PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING
UNIVERSITY OF NOTRE DAME
COLLEGE OF ENGINEERING

FUNCTIONAL OXIDE-BASED ELECTRONICS FOR LOGIC, MEMORY, AND RF APPLICATIONS

JORGE TOMÁS GÓMEZ MIR

Thesis submitted to Pontificia Universidad Católica de Chile and University of Notre Dame in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences

Advisor:
ÁNGEL ABUSLEME
SUMAN DATTA

Santiago de Chile, September, 2021

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING
UNIVERSITY OF NOTRE DAME
COLLEGE OF ENGINEERING

FUNCTIONAL OXIDE-BASED ELECTRONICS FOR LOGIC, MEMORY, AND RF APPLICATIONS

JORGE TOMÁS GÓMEZ MIR

Members of the Committee:

ÁNGEL ABUSLEME
SUMAN DATTA
ALAN SEABAUGH
CRISTOPHER HINKLE
CHRISTIAN OBERLI
PABLO ZEGERS
MARCELO GUARINI
JUAN DE DIOS ORTÚZAR

Thesis submitted to Pontificia Universidad Católica de Chile and University of Notre Dame in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences

Santiago de Chile, September, 2021

FUNCTIONAL OXIDE-BASED ELECTRONICS FOR LOGIC, MEMORY, AND RF APPLICATIONS

Abstract

by

Jorge Gómez Mir

Moore's law, which aims to double the number of transistors in the same area every 18 months, has been in full swing over the last 60 years. Almost every high-performance chip company considered moving to the next available technology node as a primary way to maximize value, however, with Moore's law slowing down, it is necessary to seek different strategies more closely aligned with the needs of each application. Without the expected device performance boost every 18 months, industries have started to look closely at each step in the production chain providing many opportunities to improve performance aside from of simply reducing the scale of transistors. This work explores and optimizes oxide-based emerging devices for logic, memory, neuromorphic computing and high frequency applications. We performed electrical characterization of several devices and developed high-fidelity, compact circuit-level models. These models bridge the different levels of the supply chain allowing us to exploit the performance of these novel devices for specific applications. For instance, for logic applications we modeled, built, and tested doped-Hafnium Dioxide based ferroelectric field effect transistors (FeFET). We then utilized these experimentally calibrated compact models to explore the phenomenon of Negative Capacitance (NC). This phenomenon can be harnessed to provide a boost in logic

transistor performance. We also proposed and experimentally demonstrated the utilization of an amorphous semiconductor oxide channel transistor using a Tungsten-doped Indium Oxide transistor. This transistor provides ultra-low leakage and is back-end-of-line (BEOL) compatible. Using these devices, we modeled, built, and tested a BEOL compatible embedded DRAM (eDRAM) with ultra-long refresh time.

Dedicated to my parents, Jorge and María Paz

CONTENTS

FIGURES

TABLES

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my advisers Dr. Angel Abusleme from PUC and Dr. Suman Datta from ND, for their valuable support, guidance and mentorship during the last five years. The continuous challenge and relevant research topics provided me with an exciting research environment that kept me motivated throughout the years. I am immensely in debt with all the people that we collaborated during these years. Sincere thanks to Professors Ioannis Vourkas, Asif Khan, Pinar Zorlutuna, Sayeef Salahuddin, Barbara de Salvo, Kai Ni and Sourav Dutta.

I would also like to thank my doctoral thesis committee at the University of Notre Dame and Pontificia Universidad Católica. Professors Juan de Dios Ortuzar, Christian Oberli, Pablo Zegers and Marcelo Guarini at PUC, and Professors Christopher Hinkle and Alan Seabough at ND. Their valuable feedback on my thesis and advice for the future have been a huge help in the last years.

The research presented here would not have been possible without the support of: *Agencia Nacional de Investigación y Desarrollo* (ANID, Chile), Semiconductor Research Corporation (SRC), Defense Advanced Research Projects Agency (DARPA), National Science Foundation (NSF), Facebook Agile Silicon Team (FAST) and Facebook Reality Labs (FRL).

My sincere gratitude to Christine Landaw, Barbara Walsh, Nicole Betti and Fernanda Kattan that, with immense patience, helped me to navigate through the intricate waters of the dual degree.

I have had the good fortune of interacting with many bright and talented friends and colleagues at PUC and ND. In particular, I would like to express my sincere

thanks to Dr. Cristóbal Alessandri, Renzo Barraza, Juan Andrés Bozzo, Pablo Walker, Matías Henriquez, Marie Carmen González and Agustín Campeny, at ICUC. Also, my sincere gratitude to Dr. Navnidhi Upadhyay, Dr. Benjamin Grisafe, Dr. Jeffrey Smith, Wriddhi Chakraborty, Matthew San Jose, Huacheng Ye, Abhishek Khanna, Akif Aabrar, Sanjukta Banerjee and Gopal Kirtania, at ND.

During this time, I also got to know a lot of amazing people outside the academic environment. This people were specially important during the lock-down months due to the COVID pandemic. I am specially grateful to all the folks in Windmoor residence and my quasi-chilean family in South Bend, that with: soccer, barbecues and hang out time, kept me going during this time.

Last, but by no means least, I would want to thanks my family, this entire endeavor would not have been possible without their continuous support.

CHAPTER 1

INTRODUCTION

> *"One truth remains: if the performance of the transistor can be improved, it can immediately and significantly improve efficiency at every level. As a result, we opine that research on new ideas for transistors will continue."*
>
> — Salahuddin et al. (2018)

1.1 Electronics at the Crossroads: The End of an Era

Globally, devices and connections are growing faster than both the population, and the Internet users (Fig. 1.1), as reported in the 2020 Cisco annual report [1] [1]. By 2023, 66% of the global population will have access to Internet (5.3 billion users), the number of devices connected to IP networks will be more than three times the population, and there will be 3.6 networked devices per capita. Computing is more personal than ever, playing a significant role in the economic, social, physical and intellectual pursuits of a human being. However, the computing dramatic growth, is threatened by power dissipation and management issues.

