Highly Energy Efficient Neuromorphic Computing Based on Memcapacitive Devices

Dissertation

zur Erlangung des
Doktorgrades der Ingenieurwissenschaften
(Dr.-Ing.)

der

Naturwissenschaftlichen Fakultät II
Chemie, Physik und Mathematik

der Martin-Luther-Universität
Halle-Wittenberg

vorgelegt von

Herrn Kai-Uwe Demasius
geb. am 15.05.1991 in Johannesburg, Südafrika

Gutachter:
Prof. Dr. Stuart S. P. Parkin
Prof. Dr. Christian Wenger
Prof. Dr. Ralf B. Wehrspohn

Tag der öffentlichen Verteidigung:
18.10.2021

Zusammenfassung

Der Datentransfer zwischen Speicher und Prozessor in digitalen von Neumann-Rechnerarchitekturen verbraucht viel Energie. Dies wird besonders kritisch für datenintensive Aufgaben, wie das Trainieren von neuronalen Netzen. Gegenwärtig werden Matrizen von resistiven nicht-volatilen Speicherbauelementen zur Implementierung von neuronalen Netzen untersucht. Diese ermöglichen hochparallele Multiplikationen und Summationen. Ebenso denkbar ist die Nutzung von (mem)-kapazitiven Bauelementen, welche den Vorteil eines niedrigeren statischen Energieverbrauchs haben, jedoch ist das niedrigere dynamische Hubverhältnis bei geringerer Skalierfähigkeit nachteilig. In dieser Arbeit wird ein CMOS-kompatibles Bauelement vorgeschlagen, welches Ladungsabschirmung ausnutzt, theoretisch simuliert und experimentell realisiert. Eine Skalierfähigkeit bis zu 45 nm-90 nm wird durch Simulationen bewiesen, wobei ein hohes dynamisches Hubverhältnis erhalten bleibt. Unter Ausnutzung einer adiabatischen Aufladung wird eine 30-300-fach bessere Energieeffizienz bei 6-8 Bit Präzision im Vergleich zu resistiven Technologien und potentiell höher als das menschliche Gehirn gezeigt. Weiterhin werden experimentelle Bauelemente und Matrizen auf der Mikrometer-Skala, sowie ein Bilderkennungsalgorithmus mit den Buchstaben “M”, “P” und “I” auf 156 synaptischen Bauelementen, demonstriert.

Abstract

Data transfer between memory and the processor in digital von Neumann architectures consumes a large amount of energy. This becomes extremely critical in modern data-intensive tasks, such as neural network training. Recently, neural networks were mapped onto arrays of resistive non volatile memory for highly parallel multiply-accumulate operations. (Mem)-capacitive devices can similarly be employed with the advantage of lower static power consumption, but they suffer from a poor dynamic range and scalability. In this thesis a CMOS-compatible (mem)-capacitive device based on charge screening is proposed, theoretically simulated and experimentally demonstrated. Scalability down to 45 nm-90 nm is shown by simulations, while retaining a large capacitance dynamic range. By using concepts of adiabatic charging it is shown that mapping neural network inference tasks with 6-8 Bit precision can be done with 30-300x better energy efficiency compared to common state-of-the-art resistive technologies and possibly greater efficiency than the human brain. Experimental devices and crossbars were fabricated at the micrometer scale and an image recognition algorithm with letters “M”, “P” and “I” is shown on 156 synaptic devices.

Contents

Abbreviations **III**

List of Figures **XIII**

List of Tables **XIV**

1 Introduction **1**

2 Fundamentals **3**

2.1 Introduction to Artificial Neural Networks 3
2.1.1 Biological Inspiration and General Working Principle 3
2.1.2 Different Topologies 4
2.1.3 Learning in Artificial Neural Networks 7
2.1.4 Challenges 10
2.2 Spiking Neural Networks 10
2.3 Physical Implementation 11
2.3.1 Common Resistive Devices 14
2.3.2 Capacitive Devices 17
2.4 Landauer Principle and Reversible Computing 19

3 Device Description and Theory **22**

3.1 General Working Principle 22
3.2 TCAD Simulation on Single Devices 24
3.3 Lateral Scalability and Read Signal Strength 28
3.4 Theoretical Limitation of Energy Efficiency 31
3.5 Spice Simulation and Estimation of Energy Efficiency 33
3.6 MNIST Simulation 39
3.7 Oxide Devices 40

4 Fabrication and Process Development **42**

4.1 First Work on SOI Wafer and Single Device Fabrication 42
4.2 Deposition Experiments of Hafnium Zirconium Oxide on Silicon 48
4.3 Process Development for Crossbar Fabrication 53
4.3.1 Aluminium Metal Layer Optimisation 54
4.3.2 Thin SU-8 Insulation Optimisation 55
4.3.3 Trench Refill Optimisation 62
4.3.4 Bond Pad Optimisation 71
4.3.5 Combination of All Processes 73
4.4 Growth of Oxide Devices 77

5 Measurement Setup and Printed Circuit Board 79
5.1 Measurement Setup for Single Devices 79
5.2 Printed Circuit Board for Crossbar Measurement 80

6 Measurement Results and Discussion 87
6.1 Capacitive Coupling Curves for Single Devices with no Memory Effect . 87
6.2 Comparison of Single Devices with Memory Effect and Different Interface Oxides 88
6.3 Single Device Measurements 90
6.4 Crossbar Measurements 93
6.5 Neuromorphic Algorithm on Crossbar 96

7 Conclusion and Outlook 99

References 100

A Appendix 116
A.1 Single Device Fabrication 116
A.2 Crossbar Fabrication 118

Acknowledgements 120

Abbreviations

AC	alternating current
ADC	analog digital converter
ALD	atomic layer deposition
ALU	arithmetic logic unit
ANN	artificial neuronal network
ASICs	application specific integrated circuits
BL	bitline
BEOL	back end of line
CMP	chemical mechanical polishing
C2F	capacitance-to-frequency
CBCM	charge-based-capacitive-measurements
CNN	convolutional neural network
CNNs	convolutional neural networks
CV	capacitance-voltage
CVD	chemical vapor deposition
DAC	digital analog converter
DAQ	data acquisition system
DC	direct current
DRAM	dynamic random access memory
EOT	equivalent oxide thickness
ESM	echo state machines
FeFET	ferroelectric field effect transistor
FeFETs	ferroelectric field effect transistors
FF	feed forward network

GPUs	graphical processing units
HDD	hard drive disk
HF	hydrofluoric acid
HZO	hafnium zirconium oxide
ICs	integrated circuits
IV	current-voltage
LRS	low resistance state
LSTM	long short-term memory
LTD	long-term depression
LTP	long-term potentiation
MAC	multiply-accumulate
MEMS	micro electro mechanical system
MFM	metal-ferroelectric-metal
MFOS	metal-ferroelectric-interfaceoxide-silicon
MNIST	Modified National Institute of Standards and Technology
MW	memory window
OPA	operational amplifier
PCB	printed circuit board
PECVD	plasma enhanced chemical vapor deposition
PLD	pulsed laser deposition
ReLU	rectified linear units
RHEED	reflection high-energy electron diffraction
RIE	reactive ion etching
RNN	recurrent neural network
RRAM	resistive random access memory

RTP	rapid thermal annealing
SEM	scanning electron microscopy
SL	shielding line
SMU	source measurement unit
SMUs	source measurement units
SNN	spiking neural networks
STDP	spike timing dependent plasticity
SOI	silicon on insulator
SONOS	silicon-oxide-nitride-oxide-silion
SRAM	static random access memory
TCAD	technology computer-aided Design
TEOS	tetraethylorthosilicate
TFT	thin film transistor
TLM	transmission line method
TPUs	tensor processing unit
WL	wordline
XRD	x-ray diffraction

List of Figures

2.1 a) Biological neural network with neuron soma, dendrites as inputs and axon as output. The synapses are arranged between axon branches and dendrites. b) Artificial neuronal network with neurons as circles and synaptic weights as arrows. The neurons sum the input values and a non-linear activation is applied. This can e.g. be a hyperbolic tangent, sigmoid or rectified linear units (ReLU) function. 4
2.2 Family tree of artificial neuronal network (ANN) with the two main classes: Feed forward and recurrent neural networks. 6
2.3 a) Supervised learning with an error function, which needs to be minimized, by using an desired output and obtained output. The network is trained with a labeled dataset (desired output). b) Unsupervised learning with unlabeled data on a self-organising map. In this case colors are sorted according to their similarities. c) Reinforcement learning with an agent, which receives rewards for certain actions on the environment. . 8
2.4 a) Circuit model of a leaky-integrate and fire neuron. The corresponding waveform of a spike is shown on the right. b) Spike-timing dependent plasticity of a pre-neuron and post-neuron spike. 11
2.5 Illustration of von Neumann bottleneck with distinct memory and processor (ALU, Control Unit). The different memory hierarchies are shown. 12
2.6 a) Fully connected neural network with the pre-neurons N_1 to N_n and post-neurons M_1 to M_n b) Matrix arrangement with conventional digital CMOS to map the neural network. c) Crossbar arrangement of resistive devices. d) Crossbar arrangement with selector in differential topology and the different input coding schemes. Similar illustrations can be found in [1, 2]. 14
2.7 a) resistive random access memory (RRAM) device with conducting filament and corresponding pinched hysteresis loop of current-voltage (IV) curve. b) Phase-change memory cell with asymmetric conductance response versus write pulse number. c) Spintronic memristor based on domain wall motion. d) ferroelectric field effect transistor (FeFET) and ferroelectric tunnel juntion based synapse. Shown are domain nucleation events and single grain switching. Similar illustrations can be found in [3, 4]. 16

2.8 Variable capacitance with changing plate distance based on: a) micro electro mechanical system (MEMS) device (a similar illustration can be found in [5]), b) metal-to-insulator transition (MIT) material in series to normal dielectric layer, c) filament formation in a memristor, d) MOS capacitance with variable depletion layer. 17

2.9 'Varying surface memcapacitor' based on memristor in series to a large area capacitor and 'Varying dielectric constant memcapacitor' based on metal-ferroelectric-metal (MFM) capacitor. Similar illustrations can be found in [6, 7]. 18

2.10 a) Irreversibility of the AND gate: It is not possible to conclude from the output "0", which input was applied. b) Reversibility of the Toffoli gate. c) Physical implementation of a reversible computer: The billiard ball computer. Ultra low friction and elastic collisions are assumed. d) Non-adiabatic charging versus adiabatic charging for a capacitor. The slow charging of the capacitor gives an advantage compared to the abrupt charging. 20

3.1 a) General Device structure with gate electrode, memory dielectric, shielding layer, passive dielectric and read-out electrode. The capacitive coupling is indicated by the blue arrow. b) Device structure with lateral p+n-n+ junction in the shielding layer. Material from [8] 23

3.2 a) Crossbar arrangement of devices for highly parallel multiply-accumulate (MAC) operations. The inputs signals are applied to the wordlines and weighted by the shielding layer and memory material. The accumulated result is read-out at the bitlines. Material from [8] 24

3.3 a) Simulated structure and dimensions. Silicon oxide (no memory effect) was used for the dielectric layers. b) capacitance-voltage (CV) curves for the capacitive coupling between gate electrode and read-out electrode. A voltage V_{AK} was applied to the p+n-n+ junction. c) Half-width broadness of the capacitive window in b) versus applied V_{AK}. d) Accumulated charge on read-out electrode versus applied gate voltage. Material from [8] 26

3.4 a) Capacitive swing for different gate length and equivalent gate oxide thicknesses (EOT). The inset shows the accumulated electron concentration and the charge rejection due to the short channel effect. b) CV curve shifting due to memory charges and c) The corresponding accumulated charge on the read-out electrode over a half period of the sinusoidal signal in b. d) Total accumulated charge over a half period for different CV curve shift voltages V_{shift}. Material from [8] 27

3.5 charge-based-capacitive-measurements (CBCM) detection circuit with measurement capacitor C_S and reference capacitor C_R (a similar illustration can be found in [9]). 29

3.6 a) Simulated structure of the 45 nm device with HfO_2 as a high-k dielectric and corresponding CV curve in b). Material from [8] 30

3.7 a) Rect pulse for the measurement current integration b) Fourier-transform of the rect pulse in a) with the indicated band limitation. 32

3.8 a) Spice circuit model of single device with parasitic capacitors and resistors. b) Gate oxide capacitance and coupling capacitance of spice simulation in comparison to Synopsys simulation. c) Coupling capacitance between gate and read-out electrode of spice simulation in comparison to Synopsys simulation. Material from [8] 34

3.9 a) Arrangement of spice model in an array along the bitline (BL)/shielding line (SL) with further parasitic elements. b) Accumulated charge on the BL over a half-period of 15 ns for different array sizes. A written state is programmed into the memory cells in this case. c) Accumulated charge on the BL over a half-period of 15 ns for different array sizes. An erased state is programmed into the memory cells in this case. d) Comparison of written and erased memory cells with regards to the total accumulated charge on the BL. Material from [8] 35

3.10 a) Active energy per memory cell W_a/N_{cells} and b) reactive energy per memory cell W_q/N_{cells} for different array length and T_{per}. c) Energy efficiency without recovery and d) energy efficiency with energy recovery, assuming $q = 20$. (The legend from a) is valid for all plots) Material from [8] 37

3.11 a) Map of implemented weights and b) applied images of handwritten numbers. Material from [8] 39

3.12 Obtained output activation compared to the calculated activation for a) the number 0, b) the number 2 and c) the number 4. Material from [8] 40

3.13 a) Depletion of n-doped semiconductor during application of a negative gate voltage. b) Fully depleted semiconductor layer, thus no further switching field increase. c) Use of positive back-gate voltage from read-out electrode during writing. 41

4.1 a) IV curves between device and handle wafer for different hydrofluoric acid (HF) etching times b) Maximum leakage current through buried oxide layer versus HF etching time c) Illustration of the under-etching with HF of the buried oxide layer. 43

4.2 a) Stepwise processing of silicon on insulator (SOI) Wafer piece. b) Holes after p-doping. c) Island yield after p-doping (first piece). d) Rests on p-doped regions after n-doping and underetching of holes. e) Island yield after p- and n- doping. 45

4.3 a) Boron concentration with 10 keV (equivalent to 45 keV BF2) and two different Doses [10]. b) Phosphorous concentration with 25 keV and two different doses [10]. 46

4.4 a) Island yield after pin diode fabrication. b) Measured CV coupling with lock-amplification. 46

4.5 Fabrication flow for single devices with the two branches a and b for the thermal oxide and the memory dielectric with HZO/TiN. In the bottom left the mask layout of a device with gate contact winded around n+ and p+ regions is shown. 47

4.6 a) P-Vg Hysteresis loop of a MFM capacitor and b) corresponding CV curve. 48

4.7 a) Ellipsometrically measured interface oxide thicknesses for a 5:1:1 SC1 solution at different temperatures. b) Ellipsometrically measured interface oxide thicknesses for a 40:2:1 SC1 solution at 36 °C. c) Interface oxide thickness for different diluted SC1 solutions at room temperature and 30 s. d) Interface oxide thickness for 10000:1:1 SC1 solution at room temperature. 49

4.8 a) CV sweeps for 5:1:1 SC1 solution for different times and b) corresponding fraction of ferroelectric MFOS capacitors. c) CV sweeps for 40:2:1 SC1 solution and d) corresponding ferroelectric fraction. e) CV sweeps for 10000:1:1 SC1 solution and f) corresponding ferroelectric fraction. 51

4.9 CV sweep of a ferroelectric memory window and the charge trapping memory window after a second CV sweep. 52

4.10 a) CV sweeps for different interface oxidation temperatures and b) corresponding fraction of charge trapping MFOS capacitors. 52

4.11 a) XRD of the hafnium zirconium oxide (HZO) films, indicating the orthorhombic phase. b) TEM images of the Si / SiO_2 / $Hf_{0.5}Zr_{0.5}O_2$ / TiN / Alu stack for different interface oxide growth conditions. 53

4.12 a) Cross-section of the crossbar arrangement with trenches and wordline metallisation and b) top view of crossbar arrangement with bitline connection, trenches and wordline. 54

4.13 a) Microscope images of aluminium holes for different temperatures of 230 nm thick aluminium. b) Microscope images at 180 °C for (110 nm) and 160 nm thick films. 55

4.14 Crosssection of the trench and hill structure and corresponding top view mask layout. 56
4.15 SU8-6001: a) IV curves of connected metal lines on hills. b) IV curves of connected metal lines on trenches. c) IV curves of neighbouring metal lines on hills. d) IV curves of neighbouring metal lines on trenches. . . 58
4.16 SU8-6000.5: a) IV curves of connected metal lines on hills. b) IV curves of connected metal lines on trenches. c) IV curves of neighbouring metal lines on hills. d) IV curves of neighbouring metal lines on trenches. . . 59
4.17 SU8-6001 with 130C hard-bake: a) IV curves of connected metal lines on hills. b) IV curves of connected metal lines on trenches. 59
4.18 Mask design of first leakage test with fully covered backside electrode. Microscope images of the alumnium dots are provided on the right-hand side. 60
4.19 IV curve of leakage current for a) SU-8 6000.5 b) SU-8 6001 c) SU-8 6005. 61
4.20 Mask design of second crossbar leakage test. Microscope images of the crossing points are provided on the right-hand side. 61
4.21 IV curves of leakage for a) SU-8 6000.5 with side pad contact b) SU-8 6000.5 with contact on crossing point c) SU-8 6001 with side pad contact d) SU-8 6001 with contact on crossing point d) SU-8 6005 with side pad contact and e) SU-8 6005 with contact on crossing point. 62
4.22 Mask layout of device structure with islands surrounded by trenched surface. 63
4.23 Trench refill with structured SU-8 6005 (8000 rpm) for a) 95 µm island + 95 µm trench width, b) 140 µm island + 150 µm trench width, c) 95 µm island + 95 µm trench width with 180 °C anneal and d) 140 µm island + 150 µm trench width with 180 °C anneal. 64
4.24 Trench refill with structured thicker SU-8 6005 (3000 rpm) for a) 95 µm island + 95 µm trench width and b) 140 µm island + 150 µm trench width. 64
4.25 Three coatings with SU-8 6001 for a) 95 µm island + 95 µm trench width and b) 140 µm island + 150 µm trench width. 65
4.26 reactive ion etching (RIE) etching of SU-8 6005 resist a) Profile of RIE etched SU-8 6005 resist and b) corresponding microscope images. . . . 66
4.27 RIE etching of SU-8 6005 resist with CF_4 plasma: a) Profile of an SU-8 step for different etching times, b) corresponding roughness after 14 min etching and c) profile of the real trenched structure for different etching times. 66

4.28 a) Mask layout of the new trench etch. b) IV connectivity for different trench widths along the n+ doped island. c) IV connectivity between the silicon islands for different trench widths. 68
4.29 Before and after spin coating for a) 10 μm trench width and 100 μm island width, b) 20 μm trench width and 100 μm island width, c) 50 μm trench width and 100 μm island width and d) 100 μm trench width and 150 μm island width. 69
4.30 Surface height difference versus trench coverage. 70
4.31 Profile for 20 μm trench width at a) wafer center and b) wafer top. . . . 70
4.32 Profile for 50 μm trench width at a) wafer center and b) wafer top. . . . 71
4.33 a) Crosssection of the CV parasitic coupling test between two pads. The n+ back contact is either connected to ground or kept floating during the measurement. b) Corresponding CV plots for grounded or floated back contact. 72
4.34 IV curve between the aluminium Bond pad and the n+ back contact: a) Aluminium bond pad deposited after O_2 etching, before wire bonding. b) Aluminium Bond pad deposited after O_2 etching, after wire bonding. c) Aluminium bond pad deposited before O_2 etching, before wire bonding. d) Aluminium bond pad deposited before O_2 etching, after wire bonding. 72
4.35 Mask design of the test wafer with a zoom in for one structure. 73
4.36 a) Connectivity along the wordlines. The current saturation is from the instrument. b) Insulation between wordline and n+ and p+ contact. c) Insulation between different wordlines. d) Insulation between device layer and bitline. 74
4.37 Process flow for crossbar devices. 76
4.38 Wirebonded crossbar chip with microscope image of the memory matrix and further zoomed-in scanning electron microscopy (SEM) images. Material from [8] 77
4.39 a) Device stack and material combination and the growth problem. b) reflection high-energy electron diffraction (RHEED) image of a grown $SrTiO_3$(ins)-$SrTiO_3$:Nb stack. 78
5.1 a) Measurement setup for single device measurements. Applied write pulse and read-out sine wave for successive programming with: b) pulse number modulation (read-out sine wave in the inset) c) pulse height modulation and d) pulse length modulation. Material from [8] 80
5.2 a) Test-printed circuit board (PCB) for selecting single Wordlines, Shielding lines and Bitlines. The PCB is mainly used for CV measurements. . 80

5.3 a) Test circuit for measuring two capacitances (C1 and C2), which are subtracted from each other. b) Corresponding picture of the PCB with the different circuit blocks as framed in a). 81
5.4 Switched capacitor approach for read-out. A negative input signal is applied to WL1, while a positive one is applied to WL2. The positive and negative measurement capacitance (C- and C+) is either connected to ground or with the amplifier by switches, which are controlled by the clock signal. The state of the clock signal is indicated by the purple point (low for the right side and high for the left side). The arrows indicate the current flow direction. 83
5.5 Output voltage V_o versus time for a) positive measurement capacitor (4.7 pF) + positive input signal b) positive measurement capacitor (4.7 pF) + negative input signal c) negative measurement capacitor (4.7 pF) + positive input signal d) negative measurement capacitor (4.7 pF) + negative input signal. The blue line is the charge injection background (init), the green line the result with capacitor, as obtained and the red line the subtracted charge injection from the measurement (green line minus blue line). The period and half period are indicated by the dashed/dotted and dotted line and furthermore the resulting voltage V_r is indicated by the star. 84
5.6 a) Total resulting voltage for different measurement capacitors and input period numbers. b) the total charge is calculated from a) and the integration capacitance. c) Slope of lines in b) versus measurement capacitors, compared to the theoretical value of charge integration. 85
5.7 a) Circuit for the crossbar measurement with 26 wordlines. b) PCB for crossbar measurement. The crossbar chip is inserted into the black socket. 86
6.1 CV curves (a) for different depletion and forward voltages on the pin-junction (V_{AK}). The applied alternating current (AC) voltage levels to obtain the sigmoid curve in b) are indicated. Material from [8] 88
6.2 a) CV curves for different interface oxide. b) Pulse number modulation of multi-level switching for different interface oxides. c) Retention of different interface oxides. Material from [8] 89
6.3 a) CV curves for 10 devices. The applied sine wave is used for displacement current measurement in c. The current is proportional to the shaded area. b) Pulse number modulation for 10 devices. c) Read-out current for the written state. d) Read-out current for the erased state. Material from [8] 91

6.4 a) Pulse number modulation for a single device. b) Pulse number modulation for different pulse heights. c) Pulse length modulation and d) Pulse height modulation. Material from [8] 92
6.5 Measured read-out current for the pulse height modulation for different pulse numbers, during the a) long-term potentiation (LTP) and b) long-term depression (LTD). Material from [8] 93
6.6 a) Retention of different stored capacitance values (here read-out current value) b) Endurance characteristics of the CV curves. 93
6.7 a) CV curve for wire-bonded or probe station measured single memory cell. b) Pulse number modulation for wire-bonded crossbar chip along one BL with all wordline (WL) being activated. c) Disturb of an erased (red line) and written (cyan and blue line) memory cell with a small voltage. 95
6.8 a) 3D plot of 'four-quadrant multiplication' with varying the input period number (N_{per}) and the programming pulse number (N_{pgr}). The output voltage is the z-axis. Cross-sections in parallel to the N_{pgr} and N_{per} axis are shown in b) and c). d) The number of positive wordlines is varied, while all non-positive WL are negative, thus showing the linear accumulation operation. Material from [8] 96
6.9 a) Pulse scheme for implementation of XNOR operation of the Manhatten update rule. b) Used training and test image set of the letters M,P and I. The test images are framed purple. Material from [8] 97
6.10 a) Number of mis-classified images for the training and test set versus training epoch. b)-d) Mean neuron activations for the three classifications over the training epochs. Material from [8] 98
A.1 Mask layers for different processes of single devices fabrication. The steps are corresponding to Fig. 4.5. 117
A.2 Mask layers for different processes of crossbar devices fabrication. The steps are corresponding to Fig. 4.37. 119

List of Tables

1 Properties of the brain and a typical desktop computer in comparison. The brain is approximately 10,000 times more energy efficient and can store more weights due to its three-dimensional structure. 12

2 Desired properties of the resistive memory device for inference, training and both. [11,12] 14

3 Summary of time delay, areal efficiency and active energy/reactive energy per cell. Note that these values include $N_{per} = 142$. From the active and reactive energy the energy efficiency (η_{rec}; η) is deduced. **all cells are erased (worst case scenario), 95% energy efficiency of power clock source ($q = 20$)* 39

4 Processing conditions for SU-8 TF 6000.5, SU-8 TF 6001 and SU-8 TF 6005. 56

5 Lift-off results on trenches and hills with different SU-8 resist thicknesses. 57

6 Results of height difference and remaining trench depth after spin coating for 10 μm, 20 μm and 50 μm trench width. 70

1 Introduction

There is considerable interest in brain-inspired computing, particularly by using artificial neural networks and their hardware implementation, a field of research often termed neuromorphic computing. Artificial neural networks, a subform of artificial intelligence, have made significant progress in recent years, fueled by the internet, which enabled the big data revolution, as well as increased hardware performance, especially with graphical processing units (GPUs).

The first computational model of a brain-inspired neural network model can be traced back to Warren McCulloch and Walter Pitts in 1943 [13]. The McCulloch-Pitts neuron is a simple threshold switch which can either output a one or zero. In the late 1940s Donald Hebb suggested the Hebbian learning rule for artificial neural networks [14], which forms the basis for spike timing dependent plasticity (STDP) (see chapter 2.2) [15]. The backpropagation algorithm was invented by Paul Werbos in 1975 [16] and is the most important learning algorithm for supervised learning. Convolutional neural networks, which are inspired by the receptive field and visual cortex of the human brain, were first used in 1989 by Yann LeCun [17]. Since the early 2010s convolutional neural networks have gained significant momentum and won several image recognition competitions [18, 19].

Current computing hardware is not well-suited for calculating neural network, due to the von Neumann architecture, which implies a separated memory and processor. Thus significant data transfer between memory and processor is necessary, which leads to a 1,000-10,000 times larger energy consumption compared to the human brain [20]. Early works on neuromorphic implementations can be traced back to Carver Mead. In his famous paper from 1990 [21] he first proposed the replacement of digital information processing with analog processing for neural networks. Use of more suitable computational primitives can lead to significantly reduced circuit overhead and power consumption. Since the intensive investigation of the memristor effect in 2008 [22] arrays of resistive devices were implemented and used for highly parallel neural network calculation.

Most recent work has been based on resistive devices and systems. There have been some theoretical proposals for memcapacitive devices [23–31], but few practical implementations [6,7,32,33]. The aim of this thesis is to theoretically analyse and practically implement a new memcapacitive device for neuromorphic computing. The roots of the device can be traced back to a sensor for electrostatic fields, which was already the topic of the authors diploma thesis [34].

The thesis is divided into the following sections: First in chapter 2 some fundamentals on ANN are introduced, including different network topologies and learning algorithms. Then the physical implementation (chapter 2.3) in the form of a crossbar arrangement and different resistive and capacitive technologies are introduced.

Thereafter a theory section (chapter 3) describes the device working principle and provides technology computer-aided Design (TCAD) and Spice simulation, in which the energy consumption and scalability is estimated. The next two sections describe the fabrication (chapter 4) of single capacitive devices and crossbars, as well as the measurement setup (chapter 5). The second last chapter (chapter 6) describes the measurement results and the image recognition algorithm implemented on a memcapacitive crossbar array. The thesis is summarized by a conclusion (chapter 7).

2 Fundamentals

In this chapter a broad overview on neuronal networks is given, as well as the hardware implementation (neuromorphic computing) of neural networks.

2.1 Introduction to Artificial Neural Networks

2.1.1 Biological Inspiration and General Working Principle

Fig. 2.1 shows the structure of a biological neuronal network in comparison to an artificial version. The biological neurons comprise of a soma, a single axon and dendrites [35–38]. Electrical signals from a pre-neuron are fed through a synaptic gap into the dendrites (the inputs) of neurons. In the soma and the axon, electrical pulses are generated by a membrane containing potassium (K^+) and sodium (Na^+) ion channels. So the calculation is done in the neurons, while the synaptic gaps are responsible for the learning. Details on the information processing can be found in chapter 2.2. The output spike voltages generated by the neurons are fed along the axon to several other synaptic terminals and thus other neurons. The synapses connect the axon with a dendrite of a neuron and has a gap of 20 nm as shown in Fig. 2.1a). One needs to distinguish between chemical synapses and electrical synapses [36]. While chemical synapses use neuro-transmitter molecules that are attached to receptors of the dentrites for communication, the electrical synapses use a direct electrical connection for communication and are much faster, but less versatile. The learning in the brain is caused by synaptic plasticity, which means connectivity can increase or decrease depending on the activity. The principle behind the plasticity is caused by several mechanisms, but generally the variation of the number of receptors for neuro-transmitters is one of the main mechanism.

On average, every neuron is connected to 10000 other neurons. An adult human brain contains approximately 14-16 billion neurons and 100-500 trillion ($1E + 14 - 5E + 14$) synapses, while it consumes only 20 W.

In contrast, an ANN [1, 38–40] (Fig. 2.1b) is clearly biologically inspired regarding its structure, but the detailed underlying information processing can be quite different. The circles in Fig. 2.1b) represent the neurons and contrary to biological networks, possess a non-linear activation function (usually a sigmoid/tanh or ReLU function). The inputs of a neuron are summed up and applied to the non-linear activation. The synapses are represented by the arrows in Fig. 2.1b) and are also called weights, which means the output of a neuron is multiplied by the weight and fed to the next neuron. The ANN forms a layered structure, where an ANN is called deep if it has several hidden layers. The structure in 2.1b) is a simple feed forward, fully connected neural network, which means that the data flow is only from one layer to the next and every pre-neuron is connected to every post-neuron. Other network topologies are explained in the next chapter 2.1.2.

The aim of a neural network is to train the weights in such a way that they represent certain patterns. The neural network can classify an input pattern (image, sound or other data) into output classes after training. Using the neural network for classification tasks after training is called inference. Different training and learning methods are explained in chapter 2.1.3.

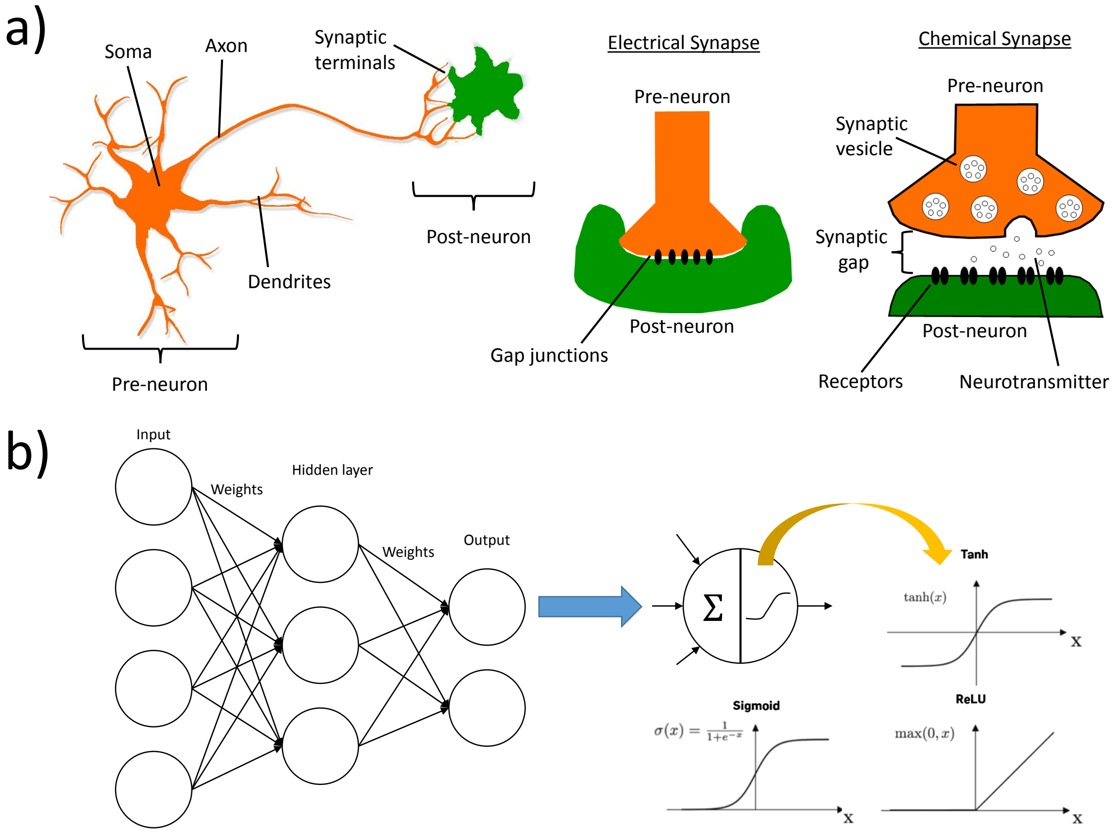


Figure 2.1: a) Biological neural network with neuron soma, dendrites as inputs and axon as output. The synapses are arranged between axon branches and dendrites. b) Artificial neuronal network with neurons as circles and synaptic weights as arrows. The neurons sum the input values and a non-linear activation is applied. This can e.g. be a hyperbolic tangent, sigmoid or ReLU function.

2.1.2 Different Topologies

Generally, neural networks can be classified into two different main classes: feed forward network (FF) and recurrent neural network (RNN) (Fig. 2.2) [41]. The two main classes can be further divided into different types. Feed forward means the information flow is only in one direction (except during back propagation training), while recurrent neural networks have feedback loops and thus a temporal weight sharing. In recurrent neural networks the timing of certain events in the past have an effect on how outputs are generated for the current state.

Within the class of feed forward networks the fully connected neural networks were already introduced in chapter 2.1.1: Every pre-neuron is connected to every post-neuron of the layers. These kind of neural networks are the simplest and oldest version, and an experimental implementation of a fully connected layer is explained in chapter 6.5. convolutional neural networks (CNNs) [17–19, 42] play an important part in image recognition, and a small filter matrix is convolved over an input image. The filter matrix contains the filter weights and is multiplied with the pixels of the input feature map and summed to one pixel of the output feature map. Due to the convolution, the weights are reused over the full image and thus a weight sharing is implemented. The aim of the convolution with a filter matrix is to extract certain features, like edges or circles. Each convolutional layer is usually followed by a pooling layer which reduces the dimension of the output feature map by averaging or taking the maximum values of a small area of the output feature map. Afterwards a ReLU activation is mostly used. The last layers of a convolutional neural network are fully connected in order to accomplish the desired classification. Convolutional neural networks are highly biologically-inspired by the visual cortex and the filter matrix has similarities to the receptive field of the retina. With modern convolutional neural networks super-human accuracy has already been achieved (e.g. ResNet-50 [19]).

Radial basis function neural networks [43,44] use a radial basis function as an activation function for the neurons. They are used for function approximation. The autoencoder [45] will be explained in the next chapter 2.1.3 in the context of unsupervised learning. They have the same number of outputs as inputs and they are trained in such a way that the output is a close representation of its original input.

As already mentioned, recurrent neural networks [41] show a much more complicated dynamic behaviour. The oldest subtype of RNN are Hopfield networks [46], which are a kind of associative memory. Every neuron is connected to the input of its neighbouring neurons, but there is no feedback to its own input. Trained Hopfield networks have a minimum energy for a certain input and thus, when a new input is applied, which is similar to the trained input, the neurons return back to the energetic optimum. (e.g. input vector [-1 1 1 -1] would return to the trained input [-1 1 1 1]). One can define a Hamiltonian for Hopfield networks to obtain the minimum energy, similar to physics. Boltzmann machines [47] are similar to Hopfield networks, but are stochastic.

In conventional recurrent neural networks the state is only stored from one moment to the next moment. There are many practical cases where an output far in the past should have a connection to the current input. This problem is solved by long short-term memory (LSTM) [48,49], a network class, which contains a forget gate, input gate and output gate, where each gate controls to what extent a cell status is considered for the current point of time or forget. LSTM have lead to wide-spread application in natural language processing [49].

Due to the fact that recurrent neural networks are difficult to train echo state machines (ESM) and reservoir computing have been developed. In this a random recurrent neural network, the reservoir ensures a complicated dynamic behaviour, while only the output layer is trained. There are widespread physical systems that can act as a reservoir [50].

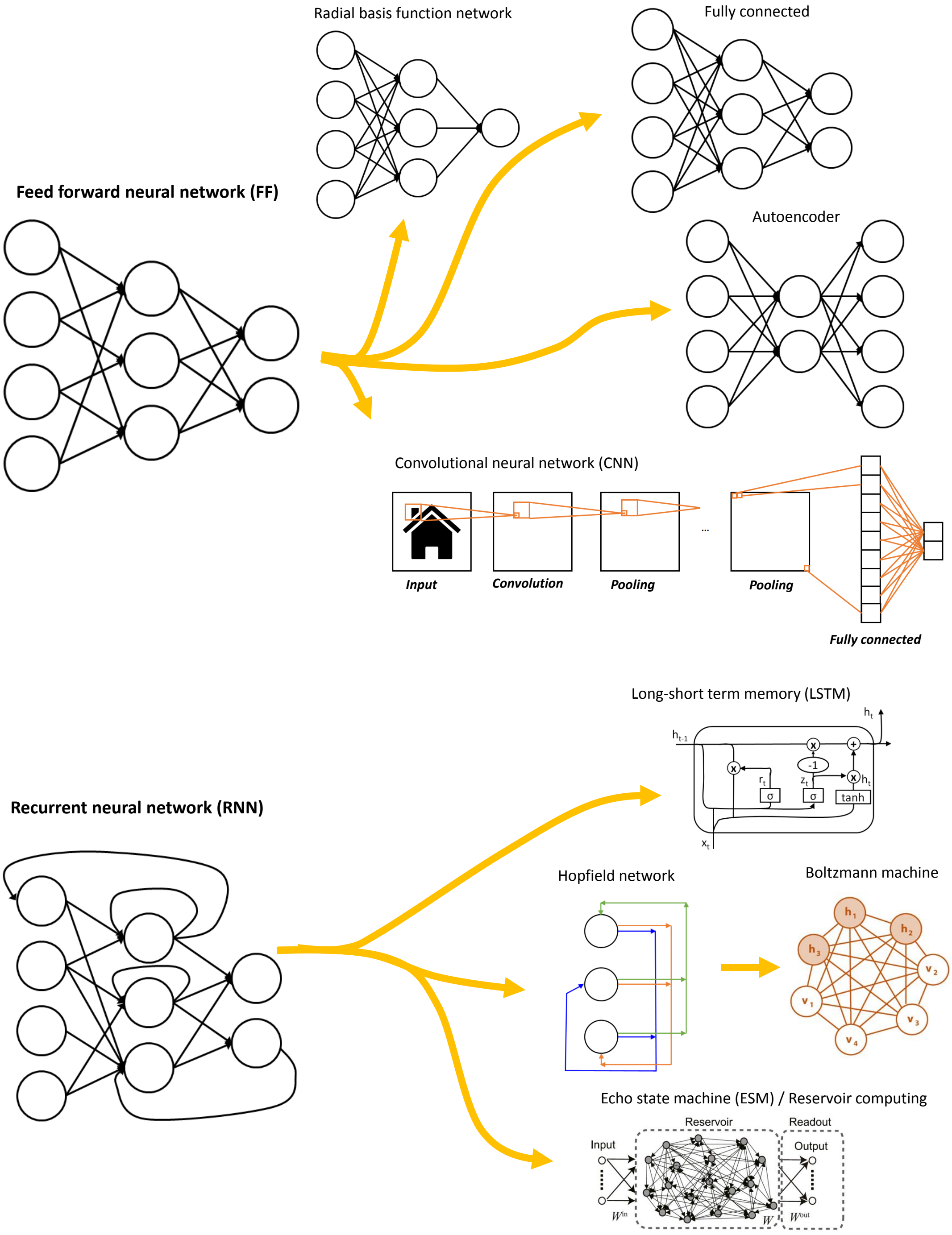


Figure 2.2: Family tree of ANN with the two main classes: Feed forward and recurrent neural networks.

2.1.3 Learning in Artificial Neural Networks

Generally, one can distinguish between supervised learning, unsupervised learning and reinforcement learning, which is a mixture of supervised and unsupervised learning.
Supervised learning is accomplished with labeled datasets, which means the weights are initalized arbitarly in the beginning, an input (e.g. image) is applied, and an error between current and desired output classification is calculated. This error is used to train the network in such a way that a minimized error is achieved on the training set of inputs. Afterwards the network is able to sort most inputs of a test set into the correct classifications. A detailed description of the backpropagation algorithm [16] is given below.
Unsupervised learning is achieved on an unlabeled dataset. The network is able to sort the inputs, due to differences in the input, into different output classifications. An example of an unsupervised algorithm is a self-organising map [39, 51], which can sort different colors into a map of similar colors (Fig. 2.3b). Self organising maps strengthen weights to neighbourhood neurons that have a similar color. Another example, the autoencoder [45], was mentioned in the previous chapter: The input layer is connected via one hidden layer to the output layer, where the input features are mapped to the same output features. The hidden layer is a lower dimensional representation of the inputs.
Reinforcement learning [52] (Fig. 2.3c) has an agent, which is usally a neural network, and the agent is taking action to an environment. The environment returns the actual status and a reward back to the agent. The aim of reinforcement learning is to maximize the cumulative reward with every action of the agent. Thus the agent learns a certain policy during the exploration of the environment.

Backpropagation training [16, 39] for supervised learning remains one of the most popular methods to train neural networks and was suggested in 1975. The general aim is to change the weights in such a way that the cost function will decrease, which can be achieved by gradient descent: The derivative of the cost function is calculated with respect to each weight and the weights and biases of each neuron are changed with the opposite sign of the derivative. One possible cost functions is the quadratic cost function:

$$K = \frac{1}{2} \cdot \sum_j \left(y_j - a_j^L \right)^2 \tag{2.1.1}$$

with K being the cost, y_i is the desired output of output neuron j and a_j is the activation of neuron j in the output layer L. The upper index describes the layer number in the neural network, while j is the index of the post-neuron and i the index of the pre-neuron.

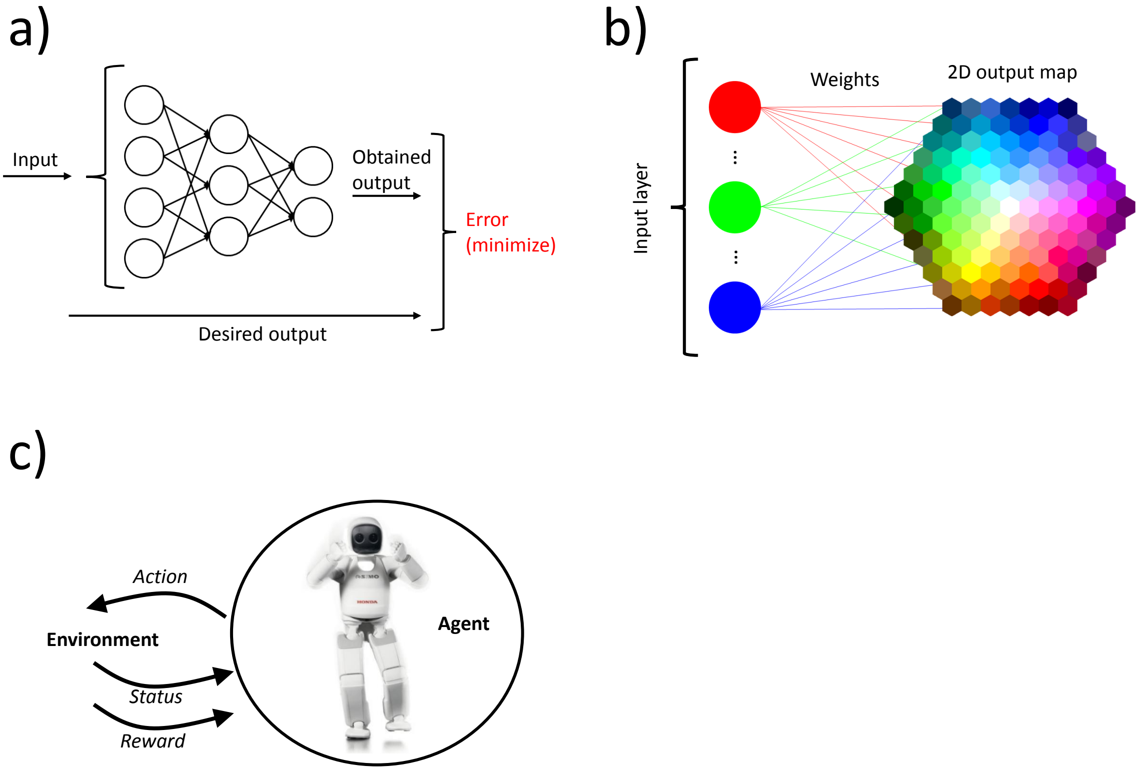


Figure 2.3: a) Supervised learning with an error function, which needs to be minimized, by using an desired output and obtained output. The network is trained with a labeled dataset (desired output). b) Unsupervised learning with unlabeled data on a self-organising map. In this case colors are sorted according to their similarities. c) Reinforcement learning with an agent, which receives rewards for certain actions on the environment.

Backpropagation training can be applied in a stochastic mode (equation 2.1.1), were the cost function is calculated for every input of the training set or in a batch mode where the cost function is averaged for all inputs of the training set:

$$K = \frac{1}{2n} \cdot \sum_{x} \sum_{j} \left(y_j - a_j^L \right)^2 \tag{2.1.2}$$

with n being the number of samples in the training set and x the individual inputs. The weighted input z^l of a neuron in the layer l is:

$$z^l = w^l a^{l-1} + b^l \tag{2.1.3}$$

so it depends on the weights w^l, the activation of the pre-neurons a^{l-1} and the bias b^l. The error at the j^{th} neuron is defined as the derivative of the cost function from the weighted input of the neuron z_j^l:

$$\delta_j^l = \frac{\partial K}{\partial z_j^l} \tag{2.1.4}$$

For the last layer L neurons the error is simply:

$$\delta^L = \nabla_a K \circ \sigma'(z^L) \tag{2.1.5}$$

With $\circ$ describing the Hadamard product and $\sigma'(z^L)$ is the derivative of the non-linear activation function (e.g. sigmoid) of the output neurons at the weighted input.
The error of the other layers is calculated by the backpropagated error, which can be proven by the chain rule:

$$\delta^l = \left(\left(w^{l+1} \right)^T \delta^{l+1} \right) \circ \sigma'(z^l) \tag{2.1.6}$$

The error is again weighted by w^{l+1}. From the errors one can make conclusions on the dependence of the cost function on the weights between the i^{th} pre-neuron and the j^{th} post-neuron, as well as on the bias of the j^{th} neuron:

$$\frac{\partial K}{\partial b_j^l} = \delta_j^l \tag{2.1.7}$$

$$\frac{\partial K}{\partial w_{ji}^l} = a_i^{l-1} \cdot \delta_j^l \tag{2.1.8}$$

These formulas can be derived from the equations 2.1.4 and 2.1.3. Thus for the weight and bias update during training one can conclude:

$$\Delta W_{ij}^l = -\alpha \cdot \delta_j^l \cdot a_i^{l-1} \tag{2.1.9}$$

$$\Delta b_j = -\alpha \cdot \delta_j^l \tag{2.1.10}$$

Where α is the learning rate and during the weight update also the output activation of the pre-neuron needs to be considered. In a special type of backpropagation algorithm, the Manhattan update rule, the weights are updated always by the same amount and only the sign of equation 2.1.9 is used:

$$\Delta W_{ij}^l = \operatorname{sgn}\left(-\alpha \cdot \delta_j^l \cdot a_i^{l-1} \right) \tag{2.1.11}$$

This algorithm is implemented in chapter 6.5.

2.1.4 Challenges

One of the big issues with ANN is the huge amount of data required to train the networks. For this reason the models are usually implemented in the cloud. Furthermore, calculation of neural networks requires a vast amount of energy. This problem is addressed in chapter 2.3. Also the networks are quite often only good for specific tasks and far away from the flexibility of the human brain.

Improvements in unsupervised learning algorithms might solve the bottleneck of available training data.

2.2 Spiking Neural Networks

Spiking neural networks [15, 53–55] are the more biologically plausible type of neural networks and often termed third generation neural networks. The implementation of neurons and the learning algorithms are different compared to perceptron like ANN. The neurons integrate the input signals, and once a certain threshold of membrane potential is achieved, the neurons send out a spike. The integration is usually implemented with capacitors, while the neurons may also have a leaky path, which is implemented with a resistor. Neurons of this type are called leaky-integrate and fire neurons. Neurons can also have even more biologically plausible circuits, like implemented in the Hodgkin-Huxley model. The learning in spiking neural networks (SNN) is most often a STDP algorithm [54], which means in short, that neurons that fire together, wire together (Hebbian learning rule). A typical STDP curve is shown in Fig. 2.4b): If the post neuron fires shortly after the pre-neuron, the weight between them is increased, while for a firing in the post-neuron shortly before the pre-neuron the weight is depressed strongly. If the time distance between the post-neuron and pre-neuron spike is very long, the spikes are uncorrelated and thus the weight is nearly unaffected.

Supervised learning is more complicated to implement in SNN, since backpropagation cannot be applied (the neuron activation function needs to be differentiable). There are some other supervised learning algorithms specific for spiking neural networks (e.g. NormAD [56, 57]).

