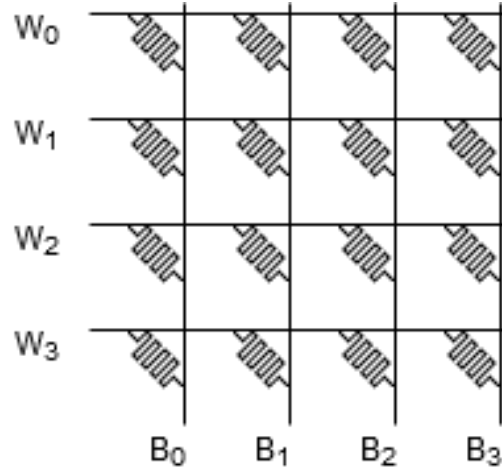
FIGURE 2.3: Diagram of a small memristor crossbar array.

Crossbar arrays are a topology used for high density array designs. Memory arrays with similar structures date back as far as 1947[72], but the fundamental concept behind the crossbar remains relevant today. In a crossbar array the devices are arranged in a grid, with the top electrode of each device connected to a horizontal line and the bottom electrode connected to a vertical line (Fig. 2.3). This allows for an array of x^2 devices to be controlled by a system with x terminals, making the size of the device the limiting factor and permitting very high density arrays. As memristors are two terminal devices, with a fairly simple layered structure this topology is an obvious choice for memristor arrays, where the devices can be formed between two metallic layers using common processes. In addition to the obvious applications of non-volatile memory, this topology can also be used to implement vector-matrix multiplication. The voltage on horizontal lines act as the vector input, and the resistive state of the devices act as the matrix values, with the current on the vertical lines as the output. As vector matrix multiplication is a fundamental part of neural network operation, this makes crossbar arrays ideally suited for implementing neuromorphic systems[73]. While versatile, this topology has significant limitations, as the non-zero resistance of the access lines results in parasitic currents passing through unselected devices[74]. It is possible to mitigate these sneak currents by adding diodes or transistors that block these paths, but doing so comes at the cost of area, complexity, and voltage constraints.

2.4 Instrumentation

Measurement of memristor characteristics is usually done using an instrument called a source measurement unit (SMU). An SMU is an instrument that can set a precise voltage/current and measure a precise current/voltage at the same time. While the

state of a memristor should change as current passes through it, in practice a memristor will retain its state until the voltage or current passes a threshold[75]. This allows for the state of the device to be measured at low voltages/currents. This is not a novel requirement, and many other groups have designed systems to this end. Wust, D. et al. developed a field programmable gate array (FPGA) based memristor prototyping environment[76], but with a maximum theoretical resolution of 740pA, this system cannot deliver more detailed characterisation tasks. Berdan, R. et al. implemented a microcontroller-based advance testing system for memristor devices[77], but the parallelism is limited. Wang, Y. et al. presented a high-speed driving system for phase change memory devices[78], with pulse width as narrow as 500ns. However, this work only has a driver side. Other works such as Merced-Grafals, E. et al. applied commercially available device analysers[79], which have limited channel numbers as well as parallelism. Such a gap in capability calls for the development of a new instrument with the parallel SMU capacity to operate large numbers of memristors and multiple analogue circuits simultaneously.

2.5   Summary

The demands of modern neurological medicine are not yet met by the information processing systems currently available. This project seeks to develop a hardware template matching system suitable for use with neurological spike signals, using ReRAM devices. To this end, this project will also seek to develop an instrumentation platform to support this project and others.

Chapter 3

Mixed signal parallel instrumentation

3.1 Introduction

This chapter covers the development and performance of the instrument used in this project (Fig. 3.1). Section 3.2 discusses the problem of instrumentation. Section 3.3 covers the design history of the subject of this chapter. Section 3.4 outlines the requirements of the system. Section 3.5 then describes in detail the design of the circuitry, followed by experimental demonstration of its performance in sections 3.6.1, 3.6.2, 3.6.3, 3.6.4, and 3.6.5. Section 3.7 reviews the performance against the specification and the predecessor system.

3.2 Instrumentation

The development of new circuits and devices relies upon the foundation of a wide array of instruments, from small components such as instrumentation amplifiers and data converters to full systems such as oscilloscopes and signal generators. These instruments define the limit of what research can be pursued, and as such the development of suitable instrumentation becomes a prerequisite tasks for projects such as this one. In this case, the limiting factor in existing instruments is channel count. Existing instruments are not without capability; the ArC ONE[80] has many channels and can produce the pulses commonly used to write to memristors[81], but only writes and reads from one pair of channels, with the rest operating in an inflexible biasing regime. More capable instruments, such as the Keithley 4200[82] have incredible accuracy and precision, but in most configurations lack the channel count required to test anything more complex than transistors and other three terminal devices. Higher channel configurations of

the 4200 can be assembled, but lack in timing precision and are as bulky as they are expensive. Other systems such as the Analog Discovery 2[83] used by the Knowm memristor characterisation platform[84] simply lack the channel count and precision of a dedicated characterisation instrument. The analogue and mixed signal circuits considered in this project call for a new instrument to act as the platform for the development. Given the ubiquity of the ArC ONE within the research group it seems reasonable to aim for a similar form factor and performance, as the parallelism that will be required for this project could potentially serve the needs of many other research projects, particularly those utilising crossbar arrays commonly found in those of the research group.

3.3   ArC TWO
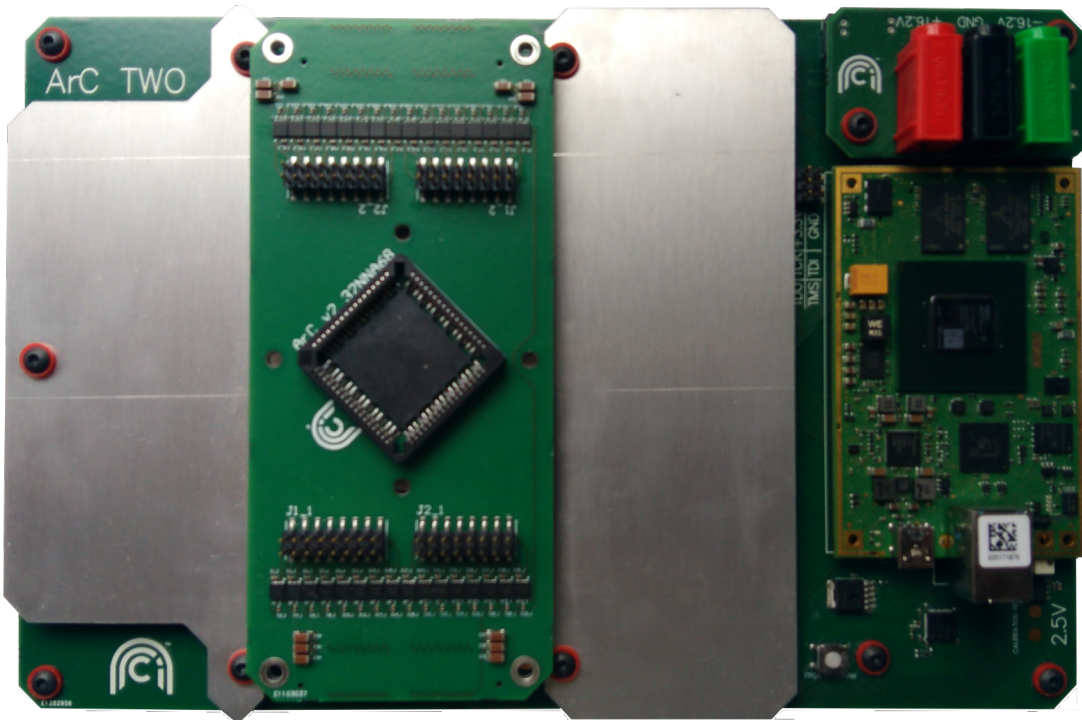
FIGURE 3.1: Photograph of the ArC TWO.

The development of this instrument proved troubled, as the complexity practically makes it a research project unto itself. Development began with a reduced prototype. With half the number of planned channels, this design allowed for the testing of different structures for the channel circuitry and experimentation with operating modes. More detail can be found in Appendix A. The prototype was followed by a full scale version (Appendix B) including almost all planned features. While capable of most of the requirements, this version was plagued with FPGA problems and thermal issues. A second version (Appendix C) addressed the thermal issues, while also solving a number

of other minor problems. Persistent FPGA issues and supply chain inconsistency forced another redesign. This last version (Appendix D) resolved the remaining issues and went into limited production. This instrument was developed in collaboration with other researchers. More information on their contribution to the FPGA configuration and software can be found in Appendix F.

3.4 Specification

As the instrument is intended to be a direct upgrade and replacement of the ArC ONE[80], it should replicate the functionality of that instrument and be capable of operating similar crossbar arrays. To permit the broad range of circuit tests that instrument cannot a high degree of flexibility is required, perhaps implemented as a large number of parallel SMUs. The goal of the broader research project is to develop low-level circuit designs to facilitate spike sorting. To this end, instrumentation developed in support of this should operate in and around the typical range of memristors and integrated devices. Memristors typically operate below 10 V with currents between 1 nA and 1 mA, and the low voltage integrated devices suitable for low power design have similar ranges. As such, each SMU channel should be capable of measuring currents between 1 nA and 1 mA, and voltages between 1 mV and 10 V. A further requirement for memristors is the need for short pulse generator circuits, as fast pulses are useful in writing to such devices[81]. The pulse drivers should be capable of producing pulses with a minimum pulse width of less than 100 ns within the operating range of the SMUs. The mixed signal nature of the planned tests required the presence of supporting digital circuits, for control of and communication with device(s) under test (DUTs). While this will most likely need only a small range of typical logic voltage levels, it may be useful to have some capability for circuits with unusual voltage levels or ground references.

3.5 Design

To fulfil the parallel operation requirements of the specification, the new instrument needed a full set of read and write hardware on every channel. This placed significant constraints on the quiescent current, with regards to both the power consumption and component count of a single channel. Therefore, the topology used needed to be quite simple.
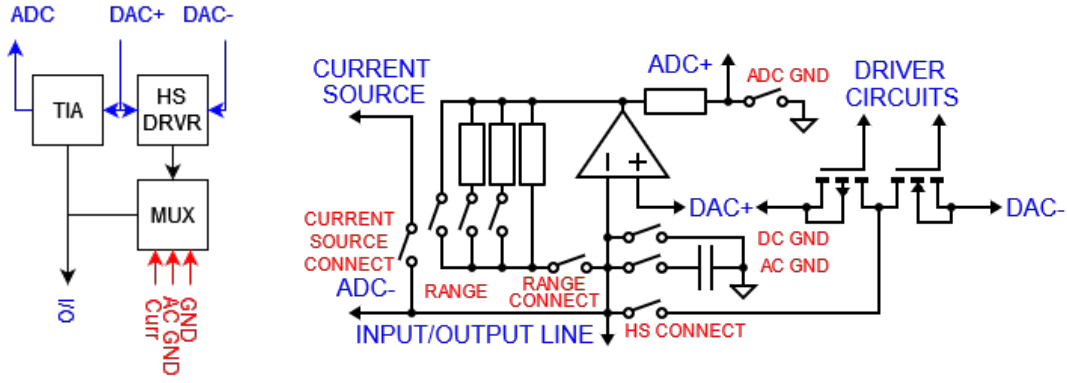
FIGURE 3.2: Left: A system-level schematic of the channel. Right: A component-level schematic of the channel architecture. Signals are labelled in blue, switches are labelled in red.

The individual channel topology (Fig. 3.2) used in the instrument is based around a transimpedance amplifier. The transimpedance amplifier is referenced to that channel's positive DAC output, allowing the virtual ground at the input of the amplifier to set to other voltage levels. The amplifier has three geometrically spaced ranging resistors ($820\,\Omega$, $110\,\mathrm{k\Omega}$, and $15\,\mathrm{M\Omega}$), plus a short circuit switch that turns the amplifier into a voltage buffer for the DAC. For current sensitive operation outside of current sensing, the ranging resistors could be disconnected from the input node. The ADC used in this design is a differential ADC. Its input terminals are connected to the input node (through a low input current bias buffer) and the output of the transimpedance amplifier, allowing it to directly read the voltage across the ranging resistor, and by extension the current flowing into the input node. When the ADC shorting switch is closed (ADC GND on Fig. 3.2) the ADC instead measures the voltage at the measurement node. Additional functions, such as the high speed driver and the unified current source are also connected to the measurement node.
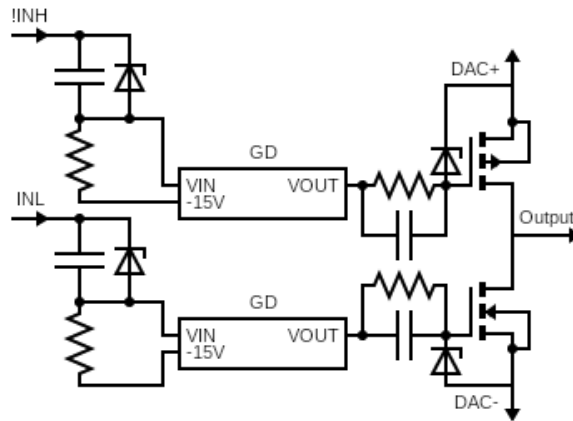
FIGURE 3.3: Concept schematic of the gate driver circuit.

Designing an arbitrary pulse generator proved a significant challenge, as there exists no IC available that can produce the voltages or edge rates called for by the specification. Instead, discrete MOSFETs were used to form a push-pull stage (Fig. 3.3). This output stage was supplied by a pair of DAC outputs. Since the pulse driver is never expected to operate simultaneously with the transimpedance amplifier, they share one of the DAC channels. To drive the gate of this floating MOSFET, a high voltage gate driver is used to produce a drive signal that spans the full supply range. This signal is trimmed using transient voltage suppression (TVS) diodes to protect the MOSFET gate, using a resistor/capacitor bias network to provide a reasonable bias current for the diode without compromising transient speed. As the input of the gate driver is referenced to the negative supply rail, a zener diode based level shifter is used to allow the circuit to be controlled by the FPGA output, which is referenced to ground. This level shifter uses another resistor/capacitor/zener network to effect a 15 V step between the control signal and the gate driver. With a large number of the GPIO pins dedicated to the serial communications, the high speed driver signals needed to be grouped. This was achieved using analogue multiplexers, as digital multiplexers with serially addressed, arbitrarily selectable outputs weren't available.
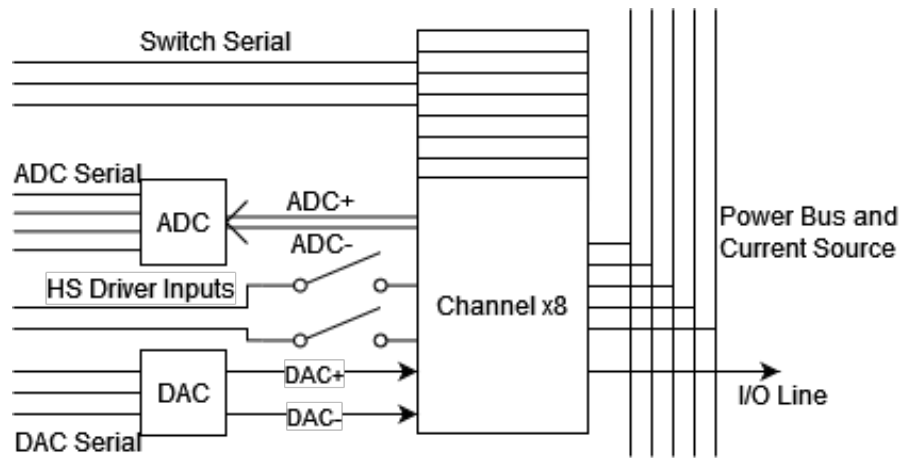
FIGURE 3.4: Concept schematic of the channel cluster, showing connections for one of the eight channels of the cluster.

To fit the necessary number of channels in a reasonable space, the channel layout was stretched out and place in a side-by-side arrangement. Power supplies rails were then run perpendicular to the channels on buried layers. Two of the resulting strips of channels were then placed back-to-back, with the grouped serial busses routed in the space between, with an FPGA module at one end to operate the many serial lines (Fig. 3.6). To simplify the control of the channels, they were grouped into clusters of eight (Fig. 3.4); each cluster being served by a single DAC, ADC, and analogue switch daisy chain (Fig. 3.5).A daughter-board, containing the test socket, connects to the board through mezzanine connectors at either edge of the instrument. Not only does this allow for compact channel placement and routing, but it also allows for the test

socket to be swapped out for alternative daughter-boards with different sockets or extra circuits.
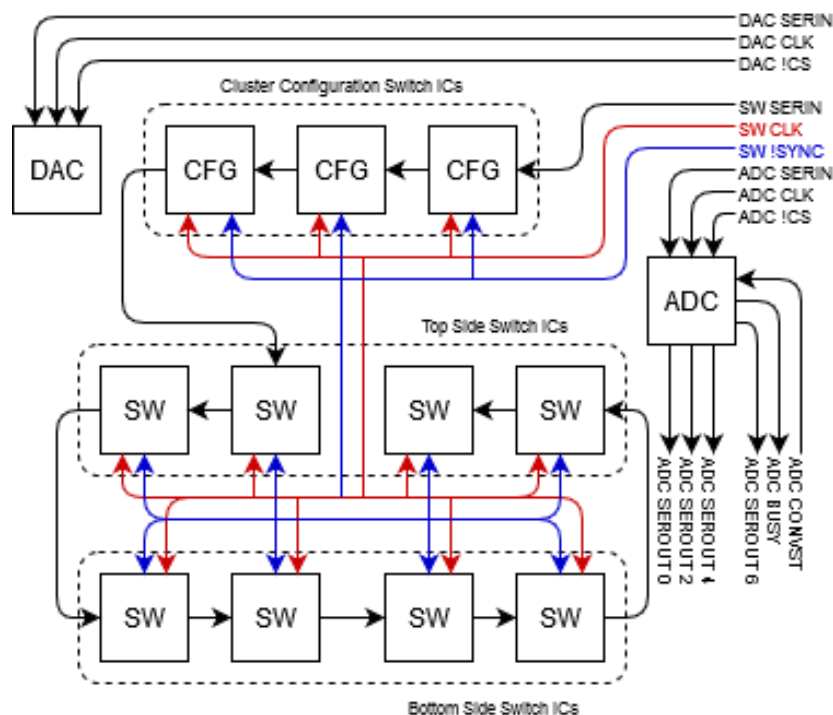
FIGURE 3.5: Diagram of the serial connections within the channel cluster, showing the daisy-chain SPI bus that weaves through all switch ICs in the cluster.
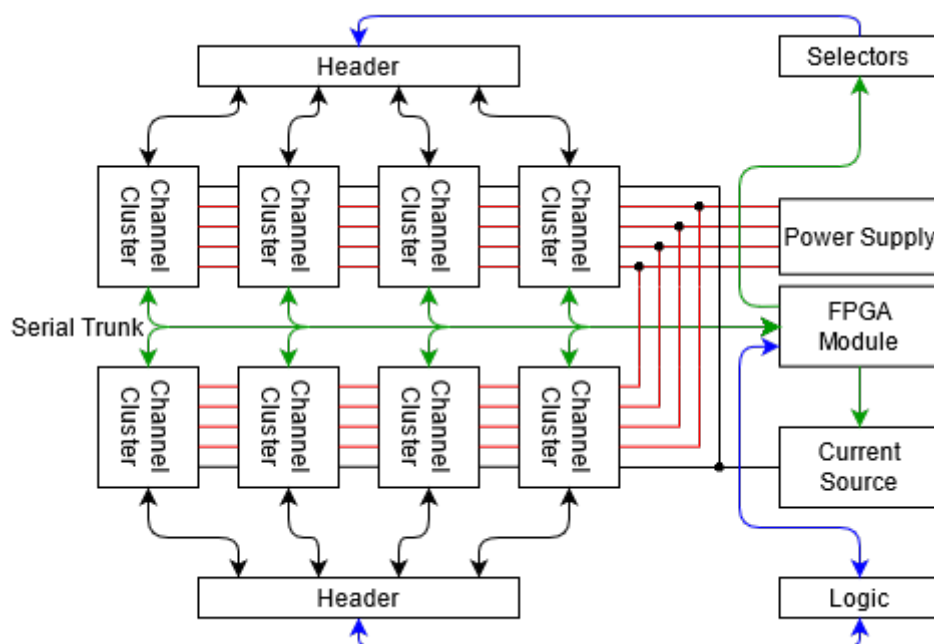
FIGURE 3.6: Block diagram of the full instrument, showing various systems as arranged on the PCB. Analogue signals are shown in black, parallel digital signals in blue, serial signals in green, and power in red.