Enabling this dramatic growth of data-intensive technologies are overall system improvements at each level of the supply chain, which target to decouple power dissipation from the computing capability growth. A paradigmatic example would be data-centers, in which a predicted explosion in power consumption made in 2010, has not occurred thanks to continuous advances in power efficiency. Between 2010

[1] Connections: 10% Compound Annual Growth Rate (CAGR), Population: 1% CAGR, and Internet Users: 6% CAGR.

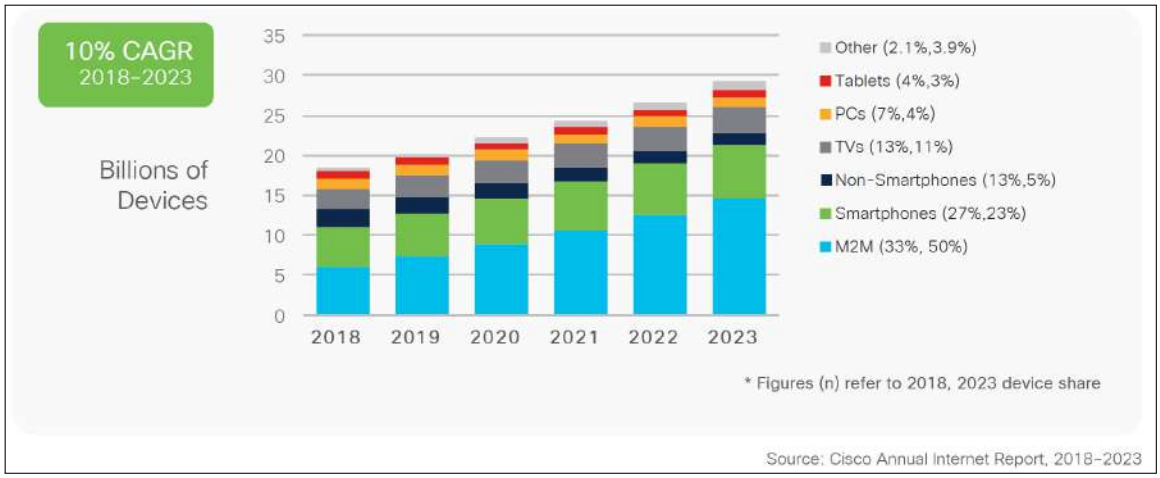


Figure 1.1. Devices and connection growth over the 2018-2023 period. Machine-to-Machine (M2M) applications such as: smart meters, video surveillance, health-care monitoring, transportation, and package or asset tracking connections, will be the fastest-growing device and connection category, growing nearly 2.4-fold during the forecast period (Figure reproduced from: [1]).

and 2018, datacenter workloads and computer instances increased by 6-fold, internet traffic increased by more than 10-fold, and data center storage capacity has increased by 25-fold. However, in the same period, the energy per computation has been reduced by four and the watts per terabyte installed has dropped by a factor of nine. Furthermore, in the same period, there was an improvement of 5-fold in the average number of compute instances hosted per sever. In summary, estimations shows that the global data center energy use, went from 194 (TWh) in 2010 to 205 (TWh) in 2018, which represent an increase of 6%, whereas global data center compute instances increased by 550%, as shown in [32].

A key backbone of this relentless technology improvement is a better transistor every 18 months, which can immediately and significantly improve efficiency at every level. Moore's law, which aims to double the number of transistors on the same

area every 18 months [33], has been in full swing over the last 60 years. In the market of high performance chips, almost every company considered moving to the next available node as a primary way to maximize value; however, as transistors and memory features are in the realm of a single digit nanometer, it becomes evident that room for further scaling in the horizontal direction is running out, and power densities have reached the limit of 100 W/cm^2, which has forced to stop increasing the clock frequency at 3GHz (Fig. 1.2 (a)). Also, the power supply voltage in microprocessors has not scaled at par with the transistor dimensions, as shown in Fig. 1.2 (b). The slowing down of transistor 2D scaling and huge power densities, jeopardizes the the future of transistor scaling, and, in consequence, the entire technology industry which benefits, at every level, from transistor scaling.

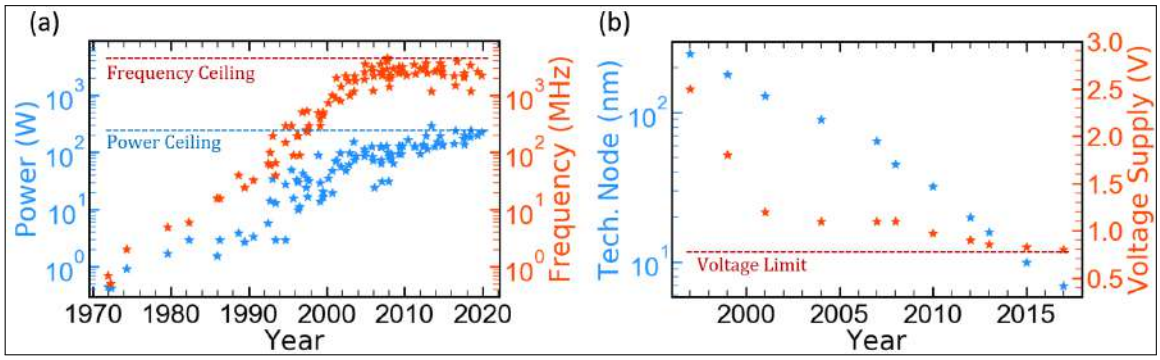


Figure 1.2. (a) Evolution of microprocessor power and frequency in the last five decades (Figure adapted from [2]), and (b) , the evolution of CMOS technology node and supply voltage in the last two decades (Figure adapted from [3]).

1.2 The Third Era of Scaling in Electronics: Hyper-Scaling

In the past six decades, the semiconductor industry has seen two distinct eras of scaling (Fig. 1.3): (a) the geometric scaling era, which reduced the transistor dimensions from the 1960s to the 2000s, and (b) the equivalent scaling era which, aided by strained silicon channel, high-κ/metal-gate stack and non-planar fin field-effect transistors, has increased the effective electron and hole velocity in the channel (strain), decreased the effective oxide thickness (high-κ dielectric), and increased the effective width of the minimum-size transistors (FinFETs).