Generally the STDP is a very localized learning algorithm, thus the accuracy of SNN has not reached the same as ANN and is practically barely used. The learning algorithm of the biological brain is not fully understood as yet to implement efficient SNN. The advantage of SNN is the lower energy consumption compared to ANN, since SNN are event-driven networks.

Spiking neural networks are not the focus of this thesis, although there are some physical realisations of some ANN that use a pulse number [58] as an input signal coding to neurons (see next chapter 2.3). These realisations are not classified as SNN.

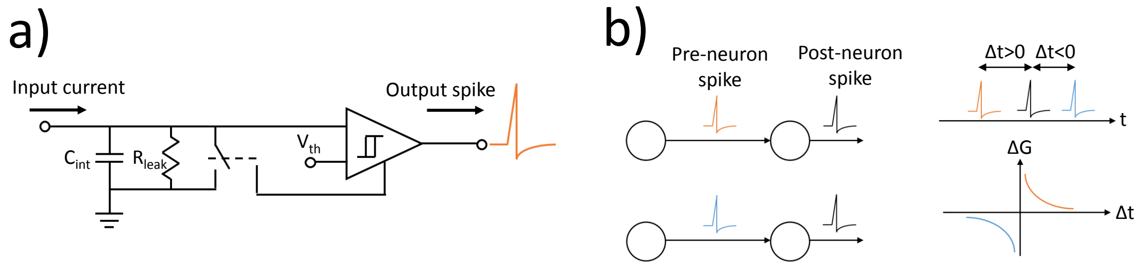


Figure 2.4: a) Circuit model of a leaky-integrate and fire neuron. The corresponding waveform of a spike is shown on the right. b) Spike-timing dependent plasticity of a pre-neuron and post-neuron spike.

2.3 Physical Implementation

The prior described neural networks are nowadays mostly simulated on conventional digital electronic circuits. Conventional digital electronics is not well suited for implementing artificial neural networks due to the following reasons:

The fact that a neural network consists of many multiplications and many weights, lead to the necessity to store some weights on a memory besides the processor (e.g. dynamic random access memory (DRAM)). Most computers are nowadays designed and built along the lines of the von Neumann-architecture (Fig. 2.5), thus the memory is strictly seperated from the arithmetic logic unit (ALU), where the calculation is taking place. Computers have a certain memory hierarchy from the registers, which are close to the ALU, to larger memory arrays, which are slower. The order is [59]: Register - static random access memory (SRAM) - DRAM - Flash - hard drive disk (HDD). In order store the weights of neural networks, which are nowadays >50 million [19], some of them need to be stored on the larger memory arrays and transferred for calculation to the ALU. The calculated result is often transferred back to the higher hierarchical memory. This data transfer consumes huge amounts of energy, because of the large physical distance between ALU and memory, a problem well-known as von Neumann bottleneck.

There are some digital computer architectures that are better suited for neural networks, like the GPUs and tensor processing unit (TPUs) from Google, due to their multi-core architecture and more parallelism. Still these processors just improve the energy efficiency by a factor of 10. For convolutional neural networks some processors have been developed that make use of weight reuse, due to the fact that the filter matrices are rather small (e.g. Eyeriss processor) [60, 61]. Most of these application specific integrated circuits (ASICs) still have energy efficiencies in the range of 1 TOPS/W-10 TOPS/W (Terraoperations per second per Watt) [20, 62] and are thus still by a factor of 10-100 less energy efficient than the human brain, besides that the number of parameters they can store per chip are very small. Tab. 1 gives a rough comparison of the brain versus computers nowadays: The brain can store much more synapses compared to the transistor count of conventional integrated circuits (ICs), but the single operations are conducted much slower in the brain and it is still much more powerful for certain tasks, like recognition. The reason for this difference is the inherent three-dimensional structure of the brain and the high parallelism in the brain. There is no distinct memory and processor in the brain and neural networks are directly physically implemented.

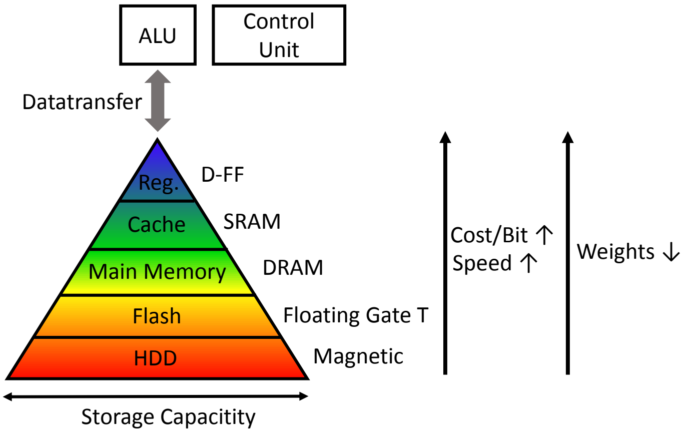


Figure 2.5: Illustration of von Neumann bottleneck with distinct memory and processor (ALU, Control Unit). The different memory hierarchies are shown.

Brain	Computer
1E+9 neurons, 1E+14 synapses	1E+9-1E+10 transistors
Max firing rate: 1×10^3 Hz	1×10^9 Hz clock speed
20 W	1000 W
10 fJ/operation	100 pJ/operation

Table 1: Properties of the brain and a typical desktop computer in comparison. The brain is approximately 10,000 times more energy efficient and can store more weights due to its three-dimensional structure.

One might think about implementing synapses/weights together with a multiplier unit in digital circuits directly and arrange them in a matrix (2.6b) for highly parallel calculation (also known as distributed architecture) [2], but for each 8 Bit register for storing the weights there are ~128 transistors necessary. With SRAM there are ~48 transistors necessary for 8 Bit. Furthermore, for a 8 Bit multiplication ~400 transistors are necessary [63], which gives a total amount of ~500 transistors for one node in the matrix. Assuming an advanced 7 nm technology node with ~100 MTr/mm^2 the footprint of this circuit would be ~2.2 μm x 2.2 μm or ~5 μm^2.

The advantage of digital circuits are their noise immunity and easy design, but the circuits themselves are very bulky. Neural Networks are to some extent inherently noise immune, and thus the precision for neural network calculation was lowered from floating point numbers to 6-8 Bit integer precision [11,60]. There are even publications on binary weight precision [64]. If we now consider an analog storage device, like a variable resistor with a memory effect, the device can accomplish a multiplication by using Ohms law and an accumulation operation with Kirchhoff´s current law if arranged in a crossbar arrangement (Fig. 2.6c) [1,65]. A fully-connected neural network (Fig. 2.6a) can be perfectly mapped to the crossbar arrangement with the pre-neurons attached to the input lines (WL) and the post-neurons attached to the output lines (BL). The advantage of this analog resistive device is the much smaller footprint, even at matured technology nodes, compared to the digital solutions (assuming 16F^2 memory footprint and 90 nm technology node: 0.13 μm^2 or 0.36 μm x 0.36 μm). Thus the area is 38 times smaller for the analog version at 90 nm technology node compared to the digital version at 7 nm technology node and much more weights can be stored per chip. A precision of 6-8 Bit is feasible for an analog weight storage and neural networks, as mentioned earlier, are inherently tolerant to some extent with respect to device-to-device variation and write/read inaccuracy [11,12]. Moreover, a single memory device is easier to integrate into the third dimension compared to circuits. The weights in a crossbar are often stored in a differential topography with two memory cells (Fig. 2.6d), thus the negative BL is subtracted from the positive BL and a 'four-quadrant multiplication' can be implemented. The differential weight configuration also makes the matrix more robust to variations and read-out errors. In addition the matrix can contain a selector, like a transistor, which is mainly used during writing/training. The input signals to the WL of the matrix can be either coded as an analog value, where a large digital analog converter (DAC) and a perfect linear device is needed in this case, or as a pulse length or pulse number [1], where the disadvantages of the prior mentioned code scheme do not exist (Fig. 2.6d). Only the latency could be higher.

Regarding inference and training, the memory cells should have several properties (see Tab. 2) of which not all properties are yet reached with common resistive technologies [11,12]:

For training the weights should be updated in a linear fashion and highly symmetric with respect to writing and erasing, because the derivative of the weight update during backpropagation training should stay constant (see chapter 2.1.3). A high endurance is necessary for training. For inference on the other side a long retention time is required. A high dynamic range (on/off ratio) is desired for both and the device should be as resistive as possible to enable low power consumption and no saturation of the total accumulated result ($R \approx 20\,\mathrm{M\Omega}$).

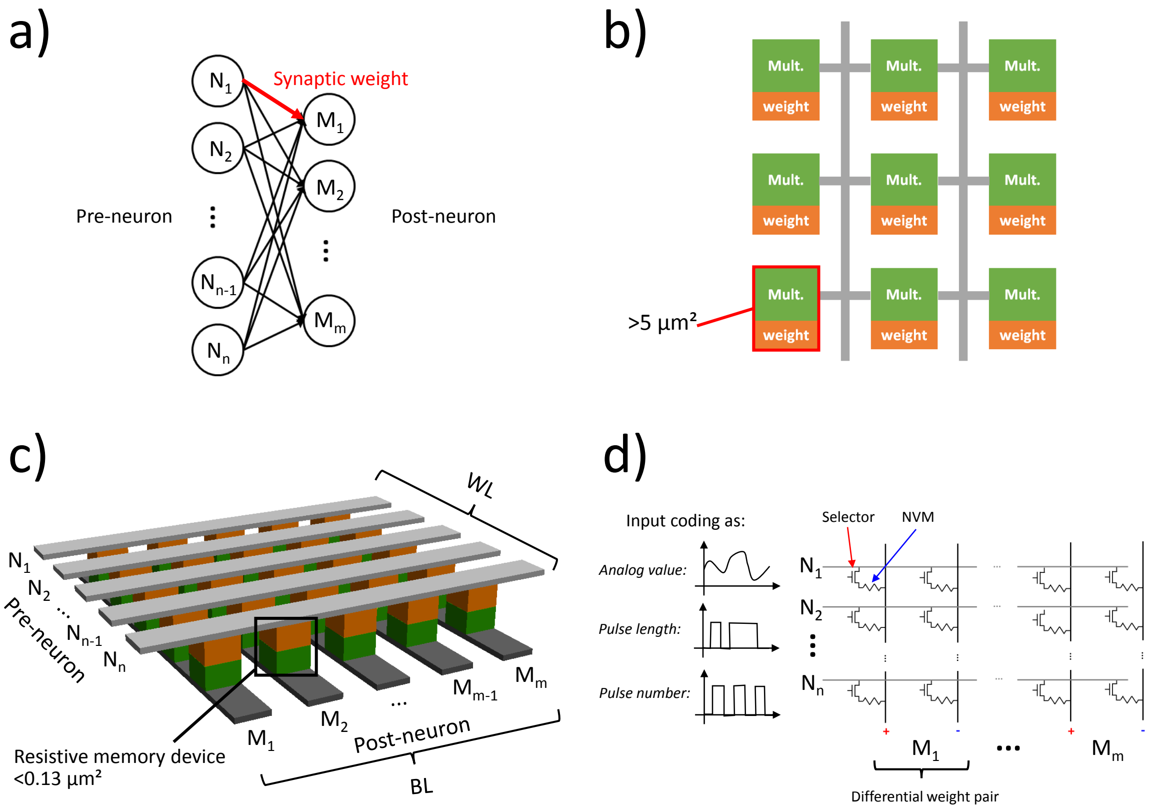


Figure 2.6: a) Fully connected neural network with the pre-neurons N_1 to N_n and post-neurons M_1 to M_n b) Matrix arrangement with conventional digital CMOS to map the neural network. c) Crossbar arrangement of resistive devices. d) Crossbar arrangement with selector in differential topology and the different input coding schemes. Similar illustrations can be found in [1,2].

Inference	Both	Training
Long retention Low read-out energy	High $R_{device} \approx 20M\Omega$ Dynamic range >1:60-1:100 Precision: 6-8 Bit	Low write/erase asymmetry: <6-10% High linearity Low noise: $\sigma < 10\%$ High endurance: >1E+9 Fast programming: 10 ns-100 ns

Table 2: Desired properties of the resistive memory device for inference, training and both. [11,12]

2.3.1 Common Resistive Devices

One of the first resistive devices used for neuromorphic computing was the floating gate transistor [66,67]. Electrons are injected and stored in an isolated poly-Si gate (floating gate) either by Fowler-Nordheim tunneling or hot carrier injection. The stored charge leads to a shifting of the threshold voltage of a MOSFET. Similarly, in a more modern version, a silicon-oxide-nitride-oxide-silion (SONOS) memory [68–71] uses charge trapping in a silicon nitride layer and Fowler-Nordheim tunneling for writing and erasing. SONOS has a higher endurance compared to floating gate transistors. Both types of charge trapping memories are too slow and energy consuming with respect to writing for training applications, but are interesting for inference-only application, since they are industrially matured.

A memristor, also known as RRAM in the context of memory technology, is a device first postulated by Leon Chua in 1971 [72], which was experimentally described in 2008 [22]. It has a typical pinched hystersis loop in the IV-curve. Although a memristor has a rigorous mathematical definition [24], sometimes other resistive memory technologies are also called a 'memristor'. There is even an ongoing discussion if the device from 2008 is really a memristor [73, 74]. The memristor is written to the low resistance state (LRS) by formation of a conducting filament, which is composed out of oxygen vacancies or metal ions, between two metal electrodes. Usually the setting is an abrupt process, while the reset can be achieved gradual, thus RRAM is a highly asymmetric device [1]. With non-localized switching, like in PCMO devices, the conductance can be tuned more gradually. The first demonstration on a memristor crossbar array was implemented for a 3x3 image recognition with Manhatten update training [75]. Recently, memristor based crossbar arrays have also been implemented with CMOS [58]. The biggest problems with memristors are their asymmetry, large device-to-device variations and slowness during gradual switching.

In a phase-change memory a resistance change in a material is achieved by a phase transition from crystalline to amorphous. The material is molten by an electric current and abruptly solidified to obtain the amorphous phase. Although phase-change memories are a very robust memory technology with large endurance, the asymmetry is one of the biggest obstacle for neuromorphic applications. There have been implementations with up to 165,000 synapses by using phase change memory arrays [76].

Regarding spintronic implementations, a tunnel junction is used, which depending on the magnetization direction of the two magnetic layers, leads to a low or high resistive state. In order to obtain a analog value storage, a domain wall motion is used in one magnetic layer and depending on the fraction of the up to the down domain an arbitrary resistance state will be adjusted [77, 78]. The advantage of spintronic realisations are the high endurance and good controllability of domain wall motion. Scalability still needs to be improved due to the large size of domain walls.

Ferroelectric resistive realisations can either be in the form of a FeFET [79–84] or a ferroelectric tunnel junction [85, 86]. The discovery of ferroelectric hafniumoxide in 2011 [87, 88] lead to highly CMOS compatible device implementations. In a FeFET the polarization charge of the ferroelectric material leads to a shifting of the threshold voltage. There is still an ongoing debate regarding the scalability of the analog storage capabilities. In the material one has to distinguish between the switching of grains due to small differences in the coercive field [81]. This type of switching is clearly limited with regards to scalability to $\sim$500 nm, due to limited grain size. The other type of switching events are nucleation events within the grains [3, 89], which are exponentially dependent on the applied voltage, and can only be measured indirectly in a nanoscaled FeFET by abrupt switching events after a certain number of pulses. Generally, FeFET show a very good symmetry and thus high training accuracies for neural networks have been simulated [83]. Ferroelectric tunnel junctions have been shown to have very high energy efficiencies and potentially a better analog value performance compared to FeFETs, because of their two dimensional coupling.

Resistive technologies are proven to achieve energy efficiencies during inference of up to 100 TOPS/W [12, 62, 68, 85] for 6-8 Bit resolution and an areal efficiency of up to 3.6 TOPS/mm^2 [90].

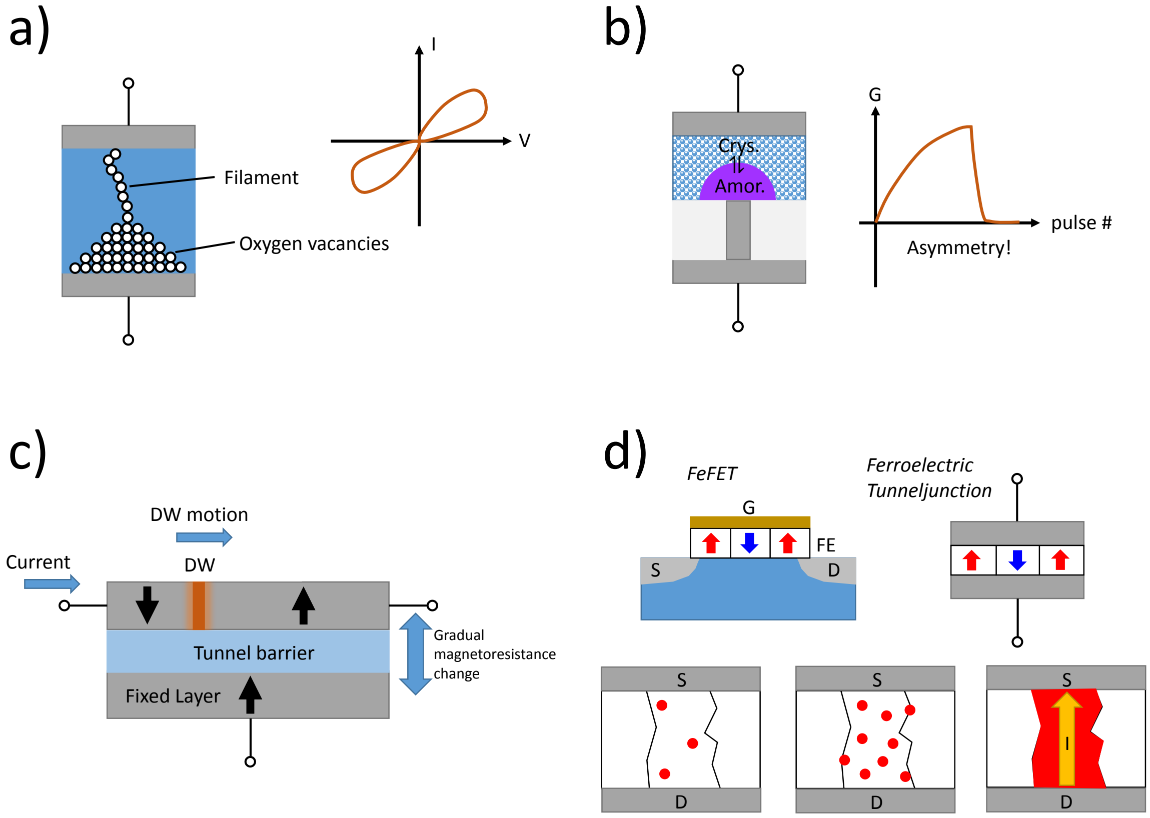


Figure 2.7: a) RRAM device with conducting filament and corresponding pinched hysteresis loop of IV curve. b) Phase-change memory cell with asymmetric conductance response versus write pulse number. c) Spintronic memristor based on domain wall motion. d) FeFET and ferroelectric tunnel juntion based synapse. Shown are domain nucleation events and single grain switching. Similar illustrations can be found in [3, 4].

2.3.2 Capacitive Devices

One can consider to use a variable capacitor along with a memory effect instead of a resistor for neuromorphic computing. Similarly to resistive devices a memcapacitor is a device with pinched hystersis loop in the charge-voltage curve [24]. There have been some theoretical proposals for memcapacitive devices [23–31], but few practical implementations [6, 7, 32, 33].

There are some reports of DRAM trenched capacitors [91, 92] as a capacitive synapse, however, these implementations only use the stored charge as a weighted value, cannot vary capacitance of the device and moreover, such devices are volatile. However, very high linearity has been obtained with trenched capacitor implementations.

A variable capacitance along with a non-volatile memory effect can be realized by either varying the distance between the capacitor plates or their surface areas, or by varying the dielectric constant of the insulating medium. Thus these devices can be devided into 'varying plate distance memcapacitors', 'varying surface area memcapacitors' and 'varying dielectric constant memcapacitors'.

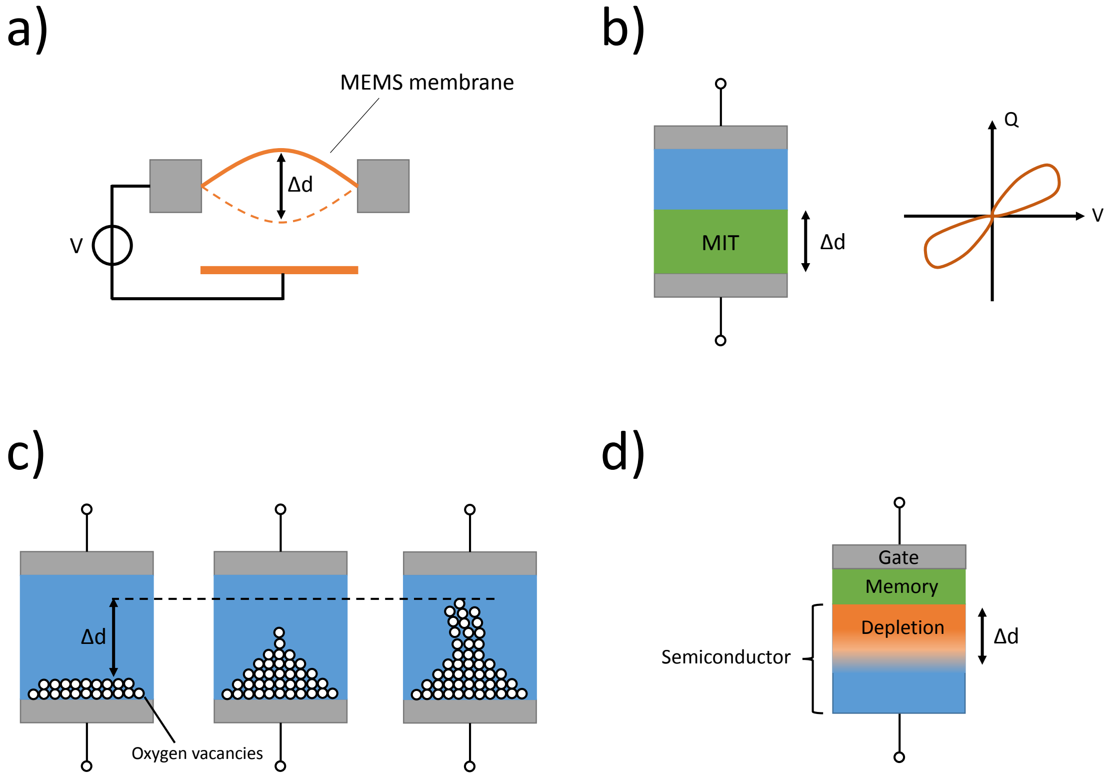


Figure 2.8: Variable capacitance with changing plate distance based on: a) MEMS device (a similar illustration can be found in [5]), b) metal-to-insulator transition (MIT) material in series to normal dielectric layer, c) filament formation in a memristor, d) MOS capacitance with variable depletion layer.

Devices with variable plate distance d can be implemented in various ways and are shown in Fig. 2.8: MEMS [5, 93]; a metal-to-insulator transition material in series with a dielectric layer [31]; changing the oxygen vacancy front in a classical memristor [27, 28]; or a simple Metal-Oxide-Semiconductor (MOS) capacitor with a memory effect [32, 33].

Every classical memristor has a parasitic memcapacitive effect due to the changing oxygen vacancy front distance [27]. Optimisation of a memristor with a low resistive effect and large capacitive effect is very difficult because of leakage currents. Generally for providing a large dynamic range with plate distance variation, one needs either a very small plate distances, where tunnel currents are dominant and thus degrade power efficiency [27, 28], or very large plate distances, which results in problems with lateral scalability and stray coupling to neighboring cells [32].

Changes in surface area have been investigated by other groups using a memristor with a much smaller area, and a normal capacitor in series with a much larger area [6] (Fig. 2.9a). The memristor, if switched off, acts as a small parasitic capacitance. In this case, the scalability of the small surface area stray capacitor (memristor) is limited by lithographic patterning, and since the series capacitor needs to be much larger, the full potential of lateral scalability cannot be exploited.

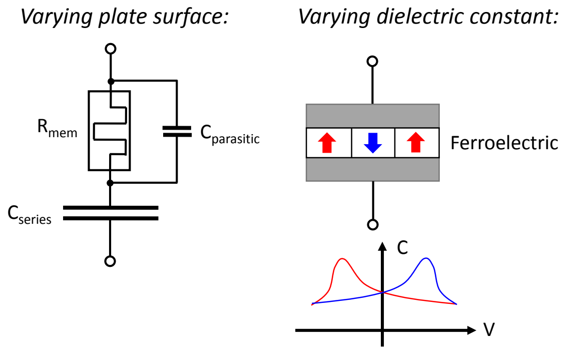


Figure 2.9: 'Varying surface memcapacitor' based on memristor in series to a large area capacitor and 'Varying dielectric constant memcapacitor' based on MFM capacitor. Similar illustrations can be found in [6, 7].

Finally, changes in the dielectric constant limit the choice of materials. Materials with very large dielectric constants are usually perovskites ($\epsilon_r > 1000$) and they are generally not CMOS compatible, which makes their use rather difficult. There is a capacitive device proposal based on a MFM capacitor with doped HfO_2 [7], which falls into the category of a 'varying dielectric constant memcapacitor', but the dynamic range is rather low and overwrite during read-out might be a problem (Fig. 2.9b). Another implementation used the change in dielectric constant of VO_2 at its metal-to-insulator transition point [94]. However, the change in dielectric constant is only visible at very high frequency (THz).

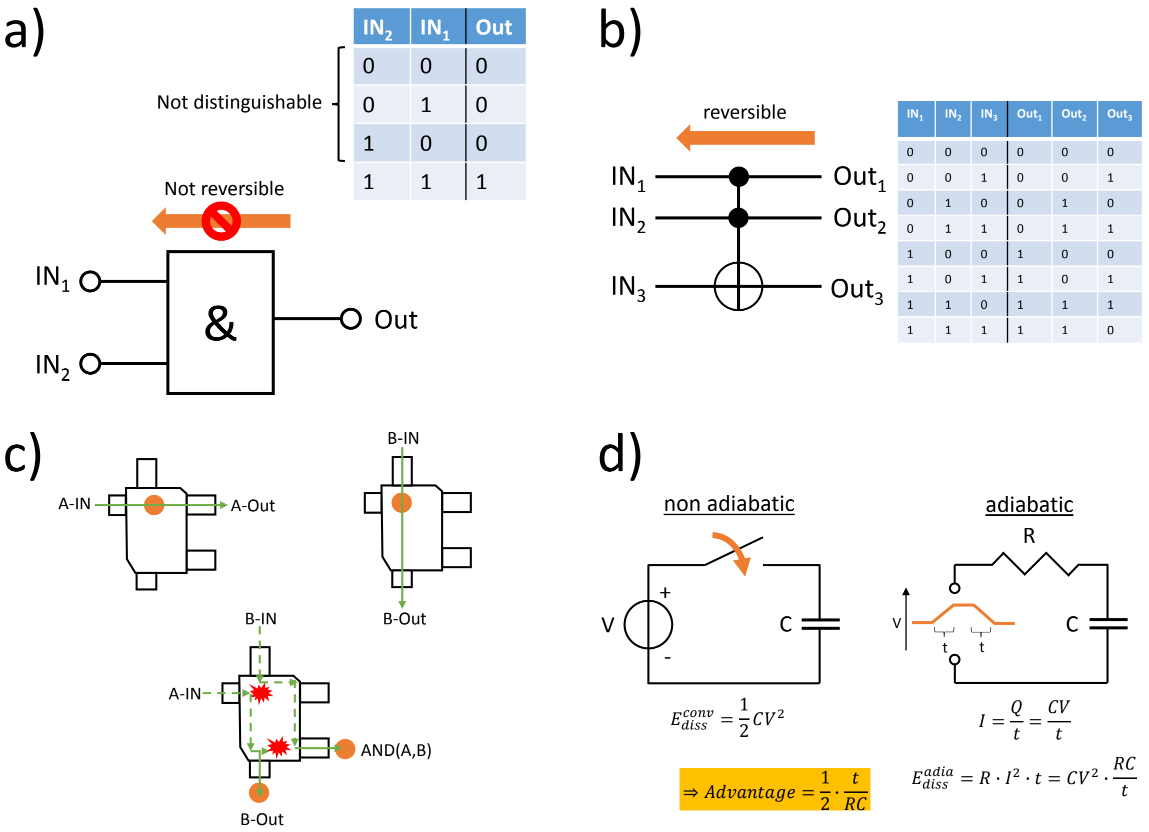
2.4 Landauer Principle and Reversible Computing

In this chapter a short overview on the physical limitations of energy efficient computation is given. In 1961 Rolf Landauer explained that information and entropy are linked and that logically irreversible computation is accompanied by an entropy increase [95]. This is also known as Landauer´s principle. In a classical AND gate as shown in Fig. 2.10a), for the zero input state there are three possible input states. Thus it is impossible to state from the output state, which input state was applied. Therefore, the information of one bit is erased and the information loss is accompanied by a minimum energy converted into heat:

$$E = k_B T \cdot \ln 2 \tag{2.4.1}$$

Where k_B is the Boltzmann constant and T the temperature. In order to circumvent the Landauer limit, reversible computing schemes were suggested by Charles Bennett in 1973 [96–98]. Special logic operations, like the Toffoli gate or Fredkin gate were developed and later adopted for quantum computation. Reversible gates have the same number of inputs as outputs in order to avoid information losses. The Toffoli Gate is similar to the AND gate, where the third input is flipped, if the first two inputs are 1 (Fig. 2.10b). The first two inputs are directly connected to the outputs and are not manipulated. With reversible computing the energy consumption can be lowered far below the Landauer limit.

A possible physical implementation of a reversible computer is the billard ball machine [99, 100] (Fig. 2.10c): Only if the balls are fed into both inputs, an AND operation is conducted, otherwise the input billard balls are directed to each output. The collisions have to be as elastic as possible and friction as low as possible in order to be as reversible as possible. Generally, physical implementations of reversible computing are about reducing friction and other non-ideal processes to reach ultra low power consumption. Another implementation used nanomechanical rods [101]. Quite often oscillators with reduced damping are used for reversible computing.


a)
Not distinguishable
Not reversible
IN$_1$
IN$_2$
&
Out
b)
reversible
IN$_1$
IN$_2$
IN$_3$
Out$_1$
Out$_2$
Out$_3$
c)
A-IN
A-Out
B-IN
B-Out
B-IN
A-IN
AND(A,B)
B-Out
d)
non adiabatic
adiabatic
V
+
-
C
R
V
t
t
C
$E_{diss}^{conv} = \frac{1}{2} CV^2$
$I = \frac{Q}{t} = \frac{CV}{t}$
$\Rightarrow Advantage = \frac{1}{2} \cdot \frac{t}{RC}$
$E_{diss}^{adia} = R \cdot I^2 \cdot t = CV^2 \cdot \frac{RC}{t}$

IN$_2$	IN$_1$	Out
0	0	0
0	1	0
1	0	0
1	1	1

IN$_1$	IN$_2$	IN$_3$	Out$_1$	Out$_2$	Out$_3$
0	0	0	0	0	0
0	0	1	0	0	1
0	1	0	0	1	0
0	1	1	0	1	1
1	0	0	1	0	0
1	0	1	1	0	1
1	1	0	1	1	1
1	1	1	1	1	0

Figure 2.10: a) Irreversibility of the AND gate: It is not possible to conclude from the output "0", which input was applied. b) Reversibility of the Toffoli gate. c) Physical implementation of a reversible computer: The billiard ball computer. Ultra low friction and elastic collisions are assumed. d) Non-adiabatic charging versus adiabatic charging for a capacitor. The slow charging of the capacitor gives an advantage compared to the abrupt charging.

Adiabatic circuits are implementations of reversible concepts in CMOS technology [97, 102–110]. CMOS is mostly about charging other gate capacitances during computation, and this charging energy is dissipated and also known as dynamic power consumption. In a classical two transistor inverter first the energy is wasted during a transition from one to zero in the upper NMOS transistor due to the fast transition, and finally during the transition from zero to one the energy stored on the capacitor is dissipated on the PMOS. Especially the abrupt charging transitions in CMOS give rise to a large amount of resistive and thus irreversible losses. On the other hand, in adiabatic circuits the capacitors are charged slower (Fig. 2.10d) with a voltage ramp applied to the circuit. Thus a power clock gating is used in adiabatic circuits and the power clock is able to recover the charges/energy of the capacitors during discharging by using an energy storage, usually an off-chip inductor. The amount of recovered energy is dependent on the efficiency of the oscillator and depends, amongst others, on the quality factor of the inductor. Typical quality factors for inductors are in the range of a few dozens to hundred [107, 109, 111].

Adiabatic circuit design helps to reduce the dynamic power consumption, but on the other hand side the circuits become slow and large [97]. Due to scaling of CMOS according to Moore´s law the subthreshold leakage power has become the most dominant power loss in advanced CMOS (<65 nm) [112]. Adiabatic circuits are not suited to solve the subthreshold leakage problem.
The use of adiabatic charging for memcapacitive crossbars will be explained further in chapter 3.5.

3 Device Description and Theory

3.1 General Working Principle

In subchapter 2.3.2 several implementations were discussed for realising memcapacitors. In order to solve the problem of limited scalability with ’varying plate distance’, or ’varying area’ in a (mem)-capacitor, a device design which is based on charge shielding is proposed in this thesis. The general structure is shown in Fig. 3.1a) and consists of a top gate electrode, a shielding layer with contacts and a back side read-out electrode, where each of these layers are separated by two dielectric layers. The capacitive coupling between the top gate electrode and bottom read-out electrode is detected during read-out and the coupling depends on the state of the shielding layer.

The lateral scalability is significantly better compared to the previously mentioned concepts, since the thickness of the layers can be readily optimized, while the on/off ratio is mainly dependent on the shielding efficiency of the shielding layer. Generally charge screening depends on the Debye screening length L_D [113, 114]:

$$L_D = \sqrt{\frac{\epsilon_0 \epsilon_r k_B T}{n^2 e^2}} \tag{3.1.1}$$

where k_B is the Boltzmann constant, T is the temperature, n is the charge carrier concentration and e is the elementary charge. The electric field drops exponentially within the shielding layer and drops to 37% within the screening length L_D when the condition, $\Psi < U_T$, applies, where ψ is the potential in the semiconductor and U_T is the thermal voltage:

$$U_T = \frac{k_B T}{e} \tag{3.1.2}$$

The Debye screening length is only a linear approximation of the following differential equation [114]:

$$\frac{d^2\Psi}{dx^2} = -\frac{e}{\epsilon_0 \epsilon_r} \cdot \left(p_0 \cdot \left[\exp\left(\frac{-\Psi}{U_T}\right) - 1 \right] - n_0 \cdot \left[\exp\left(\frac{\Psi}{U_T}\right) - 1 \right] \right) \tag{3.1.3}$$

where p_0 and n_0 are the charge carrier concentrations of holes and electrons in thermal equilibrium, respectively. Therefore, using the Debye screening length 3.1.1, given the exponential spatial dependence of the field in the material, is only a linear approximation of the non-linear differential equation 3.1.3. Especially for strong inversion and accumulation within the shielding layer, the length scales of screening become much smaller than the Debye length. This non-linearity with respect to the semiconductor potential leads to either strong shielding or fairly good transmission.

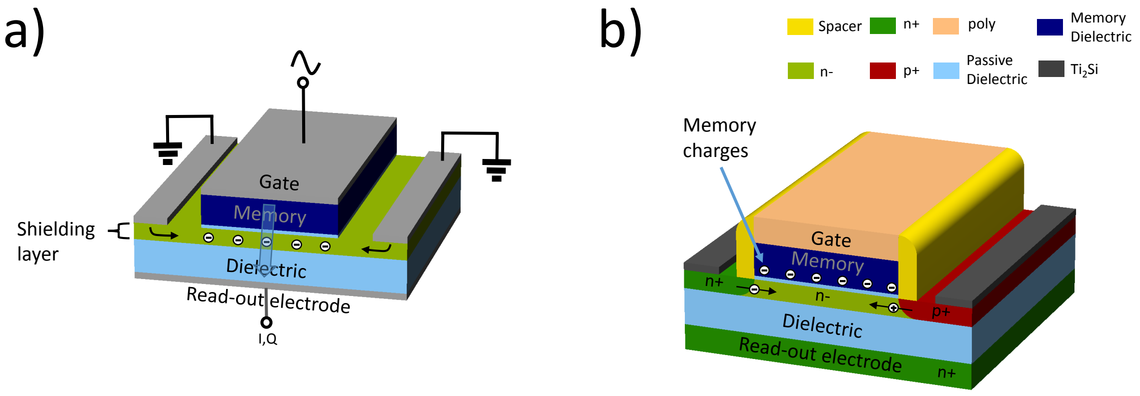


Figure 3.1: a) General Device structure with gate electrode, memory dielectric, shielding layer, passive dielectric and read-out electrode. The capacitive coupling is indicated by the blue arrow. b) Device structure with lateral p+n-n+ junction in the shielding layer. Material from [8]

There are several possibilities on how to adjust the shielding efficiency in a non-volatile manner: One option is to change the resistivity of the material, which can e.g. be implemented with a metal-to-insulator transition. Another option is to use a simple semiconductor and use a memory dielectric in the top dielectric layer. Throughout this thesis only the latter type of variable capacitance along with a non-volatile memory effect is investigated (Fig. 3.1b). This can be either a ferroelectric or a charge-trapping memory. The stored charges (either trapped charges or polarization charge) will drive the semiconductor potential either to strong inversion/accumulation or depletion and thus adjust the shielding efficiency. Furthermore, a lateral p+n-n+ junction is used in the shielding layer, which enables more functionality:

The p+ and n+ regions act as reservoirs for holes and electrons and lead to a symmetric device behaviour with respect to negative or positive gate voltages. Symmetry, especially during writing, is an important feature for neuromorphic devices, because it leads to higher training accuracies [1]. Moreover, the shielding layer can be depleted or enhanced with carriers actively by applying voltages to the p+ and n+ regions. This feature can e.g. be used for deselecting columns in a matrix arrangement, if necessary. As shown in Fig. 3.2, the single device can be arranged into a crossbar for highly-parallel MAC operations. In this case the gate electrode is connected to the WL, where input signals are applied, and the SL is connected to the n+ and p+ regions in vertical direction to the WL. The read-out electrode functions as the BL, which is parallel to the SL, and the accumulated charge out of one BL is the calculated result of accumulated multiplications (MAC) at each crossing point. The multiplication is conducted between the input signal of the WL and the state of the shielding layer, which in turn is adjusted by the memory material. The weights are encoded in the capacitance of each crossing point.

Writing of each memory cell can be achieved by applying a voltage between the SL and WL. The voltages can be chosen in such a way that the disturb level is 1/3 of the non-disturbed memory cell (see chapter 6.5). Furthermore, certain SL can also be deselected by depletion of the p+n-n+ junction, as mentioned earlier.

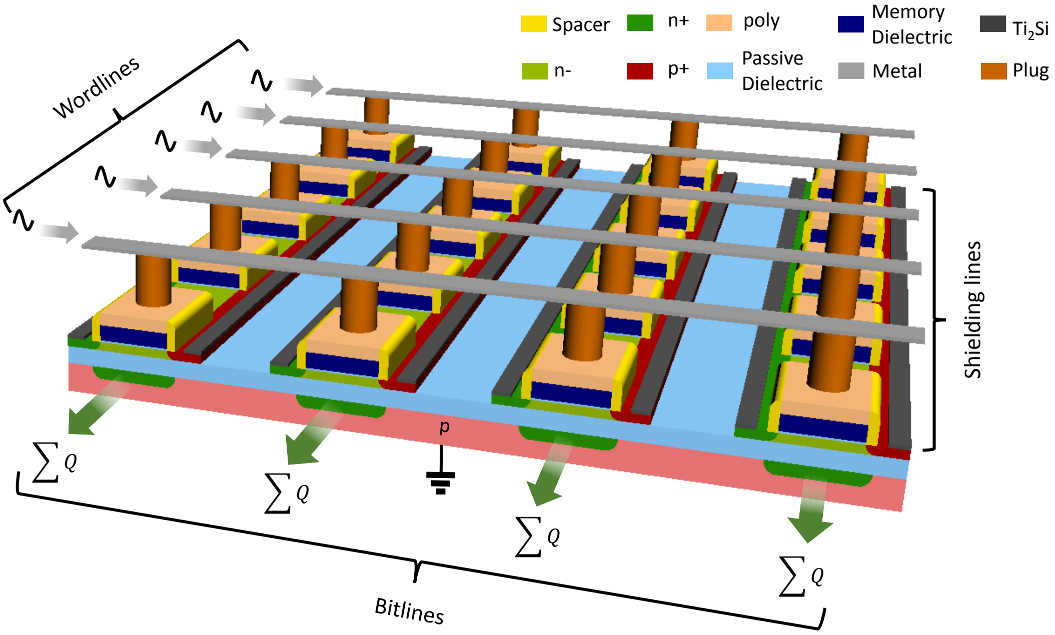


Figure 3.2: a) Crossbar arrangement of devices for highly parallel MAC operations. The inputs signals are applied to the wordlines and weighted by the shielding layer and memory material. The accumulated result is read-out at the bitlines. Material from [8]

3.2 TCAD Simulation on Single Devices

With TCAD Simulation the device behaviour can be simulated before fabricating it. A drift-diffusion simulation of the 90 nm gate length and width device in Fig. 3.3a) was performed in Synopsys. The obtained CV curves of coupling capacitance between the gate and the read-out electrode are shown in Fig. 3.3b). In this simulation the top dielectric layer had no memory effect in order to verify the general capacitive coupling behaviour. As indicated in chapter 3.1 there is a certain gate voltage region (depletion region), where electric field transmission is large. Towards higher positive or negative gate voltages the transmission is low (accumulation and inversion region). This leads to the observed capacitive coupling window in Fig. 3.3b).
The p+ and n+ junction was applied to different voltages, where V_{AK} is the voltage difference of the p+ and n+ contact. Depending as to whether the p+sn+ diode is used in forward or reverse bias, the coupling window range can be made broader (reverse) or extinguished (forward). The broadening of the curve is proportional to the applied reverse voltage of V_{AK}, which can be accounted for by the splitting of

the Fermi potential in quasi Fermi potentials, which is directly related to the applied voltage V_{AK}. The half peak width of the coupling window is plotted in Fig. 3.3c) with respect to different V_{AK} and a linear relationship becomes obvious. The reason for this proportionality can be explained from equation 3.1.3 with the hole and electron concentrations:

$$n = n_i \cdot \exp\left(\frac{\phi_n - \phi_i}{U_T}\right) \tag{3.2.1}$$

$$p = n_i \cdot \exp\left(\frac{\phi_i - \phi_p}{U_T}\right) \tag{3.2.2}$$

With ϕ_i being the intrinsic Fermi potential and ϕ_p the quasi-Fermi potentials of holes and ϕ_n of electrons. The intrinsic charge carrier concentration is described by n_i. For a semiconductor in equilibrium the quasi-Fermi potentials are equal: $\phi_p = \phi_n$. In case of a pn-junction the semiconductor is brought out of equilibrium and thus the applied voltage V_{AK} is related to the quasi-Fermi potentials:

$$V_{AK} = \phi_n - \phi_p \tag{3.2.3}$$

Due to the fact, that the applied voltages to the p+ and n+ junction are antisymmetric one can conclude:

$$\phi_n = -V_{AK}/2 \tag{3.2.4}$$

$$\phi_p = V_{AK}/2 \tag{3.2.5}$$

These terms can be put back into equation 3.1.3 and the right term rapidly increases, when the exponential terms start to dominate (accumulation and inversion). That is when the shielding takes place and the point where the exponential terms start to dominate is shifted exactly by $\pm V_{AK}/2$. This is how the broadening by approximately V_{AK} is explained.

The transmitted field through the shielding layer is proportional to the accumulated charge on the read-out electrode and the charge is shown in Fig. 3.3d) and describes a sigmoid/tanh behaviour. The saturation comes from the inversion/accumulation. The curves in Fig. 3.3b) are proportional to the derivatives of the curves in Fig. 3.3d). Sigmoid transfer functions play an important role in modeling neurons in artificial neural networks, and furthermore, adjusting their broadness by application of reverse or forward voltages provides additional functionality. A common practice during training is the dropout of certain neurons to avoid overfitting [115], which can be implemented though the forward direction of the diode. Furthermore, the application of the voltage

V_{AK} can also enable selection and deselection of memory cells, if necessary.

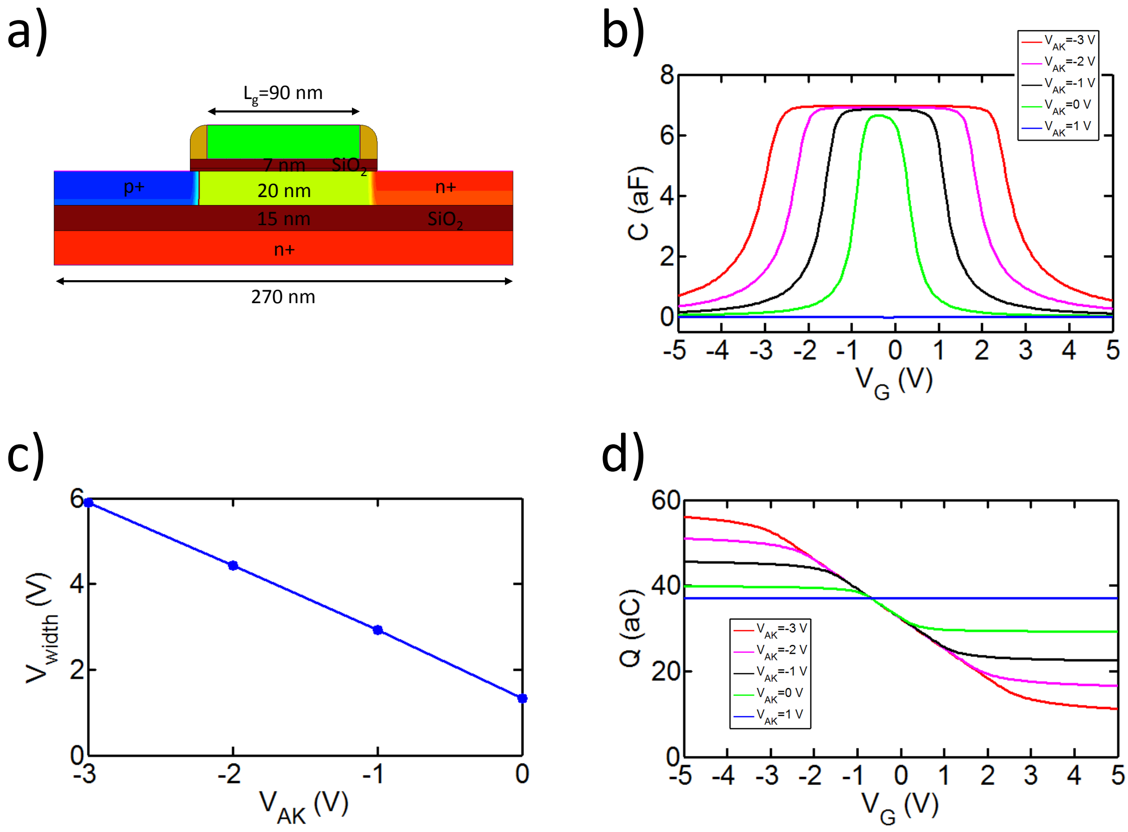


Figure 3.3: a) Simulated structure and dimensions. Silicon oxide (no memory effect) was used for the dielectric layers. b) CV curves for the capacitive coupling between gate electrode and read-out electrode. A voltage V_{AK} was applied to the p+n-n+ junction. c) Half-width broadness of the capacitive window in b) versus applied V_{AK}. d) Accumulated charge on read-out electrode versus applied gate voltage. Material from [8]

Along with the CV curves in Fig. 3.3b) the large on/off ratio can also be observed (up to 1:90). Fig. 3.4a) shows the capacitive swing for different gate length and gate oxide thicknesses. Generally, longer gate length and thinner gate oxides lead to larger on/off ratios. The reason for the degraded on/off ratio for shorter gate length is due to the short channel effect: At the border to the p+n- or n-n+ junction accumulation or inversion charges get rejected because of the space charge region. Therefore, higher gate voltages are required to overcome the rejection and less shielding charges are supplied at the border. The border region is around 5 nm wide and becomes especially relevant for shorter gate length. Decreasing the equivalent oxide thickness (EOT) increases the electric field and therefore the shielding charge rejection is reduced. Usually in transistor designs Halo-implants are used to reduce short-channel effects [114].
Including memory charges in the top gate dielectric leads to a shifting of the CV curves, as shown in Fig. 3.4b). Typical memory window (MW) for charge trapping memories

are in the range of 2.5 V to 3 V. For ferroelectric memories the MW depends on the thickness and coercive field, but is usually in the range of 1 V to 2 V for ferroelectric HfO_2 [116]. Recently, a memory window of 2.5 V was reported for thick doped hafnium oxide films [117]. Thus the memory window in Fig. 3.4b) was chosen to be 3 V.
The CV curves can be shifted in a gradual manner and a read-out sinusoidal signal is indicated between 1 V and 2 V gate voltage and thus covers the blue shifted CV curve. Fig. 3.4c) shows the accumulated charge on the read-out electrode for a half-period of the shown read-out sinusoidal signal. The far left shifted curves (orange-red) are mostly used to turn the device off, while the blue coupling window gives the largest read-out charge. Fig. 3.4d) shows the total accumulated charge over a half-period for different coupling window shifts V_{shift}.