As the current control required for sourcing current proved impractical to fit into the channel architecture, each channel was instead given a connection to a single current source for the entire board. The unified current source was designed such that it can be reconfigured to either source or sink current (Fig. 3.7). To achieve this, either an NMOS or PMOS can be connected to the ranging resistor array and setting the bias and reference voltages appropriately. As with the TIA the resistors in the range array are roughly spaced geometrically, with values of $51\,\Omega$, $220\,k\Omega$, $3.6\,M\Omega$, and $68\,M\Omega$. The gap in the resistor array that would be occupied by a $1-10\,k\Omega$ resistor is instead occupied by a $20\,k\Omega$ digital potentiometer, to provide capability for current compliance operations. The feedback of the bias op-amp adjusts the conductivity of the selected MOSFET until the current being driven causes a voltage across the selected resistor to be equal to that between the bias and reference voltage. By selecting the PMOS and setting a reference voltage above the expected output voltage, the circuits sources current. By selecting the NMOS and setting a lower reference voltage, the circuit sinks current. The advantage of this circuit is that there are fewer parasitic current paths than a design with separate source and sink circuits. To protect the gates of the MOSFETs from over-voltage condition when attempting to drive current into an open circuit, an TVS diode was added to limit gate source voltage.
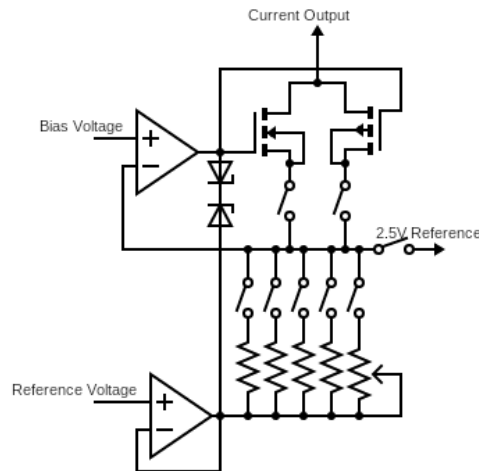
FIGURE 3.7: Concept schematic of the current source.

To accommodate unusual logic levels such as the control of the selector equipped crossbar arrays, a bank of configurable level digital outputs was added to the design (Fig. 3.8). To allow this bank to be controlled with few GPIO pins, serial controlled analogue multiplexers were used in conjunction with pull-down resistors to implement this circuit. A pair of high current op-amps act as buffer to the DAC references used; one for the high output level and one for the low. The analogue switch on each channel can be closed to bring the output to the high level, otherwise the pull-down resistor brings the output to

the low level. While slow and lacking in drive strength, this configuration allows for any value for the high and low levels in the $\pm 13.5\,\text{V}$ operating range, including low levels that are higher voltage than the high level.
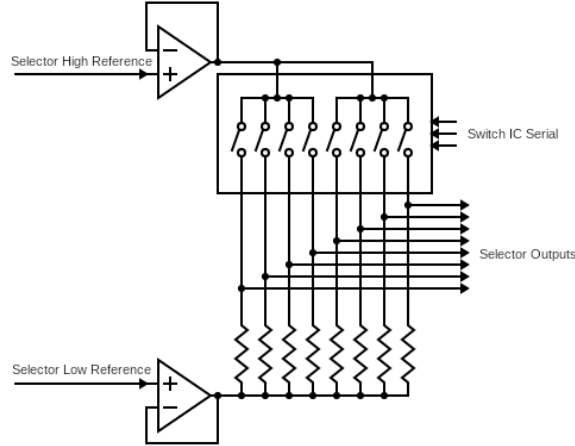
FIGURE 3.8: Concept schematic of one quarter of the selectors bank.

For more conventional digital tasks, such as control of daughterboard infrastructure and mixed signal test subjects, a second set of digital channels were added to the design. Unlike the selectors, this digital bank operates in parallel, and is implemented with level shifter ICs connected directly to the GPIO pins of the FPGA module. This limits this digital bank to ground referenced signals with a high level between $1.8 - 5\,\text{V}$, but allows the pins to act as either inputs or outputs, with speeds up to $100\,\text{MHz}$.

3.6 Performance

To assess the performance of the instrument, experiments were conducted to determine the noise floor, crossbar array read accuracy, and pulse generator performance. While the instrument has many characteristics, these ones are the most critical to its intended role.

3.6.1 Noise Floor

To measure the noise performance in current measurements, a channel was configured as a TIA with a reference of $-0.5\,\text{V}$ and a resistor connected between the channel and ground. This forced the channel to sink and measure a current of $I = 0.5/R_{TEST}$. 10k points with the default 32 sample averaging were collected. This procedure was conducted with three resistances to force the instrument to automatically select each resistor range: $2.2\,\text{k}\Omega$ to select the $820\,\Omega$ range, $16.4\,\text{k}\Omega$ to select the $110\,\text{k}\Omega$ range, and
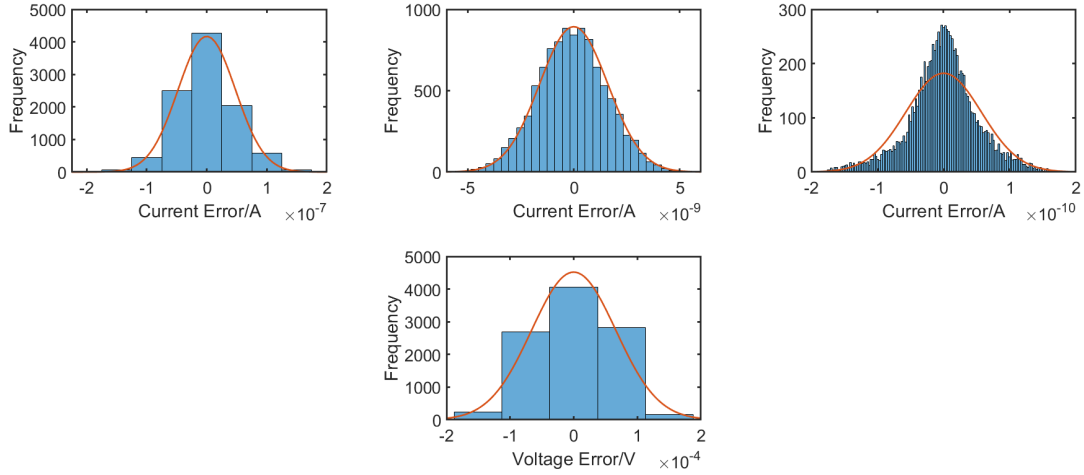
FIGURE 3.9: Histograms showing noise characteristics of the various modes of measurement. All histograms have one bin per ADC code with widths of 47.6 nA, 355 pA, 2.60 pA, and 78.1 µV respectively. Top: 10k point histograms of current read-out tests, overlaid with Gaussian distribution estimates. Top left: 820 Ω TIA range yields $\sigma = 48$ nA. Top centre: 110 kΩ TIA range yields $\sigma = 1.6$ nA. Top right: 15 MΩ TIA range yields $\sigma = 57$ pA. Bottom centre: 10k point histogram of a read-out voltage error test (V=GND), overlaid with Gaussian distribution estimate of $\sigma = 65$ µV.

open circuit to select the 15 MΩ range. To measure the noise performance in voltage measurements, the channel was grounded and 10k points with 32 sample averaging were collected. The results were then plotted on histograms and the standard deviation calculated. In the 820 Ω range, a value of of $\sigma = 48$ nA was obtained, approximately 1 LSB. In the 110 kΩ range, a value of of $\sigma = 1.6$ nA was obtained, roughly 5 LSB. In the 15 MΩ range, a value of of $\sigma = 57$ pA was obtained, roughly 22 LSB. In voltage measurements, a value of of $\sigma = 65$ µV was obtained, approximately 0.83 LSB. In all but the 15 MΩ test, the distribution of the results followed a very clean normal distribution (Fig. 3.9), but the 15 MΩ test had significant low frequency tails. This suggests that there are two or more sources of noise, with slightly different distributions. To test the possibility that this second source of noise was radiated mains interference, jumper wires were connected to the channel input to add to the antenna effect of the PCB traces. This drastically increased the size of the tails. Demounting the daughterboard to remove the antenna effect of its traces reduced the tails. Together, this suggests that radiated mains interference was the cause. Without the tails, the 15 MΩ range has a distribution of $\sigma = 38$ pA, but as the antenna effect could not be completely eliminated this could not be confirmed.

Using a reasonable worst-case error of 3σ, the proportional error was plotted as a function of measured current, using the equation $Err = 100(I_{ERR}/I_{MEAS})$ (Fig. 3.10). From this, it can be seen that current measurements of more than 16 nA can be made with 1% accuracy. Measurements above 3.4 nA and 1.7 nA can be made with 5 and 10% accuracy respectively. The calculation suggests that, at a bias voltage of 0.5 V, resistive of devices up to 300 MΩ can be measured before precision starts to degrade. With
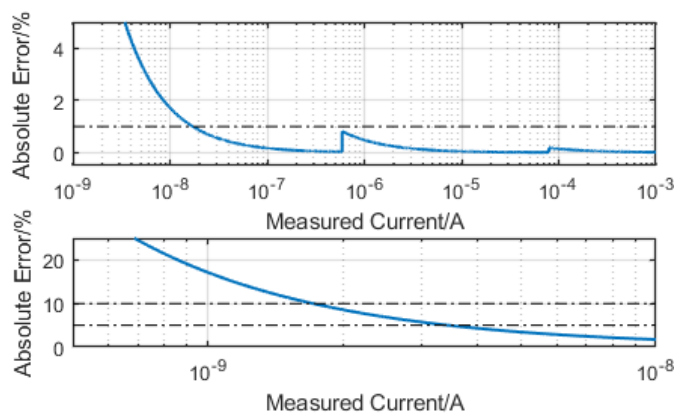
FIGURE 3.10: Graph showing predicted absolute error based on 3σ current noise error, with dotted lines to show 1, 5, and 10% error.

further averaging, it may be possible to push the maximum resistance up to $\approx 1\,\text{G}\Omega$, but averaging is an operation with diminishing returns, which imposes practical limits on this approach. Above $1\,\text{G}\Omega$, it becomes difficult to tell the difference between the resistance of the DUT and the insulation around the devices, so the precision achieved here should be more than sufficient. The effect of changing ranging resistors is clearly visible in the figure as step discontinuities in the error magnitude.

3.6.2   Crossbar Array Read Accuracy
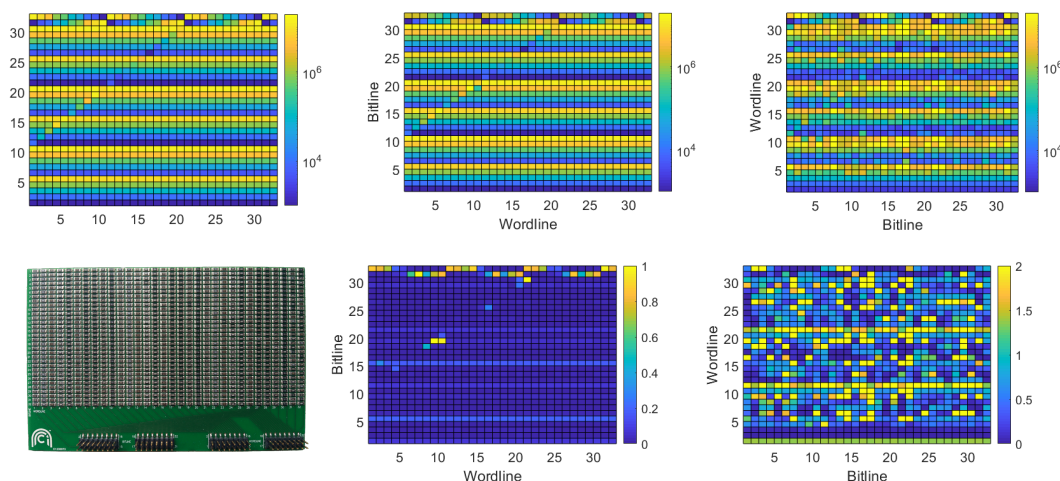
FIGURE 3.11: Array read operations for a 32x32 resistor array. Top left shows the array as designed, with resistors ranging from $1\,\text{k}\Omega$ to $15\,\text{M}\Omega$. The colourbar is scaled from $1\,\text{k}\Omega$ to $20\,\text{M}\Omega$. Top centre shows the array as read in columns and bottom centre shows the proportional error. Top right shows the array as read in rows and bottom right shows the proportional error. Bottom left is a photograph of the test array.

To assess the array read performance, a 32x32 crossbar configured array of SMD resistors was connected to the daughterboard headers through ribbon cable. The resistors were then read by biasing one of the lines and reading the current from the perpendicular

lines while the unbiased lines were grounded. Results were collected by both biasing the rows and reading the columns, and biasing a column and reading the rows. In the read procedure used here, the lines that are being biased or grounded are referred to as wordlines and the lines that are reading current are referred to as bitlines. The array used 1% resistors of $1\,k\Omega$ to $10\,M\Omega$ and 5% resistors of $15\,M\Omega$; its nominal design is shown in Fig. 3.11, Left. The proportional errors in both sets of measurements were then calculated with the following equation: $|(R_{meas} - R_{actual})/R_{actual}|$.

Even small differences in voltage between bitlines can cause non-trivial sneak currents to flow between them if both lines have a low resistance connection to an inactive wordline. The channel-to-channel voltage discrepancy is typically only $500\,\mu V$, but if the ratio between the smallest device on a bitline and the device being read is comparable to the ratio between the read voltage and the mismatch voltage then accuracy will suffer. Our test used a read voltage of $5\,V$, which gives a ratio of 10000. In a configuration where the devices on a bitline are largely of the same value (Fig. 3.11, Centre) the performance is excellent, with 802 of 1024 resistors measured with less than 5% error. Reading from the other direction (Fig. 3.11, Right), the ratio between the largest and smallest devices on most bitlines is 15000. In this configuration, only 171 of 1024 resistors were measured with less than 5% error and 758 measured with less than 100% error.

For this experiment, the DACs on instrument were manually trimmed to minimise channel-to-channel offset, but the offset was measured with the uncalibrated onboard ADCs. The ADCs have a maximum rated zero-scale error of $\pm700\,\mu V$ (typ. $\pm160\,\mu V$). As such, the channel to channel offset voltage may be higher than expected. Software control of ADC and DAC calibration should be able to mitigate this issue. Since the resolution of a voltage read operation is greater than the DAC resolution, it may be possible to measure the channel to channel offset and use deconvolution to obtain more accurate values, but this is beyond the scope of this project. The predecessor to this instrument exhibited superior array read accuracy, although it should be noted that the array used in that test had far fewer high resistance devices and no devices above $1\,M\Omega$. The difference is most likely due to it using a single biasing circuit for all unused wordlines and bitlines. This results in smaller line-to-line offset, with only the offset voltage of a single op-amp ΔV between the active and inactive lines. While the array read performance of the instrument developed here is not spectacular, it compensates for this by having much lower access resistance and superior current measurement precision. The array read performance is expected to improve as the software improves.

3.6.3 Pulse Performance

In contrast to the array read operation, the pulse generators performed well, producing well defined pulses of specified length and amplitude (Fig. 3.12). These traces were capture using a high speed oscilloscope with a $6\,GHz$, $1\,M\Omega$ active probe. The pulse
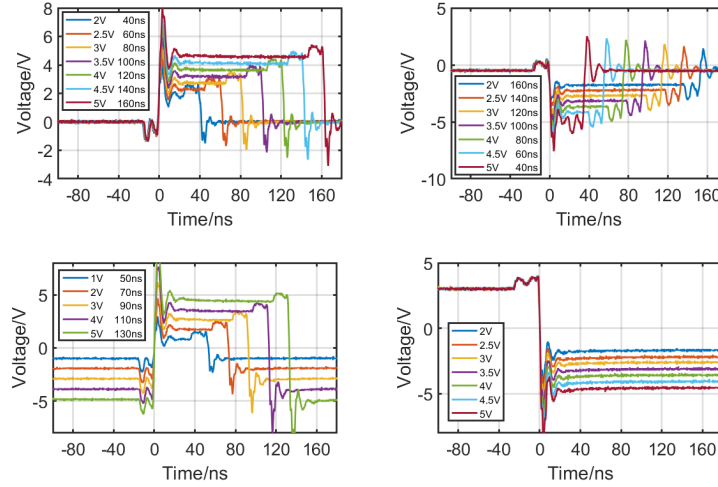
FIGURE 3.12: Oscilloscope captures of a variety of pulses produced with the high speed pulse generator. Top left: +VE pulses starting at $0\,\mathrm{V}$. Top right: -VE pulses starting at $-0.5\,\mathrm{V}$. Bottom left: +VE pulses symmetrical around $0\,\mathrm{V}$. Bottom right: Continuous pulses starting at $3\,\mathrm{V}$.

generators were otherwise unloaded. The pulses are produced by driving the either the PMOS or the NMOS in the push-pull pair, with $10\,\mathrm{ns}$ dead time during transitions in which both FETs are off. This dead time produces a notable step, just before and after a transition. This appears to be due to capacitive coupling between the output of the driver and the gates of the FETs. To account for the wide range of values the pulse circuit can be set to, the ICs driving the gate of the FETs have a very wide output swing. There is also significant overshoot of around 50%, although this is obscured by the aforementioned step in the case of small transitions. There is an RC snubber circuit in the design to mitigate this overshoot, but the capacitor in the snubber must be significantly smaller then the decoupling capacitors connected to the DAC outputs, else the charge required to fill the snubber would deplete the decoupling capacitors. This would cause much greater errors in the voltage level after a transition than the $\sim 5\%$ seen here.