Salahuddin et al. [4] argue that, thanks to design-technology co-optimization (DTCO) and process innovations, the era of equivalent scaling will continue until 2025. After 2025, scaling needs to enter into uncharted territory. In the same work, the authors argue that, after 2025, electronics will enter into a new third era which will be characterized by the functional augmentation of today's technology, which is refereed to as the Hyper-scaling era. This new era will be fueled by innovations in four areas: (i) beyond Boltzmann transistors, (ii) embedded high-performance memories beyond static random-access memory (SRAM) and dynamic random-access memory (DRAM), (iii) monolithic 3D integration of logic, memory, analog and I/O transistors, and (iv) heterogeneous integration of functionally diverse integrated circuits delivering monolithic-like performance.

1.3 Hypothesis, Objectives and Methodology of this Work

In this novel and challenging scenario the **hypothesis** of this thesis is that functional oxide electronics can enable the third era of scaling: hyper-scaling. Oxide-based emerging devices will play a key role in this new era, specifically by enabling two of the four areas: beyond-Boltzmann transistors and monolithic 3D integration of memory.

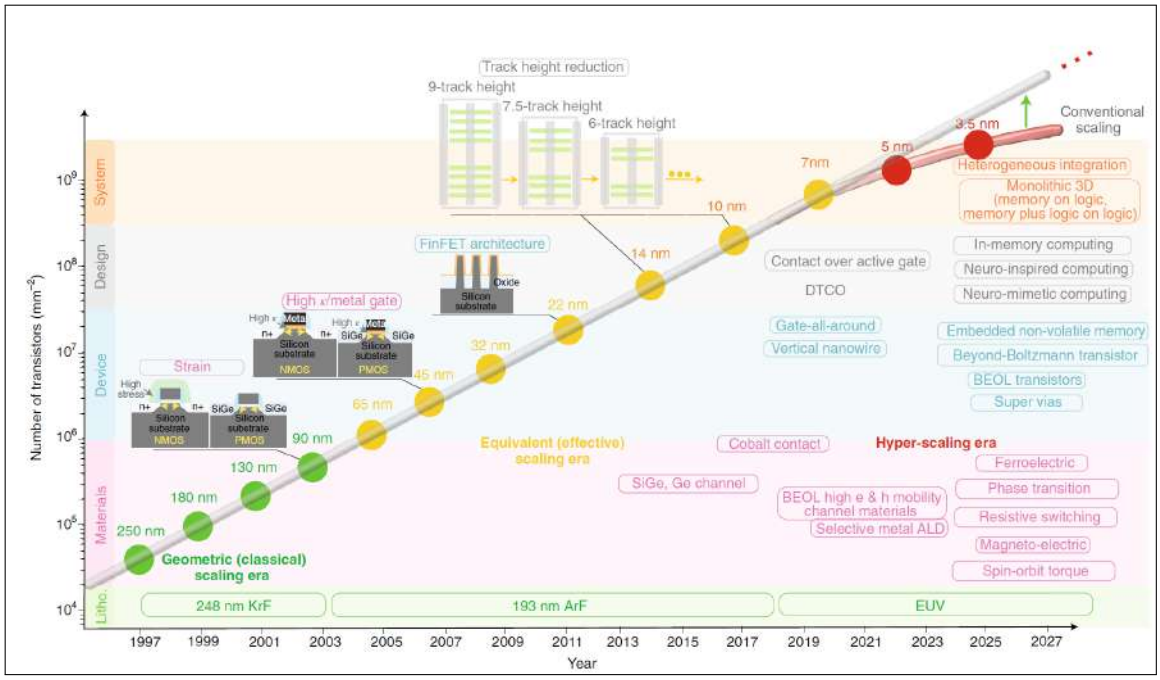


Figure 1.3. Three distinct eras of scaling: geometric scaling, equivalent scaling, and hyper scaling. Figure reproduced from [4]

.

The **objective** of this work is to theoretically understand and experimentally demonstrate how complex oxide multi-functional properties can be harnessed to augment the capabilities of traditional CMOS. The specific objectives for each one of the topics are:

1. To harness the phenomenon of Negative Capacitance to enable energy-efficient computing.

   (a) To theoretically understand the phenomenon of Negative Capacitance.

   (b) To experimentally demonstrate the phenomenon of Negative Capacitance.

   (c) To explore the performance boost obtained in transistors by incorporating Negative Capacitance.

2. To harness the properties of amorphous oxide transistors for Back-End-Of-Line (BEOL) compatible memories.

(a) To fabricate and characterize a BEOL compatible transistor.

(b) To fabricate and characterize a BEOL compatible Memory cell.

(c) To benchmark the performance of the proposed BEOL compatible memory.

The methodology followed to explore these topics is the following:

1. Develop a theoretical analysis to design the experiment.
2. Fabricate the devices considering the parameters obtained from the theoretical analysis.
3. Characterize the fabricated devices.
4. Benchmark the results with state-of-the-art literature.

Working on these two topics, we performed electrical characterization over several devices and developed compact circuit-level models. These models are the bridge between the different levels of the supply chain and allow us to explore the performance of these novel devices for specific applications.

1.4 Beyond Boltzmann transistor: Ferroelectric Negative Capacitance for High Performance Transistors

The digital dynamic power of a transistor is proportional to the squared supply voltage ($\propto V_{dd}^2$) and, the static power, is proportional to the supply voltage ($\propto V_{dd}$), hereafter the supply voltage plays a major role on the power dissipation of a chip and the ability to scale the supply voltage at par with the technology node was at the core *Dennard Scaling*, which states that as transistor get smaller, their power density stays constant over time. In section 1.1 we showed that there is a saturation of the supply voltage scaling and, in consequence, the power density is no longer constant, rather it increases with every technology node.

