Complex Oxides for Computing Beyond von Neumann
Goossens, Anouk

DOI:
10.33612/diss.766596376

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 8

Conclusions and Outlook

This thesis has taken a look at devices based on complex oxides for novel computing architectures. With the wealth of foreseeable applications of computing using memristive devices, having access to a large variety of devices with different properties is a great advantage towards implementation. In this respect, we found complex oxide devices to offer interesting properties that are easily tunable through material and device characteristics. This makes them potentially suitable for a wide range of applications.

8.1 Nb-doped SrTiO$_3$ Interface Memristors

Chapters 3-5 investigated interface memristors based on Schottky junctions on Nb-doped SrTiO$_3$ (Nb:STO). These devices show bipolar resistive switching, whereby the resistance decreases when a forward bias voltage is applied and increases under the application of a reverse bias voltage. This behaviour is fully analogue and both the set and reset operations are gradual, allowing for easy control of the resistance state. Unlike most ionic systems relying on filamentary conduction for the switching, this behaviour is present in as-fabricated devices, which is a great advantage in terms of performance and integration. The partial dielectric breakdown that occurs during typical high voltage forming steps is difficult to control and can induce greater variation between devices as well as device failure. These issues are further elevated when concerned with network integration where individual devices are harder to access and each may have to be electroformed individually. We have developed a reliable method of device fabrication that gives rise to minimal device-to-device variation and high effective fabrication yields. The device structures are also relatively simple which is beneficial for integration.

In general, these devices were found to provide high storage capacity: the memory window is large and many intermediate states can be accessed. Additionally, their durability and stability are high: the devices can be kept at ambient conditions and can withstand a large number ($>10^5$) of cycles without degrading.

We found the memristive properties to be highly tunable; in Chapter 4 we studied the effect of changes the device area and determined that areal downscaling results
in an increase in the resistance window. In Chapter 5 we looked at how the semiconductor’s doping concentration affects device performance and we find the doping concentration leads to an increase in the current in both bias directions, an increase in the resistance ratio, a greater symmetry between reverse and forward bias and faster switching.

Furthermore, our results provide insight into the physics underpinning the switching mechanisms. Through scanning transmission electron microscopy and retention measurements we confirm both ionic movement in the form of electric field-controlled oxygen vacancy migration and electronic trapping of charges to be important aspects of resistive switching. In literature, the switching mechanism at Nb:STO/metal interfaces is much debated (Section 3.2) and we find that while the resistive switching does not occur at localised filamentary regions, in our devices it also does not happen completely homogeneously over the entire device area. Instead, the enhancement of the electric field around the device perimeter gives rise to a change in the energy landscape and aids the switching in a well-controlled way. Similarly, varying the doping concentration changes the interfacial electric field and subsequently the energy landscape. This highlights the power of complex oxides as this intimate relation between the interfacial energy profile and the electric field is largely enabled by the unconventional field dependence of the dielectric permittivity of STO.

8.2 SrRuO$_3$/SrTiO$_3$-based Magnetic Tunnel Junctions

We have demonstrated the ability to manipulate the magnetic anisotropy of SrRuO$_3$ (SRO) ferromagnetic layers by the choice of substrate. Due to the close lattice match between SRO and STO, these layers can be made relatively thick and integrated into stacks without loss of strain and substrate-controlled growth. We showed that the addition of a heavy metal layer on top of SRO can provide current-induced spin-orbit torque which could be used to manipulate the magnetisation. When an STO layer is deposited between two SRO layers of different thicknesses, the magnetisations of the layers can be switched independently allowing clear parallel and antiparallel states to be defined.

Depending on how the anisotropy is tailored, such magnetic tunnel junctions have either perfect perpendicular magnetic anisotropy and a corresponding high symmetry or a tilted anisotropy giving rise to a breaking of rotational symmetry. We envision that this could allow SRO-based MTJs to be used for both probabilistic and deterministic applications. In Chapter 6 we proposed a potential device design where magnetisation can be switched using an additional Pt layer. A great advantage of this device design is its relatively straightforward structure compared to the
8.3. Outlook

It is clear that memristive architectures are an active area field of research. There are challenges that are being investigated at different levels that include material and device engineering, architecture engineering, array integration and upscaling, and development and implementation of neural network algorithms.\[5-43\]

8.3.1 Material Considerations

CMOS technology has become a critical part of current-day computing, with silicon being the most widely used semiconductor in the electronics industry. Silicon is the second most abundant element in the Earth’s crust, making it widely available and relatively inexpensive compared to alternative materials.\[44\]. It is also a stable material and can withstand high temperatures, radiation, and other harsh conditions. It is tunable: by adding dopant atoms it can become either a p-type or n-type semiconductor, allowing it to be used for diodes, transistors, and other electronic components. Finally, due to its long and widespread use, the manufacturing processes used
8. Conclusions and Outlook

in semiconductor fabrication plants are specialised to manufacture silicon-based devices allowing for ease of mass-scale fabrication at a low cost.