3.6.4 Device Characterisation

To expand upon the demonstration of the instruments capability, the ArC TWO was used to perform IV sweeps of a number of simple devices; a $10\,\mathrm{M\Omega}$ resistor, a 1N4148 diode[85], and a 2N7000 MOSFET[86] (Fig. 3.13). For the resistor, the devices was connected between two channel; one biased to a test voltage and the other biased at $0\,\mathrm{V}$ and used to measure current. The measured current-voltage relationship was plotted and found to be entirely unremarkable. The diode was measured in a similar manner, although the data above $1\,\mathrm{V}$ in forward bias was discarded as the current through the diode exceeded the rated current of the TIA. In such a saturated sate, the TIA can no longer provide enough current to bring the connected terminal to the specified voltage,
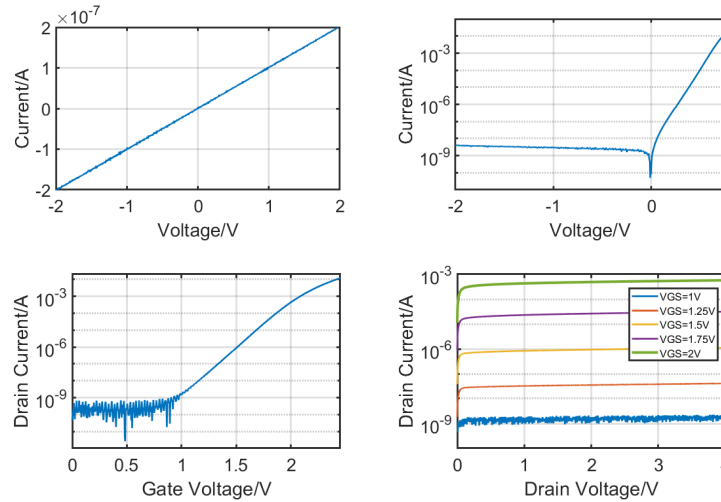
FIGURE 3.13: IV characteristics of a small selection of components. Top left: IV sweep of a $10\,\text{M}\Omega$ resistor. Top right: IV sweep of a 1N4148 diode, from $-2\,\text{V}$ to $0.75\,\text{V}$. Bottom left and right: Gate terminal drain terminal sweeps of a 2N7000 nFET.

invalidating the data. Despite this, the measured behaviour is very nearly ideal for a typical PN junction. Between $0-1\,\text{V}$ the relationship is almost perfectly exponential (presenting as linear on the logarithmic scale used here). In reverse bias the reverse leakage current is readily apparent, despite only being a few nA. The MOSFET has three terminals, rather than the two of the other devices so two IV sweeps were made; one of the drain current as a function of gate voltage, and one of the drain current as a function of drain voltage. As with the diode, the measurements of the MOSFET are similarly excellent. The measurements are noticeably noisier than those of the diode, but as the noise occurs as current levels similar to those measured in the diode test this suggests that the 2N7000 is a noisier device than the 1N4148, rather than an issue with the instrument.

3.6.5 Mixed Signal Testing
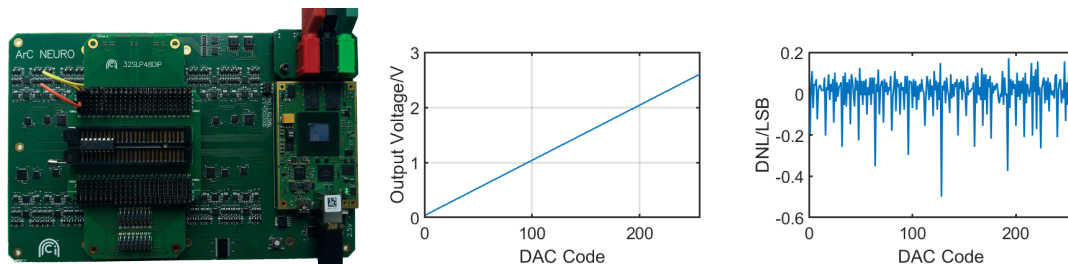
FIGURE 3.14: Results from an automated test of an AD558J DAC (Left) in $2.56\,\text{V}$ range. Centre shows the output from code 0 to code 255. Right shows the normalised differential non-linearity.

While quite capable of DC characterisation of devices, the intended role of the instrument includes more complex tests of ICs with digital and analogue functionality. To

demonstrate this capability, the ArC TWO was used to measure the nonlinearity of an AD558J DAC (Fig. 3.14). The digital IO pins were used to command a DAC output at each code, which was then measured with an analogue channel. The result was plotted, showing the expected linearity of reputably sourced parts. A more detailed analysis of the linearity shows the error not visible in the initial plot, with higher errors from more significant bits of the DAC code.

3.7   Conclusion

The instrument fulfils the requirements laid out at the start of this chapter, demonstrating the operation of its SMU channels in a variety of tasks. It compares favourably with its predecessor (Tab. 3.1), exhibiting the flexibility necessary to support future research.

	ArC ONE[80]	ArC TWO[14]
Parallel read	N	Y
Parallel write	N	Y
Channel count	32R+32W	64R/W+64D
Min. chan. current	$\pm 1\,\text{nA}$	$\pm 100\,\text{pA}$
Max. chan. current	$\pm 5\,\text{mA}$	$\pm 12\,\text{mA}$
Current sample rate	$50 - 1000\,\text{Hz}$	$833\,\text{Hz}$
Voltage resolution	$3/24\,\text{mV}$	$78\,\mu\text{V}$
Voltage sample rate	$200\,\text{kHz}$	$100\,\text{kHz}$
Min. pulse width	$90\,\text{ns}$	$40\,\text{ns}$
Pulse volt. range	$\pm 12\,\text{V}$	$\pm 13.5\,\text{V}$
Power	$4.5\,\text{W}$	$20\,\text{W}$

Table 3.1: Comparison between this instrument and its predecessor.

Chapter 4

Template matching using RRAM configurable circuits

4.1 Introduction

The last chapter covered the development of the ArC TWO, describing its design and characterising the performance of the final instrument. This chapter will discuss the operating theory behind the hardware implementation of template matching considered in this project, along with three designs for RRAM adjustable window comparators. The first design is largely a previous work[12], included both as a point of reference and as the work from which the two other designs were developed.

Section 4.2 discusses the methodology of template matching. Sections 4.3, 4.4, and 4.5 discuss the three designs considered in this chapter. Each section contains a diagram of the circuit in question and presents SPICE simulations of the expected operations. As a PCB model of the previous work was suitable for tests of the circuit in section 4.3, that section also includes measurements collected with the ArC TWO. Section 4.6 compares the presented designs, and discusses the selection of a design for further development.

4.2 Template Matching

The approach to spike classification considered in this project begins with the detection of an event. Once an event is detected, an array of sample and hold circuits are used to capture the amplitude of the waveform at regular intervals. These held samples are then tested against an array of window comparators to determine if the waveform falls within a the expected range for a known template, as shown in Figure 4.1. If enough cells in a template report hits, the detected event can be considered to match the template, and the origin of the neural spike identified. This method uses a large number of window
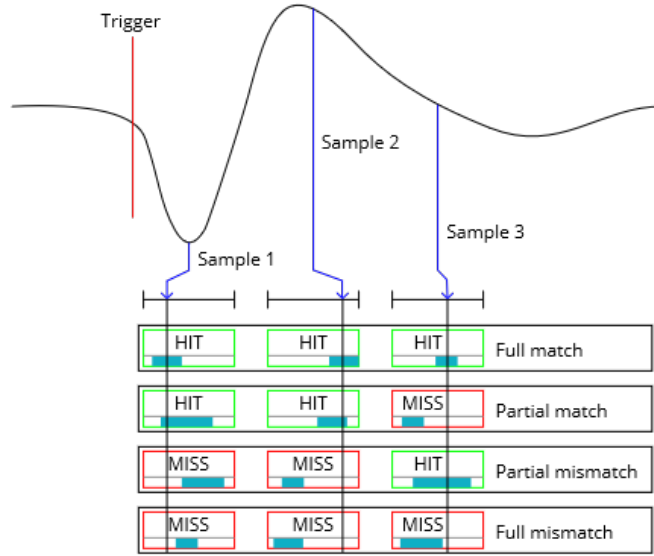
FIGURE 4.1: An example of template matching with window comparators. This diagram shows three samples captured from a waveform being tested against four templates.

comparators; with tens of samples per template and tens of templates per channel, the comparators will make up the majority of the power consumption and footprint of an integrated design.

Discuss template matching and existing window comparators

4.3   Inverter TXL

The first comparator design builds upon a past work on analogue reconfigurable logic gates[12]. These circuits source current when the input signal falls within a small window. By accumulating the charge from multiple circuits in a capacitor, the number of samples from an input that match a template can be assessed. These template pixel (TXL) cells can be tuned by changing the resistance of memristors in the cell.

The TXL cell is formed from a pair of CMOS inverters and a MOSFET current mirror (Fig. 4.2). The M1 and M2 memristors cause source degeneration in the transistors of the first inverter, altering the threshold of the logic gate as a logarithmic function of the ratio between M1 and M2. The bias current of the second inverter rises as the as the input voltage rises to the crossover point and falls again past that point. This produces a peak in the bias current that can be positioned by setting the memristors in the first inverter stage. To limit the maximum current and produce a window of consistent amplitude, a resistor is added in between the transistors of the second inverter. This current limits the inverter and caps the peak bias current, producing a window with controllable position.
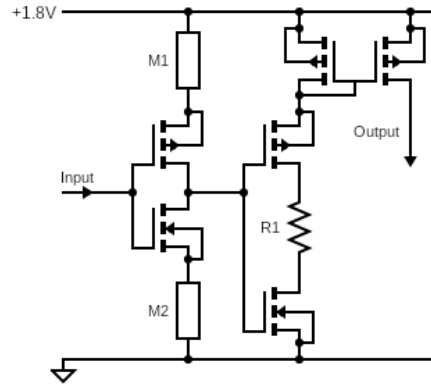
FIGURE 4.2: Circuit diagram of a TXL cell.
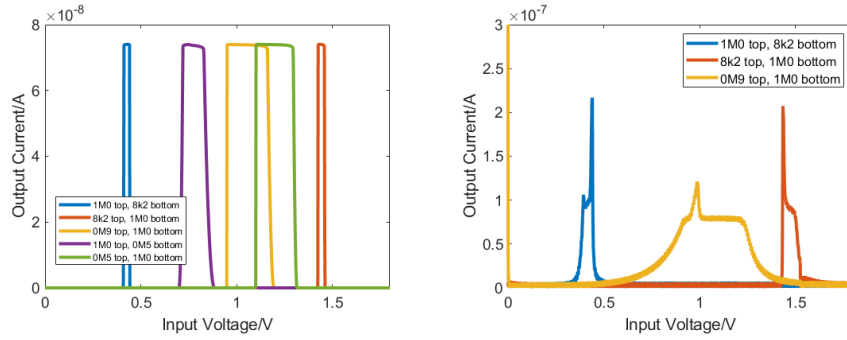
FIGURE 4.3: Simulated and real behaviour of the TXL cell, showing the output current at several settings of the top (M1) and bottom (M2) memristors. Left: simulated behaviour. Right: data collected from tests on a PCB model.

While this approach shows promise, the selectivity of the circuit is best when the window is set to a voltage near that of the supplies. The window is formed from the overlap of the transfer characteristics of the input MOSFETs. To achieve a window near the middle of the supply range, the source degeneration on both FETs must be significant. The combined increase in source degeneration causes the gain of the first stage inverter to drop dramatically, causing the output of the first stage to pass through the crossover point of the second stage more slowly. Thus, setting the window near the middle of the supply range significantly degrades the selectivity (Fig. 4.3), with both a wider window and less well defined window thresholds. This is not necessarily a significant issue, as the signal can be conditioned in the pre-amplifier stage to fall inside a window set near the supplies. Because the window of this design is always centred on the crossover point of the input inverter, the supply current will be high across the entire window, wasting considerable current.
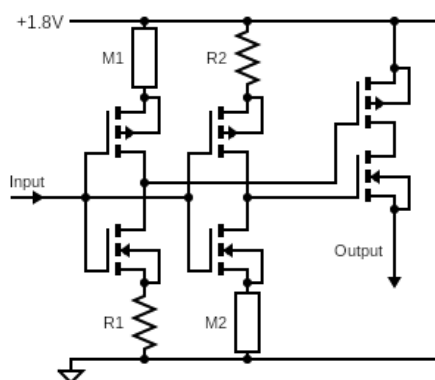
FIGURE 4.4: Circuit diagram of the Split TXL design.

4.4   Split TXL

To separate control of the high and low thresholds, the input inverter can be Split into two inverters (Fig. 4.4), with one memristor on each gate replaced with a static resistor to allow for resistance ratios above and below 1. This allows each gate to produce a step function with a controllable threshold. The outputs of the two inverters control the gates of a PMOS and an NMOS, allowing current to be sourced at the output inside a window (Fig. 4.5).
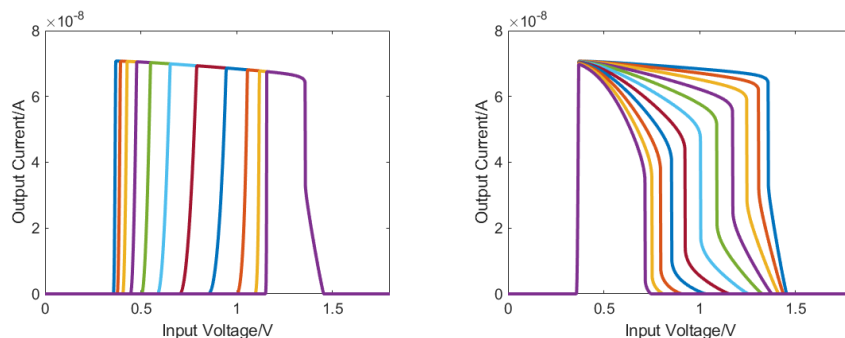
FIGURE 4.5: Simulated input output relationship of the Split TXL design. R1 and R2 are fixed at $1\,\text{M}\Omega$ while one of the memristors is swept from $100\,\text{k}\Omega$ to $10\,\text{M}\Omega$. Left: sweep of M1. Right: sweep of M2.

This approach allows for much greater control of the window width than the inverter TXL design, without compromising the control of the position of the window or the transistor count. The drawbacks are that the design calls for high value resistors and that the controlability of the thresholds is not the same for the high and low thresholds due to NMOS source degeneration at the output. The inverter controlled by M1 feeds into the gate of the output PMOS. With the PMOS against the positive supply it acts as a switch, resulting in a relatively decisive lower threshold. The M2 controlled inverter feeds into the gate of the output NMOS, which in this configuration acts as a source follower. As such, it passes the transfer characteristics of the skewed inverter

almost transparently to the ouptut. While the inverter is current limited, it is not entirely current starved, so the upper threshold has significant regions either side of the switching point where the window is not as clearly defined. The resulting asymmetry is not a critical issue, but the large footprint of integrated $1\,\mathrm{M\Omega}$ resistors is a much greater hurdle. It may be possible to replace the resistors with other devices that achieve higher resistances in a given area than polysilicon, but this has yet to be explored. In most configurations of this design, the inverters will be at their crossover points at different input values, so there should only be a small range of input values where both inverters pass significant current. However since the high current consumption is centred on the thresholds of the window, it may be possible to limit the input ranges that can cause high current by reducing the value of all resistors and memristors and reducing the time that the cell is powered.

4.5 Capacitive Subtractor TXL
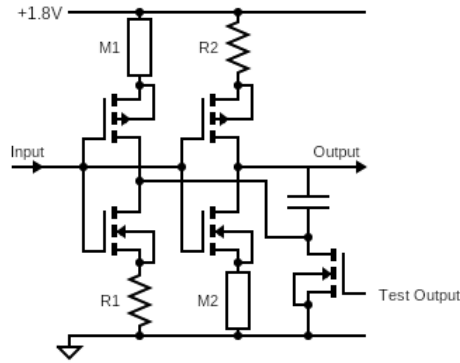
FIGURE 4.6: Circuit diagram of the capacitive subtractor design.

Instead of using a PMOS and an NMOS to compare the outputs of the inverters, a capacitor can be placed between the outputs (Fig. 4.6). When the input is within the window, the outputs take different values, storing a charge in the capacitor. The inverters can then be shut down and the capacitor grounded to test the voltage across it. The easiest way to implement this might be to park the input at the positive supply, and add a second gate to the M2 adjacent transistor to prevent it from conducting while reading.

This gives an input output relationship (Fig. 4.7) that is entirely symmetrical and just as controllable as with the Split TXL, however the output is now a voltage rather than a current. This makes it more difficult to assess an entire array of cells at once, requiring an extra transistor to turn the output back into a current, negating the lower transistor count of this design. In addition to this, the direct conversion of the inverter behaviour to an output causes similar issues to those seen in the upper threshold of the Split TXL,
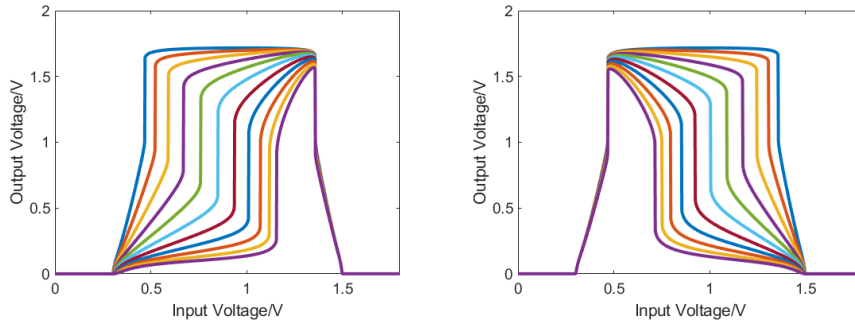
FIGURE 4.7: Simulated input output relationship of the capacitive subtractor design. R1 and R2 are fixed at $1\,\text{M}\Omega$ while one of the memristors is swept from $100\,\text{k}\Omega$ to $10\,\text{M}\Omega$. Left: sweep of M1. Right: sweep of M2.

only now on both thresholds. The capacitor in this topology needs to be larger than the typical gate capacitance, so the energy it stores should be the determining factor in the power consumption. Larger capacitors will give more accurate readings and will easier to design for, but will reduce speed and increase energy used per comparison.

4.6 Conclusions

	TXL	Split TXL	Capacitive TXL
Control	Position	Thresholds	Thresholds
Transistors	6	6	5
Memristors	2	2	2
Resistors	1	2	2
Capacitors	0	0	1
Hit energy	High	Medium	Medium
Miss energy	Low	Low	Low

TABLE 4.1: Comparison of the TXL designs described in this chapter.

To compare the designs discussed in this chapter, relevant performance characteristics such as energy consumption and area should be considered (Tab. 4.1). While the area of silicon substrate used by each design is impossible to calculate at this stage, the number of devices used in each design can be used to estimate the relative are of each design. The energy of operation is also impossible to calculate, but can be estimated based on the number of DC paths in any given situation. The original TXL uses 6 transistors, two memristors, and a static resistor. It has three DC paths: the input skewed inverter, the current limited inverter, and the output path. Because both the input inverter and the current limited inverter are at their crossing points during a hit the current during a hit is exaggerated. The Split TXL uses the same number of transistors and memristors as the original TXL, but uses an extra static resistor. While this gives it a slightly larger footprint than the original TXL, the difference in the design provides separate control of the thresholds of the window. There are also three DC paths in this design: two input inverters and the output path. As neither of the input inverters are at their

crossing points during a hit, the current used is significantly lower than that of the original TXL. The capacitive subtractor has one fewer transistor than the Split TXL, but adds a capacitor. This has little impact on the footprint as capacitors are usually implemented on metal layers above the transistors and can overlap with other devices. Compared to the Split TXL, the capacitive design has no output current path, but the capacitor must be charged/discharged so the current used is similar. The capacitive subtractor has several obvious advantages over the other TXLs. Its independent control of thresholds is an improvement over the original TXL and its internal capacitor allows the assessment of the output to conducted at a significant amount of time after the input. This could allow a design with capacitive TXLs to omit a conventional sample and hold circuit. While this design is of interest, the low capacitance required to make this design fast will result in the output being suppressed by the input capacitance of the instrument being used to measure it, so integrated versions of this topology cannot realistically be tested with available equipment. Recording the output voltage of each inverter in a Split TXL cell should allow for the behaviour of this system to be predicted from tests of the Split TXL design, so the Split TXL will be the preferred design going forwards.