The supply voltage cannot be further scaled because there exists a minimum-allowable operating supply voltage so as to prevent an unacceptable increase of the

transistor's OFF-state current while guaranteeing an acceptable transistor's ON-state current. The ON-state and OFF-state currents depend on by the Sub-threshold Slope (SS), which is the inverse of the change of current that can be obtained for a unit change in gate voltage (V_G), as shown in equation (1.1) (Fig. 1.4 (b)). The expression has two factors: m and n. The first factor (equation 1.2) also called "body factor" comes from the electrostatics of the device and can be seen as a capacitor divider between the insulator capacitor (C_{ins}) and the semiconductor capacitor (C_{semi}) and relates V_G to the surface potential (ψ_S) (Fig. 1.4 (a)). The minimum value for this expression is one. On the transport mechanism side (n), the Boltzmann distribution dictates that to increase the drain current (I_D) by one order of magnitude at 300°K, V_G has to be increased by at least 60mV, as shown in equation (1.3). Accordingly, the minimum SS value is 60mV/decade. This is a fundamental bottleneck properly termed "Boltzmann Tyranny", which doesn't allow to further reduce supply voltage of the transistors without increasing exponentially the transistor's OFF-state current (Fig. 1.4 (b,c)).

$$SS = \left(\frac{dlog(I_D)}{dV_G}\right)^{-1} = \left[\left(\frac{d\psi_S}{dV_G}\right) \times \left(\frac{dlog(I_D)}{d\psi_S}\right)\right]^{-1} = m \times n \tag{1.1}$$

$$m = 1 + \frac{C_{semi}}{C_{ins}} \tag{1.2}$$

$$n = \frac{2.3k_BT}{q} \approx 60(mV) \tag{1.3}$$

Most of the proposed solutions to improve SS aim to enhance the transport mechanism (the n factor). Examples include impact ionization metal oxide semiconductor transistors (IMOS) [34] and tunnel field-effect transistors (TFETs) [35], however these solutions have not been successful because other problems arise with these devices. For example, IMOS present poor reliability and TFETs present very low I_{ON}/I_{OFF}

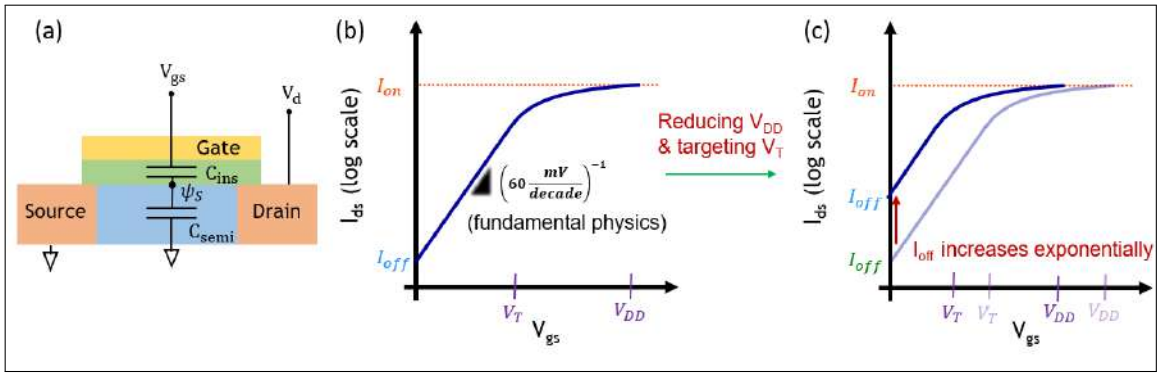


Figure 1.4. (a) Schematic of a MOSFET structure where C_{ins} is the insulator capacitance, C_{semi} is the semiconductor capacitance , and ψ_S is the surface potential. (b) Sub-threshold Slope of a transistor relates ON-state current with OFF-state current, and (c) Parallel shift of $I_d - V_g$ by reducing the supply voltage without reducing the SS.

current ratio [36]. An alternative approach is to improve the electrostatics of the device: reducing the equivalent oxide thickness (EOT) which will increase the value of C_{ins} in equation (1.2), hereafter reducing the value of the m factor. Originally SiO_2 was used as gate oxide, thus EOT is defined as the SiO_2 thickness necessary to achieve the same capacitance as a gate oxide with higher relative permittivity (κ) and thickness (d). Thus, with an SiO_2 relative permittivity of 3.9, the EOT can be calculated as shown in equation (1.4).

$$EOT = \frac{3.9}{\kappa} \times d \tag{1.4}$$

Fig. 1.5 shows the historic trend of EOT and the supply voltage. For 3nm and beyond logic nodes it is imperative to resume EOT scaling to meet the high performance, low-gate leakage and improved reliability requirements of advanced CMOS.

In 2008, it was theoretically shown that, by introducing a ferroelectric layer in the gate stack of a transistor, it is possible to create a charge boost, which will

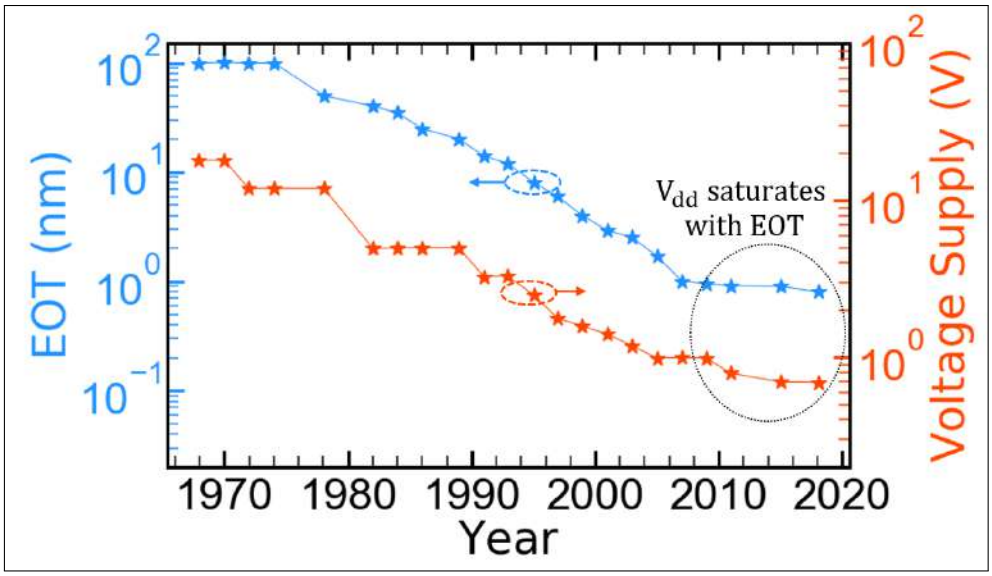


Figure 1.5. Scaling on the supply voltage and EOT of MOSFET over the last five decades. Figure adapted from [5]

show up as a negative capacitance [37]. The phenomenon of Negative Capacitance (NC) can provide a path toward continued scaling the EOT, hereafter improving the control of the gate electrode over the channel and, in consequence, reducing the SS of the transistors, which will enable the reduction of the supply voltage. NC will be explained in detailed in chapters 2 and 3, in which we demonstrate, through a comprehensive theoretical modeling and experimental characterization, a novel strategy to reduce EOT of the gate-stacks, by enhancing the dielectric constant (κ).