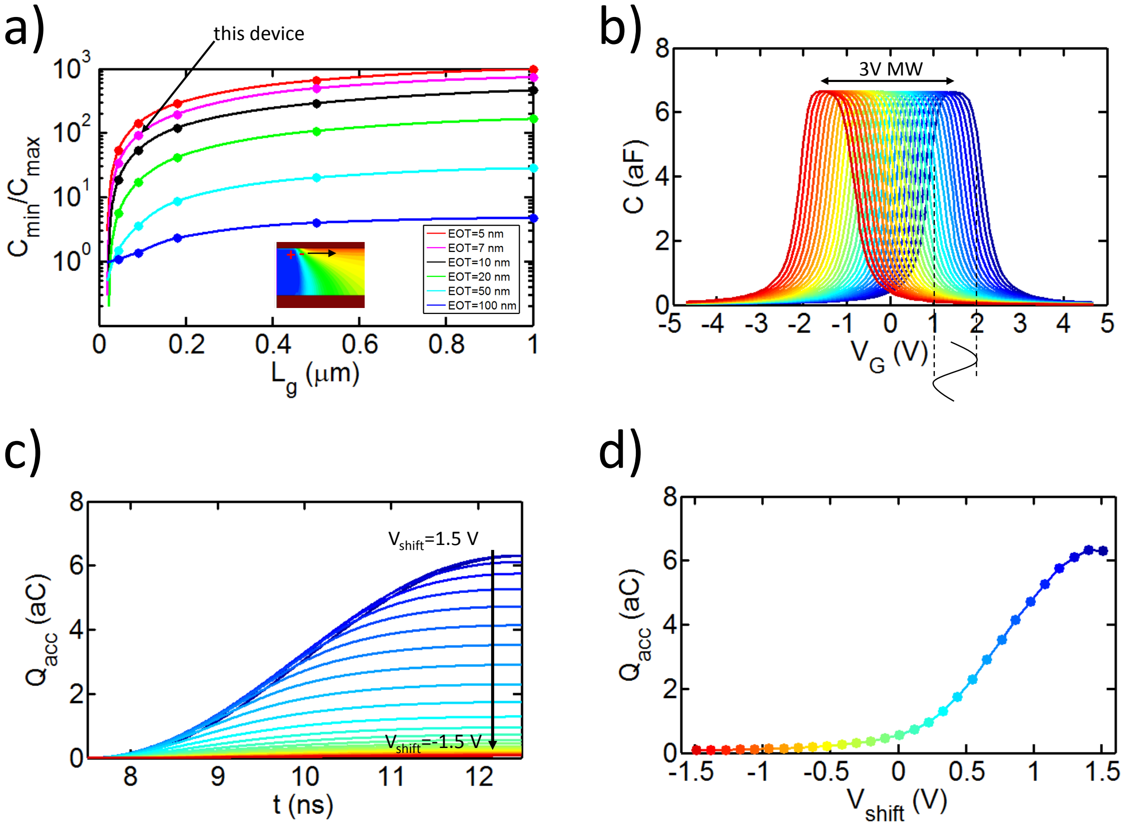


Figure 3.4: a) Capacitive swing for different gate length and equivalent gate oxide thicknesses (EOT). The inset shows the accumulated electron concentration and the charge rejection due to the short channel effect. b) CV curve shifting due to memory charges and c) The corresponding accumulated charge on the read-out electrode over a half period of the sinusoidal signal in b. d) Total accumulated charge over a half period for different CV curve shift voltages V_{shift}. Material from [8]

3.3 Lateral Scalability and Read Signal Strength

With regard to lateral scalability it is necessary to distinguish three aspects:

1) The scalability of the memory technology/material itself with regard to how many levels can be written.
2) The sensitivity of the sense amplifier.
3) The noise level of a single device during read-out.

Especially the last two aspects can be quite different for capacitive devices compared to common resistive technologies. It is well-known from other capactive memory technologies (like DRAM, ferroelectric capacitors) that three dimensional capacitors are necessary at a certain technology node due to low read-out signals [118, 119]. In the case of neuromorphic applications, one needs to consider that many memory cells are read-out at the same time and only the accumulated result is relevant for further processing, which can be much larger than the signal of single memory cells in conventional memory technology. For this reason, resistive memories for neuromorphic application are made more resistive [1, 11, 12]. Quite common resolutions for input, weight and output signals for neural networks are in the range of 4-8 bit (16-256 levels) [11, 60]. This analog-like resolution has significant influence on scalability and thus needs to be considered.

1) With respect to the first aspect, differences of a charge trapping memory and a ferroelectric memory were already explained in chapter 2.3.1. For the charge trapping memory a scalability down to 40 nm with up to 31 levels was already practically proven (SONOS memory) [68]. The scalability of a ferroelectric memory for analog storage is still unclear. On the one side, the grain size of the material can limit the multilevel switching (abrupt switching below 500 nm for FeFETs [81]), on the other side, nucleation events should lead to more levels than grain switching events. These nucleation events were only measured indirectly in ferroelectric field effect transistors (FeFETs) [4], but a one dimensional current path between source and drain is necessary there. For a two dimensional device, like a tunnel junction [85], or the device proposed here, the situation might be different.

2) Capacitive measurement techniques are quite common in the context of DNA sensing or chip interconnect measurement [9, 120–126]. Techniques for ultralow capacitance measurements include CBCM [9, 121–125], capacitance-to-frequency (C2F) conversion [127] and lock-in detection [128], where lock-in detection gives resolutions down to 65 zF. Lock-in detection will most likely not be suitable as a sense amplifier, since it occupies a relatively large chip area. CBCM resolutions down to <10 aF have been shown with a compact and easy to implement detection scheme. An example circuit for CBCM is shown in Fig. 3.5: The circuit consists of measurement capacitor C_S and a reference capacitor C_R, which are charged and discharged by two non-overlapping clock signals Φ_1 and Φ_2. Charging happens over the transistors M4 and M3, discharging over the transistors M2 and M1. Current mirrors are used to subtract the charging currents of each branch (M5+M7,M8+M6 and M10+M9). The subtracted current is finally integrated over another capacitor C_{int}, whose voltage is proportional to the measurement capacitor C_S.

Current mirrors and integration capacitors are also commonly used in sense amplifiers for classical memories [129]. Operational amplifiers with a feedback capacitor are sometimes used in the context of neuromorphic computing [130]. Since both implementations had similar sensitivities and the sensitivity was also comparable to CBCM circuits, the following calculation was done with a current mirror sense amplifier [129]:

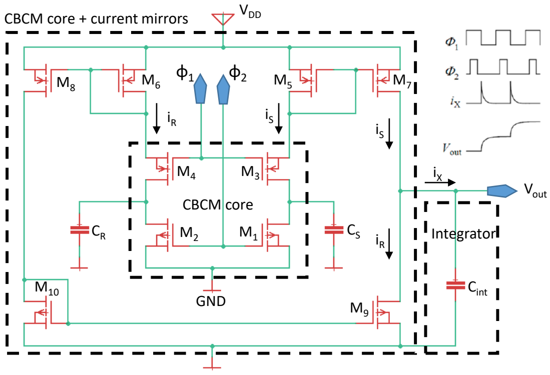


Figure 3.5: CBCM detection circuit with measurement capacitor C_S and reference capacitor C_R (a similar illustration can be found in [9]).

The charge integrating amplifier [129] can achieve up to 50 mV output voltage for 1 μA input current, within 3 ns or a calculated charge sensitivity of 16.67 mV/fC. Assuming 100 distinct values of the digitized output value (7 Bit) a voltage of ± 1.5 V must be achieved on the sense amplifier output (30 mV steps by Schmitt-Trigger or analog digital converter (ADC) [130]). The maximum/minimum output voltage corresponds according to the sensitivity to a charge of ± 90 fC. As mentioned earlier, in neuromorphic systems many memory cells are read-out at the same time, thus contributing to the charge of the sense amplifier. Usually array sizes for fully connected layers are in the range of a few thousand, for convolutional layers they can be smaller (in the range of 100). Thus for now it is assumed that 100 devices are read-out at the same time and the maximum charge of each of these memory cells is for the 90 nm devices 6.32 aC (see Fig. 3.4d). So the total charge over one half-period is for 100 devices 0.632 fC.

So in total the following number of half-periods in the input signal are necessary to achieve the desired 90 fC:

$$N_{per} = \frac{90\,\mathrm{fC}}{0.632\,\mathrm{fC}} = 142 \tag{3.3.1}$$

This number fits well into the 7-8 bit range of the input signal. Thus similarly to resistive neuromorphic systems also a pulse length or pulse number coding for the input value can be used to avoid non-linear multiplications [1] and achieve large enough output signals. The used switched capacitor approach for integrating over many periods is explained in chapter 5.2.

3) Regarding the noise immunity, one has to consider kTC noise [131] in the case of capacitive devices and the effective noise voltage v_n is:

$$v_n = \sqrt{\frac{k_B T}{C}} \tag{3.3.2}$$

For the 6.65 aF capacitance from Fig. 3.4b) one obtains an effective noise voltage of 25 mV, which is 14 times less than the effective read-out voltage (0.35 V) in Fig. 3.4b). Furthermore, one has to consider that the noise level decreases with the number of repetitive measurements by $\frac{1}{\sqrt{N_{per}}}$, where N_{per} is the number of repetitive measurements, or in this context the number of periods (142), which will result in a total noise level of 2.2 mV in this context, or 169 times less than the read-out amplitude. The noise level defines the lower limit for the resolution and it fits well into the 7-8 bit range.

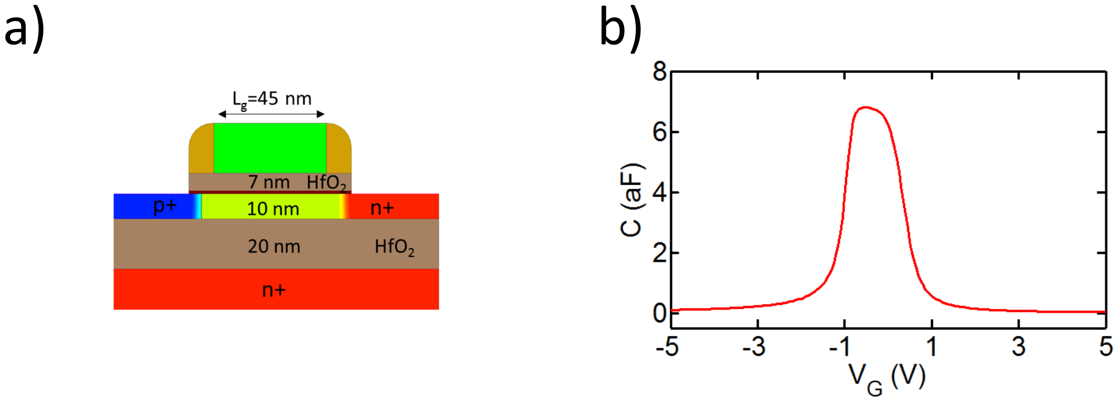


Figure 3.6: a) Simulated structure of the 45 nm device with HfO_2 as a high-k dielectric and corresponding CV curve in b). Material from [8]

In total, a capacitance per memory cell of 6.65 aF, like in 3.4b) seems sufficient to achieve good read-out performance for neuromorphic systems. Smaller capacitances may lead to problems, which means that the limit for scalability is at 90 nm gate length when using conventional dielectric layers (SiO_2). For scaling to 45 nm the SiO_2 needs to be replaced with a high-k dielectric, like HfO_2. This also includes the bottom dielectric layer, because read-out is also performed over this dielectric layer. The top high-k dielectric leads to smaller EOT and thus a reduced short channel effect (Fig. 3.4a). Fig. 3.6a) shows the simulated 45 nm device and the corresponding CV curve (Fig. 3.6b). The on/off ratio is 1:60 and slightly worse than the 90 nm device and the maximum capacitance is comparable to the 90 nm device, which ensures good detectability.

This result reveals that a scalability down to 45 nm is feasible. The further discussion in the next subchapters continues for the 90 nm device.

3.4 Theoretical Limitation of Energy Efficiency

Since the noise voltage defines the lowest possible resolution limit, and the energy consumed by the resistive or capacitive device during read-out is defined by the applied total voltage, one can make conclusions for the energy efficiency.

For resistive devices the thermal noise current is:

$$I_{n,res} = \sqrt{\frac{4k_B T \Delta f}{R}} \tag{3.4.1}$$

The total measured current ($I_{meas,res}$) consists of the noise ($I_{n,res}$) and read-out current ($I_{s,eff}$):

$$I_{meas,res} = I_{n,res} + I_{s,eff} \tag{3.4.2}$$

A measurement procedure is very similar to Nyquist–Shannon sampling [132]. A measurement and thus integration of the current from 3.4.2 leads to the following charge (Q_o):

$$Q_o = \int_t^{t+T_{meas}} I_{meas,res} dt = \int I_{meas,res} \cdot \mathrm{rect}\left(\tfrac{t}{T_{meas}}\right) dt = I_{meas,res} * \mathrm{rect}\left(\tfrac{t}{T_{meas}}\right) \tag{3.4.3}$$

As shown in the above equation the integration can be also conducted with a convolution with a rect pulse (Fig. 3.7a). The frequency spectrum of the above is obtained by Fourier transform:

$$\mathcal{F}(Q_o) = \mathcal{F}(I_{meas,res}) \cdot \mathrm{sinc}\left(\frac{t}{T_{meas}}\right) \tag{3.4.4}$$

Thus the frequency spectrum is a simple multiplication of the measured current spectrum with a sinc function, which is plotted in Fig. 3.7b). As can be seen, the spectrum of $I_{meas,res}$ is band limited by approximately the zero points of the sinc function:

$$\Delta f = \frac{1}{T_{meas}} \tag{3.4.5}$$

Inserting equation 3.4.5 into equation 3.4.1 gives:

$$I_{n,res} = \sqrt{\frac{4k_BT}{RT_{meas}}} \tag{3.4.6}$$

In order to achieve a certain number of levels (B bits), the maximum signal needs to have at least the following effective value:

$$I_{s,eff} = I_{n,res} \cdot 2^B \tag{3.4.7}$$

Using equation 3.4.7 and 3.4.6 one can conclude for the energy needed for a single MAC operation:

$$E = R \cdot I_{s,eff}^2 \cdot T_{meas} = 4k_BT \cdot 2^{2B} \tag{3.4.8}$$

This equation takes the resisitive power consumption $P = RI^2$ into account. The maximum energy efficiency η can be calculated by the inverse, and considering that one MAC operation consists out of two operations: one multiplication and one addition. It is 1842 TOPS/W for 8 bit and 29 472 TOPS/W for 6 bit.

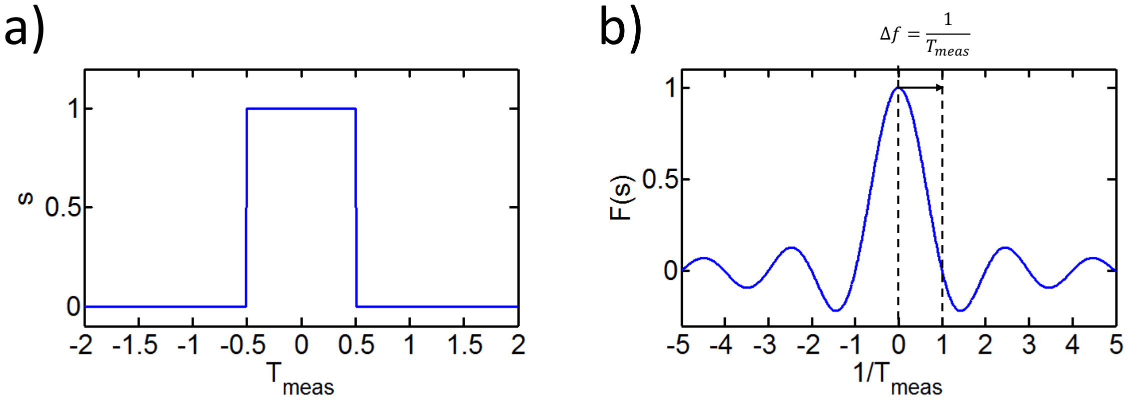


Figure 3.7: a) Rect pulse for the measurement current integration b) Fourier-transform of the rect pulse in a) with the indicated band limitation.

Many resistive devices have shot-noise, like tunnel junctions and MOSFETs in subthreshold regime, which leads to the following noise current:

$$I_{n,shot} = \sqrt{2q \cdot I_{s,eff} \cdot \Delta f} \tag{3.4.9}$$

This leads with a similar calculation as for the thermal noise to the following energy per MAC:

$$E = 2q \cdot U_{s,eff} \cdot 2^{2B} \tag{3.4.10}$$

Where $U_{s,eff}$ is the operating voltage and assuming $U_{s,eff} = 0.35\,\mathrm{V}$ one obtains 269.4 TOPS/W for 8 bit or 4310 TOPS/W for 6 bit. So the energy efficiency is significantly lower with shot noise compared to thermal noise.

On the other hand, for capacitive devices with the kTC noise (equation 3.3.2) one obtains for the necessary signal voltage:

$$V_{s,eff} = 2^B \cdot \sqrt{\frac{k_B T}{C}} \tag{3.4.11}$$

and for the energy per MAC:

$$E = \frac{1}{2} C V_{s,eff}^2 = \frac{1}{2} \cdot k_B T \cdot 2^{2B} \tag{3.4.12}$$

So for capacitive devices the value is 8 times lower compared to resistive devices with thermal noise (equation 3.4.8). Capacitive devices are more immune to noise, due to the filtering effect of capacitors. It results in 14 736 TOPS/W for 8 bit and 235 774 TOPS/W for 6 bit. Most of the energy stored on the capacitor can be recovered as explained in the next chapter, which in principle results in even higher energy efficiencies.

3.5 Spice Simulation and Estimation of Energy Efficiency

In order to simulate a full crossbar array, a Spice model of the proposed device was created and simulated in LTspice. The model is shown in Fig. 3.8a) and also includes parasitic capacitances as well as leakage and supply resistors. For the gate leakage and buried oxide leakage resistors a surface current of $1 \times 10^{-8}\,\mathrm{A/cm^2}$ at a voltage of 2 V was assumed [133]. The model also includes parasitic capacitances of the BL ($C_{box,par}$ and $C_{depl,par}$). The non-linear capacitances C_{n+} and C_{p+} contain models for the inversion and accumulation capacitances, which are derived from equation 3.1.3. As can be seen from Fig. 3.8b) the gate capacitance and read-out electrode capacitance fits well with the TCAD simulation. Fig. 3.8c) shows the coupling capacitance between gate and read-out electrode in comparison to Synopsys.
Each of these single device models was arranged in a crossbar vector along the SL/BL, where further parasitic resistors (silicide lines R_{Ti2Si}) and capacitors ($C_{box,par}$) are added (Fig. 3.9a). The parasitic elements of the wordline are negligible compared to the shielding line and bitline, due to the much lower resistivity of copper interconnects

and the dominating gate oxide capacitance in the time delay and energy consumption. The critical path regarding time delay and resistive losses is along the silicide lines and the highly n-doped BL. The thickness of the silicide line was set to 28 nm and the resistivity to 14 $\mu\Omega$cm [134]. The bitline resistivity was assumed to be 100 $\mu\Omega$cm, corresponding to highly n-doped silicon, and the thickness to be 90 nm.

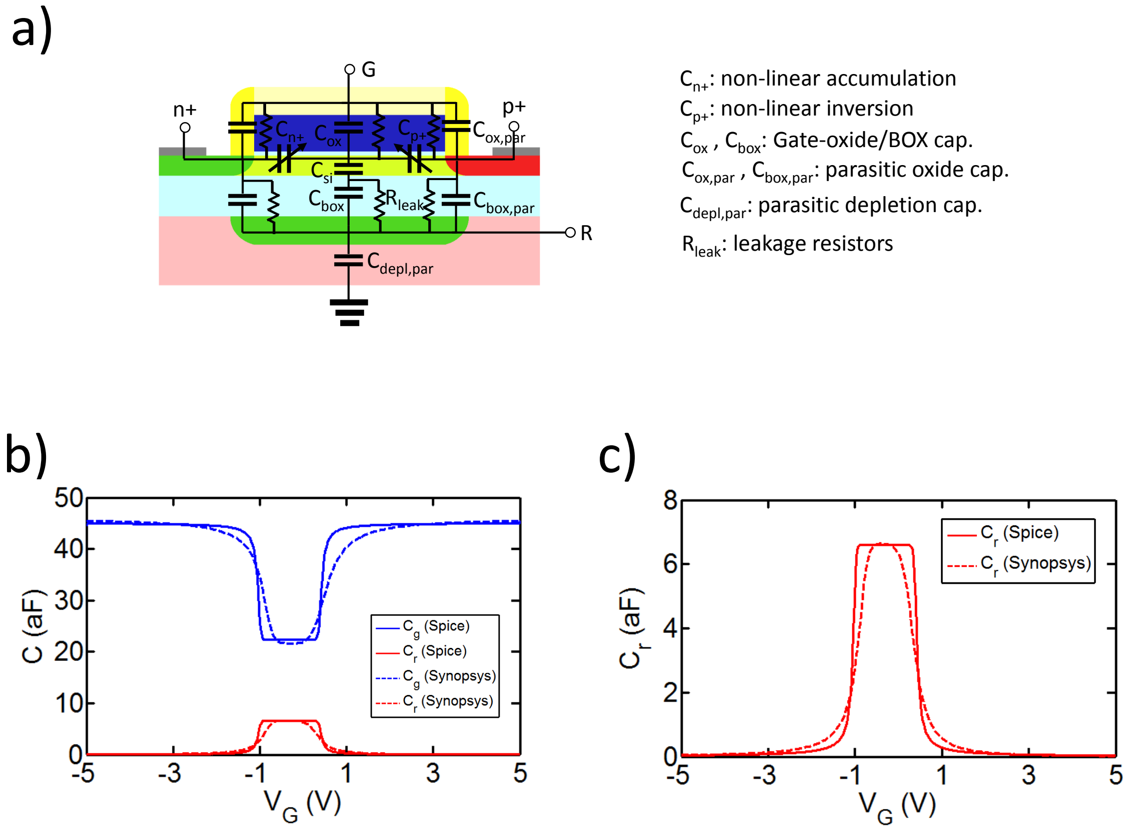


Figure 3.8: a) Spice circuit model of single device with parasitic capacitors and resistors. b) Gate oxide capacitance and coupling capacitance of spice simulation in comparison to Synopsys simulation. c) Coupling capacitance between gate and read-out electrode of spice simulation in comparison to Synopsys simulation. Material from [8]

For now, simulations in extreme cases are shown, which means that all WL are activated and all memory cells are in an erased or written state. These extreme cases will give an upper limit for time delay and a lower limit for energy efficiency for reading or inference. The energetically worst case scenario is that all WLs are activated at once and all weights are zero with a resulting shielding effect, which in turn would lead to charging in the top gate oxide. Fig. 3.9b) shows the accumulated charge on the read-out electrode over a half-period of 15 ns on the BL for the case when all cells are in a written state. The number of cells along the BL was varied and for the very long BL (2500 cells) the accumulated charge is lower, due to the fact that the resistivity of the n+-well BL is too high to supply sufficient charge.

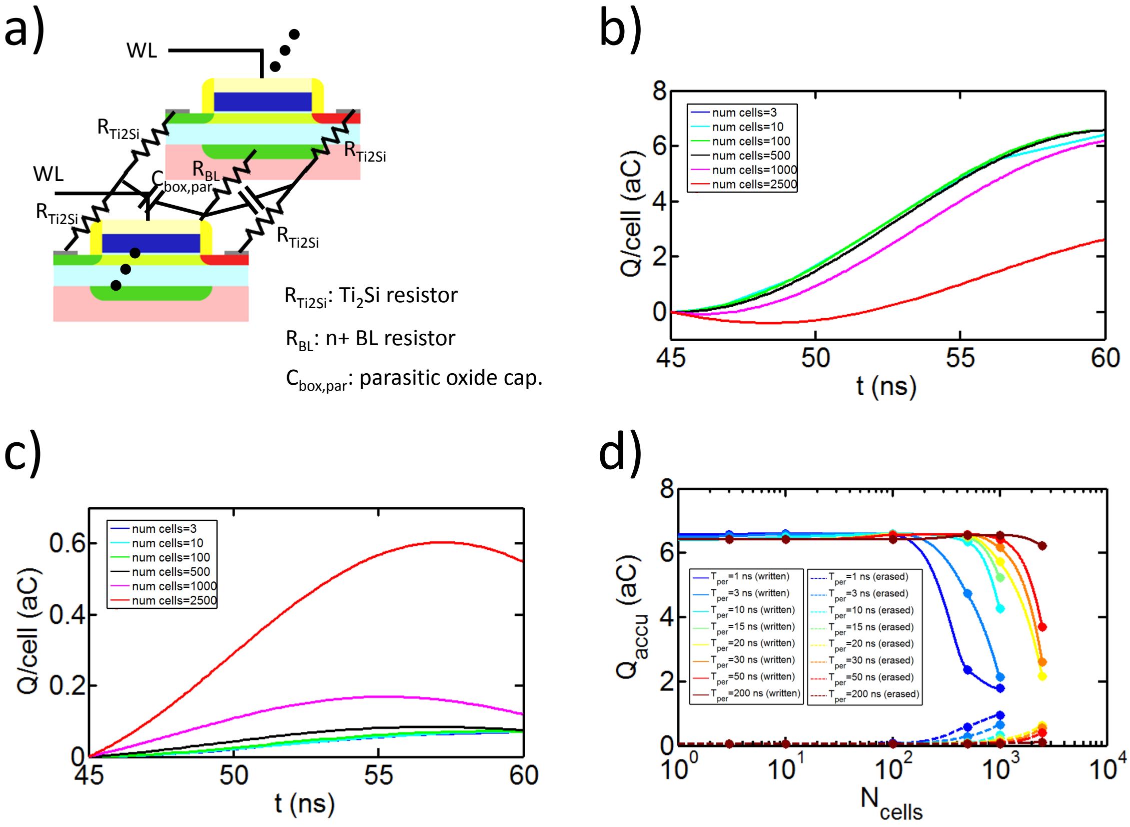


Figure 3.9: a) Arrangement of spice model in an array along the BL/SL with further parasitic elements. b) Accumulated charge on the BL over a half-period of 15 ns for different array sizes. A written state is programmed into the memory cells in this case. c) Accumulated charge on the BL over a half-period of 15 ns for different array sizes. An erased state is programmed into the memory cells in this case. d) Comparison of written and erased memory cells with regards to the total accumulated charge on the BL. Material from [8]

The charge gets lost in the parasitic capacitances. For very long BL the time delay is thus longer than 15 ns. Similarly, for the case when all memory cells are in an erased state (Fig. 3.9c), there is a significant overshoot of the accumulated charge on the BL, which is the result of not sufficient shielding charge supplied by the silicide lines. Fig. 3.9d) shows the total accumulated charge over a half-period for different period length T_{per} and number of memory cells. The on/off ratio between the erased and written state degrades for more memory cells (longer BL) and shorter T_{per}. From this graph one can deduce that a minimum T_{per} is necessary to achieve sufficient supply of charges. For 1000 cells in a row a period time of $T_{per} = 30$ ns is necessary, for 2500 cells the necessary time increases to $T_{per} = 200$ ns and for smaller arrays the time can be smaller. From the time delay (including the number of periods: $N_{per} = 142$) the areal efficiency A_η can be calculated with an area of $8F^2$ for each memory cell (one weight consists of two memory cells and one MAC out of two operations – multiplication and addition):

$$A_\eta = \frac{2}{2 \cdot 8F^2 \cdot T_{per} \cdot N_{per}} \tag{3.5.1}$$

The resulting areal efficiency is summarized in Tab. 3 for different array sizes.
The Spice simulation also gives results on the energy consumption of the crossbar during read-out (inference), and since the circuit consists of capacitors and resistors, one needs to distinguish between complex, active and reactive power or energy. The amount of active energy W_p per period can be calculated by the applied voltage to the WL and the total current flow as follows:

$$W_p = \int_t^{t+T_{per}} i_{tot}(t) \cdot v_{WL}(t) dt \tag{3.5.2}$$

The active energy per memory cell is shown in Fig. 3.10a) and increases for more memory cells and shorter periodic time T_{per}. This increase is in direct correlation to the decreased on/off ratio (Fig. 3.9d). The amount of complex energy W_s is calculated from the effective currents and voltages:

$$W_s = \sqrt{\int_t^{t+T_{per}} i_{tot}(t)^2 dt \cdot \int_t^{t+T_{per}} v_{WL}(t)^2 dt} \tag{3.5.3}$$

The reactive energy W_r is finally obtained by the active and amount of complex energy:

$$W_r = \sqrt{W_s^2 - W_p^2} \tag{3.5.4}$$

Fig. 3.10b) shows the reactive energy per memory cell for different array sizes and periodic times, and is approximately independent of these parameters. There is a difference between the written and erased state, since in an erased state the larger gate capacitance needs to be charged. The amount of reactive energy can also be estimated from the gate capacitance (Fig. 3.8b) directly. The current for an applied effective voltage of $U_{ac} = 0.5\,\mathrm{V}/\sqrt(2) = 0.35\,\mathrm{V}$ and gate capacitance of $C_g = 44.8\,\mathrm{aF}$ (Fig. 3.8) is calculated as follows:

$$I_{ac} = U_{ac} \cdot 2\pi f \cdot C_g \tag{3.5.5}$$

From this current one can deduce the reactive energy per memory cell:

$$W_r = I_{ac} U_{ac} \cdot T_{per} = U_{ac}^2 \cdot 2\pi \cdot C_g = 35.19\,\mathrm{aJ} \tag{3.5.6}$$

This value is quite close to the value obtained by the simulation (3.10b). For the energy per MAC operation the number of periods N_{per} needs to be included and also the fact that each weight composes out of two memory cells (differential weight):

$$E_{MAC} = \frac{2W_s \cdot N_{per}}{N_{cells}} \tag{3.5.7}$$

The energy efficiency η in TOPS/W can be calculated by considering that each MAC operation consists of two operations (one multiplication and one addition):

$$\eta = \frac{2}{E_{MAC}} \cdot 10^{-12} OP/TOP \tag{3.5.8}$$

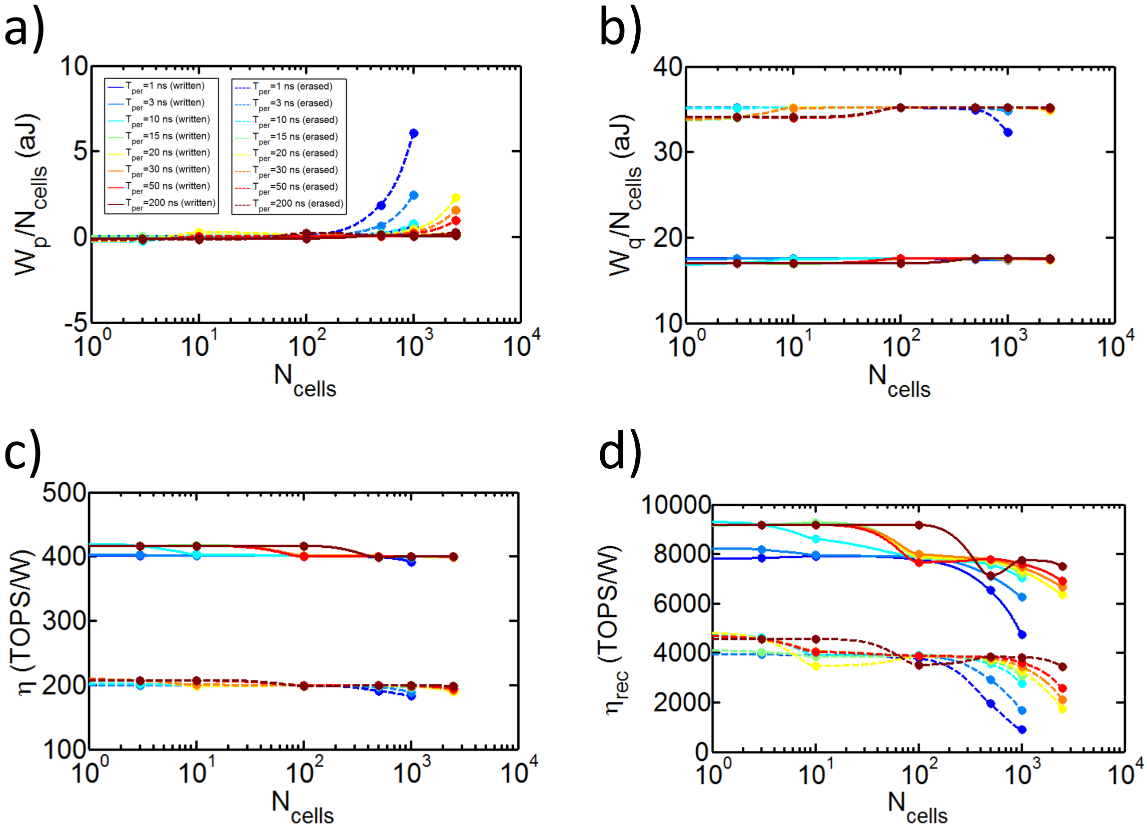


Figure 3.10: a) Active energy per memory cell W_a/N_{cells} and b) reactive energy per memory cell W_q/N_{cells} for different array length and T_{per}. c) Energy efficiency without recovery and d) energy efficiency with energy recovery, assuming $q = 20$. (The legend from a) is valid for all plots) Material from [8]

The energy efficiency η is plotted in Fig. 3.10c) and is 198.5 TOPS/W for the erased state and 400.3 TOPS/W for the written state. The efficiency stays relatively constant since most of the complex energy is reactive and it is relatively independent on the parameters.

In chapter 2.4 the concept of reversible computing [96, 97] and adiabatic switching [103–110] was already introduced. Since capacitances are also charged and discharged in this concept, a lower energy consumption can in principle be achieved by using charge recovery techniques. The adiabatic charging/discharging would reduce the reactive power component of the matrix. As already noted in previous publications, a memcapacitor is, if ideal conditions are supposed, a powerless device [24].

Regarding scalability, the proposed device does not have a subthreshold leakage power, so even for scaled versions of the device (when using high-k dielectrics) the adiabatic charging has an advantage, contrary to CMOS realizations of adiabetic circuits. Furthermore, the speed is not as that important compared to digital electronics, making the adiabatic charging for neuromorphic applications more applicable.

However, it is necessary to consider that the AC power source, which usually uses inductances as an energy storage, cannot recover the energy completely. The inductances have limited quality factors (q factor) in the order of some dozens to hundreds. In common adiabatic realizations the energy recovery of supply clock generators is of the order of 95% for harmonic signals [107, 109, 111], which means the active power is $q = 20$ times lower than the reactive power.

Using this number for the quality factor we can conclude for the energy per MAC (again two memory cells per weight):

$$E_{MAC} = 2 \cdot \frac{N_{per}}{N_{cells}} \cdot \left(\frac{W_r}{q} + W_p \right) \tag{3.5.9}$$

The energy recovery of the AC power source depends on its quality factor q and dictates how much of the reactive power needs to be supplied in order to maintain the oscillation. The amount of active energy in the oscillator is the reactive energy divided by the quality factor (equation 3.5.9) and furthermore, the active part of the crossbar array W_p needs to be added (equation 3.5.9). Fig. 3.10d) shows the energy efficiency with energy recovery (η_{rec}), assuming a quality factor of $q = 20$). For shorter period length and longer bitlines the η_{rec} decreases, due to larger resistive losses, which cannot be recovered. For 1000 cells and $T_{per} = 30\,\text{ns}$ a total energy efficiency of 3452.6 TOPS/W for the erased state and 7468.1 TOPS/W for the written state can be achieved with energy recovery. Tab. 3 also includes the energy efficiencies for different array sizes in the erased state, which is the worst case scenario.

The simulated and calculated energy efficiency is amongst the highest reported so far. For resistive systems the highest reported energy efficiencies for 6-8 Bit resolution is around 100 TOPS/W [68, 85].

So far only the read-out energy, which is relevant for inference tasks, has been investigated. For neural network training read-out and writing is important, thus also the write energy needs to be considered for non-inference tasks. The write energy depends on the underlying memory principle and is very low for ferroelectric memories (fJ regime) and close to biological values. On the other hand, charge trap memories have a much higher write energy due to the voltage-time dilemma.

Array size	Period T_{per} [tot. delay ($N_{per} = 142$)]	A_η ($TOPS/mm^2$)	$W_r(fJ/cell)$* [$W_p(fJ/cell)$]*	$\eta_{rec}(TOPS/W)$* [$\eta(TOPS/W)$]*
100x100	1 ns [142 ns]	108.7 TOPS/mm^2	5 fJ/cell [0.015 fJ/cell]	3782.2 TOPS/W [199.51 TOPS/W]
500x500	15 ns [2.13 μs]	7.25 TOPS/mm^2	5 fJ/cell [0.022 fJ/cell]	3676.8 TOPS/W [199.19 TOPS/W]]
1000x1000	30 ns [4.25 μs]	3.62 TOPS/mm^2	5 fJ/cell [0.04 fJ/cell]	3452.6 TOPS/W [198.54 TOPS/W]
2500x2500	200 ns [28.4 μs]	0.54 TOPS/mm^2	5 fJ/cell [0.039 fJ/cell]	3461.7 TOPS/W [198.59 TOPS/W]

Table 3: Summary of time delay, areal efficiency and active energy/reactive energy per cell. Note that these values include $N_{per} = 142$. From the active and reactive energy the energy efficiency (η_{rec}; η) is deduced.

**all cells are erased (worst case scenario), 95% energy efficiency of power clock source* ($q = 20$)

3.6 MNIST Simulation

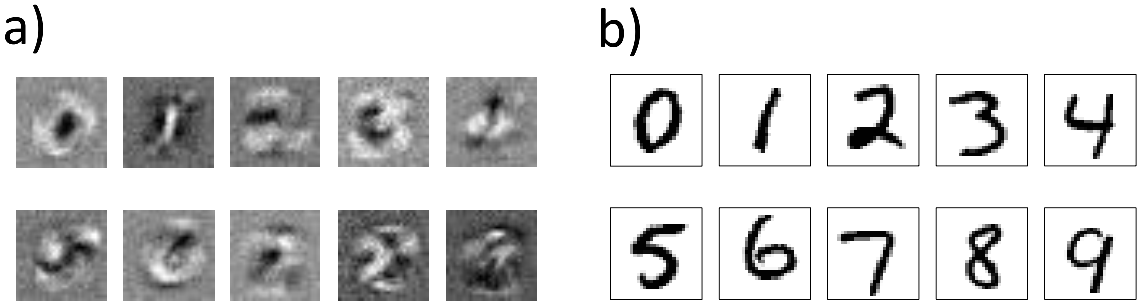


Figure 3.11: a) Map of implemented weights and b) applied images of handwritten numbers. Material from [8]

The energy efficiency in the last subchapter was simulated for the worst case scenario, when all WL are activated and all memory cells were in the erased or written state. This is not a realistic scenario, since e.g. in image recognition tasks the input feature map and the weight values contain patterns. Thus sparsity has to be included. For this reason a simple one-layer perceptron of a Modified National Institute of Standards and Technology (MNIST) database was simulated. The trained weights (Fig. 3.11a) were implemented on the arrays and the pixels of the handwritten numbers 0-9 (Fig. 3.11b) were applied to the WL. The obtained output activation for the numbers 0,2 and 4 is comparable to the directly calculated ones (Fig. 3.12a-c), showing sufficient precision of the device. In this case an average energy efficiency of 29 600 TOPS/W is achieved with charge recovery. Without recovery, the efficiency amounts to 1702 TOPS/W for MNIST.

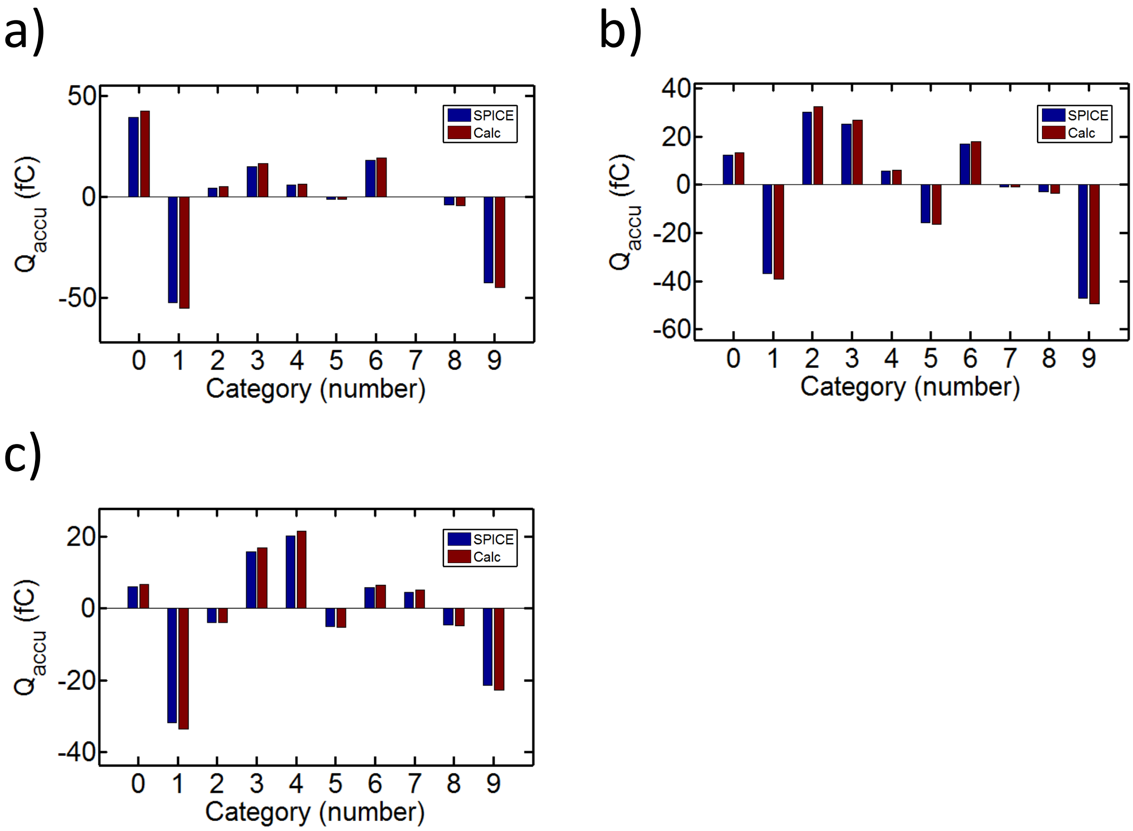


Figure 3.12: Obtained output activation compared to the calculated activation for a) the number 0, b) the number 2 and c) the number 4. Material from [8]

3.7 Oxide Devices

Growing the device stack (Metal-Dielectric-Semiconductor-Dielectric-Metal structure) fully out of oxides has several advantages: There is no interface oxide between semiconductor and ferroelectric material, when using ferroelectric memories, thus the endurance and ferroelectric behaviour might be better. Moreover, the devices can be grown at lower temperature, thus enabling back-end-of-line integration. Another advantage is the easier 3D integration, since several crossbar structures can be grown on top of each other.

As an oxide semiconductor material several oxides are possible, e.g. $SrTiO_3$,TiO_2 or Indium-Gallium-Zinc-Oxide (IGZO).

One disadvantage is that in most oxide semiconductors there is no possibility to achieve p- and n-doping, thus the device must be implemented without lateral pin-junction. This has several implications, amongst others, there is no possibility to modulate the shielding layer with a voltage V_{AK}. The device might behave more non-symmetric since only one carrier type (usually electrons) can be injected for shielding. Especially during writing with a negative voltage (assumption n-doped semiconductor) there are only depletion charges in the shielding layer (Fig. 3.13a). Once the shielding layer is completely depleted no further charge can be supplied (Fig. 3.13b). This problem can solved by using back-gating with a positive voltage from the read-out electrode (Fig. 3.13c).

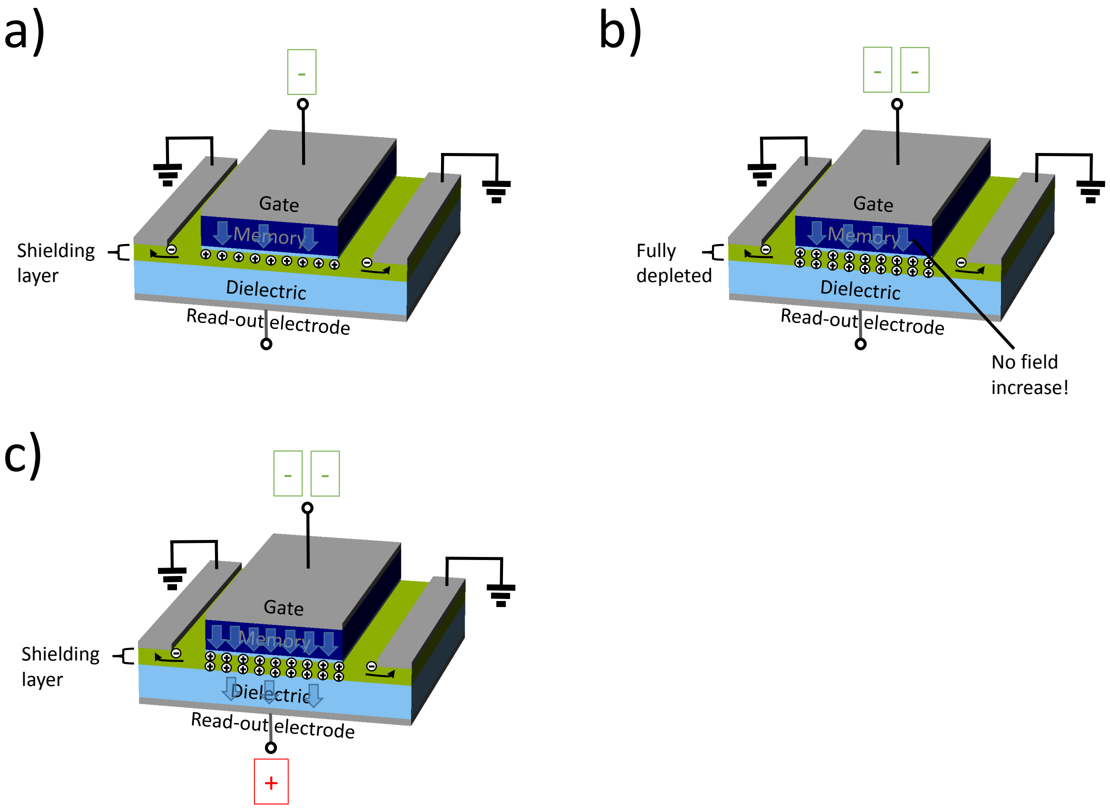


Figure 3.13: a) Depletion of n-doped semiconductor during application of a negative gate voltage. b) Fully depleted semiconductor layer, thus no further switching field increase. c) Use of positive back-gate voltage from read-out electrode during writing.

4 Fabrication and Process Development

The fabrication of the capacitive device described in the last chapter can be executed by using SOI wafers [135–140]. These wafers have a single crystalline silicon layer (device layer) on top of an insulating silicon oxide layer (buried oxide). Throughout the next subchapters the steps for the process development are introduced until the fabrication of a full crossbar arrangement of the capacitive devices is explained.

During the first stage of the investigation multiple problems were encountered with leakage currents and pin holes in the buried oxide, which were mainly solved by using a different doping method and etching of islands on the device layer of the SOI wafer. The improvement will be explained in chapter 4.1. These first devices had a normal dielectric layer instead of a memory layer and were thus used to investigate fundamental capacitive coupling properties.

Thereupon experiments were conducted on the deposition and improvement of the memory stack on normal silicon wafers by using atomic layer deposition (ALD) of hafnium zirconium oxide (chapter 4.2). The main challenge was the improvement of interface oxide to achieve either a good charge trapping memory or a ferroelectric memory with a minimum amount of charge trapping.

The upcoming chapter 4.3 includes the improvement of the crossbar fabrication. These include a trench refilling with SU-8 resist and second metallisation layer compared to the single devices of the prior chapters. Trenches were etched in order to seperate the read-out electrodes/bitlines of the crossbar.

The subchapter 4.3.5 summarizes the full fabrication procedure for crossbars with its corresponding mask layers.

The last chapter 4.4 includes some fabrication trials for oxide devices with pulsed laser deposition (PLD).

4.1 First Work on SOI Wafer and Single Device Fabrication

The first fabrication trials on SOI wafer were conducted with a solid diffusion source [141, 142] for doping because there was no access to ion implantation and it is an easy method to implement. For p-doping a boron oxide slice was put in close proximity to the wafer in a furnace process. The boron oxide is deposited during heating on the wafer and the boron oxide is reduced with the following reaction:

$$2\,\mathrm{B_2O_3} + 3\,\mathrm{Si} \longrightarrow \mathrm{B} + \mathrm{SiO_2}$$

Similarly the phosphor doping was achieved by using a spin-on-dopant [143] from Honeywell, which contained phosphorous oxide. The doping was done through a hardmask

(CVD silicon dioxide).

The biggest issue was the large leakage current through the buried oxide layer, which was mainly attributed to the doping method used. The etching of silicon islands on the device layer prior to doping tests improved device yield. Generally, pin-holes in the buried oxide layer are a common cause of failure in SOI based devices, and the probability of a pin-hole within a device can be reduced by the silicon surface reduction per device, which is achieved by island etching. Secondly, it was investigated if HF etching performed subsequent to silicon island etching had any influence on leakage current characteristics. The outcome of these investigations was that HF etching significantly degrades the leakage properties, most likely as a result of defects caused by underetching of the buried oxide layer below the silicon island (Fig. 4.1). Long HF etching times were used after the boron and phosphorous doping with the solid source diffusion method in order to remove rests from the source material on the wafer. It would appear to be a better option to perform the doping and removal of source material before the island etching, which in turn should be done at the end of the process.