Chapter 5

Split TXL

5.1 Introduction

The last chapter explored the design of window comparators for use in template matching systems. Several designs were considered, with a skewed inverter based design being identified as the best candidate for further development. This chapter will discuss that development, presenting measurements of PCB models of the design with both resistors and memristive devices. It will then investigate the suitability of the design for integration, testing three variants of the circuit, estimating the power consumption. An array of TXL will then be used to demonstrate a simple template matching task.

Section 5.2 outlines the objectives of this design work and section 5.3 describes the circuit under test. Sections 5.4.1 and 5.4.2 present the results of tests on resistor array and memristor controlled PCB implementations of the circuit. Section 5.5.1 discusses the reworking of the design for integration and section 5.5.2 compares the results of this integrated simulation with the physical models. Sections 5.5.3 and 5.5.4 present simulations of the test energy and process variation. Section 5.6 presents the operation of the array of resistor controlled TXLs in a simple template matching task. Section 5.7 will the discuss the suitability of the design in its intended role.

5.2 Objective

The goal of this development is to design and demonstrate a low energy window comparator suitable for use within integrated template matching systems, such as might be found in a neural spike sorting system. Unlike a conventional comparator which responds to an input above a threshold, a window comparator responds to an input between two thresholds. For this comparator to be useful in a template matching application it will need to be tunable, so that an array can be tuned to recognise a specific shape of spike.

This can be done in two ways: controlling the position and width of the window, or controlling the thresholds. The previous work [12] used the former method and ran into difficulties with the transition between hit and miss with wider windows, so this design will pursue independent threshold control. It also used two tuning elements to control a single parameter, with would require two sets of writing infrastructure for each parameter. This design should reduce that to one tuning element per threshold. This method of control should make the control of a template easier, since a hypothetical future control system can set the input at a target level and then adjust a single element until the output changes. For integration with a sample/hold circuit and other related analogue front-end circuits, this comparator will need to have the maximum possible input range. It will also need to be implementable in a similar technology node to existing analogue and digital circuits and not consume excessive amoounts of energy.

5.3   Design
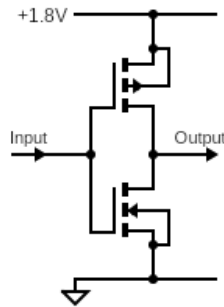
FIGURE 5.1: A diagram of the split TXL circuit.

An inverter (Fig. 5.1) is a simple logic gate that pits the current of a PMOS and an NMOS against each other. When the input voltage is high, the current flowing into the midpoint through the NMOS exceeds the current flowing out through the PMOS, and the loss of charge causes the voltage to drop. When the input voltage is low, the PMOS current is greater and the midpoint voltage rises. Even a small imbalance between the currents will cause the midpoint to swing to the supply rails, so the transition between these state is usually very abrupt. In most designs, this threshold is set roughly halfway between the supplies, as this provides good noise immunity for digital designs. To achieve this, the transconductance of the PMOS and NMOS are adjusted (by altering MOSFET width/length) to match, ensuring that the FETs have the same current at the same gate-source voltage. In this application the threshold needs to be adjustable, which requires that the transconductance of at least one of the FETs be controllable. This cannot be done by altering the W/L ratio of the FETs in the field for obvious reasons, and converting the FETs into switchable transistor arrays would require vast area to

implement, so an adjustable single element must be added. An alternative might be to replace the FETs with flash memory cells instead, with the trapped charge adjusting the effective gate-source voltage, although this requires large floating gate transistors and their high voltage writing circuits. The difficulty in writing multiple states would also limit adjustment to a dozen or so possible states, which is insufficient for an application that might require narrow match windows. Instead, the transconductance of the FETs can be adjusted through source degeneration, where resistive elements are added between the source terminals of each FET and their respective supply rails. As the current through a FET rises, the voltage drop across the resistive element reduce the effective gate-source voltage, providing negative feedback and a lower effective gain. The reduced gain alters the input voltage at which the PMOS and NMOS match current, an therefore the threshold.
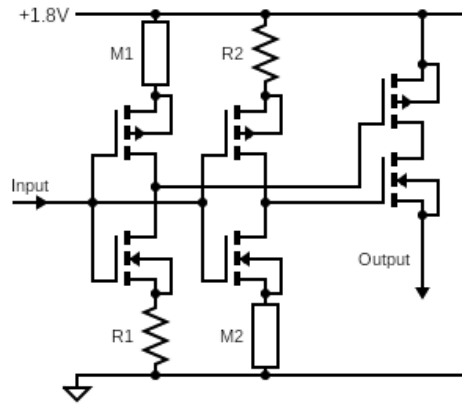
FIGURE 5.2: A diagram of the split TXL circuit.

The window comparator circuit considered (Fig. 5.2) uses a pair of skewed inverters with differing thresholds and uses an XOR mode transmission gate to compare their outputs. As the threshold of each inverter is dependent on the ratio between the top and bottom degeneration resistors and not the total resistance of the branch, the threshold can be controlled by altering the state of only one of them. As such, this design uses a single adjustable RRAM device to control each threshold, with the opposing degeneration being provided by a static balancing resistor. When the input is above the threshold of the M1 branch inverter and below the threshold of the M2 branch inverter, both the PMOS and NMOS of the output stage pass current, producing an output signal. While this configuration permits the independent control of each threshold with a single resistive device, the dynamic range required is much higher than in the previous designs. This is because the adjustable element must go both above and below the balancing resistor by a factor of 10 to achieve a reasonable range of threshold values. As DC paths remain in this design, the energy per test may not be as low as alternative designs with adiabatic operation, but should offer improved controlability.

5.4   Model

Existing models of memristive devices are limited in scope and detail, so simulations of
a TXL circuit are somewhat difficult. Analysis of this circuit must therefore be done
with a physical model. In theory, the state of the memristors used should only matter
at two points: the current at which the inverter switches, and at zero current. With the
low supply voltage and the series resistance, the current through the memristor should
never be sufficient to significantly alter its state. Because of this, models of TXL using
linear resistors should be a reasonable approximation of the intended circuit, but this
assumption must be tested. To this end, two ArC TWO daughterboards were made with
arrays of discreet component TXL circuits. The discreet component implementations of
this circuit use the SSM6L36TU,LF[87], along with $1\,M\Omega$ balancing and output resis-
tors. The FETs used are small signal MOSFETs with threshold voltages comparable to
integrated FETs. The high value balancing resistor result in low supply current in the
skewed inverters, allowing the inverters to switch a little closer to the supplies, reduc-
ing the required headroom. For the adjustable element, the resistor model (Fig. 5.3)
used a multiplexer switched array of 16 resistors, geometrically spaced between $100\,k\Omega$
and $10\,M\Omega$. The memristor model (Fig. 5.6) used a PLCC68 package of $30 \times 30\,\mu m$
monolayer titanium oxide memristors.

5.4.1   Resistor Model
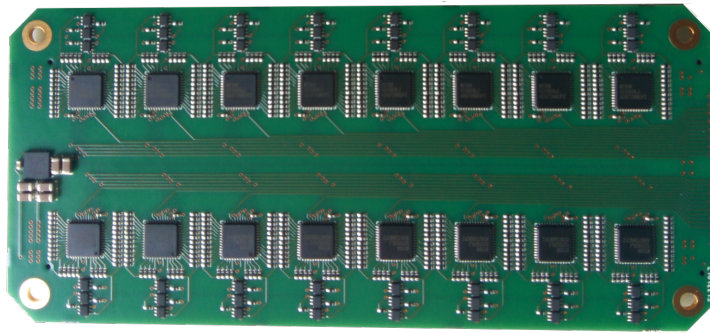
FIGURE 5.3: A photograph of the resistor version of the PCB model of the TXL.

The tests on both models was conducted with the ArC TWO. In the resistor model, the
TXL circuit was supplied at $1.8\,V$ and the input swept from $0\,V$ to the supply voltage.
The current was measured at every DAC increment, for a total of 5898 samples.

To asses how the window responded to changes in the control devices, the current output
is denoised with a 50 sample moving average filter and the differential calculated. The
window width is calculated as the difference in input voltage between the minimum
and maximum of the derivative (Fig. 5.4). This method was chosen as it provides a
meaningful result in circumstances where the window peak is not constant, where a
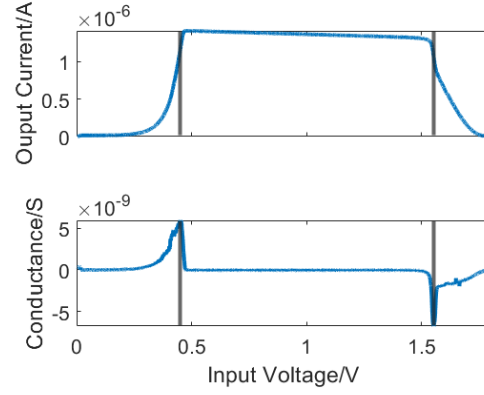
FIGURE 5.4: A diagram of the method used to determine window width. In this case the window width is calculated as 1.1 V.

more common method such as full width at half maximum (FWHM) might not. The differential method typically gives a slightly more pessimistic valuation of the circuit performance than FWHM.
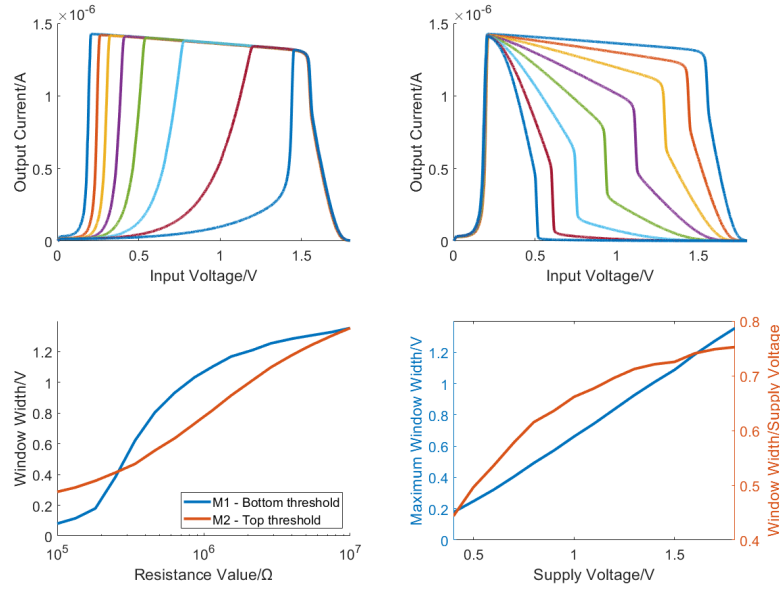
FIGURE 5.5: Top Left: A graph of the IV characteristics of the PCB model of the TXL circuit, sweeping M1 between $10\,\mathrm{M\Omega}$ and $100\,\mathrm{k\Omega}$. Top right: A graph of an identical sweep of M2. Bottom right: A graph of the width of the window as a function of resistor state. Bottom left: A graph of the maximum window width as a function of the supply voltage.

As can be seen in Figure 5.5, the physical model of the TXL gives comparable characteristics to the preliminary simulations seen in Figure 4.5. The operation of the MOSFETs in a low current regime reduces their effective gain, so the model is somewhat less sharp in its output transitions, but the required headroom is smaller. The upper threshold was measured to be almost linear with the logarithm of the control resistor. The lower threshold displayed significant nonlinearity, with a pronounced s-curve. Additionally, the maximum window width was calculated for a range of supply voltages. This was

found to be linear, implying that the headroom required is a function of the threshold value of the MOSFETs used, regardless of the operating current.

5.4.2   Memristor Model
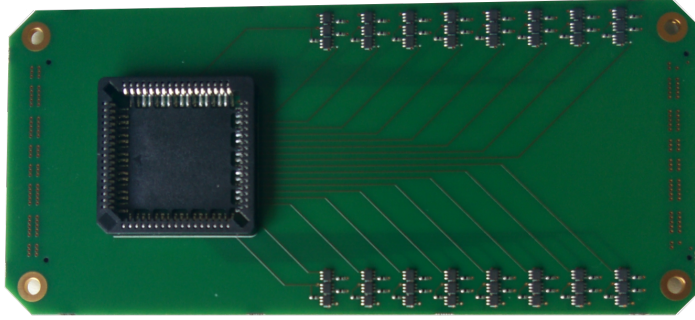
FIGURE 5.6: A photograph of the memristor version of the PCB model of the TXL.

For the memristor tests, the devices were set to a new state with $100 - 500\,\mu$s pulses and then read at $250\,$mV, before performing a sweep of the input voltage in $1\,$mV steps, reading the current at each step for a total of 1801 samples per trace. For the purposes of calculating the threshold position, the current output curve was denoised with the same 50 sample moving average filter. In this case, the same 50 samples represents a larger portion of the trace, but this was required as some of the tests produced exceptionally noisy results. The memristor version of this experiment proved much more challenging to run. The monolayer titanium oxide memristors used in this test are an early memristor technology and come with many of the expected teething issues, presenting difficulty setting new resistive states and holding existing ones. During writing, the state of the devices would relax back towards the previous state and in some tests, the state of the control device would fluctuate. In the sweep of M1, some tests saw the memristor taking random values within a small range, as can be seen in the top right of Figure 5.7. In the sweep of M2, a test at high resistive state saw the control device switching between two distinct values, as can be seen in the outermost trace of the top left of Figure 5.7. This may be indicative of the formation (and disruption) of conductive filaments within the titanium oxide layer, as this is one of the hypothesised mechanisms of the memristive behaviour of these devices[88]. Attempts were made to run tests using bilayer titanium oxide memristors, but the devices used could not be reliably set to resistive states above $500\,$kΩ, which rendered them useless for the bias conditions of this model.

Due to the difficulties holding a device at a specific state, instead of assessing the window width, the position of each threshold was calculated. While this is less useful in characterising the circuit, it prevents issues with setting one threshold from affecting the assessment of the other. Despite this, the tests on the bottom threshold showed a similar s-curve shape to the resistor model. The shape of the window produce is also
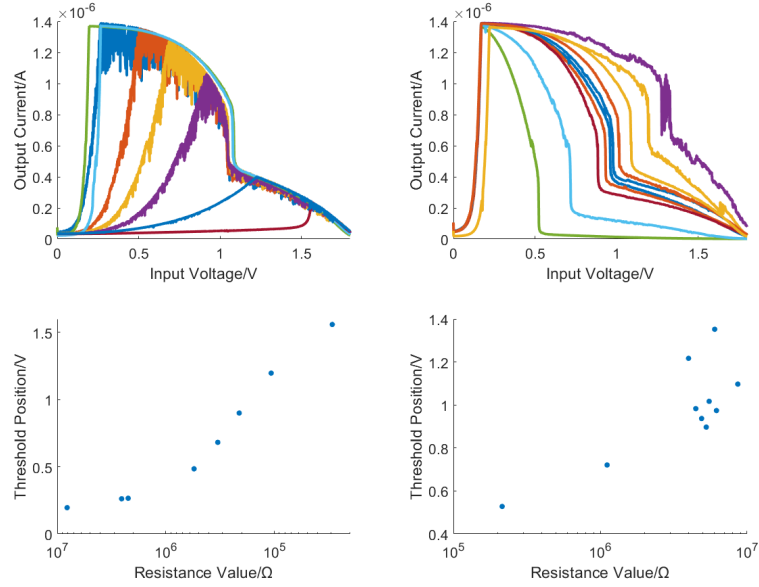
FIGURE 5.7: Top left: A graph of the IV characteristics of the memristor model, sweeping M1 from 8 MΩ to 30 kΩ, with M2 at 6-8 MΩ. Top right: A graph of a similar sweep of M2 from 200 kΩ to 8.6 MΩ, with M1 at 5.6-6.5 MΩ. Bottom left: A scatter plot of the window width as a function of the memristor state. Note that the X axis is reversed compared to the bottom right figure. Bottom right: A scatter plot of the window width as a function of memristor state.

very similar to the shape obtained in the resistor model. The upper threshold was not so well behaved, showing significant loss in the sharpness of the transition. Further, there was a loss of monotonicity at high resistive states, although this may be due to changes in the state of the M2 device between the measurement taken at the start of each test and the state it took during the test. Despite the irregularities, the traces produce in the tests of the memristor model are clearly of comparable shape and range of position to the results of the resistor model. From this, it can be concluded that resistor models of the circuit can be assumed a reasonable model of the circuit for the further experiments assessing characteristics not explicitly related to the threshold shape, such as power consumption and footprint area.

5.5 Integrated Simulation

5.5.1 Design

While useful for assessing the input-output relationship with novel devices, the discreet component model is somewhat limited. To estimate the power consumed by the circuit, an integrated circuit model of the design was produced in Cadence Virtuoso using the TSMC180BCD product development kit. This technology node uses planar MOSFETs, for either 1.8 V or 5 V. As integrated design provides much greater control of the specifics of the transistors used, three versions of the circuit were designed, with different sizes

for the transistor used in the two inverters (Tab. 5.1): one with minimum size input transistors, one with wide input transistors, and one with minimum size native NMOS and conventional PMOS input transistors. In all cases the two transistors on the output branch were minimum size. All transistors are 1.8 V devices. While it may be possible to replace the balancing resistors with static memristors, the assumption is that polysilicon resistors will be used. This constraint makes the 1 MΩ balancing resistors of the PCB model wildly impractical, due to the area required. The simulations shown here were conducted with 10 kΩ balancing resistors and adjustment resistors of 1 kΩ to 100 kΩ. While this strips the circuit of a significant portion of its range, it allows for the recovery of much of the sharpness that is now expected to be lost when implemented with memristors. It should be noted that a reduced range is not necessarily a problem, as the circuits before the TXL in the signal chain can be designed to output in the operating range of the TXL, and may even have their own headroom requirements that make a wider window unnecessary.

	Minimum	Wide	Native
NMOS Width	220 nm	1 µm	420 nm
NMOS Length	180 nm	180 nm	500 nm
PMOS Width	220 nm	1 µm	220 nm
PMOS Length	180 nm	180 nm	180 nm

TABLE 5.1: Chart of transistor sizes in the three designs that were simulated.

5.5.2   Assessment

To assess the window width, static DC simulations of the circuits were run. The circuits were supplied at 1.8 V and the inputs swept from 0 V to the supply voltage, measuring the current measured at the output. This was repeated for geometrically spaced values of 1 kΩ to 100 kΩ for the adjustable element. In all cases the output was connected to a 100 kΩ dummy load. The window widths were calculated using the same differential method mentioned earlier.