We have performed electrical characterization and developed compact circuit-level models for doped-Hafnium Dioxide (HfO_2) based ferroelectric negative capacitance field effect transistor (NCFET). Experimentally we demonstrate higher-κ response in NCFET compared to baseline, which allows EOT scaling without mobility degradation and improved gate-stack reliability. We also demonstrate a boost in saturation current ($I_{D,SAT}$) and an improvement in transit frequency (f_T). All this indicates the

potential of NCFET to be integrated in the next-generation advanced CMOS nodes to resume EOT scaling.

1.5 Monolithic 3D integration: Double-Gate W-Doped Amorphous Indium Oxide Transistors for Monolithic 3D Capacitorless Gain Cell DRAM

Recent years have witnessed a rise in domain specific accelerators designed for graphics, deep learning, bioinformatics, image processing and other tasks ([38], [39]). In addition to the obvious benefit of specialization of logic cores and high level of parallelism, a key source of the acceleration comes from availability of local memories for individual cores and large shared buffer memory on a monolithic chip. AI Accelerators have created an increasing need of high-density, high-performance and low stand-by power memory alternatives to traditional SRAM. These requirements get magnified for today mobile applications in which power dissipation and size constraints play a major role on the design space. Each Deep Neural Network (DNN) has multiple layers which imply to convolve each image with multiple filters (Fig. 1.6 (a)). At the core of each convolution we find the multiply-accumulate (MAC) operation which requires three memory reads (filter weight, fmap activation and partial sum) and one memory write (for partial sum update). To achieve high accuracy, we need to pay a big computational and memory cost, as shown in Fig. 1.6 (b), in which the accuracy is related to the computation-memory requirements for different image recognition DNNs.

The DNN's memory requirements define a complex design space which implies to carefully balance: delay, energy, cost and capacity. Normally, a hierarchical approach is chosen to accomplish these requirements in which different memory types and capacities are chosen at each level to correctly balance the system performance. Fig. 1.7 shows an example of an accelerator with an specific memory hierarchy. The memory hierarchy may vary for different accelerators. At the core of the accelerator

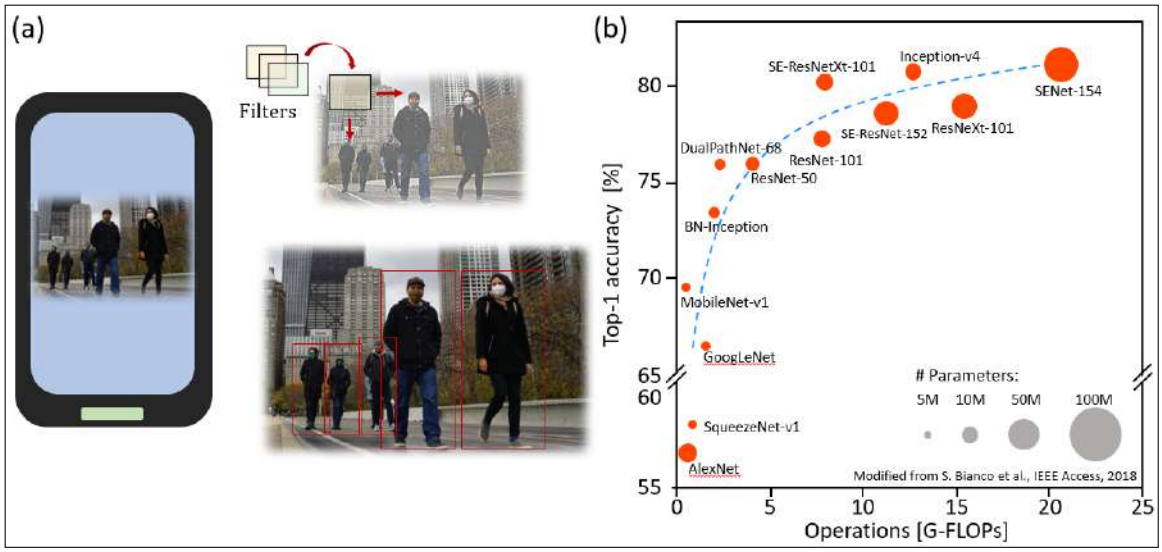
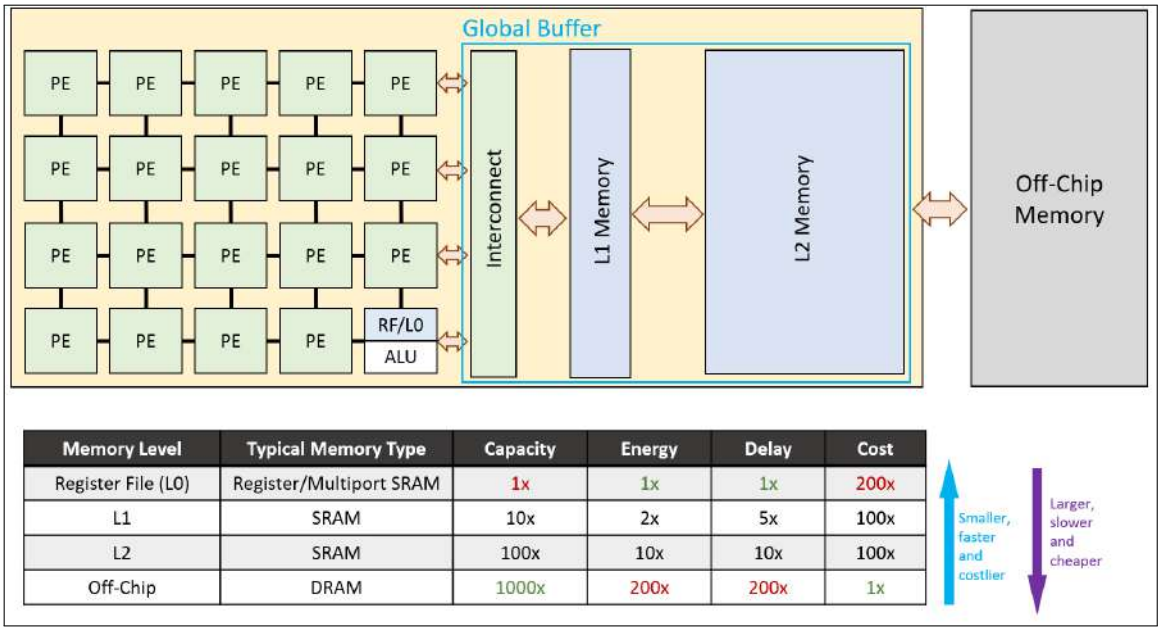