While memristors tend to have relatively simple device structures, the manufacturing costs are typically high because the custom back-end-of-line interconnections fall outside the standard processing used for CMOS technology. As a result, the production of memristive devices on a commercial scale is still costly[45]. Hence, while the operating costs and power consumption of neuromorphic systems may be substantially lower, their manufacturing costs are likely to be significantly higher than a system utilising the von Neumann architecture.

The material most central to the work presented in this thesis is SrTiO$_3$; all the constituents elements of SrTiO$_3$ are naturally abundant and are easily sourceable[46]. Consequently, the environmental impact associated with mining and extraction is relatively low. Additionally, it has relatively low toxicity and does not pose a significant risk to human health or the environment.

Due to the versatile additional functionalities that complex oxides can provide to the electronics industry, the compatibility of these materials with silicon is an active area of research. In particular, many works focus on growing STO on silicon. This was first realised by McKee et al. in 1998 using molecular beam epitaxy (MBE)[47] and has since been extended by many others[48]. The possibility of epitaxial growing near single-crystal STO on large 8-inch Si wafers has also been experimentally demonstrated[49]. In addition to MBE growth, PLD growth of STO on Si is also being researched[50].

8.3.2 Applications

The von Neumann architecture utilises sequential processing while neuromorphic computing is optimised for parallel processing. This naturally means the architectures are inherently suitable for different tasks.

Sequential processing is more suitable for tasks that involve processing data in a specific order and making decisions based on the results of previous operations. For example, sorting a list of numbers requires operations to be carried out sequentially. Similarly, searching for items in a list and prioritising tasks typically involves processing data in a specific order. At the same time, the digital representation of data provides high precision, reliability and determinacy, making traditional computers more adept at performing precise calculations[51].

Parallel computing, on the other hand, is more suited for tasks that can be divided into independent subtasks that can be executed simultaneously. This is particularly useful for applications such as image and video processing where multiple tasks, such as feature extraction, filtering and segmentation can be carried out independently. Data analysis tasks, such as classification, clustering and pattern recognition.
Optimisation problems, such as finding the shortest path in a graph. Neuromorphic systems can also efficiently process unstructured data and perform image recognition and classification of noisy datasets. They are adaptable which allows them to carry out signal processing resilient to noise and stochasticity and perform learning in hardware[51].

An advantage of the von Neumann architecture is its flexibility so that it can be programmed to perform a wide range of tasks. Hence, computers are general-purpose machines that can execute a wide variety of tasks, albeit at a significant energy cost for certain operations. For these reasons, it is unlikely that neuromorphic systems could completely replace conventional computing platforms. Instead, the best solution is to combine the approaches for different tasks.

Ternary content addressable memory (TCAM)

Ternary content addressable memory (TCAM) is an important type of memory heavily used for the internet to perform fast and efficient packet forwarding and filtering. CMOS-based TCAM cells typically have high energy consumption. Recently, a TCAM design in which part of the transistors are replaced by memristors has been proposed. Simulations based on experimental data from Nb:STO interface memristors suggest that these devices could significantly lower the power consumption and double the resource density compared to the traditional architectures[52, 53]. These results underline TCAM as one of the potential applications of the devices investigated here.

8.3.3 Integration

Depending on the desired application, the required memristor properties may differ and it is unlikely that a single memristor will be suitable for all envisioned implementations. However, a large number of memristive materials and devices have been investigated, and consequently, a vast amount of device characteristics are available. Hence, at this stage, the availability of memristive properties is not a main factor hindering implementation and depending on the application it is likely that a suitable material can be identified. So far the focus has largely been on studying individual devices and the connectivity between devices needs to be studied. A significant challenge now is building networks and achieving high degrees of interconnectivity between devices. Neurons in the human brain typically have around $10^3$-$10^4$ connections, which is enabled by the brain’s three-dimensional structure. This level of interconnection is beyond the reach of current electronics, as CMOS technology is mostly confined to two-dimensional structures[9, 13]. Hence, there should be a focus on realising architectures that can integrate memristive devices in
networks.

In this respect, both connections between memristive devices, as well as connections to the chip as a whole need to be considered. One of the leading candidates connecting memristive elements is crossbar arrays. A crossbar array consists of a grid of intersecting vertical and horizontal wires with memristive devices located at the intersection points between a pair of perpendicular lines. The rows and columns can be addressed independently, allowing any memristor to be accessed. They are highly scalable: they can easily be expanded in two dimensions by adding more rows and columns and in three dimensions by stacking multiple layers of arrays on top of each other with each layer connected to the layer above and below it through vertical interconnects.