In the circuit with minimum size input transistors, the window width was drastically reduced. Further to this, the range of values for the lower thresholds was roughly half the maximum window width (Fig. 5.8↖). While the transitions were sharper, this reduced range represents a catastrophic loss of functionality, as a narrow window can only be set in a 100 mV range. The wide input stage circuit performed much better, losing less maximum window width and retaining most of the threshold sweep of the PCB model (Fig. 5.8↑). The sharpness of the transitions was also superior to the minimum width circuit, due to the higher gain of the input FETs. The native input stage circuit displayed a wider maximum window width than even the wide circuit, although the minimum size PMOS left it with a lower threshold range comparable to the minimum circuit (Fig. 5.8↗). The sharpness of transitions in the native circuit were noticeably worse. As one might expect from the lower threshold voltages of native devices, the
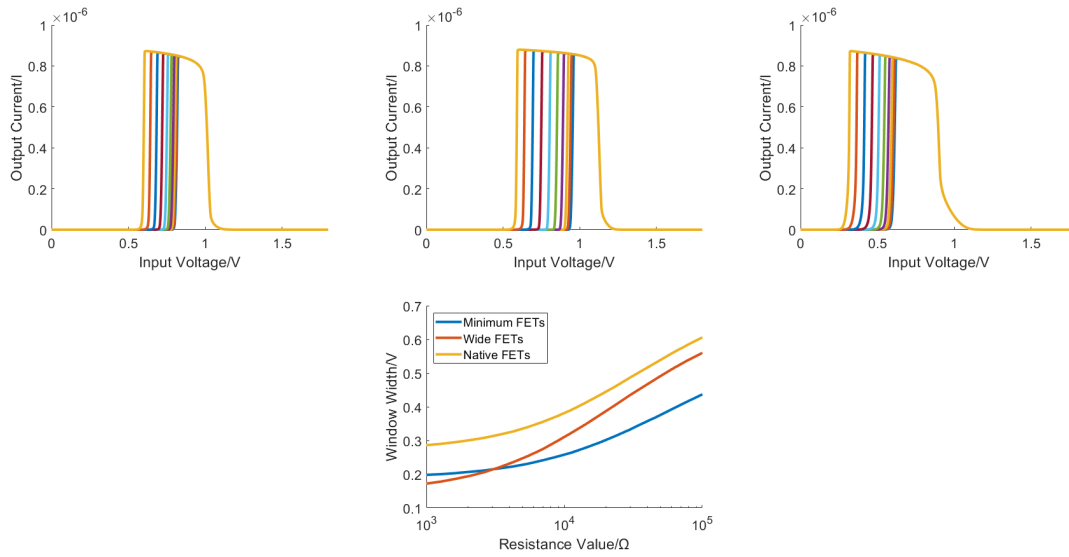
FIGURE 5.8: Top left: A graph of the IV characteristics of the minimum circuit, sweeping M1 from $1\,\mathrm{k\Omega}$ to $100\,\mathrm{k\Omega}$. Top centre: A graph of an identical sweep of M1 with the wide circuit. Top right: A graph of an identical sweep of M1 with the native circuit. Bottom: A graph of the window width of all three circuits as a function of M1.

window of the native circuit was much loser to $0\,\mathrm{V}$ than the others, which were almost centred in the supply range. In all circuits, plotting the window width as a function of the logarithm of M1 showed less nonlinearity than lower threshold of the PCB model.
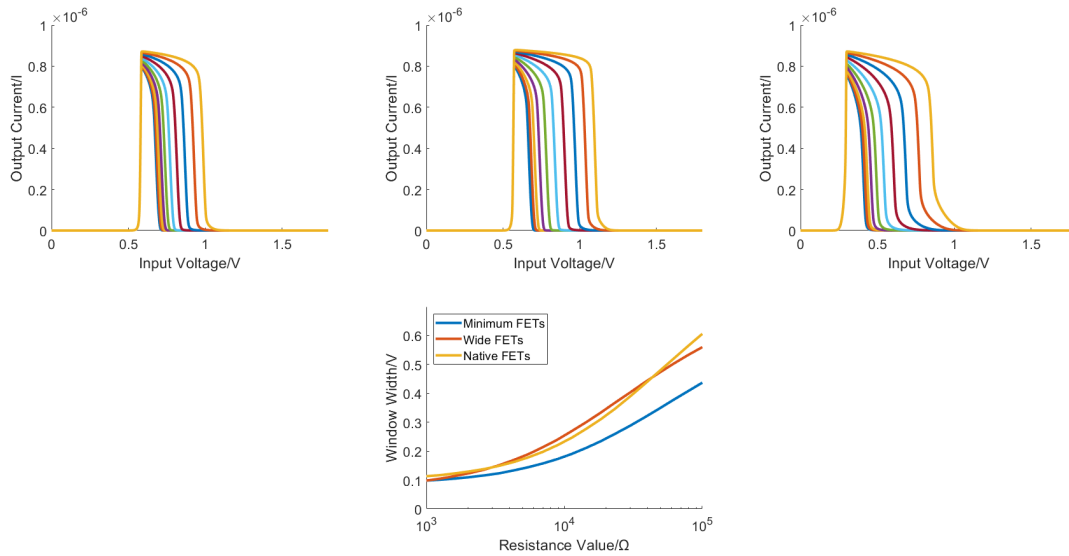
FIGURE 5.9: Top left: A graph of the IV characteristics of the minimum circuit, sweeping M2 from $1\,\mathrm{k\Omega}$ to $100\,\mathrm{k\Omega}$. Top centre: A graph of an identical sweep of M2 with the wide circuit. Top right: A graph of an identical sweep of M2 with the native circuit. Bottom: A graph of the window width of all three circuits as a function of M2.

The differences in the behaviour of the top threshold between the three circuits was less pronounced. All three were able to bring the top threshold to within $100\,\mathrm{mV}$ of the bottom threshold. The only significant differences were the inability of the minimum

circuit to reach the maximum window widths of the other two, and the less defined transitions of the native circuit (Fig. 5.9).

5.5.3   Energy
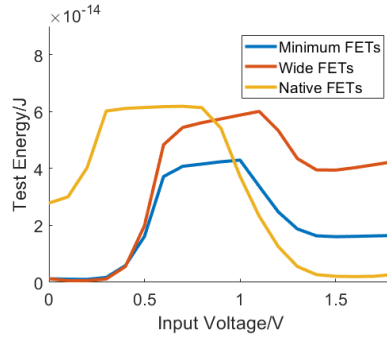
FIGURE 5.10: A graph of the test energy of a single simulated TXL cell, as a function of input voltage.

To estimate the energy required to test a single sample, a second gate was added to each of the output transistors as an output enable control, and a transient analysis simulation was run. In this test, the circuit began with the supply energised, the output disabled, and the input parked at either supply or ground. The minimum and wide circuits were parked ant ground and the native circuit parked at supply. The input was brought to the test value and held there for 2.35 ns to settle. The enable signal was then pulsed for 450 ps. 200 ps after the enable pulse ends, the input is returned to its parked state. The energy was calculated by integrating the sum of the instantaneous power at the input, the enable input, and the supply (Fig. 5.10). As might be expected for an analysis that includes the output current, the test energy of a hit was the peak energy in all cases. Curiously, the miss energy on the opposite side of the window to the parked state was substantially higher than the miss energy on the same side of the window. This is likely because the act of bringing the input signal to the test voltage and back causes the input to pass through both input inverter's crossover points. This would cause a small surge of current each time, resulting in the observed plateau.

Further simulations were run at relevant process corners. A corner is a set of environmental or process conditions that might affect circuit performance. Each of the circuits was simulated at four corner conditions: room temperature, body temperature, P/N-MOS above specified gain (Fast/Fast), and P/NMOS below specified gain (Slow/Slow) (Tab. 5.2←). A hit and a miss voltage were chosen for each circuit, to give estimates for a typical use case. The variation with temperature was minimal, but the effect of process variation was substantial, with a Fast/Fast hit dissipating almost twice the energy of a Slow/Slow hit. The process variation also cause significant variation in the window width, with with variation of as much as 20% (Tab. 5.2→). As the name suggests,

	Minimum	Wide	Native
Hit Voltage	900 mV	900 mV	600 mV
25 °C	42.36 fJ	57.44 fJ	61.75 fJ
37 °C	42.92 fJ	58.30 fJ	62.41 fJ
Fast/Fast Corner	61.26 fJ	76.11 fJ	86.51 fJ
Slow/Slow Corner	29.34 fJ	44.48 fJ	44.55 fJ
Miss Voltage	1.3 V	1.5 V	1.2 V
25 °C	18.75 fJ	39.43 fJ	12.54 fJ
37 °C	19.39 fJ	39.78 fJ	13.77 fJ
Fast/Fast Corner	27.40 fJ	40.88 fJ	27.88 fJ
Slow/Slow Corner	15.22 fJ	37.99 fJ	4.738 fJ

Maximum Window Width at	Minimum	Wide	Native
25 °C	436 mV	558 mV	606 mV
37 °C	441 mV	568 mV	610 mV
Fast/Fast Corner	523 mV	630 mV	690 mV
Slow/Slow Corner	341 mV	482 mV	511 mV

TABLE 5.2: Left: A chart of test energies at typical hit/miss voltages for several process corners. Note that the difference in hit/miss voltage between the circuit variants make this data less useful for comparing the variants. It is more intended to show the response of each circuit to process variation and temperature. Right: A chart of the maximum window width at several process corners

a corner is the extreme edge of what one can expect out of the technology, but the variation on display here is sufficient enough to warrant further investigation.

5.5.4 Process Variation

To obtain a clearer picture of the impact of process variation, a Monte Carlo analysis was conducted. This approach uses a large number of simulations with randomised process variation to provide a statistical model of the expected impact. In this test, the characteristic being monitored is the maximum window width, calculated using the same differential method used earlier. 250 simulations were run for each circuit.

	Minimum	Wide	Native
μ	436 mV	559 mV	606 mV
σ	29.9 mV	23.6 mV	33.0 mV
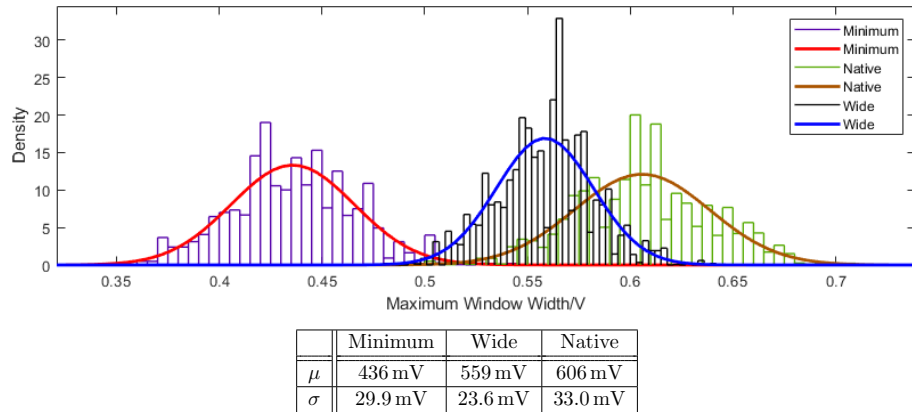
FIGURE 5.11: Top: A histogram plot and fit of a 250 point Monte Carlo simulation of the maximum window width. Bottom: A table of the fit parameters, assuming a normal distribution.

The resulting histograms of window width show a significant level of variation in the minimum and native circuits (Fig. 5.11). If an error in fabrication causes a gate element to be 20 nm narrower than intended, the proportional error will be much larger on a minimum size FET than a wider one, so greater variance on the minimum size circuit is to be expected. Why the native circuit has slightly greater variance is not entirely clear,

as the larger minimum size of the native FETs should result in marginally less variance, but the low doping level of the native devices likely makes them more vulnerable to fabrication errors there. The wide circuit showed the least variance, both in absolute terms and in proportion of average maximum window width.

5.6   TXL Template

To demonstrate the use of the circuit discussed in this chapter, the resistor array PCB model (Fig. 5.3) was used to run tests of a simulated template matching system. A set of four waveforms were extracted from a paper on neural analysis[89] and scaled to an appropriate range for the TXL circuit. Then a script was used to pick samples from them (Fig. 5.12), as if it were a sample hold circuit. Sixteen samples were taken from each waveform at a frequency of 25 kHz, starting when the signal deviates from its initial value by more than 100 mV. Using the data collected in section 5.4.1, appropriate resistor configurations were chosen to fit the window of each TXL to the corresponding sample from the target waveform. The array of TXL cells fitted to a specific waveform make a template, against which the sample spikes can be tested by biasing the TXL inputs to the collected sample voltages and summing the current outputs. All four spikes were tested against all four templates (Fig. 5.12).
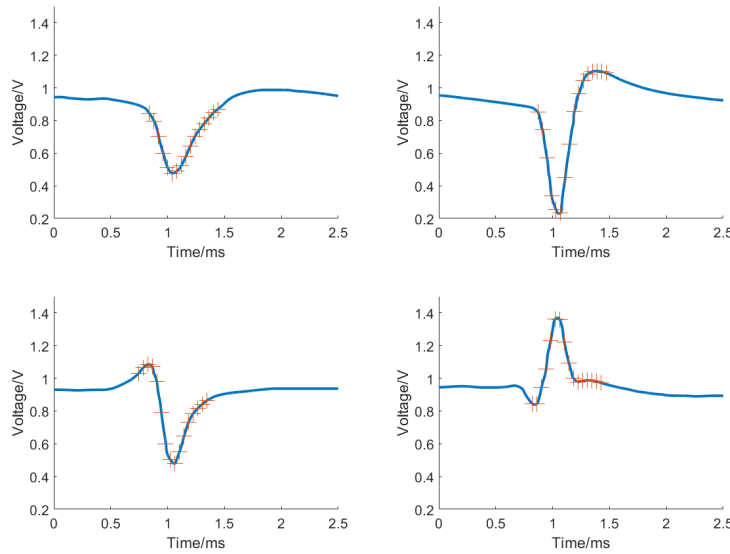
FIGURE 5.12: Graphs of the sample neural spikes used in the array test, with the samples collected marked in orange. Top left: Sample spike 1. Top right: Sample spike 2. Bottom left: Sample spike 3. Bottom right: Sample spike 4.

In all cases, the target waveform elicited greater current output than any other spike, but the ratio between the match and the nearest mismatch was less than expected. Further to this, the nearest mismatch on some templates came close to the match of others, although the lowest current match was still greater than the highest current

mismatch (Fig. 5.13). These behaviours are due to two factors; the large minimum width of windows near the middle of the input range, and the sub-optimal sharpness of the window of this particular implementation. The wide mid-range window allows for waveforms with close sample voltages, such as in the latter third of spikes 1 and 3, to achieve partial matches, reducing the margin of a successful match. The low sharpness of the window transitions in this model causes the current output within the bounds of the window to be noticeably lower when the upper threshold is set to a lower voltage (Fig. 5.5). Because of this, the total current in templates for waveforms with many low voltage samples is lower than in templates for waveforms with mostly mid or high voltages. While these two effects do not seem to have much overlap in this example, it is not difficult to imagine situations where such combined problems cripple the accuracy of the system. The superior granularity of memristors and the sharper windows of an integrated implementation will likely mitigate these issues, but if they persist measures could be taken to reduce the impact. To help distinguish between similar spikes, the number of samples taken from the recovery portion of the spike could be reduced by increasing the sample rate or reducing the number of samples, restricting the template to the more distinct section of the spike around the peak. To avoid the issue of some templates producing lower match current, the template matching channel could have an independent threshold for each of the templates in its design, isolating the low current templates from the threshold requirements of the high current templates. It may also be possible to adjust the weight of each TXL in a template by replacing the current limiting output resistor with a memristor. This would allow cells with low expected match current to be amplified and less relevant samples from the post-peak recovery to be suppressed, although this may not be practical as writing to a device that is not anchored on one end by a supply might prove challenging.
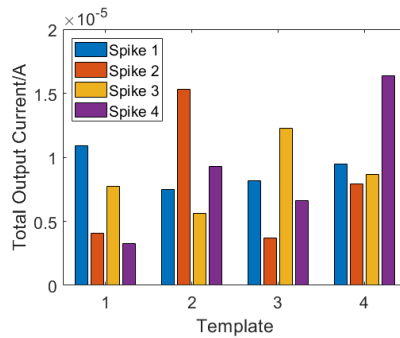
FIGURE 5.13: A graph of the test energy of a single simulated TXL cell, as a function of input voltage.

A further model of this template matching channel was planned, with a PCB designed and fabricated for a complete channel (Fig. 5.14). This circuit featured two seven sample templates, along with supporting sample hold and trigger circuitry. The planned tests included template matching of sample waveforms with memristors instead of resistor arrays, and sorting of neural spike signals from live tissue samples. Unfortunately, the

pandemic related disruption slowed the production and characterisation of the newer memristor devices needed to run these test and they could not be completed in a timely fashion.
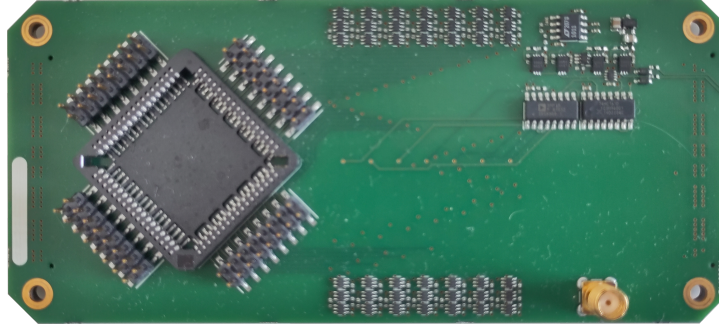
FIGURE 5.14: A photograph of the TXL channel demonstrator.

5.7   Conclusion

For a hypothetical spike sorting array of three templates with twenty samples each, operating at body temperature, the data obtained in the simulations can be used to estimate the total power dissipation. If it can be assumed that a detected event causes a full match on one template and a two-fifths match on the others, the typical energy expended to classify an event is 3.05 nJ, giving a power consumption of 122 nW under an average event frequency of 40 Hz[20]. Pairing this array with a suitable sample and hold circuit[90] gives a total power of 179 nW, assuming all 20 sample hold circuits are shut down outside of a 2 ms window around a detected event. A system of comparable scale presented in 2019[9] claims to use roughly one third the power estimated here at 64 nW, although that work ignores the ADC power consumption and uses a far lower energy CMOS technology node, allowing it to operate on a much lower voltage supply. An appropriate 10-bit, 100 kHz sample rate ADC[91] for this ASIC consumes 1.3 µW and can serve four ASIC channels, for 389 nW total power per sorting channel. This gives a much more favourable comparison for the work covered in this chapter, at around half the power of a comparable ASIC system. It should be noted though, that the energy estimated here only considers the power consumption of the TXL array, and not the peripheral systems such as event detection. Further, the timings proposed in the simulations are somewhat optimistic. Driving the enable signals of more than one TXL at those speeds may prove challenging, even with a faster technology than TSMC180BCD. Despite this, the data collected here shows that this circuit has significant potential. With the development of more controllable ReRAM devices, such as SnO_2 or HfO_2 based devices[70], an integrated pattern matching system is within reach.

Chapter 6

Future Directions and Conclusions

6.1 Conclusions

This thesis has presented the development of an instrumentation platform and then shown the testing of a new memristor tuned window comparator topology. The developed instrument has been characterised and shown to be capable of the required tasks. This thesis has discussed topologies for a circuit that fulfils the requirements, and then conducted a detailed analysis of the selected design, showing performance with physical memristive devices and estimating power dissipation.