Figure 1.6. (a) Representation of the workloads that state-of-the-art DNNs have to compute in energy-compute constrained mobile devices, (b) Computation-memory requirements for state-of-the-art DNNs (Figure reproduced from [6])

we find the Processing Element (PE) that consists on an ALU with a few kilobytes of register file (RF) memory to which the ALU has a direct and fast access. Register File may be an array of registers or multiple ports SRAM. The PE fetch its data from other PEs or from a shared Multi-Bank L1 memory, which, although it is slower than L0, it is still a very fast memory. When the data is not available at L1 cache, L1 has to fetch the data from L2, which basically trades energy and speed by capacity. L2 cache at the same time has to fetch its data from the off chip DRAM memory, which has a big capacity at a cost of a bigger delay and energy cost compared to the On-chip memories.

To relax the memory hierarchy trade-off, chip designers are looking for high capacity, high performance On-chip memory. A promising candidate is embedded DRAM (eDRAM) which is a perfect gap filler between SRAM and Off-chip DRAM, as shown



Memory Level	Typical Memory Type	Capacity	Energy	Delay	Cost
Register File (L0)	Register/Multiport SRAM	1x	1x	1x	200x
L1	SRAM	10x	2x	5x	100x
L2	SRAM	100x	10x	10x	100x
Off-Chip	DRAM	1000x	200x	200x	1x

Figure 1.7. Example of memory hierarchy of a generic DNN accelerator. Closer to the PE the memories are faster and more energy efficient; however, with a higher cost and a lower capacity. The memories that are farther away from the PEs reduce cost and increase capacity, but have a higher delay and are less energy efficient. Data compiled from multiple sources: [7], [8] and [9]

in Fig. 1.8 (a). However, eDRAM is facing its own scaling challenges which force to use a relaxed pattern in the eDRAM regions, for example, Intel GT3e graphics processing unit was fabricated using 22nm Tri-Gate transistor technology node, however the eDRAM relaxed pattering was about three generations behind the peripheral logic [11]. Fig. 1.8 (b) shows the schematic of a eDRAM bit-cell with two cylinder capacitors. The eDRAM technology has not been able to scale without decreasing the storage capacitance, although the aspect ratio (A/R) of the cell capacitor is increasing to the fabrication limit, as shown in Fig. 1.8 (c). Lower storage capacitance impacts the retention time of the eDRAM, thus the refreshing frequency increases. It is extremely hard to scale conventional eDRAM/DRAM below the 10nm technology

node.

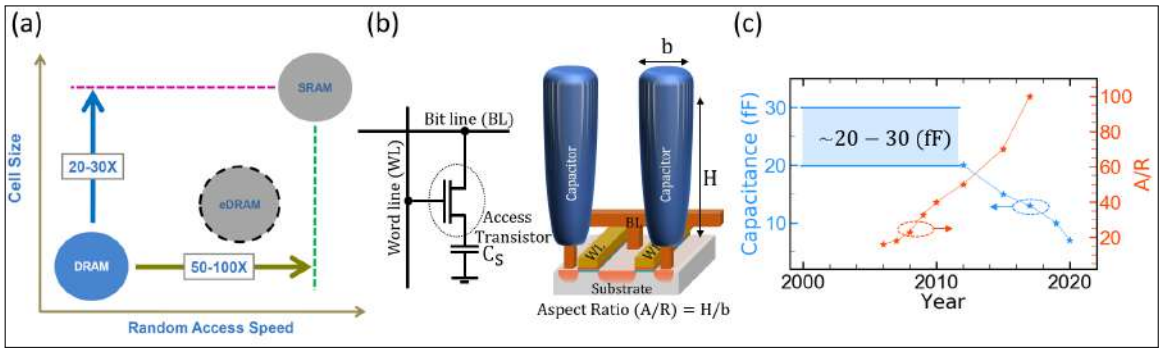


Figure 1.8. (a) eDRAM is a performance Gap Filler between Off-Chip DRAM and SRAM (Fig. reproduced from [10]) (b) Schematic of DRAM bit-cell, and (c) Capacitor aspect ratio and capacitance scaling. Data compiled from multiple sources: [11], [12], [13], [14], [15] and [16]

Novel technologies are emerging to replace conventional eDRAM in smaller nodes. Thin film transistors (TFT) based capacitor-less two-transistor gain cell embedded DRAM is a promising candidate due to its potential for higher density, lower power consumption, higher endurance than eNVRAM and their scaling potential to advanced nodes. Gain cells are dynamic memory cells made of two logic transistors, where the second transistor is used not only to increase the in-cell storage capacitance but also to amplify the readout charge via the transconductance gain, resulting in a non-destructive read. In chapter 4 we demonstrate back-end-of-line (BEOL) compatible W-doped amorphous In_2O_3 double-gate field-effect transistors (DG IWO FET) that exhibit:(a) excellent subthreshold slope (SS), (b) high $I_{D,SAT}$, and (c) high I_{ON}/I_{OFF} ratio. All these characteristics make IWO FET a perfect candidate for

low-energy, high density and high performance memories. Based on these transistors, we experimentally demonstrate an IWO FET based capacitorless 2T gain cell embedded DRAM (eDRAM) ideal for Monolithic 3D (M3D) integration exhibiting: (a) extremely low cell level leakage and (b) high retention time (Fig. 1.9).