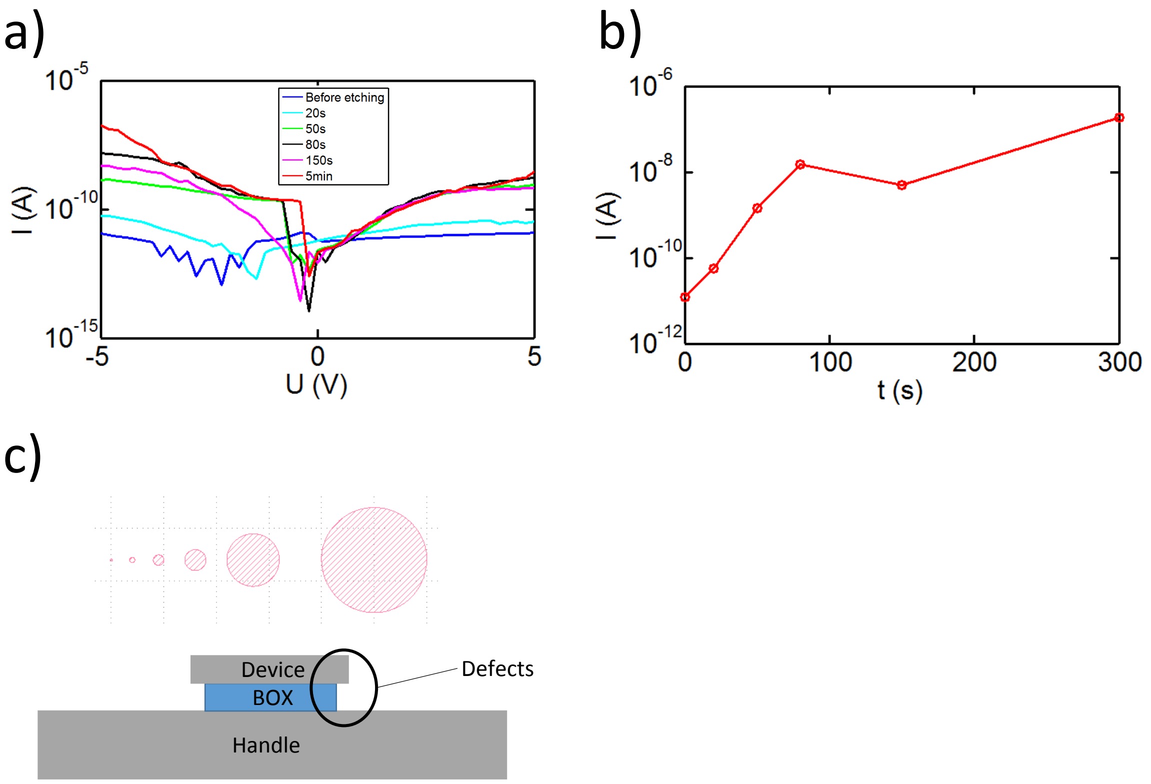


Figure 4.1: a) IV curves between device and handle wafer for different HF etching times b) Maximum leakage current through buried oxide layer versus HF etching time c) Illustration of the under-etching with HF of the buried oxide layer.

In spite of this procedure, leakage properties of the fabricated devices were still unsatisfactory. One of the main causes was that during the diffusion process of the dopants oxygen is introduced to ease the removal the solid source rests on the wafer. This leads to significant oxidation of device silicon and subsequent removal of the same during HF etching. As a consequence, tests without oxygen introduction were tried in order to reduce silicon consumption during diffusion.

For the purpose of testing the pin diode fabrication on the SOI wafer, a stepwise approach was followed (Fig. 4.2a). In the first step, the islands were etched and tested after p-doping with the first piece. In the next step, the second piece was tested with p- and n-doping. In the final step, the third piece was tested with the full fabrication scheme, including gate-oxidation. Each piece was tested regarding buried oxide leakage current. After the p-doping, 44.6 % had satisfying leakage properties and the undoped island had a yield of 69.4 %. After the p-doping, already holes became visible of approximately 1 μm in size, which are 100-200 nm deep (Fig. 4.2b)). Most likely the silicon was reduced totally on these parts during the reaction with the boron oxide. After the n-doping, the p-doped regions reacted with the chemical vapor deposition (CVD) oxide and could not be removed, even with long HF etching times. Furthermore, underetching of the prior mentioned holes became visible (Fig 4.2d). In the end, the yield of the pin junction islands was only 11.2 % and not sufficient to proceed (Fig. 4.2e). The reaction between the dopand and the long HF etching times to remove the dopand source made this method unsuitable for SOI wafers.

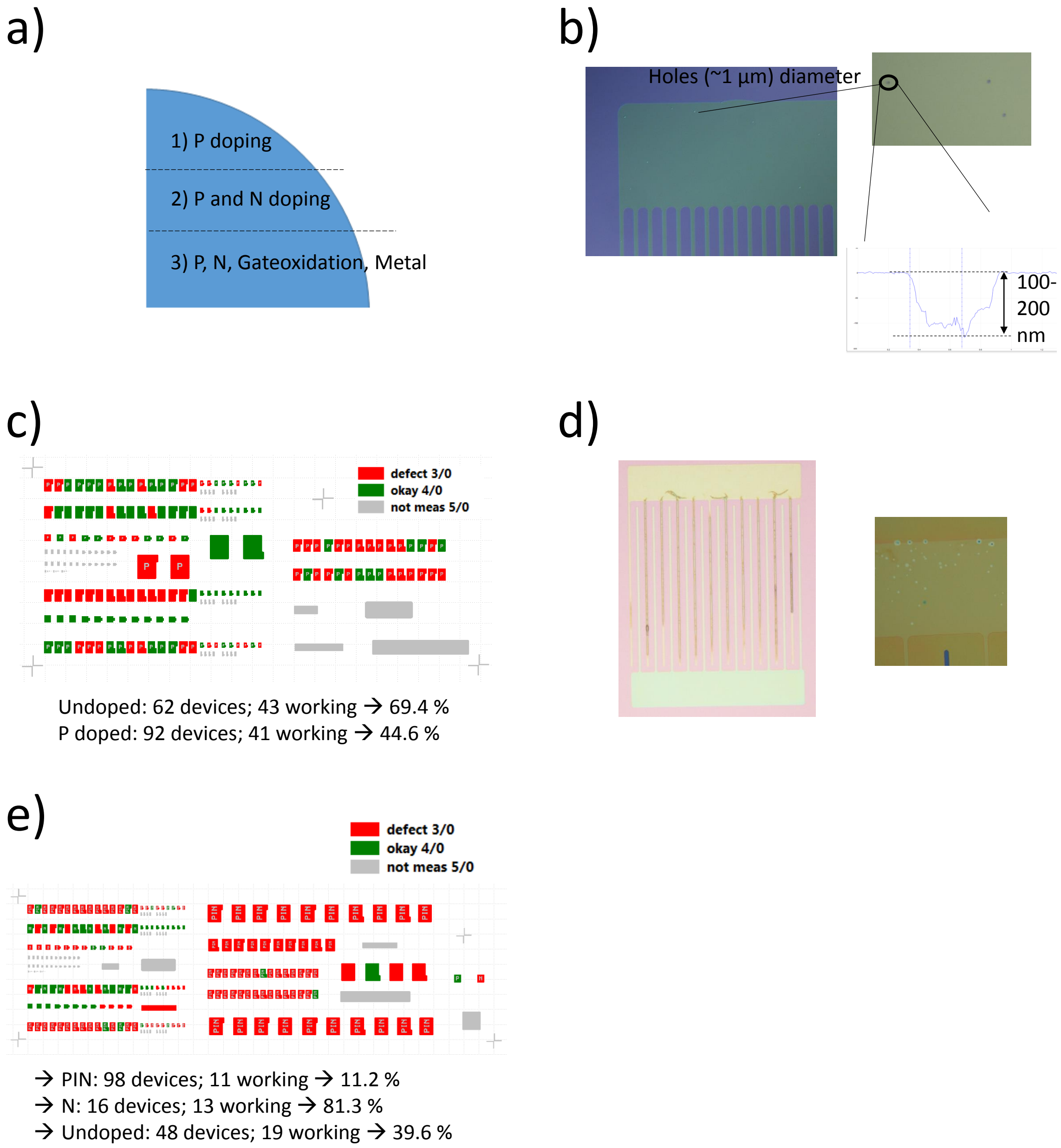

a)
1) P doping
2) P and N doping
3) P, N, Gateoxidation, Metal
b)
Holes (~1 µm) diameter
100-200 nm
c)
defect 3/0
okay 4/0
not meas 5/0
Undoped: 62 devices; 43 working → 69.4 %
P doped: 92 devices; 41 working → 44.6 %
d)
e)
defect 3/0
okay 4/0
not meas 5/0
→ PIN: 98 devices; 11 working → 11.2 %
→ N: 16 devices; 13 working → 81.3 %
→ Undoped: 48 devices; 19 working → 39.6 %

Figure 4.2: a) Stepwise processing of SOI Wafer piece. b) Holes after p-doping. c) Island yield after p-doping (first piece). d) Rests on p-doped regions after n-doping and underetching of holes. e) Island yield after p- and n- doping.

Ion implantation is the state-of-the-art doping method [10, 144, 145] and applied for fabricating the capacitive devices after tests with solid source diffusion. Due to the thin device layer, the energy of ion implantation was chosen low enough in order to protect the buried oxide. The dose was chosen in such a way that high doping concentrations could be achieved, whilst amorphisation of the device layer is kept at an acceptable level [146, 147]. As can be seen from Fig. 4.3, the parameters for the BF2 implantation were: E=45 keV (equivalent to 10 keV B) and D=$2 \times 10^{14}\ 1/\mathrm{cm}^2$. For the phosphorous implantation the parameters were: E=25 keV and D=$2 \times 10^{14}\ 1/\mathrm{cm}^2$.

Hence the maximum doping concentration is around $2 \times 10^{19}\,1/\mathrm{cm}^3$ [10] and the amorphous layer thickness is 50 nm [146].

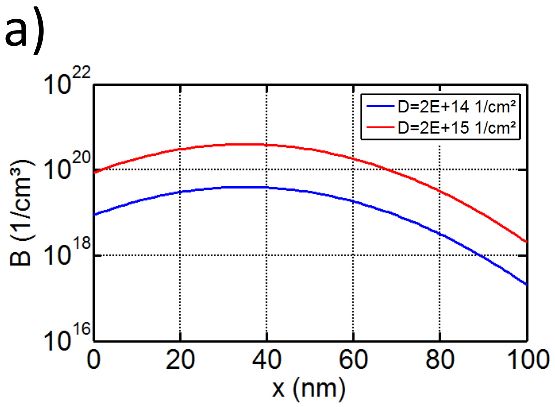


Figure 4.3: a) Boron concentration with 10 keV (equivalent to 45 keV BF2) and two different Doses [10]. b) Phosphorous concentration with 25 keV and two different doses [10].

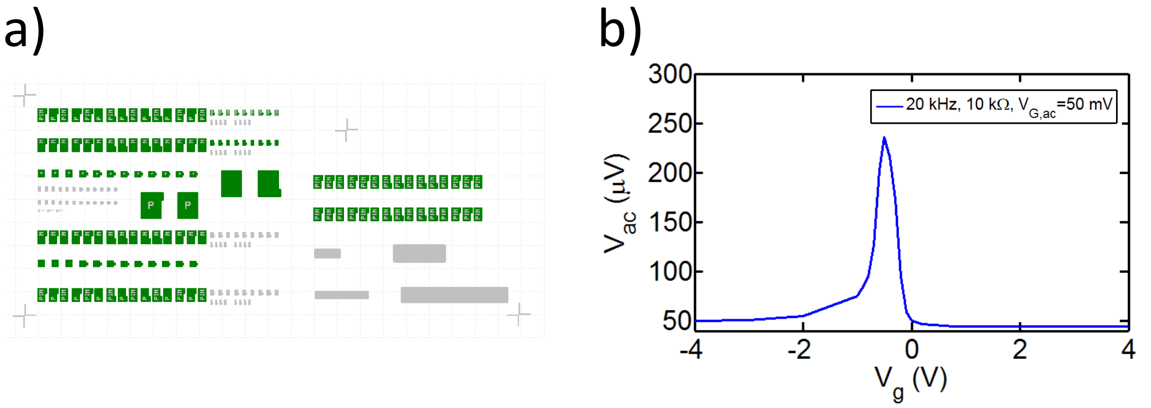


Figure 4.4: a) Island yield after pin diode fabrication. b) Measured CV coupling with lock-amplification.

Nearly 100 % yield was achieved with ion implantation and for the first time the capacitive coupling window became visible (Fig. 4.4). In later fabrication schemes a higher implantation dose was used ($2 \times 10^{15}\,1/\mathrm{cm}^2$) in order to achieve better electron and hole injection in the pin-Diode. In this case the thickness of the remaining crystalline layer 10 nm-20 nm was sufficiently large to enable recrystallisation during activation. Furthermore a metallic supply line on the n+ and p+ fingers was used to ensure sufficient charge shielding. The Gate electrode was winded around n+ and p+ fingers to increase the capacitance of the device (see masklayout in Fig. 4.5).

The process flow for single device fabrication is shown in Fig. 4.5. Depending on, if the ferroelectric memory material is integrated or devices with normal silicon dioxide gate oxide are fabricated the process flow divides into two path. The different mask layers for each processing step are also shown in Appendix A.1.

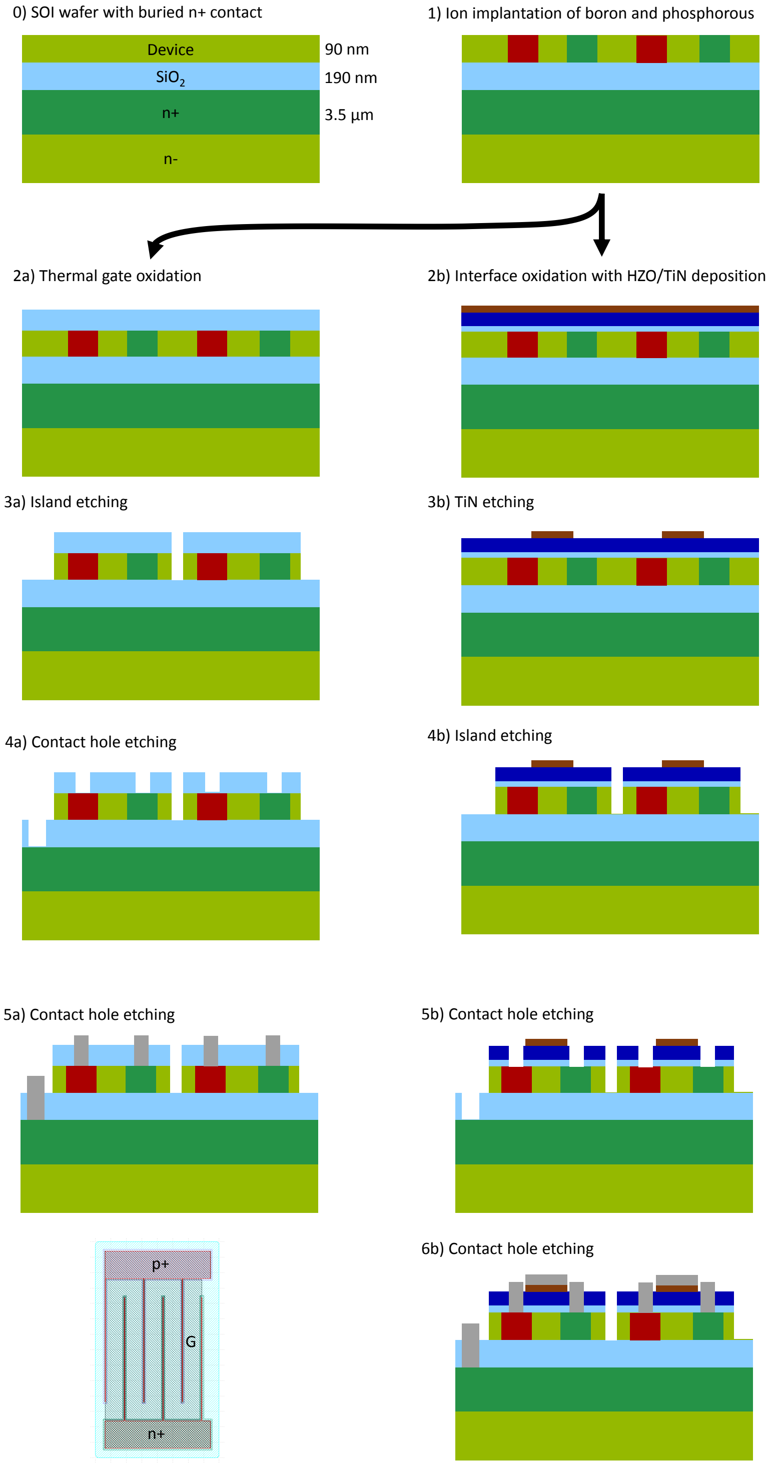


Figure 4.5: Fabrication flow for single devices with the two branches a and b for the thermal oxide and the memory dielectric with HZO/TiN. In the bottom left the mask layout of a device with gate contact winded around n+ and p+ regions is shown.

4.2 Deposition Experiments of Hafnium Zirconium Oxide on Silicon

In order to optimize the properties of the memory material (ferroelectric HZO [148]) MFM and metal-ferroelectric-interfaceoxide-silicon (MFOS) capacitors were fabricated and measured. The ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ (HZO) film was deposited with ALD by Namlab gGmbH and was already an optimized process [87, 148]. TiN was used as a metal electrode as it promotes the orthorombic phase of the film [87] during the 500 °C N_2 anneal (20 s). The precursors HyALD and ZyALD were used during ALD deposition and ozon was used as an oxygen source. The deposition temperature was 280 °C. The HZO thickness was 15 nm. The TiN electrode was deposited by sputtering and had a thickness of 12 nm.

As can be seen from Fig. 4.6 sufficient hystersis loops were obtained with MFM capacitors and thus showing ferroelectric behaviour. The remanent polarization with $12\,\mu C/cm^2$ is consistent with literature values [148].

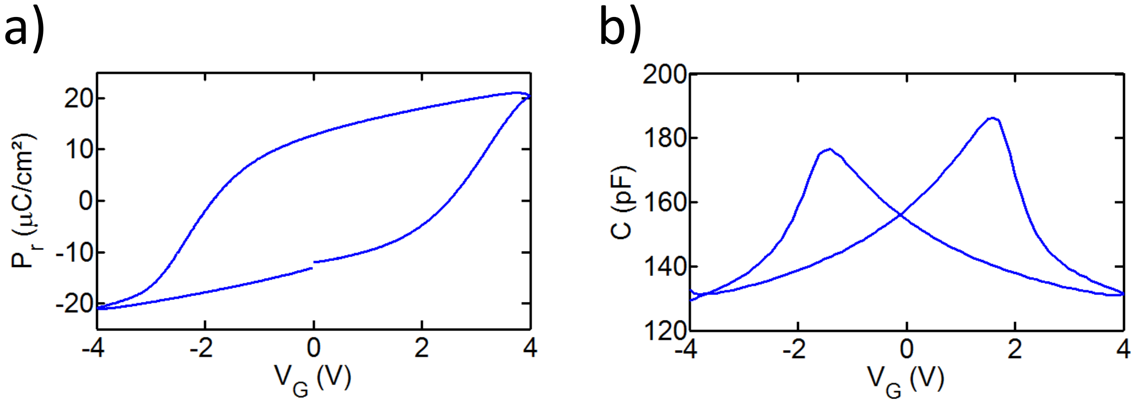


Figure 4.6: a) P-Vg Hysteresis loop of a MFM capacitor and b) corresponding CV curve.

The optimization of the interface oxide for the fabrication of MFOS capacitors is more complicated. Various interface oxide thicknesses have to be tested. Usally the interface oxide is grown by a mixture of ammoniumydroxid and hydrogenperoxide (also known as SC1 solution) and subsequently nitridated by an ammonia anneal [149]. The obtained SiON has a higher dielectric constant than normal SiO_2 and enables a lower electric field drop during ferroelectric switching on the interface oxide. This in turn improves the charge trapping characteristics of the obtained stack. Charge trapping leads to a CV curve shifting in the opposite direction compared to ferroelectric switching and thus both effects are contrary [150–152]. With the capabilities during this thesis there was no option for nitridation and thus pure SC1 or thermal SiO_2 has to be used.

Different SC1 concentrations at various temperatures and treatment times [153, 154] were first investigated by ellipsometry. After HF etching of the 1 inch test wafers, the DI water rolled off, but a thickness of 0.7 nm was measured. Fig. 4.7a) shows the thickness of a H_2O:NH_4OH:H_2O_2 (5:1:1) solution for different temperature. The equilibrium thickness stays around 1 nm, no matter how long the treatment was and at which temperature. This result reveals that the equilibrium is reached very fast. For a 40:2:1 solution the situation is not different, but the thickness is higher (Fig. 4.7b). Tests at various concentration levels (80:1:1, 340:1:1, 1280:1:1 and 10000:1:1) for 30 s show (Fig. 4.7c)) that a good control over the oxide thickness can be achieved for ultra high diluted SC1 (1280:1:1 and 10000:1:1). Therefore, Fig. 4.7d) shows the thickness for different times with 10000:1:1 solution.

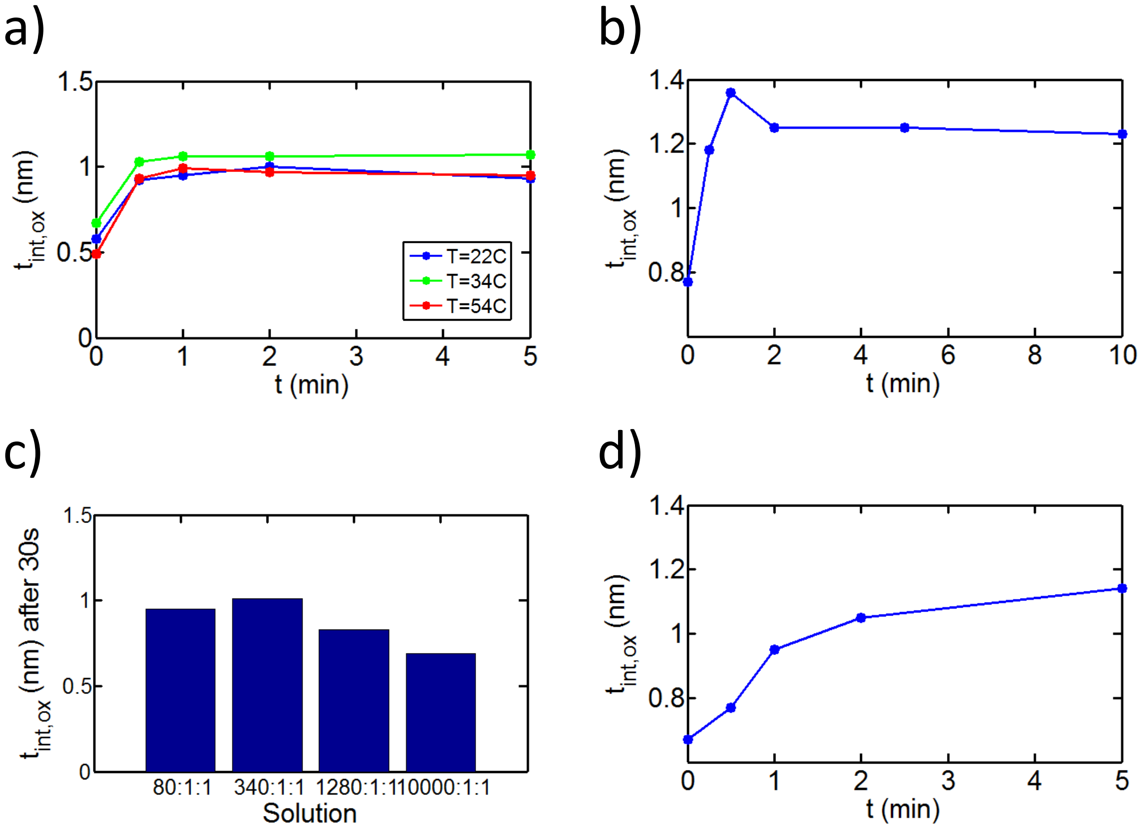


Figure 4.7: a) Ellipsometrically measured interface oxide thicknesses for a 5:1:1 SC1 solution at different temperatures. b) Ellipsometrically measured interface oxide thicknesses for a 40:2:1 SC1 solution at 36 °C. c) Interface oxide thickness for different diluted SC1 solutions at room temperature and 30 s. d) Interface oxide thickness for 10000:1:1 SC1 solution at room temperature.

MFOS capacitances were built for the solutions 5:1:1, 40:2:1 and 10000:1:1. Furthermore thermal oxides were tested: First a SC1 in 5:1:1 solution at 53 °C was applied, and the oxidation was done with 600 °C for 30 min, 700 °C for 30 min and 750 °C for 10 min, which results in the thicknesses 1.63 nm, 2.62 nm and 2.62 nm. Tests were also performed on HZO grown directly on HF etched silicon. It is expected that the interface oxide grows by the Ozon [155–157] during ALD and the thermal annealing step for crystallisation.

Fig. 4.8a) shows the CV curves for 2 min and 5 min SC1 (5:1:1) solution. The solid line indicates the CV sweep from −5 V to 5 V and the dashed line is the vice-versa CV sweep. The threshold voltage is determined by the tangent at the deflection point and its intersection with the x-axis. If the CV curve shifts to the left or right for the backward CV sweep a ferroelectric or charge trapping behaviour can be measured. For the shown CV curves in Fig. 4.8a), the plot for 2 min reveals a ferroelectric behaviour and for 5 min both effects are balanced. Measurement over many dots (10) shows a ferroelectric behaviour for 60 % of the devices (2 min) (Fig. 4.8b). For 5 min a ferroelectric yield of 80 % was achieved. Devices with HZO grown directly on HF etched silicon mostly show charge trapping behaviour.

For 40:2:1 solutions, the ferroelectric behaviour was most visible at 1 min, 2 min and 5 min (Fig. 4.8c-d). Similarly, for 10000:1:1 solutions the charge trapping behaviour was dominant at either very short or very long SC1 times (Fig. 4.8e-f). It was generally very difficult to obtain reproducible results with respect to ferroelectric behaviour and during a second forward and backward CV sweep the ferroelectric memory window disappeared (Fig. 4.9). One possible explanation is the large voltage stress during the rather slow CV sweeps. Experiments by other groups on FeFET have been done with fast pulses [150, 152]. Furthermore, variations in the trap density of the $Hf_{0.5}Zr_{0.5}O_2$ film could have an influence on the reproducibility.

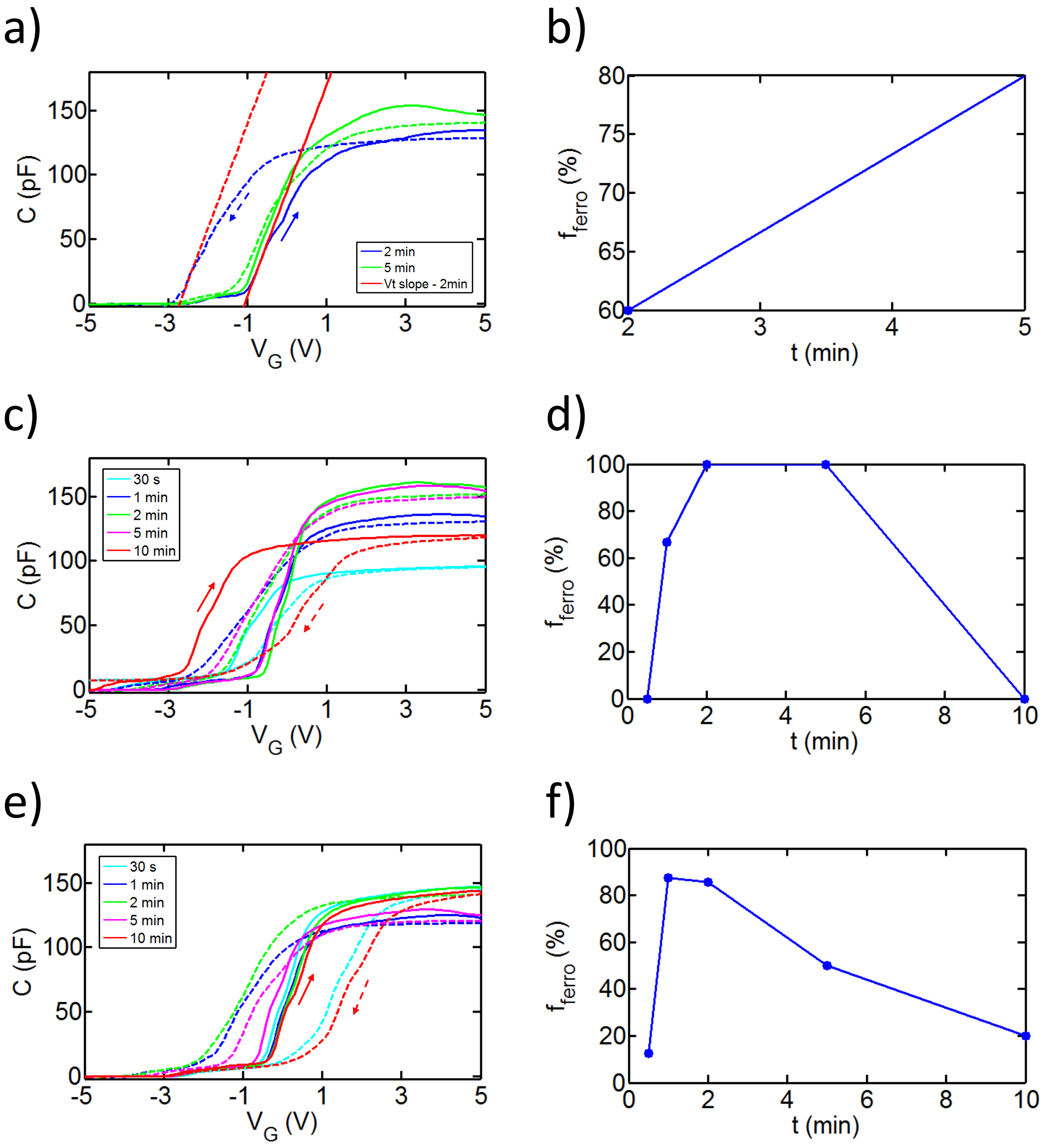


Figure 4.8: a) CV sweeps for 5:1:1 SC1 solution for different times and b) corresponding fraction of ferroelectric MFOS capacitors. c) CV sweeps for 40:2:1 SC1 solution and d) corresponding ferroelectric fraction. e) CV sweeps for 10000:1:1 SC1 solution and f) corresponding ferroelectric fraction.

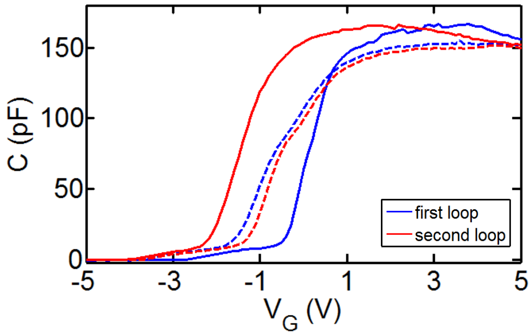


Figure 4.9: CV sweep of a ferroelectric memory window and the charge trapping memory window after a second CV sweep.

In most cases, a charge trapping behaviour was obtained for the thermally grown interface (Fig. 4.10). The 700 °C MFOS capacitors had a smaller memory window and in some cases a ferroelectric shifting. The reason for the varying behaviour of the 700 °C sample is unknown. However, the thermally grown interface layers are of interest in order to realize charge trapping memories. It is expected that the polarization charge of the ferroelectric layer assists the charge trapping behaviour [151]. It is noteworthy that a very stable memory window of ~2.8 V was obtained for the 750 °C sample.

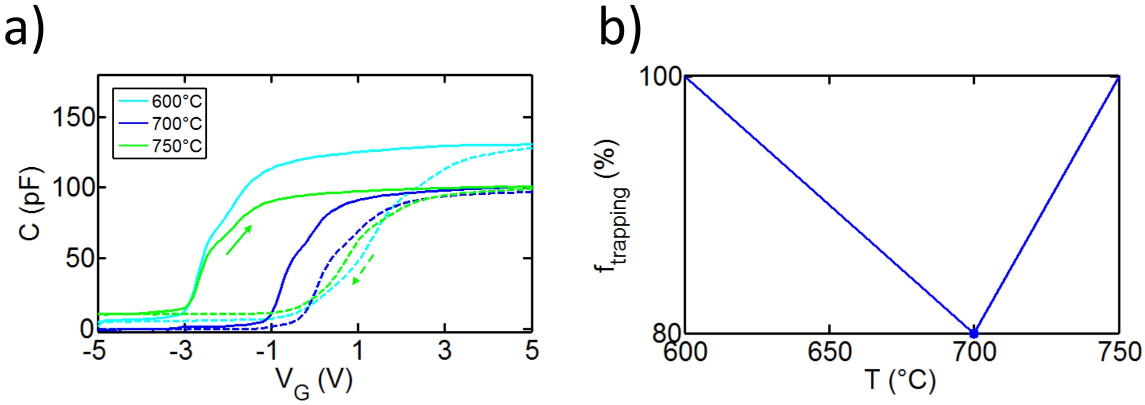


Figure 4.10: a) CV sweeps for different interface oxidation temperatures and b) corresponding fraction of charge trapping MFOS capacitors.

From Fig. 4.11a) the x-ray diffraction (XRD) pattern of the ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ film is shown and an orthorhombic phase is proven. Further, the TEM images of some samples are shown in Fig. 4.11b) and the interface oxide thicknesses are nearly consistent with Fig. 4.7.

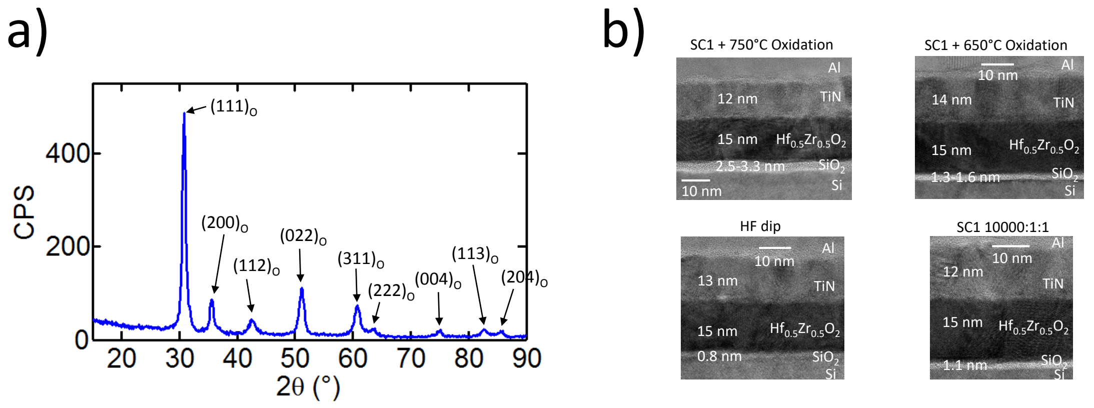


Figure 4.11: a) XRD of the HZO films, indicating the orthorhombic phase. b) TEM images of the Si / SiO_2 / $Hf_{0.5}Zr_{0.5}O_2$ / TiN / Alu stack for different interface oxide growth conditions.

In summary, different SC1 concentrations and times were tested and it seemed difficult to obtain ferroelectric behaviour in a reproducible manner. The 10000:1:1 solutions was best suited to obtain good control over interface oxide thicknesses. Thermally grown interface oxides seemed good to obtain a charge trapping memory. For the single and crossbar devices the following conditions were tested: SC1 5:1:1, SC1 10000:1:1, HF etched, 650 °C and 750 °C thermal oxide.

4.3 Process Development for Crossbar Fabrication

In comparison to chapter 4.1, the fabrication of crossbar devices requires the separation of the bottom electrode, which consists of a highly n-doped (n+) epitaxial layer on the handle of the SOI Wafer. The seperated n+ layer will form the Bitline of the crossbar arrangement. Furthermore, there is a need to connect the gate electrodes of single memory cells to form the wordline of the crossbar arrangement (Fig. 4.12). Since the gate electrodes are surrounded by n+ and p+ metal lines to the left and right on the device layer, the metallisation layer of the wordline needs some insulation from a first metallisation layer and will form a second metallisation layer.

The seperation of the 3.5 μm eptaxial n+ region on the handle wafer can be achieved by RIE. For the connection and structuring of the wordline there is a necessity to have a relatively planar surface. Therefore, the trenches etched by RIE need to be refilled. In conventional ICs manufacturing one would use a tetraethylorthosilicate (TEOS) oxide in back end of line (BEOL) processing and chemical mechanical polishing (CMP) to achieve inter metal insulation and planarisation. Since both were not available, a negative resist called SU-8 was used to achieve trench planarisation and insulation. This resist is usually used in MEMS and microfluidic systems to deposit a permanent resist layer as an active part of fabricated devices with high biocompatibility. The SU-8 resist, if properly hard-baked and cross-linked, is very stable and resistant to further chemical processes [158]. The cross-section of the fabricated crossbar arrangements is shown in Fig. 4.12 and a thin SU-8 Layer (1.4 μm) is used for insulation of the metal layers and thick SU-8 layer (4 μm-5 μm) for the trench refill. Consequently, the process optimisation is divided into the chapters: Aluminium metal layer optimisation (4.3.1), Thin SU-8 insulation optimisation (4.3.2), Trench refill optimisation (4.3.3), Bond pad optimisation (4.3.4) and the Combination of all processes (4.3.5).

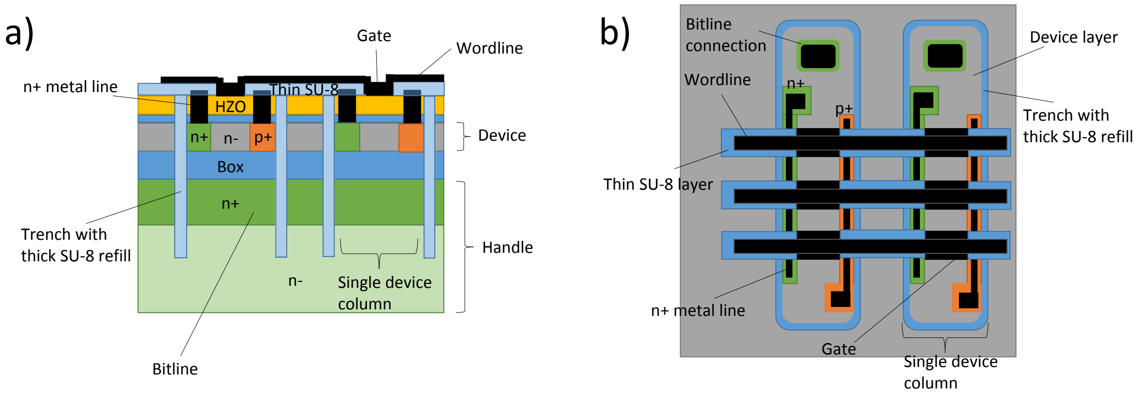


Figure 4.12: a) Cross-section of the crossbar arrangement with trenches and wordline metallisation and b) top view of crossbar arrangement with bitline connection, trenches and wordline.

4.3.1 Aluminium Metal Layer Optimisation

To start with, aluminium metallisation obtained from TU Dresden was used, which was e-beam evaporated. Later, the metallisation was grown at the Max Planck Institute by sputtering. Problems occurred with the metallisation for thick layers (230 nm) during heating to 180 °C. The aluminium appeared to chip away in some areas, which resulted in holes in the aluminium layer. The reason for this could be found in the tension which developed between the aluminium layer and the underlying layer due to differences in thermal expansion. Different temperatures and three different aluminium thicknesses (230 nm, 160 nm, 110 nm) were used to obtain an optimal aluminium thickness and its corresponding maximum temperature budget. Fig. 4.13a) shows microscope images of the aluminium surface for different annealing temperatures and a thickness of 230 nm. Temperatures over 150 °C lead to the mentioned holes and the situation deteriorates for 180 °C, whereas slow heating to 180 °C seems to improve the hole formation slightly. Thin aluminium layers (110 nm) are resistant against hole formation even at 180 °C. The hole formation is slightly visible for 160 nm films at 180 °C, but still acceptable.

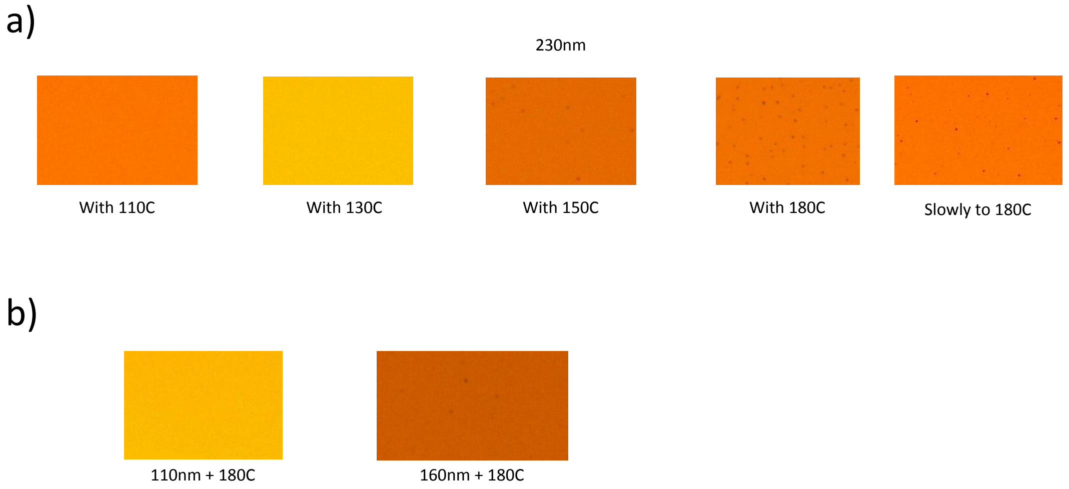


Figure 4.13: a) Microscope images of aluminium holes for different temperatures of 230 nm thick aluminium. b) Microscope images at 180 °C for (110 nm) and 160 nm thick films.

The conclusion from this experiment is that aluminium layers should be kept as thin and temperatures as low as possible.

Furthermore, a connection problem occurred between the first and second metallisation layer (contact between gate and WL). This is caused by an aluminium oxide layer which natively grows in air when taking out the sample from the chamber between the two metal layer growths. Moreover, this aluminium oxide layer is getting thicker during the SU-8 resist etch back in O_2 plasma (see chapter 4.3.3). An etching step was done before the deposition of the second aluminium layer in the same chamber with the same mask. The etching step etches approximately 7 nm into the aluminium oxide/aluminium and ensures aluminium oxide removal.

4.3.2 Thin SU-8 Insulation Optimisation

The thin SU-8 layer after the trench refill needs to fulfil two tasks: Firstly, the wordline metal needs to be connected throughout the full length, even at the holes/edges where it is connected to the gates of the devices. Secondly, the SU-8 should have good insulation properties to avoid any leakage or connection between the wordline and the n+ or p+ metal lines.

With regard to the first issue, connectivity tests were performed along SU-8 hills and trenches (Fig. 4.14). The aluminium films were structured by lift-off. Three different viscosities and therefore thicknesses of SU-8 resist were tried: SU-8 TF 6000.5 (0.5 μm), SU-8 TF 6001 (1.4 μm) and SU-8 TF 6005 (4.3 μm). The metal lines had different width and the trench width was also varied. Tab. 4 reveals the processing parameters for the three different SU-8 resists.

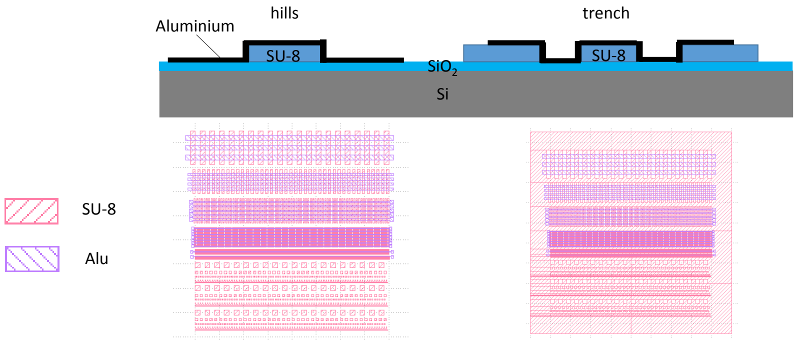


Figure 4.14: Crosssection of the trench and hill structure and corresponding top view mask layout.

SU-8 TF 6000.5	SU-8 TF 6001	SU-8 TF 6005
• Spin coat adhesion agent: AR300-800 • 60s; 4000 rpm; • Heat at 180C for 60s; • Spin coat SU8: 5s 500rpm, 40s 3000rpm; • Softbake: 1.5 min, 110C; • Expose: 350 mJ/cm²; • Post-exposure bake: 2 min 110C; • Develop: 3 min, MR-dev 600; • Aceton + IPA + DI Water for 30 s to remove rests; • Hardbake: Start at 110C –> 180C for 1 h; • –> 0.49 µm; resolution: 5 µm	• Spin coat: 5s 500rpm, 40s 2000rpm; • Softbake: 2.5 min, 110C; • Expose: 350 mJ/cm²; • Post-exposure bake: 2 min 110C; • Develop: 3 min, MR-dev 600; • Aceton + IPA + DI Water for 30 s to remove rests; • Hardbake: Start at 110C –> 180C for 1 h; • –> 1.4 µm; resolution: 5 µm	• Spin coat: 5s 500rpm, 40s 8000rpm; • Softbake: 4 min, 110C; • Expose: 350 mJ/cm²; • Post-exposure bake: 2 min 110C; • Develop: 3 min, MR-dev 600; • Aceton + IPA + DI Water for 30 s to remove rests; • Hardbake: Start at 110C –> 180C for 1 h; • –> 4.3 µm; resolution: 5 µm

Table 4: Processing conditions for SU-8 TF 6000.5, SU-8 TF 6001 and SU-8 TF 6005.

The metal lines were made out of 230 nm thick aluminium and structured by using Lift-off of overdeveloped negative resist. The resulting microscope images for the different SU-8 thicknesses are shown in Tab. 5.

The lift-off worked well for the ($0.5\,\mu$m) and ($1.4\,\mu$m) resist. Due to the fact that the negative resist for lift-off had a thickness of ($1.4\,\mu$m), the process did not work out well for the ($4.3\,\mu$m) thick SU-8.
In Fig. 4.15a) and b) the IV plots of connected metal lines are revealed for SU-8 TF 6001. The connectivity on SU-8 TF 6005 was not good enough and thus is not shown here. It is apparent that if all metal lines are connected a very good yield can be achieved. In Fig. 4.15c) and d) the IV plots of neighbouring metal lines is shown and for SU-8 TF 6001 the metal lines are well seperated.
The connectivity of the SU8-6000.5 resist (Fig. 4.16a) and b) was also good, only the seperation of neighbouring metal lines (Fig. 4.16c and d) was not sufficient in some cases, most likely due to bad insulation of the oxidized silicon substrate.

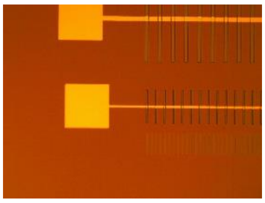
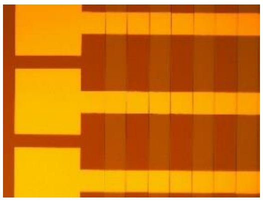
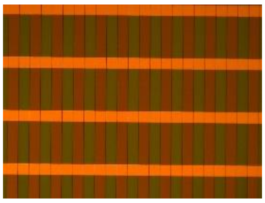
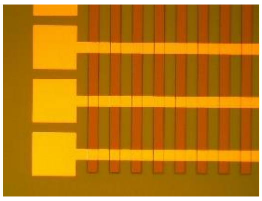
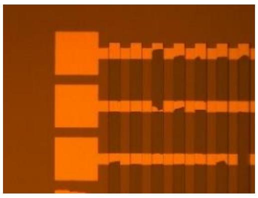
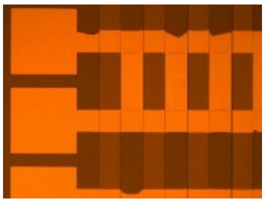
	hills	trenches
SU-8 TF 6000.5		
SU-8 TF 6001		
SU-8 TF 6005		

Table 5: Lift-off results on trenches and hills with different SU-8 resist thicknesses.

So far the SU-8 resist was hard-baked at 180 °C and the influence of the hard-bake on connectivity was not investigated. In 4.17 the connectivity for SU-8 6001 and 130 °C hard-bake is revealed and it becomes obvious that the many metal lines are not connected any more, especially for the trenched structures. It seems that hard-baking at 180 °C is rounding the edges of the resist, due to the fact that the glass transition temperature is exceeded. The glass transition temperature of SU-8 resist is in the range of 150 °C-200 °C [158]. Furthermore, thin aluminium (110 nm) was tested as a metallisation, and with which good connectivity of the metal lines was also achieved. In order to ensure good connectivity without risk the thick aluminium was used for the second metallisation layer.