The intended application of the subject of this thesis necessarily invites comparison to two similar projects: a CMOS ASIC for sorting digitised spike signals by D. Valencia et al. [9] , and a window comparator for use in analogue content addressable memory (CAM) by Can Li et al. [34].

Compared to the ASIC, this work operates at favourable power consumption, with around half the power for a channel of similar scale. This improvement arises from the lack of need for an ADC or other analogue circuitry between the amplified signals and sorting system. The potential improvement on an already high performance system offers greater access to neural signals in future neurological research.

The comparison to the content addressable memory is much less favourable, with around one hundred times the test energy per cell in the work presented in this thesis. The content addressable memory cell manages this by using a match line, which stores a small amount of energy in its capacitance. If any of the cells on a match line detect a miss, they deplete the match line. This topology has few DC current paths and handles only the energy required to charge the match line, allowing for impressively low power figures. The downside to this approach is that any cell detecting a miss depletes the

match line entirely, there can be no partial matches. For an application as prone to subject matter irregularity and noise as neurological medicine this constraint will likely limit the maximum accuracy of such a system. In addition to this, the CAM cell used as comparison requires much greater changes in memristor value to achieve its full threshold swing, around five times the proportional changed used in this work. This work also achieves far sharper thresholds than the CAM cell, so while it will likely never reach the same energy performance it remains a relevant development.

To conclude, the circuit developed in this project shows potential for further research, with existing material for expanding the scope to tests with in vitro neural tissue samples.

6.2   Future Directions

6.2.1   TXL

The results obtained in this project show the potential for further development of the Split TXL. Tools for continuing this research were developed during this project, but tests with them could not be completed due to poor ReRAM availability. An initial and obvious direction to pursue would be the completion of these tests. Alternative memristor and transistor technologies might enable the lower supply voltages required to give this circuit a decisive advantage over competing research. With a proper demonstrator channel, an integrated implementation of the circuit could be pursued. Such a system would have broad modularity, as it is largely input agnostic, requiring only an analogue signal in the input range and an accompanying trigger signal. Because of this it could remain relevant regardless of development in research of supporting systems, such as neural amplifiers[11].

6.2.2   Further development of ArC TWO

The flexibility of the instrument developed for this project not only supports the future development of the TXL, but also enables further research into ReRAM technology. One of the fields of interest in the research group is the development of selector enabled crossbar arrays, which allow for the operation of high density memory arrays without the need to be concerned about the sneak path currents. New daughterboard adaptors could be produced for any package requirement. Further, the daughterboards can be fitted with circuits to augment the capabilities of the instrument. An example of this might be a sub ns pulse generator which could be used in research pursuing lower energy memory technologies.

6.2.2.1 Pulse Generator

The response of memristors to writing pulses suggests that shorter, higher voltage pulses may be able to write to the memristor with lower energy. Reducing the write energy by using ps scale pulses could have significant benefits for the overall power consumption and controllability of any memristor based system. A better pulsing scheme might have alleviated some of the issues found with memristor operation in chapter 5. Testing of an SRD switched Blumlein pulse generator found that the pulse generator did not perform to expectations (Fig. 6.1). An alternative design based on a varactor loaded NLTL is under consideration. Simulations of the NLTL pulse generator show promise, but have yet to be conducted with models of real varactors. Once this has been done, a PCB prototype will be produced and tested. There is presently little study of the write behaviours of memristors with sub ns write pulses; experiments in this area could shed a deeper light on the exact functioning of these devices, opening new avenues of research.
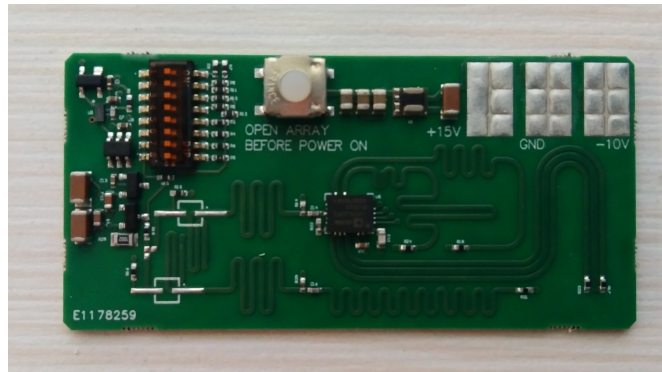
FIGURE 6.1: A photograph of the SRD-based pulse generator.

Appendix A

ArC Neuro Prototype

A.1 Overview
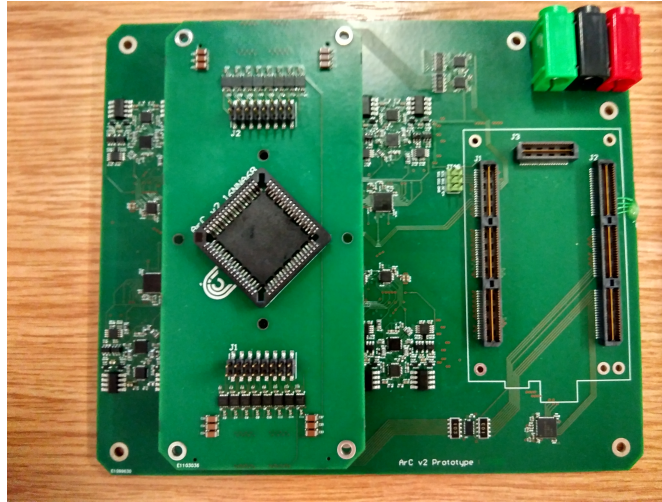
The initial planning of the parallel SMU instrument called for a pair of systems, a 64-channel instrument for 32x32 crossbar arrays and a half scale 32-channel instrument for 16x16 crossbar arrays. The first prototype of this project was built as a half scale instrument (Fig. A.1). This prototype lacked the power management circuitry of later versions, instead receiving $\pm 15\,\text{V}$ directly from a benchtop supply. Two versions of the high speed driver circuits were tested in this design, one based on comparators and another based on gate drivers with level shifters, with each cluster of eight channels featuring a mix of the two.

A.2   Software

A.2.1

Appendix B

ArC Neuro Block 1
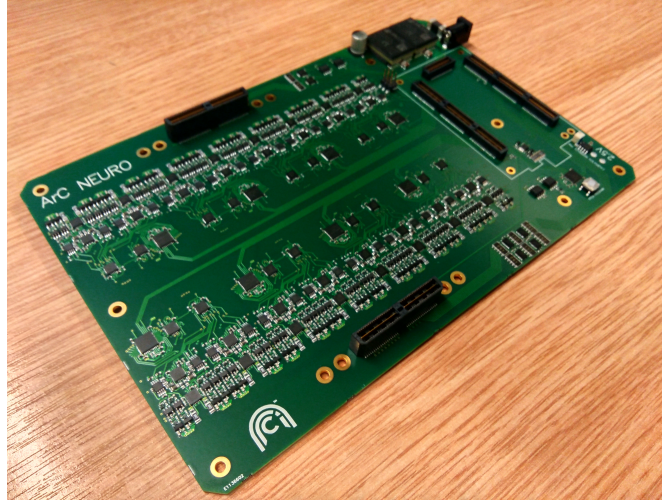
FIGURE B.1: Photograph of the block 1 ArC Neuro.

B.1 Overview

The first version of the 64-channel instrument (Fig. B.1) implemented a redesigned channel cluster with four fifths the width. For this revised cluster, a new high speed driver circuit was designed and tested (Fig. B.2). This driver circuit used the same gate drivers as the driver circuit on the prototype, but with a better level shifter to translate between the ground referenced FPGA and the $-15\,\mathrm{V}$ referenced gate drivers. A snubber circuit was added to mitigate the ringing of this new circuit. To simplify the cluster, the redesign uses only one type of analogue switch array, rather than two. The ADCs selected for this design featured differential inputs suitable to the operating range of the instrument, although 16-bit ADCs were used as the 18-bit version of the part was not available at the time. The block one also implements the remaining planned peripheral circuits that were missing from the prototype, adding arbitrary voltage supplies for the DUT and a calibration reference on the current source circuit.
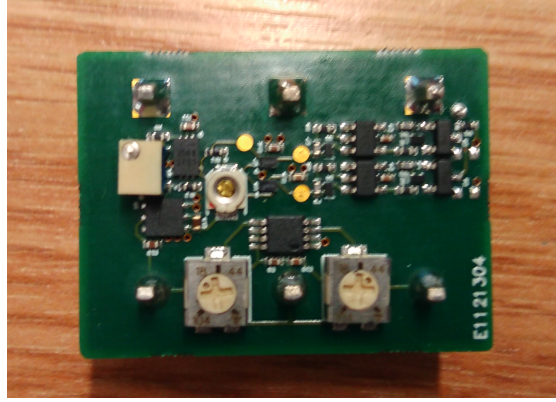
FIGURE B.2: A photograph of the PCB design used to test the revised high speed driver circuit.

B.2 Design Issues
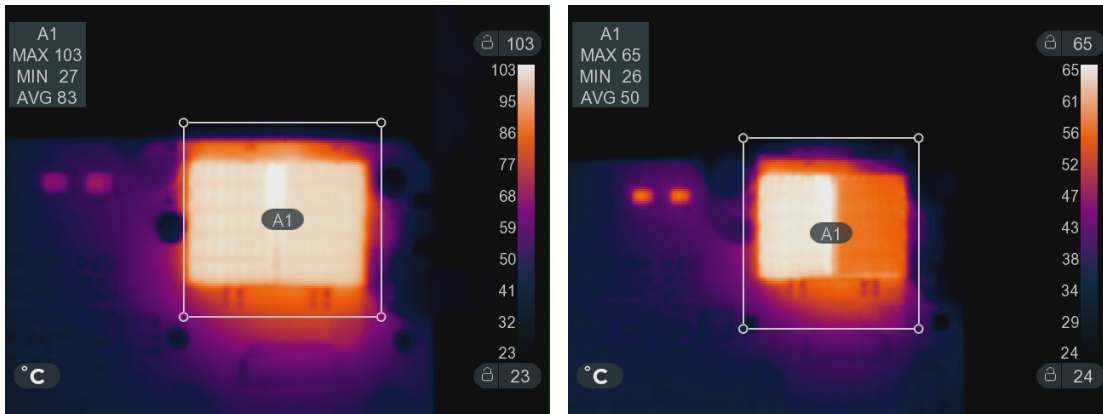
FIGURE B.3: Left: A thermal photograph of the DC/DC modules at equilibrium temperature, with heatsinks fitted. Right: A thermal photograph of the DC/DC modules at equilibrium temperature, with both heatsinks and fan fitted.

The isolated DC/DC modules used to generate the $\pm 16.5\,\text{V}$ supplies from the $18\,\text{V}$ input proved unsuited to the application. Due to the unusually high quiescent power consumption of the DC/DC module and the close placement of two such parts, the modules exceeded their thermal ratings and shut down (Fig. B.3). This issue proved to be surmountable; with the addition of heatsinks and a fan, the operating temperature was brought down to an acceptable level (Fig. B.3). In addition to being excessively hot, the supply circuitry was also very noisy (Fig. B.4). To complicate the matter further, no alternative supply option was present in this design, and many tests were run using a board with wires soldered onto exposed metal on the outputs of the DC/DC modules. The thermal issues did not end with the system's power supply; the op-amps used to buffer the peripheral DAC had insufficient thermal conductivity, with the op-amp supplying the arbitrary logic circuit reaching over $100\,^{\circ}\text{C}$ under normal operation.
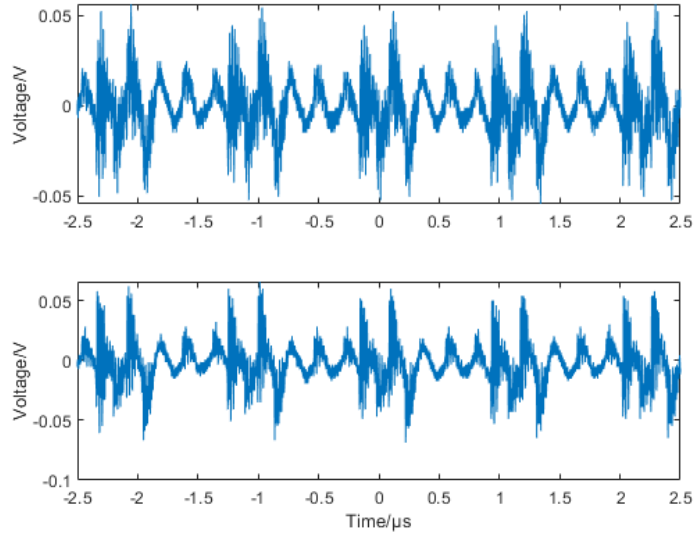
FIGURE B.4: Oscilloscope capture of the noise on the ±15 V supplies.

In the block 1, there were four serial daisy-chains that control the TIA ranging and channel function switches, one each for top side and bottom side on each half of the board, with an additional serial daisy-chain for the high speed driver signal and ADC grounding switches of each channel. While this allowed for very intuitive PCB layout, it causes complications for the automatic TIA range selection operation. The next range state needed to be determined for all channels on a given serial chain before the switches on that chain can be updated. In addition to this, a common serial clock was used for the ADCs and another for the DACs of each row. This mix of cluster, location, and component type serial groups tied the functionality of the entire instrument into two groups, crippling the intended parallel functionality of the instrument. Further frustrating efforts to get this working, the FPGA module used in the instrument produced a wide range of undiagnosable bugs, with the FPGA failing to initialise when mounted and circuits within the FPGA not behaving as specified. In some cases, identical copies of a given block within the FPGA design would respond differently to commands and measurements. The initialisation issues were resolved by powering the FPGA from the 5 V supply for the ADCs/DACs, but this did not resolve the other problems. These issues would later be found to be a result of undocumented supply requirements of the FPGA module. This version also displayed ringing in its $110\,\mathrm{k\Omega}$ TIA range. One board was modified with compensation capacitors in parallel with the $110\,\mathrm{k\Omega}$ resistor to mitigate this.

Appendix C

ArC Neuro Block 2

FIGURE C.1: Photograph of the block 2 ArC Neuro with FPGA module and daughterboard mounted.

C.1 Overview

For the block 2 version of the instrument (Fig. C.1), a new power supply circuit was designed (Fig. C.2). This circuit replaced the isolated DC/DC modules with non isolated modules; a SEPIC and a buck converter. Not only did these modules have higher efficiencies at the intended operating power, but they also had better thermal conductivity, and were small enough to fit beneath an EMI shield. A set of solder terminals and accompanying reverse protection circuitry was added to permit the use of alternative supply options.
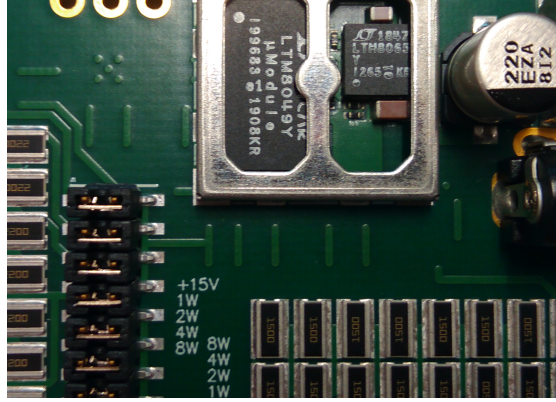
FIGURE C.2: Photograph of the test PCB for the revised power supply circuitry, including load banks. The lid of the EMI shield has been removed to show the circuitry underneath.

As might be expected from such differences, the new supply configuration performed significantly better; while the EMI shield precluded the use of forced convection, the peak temperature was 25 °C lower than the previous version (Fig. C.3). The new peak temperature was not quite as low as with the fan on the previous version, but dropping the requirement for a noisy load on the supply was considered a worthwhile trade-off.
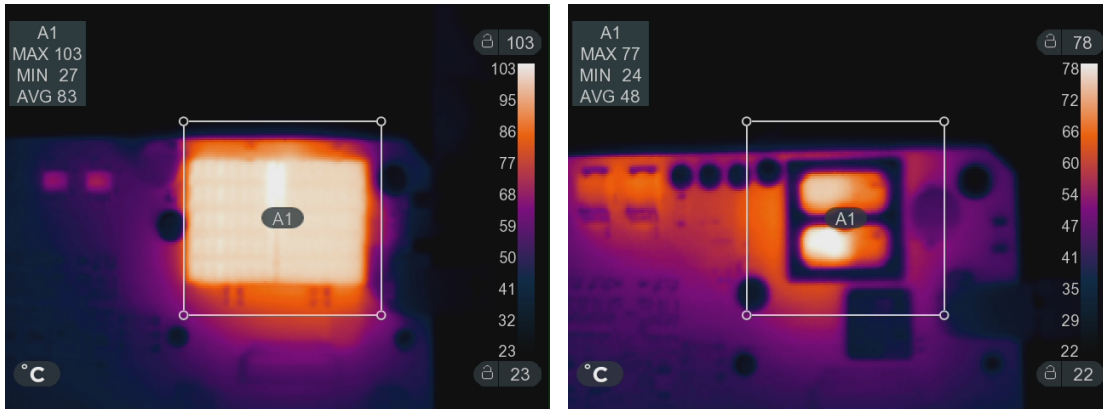
FIGURE C.3: Left: A thermal photograph of the DC/DC modules at equilibrium temperature, with heatsinks fitted. Right: A thermal photograph of the revised power supply circuitry at equilibrium temperature. The lid of the EMI shield has been removed to allow for accurate thermography.

The removal of the fan, combined with the higher performance DC/DC modules and EMI shield resulted in significantly less noise on the ± 15 V supplies. The revision of the supply circuit reduced the peak-to-peak amplitude by almost two thirds, and reduced the fundamental frequency from 1 MHz to 25 kHz (Fig. C.4).
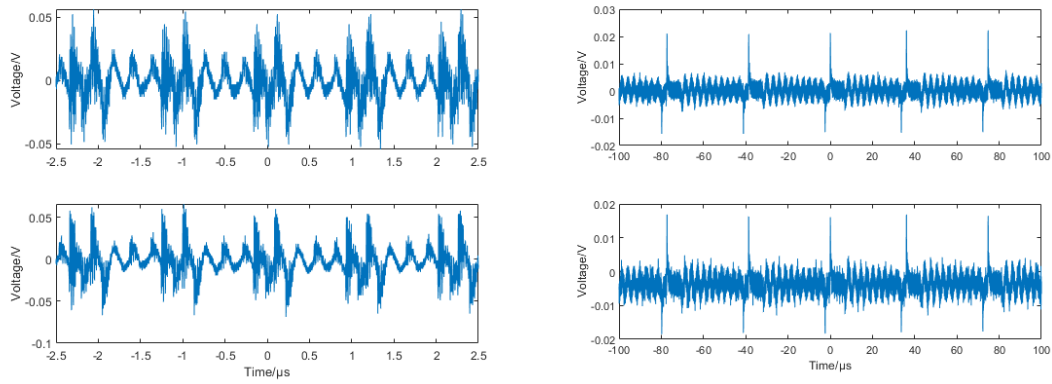
FIGURE C.4: Left: An oscilloscope capture of the noise on the $\pm15\,$V supplies of the block 1. Right: An oscilloscope capture of the noise on the $\pm15\,$V supplies of the block 2.