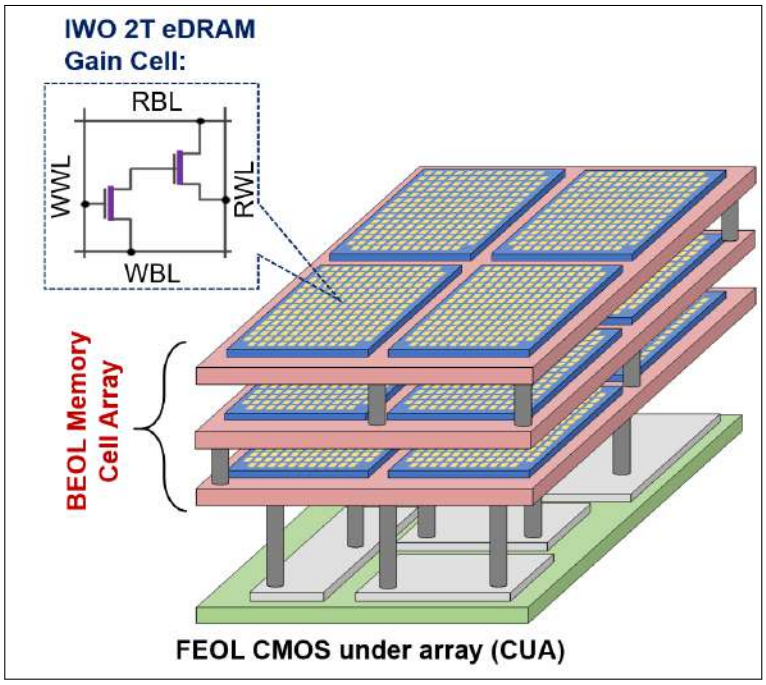


Figure 1.9. Monolithic-3D integration of Tungsten-doped Indium Oxide (IWO) channel FET based 2T-eDRAM

On the model side, we extend the semi-empirical physics-based Virtual Source (VS) model for IWO transistors [40]. We incorporate gate voltage (V_{GS}) modulated Schottky diodes in series with an intrinsic FET to capture the V_{GS} dependent source-drain contact resistance in IWO FET [41]. We implemented this model in Spice compatible Verilog-A language, and we used it to understand the performance of our

memory cells at an array level.

1.6 Organization of the Thesis

This thesis describes the research done on two promising functional oxide-based emerging nanoelectronics devices which will contribute to the third scaling era: hyperscaling. NCFETs and BEOL compatible IWO FETs are running candidates to play a role in the next generation of low-power logic, ultra-dense low-power memory, and high frequency applications. The thesis is divided into five chapters:

1. *Chapter 1:* Present the motivation for this research, describes the landscape in which this research takes place and introduces some key concepts to understand the relevance of this work.
2. *Chapter 2:* Explains the current understanding of the phenomenon of negative capacitance and introduces our alternative, although equivalent, explanation of this phenomenon. Based on this new understanding we propose a spice-compatible circuit-based compact model which also captures the switching dynamics of ferroelectrics. We use this model to target some of the bigger concern about negative capacitance such as capacitance matching and the ferroelectric multi-domain scenario.
3. *Chapter 3:* In this chapter we validate our theoretical results on negative capacitance. We first explain in detail the measurement methodology of negative capacitance and then we show experimental results of metal-ferroelectric-insulator-metal (MFIM) structures, demonstrating how negative capacitance arises in these structures. Then, we introduce a ferroelectric layer into the gate stack of a transistor and show how it can be a path towards continue scaling the Equivalent Oxide Thickness (EOT).
4. *Chapter 4:* Here we introduce our W-Doped Amorphous Indium Oxide Transistors which are BEOL compatible, thus allowing Monolithic 3D integration. We experimentally demonstrate the excellent performance of our devices and the 2T-eDRAM memory performance using these devices. We develop Spice compatible Verilog A model which allows to further explore the capabilities of these devices at an array level.
5. *Chapter 5:* Summarizes the key results of this thesis and propose suggestions for future work.

This thesis contains multiple demonstrations, especially in chapter 2. In order to highlight the key results and the main take away of each chapter, all the analytic

demonstrations were moved to the appendix section, in which each step is explained in detail.

CHAPTER 2

THEORY OF FERROELECTRIC NEGATIVE CAPACITANCE FOR HIGH PERFORMANCE TRANSISTORS

> *"The voltage transformer action can be understood intuitively as the result of an effective negative capacitance provided by the ferroelectric capacitor that arises from an internal positive feedback that in principle could be obtained from other microscopic mechanisms as well. Unlike other proposals to reduce SS, this involves no change in the basic physics of the FET and thus does not affect its current drive or impose other restrictions."*
>
> — Sayeef Salahuddin and Supriyo Datta (2008)

A ferroelectric material is an insulator with two or more discrete stable states of different nonzero electric polarization. In a ferroelectric system it is possible to switch between these states with an applied electric field larger than the coercive field. The hysteretic nature of polarization was established in 1921 through the work of Joseph Valasek with Rochelle salt [42]; however, the recent discovery of ferroelectricity in the CMOS-compatible HfO_2 material system [43] has lead to a more intensive research in a variety of applications including memory, steep slope transistors, and neuromorphic computing.

In this chapter we will first introduce the Landau Theory of Polarization Switching, traditionally used to model ferroelectrics (see for example [44] and [17]). Secondly, we will introduce our own circuit based compact model and demonstrate that it is equivalent to the Landau Theory of Polarization Switching. Then, we will use our model to explore the Negative Capacitance (NC) phenomenon, which arises when a ferroelectric layer is added to the gate stack of a transistor, and can be used to

reduce the sub-threshold slope below 60 mV/decade. Finally, we will experimentally explore the impact of the NC phenomenon on the transit frequency (f_T) of Negative Capacitance Field Effect Transistors (NCFETs).