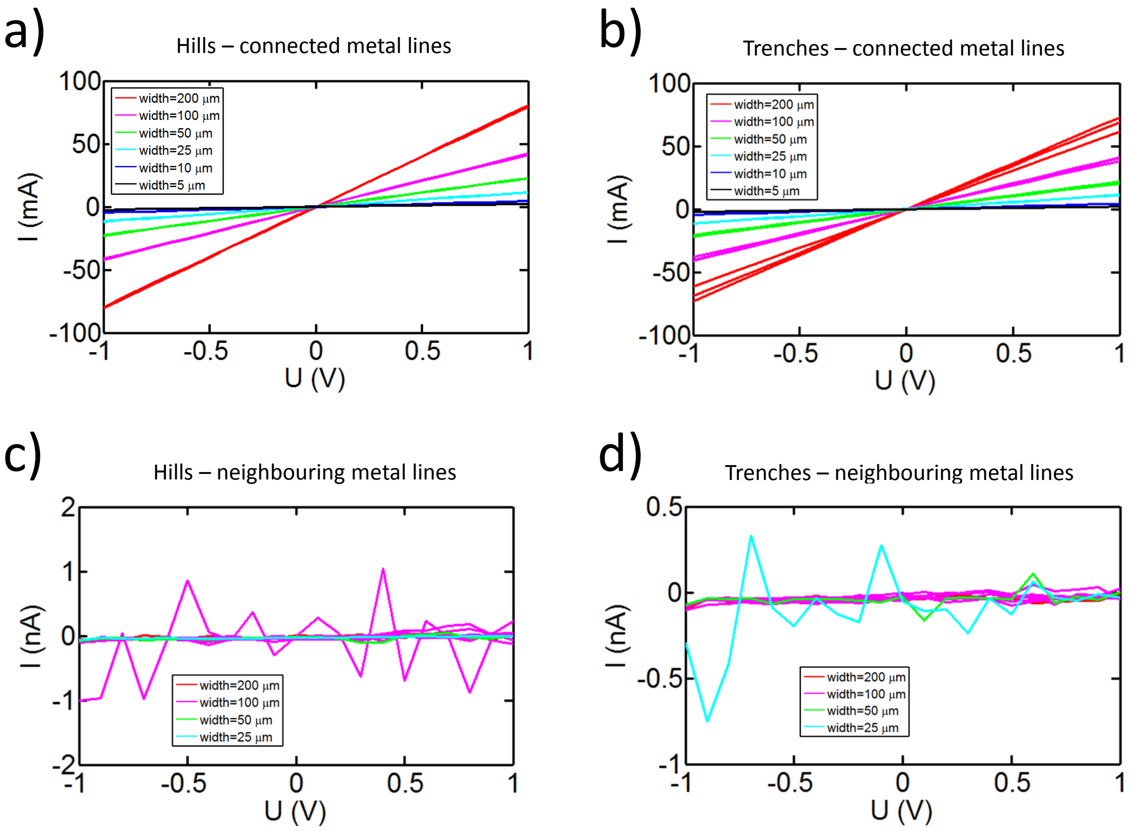


Figure 4.15: SU8-6001: a) IV curves of connected metal lines on hills. b) IV curves of connected metal lines on trenches. c) IV curves of neighbouring metal lines on hills. d) IV curves of neighbouring metal lines on trenches.

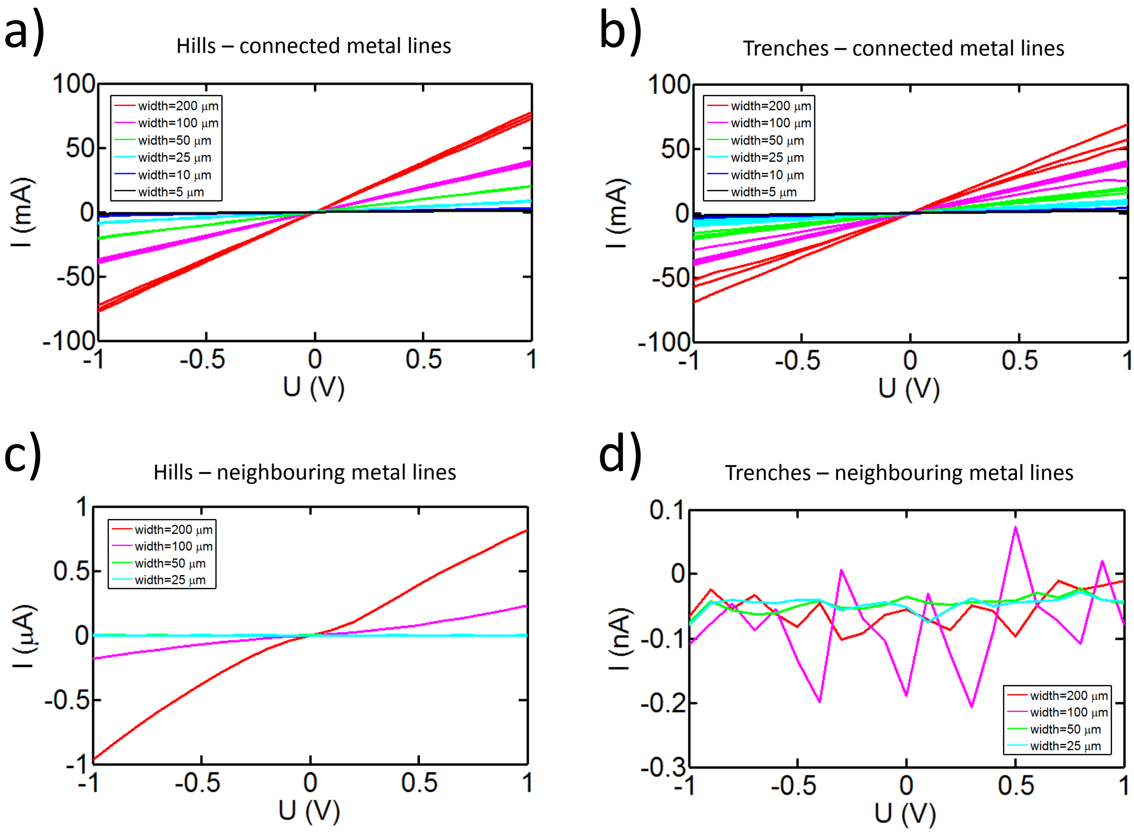


Figure 4.16: SU8-6000.5: a) IV curves of connected metal lines on hills. b) IV curves of connected metal lines on trenches. c) IV curves of neighbouring metal lines on hills. d) IV curves of neighbouring metal lines on trenches.

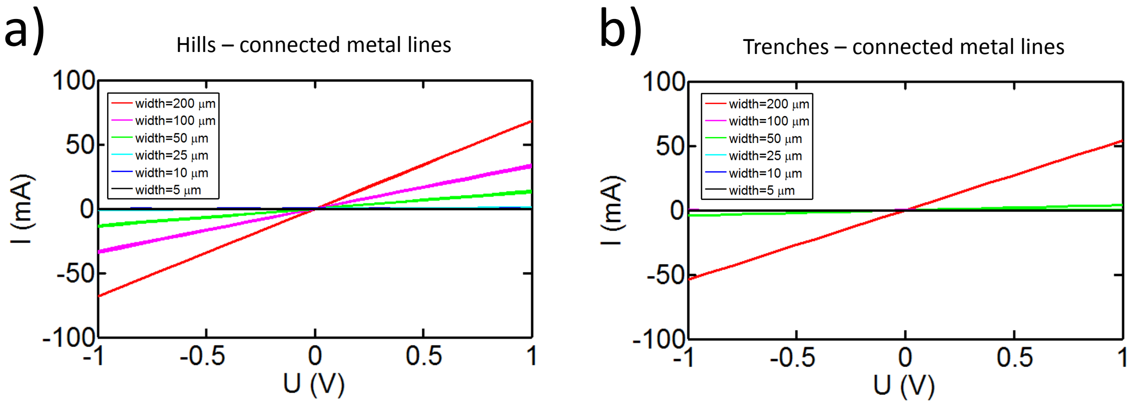


Figure 4.17: SU8-6001 with 130C hard-bake: a) IV curves of connected metal lines on hills. b) IV curves of connected metal lines on trenches.

Besides connectivity of the wordline along the edges, the SU-8 insulating layer should also enable good insulation between the wordline and n+ and p+ metal lines. For this purpose leakage tests of the SU-8 6000.5, SU-8 6001 and SU-8 6005 resist were performed with a structure shown in Fig. 4.18. There is a fully covered aluminium back-electrode and aluminium dots of different size on top of the SU-8 resist. As can be seen from Fig. 4.19 nearly all dots showed bad leakage properties with a slight improvement for the thick SU-8 resist (6005). The reason for this problem was further narrowed down to the tip contact with a leakage test structure as shown in Fig. 4.20. The structure consists of a crossbar where top and buttom aluminium electrodes cross each other with different areas. As can be seen from Fig. 4.21 the leakage current was in the range of <1 nA for nearly all sizes of crossing area, showing the good insulating properties of the SU-8 resist even for high voltages (20 V) when the aluminium lines are contacted from the side pads. Only for SU-8 6000.5 there was one large crossing area that had large leakage currents. The leakage current drastically increased with contacting the measurement probe tip directly on the crossing area, especially with higher pressure. Due to the fact that the SU-8 resist, even if hard-baked, is still a soft material there is a possibility of field strength peaks at the probe tip, which might cause the observed short circuit. In conclusion, the SU-8 resist shows good insulating properties if not directly contacted with the probe tip, which is not the case for the crossbar structure.

The conclusion from this process optimisation chapter is that the SU-8 6001 resist is used as insulating layer and 230 nm thick aluminium is used as second metallisation for crossbar fabrication, since best connectivity for the wordline was achieved with SU-8 6001 and the insulating properties are sufficient.

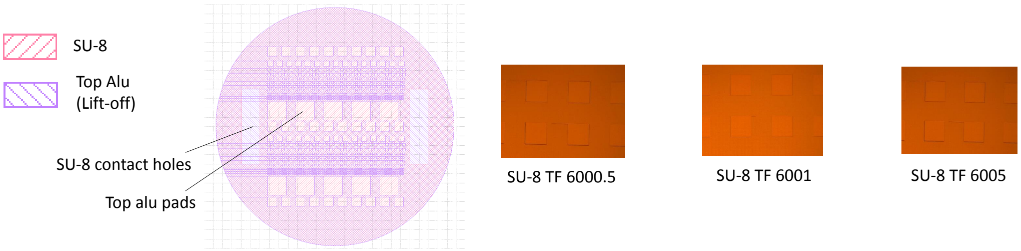


Figure 4.18: Mask design of first leakage test with fully covered backside electrode. Microscope images of the alumnium dots are provided on the right-hand side.

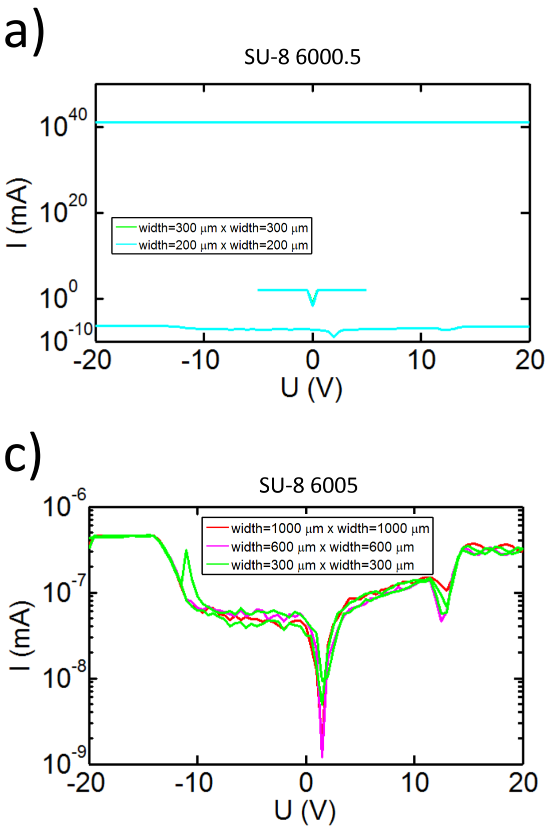


Figure 4.19: IV curve of leakage current for a) SU-8 6000.5 b) SU-8 6001 c) SU-8 6005.

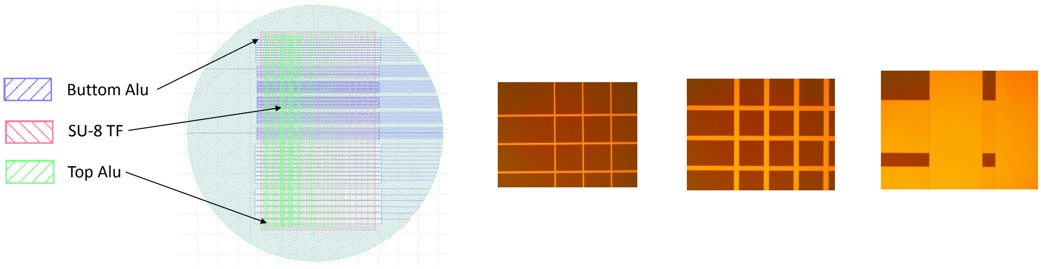


Figure 4.20: Mask design of second crossbar leakage test. Microscope images of the crossing points are provided on the right-hand side.

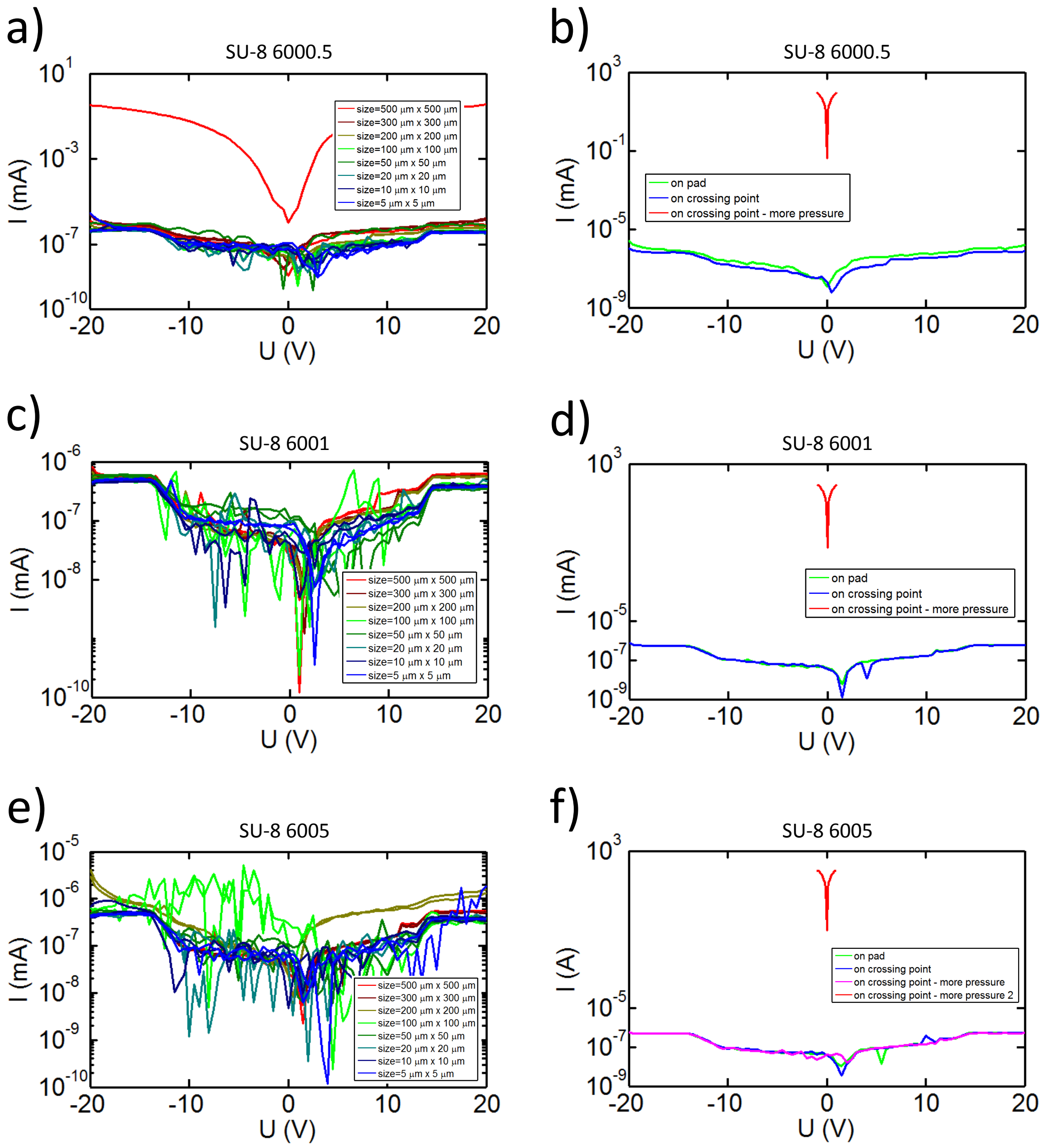


Figure 4.21: IV curves of leakage for a) SU-8 6000.5 with side pad contact b) SU-8 6000.5 with contact on crossing point c) SU-8 6001 with side pad contact d) SU-8 6001 with contact on crossing point d) SU-8 6005 with side pad contact and e) SU-8 6005 with contact on crossing point.

4.3.3 Trench Refill Optimisation

In order to seperate the buttom electrode, the 3.5 μm thick n+ epitaxial layer on the handle wafer of the SOI wafer needs to be etched. The seperation is achieved by a deep trench of 4.4 μm-7 μm depth. In the beginning the silicon device islands were seperated by wide trenches as shown in Fig. 4.22. The islands (95 μm, 110 μm, 140 μm) and trenches (105 μm, 140 μm, 155 μm) between them had different widths. The aim of this chapter is to describe the effect of planarisation on the trenched surface with resist, in order to deposit and structure the second metallisation layer.

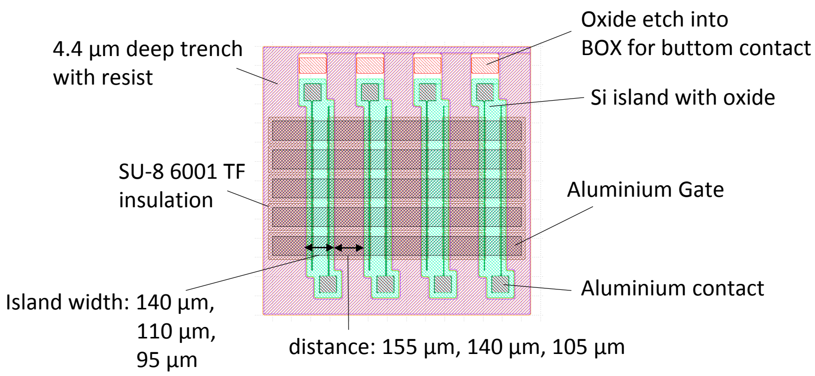


Figure 4.22: Mask layout of device structure with islands surrounded by trenched surface.

The structure of Fig. 4.22 was spin coated with 4.4 μm thick resist and structured so that only resist was left over in the trench. Fig. 4.23 reveals the obtained profile. The following characteristics become obvious: 1) The trench surface is approximately 1 μm higher than the silicon island. 2) The variation of thickness on the trench part is larger for the 150 μm trench width than the 95 μm trench width. 3) Annealing at 180 °C did not change anything on the profile. Obviously, this process did not lead to sufficient planarisation of trenches.

Spin coating of the resist with a subsequent ultrasonic step created holes in the resist, which were closed during soft-bake. No improvement was achieved with this experiment.

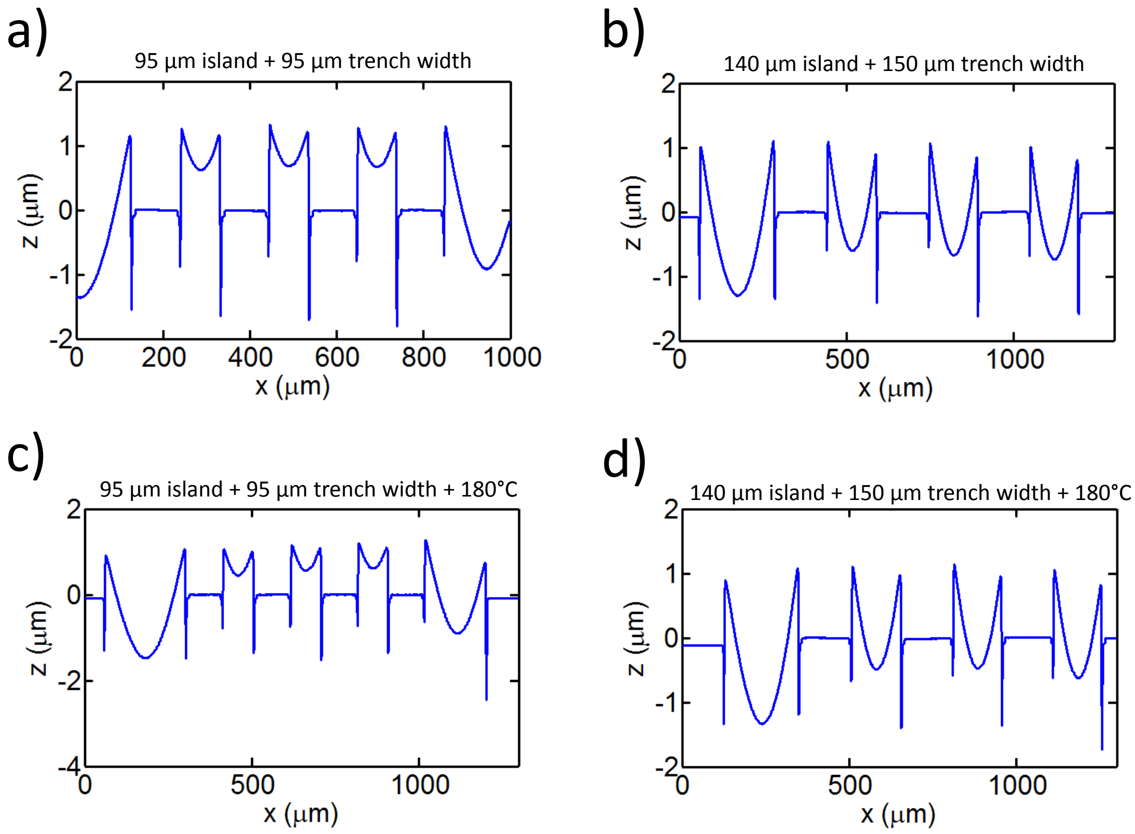


Figure 4.23: Trench refill with structured SU-8 6005 (8000 rpm) for a) 95 µm island + 95 µm trench width, b) 140 µm island + 150 µm trench width, c) 95 µm island + 95 µm trench width with 180 °C anneal and d) 140 µm island + 150 µm trench width with 180 °C anneal.

Using a thicker SU-8 resist (SU-8 6005 with 3000 rpm, 5.5 μm thickness) leads to higher trench surfaces (3 μm), but the trenches were more planar, especially for the 95 μm trench width (Fig. 4.24).

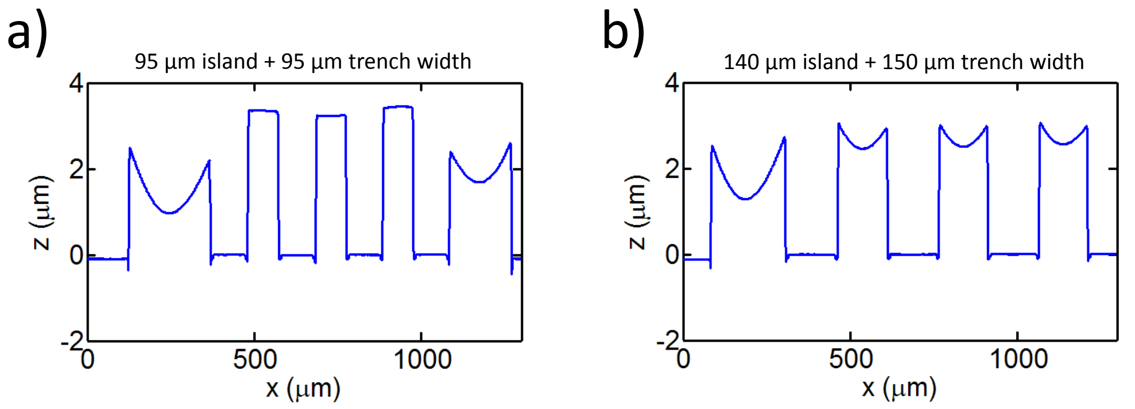


Figure 4.24: Trench refill with structured thicker SU-8 6005 (3000 rpm) for a) 95 µm island + 95 µm trench width and b) 140 µm island + 150 µm trench width.

It becomes obvious that the resist is too thick when using SU-8 6005. For this reason experiments with several subsequent coatings of SU-8 6001 (1.4 μm) were conducted. In Fig. 4.25 the obtained profiles are shown for 3 coatings of SU-8 6001, but the structures obtained were to inhomogeneous. Once a certain inhomogeneity is obtained with a first coating the inhomogeneity gets worse for the next coating.
Experiments with a doctor blade method lead to very high variations of resist thickness, especially towards the edges, and were not successful. Furthermore, planarisation resist from Brewer Science (M10-44) was tested, but the results were not satisfying [159].

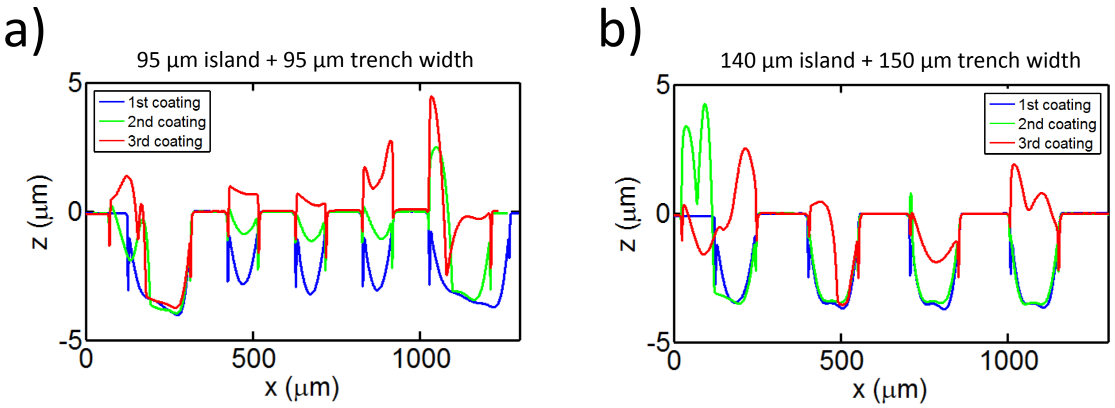


Figure 4.25: Three coatings with SU-8 6001 for a) 95 µm island + 95 µm trench width and b) 140 µm island + 150 µm trench width.

The conclusion from these experiments was that it is better to cover the whole area of islands and trenches with SU-8 resist and then perform an etch-back with RIE [160]. In the beginning only RIE with O_2 plasma was tried and the results after different lengths in etching times are shown in Fig. 4.26 with corresponding microscope images. Long etching times were necessary to remove the resist rests on certain parts of the island, which leads also to relativily deep trenches of 2 μm. Furthermore, the roughness during O_2 etching was quite high. Other groups have shown that etching of SU-8 resist with less roughness can be accomplished in a CF_4 etching process [161,162].

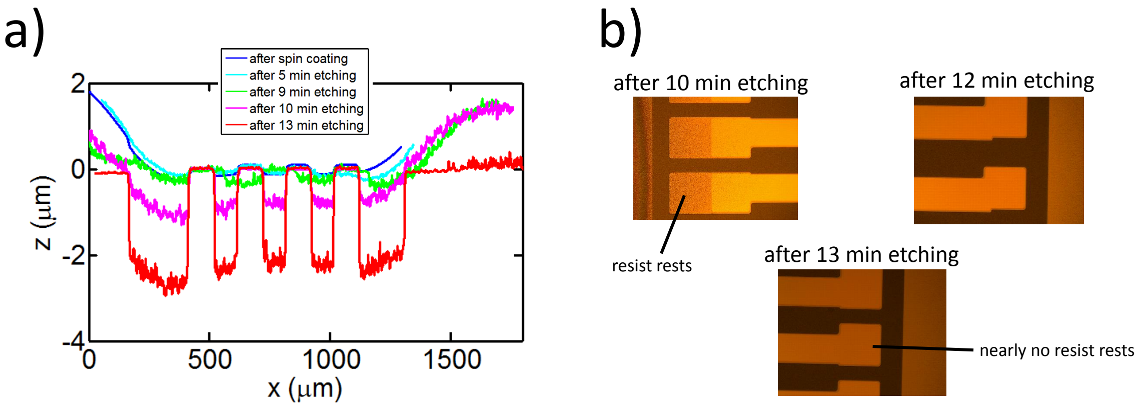


Figure 4.26: RIE etching of SU-8 6005 resist a) Profile of RIE etched SU-8 6005 resist and b) corresponding microscope images.

Fig. 4.27a) shows the etching of a SU-8 step in CF_4 plasma and Fig. 4.27b) the corresponding roughness after 13 min of etching. Fig. 4.27c) reveals the etch-back of the SU-8 resist on a trenched structure, and compared to the etching with O_2 plasma (Fig. 4.26), the roughness as well as the trench depth achieved after etching (0.8 μm-1.5 μm) are much lower.

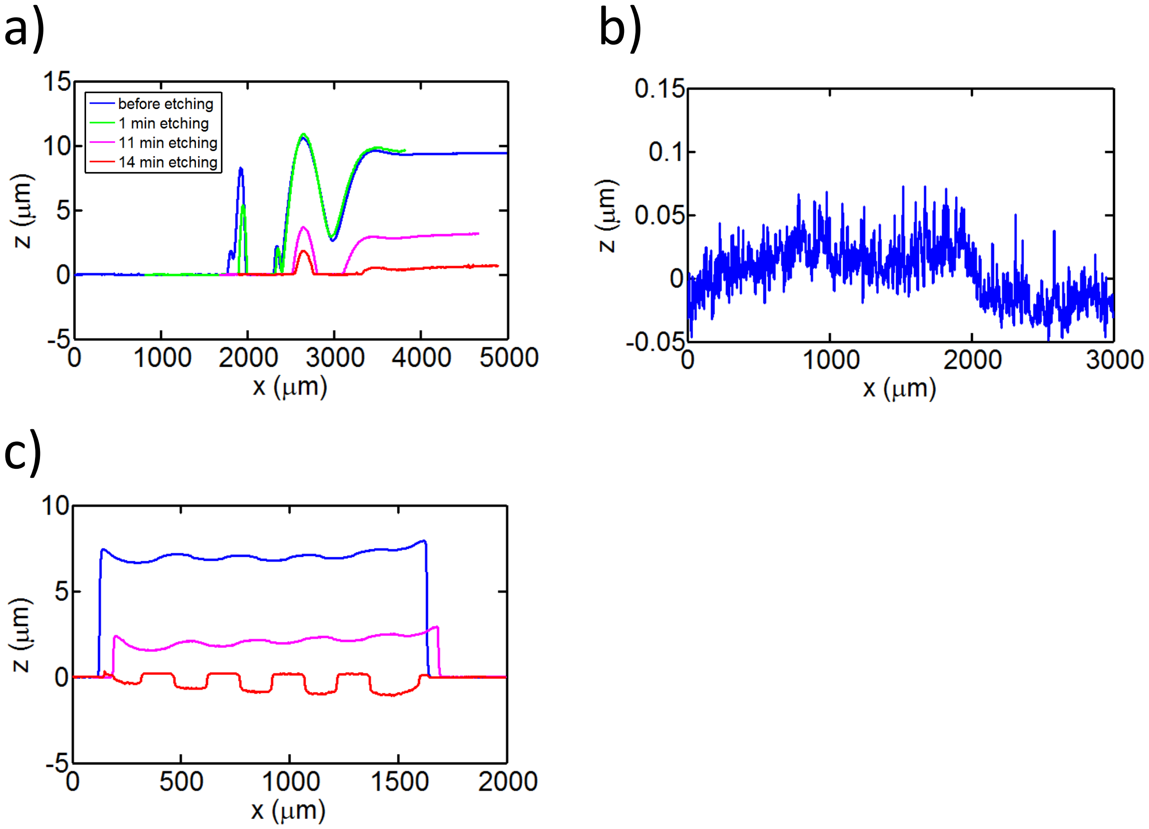


Figure 4.27: RIE etching of SU-8 6005 resist with CF_4 plasma: a) Profile of an SU-8 step for different etching times, b) corresponding roughness after 14 min etching and c) profile of the real trenched structure for different etching times.

Although the CF_4 etching improves the roughness, the problem remains that CF_4 attacks silicon. Therefore, firstly most of the resist is etched with CF_4 until less than $1.5\,\mu$m is left and secondly the left over resist thickness is etched with normal O_2 plasma. Moreover, more narrow trenches are used, which should reduce the thickness variation of the spin coated resist as a lower fraction of the total surface is covered by the deep trench. This was already observed in Fig. 4.23 and 4.24 ($95\,\mu$m versus $150\,\mu$m). Also the trench width was irregular, as can be seen from Fig. 4.22.

The mask layout of the new test vehicle is shown in Fig. 4.28a): The trench surrounded the island with a constant width of $10\,\mu$m, $20\,\mu$m, $50\,\mu$m and $100\,\mu$m. The device and buried oxide layer of the SOI wafer was etched away prior to the trench etching, which made it possible to measure the connecting resistance along the n+ silicon island and the separation resistance between the n+ silicon island with the trenches in the current path. These two IV curves are shown in Fig. 4.28b) and c) respectively. The ratio between connecting currrent at 1 V and disconnecting current at 1 V were 5817 for the $10\,\mu$m trench, 7023 for the $20\,\mu$m trench and 70818 for the $50\,\mu$m trench. Since the current/charge of the read-out line is usually detected by an low input-impedance amplifier, most of the charge is transferred to desired amplifier even at parasitic sneak paths in the range of 1:1000. So insulation of the read-out electrodes should be sufficient. Furthermore, it was later recognized that increasing the trench depth by $2.5\,\mu$m to $7\,\mu$m can increase the disconnecting resistance by 2 orders of magnitude.

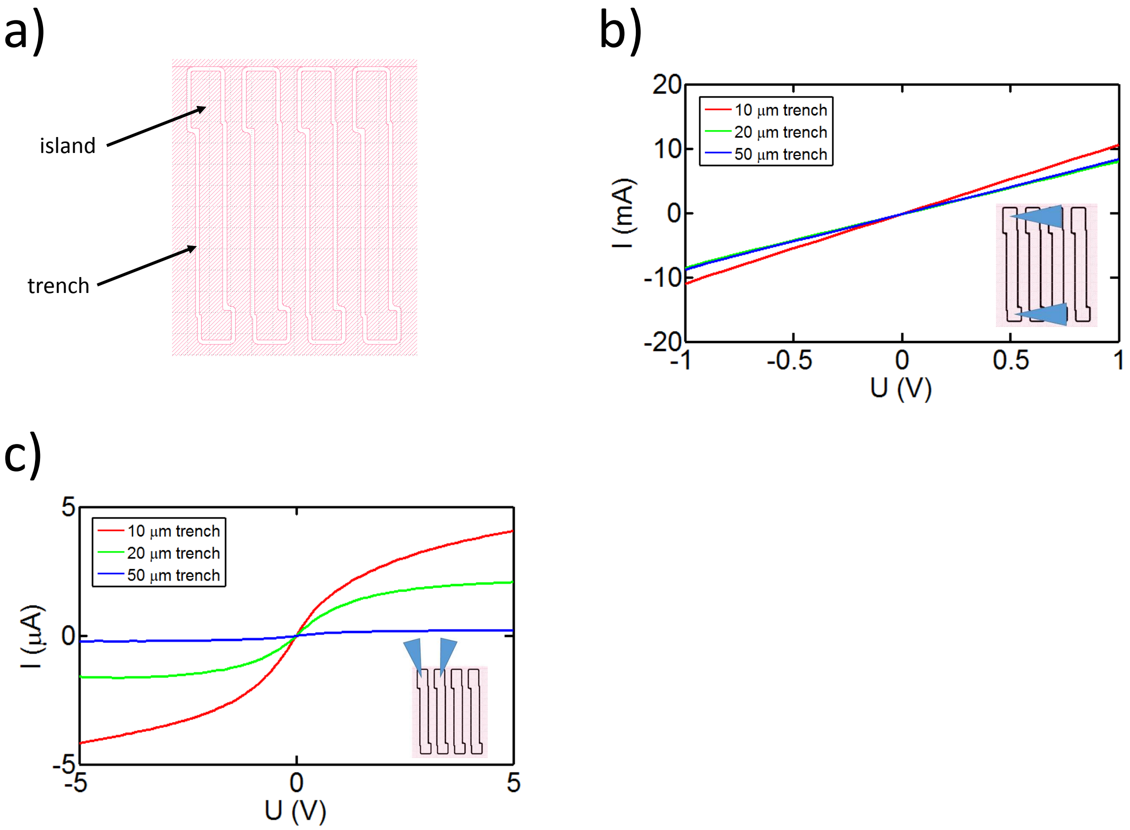


Figure 4.28: a) Mask layout of the new trench etch. b) IV connectivity for different trench widths along the n+ doped island. c) IV connectivity between the silicon islands for different trench widths.

The spin coating results for 4.5 μm thick (SU-8 6005, 8000 rpm), 1.4 μm thick (SU-8 6001, 2000 rpm) and 8.7 μm thick (SU-8 6005, 1000 rpm) resist are shown in Fig. 4.29. The SU-8 6001 resist was only tested for the 10 μm wide trenches and relatively large variations on the trenches become visible. For the other trench width only SU-8 6005, 8000 rpm and SU-8 6005, 1000 rpm was tried. The remaining trench depth difference remained the same for both resists for the 20 μm and 50 μm trench width. Only the 100 μm wide trenches had a different trench depth between the SU-8 6005, 8000 rpm and SU-8 6005, 1000 rpm.

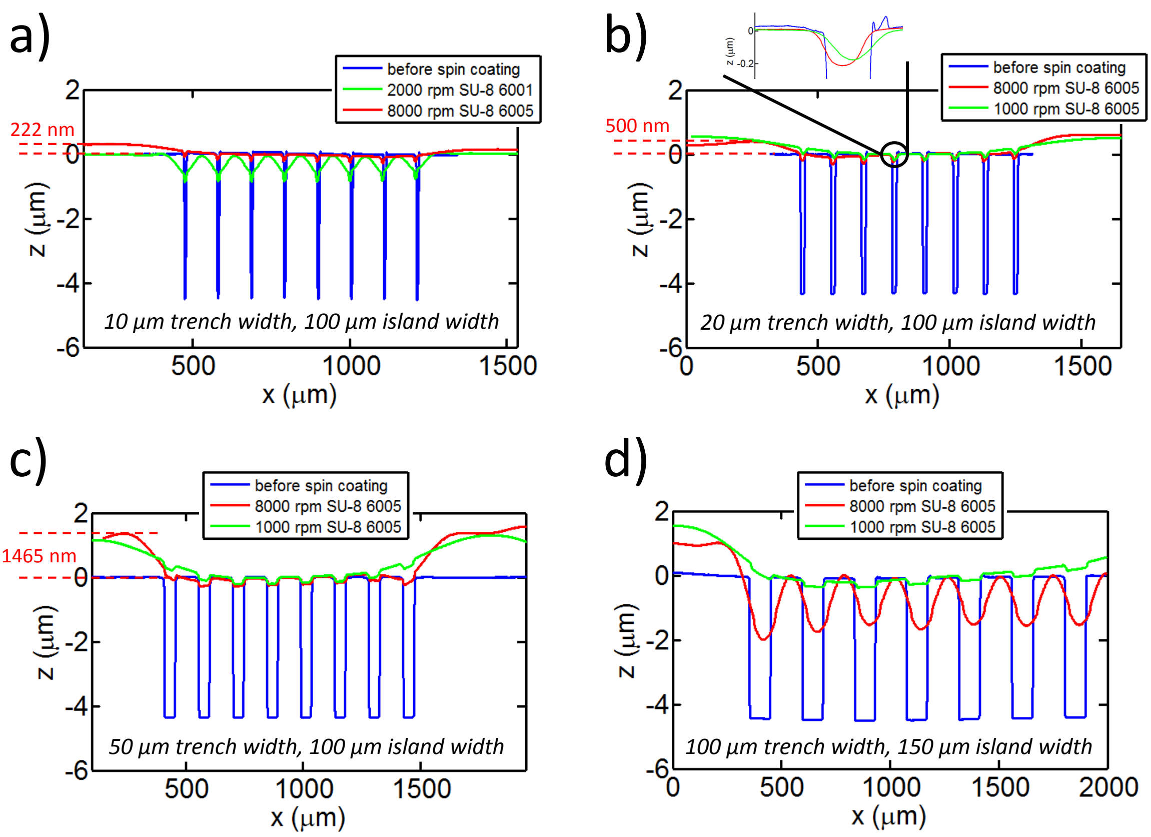


Figure 4.29: Before and after spin coating for a) 10 μm trench width and 100 μm island width, b) 20 μm trench width and 100 μm island width, c) 50 μm trench width and 100 μm island width and d) 100 μm trench width and 150 μm island width.

The height difference of the surfaces with trenches and without trenches is obviously different between the different trench widths and is marked in Fig. 4.29 and Tab. 6. As already stated, the remaining trench depth after spin-coating for different trench widths did not change much, but as the relative percentage of trenched surface increases, the spin coated surface is deeper compared to the surface that is not covered with trenches. This surface height difference is plotted against the trench coverage of the surface and a linear relationship becomes visible (Fig. 4.30). Deeper trenches lead to a more pronounced surface height difference in a proportional manner, as proven for 7 μm deep trenches. Generally, the average height difference of a surface with trenches and without trenches is $d_{tr} \cdot c$, where d_{tr} is the trench depth and c is the trench coverage. From Fig. 4.30 one can conclude a proportionality factor of 0.68 for the surface height difference Δh:

$$\Delta h = d_{tr} \cdot c \cdot 0.68 \tag{4.3.1}$$

Trench width d_{tr} (μm)	Trench coverage c (%)	Surface height difference Δh (nm)	Trench depth after spin coating (nm)
10 μm	10 %	222 nm	200 nm
20 μm	20 %	500 nm	230 nm
50 μm	50 %	1456 nm	220 nm

Table 6: Results of height difference and remaining trench depth after spin coating for 10 μm, 20 μm and 50 μm trench width.

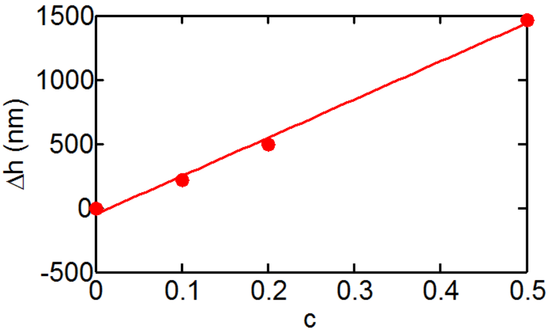


Figure 4.30: Surface height difference versus trench coverage.

The 10 μm trench width has the lowest variation in resist thickness (Fig. 4.29), but current separation between the islands does not appear to be optimal (Fig. 4.28c). As a compromise for low resist thickness variation and high current separation the 20 μm and 50 μm wide trenches were further tested for the etch-back of the resist. The etching was done in a two step etch process, first with CF_4 and then in O_2. Fig. 4.31 reveals the resist profiles for 20 μm wide trenches at different etching times. The final remaining trench depth at the center of the wafer is 0.53 μm and 0.83 μm at the top of the wafer. A thinner resist thickness towards the edge of the wafer seems to be the reason for the difference in the remaining trench.

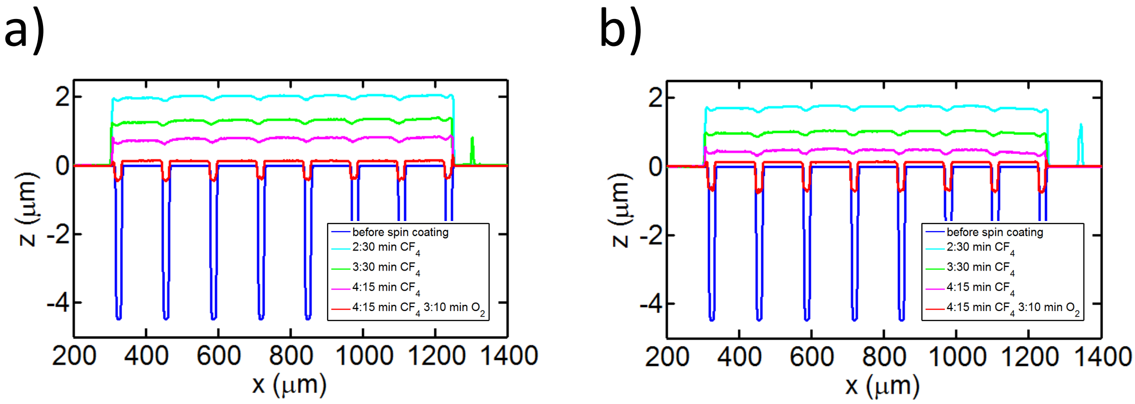


Figure 4.31: Profile for 20 μm trench width at a) wafer center and b) wafer top.

For the 50 μm wide trenches a remaining trench depth of 0.78 μm (center) and 2 μm (top) was obtained (Fig. 4.32). The 20 μm wide trenches seem better suited and were used from now on.

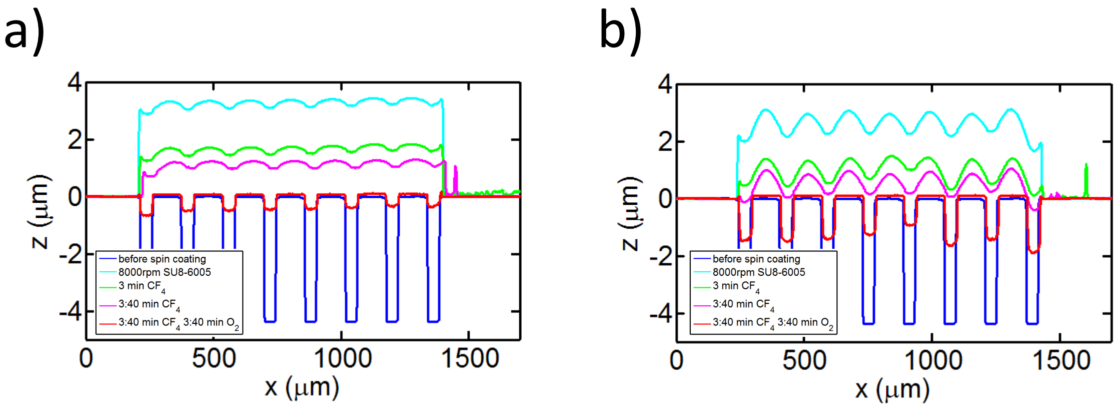


Figure 4.32: Profile for 50 μm trench width at a) wafer center and b) wafer top.

4.3.4 Bond Pad Optimisation

The bond pads at the border of the chip cannot be on the insulating SU-8 6001 resist, since the material is too soft for wire bonding, although there were some reports on the optimisation of bonding on SU-8 resist [163]. Therefore, the aluminium bond pads need to be on top of the buried oxide.

One possible failure might be the parasitic capacitive coupling between the bond pads due to the n+ doped epitaxial layer as a connector. This coupling can be avoided by grounding the epitaxial layer, as shown in Fig. 4.33.

Furthermore, due to the O_2 etching of the SU-8 resist in the process step described in chapter 4.3.3, the buried oxide is attacked and degraded with respect to its leakage properties at unprotected parts. Growing aluminium on top of this buried oxide will lead to leaky bond pads, especially after wire bonding (Fig. 4.34a and b). Therefore, the bond pads are deposited during the first aluminium metallisation before the SU-8 etch-back. Thus the buried oxide is protected by the bond pad during O_2 etching and the leakage properties are much better, even after wire bonding (Fig. 4.34c and d).

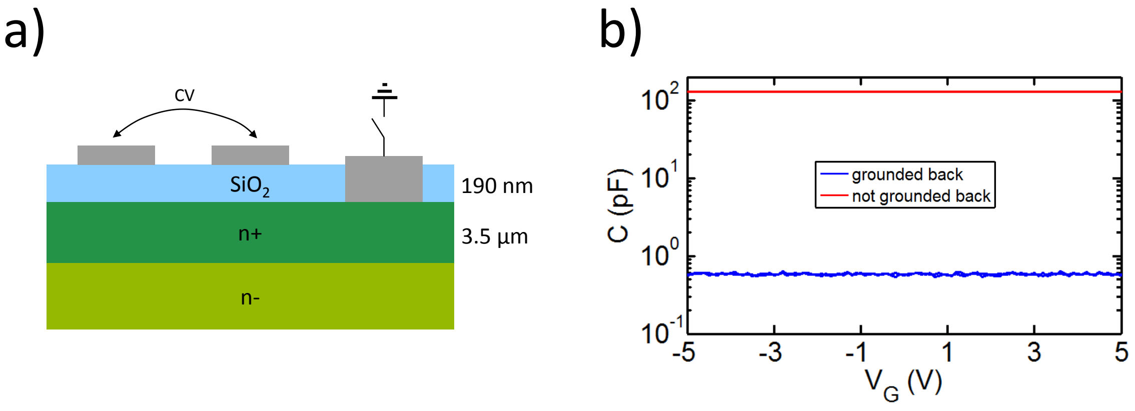


Figure 4.33: a) Crosssection of the CV parasitic coupling test between two pads. The n+ back contact is either connected to ground or kept floating during the measurement. b) Corresponding CV plots for grounded or floated back contact.