**Crossbar**

![Crossbar diagram](image)

**Figure 8.1: Crossbar fabrication.** (a)-(b) Deposition of a thin layer of Nb:STO on STO substrate. (c)-(d) Lithography and etching to define an array of vertical bottom electrodes. (e)-(f) Lithography and Al₂O₃ deposition to define insulation layer and define memristive contact areas. (g)-(h) Lithography and Co/Au deposition to define an array of horizontal top electrodes.

We propose a device design to integrate the interface memristors discussed in this thesis in a crossbar array on an STO substrate; the fabrication protocol is outlined in Fig. 8.1. First, Nb:STO is deposited using PLD (Fig. 8.1(a)-(b)). It has been shown that high-quality epitaxial films of Nb-doped SrTiO₃ can be deposited on STO (001) substrates. Due to the close lattice match, these films can be grown sufficiently
thick (up to 200 nm) while maintaining coherently strained and with a homogeneous distribution of dopant atoms\cite{54-56}.

After deposition, lithography and ion beam etching can be used to selectively remove parts of the Nb:STO film to define parallel columns that will serve as the bottom electrodes (Fig. 8.1(c)-(d)). A second lithography step defines an array of circles of resist that serve to define the cross-point contact areas (Fig. 8.1(e)). Al$_2$O$_3$ can then be deposited and after lift-off an insulation layer is present that will prevent electrical contact between the top and bottom electrode lines everywhere other than the intended cross point (Fig. 8.1(f)). This step should ensure leaving openings in the Al$_2$O$_3$ layer to contact the bottom electrodes. Subsequently, a third lithography step can be performed to define a set of parallel rows. This is followed by the deposition of for example Co/Au to define both the Schottky contacts at the circular gaps in the insulation layer, as well as the metallic electrode lines (Fig. 8.1(g)-(h)).

### 8.3.4 Other elements

![Figure 8.2: Biological neural network. Schematic depiction of different components that make up the network in the brain and their respective functions.](image)

Currently, the most widely considered computational building blocks for neuromorphic systems are synapses and neurons. The architecture of the brain, however, is substantially more complex than this and includes units such as axons, dendrites, and other neural structures, each of which contributes to the performance of the system and aids in receiving external inputs and processing and transmitting them and generating an output or a response in an energy-efficient way.

Neurons are active elements that produce electric signals (typically spikes) in re-
response to input signals. The neuron receives such input excitations through dendrites, which are the receiving connections. This is usually a non-linear process whereby a spike is only generated when the excitation voltage exceeds a threshold. This process tends to be time-dependent and the neuron is more likely to fire if the frequency of the excitation signals is high. The axon hillock of the neuron behaves like a leaky capacitor that can integrate incoming signals while slowly discharging over time. It will produce a spike if the integrated voltage exceeds a threshold. This spike is transmitted through the network through axons. These connections themselves can also implement information processing tasks. Synapses form the connection point between axons and dendrites from other neurons. The synaptic connections control the transmission strength of the outgoing signals to each dendrite, effectively how strongly two neurons are connected. These weights are plastic and can change over time allowing the system to learn [57–59].

In addition to the neural network itself, the human body is equipped with a large number of receptors that can interact with the outside world and translate this into signals that can be processed by the brain. For example, when a receptor in the skin is stimulated, a signal is transmitted through the spinal cord to the brain. The brain then interprets the signal and creates a perception of touch. This allows us to sense and respond to a wide range of tactile stimuli in our environment [60].

While our computing systems probably do not aim to be a one-to-one match, and rather the goal will be to selectively copy the biological system in a brain-inspired computing approach, interfacing this system with the outside world is an important consideration. Hence, attention should also be put on peripheral circuitry.

While sometimes the training can be done (partially) on-chip, in many cases the weights are calculated offline using an external computer at a significant energy cost. The weights are then programmed onto the chip at which point it can operate at low power. This can also mean that while the chip itself is small and energy efficient, the entire system including the external circuitry required to address the chip is bulky.

In this thesis, we also considered suitable models to describe device behaviour. This is an important step towards finding suitable learning algorithms that can eventually be used to train devices in integrated circuits on a chip. We found that when tested in a simulated neural network good performance on a machine learning task can be achieved. This is a promising result towards performance in hardware. It will be likely, however, that the simulated results will not match exactly with performance in hardware. Therefore, alongside developing the hardware new algorithms for application with memristive computing systems should be developed to match the performance. To overcome these challenges the continued efforts and expertise of researchers working in material science, computing science, mathematics, artificial intelligence and circuit design will be crucial.
Bibliography


8. Conclusions and Outlook