C.2   Design Issues

The serial structure of the block 2 was revised. In this version, all the switch ICs on a cluster were moved to a single serial daisy-chain. This resulted in three serial busses for each cluster; one for the ADC, one for the DAC, and one for all analogue switches in the cluster. While this did not render each channel completely independent, achieving greater granularity would require separate ADC and DAC ICs for each channel, which would require prohibitively large PCB area. The common ADC serial clock of the previous version was broken up and a dedicated clock signal provided for each ADC, although the common clock for the DACs was retained.
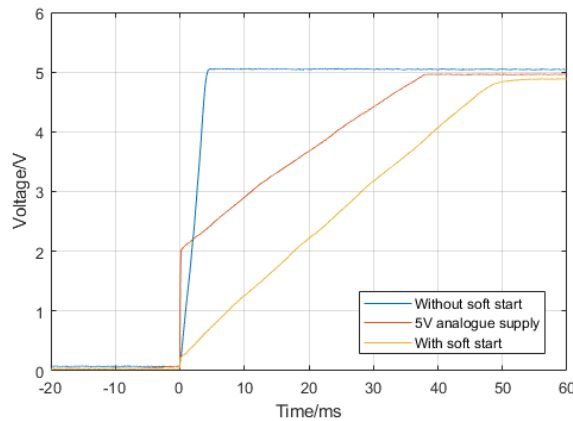
FIGURE C.5:   Oscilloscope capture of the startup curves of different FPGA supply solutions.

To address the FPGA supply issues that had started to become apparent with the block 1, the block 2 implemented a dedicated 5 V supply for the FPGA. As the supply issue was not yet understood, this design choice served only to reintroduce the initialisation issues

of the block 1. An in-depth investigation was carried out to determine the source of this fault by process of elimination. It was found that the FPGA module was sensitive to the rise time of the power suppl, with a rise time of over 40 ms being required to reliably initialise the FPGA (Fig. C.5). While the 5 V analogue circuit supply does fulfil this requirement, a linear regulator in the path caused a step from ground to 2 V. While no confirmed, this is likely due to latch-up of the FPGA due to improper sequencing of the various supply circuits within the module, as the FPGA did not misbehave when not mounted with capacitive loads on its GPIO pins. Further investigation into this fault was considered, but determined to be outside the scope of this project. To bring a block 2 board to proper functionality, a fragment of PCB containing a DC/DC module that happened to feature soft start capability was cut from a block 1 board and affixed to a block 2 board (Fig. C.6). The modified board operated reliably, both in regards to FPGA initialisation and behaviour of circuits within the FPGA configuration.
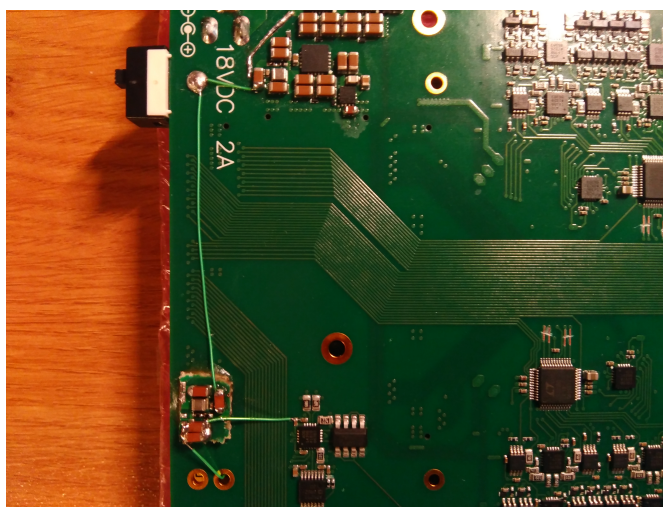
FIGURE C.6: A photograph of a block 2 board that has been modified with a DC/DC module with soft start.

To resolve the absurdly high temperature of the arbitrary logic supply, the op-amp was replaced with a linear regulator, which was set using a bank of analogue switches. This supply proved ineffective, and was neither stable nor configurable, likely due to the effect of the analogue switch parasitic characteristics on the closed-loop control of the linear regulator.

Appendix D

ArC Neuro Block 3

FIGURE D.1: A photograph of the block 3 ArC Neuro with FPGA module, daughter-board, and power supply module mounted.

D.1 Overview

The block 3 represents the pre-production state of the instrument (Fig. D.1). It addresses the FPGA supply requirements identified in the block 2 by swapping the dedicated supply circuit of the block 2 for a comparable one with soft start capability. As the supply of DC/DC modules was unreliable at the time the block 3 was produced, the majority of the supply circuitry was moved to a separate daughterboard to modularise the supply (Fig. D.2). In addition to allowing for the rapid and independent redesign of the DC/DC circuitry, it also allowed the auxiliary supply solder terminals to be replaced with a full size set of 4mm plugs. To prevent user error from causing damage to the instrument, reverse bias and UVLO/OVLO protection circuits were added to this daughterboard. While some minor changes were made to the high speed driver level shifters for the production model, no significant alterations were made for the production model.
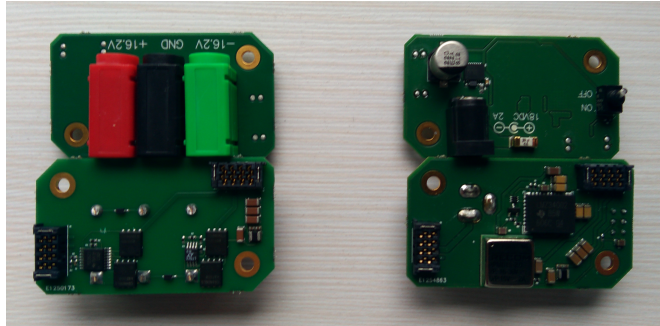
FIGURE D.2: A photograph showing a 4 mm supply module on the left and an $18\,\mathrm{V}_{DC}$ module on the right.

Appendix E

ArC TWO Daughterboards

E.1 32NNA68
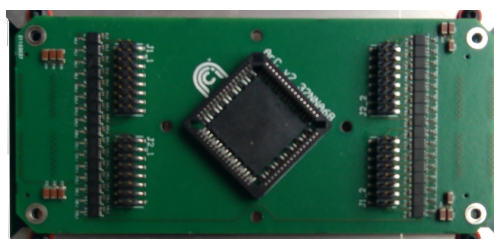
FIGURE E.1: A photograph of the 32NNA68 daughterboard.

The 32NNA68 is the default daughterboard for the ArC TWO (Fig. E.1). It has a socket for the PLCC68 package used for selectorless crossbar array in the research group and has a bank of electronically disconnectable header pins for use with external circuits and instruments such as a probe station. A half scale version for use with the prototype was made with the designation 16NNA68, but was never used.

E.2 32SLP48DIP
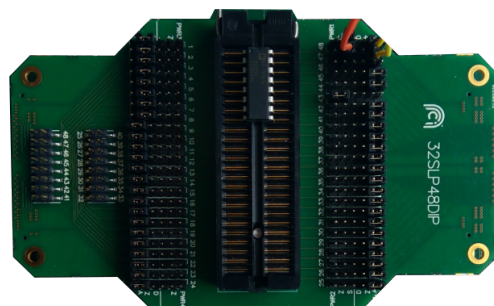
FIGURE E.2: A photograph of the 32SLP48DIP daughterboard.

The 32SLP48DIP is an alternative to the 32NNA68 that mounts a DIP/ZIF socket instead of the PLCC (Fig. E.2). It only connects to 48 of the analogue channels, but permits the connection of the DUT supply circuits, along with the selectors and digital IO channels. Selecting which function each pin of the ZIF socket is assigned is done by placing a jumper on the associated connection header.

E.3   32BNC12/32SMA32
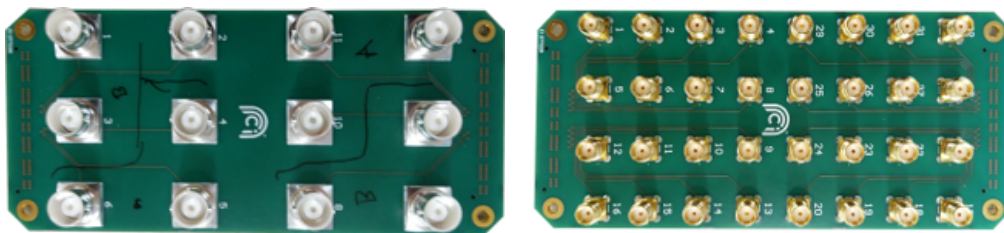
FIGURE E.3: Left: A photograph of the 32BNC12 daughterboard. Right: A photograph of the 32SMA32 daughterboard.

At the request of another researcher, two variants of coaxial daughterboards were made (Fig. E.3). These designs act as simple breakout boards for the analogue channels, allowing connection to coaxial cable systems.

E.4   32NNA68VAR
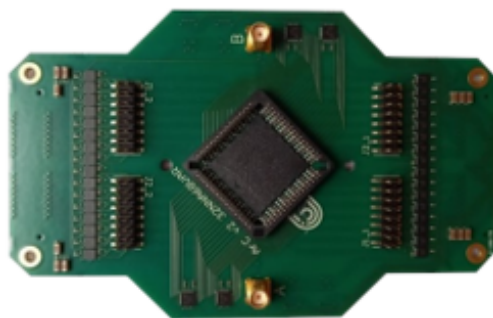
FIGURE E.4: A photograph of the 32NNA68VAR daughterboard.

To address the needs of a researcher trying to bias DUTs from an external coaxial source, a variant of the 32NNA68 daughterboard was design that adds a pair of $1:64$ multiplexers on either side of the PLCC socket, which are operated through the digital IO of the instrument (Fig. E.4). This daughterboard retains the functionality of the 32NNA68, although the input of the header disconnect switches is inverted, due to IC supply issues related to the pandemic.

E.5 TXL daughterboards

Three daughterboards were produced for the tests on the Split TXL circuit covered in this thesis (Fig. E.5). More details can be found in chapter 5.
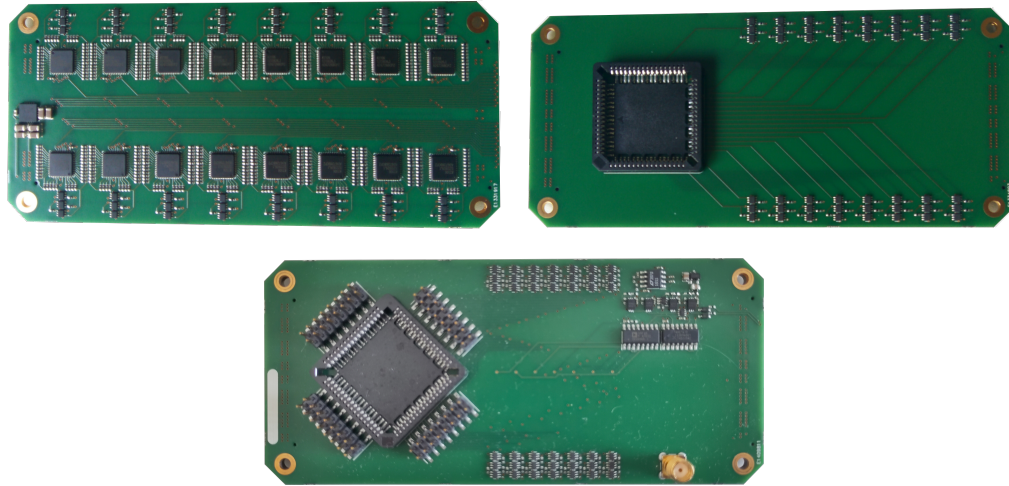
FIGURE E.5: Top Left: A photograph of the resistor version of the PCB model of the TXL. Top Right: A photograph of the memristor version of the PCB model of the TXL. Bottom: A photograph of the TXL channel demonstrator.

Appendix F

ArC TWO Contributions

F.1   FPGA Configuration
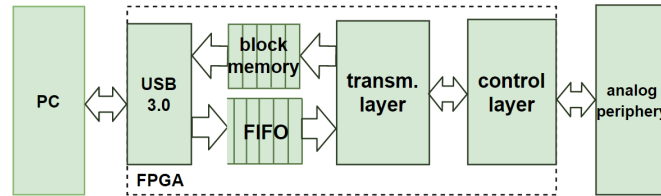
FIGURE F.1: A diagram of the architecture of the ArC TWO control system.

The FPGA configuration used by the ArC TWO was designed by Jinqi Huang (Fig. F.1).
It operates by processing 256-bit instructions from a first-in first-out (FIFO) command
buffer. Each instruction corresponds with a low-level functionality (Tab. F.1), such as
commanding a reading or connecting a channel to ground. The instruction is processed
by a decoder that routes the data contained in the instruction to the appropriate serial-
parallel interface (SPI) block and associated state machine. The SPI blocks send the
requested information to the ICs on the board, effecting the instruction. Readings
generated by instructions are stored in the RAM of the FPGA module, and read back
at the request of the host computer. While this architecture provides great flexibility,
the passive nature of the its theory of operation somewhat limits the capability of the
instrument.

F.1.1   AMP PRP

During development, it was found that the TIA design was flawed. To reduce the
access resistance of the TIA, the switch that connects the TIA disconnects the feedback,
rather than the input. As such, when a channel is set to open circuit of pulse mode,
the TIA op-amp is open loop and immediately saturates. When closing the feedback

loop to return to voltage bias or current read mode, the output voltage of the op-amp takes time to recover, appearing on the channel as a transient. This transient can be many microseconds long and reach voltages in excess of 12 V, which is problematic for the use of the instrument in the intended voltage sensitive applications. To mitigate this, an instruction was added that steps backwards through the ranges of the TIA to slow the response of the channel to the recovering op-amp output. Future expansion of the instruction set has been considered, to include instructions that make decisions or generate sequences of instructions required for more complex operations, although such expansion remains outside the scope of this project at this time.

Opcode	Instruction	Functionality
0x00001	LD VOLT	Loads 8 16-bit voltage values to registers in the FPGA to be sent to the DACs on the next UP DAC.
0x00002	UP DAC	Prompts a serial frame that sends the values written by LD VOLT to their respective DACs.
0x00004	C READ	Begins a current read operation with an autoranging routine on the channels specified in a 64-bit bitmask.
0x00008	V READ	Takes a single or 32 sample averaged voltage reading on the channels specified in a 64-bit bitmask.
0x00010	UP SEL	Sets selector channels specified in the 32-bit bitmask to a high/low state.
0x00020	UP LGC	Sets arbitrary logic channels specified in the 32-bit bitmask to a high/low state.
0x00040	UP CH	Sets the function (open circuit/voltage bias/HS pulse) of the analogue channels and configures the current source.
0x00080	CLEAR	Opens all switches and sets all DACs to 0V, effectively a software reset.
0x00100	HS CON	Sets the timer registers for the HS pulse signals of each cluster.
0x00200	UP SEL	Begins a pulse on the specified cluster of specified polarity, with length determined by HS CON.
0x00400	MOD CH	Sets the state of the GND/AC GND/Current Source switches on each channel.
0x01000	LD OFF	Loads 8 16-bit voltage values to registers to offset values being sent to the DACs. Intended for calibration.
0x02000	DELAY	Adds a delay before the execution of the next command.
0x04000	DAC RNG	Changes the rage of specified DAC channels to either ±10 V or ±20 V(±13.5 V)
0x08000	HS PAT	Begins an arbitrary series of high/low pulse states with configurable high/low durations.
0x10000	AMP PRP	Steps the specified analogue channels through their ranges, from most to least sensitive.

TABLE F.1: Chart of the ArC TWO instruction set.

F.2   Software

The instruction set is supported by a custom Python library written by Spyros Stathopoulos. The library implements basic operations, such as voltage bias and ramp operations, by stringing together instructions and loading them in batches to the FIFO buffer of the FPGA. It also converts values to and from the hexadecimal format used by the instructions and readouts, automatically compensating for the gain of the TIA. In addition to this, the library also tracks which channels have been set to open circuit, and automatically inserts an AMP PRP and associated instructions to suppress transients.

Bibliography

[1] Pierre Yger, Giulia Spampinato, Elric Esposito, Baptiste Lefebvre, Stéphane Deny, Christophe Gardella, Marcel Stimberg, Florian Jetter, Günther Zeck, Serge Picaud, Jens Duebel, and Olivier Marre. A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. *eLife*, 7, 03 2018.

[2] David Carlson and Lawrence Carin. Continuing progress of spike sorting in the era of big data. *Current Opinion in Neurobiology*, 55:90–96, 04 2019.

[3] TM Seese, H Harasaki, GM Saidel, and CR Davies. Characterization of tissue morphology, angiogenesis, and temperature in the adaptive response of muscle tissue to chronic heating. *Laboratory investigation; a journal of technical methods and pathology*, 78(12):1553—1562, December 1998.

[4] Pan Ke Wang, Sio Hang Pun, Chang Hao Chen, Elizabeth A. McCullagh, Achim Klug, Anan Li, Mang I. Vai, Peng Un Mak, and Tim C. Lei. Low-latency single channel real-time neural spike sorting system based on template matching. *PLOS ONE*, 14(11):1–30, 11 2019.

[5] Isha Gupta, Alexantrou Serb, Ali Khiat, Maria Trapatseli, and Themistoklis Prodromakis. Spike sorting using non-volatile metal-oxide memristors. *Faraday Discuss.*, 213:511–520, 2019.

[6] Majid Zamani, Dai Jiang, and Andreas Demosthenous. An adaptive neural spike processor with embedded active learning for improved unsupervised sorting accuracy. *IEEE Transactions on Biomedical Circuits and Systems*, 12:665–676, 2018.

[7] A. Eftekhar, E. P. Sivylla, and G. C. Timothy. Towards a next generation neural interface: Optimizing power, bandwidth and data quality. In *2010 Biomedical Circuits and Systems Conference (BioCAS)*, pages 122–125, Nov 2010.

[8] Hernan Rey, Carlos Pedreira, and Rodrigo Quian. Past, present and future of spike sorting techniques. *Brain Research Bulletin*, 65, 04 2015.

[9] Daniel Valencia and Amirhossein Alimohammad. An efficient hardware architecture for template matching-based spike sorting. *IEEE Transactions on Biomedical Circuits and Systems*, 13(3):481–492, 2019.

[10] Isha Gupta, Alexantrou Serb, Ali Khiat, Ralf Zeitler, Stefano Vassanelli, and Themistoklis Prodromakis. Real-time encoding and compression of neuronal spikes by metal-oxide memristors. *Nature Communications*, 7(1):12805, Sep 2016.

[11] Jiaqi Wang, Alexantrou Serb, Christos Papavassiliou, Sachin Maheshwari, and Themis Prodromakis. Analysing and measuring the performance of memristive integrating amplifiers. *International Journal of Circuit Theory and Applications*, 49(11):3507–3525, 2021.

[12] Alexander Serb, Ali Khiat, and Themis Prodromakis. Seamlessly fused digital-analogue reconfigurable computing using memristors. *Nature Communications*, 9, 12 2018.

[13] P. Foster, J. Huang, A. Serb, T. Prodromakis, and C. Papavassiliou. An fpga based system for interfacing with crossbar arrays. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, 2020.

[14] Patrick Foster, Jinqi Huang, Alex Serb, Spyros Stathopoulos, Christos Papavassiliou, and Themis Prodromakis. An fpga-based system for generalised electron devices testing. *Scientific Reports*, 2022.