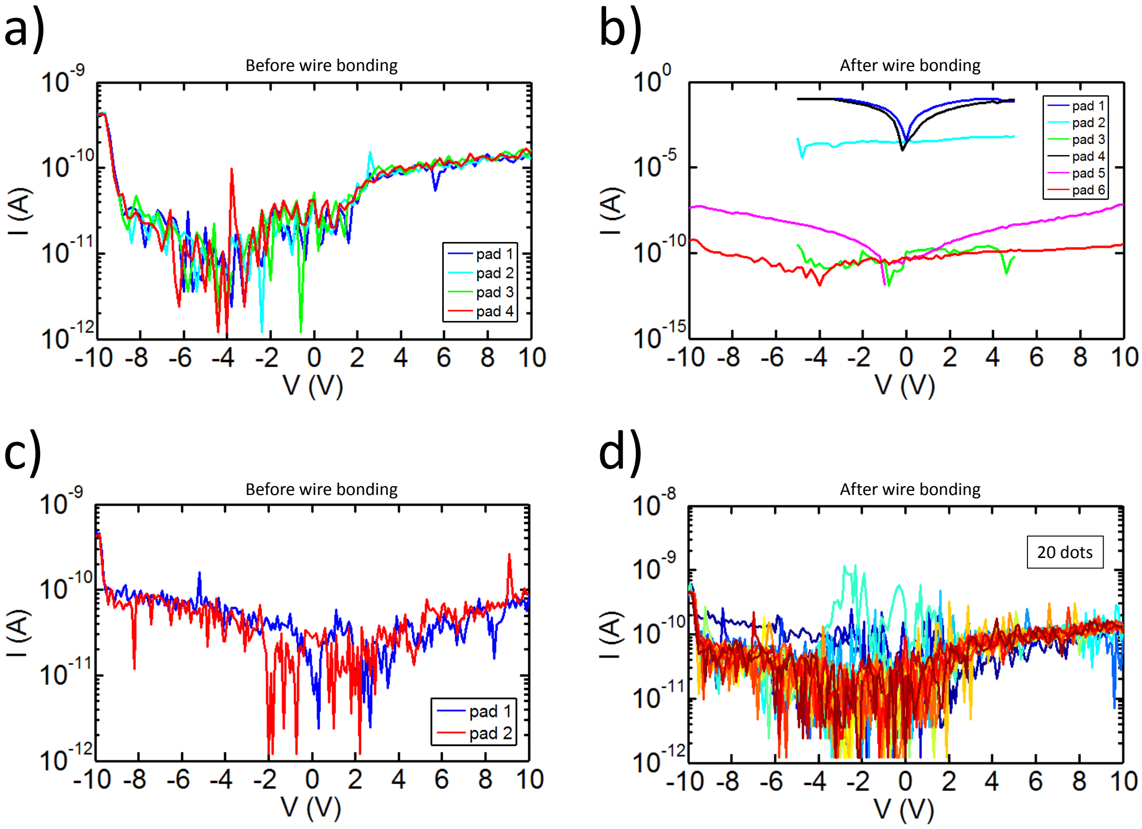


Figure 4.34: IV curve between the aluminium Bond pad and the n+ back contact: a) Aluminium bond pad deposited after O_2 etching, before wire bonding. b) Aluminium Bond pad deposited after O_2 etching, after wire bonding. c) Aluminium bond pad deposited before O_2 etching, before wire bonding. d) Aluminium bond pad deposited before O_2 etching, after wire bonding.

4.3.5 Combination of All Processes

Finally, all processes, namely first metal layer, trench etch and refill and thin SU-8 insulation with second metal layer, were combined on a test wafer to check electrical separation and connectivity. Fig. 4.35 shows the mask design of all layer and a corresponding zoom-in.

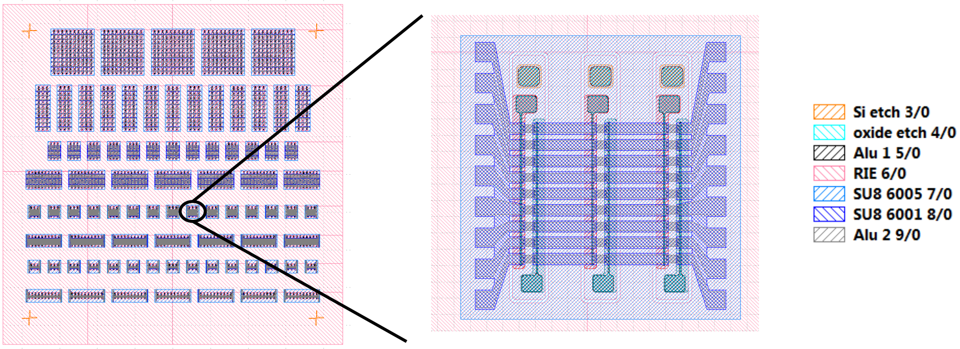


Figure 4.35: Mask design of the test wafer with a zoom in for one structure.

A thermal oxide was used for these test wafers instead of hafnium oxide. In Fig. 4.36a) the connectivity along the wordline of the crossbar test fabrication is shown for 9 wordlines. The connectivity was tested for 45 WLs and all were connected, so the WL-yield is expected to be around 100 %. Furthermore, the insulation of the wordlines with the p+ and n+ supply lines is proven in Fig. 4.36b) and the insulation between wordlines in Fig. 4.36c). Fig. 4.36d) shows the current between the device layer and the bitline and thus proves the good insulating properties of the buried oxide.

Thus generally good WL connectivity and insulation between disconnected lines was achieved.

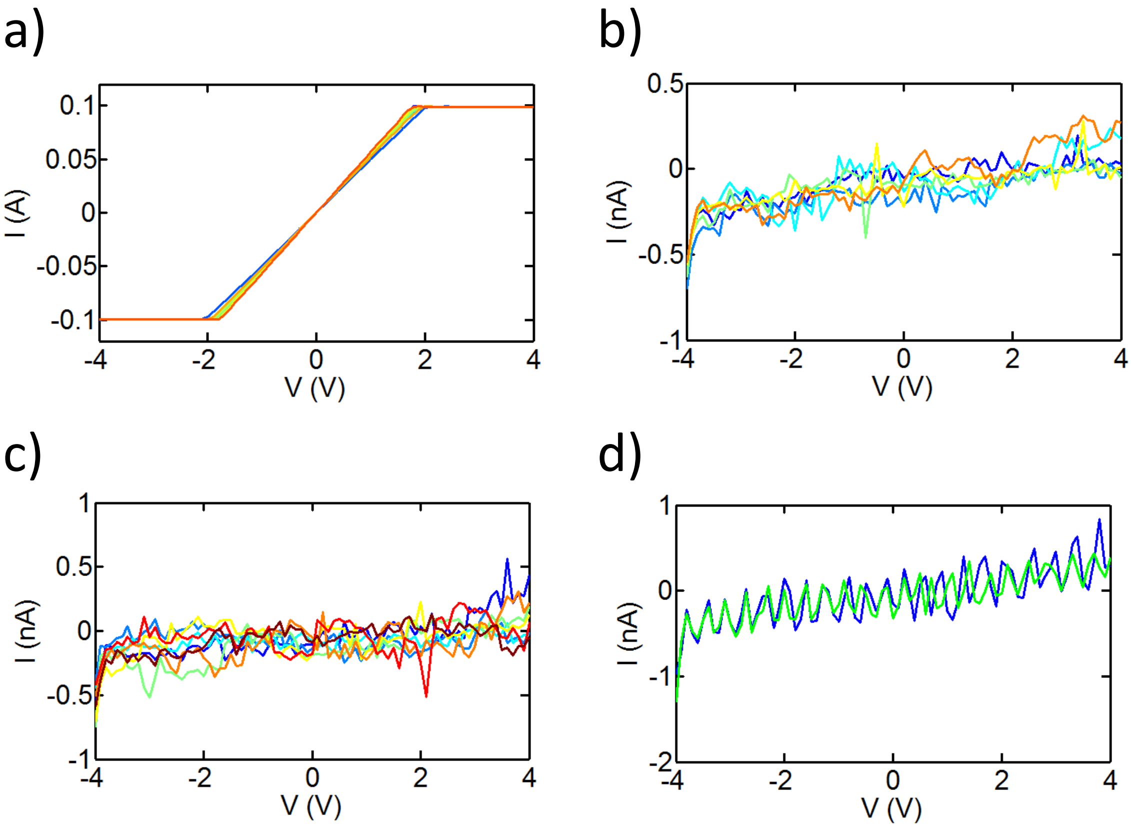
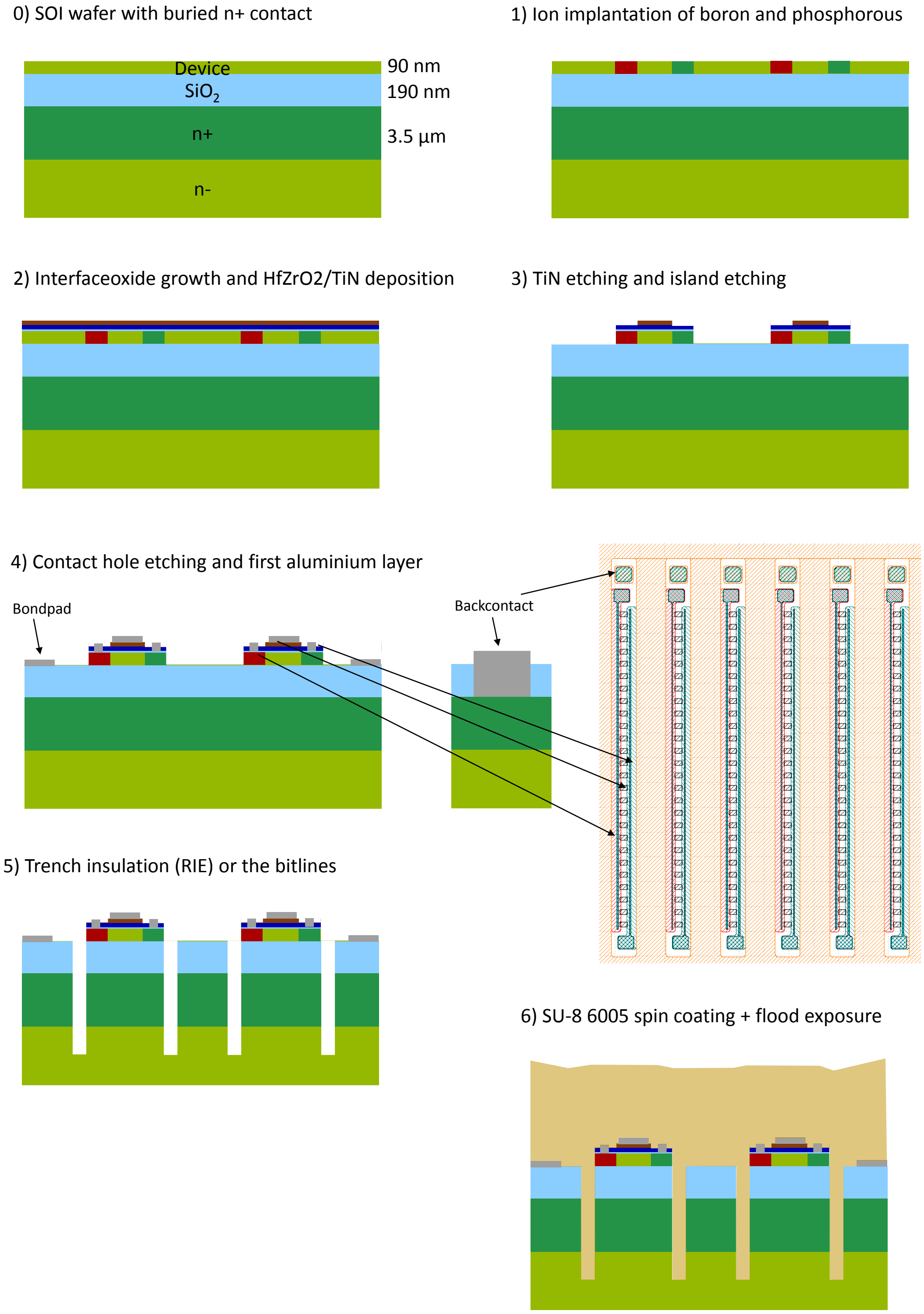


Figure 4.36: a) Connectivity along the wordlines. The current saturation is from the instrument. b) Insulation between wordline and n+ and p+ contact. c) Insulation between different wordlines. d) Insulation between device layer and bitline.

The process flow for the full crossbar fabrication with the combination of the prior explained processes is shown in Fig. 4.37: At first the ion implantation is done with a subsequent activation anneal. The interface oxide and the hafnium zirconium oxide/TiN growth is followed. The TiN is firstly structured and then the silicon islands of the device layer were etched. The contact holes into the hafnium zirconium oxide for the n+ and p+ contacts were etched by ion beam etching and subsequent HF-etch. The first aluminium layer was deposited and the bond pads were also structured by lift-off in the next step. Up to this point the fabrication of crossbar devices is very similar to the single device fabrication.

The next step is the deep trench etching of 7 μm and the refill with thick SU-8 6005 resist. Then the SU-8 6001 resist was structured as an insulating layer. The last aluminium layer was structured by lift-off and an ion beam etch step before deposition was done for aluminium oxide removal of the first aluminium layer.

The distance of the n+ and p+ region (gate length) of the fabricated devices was 50 μm and 100 μm. The width of each memory cell was the same. The GND plate in Fig. 4.37 is the trapezoidal surface in the mask design and decreases parasitic capacitive coupling between the bond pads. Appendix A.2 shows the mask layout for each process step.



0) SOI wafer with buried n+ contact

1) Ion implantation of boron and phosphorous

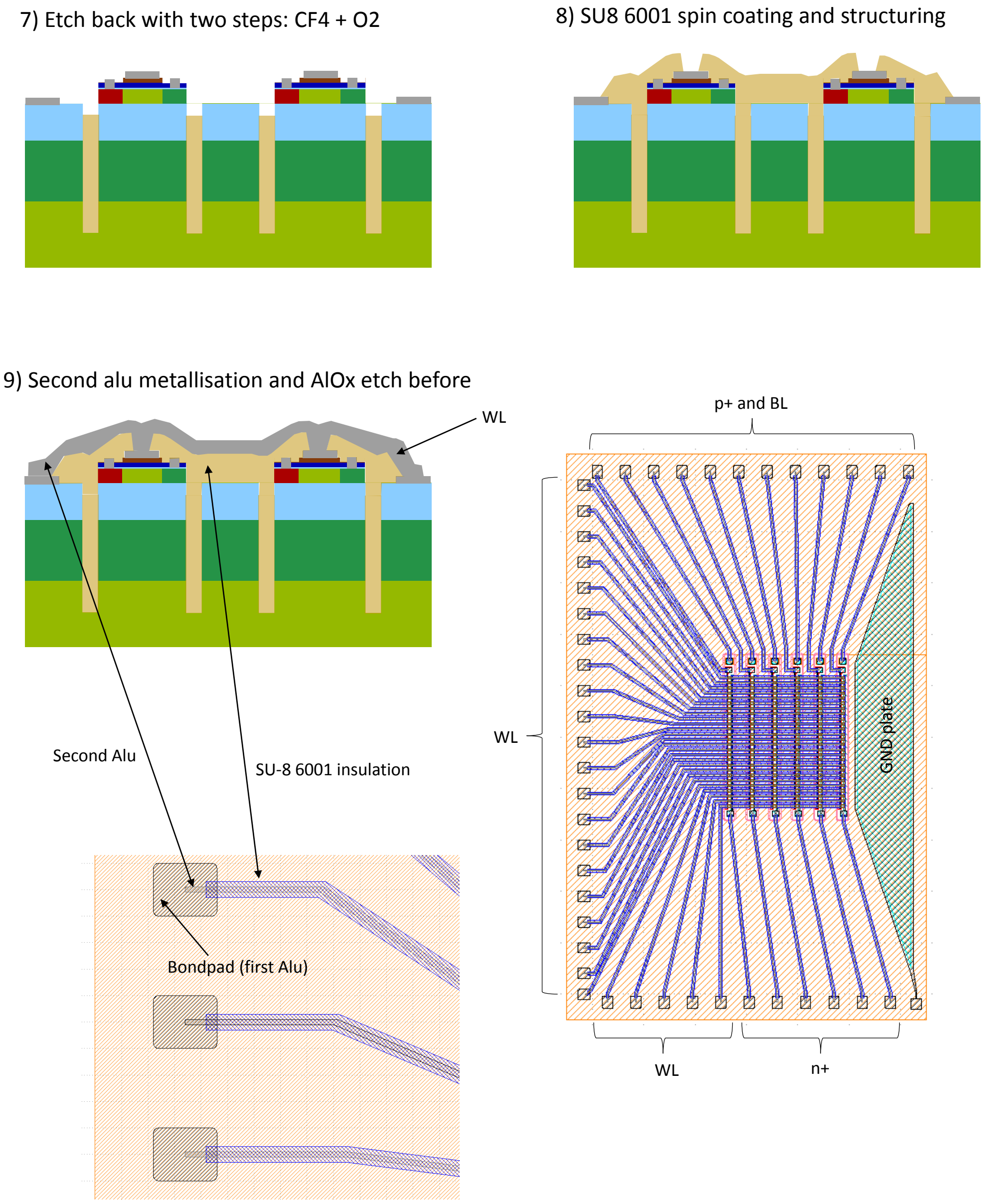
2) Interfaceoxide growth and HfZrO2/TiN deposition

3) TiN etching and island etching

4) Contact hole etching and first aluminium layer

5) Trench insulation (RIE) or the bitlines

6) SU-8 6005 spin coating + flood exposure



7) Etch back with two steps: CF4 + O2

8) SU8 6001 spin coating and structuring

9) Second alu metallisation and AlOx etch before

Figure 4.37: Process flow for crossbar devices.

Several crossbar chips with either 50 μmx50 μm or 100 μmx100 μm memory cell size and either 10 wordlines or 26 wordlines were fabricated on a sample. These chips are diced out and wire bonded on a chip carrier for further measurement. Fig. 4.38 shows the wire bonded chip and a microscope image of the memory matrix. The images of the memory cells are further zoomed-in to single memory cells by SEM images.

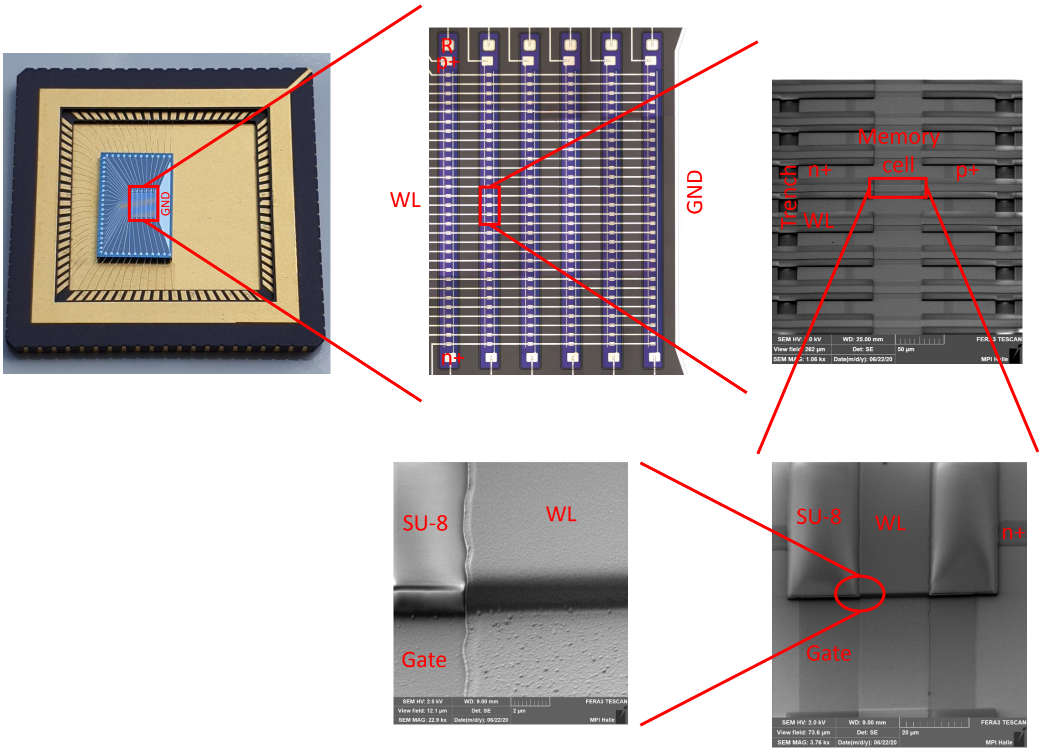


Figure 4.38: Wirebonded crossbar chip with microscope image of the memory matrix and further zoomed-in SEM images. Material from [8]

4.4 Growth of Oxide Devices

Growing the Metal-Oxide-Semiconductor-Oxide-Metal structure fully out of oxides has several advantages: There is no interface oxide between semiconductor and ferroelectric, thus the endurance and ferroelectric behaviour might be better. Moreover, the devices can be grown at lower temperature, thus enabling back end of line integration. Another advantage is the easier 3D integration, since several crossbar structures can be grown on top of each other.

As an oxide semiconductor material several oxides are possible, e.g. $SrTiO_3$,TiO_2 or Indium-Gallium-Zinc-Oxide (IGZO). In the growth experiments mainly PLD grown $SrTiO_3$ films were investigated. $SrTiO_3$ or $SrZrO_3$ can also be used as an insulating layer. The ferroelectric film can be also made of a perovskite material, like BaTiO3 or PZT.

Fig. 4.39a) shows the PLD grown stacks, which were investigated in this thesis. Fig.4.39b) shows the obtained RHEED pattern for the $SrTiO_3$ stack. For the purpose of this thesis only stacks without ferroelectric material were tested (e.g. $SrTiO_3$ (ins) - $SrTiO_3$ (conducting) - $SrTiO_3$ (ins)).

Generally superior insulating $SrTiO_3$ were obtained for growth and subsequent annealing in high oxygen partial pressure. For growing n-doped $SrTiO_3$ films a lower oxygen partial pressure is necessary in order to generate oxygen vacancies [164]. Growing an n-doped layer on top of an insulating $SrTiO_3$ seems difficult, since the insulating $SrTiO_3$ becomes too leaky when exposed to lower oxygen partial pressure during the conducting $SrTiO_3$ growth. Growing Niobium doped $SrTiO_3$:Nb can increase the conductivity whilst increasing the growth pressure [165, 166]. High doping concentrations (10 %) are necessary in order to obtain good conductivity. But most likely due to oxidation of Nb to Nb_2O_5 [167] the $SrTiO_3$:Nb also becomes insulating at high oxygen partial pressure growth or annealing. In summary the growth window for obtaining a good insulating $SrTiO_3$ and a conducting $SrTiO_3$:Nb layer on top seemed very narrow, and probably does not even exist.

For this reason experiments with $SrZrO_3$ were conducted, which has a large band gap (5.7 eV) [167]. The insulating properties seemed to be better, but still not good enough. Also it was tough to obtain flat epitaxial growth of $SrZrO_3$ for thicker films.

A better solution in future would be to grow the oxides with ALD. There is some published work in the context of thin film transistor (TFT) with TiO_2 [168–171]. The insulating layers can be made out of normal high-k dielectrics (e.g. HfO_2). Furthermore there is a publication on a ferroelectric TFT with IGZO, including a back Gate [172]. The advantage of ALD grown devices would be the lower growth temperature and the films are still very smooth, even though they are not epitaxial.

Due to the difficulties with the PLD grown films the oxide approach was not further followed throughout this thesis.

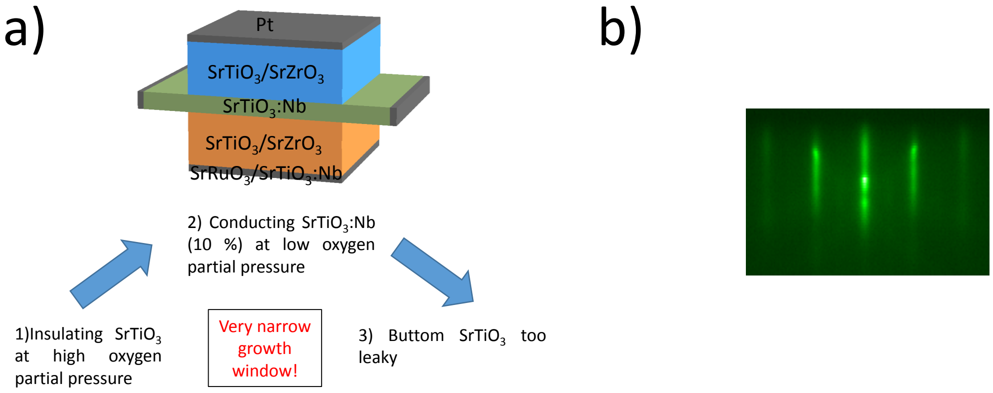

a)
Pt
$SrTiO_3/SrZrO_3$
$SrTiO_3$:Nb
$SrTiO_3/SrZrO_3$
$SrRuO_3/SrTiO_3$:Nb
2) Conducting $SrTiO_3$:Nb (10 %) at low oxygen partial pressure
1)Insulating $SrTiO_3$ at high oxygen partial pressure
Very narrow growth window!
3) Buttom $SrTiO_3$ too leaky
b)

Figure 4.39: a) Device stack and material combination and the growth problem. b) RHEED image of a grown $SrTiO_3$(ins)-$SrTiO_3$:Nb stack.

5 Measurement Setup and Printed Circuit Board

In this chapter the concepts for reading and writing single and multiple memory cells are explained. In the first sub-chapter 5.1 the measurement setup for single devices is explained. This includes CV measurement and write/read-out with pulses and larger AC voltages.

For the measurement of crossbar devices and implementation of a neuromorphic algorithm many different signals need to be applied to several pins and read-out. For this purpose a PCB was developed and it was controlled with a data acquisition system (DAQ) (see sub-chapter 5.2) (National Instruments USB-6363).

5.1 Measurement Setup for Single Devices

The general measurement setup is shown in Fig. 5.1a). The n+ and p+ contacts are mostly connected to ground, but can be also supplied to a DC (direct current (DC)) voltage in order to deplete or enhance the pin junction with carriers. The DC voltage was generated by two Keithley 2635B source measurement units (SMUs). The gate electrode is subjected to an harmonic AC voltage during read-out. Furthermore, a DC bias is applied in series. During a CV sweep the DC voltage is changed, while the AC voltage is kept constant. Both voltages (AC and DC bias) are generated by an Agilent 33500B function generator. Usual frequencies were in the range of 1 kHz to 10 kHz. For a normal CV sweep a lock-in amplifier SR830 from Stanford Research Instruments was connected to the read-out electrode and locked to the signal of the function generator. For displacement current measurements a current pre-amplifier SR570 from Stanford Research Instruments was connected to the read-out electrode and the current was visualized by a DSO5052A oscilloscope.

During the write operations voltage pulses were applied to the gate electrode, while the p+ and n+ pads were grounded. The current on the read-out electrode was ignored or it was not connected. Different write modes were tested, namely: pulse number modulation, pulse length modulation and pulse height modulation, as shown in Fig. 5.1b)-d). Between each pulse a read-out with the harmonic AC voltage was performed. The instruments were controlled by Labview.

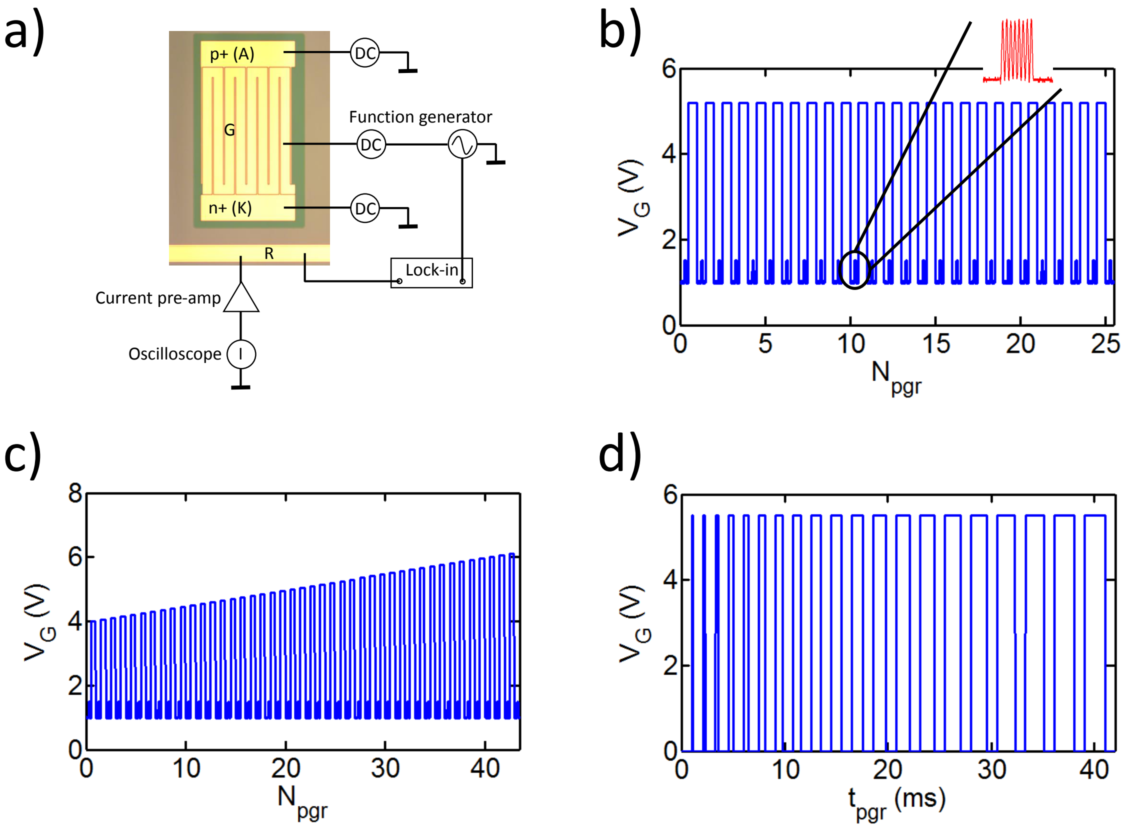


Figure 5.1: a) Measurement setup for single device measurements. Applied write pulse and read-out sine wave for successive programming with: b) pulse number modulation (read-out sine wave in the inset) c) pulse height modulation and d) pulse length modulation. Material from [8]

5.2 Printed Circuit Board for Crossbar Measurement

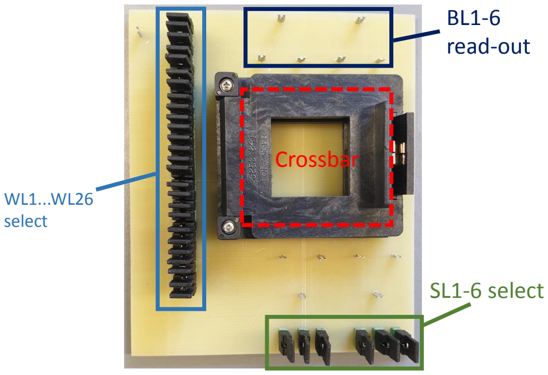


Figure 5.2: a) Test-PCB for selecting single Wordlines, Shielding lines and Bitlines. The PCB is mainly used for CV measurements.

For crossbar measurements many wordlines need to be activated at the same time and many bitlines need to be read-out at once. For this purpose firstly a PCB shown in Fig.

5.2 was designed, in which single wordlines can be connected, but only single bitlines can be measured with e.g. a lock-in amplification. This PCB was only used for quick tests.

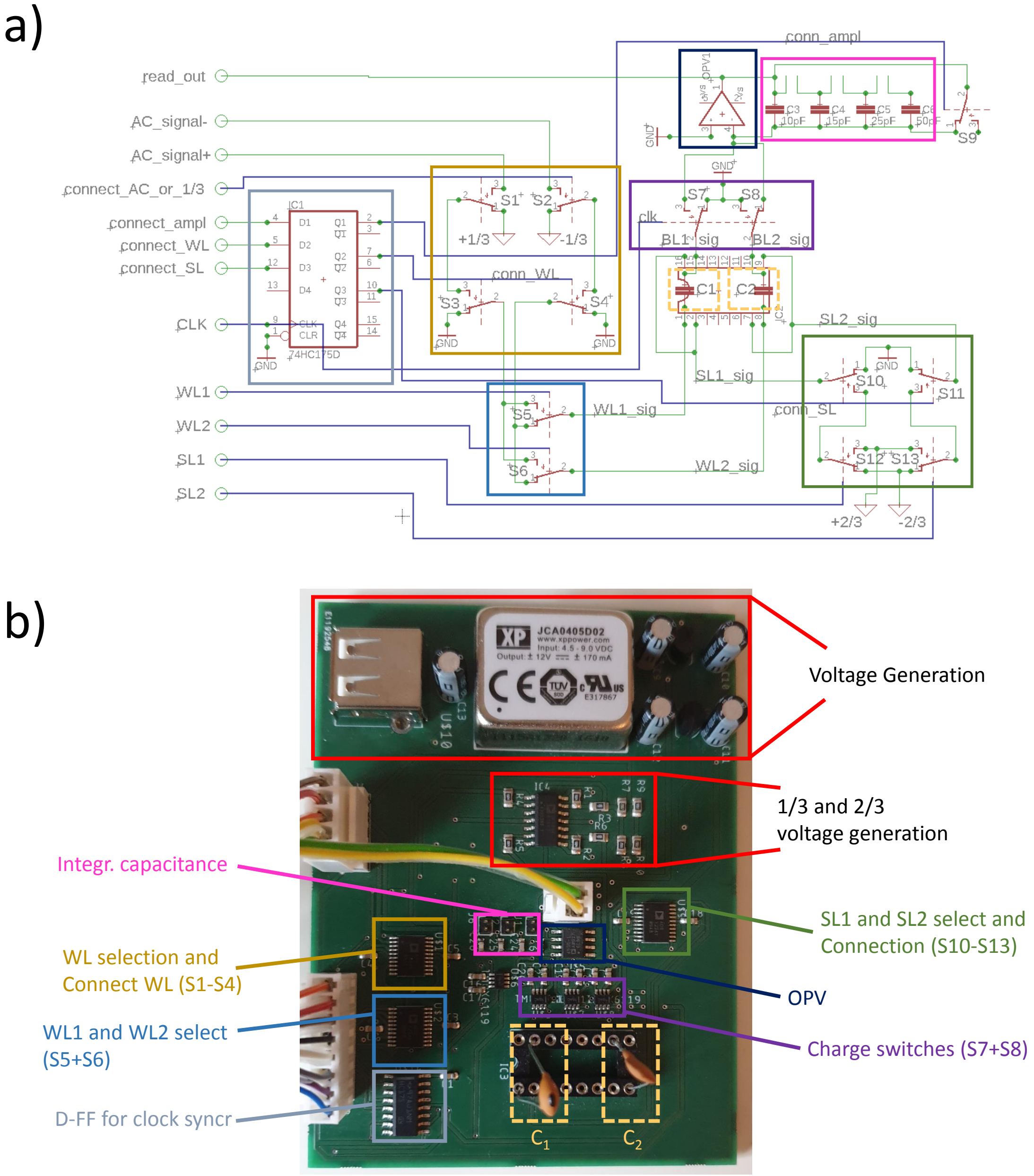


Figure 5.3: a) Test circuit for measuring two capacitances (C1 and C2), which are subtracted from each other. b) Corresponding picture of the PCB with the different circuit blocks as framed in a).

Furthermore, in a second version a DAQ was used and sense amplifiers for every differential bitline were integrated onto a PCB. A charge integration method is usually used for measuring capacitances. A switched capacitor approach was chosen for the implementation. With every input period number a new integration cycle is started.

An operational amplifier with a feed back capacitor is used for charge integration and the integration capacitance is charged up with every half period of input read-out. The general working principle of the circuit was first tested on smaller scale with two conventional capacitors. The circuit is shown in Fig. 5.3a) with a corresponding picture of the PCB in Fig. 5.3b).

The circuit generally consists out of analog multiplexers for the WL signal selection, for the charge seperation of the read-out (BL) and for the SL signal selection. Furthermore, there is a voltage transformer for the 1/3 and 2/3 voltage generation, which is used during writing. These voltages are applied to the WL and SL respectively. The first multiplexer (S1-S2) is used to select either an AC signal, which is usually used for reading, or the 1/3 voltage (for writing). The out coming signal is applied to the WL or the WLs are connected to ground (S3-S4). In order to always apply a full period the ’connect WL’ signal is controlled by a D-FlipFlop, whereas the clock signal is in phase to the AC signal. The positive and negative signal is applied to the WL selectors (S5 and S6), which chose to apply either the positive or negative signal (black or white pixel in image recognition). For the SL either a -2/3 or 2/3 signal can be applied, or they are connected to ground (during reading). This is accomplished with the switches S10-S13.

The scheme for read-out is shown in Fig. 5.4: Generally, four cases need to be considered: positive input signal + positive weight; negative input signal + positive weight; positive input signal + negative weight and negative input signal + negative weight. These four cases should enable a ’four-quadrant multiplication’. Furthermore, a differential weight topology is used, where one memory cell/capacitor is meant to be of a positive type and another memory cell/capacitor as a negative type. The amount of the input signal (applied to the WL) is encoded as the length of the sine wave or as number of periods, where negative and positive signals are 180° phase shifted. For the positive input signal the clock signal is defined high for the rising edge (the WL2 signal in Fig. 5.4). For the negative input signal the clock is defined high for the falling edge (the WL1 signal in Fig. 5.4). The clock controls the state of the two switches which are connected between the memory cell/capacitors and the amplifier. In Fig. 5.4 two states are shown (the low clock signal on the left side and the high clock signal on the right side, as indicated by the purple point). For a low clock signal the switches are in the right position and hence the C+ capacitance is connected to the amplifier, while the C- capacitance is connected to ground. Since the WL2 signal is falling during the zero clock signal (positive signal) the capacitance is charged up negatively and so also is the integration capacitance of the amplifier (Cint). Due to the inverting nature of the amplifier, a positive output voltage is obtained, as desired (positive input signal + positive capacitance). The negative capacitance C- is charged up positively at the same time, due to the rising signal at zero clock (negative input), but all the current flows to ground. In case of a positive clock signal (right side in Fig. 5.4) the C- is charged up negatively and the capacitance is now connected to the amplifier. Due to the inverting nature the amplifier a positive output voltage is again obtained, as desired, since negative input signal (WL1) and negative memory cell/capacitance should give a positive result. The other two mentioned cases are obtained similarly, thus two simple switches are sufficient to obtain a 'four-quadrant multiplication'. Furthermore, only one period is shown in Fig. 5.4, but for more periods the integration capacitor is charged up more and more, thus the amount of output voltage increases with the period number. The amplfiers have a very low input impedance due to the virtual ground of the **OPV!** (**OPV!**).

For the analog multiplexers of the WL and SL a ADG1234 from Analog Devices was used. For the amplifier/charge switches a TMUX6119 from Texas Instruments was used and the **OPV!**s were a ADA4530 from Analog Devices. For the analog multiplexers an ultra low charge injection is very important in order to ensure low disturb levels from the clock signal.

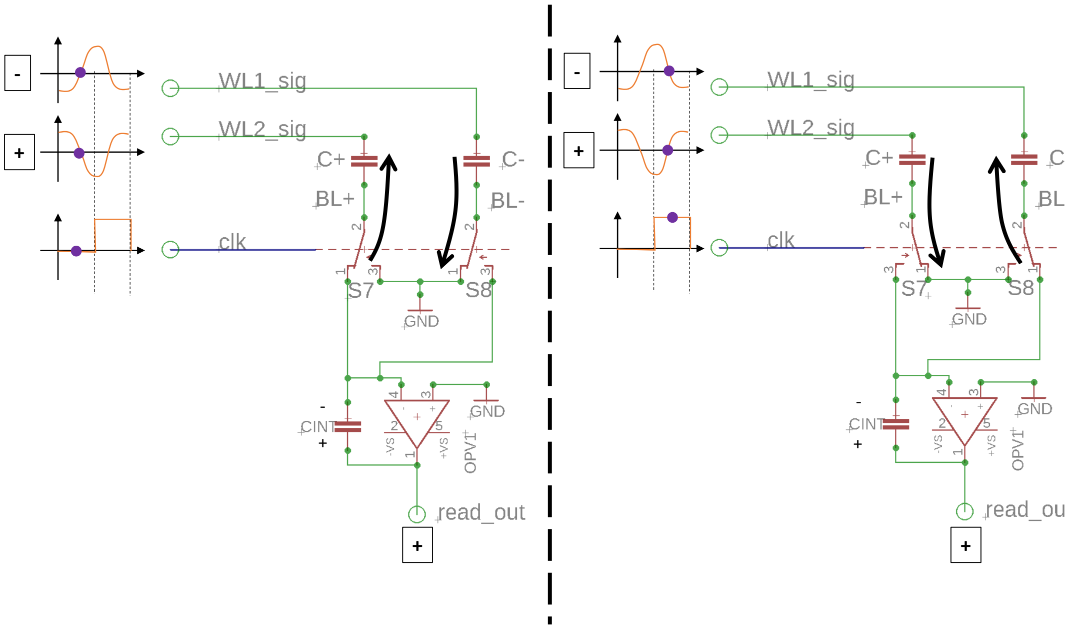


Figure 5.4: Switched capacitor approach for read-out. A negative input signal is applied to WL1, while a positive one is applied to WL2. The positive and negative measurement capacitance (C- and C+) is either connected to ground or with the amplifier by switches, which are controlled by the clock signal. The state of the clock signal is indicated by the purple point (low for the right side and high for the left side). The arrows indicate the current flow direction.

Measurement results of the test capacitor circuit are shown in Fig. 5.5, where the blue lines are cases without any test capacitor (C1 and C2). A total of 5 periods is applied as input signal. Even without a measurement capacitor steps are visible, which are caused by the charge injection of the clock signal in the switches. The measured signal with capacitors including charge injection are the green lines, and the red lines are the relevant lines after subtracting the charge injection curves. Generally, the four cases of a positive and negative input signal for a positive capacitance (a+b) and a negative capacitance (c+d) are shown. The period length of 1 ms is indicated by the dashed/dotted line, while the half period is indicated by the dotted line. The charge integration only happens over a half period, and for a positive capacitance the integration on the amplifier happens at the beginning of a new period (Fig. 5.5a+b). For the negative capacitance the integration happens at the end of a period (Fig. 5.5c+d).

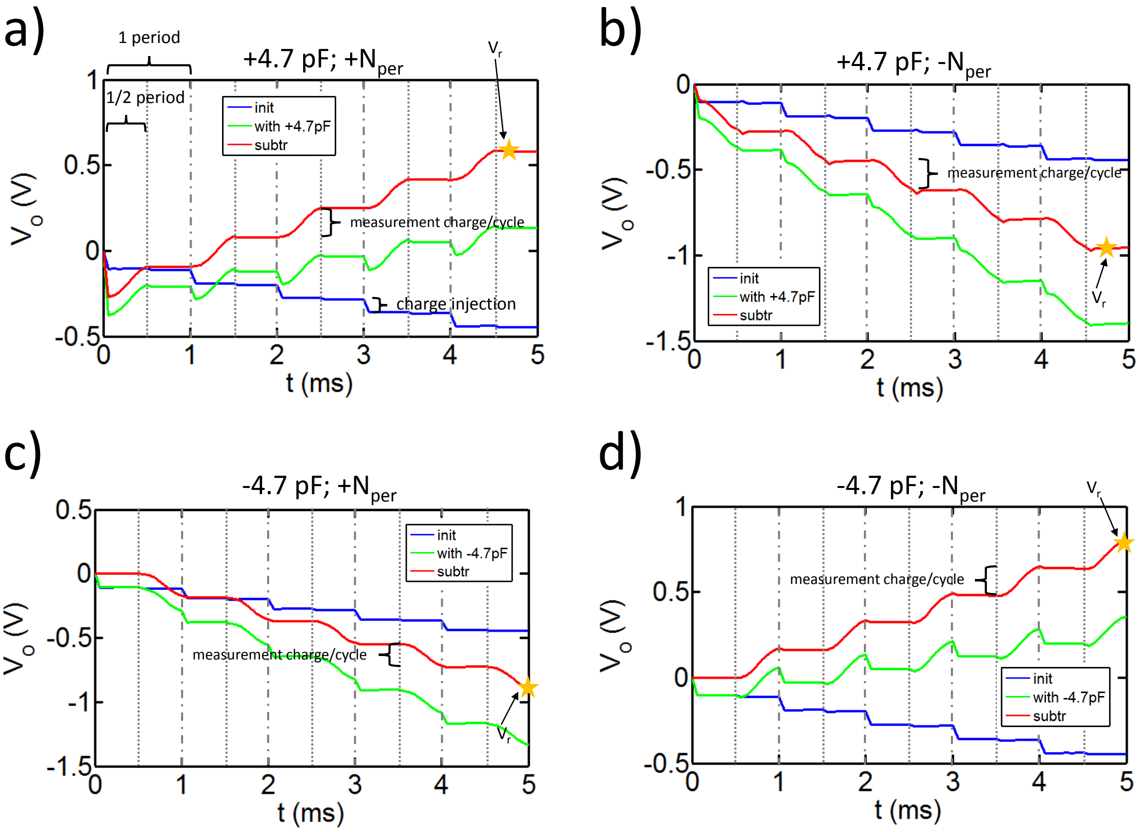


Figure 5.5: Output voltage V_o versus time for a) positive measurement capacitor (4.7 pF) + positive input signal b) positive measurement capacitor (4.7 pF) + negative input signal c) negative measurement capacitor (4.7 pF) + positive input signal d) negative measurement capacitor (4.7 pF) + negative input signal. The blue line is the charge injection background (init), the green line the result with capacitor, as obtained and the red line the subtracted charge injection from the measurement (green line minus blue line). The period and half period are indicated by the dashed/dotted and dotted line and furthermore the resulting voltage V_r is indicated by the star.

Fig. 5.6a) reveals the resulting voltage after different period numbers (N_{per}). The resulting output voltage is indicated by the star in Fig. 5.5. Negative period number inputs (N_{per}) are 180° phase shifted to the positive ones, as explained above. From the output voltage the total injected charge can be calculated from the integration capacitance, which was 10 pF for the 0.25 pF and 1 pF measurement capacitor (for the 2.2 pF and 4.7 pF measurement capacitor it was 25 pF), which is shown in Fig. 5.6b). The slope of the curves is shown in Fig. 5.6c) and it is close to the theoretical value. From these measurements one can conclude that a linear 'four-quadrant multiplication' is feasible with the switched capacitor approach and a measurement resolution down to 100 fF-250 fF is possible. The linearity is an important feature for neuromorphic systems in the weight multiplication.

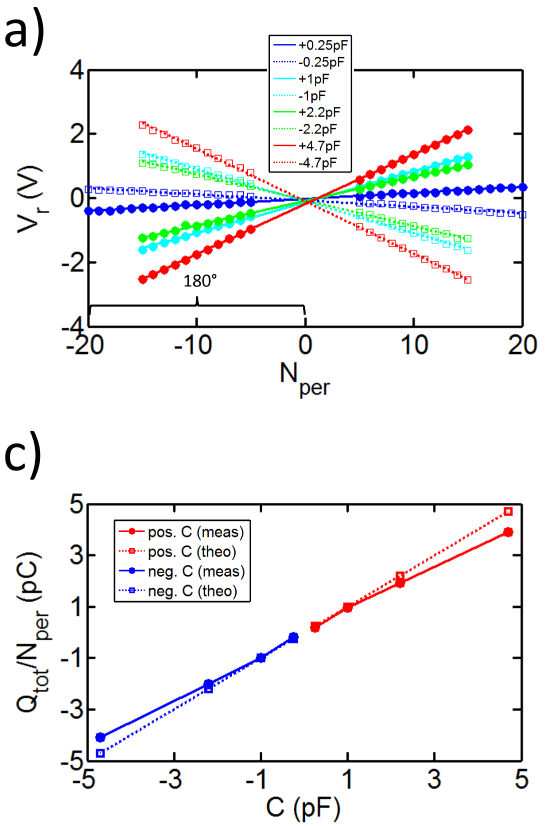


Figure 5.6: a) Total resulting voltage for different measurement capacitors and input period numbers. b) the total charge is calculated from a) and the integration capacitance. c) Slope of lines in b) versus measurement capacitors, compared to the theoretical value of charge integration.

For the crossbar chips another PCB was developed, which is similar to the one in Fig. 5.3. In this case there are 26 wordlines (5x5 image + 1 bias), six bitlines and shielding lines to be considered (Fig. 5.7a). Hence there are 26 WL selection switches and 6 SL selection units. Since each operational amplifier is used for two differential BLs, there are three operational amplifiers. A picture of the PCB is shown in Fig. 5.7b).