[15] Patrick Foster, Alex Serb, and Themis Prodromakis. A dual threshold analogue content addressable memory. *arXiv preprint arXiv:2303.02651*, 2023.

[16] Alexandra V. Ulyanova, Carlo Cottone, Christopher D. Adam, Kimberly G. Gagnon, D. Kacy Cullen, Tahl Holtzman, Brian G. Jamieson, Paul F. Koch, H. Isaac Chen, Victoria E. Johnson, and John A. Wolf. Multichannel silicon probes for awake hippocampal recordings in large animals. *Frontiers in Neuroscience*, 13:397, 2019.

[17] T. Horiuchi, T. Swindell, D. Sander, and P. Abshier. A low-power cmos neural amplifier with amplitude measurements for spike sorting. In *2004 IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 4, pages IV–29, 2004.

[18] Sarah Gibson. Neural spike sorting in hardware: From theory to practice, 2012.

[19] Adam Marblestone, Bradley Zamft, Yael Maguire, Mikhail Shapiro, Thaddeus Cybulski, Joshua Glaser, Dario Amodei, P. Benjamin Stranges, Reza Kalhor, David Dalrymple, Dongjin Seo, Elad Alon, Michel Maharbiz, Jose Carmena, Jan Rabaey, Edward Boyden, George Church, and Konrad Kording. Physical principles for scalable neural recording. *Frontiers in Computational Neuroscience*, 7, 2013.

[20] David B. Salkoff, Edward Zagha, Özge Yüzgeç, and David A. McCormick. Synaptic mechanisms of tight spike synchrony at gamma frequency in cerebral cortex. *Journal of Neuroscience*, 35(28):10236–10251, 2015.

[21] Bo Wang, Wei Ke, Jing Guang, Guang Chen, Luping Yin, Suixin Deng, Quansheng He, Yaping Liu, Ting He, Rui Zheng, Yanbo Jiang, Xiaoxue Zhang, Tianfu Li,

Guoming Luan, Haidong D. Lu, Mingsha Zhang, Xiaohui Zhang, , and Yousheng Shu1. Firing frequency maxima of fast-spiking neurons in human, monkey, and mouse neocortex. *Front Cell Neuroscience*, October 2016.

[22] Gaute T. Einevoll, Christoph Kayser, Nikos K. Logothetis, and Stefano Panzeri. Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nature Reviews Neuroscience*, 14(11):770–785, Nov 2013.

[23] Yoshinao Kajikawa and Charles E Schroeder. How local is the local field potential? *Neuron*, 72(5):847–858, 2011.

[24] T. Lenarz. Cochlear implant - state of the art. *GMS Curr Top Otorhinolaryngol Head Neck Surg.*, 16:Doc04, 2018.

[25] F. Hashemi Noshahr, M. Nabavi, and M. Sawan. Multi-channel neural recording implants: A review. *Sensors (Basel)*, 20(3):904, 2020.

[26] Edwin M. Maynard, Craig T. Nordhausen, and Richard A. Normann. The utah intracortical electrode array: A recording structure for potential brain-computer interfaces. *Electroencephalography and Clinical Neurophysiology*, 102(3):228–239, 1997.

[27] M. Malik, M. Saeed, and A. Kamboh. Automatic threshold optimization in nonlinear energy operator based spike detection. In *Annu Int Conf IEEE Eng Med Biol Soc.*, volume 2016, pages 774–777, 2016.

[28] T. Chen, K. Chen, Z. Yang, K. Cockerham, and W. Liu. A biomedical multiprocessor soc for closed-loop neuroprosthetic applications. In *2009 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, pages 434–435,435a, 2009.

[29] Takashi Sato, Takafumi Suzuki, and Kunihiko Mabuchi. Fast automatic template matching for spike sorting based on davies-bouldin validation indices. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3200–3203, 2007.

[30] Keven J. Laboy-Juárez, Sei Ahn, and Daniel E. Feldman. A normalized template matching method for improving spike detection in extracellular voltage recordings. *bioRxiv*, 2018.

[31] T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc.*, 374:20150202, 2016.

[32] C. Rocío Caro-Martín, José Delgado-García, Agnès Gruart, and Raudel Sánchez-Campusano. Spike sorting based on shape, phase, and distribution features, and k-tops clustering with validity and error indices. *Scientific Reports*, 8, 12 2018.

[33] M. J. Bak and E. M. Schmidt. An improved time-amplitude window discriminator. *IEEE Transactions on Biomedical Engineering*, BME-24(5):486–489, 1977.

[34] Can Li, Catherine Graves, Xia Sheng, Darrin Miller, Martin Foltin, Giacomo Pedretti, and John William Strachan. Analog content-addressable memories with memristors. *Nature Communications*, 11, 04 2020.

[35] Zichen Fan, Ziru Li, Bing Li, Yiran Chen, and Hai Helen Li. Red: A reram-based deconvolution accelerator. *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1763–1768, 2019.

[36] Mohammed Zidan, Yeonjoo Jeong, Jong Shin, Chao Du, and Wei Lu. Field-programmable crossbar array (fpca) for reconfigurable computing. *IEEE Transactions on Multi-Scale Computing Systems*, PP, 06 2017.

[37] H. Chen, C. Chen, and W. Hwang. An efficient hardware circuit for spike sorting based on competitive learning networks. *Sensors (Basel)*, 17(10):2232, 2017.

[38] Zhengwu Liu, Jianshi Tang, Bin Gao, Peng Yao, Xinyi Li, Dingkun Liu, Ying Zhou, He Qian, Bo Hong, and Huaqiang Wu. Neural signal analysis with memristor arrays towards high-efficiency brain–machine interfaces. *Nature Communications*, 11:4234, 08 2020.

[39] Akira Goda. Recent progress on 3d nand flash technologies. *Electronics*, 10(24), 2021.

[40] N. Papandreou, A. Pantazi, A. Sebastian, M. Breitwisch, C. Lam, H. Pozidis, and E. Eleftheriou. Multilevel phase-change memory. In *2010 17th IEEE International Conference on Electronics, Circuits and Systems*, pages 1017–1020, 2010.

[41] Yu-Der Chih, Yi-Chun Shih, Chia-Fu Lee, Yen-An Chang, Po-Hao Lee, Hon-Jarn Lin, Yu-Lin Chen, Chieh-Pu Lo, Meng-Chun Shih, Kuei-Hung Shen, Harry Chuang, and Tsung-Yung Jonathan Chang. 13.3 a 22nm 32mb embedded stt-mram with 10ns read speed, 1m cycle write endurance, 10 years retention at 150˚c and high immunity to magnetic field interference. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pages 222–224, 2020.

[42] S. Dünkel, M. Trentzsch, R. Richter, P. Moll, C. Fuchs, O. Gehring, M. Majer, S. Wittek, B. Müller, T. Melde, H. Mulaosmanovic, S. Slesazeck, S. Müller, J. Ocker, M. Noack, D.-A. Löhr, P. Polakowski, J. Müller, T. Mikolajick, J. Höntschel, B. Rice, J. Pellerin, and S. Beyer. A fefet based super-low-power ultra-fast embedded nvm technology for 22nm fdsoi and beyond. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 19.7.1–19.7.4, 2017.

[43] Wei Wu, Huaqiang Wu, Bin Gao, Peng Yao, Xiang Zhang, Xiaochen Peng, Shimeng Yu, and He Qian. A methodology to improve linearity of analog rram for

neuromorphic computing. In *2018 IEEE Symposium on VLSI Technology*, pages 103–104, 2018.

[44] Chao Li, Wendy Fan, Bo Lei, Daihua Zhang, Song Han, Tao Tang, Xiaolei Liu, Zuqin Liu, Sylvia Asano, Meyya Meyyappan, Jie Han, and Chongwu Zhou. Multi-level memory based on molecular devices. *Applied Physics Letters*, 84:1949 – 1951, 04 2004.

[45] Seung Soo Kim, Soo Kyeom Yong, Whayoung Kim, Sukin Kang, Hyeon Woo Park, Kyung Jean Yoon, Dong Sun Sheen, Seho Lee, and Cheol Seong Hwang. Review of semiconductor flash memory devices for material and process issues. *Advanced Materials*, n/a(n/a):2200659, 2022.

[46] Scott W. Fong, Christopher M. Neumann, and H.-S. Philip Wong. Phase-change memory—towards a storage-class memory. *IEEE Transactions on Electron Devices*, 64(11):4374–4385, 2017.

[47] Kevin Garello, Farrukh Yasin, and Gouri Sankar Kar. Spin-orbit torque mram for ultrafast embedded memories: from fundamentals to large scale technology integration. In *2019 IEEE 11th International Memory Workshop (IMW)*, pages 1–4, 2019.

[48] Min-Che Hsieh, Yu-Cheng Liao, Yung-Wen Chin, Chen-Hsin Lien, Tzong-Sheng Chang, Yue-Der Chih, Sreedhar Natarajan, Ming-Jinn Tsai, Ya-Chin King, and Chrong Jung Lin. Ultra high density 3d via rram in pure 28nm cmos process. In *2013 IEEE International Electron Devices Meeting*, pages 10.3.1–10.3.4, 2013.

[49] A. Pirovano, A. Redaelli, F. Pellizzer, F. Ottogalli, M. Tosi, D. Ielmini, A.L. Lacaita, and R. Bez. Reliability study of phase-change nonvolatile memories. *IEEE Transactions on Device and Materials Reliability*, 4(3):422–427, 2004.

[50] Min-Kyu Kim, Ik-Jyae Kim, and Jang-Sik Lee. Cmos-compatible ferroelectric nand flash memory for high-density, low-power, and high-speed three-dimensional memory. *Science Advances*, 7(3):eabe1341, 2021.

[51] Shyh-Shyuan Sheu, Pei-Chia Chiang, Wen-Pin Lin, Heng-Yuan Lee, Pang-Shiu Chen, Yu-Sheng Chen, Tai-Yuan Wu, Frederick T. Chen, Keng-Li Su, Ming-Jer Kao, Kuo-Hsing Cheng, and Ming-Jinn Tsai. A 5ns fast write multi-level non-volatile 1 k bits rram memory with advance write scheme. In *2009 Symposium on VLSI Circuits*, pages 82–83, 2009.

[52] S Pookpanratana, H Zhu, E. Bittle, Sean Natoli, T Ren, Curt Richter, Q Li, and C Hacker. Non-volatile memory devices with redox-active diruthenium molecular compound. *Journal of physics. Condensed matter : an Institute of Physics journal*, 28:094009, 02 2016.

[53] Marcello Calabrese, Carmine Miccoli, Christian Monzio Compagnoni, Luca Chiavarone, Silvia Beltrami, Andrea Parisi, Sebastiano Bartolone, Andrea L. Lacaita, Alessandro S. Spinelli, and Angelo Visconti. Accelerated reliability testing of flash memory: Accuracy and issues on a 45nm nor technology. In *Proceedings of 2013 International Conference on IC Design & Technology (ICICDT)*, pages 37–40, 2013.

[54] Sadahiko Miura, Koichi Nishioka, Hiroshi Naganuma, T. V. A. Nguyen, Hiroaki Honjo, Shoji Ikeda, Toshinari Watanabe, Hirofumi Inoue, Masaaki Niwa, Takaho Tanigawa, Yasuo Noguchi, Toru Yoshizuka, Mitsuo Yasuhira, and Tetsuo Endoh. Scalability of quad interface p-mtj for 1x nm stt-mram with 10-ns low power write operation, 10 years retention and endurance ¿ 10^{11}. *IEEE Transactions on Electron Devices*, 67(12):5368–5373, 2020.

[55] Shigeki Sakai and Mitsue Takahashi. Recent progress of ferroelectric-gate field-effect transistors and applications to nonvolatile logic and fenand flash memory. *Materials*, 3(11):4950–4964, 2010.

[56] Furqan Zahoor, Tun Zainal Azni Zulkifli, and Farooq Ahmad Khanday. Resistive random access memory (rram): an overview of materials, switching mechanism, performance, multilevel cell (mlc) storage, modeling, and applications. *Nanoscale Research Letters*, 15(1):90, Apr 2020.

[57] Simona Boboila and Peter Desnoyers. Write endurance in flash drives: Measurements and analysis. In *FAST*, pages 115–128, 2010.

[58] Zhiming Liu, Amir A. Yasseri, Jonathan S. Lindsey, and David F. Bocian. Molecular memories that survive silicon device processing and real-world operation. *Science*, 302(5650):1543–1545, 2003.

[59] F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka. A new flash e2prom cell using triple polysilicon technology. In *1984 International Electron Devices Meeting*, pages 464–467, 1984.

[60] Manuel Le Gallo and Abu Sebastian. An overview of phase-change memory device physics. *Journal of Physics D: Applied Physics*, 53(21):213002, mar 2020.

[61] T. Miyazaki and N. Tezuka. Giant magnetic tunneling effect in fe/al2o3/fe junction. *Journal of Magnetism and Magnetic Materials*, 139(3):L231–L234, 1995.

[62] B.N. Engel, J. Akerman, B. Butcher, R.W. Dave, M. DeHerrera, M. Durlam, G. Grynkewich, J. Janesky, S.V. Pietambaram, N.D. Rizzo, J.M. Slaughter, K. Smith, J.J. Sun, and S. Tehrani. A 4-mb toggle mram based on a novel bit and switching method. *IEEE Transactions on Magnetics*, 41(1):132–136, 2005.

[63] A V Khvalkovskiy, D Apalkov, S Watts, R Chepulskii, R S Beach, A Ong, X Tang, A Driskill-Smith, W H Butler, P B Visscher, D Lottis, E Chen, V Nikitin, and

M Krounbi. Basic principles of stt-mram cell operation in memory arrays. *Journal of Physics D: Applied Physics*, 46(7):074001, jan 2013.

[64] Cheol Seong Hwang and Thomas Mikolajick. 11 - ferroelectric memories. In Blanka Magyari-Köpe and Yoshio Nishi, editors, *Advances in Non-Volatile Memory and Storage Technology (Second Edition)*, Woodhead Publishing Series in Electronic and Optical Materials, pages 393–441. Woodhead Publishing, second edition edition, 2019.

[65] L. O. Chua and Sung Mo Kang. Memristive devices and systems. *Proceedings of the IEEE*, 64(2):209–223, 1976.

[66] L. Chua. Memristor-the missing circuit element. *IEEE Transactions on Circuit Theory*, 18(5):507–519, 1971.

[67] Dmitri B. Strukov, Gregory S. Snider, Duncan R. Stewart, and R. Stanley Williams. The missing memristor found. *nature*, 453(7191):80–83, May 2008.

[68] S. Vongehr and X. Meng. The missing memristor has not been found. *Sci Rep.*, 5:11657, 2015.

[69] J Joshua Yang, Feng Miao, Matthew D Pickett, Douglas A A Ohlberg, Duncan R Stewart, Chun Ning Lau, and R Stanley Williams. The mechanism of electroforming of metal oxide memristive switches. *Nanotechnology*, 20(21):215201, may 2009.

[70] H. Y. Lee, P. S. Chen, T. Y. Wu, Y. S. Chen, C. C. Wang, P. J. Tzeng, C. H. Lin, F. Chen, C. H. Lien, and M. . Tsai. Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust hfo2 based rram. In *2008 IEEE International Electron Devices Meeting*, pages 1–4, 2008.

[71] J. Rao, Z. Fan, L. Hong, S. Cheng, Q. Huang, J. Zhao, X. Xiang, E.-J. Guo, H. Guo, Z. Hou, Y. Chen, X. Lu, G. Zhou, X. Gao, and J.-M. Liu. An electroforming-free, analog interface-type memristor based on a srfeox epitaxial heterojunction for neuromorphic computing. *Materials Today Physics*, 18:100392, 2021.

[72] C. Evans. Conversation: Jay w. forrester. *Annals of the History of Computing*, 5(3):297–301, 1983.

[73] F. Merrikh-Bayat, S. B. Shouraki, and A. Rohani. Memristor crossbar-based hardware implementation of the ids method. *IEEE Transactions on Fuzzy Systems*, 19(6):1083–1096, 2011.

[74] Y. Jeong, M. A. Zidan, and W. D. Lu. Parasitic effect analysis in memristor-array-based neuromorphic systems. *IEEE Transactions on Nanotechnology*, 17(1):184–193, Jan 2018.

[75] Sherif Amer, Ahmed Madian, Hany ElSayed, and Ahmed Emara. Effect of the memristor threshold current on memristor-based min-max circuits. In *2016 5th International Conference on Modern Circuits and Systems*, pages 1–4, 05 2016.

[76] Daniel Wust et al. Prototyping memristors in digital system with an fpga-based testing environment. In *2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, pages 1–7, 2017.

[77] R Berdan et al. A μ-Controller-Based System for Interfacing Selectorless RRAM Crossbar Arrays. *IEEE Transactions on Electron Devices*, 62(7), 2015.

[78] Yuhan Wang et al. High speed test system of current pulse for phase change memory devices. *Journal of Physics: Conference Series*, 1237:042064, 2019.

[79] Emmanuelle Merced-Grafals et al. Repeatable, accurate, and high speed multi-level programming of memristor 1t1r arrays for power efficient analog computing applications. *Nanotechnology*, 27:365202, 2016.

[80] Radu Berdan, Alexander Serb, Ali Khiat, Anna Regoutz, C. Papavassiliou, and Themis Prodromakis. A μ-controller-based system for interfacing selectorless rram crossbar arrays. *Electron Devices, IEEE Transactions on*, 62:2190–2196, 07 2015.

[81] Olufemi Akindele Olumodeji and Massimo Gottardi. A pulse-based memristor programming circuit. *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, 2017.

[82] Keithley. *Semiconductor Characterization System Technical Data, Keithley 4200-SCS.*, 2009.

[83] DIGILENT. *Analog Discovery 2™ Reference manual*, 2015.

[84] Knowm. *Memristor Discovery Manual*, 2019.

[85] Diodes Inc. *1N4148/1N4448 Datasheet*, 2008.

[86] STMicroelectronics. *2N7000 Datasheet*, 2008.

[87] Toshiba. *SSM6L36TU LF Datasheet*, 2014.

[88] Andrea Zaffora, Francesco Di Franco, Roberto Macaluso, and Monica Santamaria. 17 - tio$_2$ in memristors and resistive random access memory devices. In Francesco Parrino and Leonardo Palmisano, editors, *Titanium Dioxide (TiO$_2$) and Its Applications*, Metal Oxides, pages 507–526. Elsevier, 2021.

[89] Shi H. Sun, Ali Almasi, Molis Yunzab, Syeda Zehra, Damien G. Hicks, Tatiana Kameneva, Michael R. Ibbotson, and Hamish Meffin. Analysis of extracellular spike waveforms and associated receptive fields of neurons in cat primary visual cortex. *The Journal of Physiology*, 599(8):2211–2238, 2021.

[90] Tasnim B. Nazzal and Soliman A. Mahmoud. Low-power bootstrapped sample and hold circuit for analog-to-digital converters. In *2016 IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1–4, 2016.

[91] Seon-Kyoo Lee, Seung-Jin Park, Hong-June Park, and Jae-Yoon Sim. A 21 fj/conversion-step 100 ks/s 10-bit adc with a low-noise time-domain comparator for low-power sensor interface. *IEEE Journal of Solid-State Circuits*, 46(3):651–659, 2011.