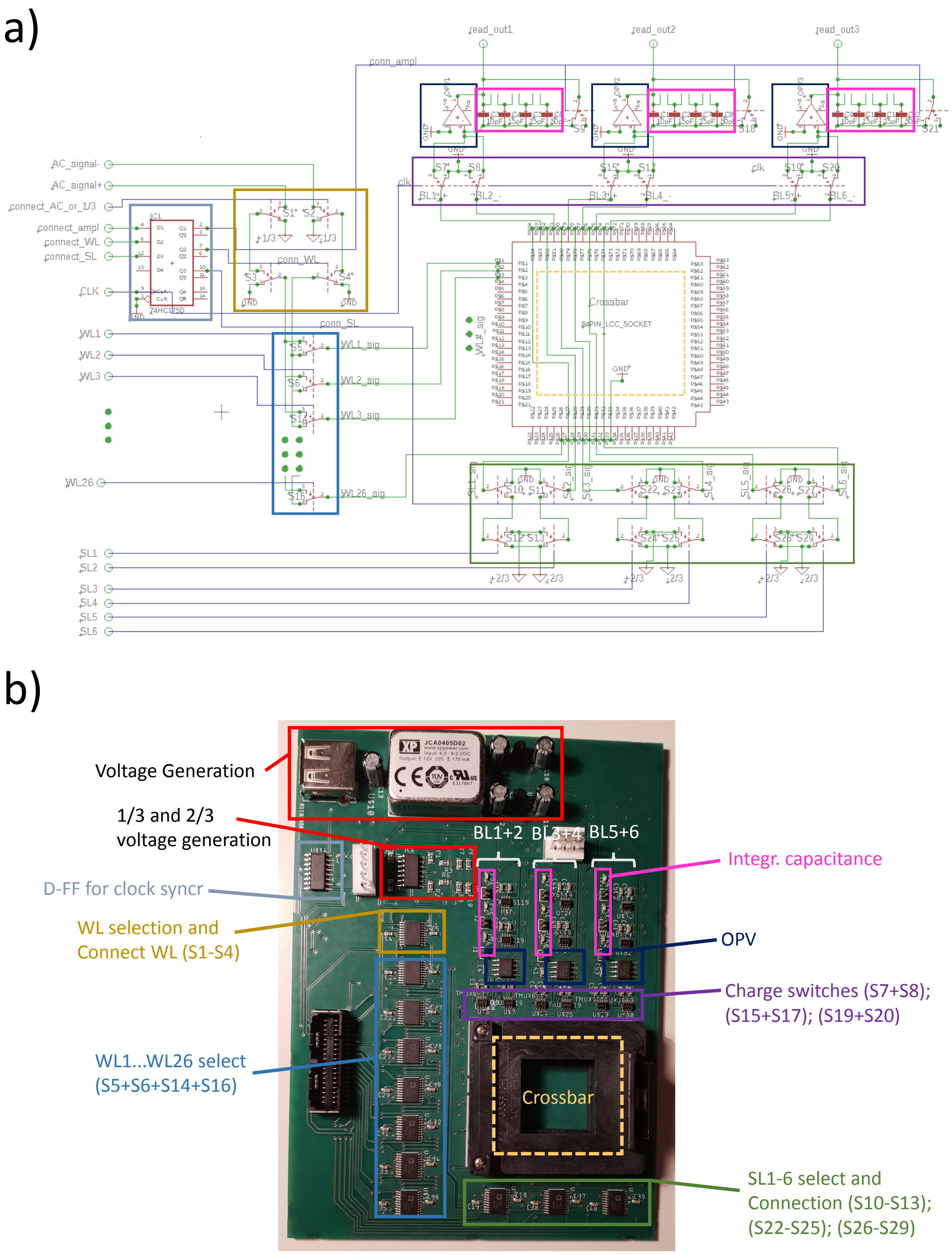


Figure 5.7: a) Circuit for the crossbar measurement with 26 wordlines. b) PCB for crossbar measurement. The crossbar chip is inserted into the black socket.

6 Measurement Results and Discussion

In this chapter the results obtained with the measurement setups described in chapter 5 are presented and discussed. In the first sub-section, some results on single devices with normal thermal oxide and no memory effect are shown. Thereafter, different interface oxides are compared for single devices, and more detailed measurement on the single device with the best interface oxide are presented in the subsequent chapter. The final chapters include crossbar measurements and the performed algorithm on it.

6.1 Capacitive Coupling Curves for Single Devices with no Memory Effect

A general coupling curve has already been shown in Fig. 4.4b) in chapter 4.1. This curve had a poor on/off ratio (or dynamic swing) due to insufficient charge injection in the shielding layer. Therefore, metal lines along the p+ and n+ fingers of the structure in Fig. 4.5 were created and higher doping concentrations for the p+ and n+ regions were used.

Fig. 6.1a) shows the obtained CV curves measured by the setup in Fig. 5.1a). The p+ and n+ contacts were subjected to different DC-voltages during the CV sweep in order to deplete or enhance the pin-junction. The voltages were antisymmetric, so e.g. $V_{AK} = -4\,\mathrm{V}$ means that $2\,\mathrm{V}$ was applied to the n+ region and $-2\,\mathrm{V}$ to the p+ regions. This ensures that the voltage within the intrinsic or slightly doped region is still near to zero, due to the symmetry of the junction. This means that the CV curves should stay nearly centred at the same position, as observed in Fig. 6.1a). As can be expected from the discussion on theory in chapter 3.2 the CV curves get broader for depletion and can be virtually switched off for the forward direction. The remaining two small peaks for the forward direction are attributed to the recombination of the carriers within the rather long intrinsic/slightly doped region ($15\,\mu\mathrm{m}$), with the result that not the entire channel length has shielding charge. This explanation remains an assumption.

The curves in Fig. 6.1a) are the derivatives of the sigmoid/tanh electric field coupling curves. In order to obtain the sigmoid/tanh curves, an AC voltage with increasing peak-to-peak amplitude (V_{AC}) and fixed maximum voltage to the right side in Fig. 6.1a) is applied, while the effective current (I_{read}) value is read-out. The obtained sigmoid curve is shown in Fig. 6.1b). The effective current value depends on the surface, which is covered by the AC signal under the CV curve (shaded area in Fig. 6.1a). So the measurement method is basically an integration.

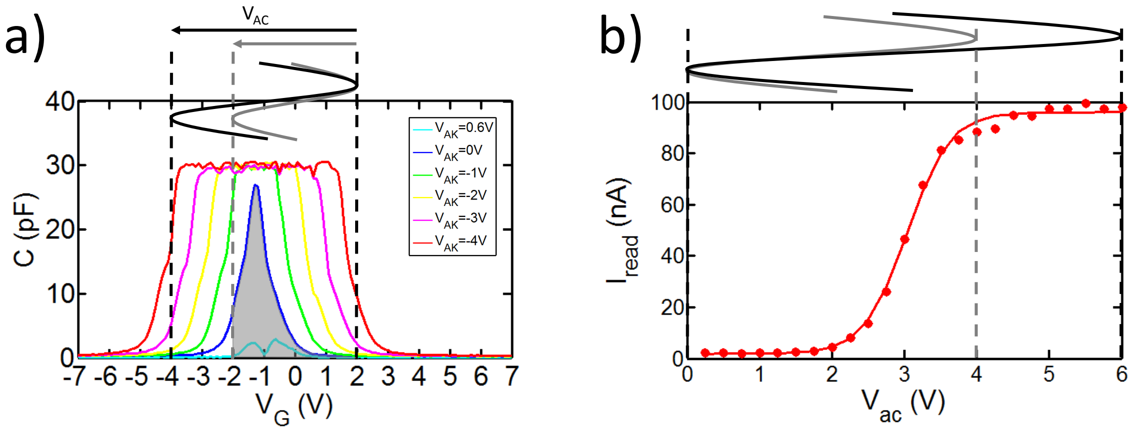


Figure 6.1: CV curves (a) for different depletion and forward voltages on the pin-junction (V_{AK}). The applied AC voltage levels to obtain the sigmoid curve in b) are indicated. Material from [8]

As mentioned earlier, the sigmoid transfer functions play an important role in modeling neurons in artificial neural networks, and furthermore, adjusting their broadness by application of reverse or forward voltages provides additional functionality. A common practice during training is the dropout of certain neurons to avoid over fitting [115], which can be implemented through the forward direction of the diode. Furthermore, the application of the voltage V_{AK} can also enable selection and deselection of memory cells, if necessary. Also, the high ratio between maximum and minimum capacitance becomes obvious (1:100 in this case).

6.2 Comparison of Single Devices with Memory Effect and Different Interface Oxides

The best-suited interface oxides mentioned in chapter 4.2 for MFOS capacitors were tested on single devices, namely: SC1 5:1:1, SC1 10000:1:1, HF etched, 650 °C and 750 °C thermal oxide.

Fig. 6.2 shows a comparison of the obtained CV curves (a), the pulse number modulated writing and erasing (b) and the data retention (c). The scheme for pulse number modulation is shown in Fig. 5.1b) in chapter 5.1.

From the CV curves it becomes visible that all memories were in a charge trapping regime, due to the shifting direction of the forward (-5 V to 5 V) and backward (5 V to -5 V) sweep (from the left to the right). No ferroelectric behaviour was obtained. All the CV curves were relatively broad compared to the 750 °C thermal oxide, which is an indication of large interface trap state density. The reason for the second small side peak in some curves is unknown. Generally, the most stable and best charge trapping CV curves were obtained for the 750 °C thermal oxide device.

With regard to the analog write performance for all devices the typical exponential LTP and LTD were measured (Fig. 6.2b):

$$C_{LTP} = C_{min} + \Delta C \cdot \left(1 - \exp\left(\frac{-N_{pgr}}{\beta_{pgr}}\right)\right) \tag{6.2.1}$$

$$C_{LTD} = C_{min} - \Delta C \cdot \left(1 - \exp\left(\frac{-N_{er}}{\beta_{er}}\right)\right) \tag{6.2.2}$$

with the number of programming or erase pulses N_{pgr} and N_{er}, the stretching factor β_{pgr}, β_{er} and the C_{min} for minimum and C_{max} for maximum capacitance. The ΔC describes the maximum change in capacitance.

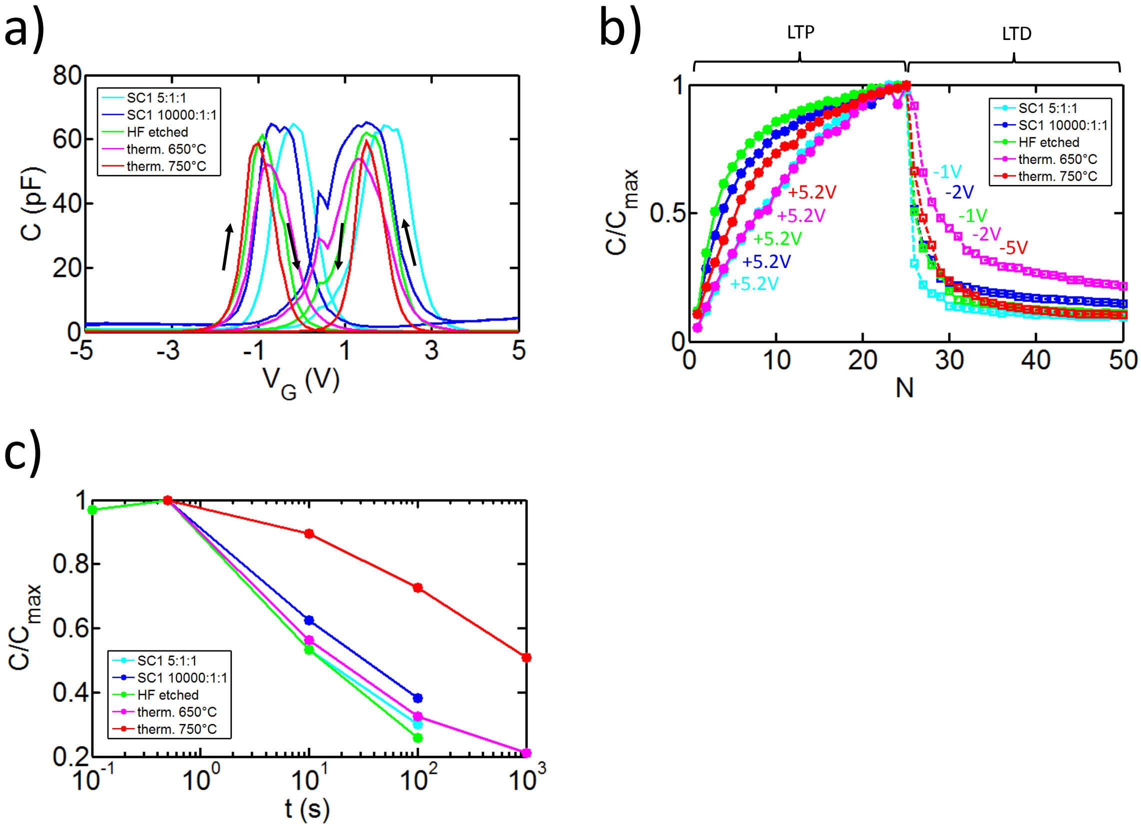


Figure 6.2: a) CV curves for different interface oxide. b) Pulse number modulation of multi-level switching for different interface oxides. c) Retention of different interface oxides. Material from [8]

From Fig. 6.2b) one can conclude an asymmetry for the LTP and LTD curve. More importantly, the asymmetry of write and erase pulse height is much more pronounced for the thinner interface oxide thicknesses (−1 V to −2 V for erase in the case of SC1 5:1:1, SC1 10000:1:1, HF etched and 650 °C samples, compared to −5 V for the 750 °C thermal oxide sample). In summary, very small negative voltages are sufficient to reject

the trapped electron charge from the HZO layer, which is an indication of very shallow traps. The thicker interface oxide of the 750 °C sample reduces this fast de-trapping. This is also evidenced by the retention curves, as shown in Fig. 6.2c). The 750 °C sample shows much longer retention times and thus slower de-trapping of electrons. As mentioned in chapter 4.2, it is expected that the polarization charge of the ferroelectric layer assists the charge trapping behaviour [151].
From these measurements the decision was made to continue measurements on single devices and crossbar devices with 750 °C interface oxides and no longer consider the other interface oxides any more.

6.3 Single Device Measurements

The chosen device with 750 °C interface oxide was measured in more detail. Fig. 6.3a) shows the CV curve for ten different devices and a very stable memory window of 2.66 V ± 0.017 V was obtained. The corresponding pulse number modulation curves for the ten different devices are shown in 6.3b), where some deviation was visible, but it was good enough to later implement a simple learning algorithm on a crossbar (see chapter 6.4). The on/off ratio can be determined from the measured displacement current (Fig. 6.3c)+d)). It turns out to be 1:1478, which is consistent with the theoretical TCAD simulation for micron scaled gate length (Fig. 3.4a). The applied read-out sine wave for displacement current measurement was between 1 V and 1.5 V, where the shaded area in Fig. 6.3a) is proportional to the measured current in the written state (Fig. 6.3c).

Fig. 6.4 shows the measured different analog write/erase modes: pulse number modulation, pulse length modulation, pulse height modulation. The applied voltage pulses are in accordance to Fig. 5.1. The pulse number modulation in Fig. 6.4a) and b) shows the prior mentioned exponential LTP and LTD curves, and the saturation value and steepness of the exponential increase can be adjusted by the used pulse height.
For the pulse length modulation (Fig. 6.4c), similar curves compared to the pulse number modulation, are measured. The LTD curve is slightly more symmetric to the LTP curve for the pulse length modulation.

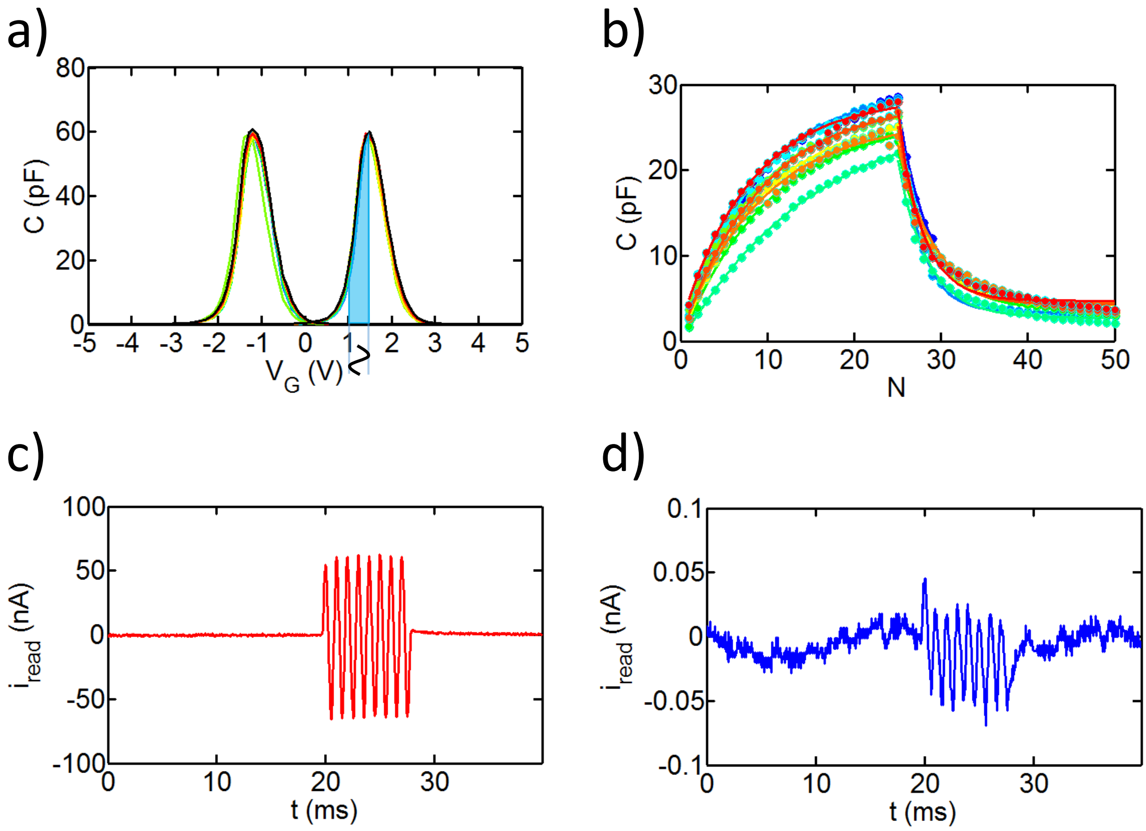


Figure 6.3: a) CV curves for 10 devices. The applied sine wave is used for displacement current measurement in c. The current is proportional to the shaded area. b) Pulse number modulation for 10 devices. c) Read-out current for the written state. d) Read-out current for the erased state. Material from [8]

With regard to the pulse height modulation, a much more symmetric and linear behaviour was obtained. This is highly beneficial for implementing neuromorphic algorithms [1].

Generally, the resulting curves exhibit some similarities to those obtained from pure ferroelectric switching [82], indicating the ferroelectric assistance in the memory storage process.

The measured read-out current during the pulse height modulation is shown in Fig. 6.5 and the pinch-off of the current due to screening is clearly observed. The capacitances from Fig. 6.4 are calculated from the effective read-out current as follows:

$$C = \frac{1}{V \cdot T_{per}} \sqrt{\int_{t}^{t+T_{per}} i(t)^2} \tag{6.3.1}$$

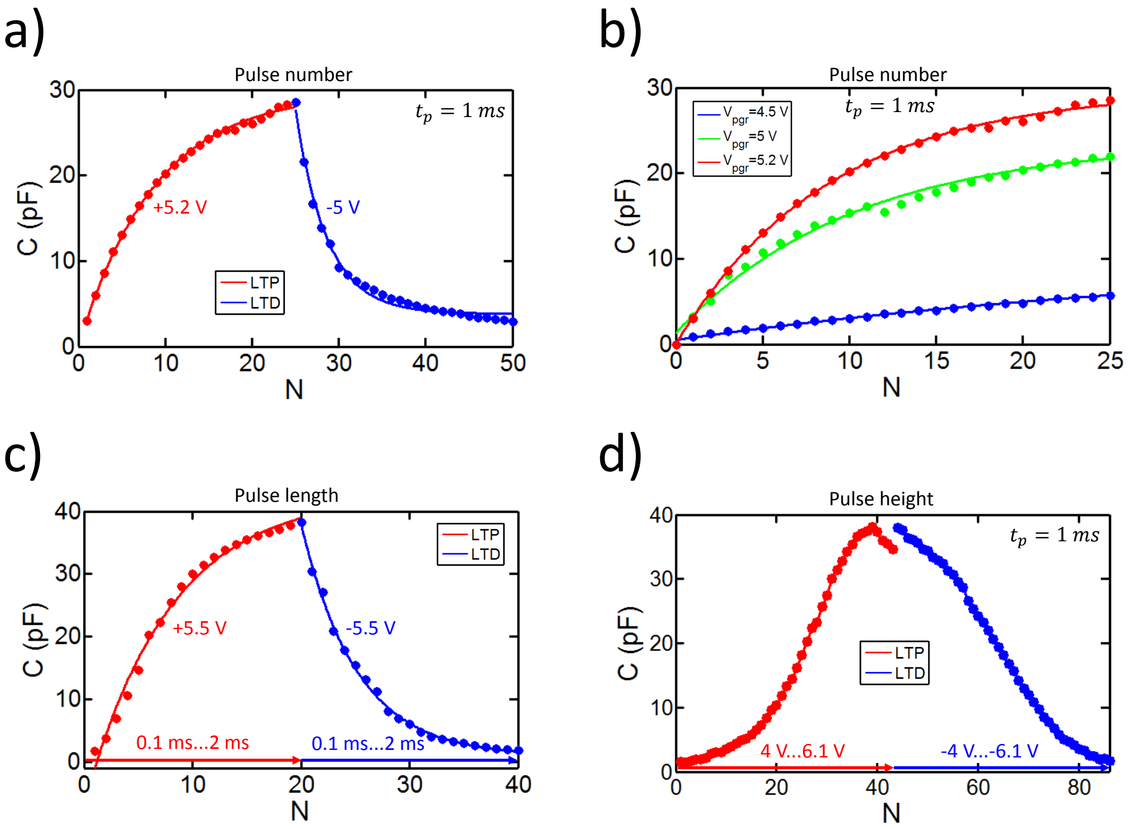


Figure 6.4: a) Pulse number modulation for a single device. b) Pulse number modulation for different pulse heights. c) Pulse length modulation and d) Pulse height modulation. Material from [8]

The retention for different stored levels (in this case read-out current) is shown in Fig. 6.6a), and an increased retention for lower capacitance values is observed. With regard to the endurance, a broadening of the CV curves is measured for 1E+8 cycles (Fig. 6.6b), which indicates a decreased quality of the interface states. For 1E+5 cycles a reduced memory window is measured. The total endurance is estimated to be in the range of 1E+5 to 1E+6 cycles for abrupt writing/erasing events as used in the measurement of Fig. 6.6b). This value is consistent to other charge trap memories and FeFETs.

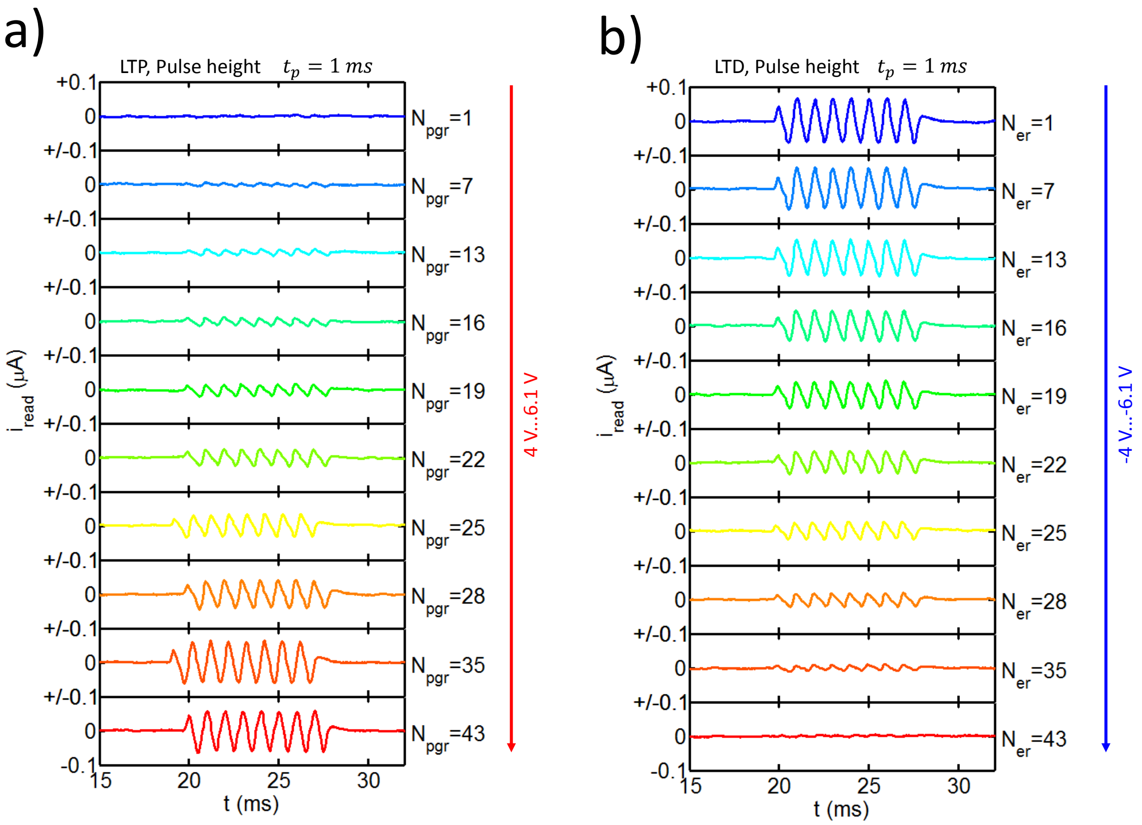


Figure 6.5: Measured read-out current for the pulse height modulation for different pulse numbers, during the a) LTP and b) LTD. Material from [8]

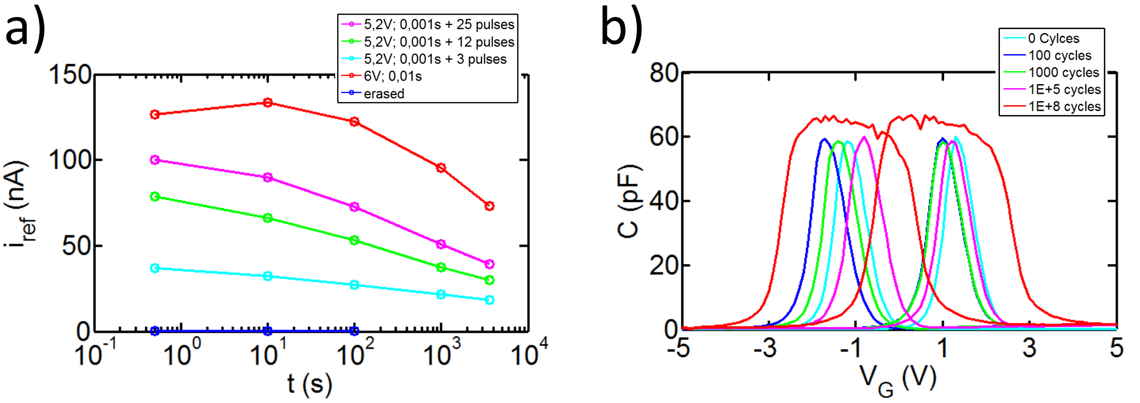


Figure 6.6: a) Retention of different stored capacitance values (here read-out current value) b) Endurance characteristics of the CV curves.

6.4 Crossbar Measurements

As mentioned earlier, the crossbar consisted of 26 wordlines and 6 bitlines, where a differential weight topology was used. Therefore, a 5x5 image with a bias (=26 inputs) can be classified into three output classes (6/2).

The fabricated crossbars (Fig. 4.38) were first tested with regards to CV curves with the PCB shown in Fig. 5.2. Also, CV curves from a normal probe station were measured and compared (Fig. 6.7a). The measurements in Fig. 6.7a) were performed on a single memory cell, which means that the AC signal was applied to one wordline, while one bitline was measured. The maximum capacitance of a single memory cell is at 300 fF, while there is a huge background for the wire bonded chip, which is caused by the parasitic capacitances of the wire. The conclusion is that there is minimum capacitance for a single memory cell if the sense amplifier is off-chip, like in this case. Smaller capacitances may become very difficult to detect off-chip due to parasitic capacitances.

The LTP and LTD curves for the wire-bonded chip for one bitline, where all wordlines are connected to the same AC signal, is shown in Fig. 6.7b). The parasitic background capacitance also becomes visible, but generally the curves are comparable to Fig. 6.4a). An important feature for writing a crossbar is the disturb of the memory cell to half selected cells. During programming some cells are subjected to a disturb voltage, where 1/3 of the programming voltage is the minimum possible voltage. Fig. 6.7c) shows the overwriting of erased (red curve) or written (cyan and blue curve) cells with respect to a certain positive (2 V) or negative (−1.5 V,−2 V) disturb voltages. If the programming voltage is 5 V, the disturb level will be 1.67 V. As can be seen, the disturbance of the erased cell for small positive voltages is nearly zero, whereas there is some disturbance for the written cell and negative voltages. The decay might also be caused by the limited retention, but generally overwriting at these small voltages is much smaller compared to other interface oxides (see Fig. 6.2b). The disturb is low enough for implementing a simple algorithm, as described in the next sub-chapter.

Furthermore, the 'four-quadrant multiplication' was tested, as shown in Fig. 6.8: The input period number (N_{per}) and the number of programming pulses (N_{pgr}), which adjust the actual weight, were varied in positive and negative values, while the output voltage is read-out. For a positive period number (N_{per}) all wordlines were activated with no phase shift, and for a negative input signals all wordlines were activated by a 180° phase shifted sine signal. For positive programming pulses (N_{pgr}) the positive BL was changed, while the negative BL was kept in an erased state (vice versa for negative programming pulses). This measurement was done with the PCB in Fig. 5.7.

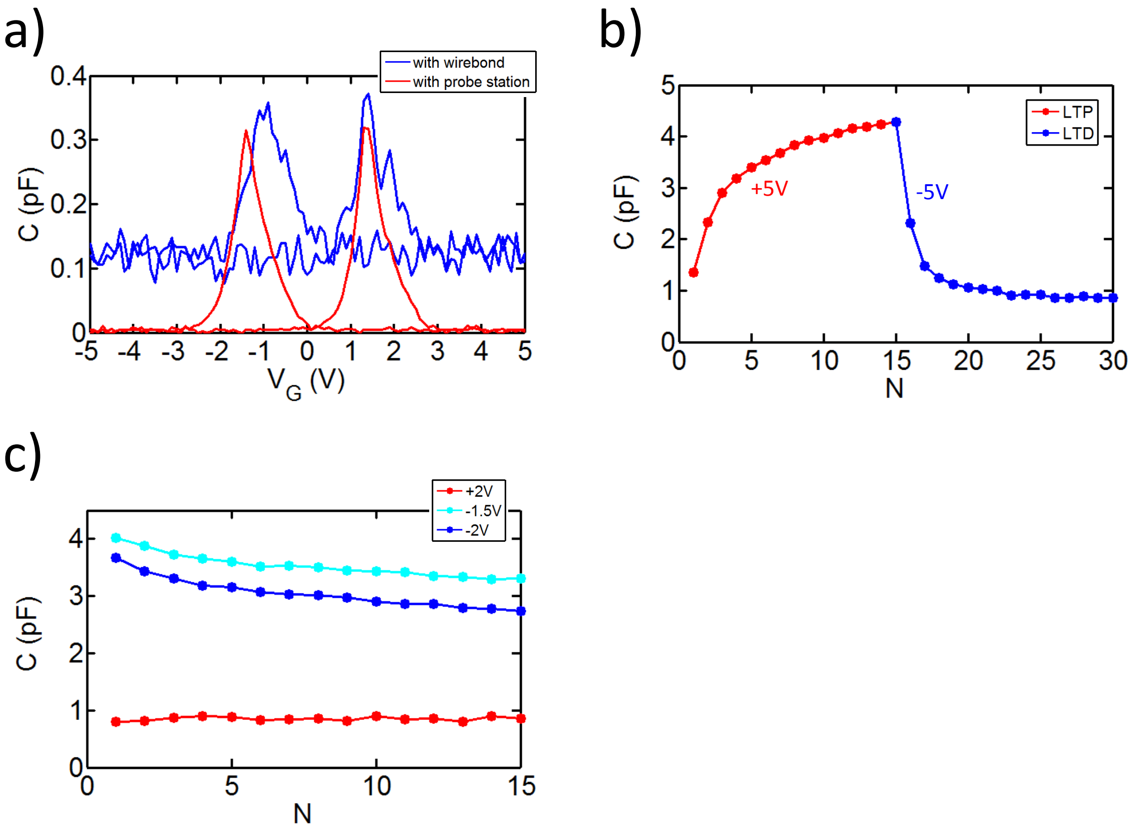


Figure 6.7: a) CV curve for wire-bonded or probe station measured single memory cell. b) Pulse number modulation for wire-bonded crossbar chip along one BL with all WL being activated. c) Disturb of an erased (red line) and written (cyan and blue line) memory cell with a small voltage.

As can be seen in Fig. 6.8a) the output voltage is negative, if either N_{per} or N_{pgr} is negative. If both are positive or negative, a positive output voltage is realized. Fig. 6.8b) shows a cross section parallel to the programming pulse number, N_{pgr}, and the same LTP curves as before are obtained. Contrary, Fig. 6.8c) is a cross section parallel to the input period number, N_{per}, and a desired linear behaviour, as before (Fig. 5.6a), is observed. So Fig. 6.8a)-c) prove the 'four-quadrant multiplication' on a crossbar. Furthermore, in Fig. 6.8d) the number of positive wordlines is varied, and a linear accumulation operation is proven by this, since the signals at each cross-section of the bitline is summed up and the input at each cell is determined by the input signal to the WL.

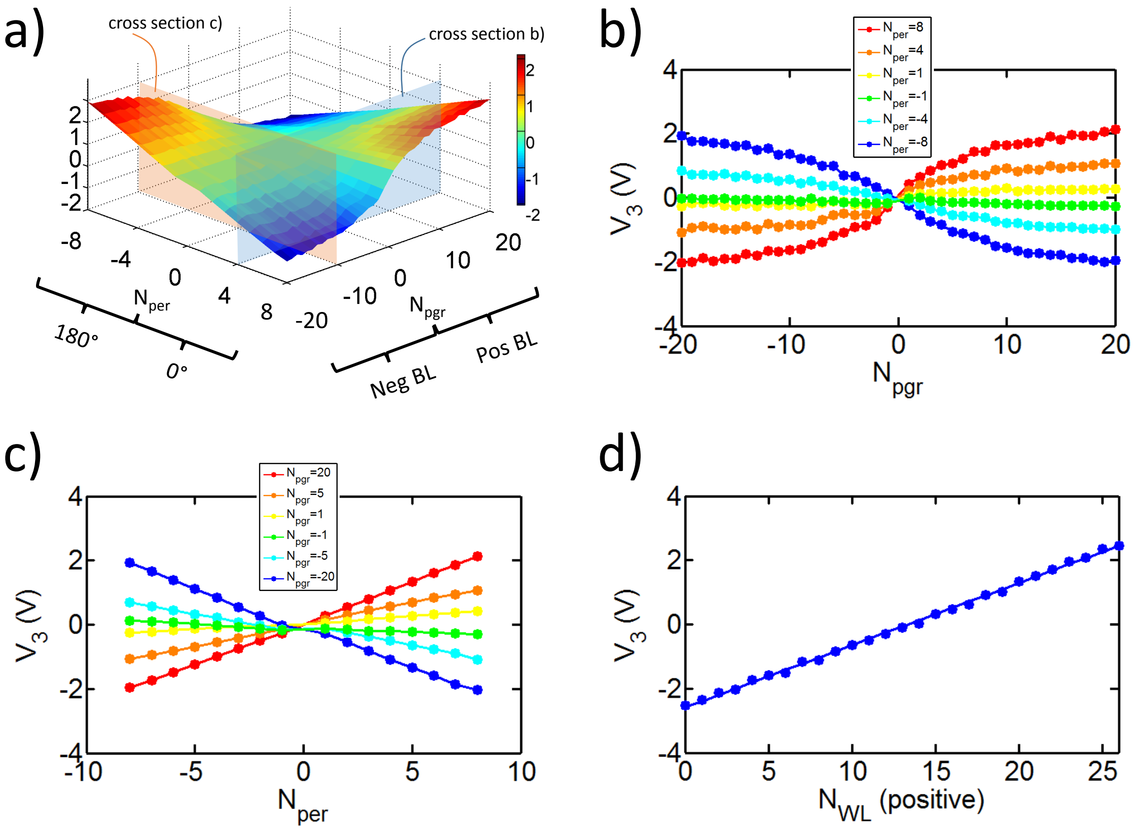


Figure 6.8: a) 3D plot of 'four-quadrant multiplication' with varying the input period number (N_{per}) and the programming pulse number (N_{pgr}). The output voltage is the z-axis. Cross-sections in parallel to the N_{pgr} and N_{per} axis are shown in b) and c). d) The number of positive wordlines is varied, while all non-positive WL are negative, thus showing the linear accumulation operation. Material from [8]

6.5 Neuromorphic Algorithm on Crossbar

In the last chapter a linear multiply-accumulation operation was verified on the crossbar, thus a neural network can be implemented on it. The first 25 wordlines enable a vectorized input feature map for images of 5x5 pixels, thus one single fully-connected layer is carried out. Dark pixels are represented by positive values and bright pixels by negative values. The bias input is mapped to the 26th Wordline.
Regarding the implemented training algorithm, Manhatten update rule [75, 173] was chosen, due to its simplified training procedure (see chapter 2.1.3). The weight update is described by the following equation:

$$\Delta W_{ij} = \operatorname{sgn}\left(-\alpha \cdot \delta_j(n) \cdot X_i(n)\right) \tag{6.5.1}$$

With α the learning rate, $\delta_j(n)$ the backpropagated error on the BL and $X_i(n)$ the actual input to the wordline. The weight update is positive if either the error and the WL input are both positive or negative. Thus this relationship can be described by a XNOR operation and pulse scheme, which is applied to the shielding line and the wordline (Fig. 6.9a). The differential signal that is applied to each memory cell in Fig. 6.9a) describes exactly the XNOR operation. The programming is taking place with the same amplitude, thus a pulse number modulation is used during training. Also, the disturb level voltages of 1/3 of the programming voltages are visible, which are low enough to effectively prevent overwriting of cells in the same column or row (the memory cell acts as selector itself, Fig. 6.7c).

As a 5x5 image recognition task, the letters M, P and I were chosen and one pixel in each of the samples was flipped, which results in a total set of 78 samples. These pseudo-images were separated into a test and training set, whereas the test images are indicated by a blue frame in Fig. 6.9b). The resulting number of mis-classified images versus training epochs for the training and test images is shown in Fig. 6.10a). As can be seen from the figure, the number decreases rapidly after one training epoch and stays almost zero throughout the training epochs. Mis-classifications after epoch 1 are caused by the very similar expected value for individual pre-synaptic neurons for features M and P. Measurements also confirm the more stable results for classification of I, as shown in Fig. 6.10d). Fig. 6.10b)-d) shows the obtained mean neuron activations for the three classifications over the training epochs. The results are in accordance with other studies [58, 75] and were confirmed by computer simulations.

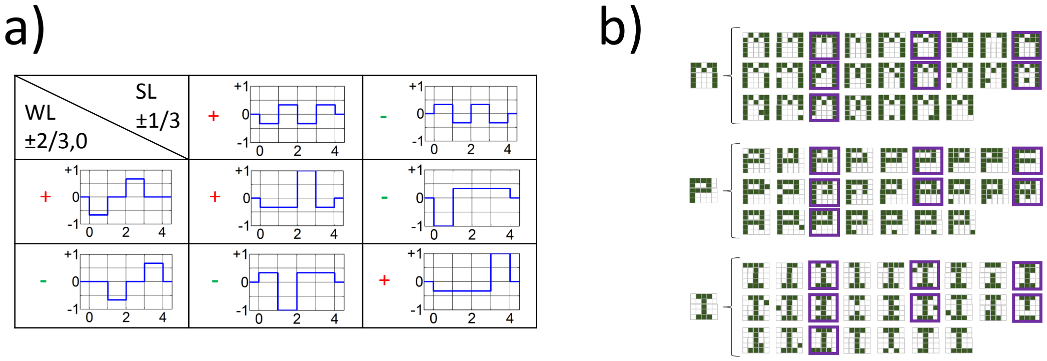


Figure 6.9: a) Pulse scheme for implementation of XNOR operation of the Manhatten update rule. b) Used training and test image set of the letters M,P and I. The test images are framed purple. Material from [8]

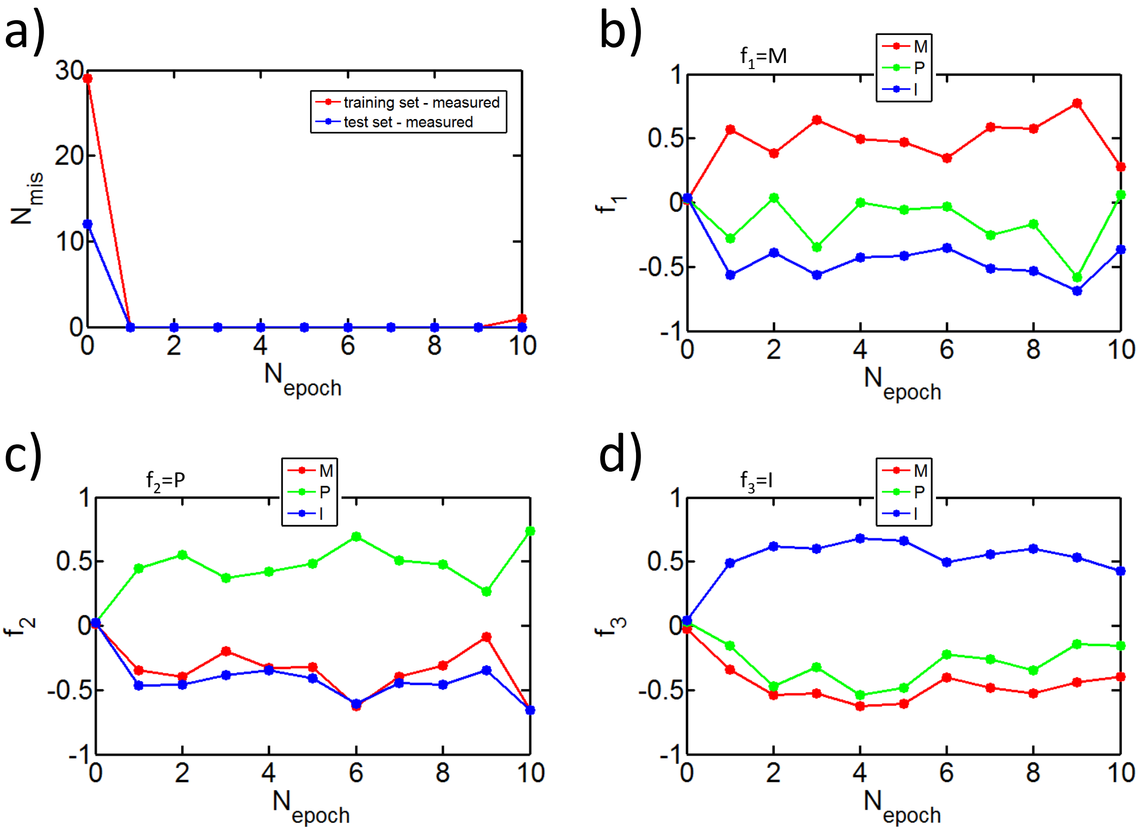


Figure 6.10: a) Number of mis-classified images for the training and test set versus training epoch. b)-d) Mean neuron activations for the three classifications over the training epochs. Material from [8]

7 Conclusion and Outlook

The aim of this dissertation was to prove a new device concept of a memcapacitive synapse theoretically and experimentally. The concept is based on a charge screening mechanism. In the theory section, the fundamental capacitive coupling curves were shown with TCAD simulations, and a high dynamic range was simulated. The high dynamic range was proven for devices down to 45 nm, thus proving the scalability of the concept. With Spice simulation a crossbar arrangement was simulated and very high energy efficiencies were obtained by using the concept of adiabatic charging with a harmonic signal. This enables the combination of reversible computing and neuromorphic computing. The energy efficiency of the human brain is estimated to be in the range of 10fJ/operation [174] (or 100 TOPS/W). With this technological approach one might beat this limit with an energy efficiency of several 1000 TOPS/W-10 000 TOPS/W. As mentioned earlier, the best known resistive devices can achieve up to 100 TOPS/W [12, 62, 68, 85]. Furthermore, the proposed technology is highly CMOS compatible and can be fabricated by state-of-the-art processes.

The process development of crossbar devices on a SOI wafer was challenging, especially due to deep trenches for the BL seperation and the two metal layers, but successful in the end.

Experimentally similar capacitive coupling curves, as in the theory section, were obtained. The high on/off ratio was also proven experimentally for micrometer scaled devices. Furthermore a 5x5 image recognition task was demonstrated using an experimental crossbar array with 156 memory cells and PCB with a switched capacitor circuit approach.

Future directions for research can be in the further scaling of experimental devices (nanometer regime). Experimental measures of the energy efficiency are especially meaningful for nanoscaled devices. An interesting approach is the BEOL integration based on entirely oxide grown devices, also for achieving three dimensional integration.

References

[1] H. Tsai, S. Ambrogio, P. Narayanan, R. M. Shelby, and G. W. Burr, “Recent progress in analog memory-based accelerators for deep learning,” *Journal of Physics D: Applied Physics*, vol. 51, no. 28, pp. 1–27, 2018.

[2] P. A. Merolla, J. V. Arthur, R. Alvarez-icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, “A million spiking-neuron integrated circuit with a scalable communication network and interface,” *Sciencemag.Org*, vol. 345, no. 7812, pp. 668–673, 2014.

[3] H. Mulaosmanovic, T. Mikolajick, and S. Slesazeck, “Accumulative Polarization Reversal in Nanoscale Ferroelectric Transistors,” *ACS Applied Materials and Interfaces*, vol. 10, no. 28, pp. 23997–24002, 2018.

[4] H. Mulaosmanovic, J. O. Ker, S. Mulle, U. Schroeder, J. Müller, P. Polakowski, S. Flachowsky, R. Van Bentum, T. Mikolajick, and S. Slesazeck, “Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors,” *ACS Applied Materials and Interfaces*, vol. 9, no. 4, pp. 3792–3798, 2017.

[5] A. A. Emara, M. M. Aboudina, and H. A. Fahmy, “Non-volatile low-power crossbar memcapacitor-based memory,” *Microelectronics Journal*, vol. 64, no. November 2016, pp. 39–44, 2017.

[6] Z. Wang, M. Rao, J. W. Han, J. Zhang, P. Lin, Y. Li, C. Li, W. Song, S. Asapu, R. Midya, Y. Zhuo, H. Jiang, J. H. Yoon, N. K. Upadhyay, S. Joshi, M. Hu, J. P. Strachan, M. Barnell, Q. Wu, H. Wu, Q. Qiu, R. S. Williams, Q. Xia, and J. J. Yang, “Capacitive neural network with neuro-transistors,” *Nature Communications*, vol. 9, no. 1, pp. 1–10, 2018.

[7] Q. Zheng, Z. Wang, N. Gong, Z. Yu, C. Chen, Y. Cai, Q. Huang, H. Jiang, Q. Xia, and R. Huang, “Artificial Neural Network Based on Doped HfO2 Ferroelectric Capacitors with Multilevel Characteristics,” *IEEE Electron Device Letters*, vol. 40, no. 8, pp. 1309–1312, 2019.

[8] K.-U. Demasius, A. Kirschen, and S. Parkin, “Energy-efficient memcapacitor devices for neuromorphic computing,” *Nature Electronics*, vol. 4, no. 10, pp. 748–756, 2021.

[9] S. Forouhi, R. Dehghani, and E. Ghafar-Zadeh, “Toward high throughput Core-CBCM CMOS capacitive sensors for life science applications: A novel current-mode for high dynamic range circuitry,” *Sensors*, vol. 18, no. 10, pp. 1–29, 2018.

[10] A. M. Geyer, “Ion Implantation,” vol. 2, 2009.

[11] M. Zhao, B. Gao, J. Tang, H. Qian, and H. Wu, “Reliability of analog resistive switching memory for neuromorphic computing,” *Applied Physics Reviews*, vol. 7, no. 1, pp. 1–18, 2020.

[12] T. Gokmen and Y. Vlasov, “Acceleration of deep neural network training with resistive cross-point devices: Design considerations,” *Frontiers in Neuroscience*, vol. 10, no. JUL, pp. 1–13, 2016.

[13] W. S. McCulloch and W. Pitts, “A Logical Calculus of the Ideas Immanent in Nervous Activity,” *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[14] D. O. Hebb, *The Organization of Behavior; A Neuropsychological Theory.* John Wiley and Sons, Inc., 1949.

[15] J. Vreeken, “Spiking neural networks , an introduction,” *Computing*, vol. 7, no. 3, pp. 1–5, 2002.

[16] P. Werbos, *The Roots of Backpropagation.* John Wiley & Sons Inc., 1994.

[17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to digit recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNET Classification with Deep Convolutional Neural Networks,” *Advances in neural information processing systems*, vol. 25(2), pp. 1–9, 2012.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2016.

[20] M. Shafique, T. Theocharides, C. S. Bouganis, M. A. Hanif, F. Khalid, R. Hafiz, and S. Rehman, “An overview of next-generation architectures for machine learning: Roadmap, opportunities and challenges in the IoT era,” *Proceedings of the 2018 Design, Automation and Test in Europe Conference and Exhibition, DATE 2018*, vol. 2018-Janua, pp. 827–832, 2018.

[21] C. Mead, “Neuromorphic Electronic Systems,” *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.

[22] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, “The missing memristor found,” *Nature*, vol. 453, no. 7191, pp. 80–83, 2008.

[23] J. Martinez-Rincon, M. Di Ventra, and Y. V. Pershin, “Solid-state memcapacitive system with negative and diverging capacitance,” *Physical Review B - Condensed Matter and Materials Physics*, vol. 81, no. 19, pp. 1–7, 2010.

[24] M. Di Ventra, Y. V. Pershin, and L. O. Chua, “Circuit elements with memory: Memristors, memcapacitors, and meminductors,” *Proceedings of the IEEE*, vol. 97, no. 10, pp. 1717–1724, 2009.

[25] S. J. Dat Tran and C. Teuscher, “Memcapacitive devices in logic and crossbar applications,” *International Journal of Unconventional Computing*, vol. 13, no. 1, pp. 35–57, 2017.

[26] M. Krems, Y. V. Pershin, and M. Di Ventra, “Ionic memcapacitive effects in nanopores,” *Nano Letters*, vol. 10, no. 7, pp. 2674–2678, 2010.

[27] M. G. Mohamed, H. Kim, and T. W. Cho, “Modeling of Memristive and Memcapacitive Behaviors in Metal-Oxide Junctions,” *Scientific World Journal*, vol. 2015, pp. 1–16, 2015.

[28] M. G. Ahmed, K. Cho, and T. W. Cho, “Memristance and memcapacitance modeling of thin film devices showing memristive behavior,” *International Workshop on Cellular Nanoscale Networks and their Applications*, pp. 1–5, 2012.

[29] D. Biolek, M. Di Ventra, and Y. V. Pershin, “Reliable SPICE simulations of memristors, memcapacitors and meminductors,” *Radioengineering*, vol. 22, no. 4, pp. 945–968, 2013.

[30] Y. V. Pershin and M. Di Ventra, “Memcapacitive neural networks,” *Electronics Letters*, vol. 50, no. 3, pp. 141–143, 2014.

[31] A. K. Khan and B. H. Lee, “Monolayer MoS2 metal insulator transition based memcapacitor modeling with extension to a ternary device,” *AIP Advances*, vol. 6, no. 9, pp. 1–7, 2016.

[32] D. Kwon and I. Y. Chung, “Capacitive Neural Network Using Charge-Stored Memory Cells for Pattern Recognition Applications,” *IEEE Electron Device Letters*, vol. 41, no. 3, pp. 493–496, 2020.

[33] T. You, L. P. Selvaraj, H. Zeng, W. Luo, N. Du, D. Bürger, I. Skorupa, S. Prucnal, A. Lawerenz, T. Mikolajick, O. G. Schmidt, and H. Schmidt, “An Energy-Efficient, BiFeO 3 -Coated Capacitive Switch with Integrated Memory and Demodulation Functions,” *Advanced Electronic Materials*, vol. 2, no. 3, pp. 1–9, 2016.

[34] K.-U. Demasius, *Entwicklung und Herstellung eines Halbleitersensors zur Messung elektrostatischer Felder mittels Ladungsträgermodulation in pin- Übergängen.* Diplomarbeit, TU Dresden, 2016.

[35] D. Purves, *Neuroscience.* Oxford University Press, 5th ed., 2011.

[36] S. Thanapitak, *Bionics Chemical Synapse.* PhD thesis, Imperial College London, 2011.

[37] D. Röttger, *Reconstruction and Visualization of Neuronal Pathways with Applications in Neuroscience.* Dissertation, Uni Koblenz, 2012.

[38] Y. Jiang, C. Yang, J. Na, G. Li, Y. Li, and J. Zhong, "A brief review of neural networks based learning and control and their applications for robots," *Complexity*, vol. 2017, 2017.

[39] J. Heaton, *Artificial Intelligence for Humans - Deep Learning and Neural Networks.* Heaton Research Inc., 2015.

[40] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[41] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. E. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, p. e00938, 2018.

[42] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proc. OF THE IEEE*, vol. 86(11), pp. 2278–2324, 1998.

[43] R. D. Jones, G. W. Flake, Y. Lee, P. S. Lewis, and S. Qian, "Function Approximation and Time Series Prediction With Neural Networks," *Proceedings of the International Joint Conference on Neural Networks*, 1989.

[44] T. Poggio and F. Girosi, "Network for Approximation and Learning," *Proc. IEEE*, vol. 78, no. 9, pp. 1481–1497, 1990.

[45] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[46] S. Sathasivam and W. A. T. Wan Abdullah, "Logic Learning in Hopfield Networks," *Arxiv*, vol. 2, no. 3, pp. 1–8, 2008.

[47] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive Science*, vol. 9, no. 1, pp. 147–169, 1985.

[48] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[49] D. Lee, M. Lim, H. Park, Y. Kang, J. S. Park, G. J. Jang, and J. H. Kim, “Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus,” *China Communications*, vol. 14, no. 9, pp. 23–31, 2017.

[50] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, “Recent advances in physical reservoir computing: A review,” *Neural Networks*, vol. 115, pp. 100–123, 2019.

[51] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[52] L. P. Kaelbling, “Reinforcement Learning: A Survey Leslie,” *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[53] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[54] D. H. Kang, H. G. Jun, K. C. Ryoo, H. Jeong, and H. Sohn, “Emulation of spike-timing dependent plasticity in nano-scale phase change memory,” *Neurocomputing*, vol. 155, pp. 153–158, 2015.

[55] U. Markowska-Kaczmar and M. Koldowski, “Spiking neural network vs multilayer perceptron: who is the winner in the racing car computer game,” *Soft Computing*, vol. 19, no. 12, pp. 3465–3478, 2015.

[56] S. R. Kulkarni and B. Rajendran, “Spiking neural networks for handwritten digit recognition—Supervised learning and network optimization,” *Neural Networks*, vol. 103, no. April, pp. 118–127, 2018.

[57] F. Ponulak and A. Kasiński, “Supervised learning in spiking neural networks with ReSuMe: Sequence learning, classification, and spike shifting,” *Neural Computation*, vol. 22, no. 2, pp. 467–510, 2010.

[58] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, “A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations,” *Nature Electronics*, vol. 2, no. 7, pp. 290–299, 2019.

[59] T. Noergaard, *Embedded Systems Architecture.* Elsevier, 2013.

[60] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[61] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, “Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,” *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.

[62] Z. Zou, Y. Jin, P. Nevalainen, Y. Huan, J. Heikkonen, and T. Westerlund, “Edge and Fog Computing Enabled AI for IoT-An Overview,” *Proceedings 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems, AICAS 2019*, pp. 51–56, 2019.

[63] C. Senthilpari, A. K. Singh, and K. Diwakar, “Design of a low-power, high performance, 8×8 bit multiplier using a Shannon-based adder cell,” *Microelectronics Journal*, vol. 39, no. 5, pp. 812–821, 2008.

[64] M. Courbariaux, Y. Bengio, and J.-P. David, “Binary Connect - Training Deep Neural Networks with binary weights during propagations,” *Advances in Neural Information Processing Systems*, no. 28, p. 10, 2015.

[65] D. Ielmini and H. S. Wong, “In-memory computing with resistive switching devices,” *Nature Electronics*, vol. 1, no. 6, pp. 333–343, 2018.

[66] M. Holler, S. Tam, H. Castro, and R. Benson, “Electrically trainable artificial neural network (ETANN) with 10240 ’floating gate’ synapses,” *IJCNN Int Jt Conf Neural Network*, pp. 191–196, 1989.

[67] C. Diorio, P. Hasler, and B. A. Minch, “A single transistor silicon synapse,” *IEEE Transactions on Electron Devices*, vol. 43, no. 11, p. 19721980, 1996.

[68] V. Agrawal, V. Prabhakar, K. Ramkumar, L. Hinh, S. Saha, S. Samanta, and R. Kapre, “In-Memory Computing array using 40nm multibit SONOS achieving 100 TOPS/W energy efficiency for Deep Neural Network Edge Inference Accelerators,” in *IEEE International Memory Workshop*, vol. -, pp. 1–4, 2020.

[69] S. Agarwal, D. Garland, J. Niroula, R. B. Jacobs-Gedrim, A. Hsia, M. S. Van Heukelom, E. Fuller, B. Draper, and M. J. Marinella, “Using Floating-Gate Memory to Train Ideal Accuracy Neural Networks,” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 5, no. 1, pp. 52–57, 2019.

[70] C. H. Bennett, T. P. Xiao, R. Dellana, B. Feinberg, S. Agarwal, M. J. Marinella, and S. N. Laboratories, “Device-aware inference operations in SONOS nonvolatile memory arrays,” *Arxiv*, 2020.

[71] L. Fick, D. Blaauw, D. Sylvester, S. Skrzyniarz, M. Parikh, and D. Fick, “Analog in-memory subthreshold deep neural network accelerator,” *Proceedings of the Custom Integrated Circuits Conference*, vol. 2017-April, 2017.

[72] L. Chua, “Memristor - The Missing Circuit Element,” *IEEE Transactions on circuit theory*, vol. 18, 1971.

[73] S. Vongehr and X. Meng, “The missing memristor has not been found,” *Scientific Reports*, vol. 5, pp. 1–7, 2015.

[74] J. Kim, Y. V. Pershin, M. Yin, T. Datta, and M. Di Ventra, “An Experimental Proof that Resistance-Switching Memory Cells are not Memristors,” *Advanced Electronic Materials*, vol. 6, no. 7, pp. 1–6, 2020.

[75] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” *Nature*, vol. 521, no. 7550, pp. 61–64, 2015.

[76] G. W. Burr, R. M. Shelby, S. Sidler, C. Di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, “Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element,” *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3498–3507, 2015.

[77] J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, “Neuromorphic spintronics,” *Nature Electronics*, vol. 3, no. 7, pp. 360–370, 2020.

[78] S. Lequeux, J. Sampaio, V. Cros, K. Yakushiji, A. Fukushima, R. Matsumoto, H. Kubota, S. Yuasa, and J. Grollier, “A magnetic synapse: Multilevel spin-torque memristor with perpendicular anisotropy,” *Scientific Reports*, vol. 6, no. July, pp. 1–7, 2016.

[79] T. S. Böescke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, “Ferroelectricity in hafnium oxide: CMOS compatible ferroelectric field effect transistors,” *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 547–550, 2011.

[80] 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,” *Technical*

Digest - International Electron Devices Meeting, IEDM, vol. 1, pp. 19.7.1–19.7.4, 2018.

[81] H. Mulaosmanovic, J. Ocker, S. Muller, M. Noack, J. Muller, P. Polakowski, T. Mikolajick, and S. Slesazeck, "Novel ferroelectric FET based synapse for neuromorphic systems," in *Symposium on VLSI Technology*, vol. -, pp. T176–T177, 2017.

[82] M. Jerry, S. Dutta, A. Kazemi, K. Ni, J. Zhang, P. Y. Chen, P. Sharma, S. Yu, X. S. Hu, M. Niemier, and S. Datta, "A ferroelectric field effect transistor based synaptic weight cell," *Journal of Physics D: Applied Physics*, vol. 51, no. 43, p. aad6f8, 2018.

[83] M. Jerry, P. Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, and S. Datta, "Ferroelectric FET analog synapse for acceleration of deep neural network training," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 6, no. c, pp. 6.2.1–6.2.4, 2018.

[84] H. Mulaosmanovic, E. Chicca, M. Bertele, T. Mikolajick, and S. Slesazeck, "Mimicking biological neurons with a nanoscale ferroelectric transistor," *Nanoscale*, vol. 10, no. 46, pp. 21755–21763, 2018.

[85] R. Berdan, T. Marukame, K. Ota, M. Yamaguchi, M. Saitoh, S. Fujii, J. Deguchi, and Y. Nishi, "Low-power linear computation using nonlinear ferroelectric tunnel junction memristors," *Nature Electronics*, vol. 3, no. 5, pp. 259–266, 2020.

[86] Z. Wang, W. Zhao, W. Kang, Y. Zhang, J. O. Klein, and C. Chappert, "Ferroelectric tunnel memristor-based neuromorphic network with 1T1R crossbar architecture," *Proceedings of the International Joint Conference on Neural Networks*, pp. 29–34, 2014.

[87] T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, "Ferroelectricity in hafnium oxide thin films," *Applied Physics Letters*, vol. 99, no. 10, pp. 0–3, 2011.

[88] J. Müller, T. S. Böscke, U. Schröder, S. Mueller, D. Bräuhaus, U. Böttger, L. Frey, and T. Mikolajick, "Ferroelectricity in simple binary ZrO 2 and HfO 2," *Nano Letters*, vol. 12, no. 8, pp. 4318–4323, 2012.

[89] H. Mulaos, J. O. Ker, S. Mulle, U. Schroeder, J. Müller, P. Polakowski, S. Flachowsky, R. Van Bentum, T. Mikolajick, and S. Slesazeck, "Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors," *ACS Applied Materials and Interfaces*, vol. 9, no. 4, pp. 3792–3798, 2017.

[90] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60–67, 2018.

[91] Y. Li, S. Kim, and X. Sun, "Capacitor-based Cross-point Array for Analog Neural Network with Record Symmetry and Linearit," *Symposium on VLSI Technology Digest of Technical Papers*, pp. 25–26, 2018.

[92] S. Kim, T. Gokmen, H. M. Lee, and W. E. Haensch, "Analog CMOS-based resistive processing unit for deep neural network training," *Midwest Symposium on Circuits and Systems*, vol. 2017-Augus, no. i, pp. 422–425, 2017.

[93] T. K. Tsang and M. N. El-Gamal, "Micro-electromechanical variable capacitors for RF applications," *Midwest Symposium on Circuits and Systems*, vol. 1, no. 2, pp. 177–186, 2002.

[94] T. Driscoll, H. T. Kim, B. G. Chae, B. J. Kim, Y. W. Lee, N. M. Jokerst, S. Palit, D. R. Smith, M. Di Ventra, and D. N. Basov, "Memory metamaterials," *Science*, vol. 325, no. 5947, pp. 1518–1521, 2009.

[95] R Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM Journal of Research and Development*, vol. 5, no. July, pp. 191, 183, 1961.

[96] C. H. Bennett, "Logical Reversibility of Computation.," *IBM Journal of Research and Development*, vol. 17, no. 6, pp. 525–532, 1973.

[97] M. P. Frank, "The Future of Computing Depends on Making It Reversible," in *IEEE Spectrum*, vol. -, pp. 1–7, 2017.

[98] M. P. Frank, "Introduction to Reversible Computing," *Proceedings of the 2nd conference on Computing frontiers*, pp. 385–390, 2005.

[99] E. Fredkin and T. Toffoli, "Conservative Logic," *Plenum Publishing Corporation*, vol. 21, pp. 219–253, 1982.

[100] A. Adamatzky, *Collision-Based Computing.* Springer, 2002.

[101] J. S. Wenzler, T. Dunn, T. Toffoli, and P. Mohanty, "A nanomechanical fredkin gate," *Nano Letters*, vol. 14, no. 1, pp. 89–93, 2014.

[102] M. Sanadhya and M. V. Kumar, "Recent Development in Efficient Adiabatic Logic Circuits and Power Analysis with CMOS Logic," *Procedia Computer Science*, vol. 57, pp. 1299–1307, 2015.

[103] R. Lal, W. Athas, and L. Svensson, “Low-power adiabatic driver system for AMLCDs,” *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, pp. 198–201, 2000.

[104] H. S. Raghav, V. A. Bartlett, and I. Kale, “Investigation of stepwise charging circuits for power-clock generation in Adiabatic Logic,” *2016 12th Conference on Ph.D. Research in Microelectronics and Electronics, PRIME 2016*, pp. 1–4, 2016.

[105] W. Athas, N. Tzartzanis, W. Mao, L. Peterson, R. Lal, K. Chong, J.-s. Moon, L. Svensson, and M. Bolotski, “The Design and Implementation of a Low-Power Clock-Powered Microprocessor,” vol. 35, no. 11, pp. 1561–1570, 2000.

[106] R. K. Yadav, A. K. Rana, S. Chauhan, D. Ranka, and K. Yadav, “Adiabatic technique for energy efficient logic circuits design,” in *2011 International Conference on Emerging Trends in Electrical and Computer Technology*, vol. -, pp. 776–780, IEEE, 2011.

[107] V. S. Kanchana Bhaaskaran, “Energy recovery performance of quasi-adiabatic circuits using lower technology nodes,” in *India International Conference on Power Electronics, IICPE 2010*, vol. -, pp. 1–7, IEEE, 2011.

[108] C. H. Ziesler, S. Kim, and M. C. Papaefthymiou, “A resonant clock generator for single-phase adiabatic systems,” *Proceedings of the International Symposium on Low Power Electronics and Design, Digest of Technical Papers*, pp. 159–164, 2001.

[109] D. Maksimović, V. G. Oklobdžija, B. Nikolić, and K. W. Current, “Clocked CMOS adiabatic logic with integrated single-phase power-clock supply: Experimental results,” *High-Performance System Design: Circuits and Logic*, vol. 8, no. 4, pp. 255–259, 1999.

[110] J. Fischer, E. Amirante, A. Bargagli-Stoffi, P. Teichmann, D. Gruber, and D. Schmitt-Landsiedel, “Power Supply Net for Adiabatic Circuits,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3254, pp. 413–422, 2004.

[111] Y. Ye and K. Roy, “QSERL: Quasi-Static Energy Recovery Logic,” *IEEE Journal of Solid-State Circuits*, vol. 36, no. 2, pp. 239–248, 2001.

[112] N. S. Kim and T. Austin, “Leakage current: Moore´s Law Meets Static Power,” *IEEE Computer Society*, 2003.

[113] E. Hückel, “Zur Theorie der Elektrolyte.,” *Naturwissenschaften*, 1924.

[114] S. Sze, *Physics of Semiconductor Devices.* John Wiley & Sons Inc., 3 ed., 2007.

[115] N. Srivastava and Geoffrey Hinton, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[116] T. Ali, P. Polakowski, S. Riedel, T. Büttner, T. Kämpfe, M. Rudolph, B. Pätzold, K. Seidel, D. Löhr, R. Hoffmann, M. Czernohorsky, K. Kühnel, X. Thrun, N. Hanisch, P. Steinke, J. Calvo, and J. Müller, "Silicon doped hafnium oxide (HSO) and hafnium zirconium oxide (HZO) based FeFET: A material relation to device physics," *Applied Physics Letters*, vol. 112, no. 22, 2018.

[117] H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, "Ferroelectric FETs With 20-nm-Thick HfO2 Layer for Large Memory Window and High Performance," *IEEE Transactions on Electron Devices*, vol. 66, no. 9, pp. 3828–3833, 2019.

[118] K. Kim, C. G. Hwang, and J. G. Lee, "DRAM technology perspective for gigabit era," *IEEE Transactions on Electron Devices*, vol. 45, no. 3, pp. 598–608, 1998.

[119] J. Muller, T. S. Boscke, S. Muller, E. Yurchuk, P. Polakowski, J. Paul, D. Martin, T. Schenk, K. Khullar, A. Kersch, W. Weinreich, S. Riedel, K. Seidel, A. Kumar, T. M. Arruda, S. V. Kalinin, T. Schlosser, R. Boschke, R. Van Bentum, U. Schroder, and T. Mikolajick, "Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories," *Technical Digest - International Electron Devices Meeting, IEDM*, no. 9, pp. 280–283, 2013.

[120] N. Gemma, S. I. O'Uchi, H. Funaki, J. Okada, and S. Hongo, "CMOS integrated DNA chip for quantitative DNA analysis," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pp. 2288–2297, 2006.

[121] B. Froment, F. Paillardet, M. Bely, J. Cluzel, E. Granger, M. Haond, and L. Dugoujon, "Ultra low capacitance measurements in multilevel metallisation CMOS by using a built-in electron-meter," *Technical Digest - International Electron Devices Meeting*, pp. 897–900, 1999.

[122] Y. W. Chang, H. W. Chang, C. H. Hsieh, H. C. Lai, T. C. Lu, W. Ting, J. Ku, and C. Y. Lu, "A novel simple CBCM method free from charge injection-induced errors," *IEEE Electron Device Letters*, vol. 25, no. 5, pp. 262–264, 2004.

[123] F. Widdershoven, A. Cossettini, C. Laborde, A. Bandiziol, P. P. Van Swinderen, S. G. Lemay, and L. Selmi, "A CMOS Pixelated Nanocapacitor Biosensor Plat-

form for High-Frequency Impedance Spectroscopy and Imaging," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 6, pp. 1369–1382, 2018.

[124] G. Nabovati, E. Ghafar-Zadeh, A. Letourneau, and M. Sawan, "Towards High Throughput Cell Growth Screening: A New CMOS 8 × 8 Biosensor Array for Life Science Applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 2, pp. 380–391, 2017.

[125] I. Evans and T. York, "Microelectronic capacitance transducer for particle detection," *IEEE Sensors Journal*, vol. 4, no. 3, pp. 364–372, 2004.

[126] A. Romani, N. Manaresi, L. Marzocchi, G. Medoro, A. Leonardi, L. Altomare, M. Tartagni, and R. Guerrieri, "Capacitive sensor array for localization of bioparticles in CMOS lab-on-a-chip," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 47, 2004.

[127] M. S. Lu, Y. C. Chen, and P. C. Huang, "5×5 CMOS capacitive sensor array for detection of the neurotransmitter dopamine," *Biosensors and Bioelectronics*, vol. 26, no. 3, pp. 1093–1097, 2010.

[128] P. Ciccarella, M. Carminati, M. Sampietro, and G. Ferrari, "Multichannel 65 zF rms Resolution CMOS Monolithic Capacitive Sensor for Counting Single Micrometer-Sized Airborne Particles on Chip," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2545–2553, 2016.

[129] T. Kern, "Symmetric differential current sense amplifier," 2010.

[130] D. Kadetotad, Z. Xu, A. Mohanty, P. Y. Chen, B. Lin, J. Ye, S. Vrudhula, S. Yu, Y. Cao, and J. S. Seo, "Parallel architecture with resistive crosspoint array for dictionary learning acceleration," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 2, pp. 194–204, 2015.

[131] H. Nyquist, "Thermal agitation of electric charge in conductors," *Physical Review*, vol. 32, no. 1, pp. 110–113, 1928.

[132] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IEEE*, vol. 86, no. 2, pp. 447–457, 1998.

[133] R.E. Sah, *Silicon Nitride, Silicon Dioxide Thin Insulating Films, and other Emerging Dielectrics VIII.* The Electrochemical Society, 2005.

[134] K. C. Saraswat, "Interconnections : Silicides," tech. rep., 2006.

[135] G. K. Celler and S. Cristoloveanu, "Frontiers of silicon-on-insulator," *Journal of Applied Physics*, vol. 93, no. 9, pp. 4955–4978, 2003.

[136] C. Maleville and C. Mazuré, "Smart-Cut® technology: From 300 mm ultrathin SOI production to advanced engineered substrates," *Solid-State Electronics*, vol. 48, no. 6, pp. 1055–1063, 2004.

[137] C. Maleville, B. Aspar, T. Poumeyrol, H. Moriceau, M. Bruel, A. J. Auberton-Hervè, and T. Barge, "Wafer bonding and H-implantation mechanisms involved in the Smart-cut® technology," *Materials Science and Engineering B*, vol. 46, no. 1-3, pp. 14–19, 1997.

[138] D. Munteanu, C. Maleville, S. Cristoloveanu, H. Moriceau, B. Aspar, C. Raynaud, O. Faynot, J. L. Pelloie, and A. J. Auberton-Hervé, "Detailed characterization of Unibond material," *Microelectronic Engineering*, vol. 36, no. 1-4, pp. 395–398, 1997.

[139] M. Bruel, B. Aspar, B. Charlet, C. Maleville, T. Poumeyrol, A. Soubie, A. J. Auberton-Herve, J. M. Lamure, T. Barge, F. Metral, and S. Trucchi, "'Smart cut': a promising new SOI material technology," *IEEE International SOI Conference*, no. October 1993, pp. 178–179, 1995.

[140] M. Bruel, B. Aspar, and A. J. Auberton-Hervé, "Smart-cut: A new silicon on insulator material technology based on hydrogen implantation and wafer bonding," *Japanese Journal of Applied Physics, Part 1: Regular Papers and Short Notes and Review Papers*, vol. 36, no. 3 SUPPL. B, pp. 1636–1641, 1997.

[141] S. P. Castro, *Characterization of the Boron Doping Process using Boron Nitride Solid Source Diffusion.* Master of science, North Carolina State University, 1999.

[142] P. Negrini, A. Ravaglia, and S. Solmi, "Boron Predeposition in Silicon Using BBr3," *J. Electrochem. Soc*, pp. 609–613, 1978.

[143] T. Nguyen Nhu, *Spin-On Glass: Materials and Applications in Advanced IC Technologies.* 1999.

[144] H. Müller, H. Ryssel, and I. Ruge, "A New Method for Boron Doping of Silicon by Implantation of BF2-Molecules," *Ion Implantation in Semiconductors*, pp. 85–95, 1971.

[145] R. G. Wilson, "Boron, fluorine, and carrier profiles for B and BF2 implants into crystalline and amorphous Si," *Journal of Applied Physics*, vol. 54, no. 12, pp. 6879–6889, 1983.

[146] S. Tian, M. F. Morris, S. J. Morris, B. Obradovic, G. Wang, A. F. Tasch, and C. M. Snell, "A detailed physical model for ion implant induced damage in sil-

icon," *IEEE Transactions on Electron Devices*, vol. 45, no. 6, pp. 1226–1238, 1998.

[147] M. Y. Tsai and B. G. Streetman, "Recrystallization of implanted amorphous silicon layers. I. Electrical properties of silicon implanted with BF+2 or Si ++B+," *Journal of Applied Physics*, vol. 50, no. 1, pp. 183–187, 1979.

[148] J. Müller, T. S. Böscke, U. Schröder, S. Mueller, D. Bräuhaus, U. Böttger, L. Frey, and T. Mikolajick, "Ferroelectricity in simple binary ZrO 2 and HfO 2," *Nano Letters*, vol. 12, no. 8, pp. 4318–4323, 2012.

[149] T. Ali, P. Polakowski, S. Riedel, T. Buttner, T. Kampfe, M. Rudolph, B. Patzold, K. Seidel, D. Lohr, R. Hoffmann, M. Czernohorsky, K. Kuhnel, P. Steinke, J. Calvo, K. Zimmermann, and J. Muller, "High Endurance Ferroelectric Hafnium Oxide-Based FeFET Memory Without Retention Penalty," *IEEE Transactions on Electron Devices*, vol. 65, no. 9, pp. 3769–3774, 2018.

[150] E. Yurchuk, J. Muller, S. Muller, J. Paul, M. Pesic, R. Van Bentum, U. Schroeder, and T. Mikolajick, "Charge-Trapping Phenomena in HfO2-Based FeFET-Type Nonvolatile Memories," *IEEE Transactions on Electron Devices*, vol. 63, no. 9, pp. 3501–3507, 2016.

[151] H. Ji, Y. Wei, X. Zhang, and R. Jiang, "Improvement of charge injection using ferroelectric Si:HfO2 as blocking layer in MONOS charge trapping memory," *IEEE Journal of the Electron Devices Society*, vol. 6, no. 1, pp. 121–125, 2018.

[152] H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, "Switching and Charge Trapping in HfO 2 -based Ferroelectric FETs : An Overview and Potential Applications," pp. 18–20, 2020.

[153] S. W. Lim, "Control of SC1 Wet Cleaning Process for Nano-Scale Gate Oxide Integrity," *Japanese Journal of Applied Physics, Part 1: Regular Papers and Short Notes and Review Papers*, vol. 42, no. 8, pp. 5002–5009, 2003.

[154] S. Petitdidier, V. Bertagna, N. Rochat, D. Rouchon, P. Besson, R. Erre, and M. Chemla, "Growth mechanism and characterization of chemical oxide films produced in peroxide mixtures on Si(100) surfaces," *Thin Solid Films*, vol. 476, no. 1, pp. 51–58, 2005.

[155] C. K. Fink, K. Nakamura, S. Ichimura, and S. J. Jenkins, "Silicon oxidation by ozone," *Journal of Physics Condensed Matter*, vol. 21, no. 18, 2009.

[156] T. Nishiguchi, Y. Morikawa, M. Miyamoto, H. Nonaka, and S. Ichimura, "Enhanced oxidation of silicon using a collimated hyperthermal ozone beam," *Applied Physics Letters*, vol. 79, no. 3, pp. 382–384, 2001.

[157] K. Koike, K. Izumi, S. Nakamura, G. Inoue, A. Kurokawa, and S. Ichimura, "Synthesis of silicon dioxide film using high-concentration ozone and evaluation of the film quality," *Journal of Electronic Materials*, vol. 34, no. 3, pp. 240–247, 2005.

[158] R. Feng and R. J. Farris, "Influence of processing conditions on the thermal and mechanical properties of SU8 negative photoresist coatings," *Journal of Micromechanics and Microengineering*, vol. 13, no. 1, pp. 80–88, 2003.

[159] D. Bai, M. Fowler, C. Planje, and X. Shao, "Planarization of deep structures using self-leveling materials," *Proceedings - 2012 45th International Symposium on Microelectronics, IMAPS 2012*, pp. 79–83, 2012.

[160] J.-B. Yoon, G. Y. Oh, C.-H. Han, E. Yoon, and C.-K. Kim, "Planarization and trench filling on severe surface topography with thick photoresist for MEMS," *Micromachining and Microfabrication Process Technology IV*, vol. 3511, no. August 1998, pp. 297–306, 1998.

[161] K. H. Rasmussen, S. S. Keller, F. Jensen, A. M. Jorgensen, and O. Hansen, "SU-8 etching in inductively coupled oxygen plasma," *Microelectronic Engineering*, vol. 112, pp. 35–40, 2013.

[162] G. Hong, A. S. Holmes, and M. E. Heaton, "SU8 resist plasma etching and its optimisation," *DTIP 2003 - Design, Test, Integration and Packaging of MEMS/MOEMS 2003*, no. October, pp. 268–271, 2003.

[163] D. Sameoto, S. W. Lee, and M. Parameswaran, "Electrical interconnection through optimized wirebonding onto SU-8 structures and actuators," *Journal of Micromechanics and Microengineering*, vol. 18, no. 7, 2008.

[164] A. Ohtomo and H. Y. Hwang, "Growth mode control of the free carrier density in SrTi O3-δ films," *Journal of Applied Physics*, vol. 102, no. 8, pp. 1–6, 2007.

[165] T. Tomio, H. Miki, H. Tabata, T. Kawai, and S. Kawai, "Control of electrical conductivity in laser deposited SrTiO3 thin films with Nb doping," *Journal of Applied Physics*, vol. 76, no. 10, pp. 5886–5890, 1994.

[166] K. Fukushima and S. Shibagaki, "Nb doped SrTiO3 thin films deposited by pulsed laser ablation," *Thin Solid Films*, vol. 315, no. 1-2, pp. 238–243, 1998.

[167] X. B. Lu, G. H. Shi, J. F. Webb, and Z. G. Liu, “Dielectric properties of SrZrO3 thin films prepared by pulsed laser deposition,” *Applied Physics A: Materials Science and Processing*, vol. 77, no. 3-4, pp. 481–484, 2003.

[168] J. W. Park, D. Lee, H. Kwon, and S. Yoo, “Improvement of on-off-current ratio in TiOx active-channel TFTs using N2O plasma treatment,” *IEEE Electron Device Letters*, vol. 30, no. 4, pp. 362–364, 2009.

[169] N. Zhong, J. J. Cao, H. Shima, and H. Akinaga, “Effect of annealing temperature on TiO 2-based thin-film-transistor performance,” *IEEE Electron Device Letters*, vol. 33, no. 7, pp. 1009–1011, 2012.

[170] J. Zhang, P. Cui, G. Lin, Y. Zhang, M. G. Sales, M. Jia, Z. Li, C. Goodwin, T. Beebe, L. Gundlach, C. Ni, S. McDonnell, and Y. Zeng, “High performance anatase-TiO2 thin film transistors with a two-step oxidized TiO2 channel and plasma enhanced atomic layer-deposited ZrO2 gate dielectric,” *Applied Physics Express*, vol. 12, no. 9, p. 96502, 2019.

[171] J. J. Yang, N. P. Kobayashi, D. a. a. Ohlberg, Z. Li, and R. S. Williams, “Engineering oxygen vacancies in atomic layer deposited TiO 2 films for Memristive switches,” vol. 429, no. V, p. 2008, 2008.

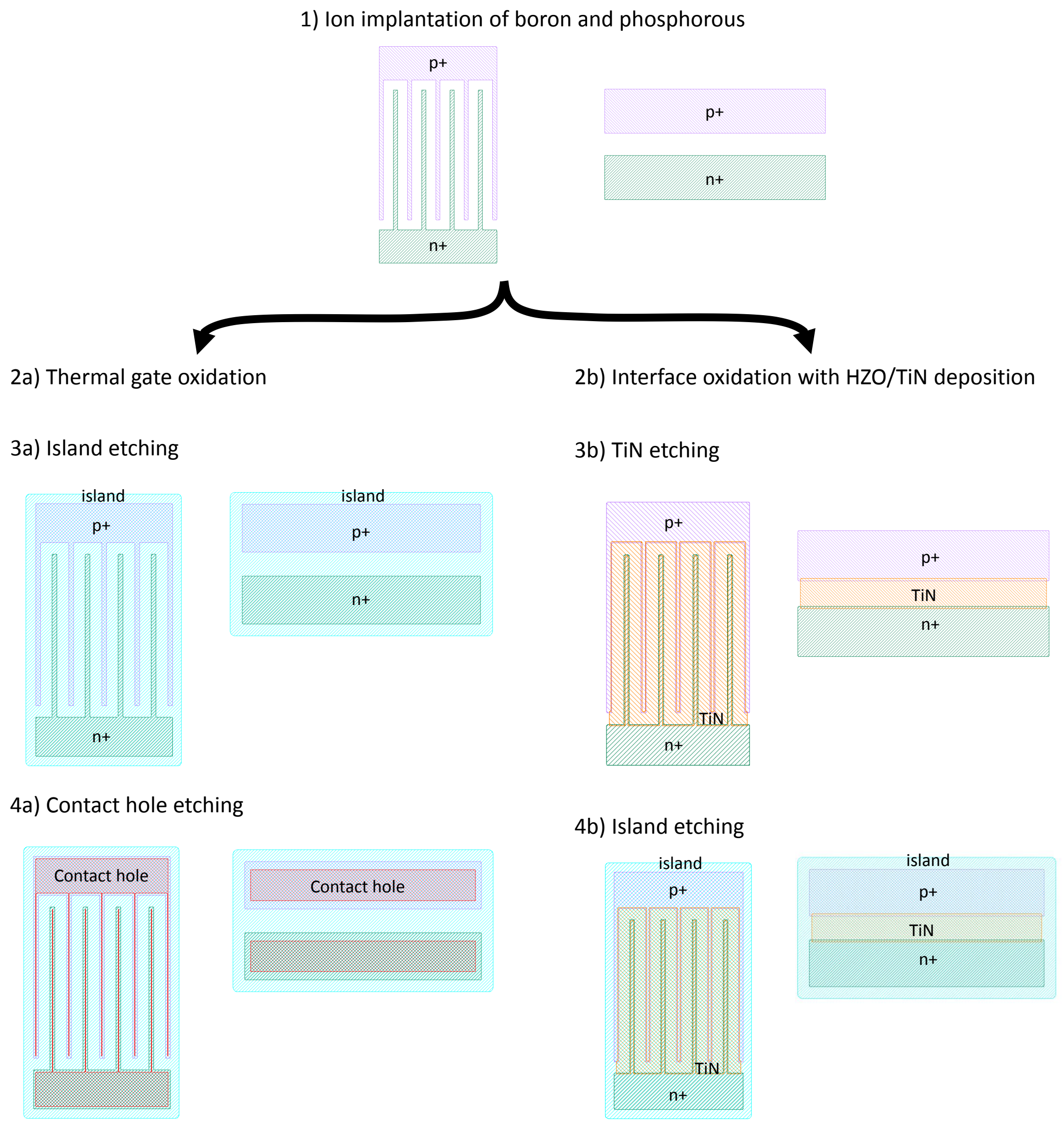
[172] F. Mo, Y. Tagawa, C. Jin, M. Ahn, T. Saraya, T. Hiramoto, and M. Kobayashi, “Experimental Demonstration of Ferroelectric HfO2 FET with Ultrathin-body IGZO for High-Density and Low-Power Memory Application,” *Digest of Technical Papers - Symposium on VLSI Technology*, vol. 2019-June, pp. T42–T43, 2019.

[173] E. Zamanidoost, F. M. Bayat, D. Strukov, and I. Kataeva, “Manhattan rule training for memristive crossbar circuit pattern classifiers,” in *Proc. IEEE*, vol. -, pp. 1–6, IEEE, 2015.

[174] W. Xu, S. Y. Min, H. Hwang, and T. W. Lee, “Organic core-sheath nanowire artificial synapses with femtojoule energy consumption,” *Science Advances*, vol. 2, no. 6, pp. 1–7, 2016.

A Appendix

A.1 Single Device Fabrication



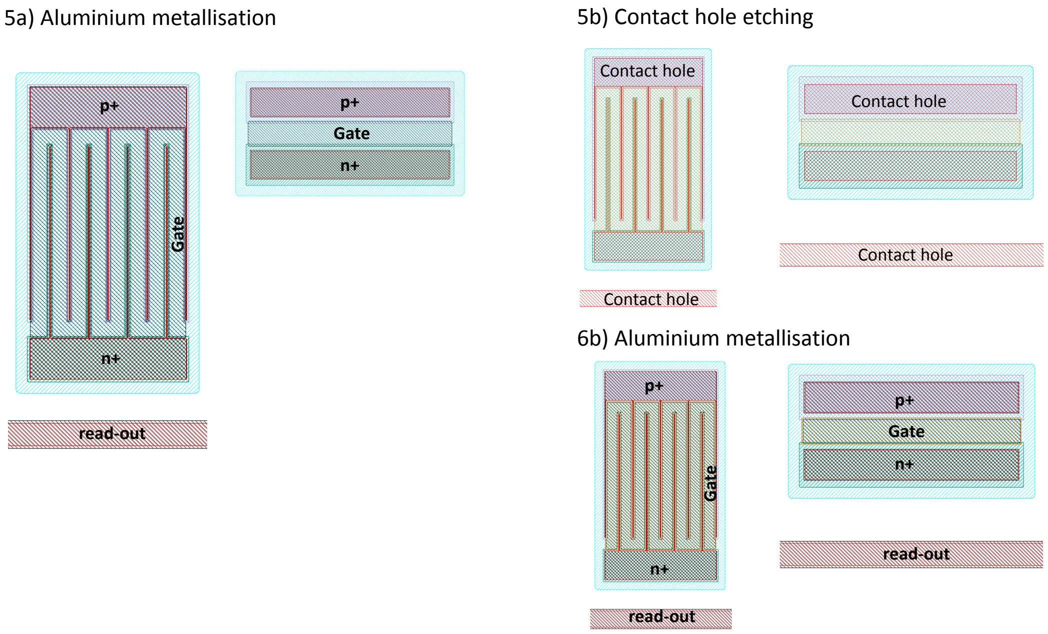
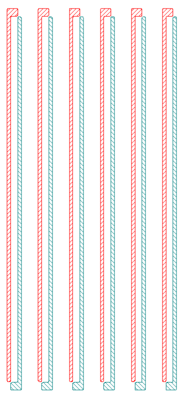


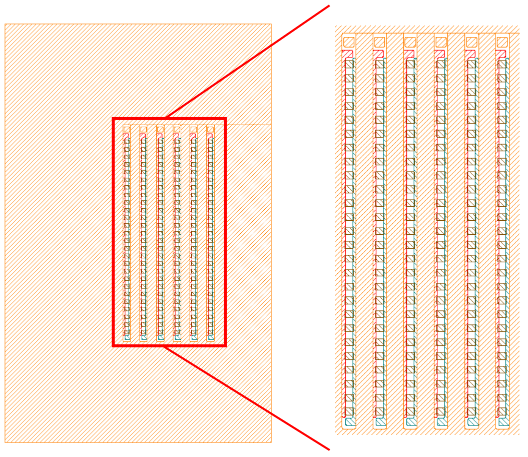
Figure A.1: Mask layers for different processes of single devices fabrication. The steps are corresponding to Fig. 4.5.

A.2 Crossbar Fabrication

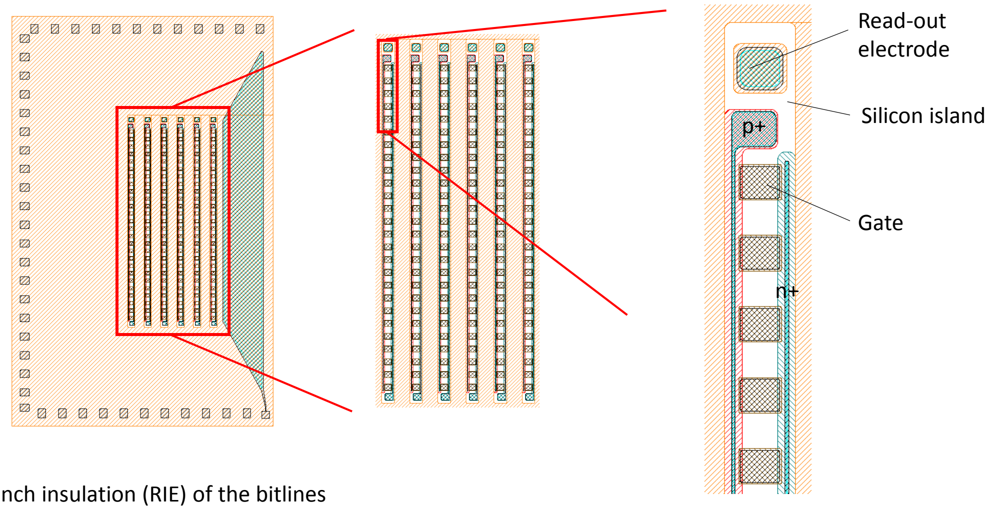
1) Ion implantation of boron and phosphorous



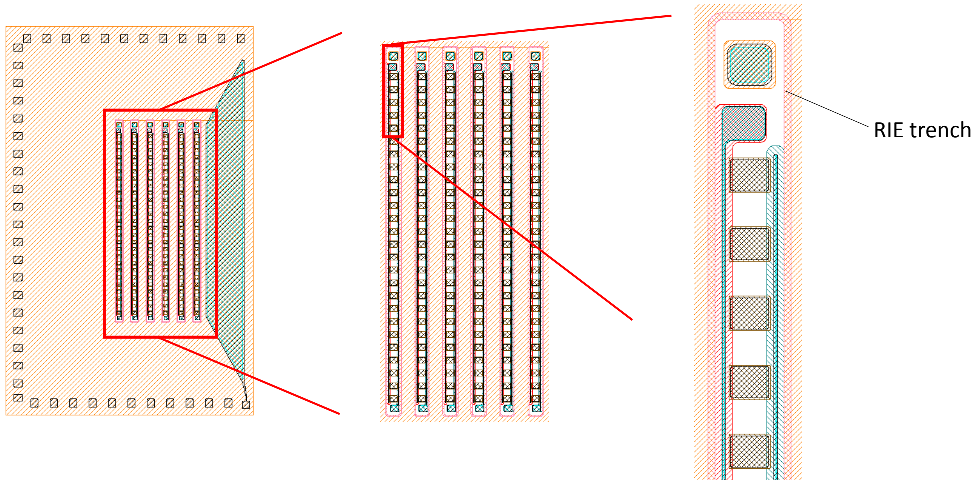
3) TiN etching and island etching



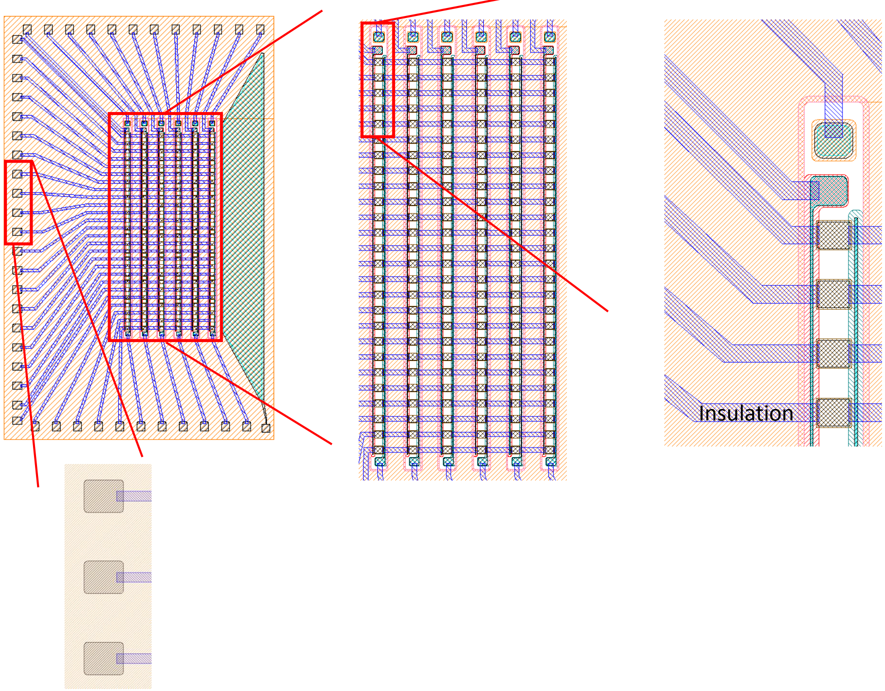
4) Contact hole etching and first aluminium layer



5) Trench insulation (RIE) of the bitlines



8) SU8 6001 spin coating and structuring



9) Second Alu metallisation and AlOx etch before

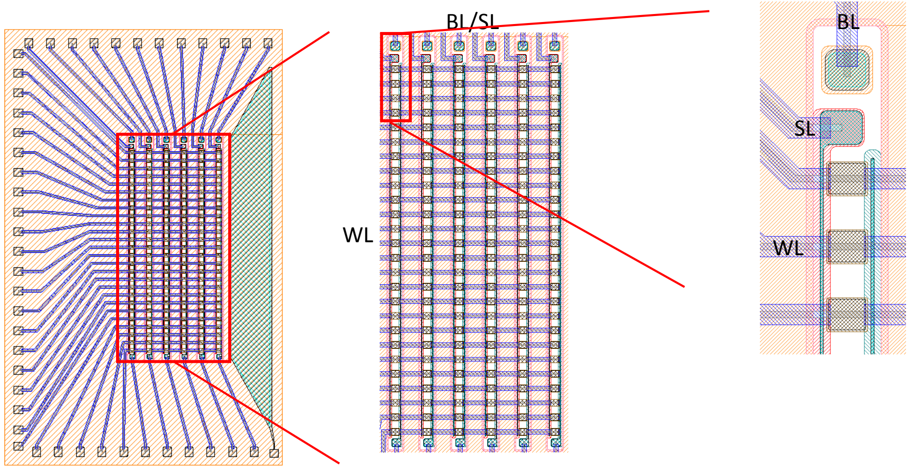


Figure A.2: Mask layers for different processes of crossbar devices fabrication. The steps are corresponding to Fig. 4.37.

Acknowledgements

I would like to express my deepest appreciation to the supervisor of the thesis, Prof. Dr. Stuart S. P. Parkin. I met him as a highly enthusiastic and creative person, which I will keep in mind throughout my future career. Furthermore, I wish to acknowledge his tolerance and understanding to allow me to work independently on my own topic.

I would like to thank all employees and fellow-students at the Max Planck Institute of Microstructure Physics for the good work environment. Special thanks are addressed to Alessandro Fumarola for sharing his knowledge about neuromorphic computing, which was new to me in the beginning of my PhD. I want to thank Fang Gao for keeping the clean room well organised and Jaechun Jeon for maintaining the Scia system in a good shape. I also gratefully acknowledge the support by the electronics workshop, especially Detlef Proske. Many thanks also to the administration and especially to Simone Jäger for organising the group.

Furthermore, I want to thank the Institute of Semiconductors and Microsystems at TU Dresden for access to the clean room and the Peter Grünberg Institute in Jülich for their support, especially at the start of the research work into this PhD. I want to acknowledge Karola Richter (TU Dresden) for RIE etching, Claudia Richter from Namlab gGmbH for the hafnium zirconium oxide deposition and Jan Krügener of the University of Hannover for the ion implantation. Moreover, I want to thank Dr. Karl-Heinz Stegemann for fruitful discussion.

I want to thank Aron Kirschen for his support in writing the paper as well as his ongoing encouragement throughout the work on this PhD thesis.

I want to thank my friend Christian Saß for his continuous support and for the joy of living alongside the PhD.

Many thanks also to my sister, Wiebke Demasius, and my brother-in-law, Achim Salamon, for their continuous support.

My deepest gratitude goes to my parents for their continuous support and for making my education possible.

Academic Curriculum Vitae

First Name:	Kai-Uwe
Family Name:	Demasius
Gender:	Male
Date of Birth:	May 15th, 1991
Place of Birth:	Johannesburg, South Africa
Nationality:	German

Education

10/2016-09/2020	**PhD Student at Max Planck Institute of Microstructure Physics** Supervisor: Prof. Dr. Stuart S. P. Parkin
09/2014-02/2015	**Research Internship** IBM Almaden Research Center, San José, USA Topic: *Spin-Hall materials* *DAAD scholarship*
10/2011-09/2016	**Dipl.-Ing. in Electrical Engineering** Dresden University of Technology, Germany Diploma thesis: *Entwicklung und Herstellung eines Halbleitersensors zur Messung elektrostatischer Felder mittels Ladungsträgermodulation in pin-Übergängen (Mark: 1.1)*
06/2010	**Abitur** Domschule, Schleswig, Germany

Ort, Datum

Kai-Uwe Demasius

Publications and Conference Contributions

Publications

K.-U. Demasius: Elektrostatischer Halbleitersensor. DE102010045363B4, issued September 14, 2010

K.-U. Demasius, T. Phung, W. Zhang, B.P. Hughes, S.-H. Yang, A. Kellock, W. Han, A. Pushp, S.S.P. Parkin: Enhanced spin–orbit torques by oxygen incorporation in tungsten films. Nature Communications, 7, 10644 (2016)

K.-U. Demasius, A. Kirschen: Matrix mit kapazitiver Steuerungsvorrichtung. DE102016012071A1, issued October 10, 2016

K.-U. Demasius, A. Kirschen, S.S.P Parkin: Extremely Energy-Efficient Memcapacitor Devices for Neuromorphic Computing. Nature Electronics, 4,748–756 (2021)

Conference contributions

K.-U. Demasius, S.S.P. Parkin: Cognitive and memory devices based on Debye length modulation. DPG-Frühjahrstagung der Sektion Halbleiterphysik, 2017

Eidesstattliche Erklärung

Hiermit erklräre ich, Kai-Uwe Demasius, dass ich die vorliegende Arbeit mit dem Titel: **Highly Energy Efficient Neuromorphic Computing Based on Memcapacitive Devices** in allen Teilen selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle wörtlich oder sinngemäß übernommenen Textstellen habe ich als solche kenntlich gemacht.

Ferner liegen gegen mich weder gerichtliche Vorstrafen vor, noch sind staatsanwaltliche Ermittlungen oder Disziplinarverfahren eingeleitet worden.

Des Weiteren erkläre ich hiermit, dass ich bisher keine andere Arbeit zur Promotion eingereicht noch mit einer anderen Arbeit den Versuch zur Promotion unternommen habe.

Ort, Datum	Kai-Uwe Demasius