2.1 Landau Theory of Polarization Switching

A common approach to model the transitions of a ferroelectric material between the discrete stable states is to use Landau theory (see for example [45]), that is a phenomenology that serves as a conceptual bridge between the microscopic models and the observed macroscopic phenomena. In general, the thermodynamic state of any system in equilibrium can be completely specified by the values of specific variables and for ferroelectrics. These include: the temperature (T), the polarization (P), the electrical field (E), the strain(η), and the stress (σ) [45]. Landau characterized the phase transition in terms of an order parameter, one of the variables that describes the thermodynamic state: it is zero in the high symmetry phase and changes continuously to a finite value when the symmetry is lowered [45]. In the case of ferroelectrics, Landau used polarization (P) or equivalently polarization charge (Q_P). The free energy (U) is expanded as a power series of Q_P, as shown in Landau-Khalatnikov (LK) equation (2.1). Also, in order to reproduce the experimental results more easily, we replace the electric field across the ferroelectric (E_{fe}) by the voltage across the ferroelectric (V_{fe}).

$$U = \frac{\alpha}{2}Q_P^2 + \frac{\beta}{4}Q_P^4 - Q_P V_{fe} \tag{2.1}$$

where α and β are anisotropy constants. To understand the dynamics of the ferroelectric, the rate of change in polarization can be described using equation (2.2).

$$\rho\frac{dQ_P}{dt} = -\frac{dU}{dQ_P} \tag{2.2}$$

where ρ is the frictional inertia or internal resistance of the system. Finally combining (2.1) and (2.2), we can derive equation (2.3) which describes the dynamics of the system.

$$V_{fe} = \rho \frac{dQ_P}{dt} + \left(\alpha Q_P + \beta Q_P^3\right) \tag{2.3}$$

In equilibrium ($\frac{\mathrm{dU}}{\mathrm{dQ_P}} = \frac{\mathrm{dQ_P}}{\mathrm{dt}} = 0$), equation (2.3) becomes equation (2.4).

$$V_{fe} = \alpha Q_P + \beta Q_P^3 \tag{2.4}$$

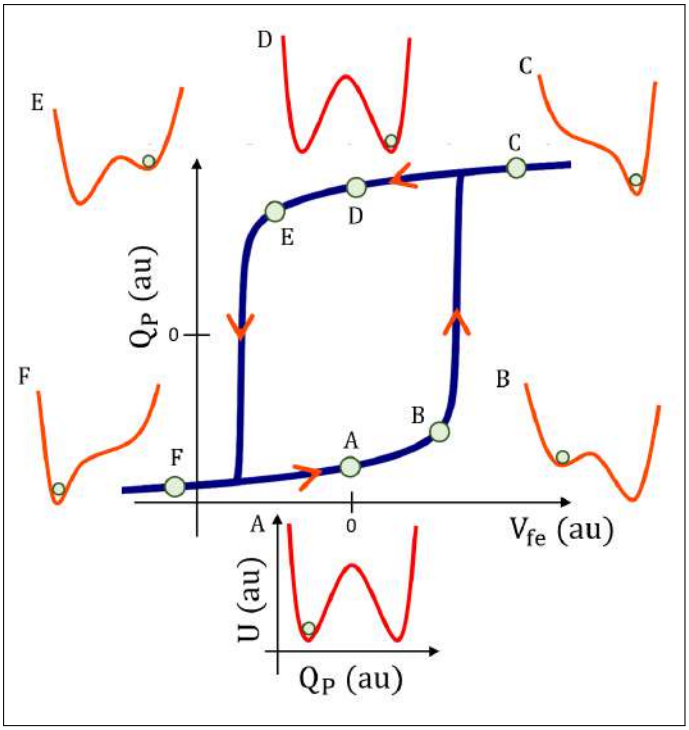


Figure 2.1. Energy landscapes (red) at different points on the hysteresis curve (blue) of the polarization-voltage characteristics of a ferroelectric capacitor. Figure adapted from [17])

Figure 2.1 shows a schematic of the evolution of the energy landscape at different voltages across the ferroelectric and the correspondence with the polarization-voltage dynamics.

To derive the anisotropy parameteres we follow the procedure presented in [44], shown in Appendix A. The anisotropy values are shown in equation (2.5) in which V_C is the coercive voltage: threshold voltage that causes the polarization to switch, and Q_0 is the remnant polarization: polarization that remains in absence of electric field. Both quantities can be obtained directly from experimental measurements.

$$\alpha = -\frac{3\sqrt{3}}{2}\frac{V_c}{Q_0} \ , \ \beta = \frac{3\sqrt{3}}{2}\frac{V_c}{Q_0^3} \tag{2.5}$$

Now replacing the anisotropy parameters in equation (2.4) and defining $Q_N = \frac{Q_P}{Q_0}$, we get the equilibrium relation shown in equation (2.6), which only depends on parameters that can be obtained experimentally.

$$V_{fe} = \frac{-3\sqrt{3}}{2}\frac{V_c}{Q_0}Q_P + \frac{3\sqrt{3}}{2}\frac{V_c}{Q_0^3}Q_P^3 \tag{2.6}$$

In this work we want to understand the interaction of the ferroelectric emerging technology with different circuital elements. Landau theory captures the ferroelectric dynamics; however, on a system perspective, it does not give insight about the polarization dynamics and, eventually, it is necessary to rely heavily in numerical solvers. In the next section we will propose a Landau-equivalent circuit-compatible compact model of ferroelectric materials, which, while providing an accurate description of the ferroelectrics dynamics, it also provides intuitive insights and therefore allows a better understanding of the system dynamics.

2.2 Positive Feedback Model of Polarization Switching

Our proposed modeling framework is based on the microscopic mechanism of domain switching, built upon the interaction of the electric-dipoles with the local electric field [46]. According to the microscopic picture of ferroelectric domain switching, the dipole moment is affected by the local electric field (E_{local}) that has contributions from the external electric field (E_{FE}) and the dipolar-interaction from the neighbors (E_{dipole}) (Figure 2.2 (a)). Because the latter is proportional to Q_P, and Q_P is proportional to E_{local}, this creates a positive feedback-loop as is shown in Figure 2.2 (b) in which E_c is the coercive electric field and K_1 and K_2 are constants related to the anisotropy parameters of Landau theory.

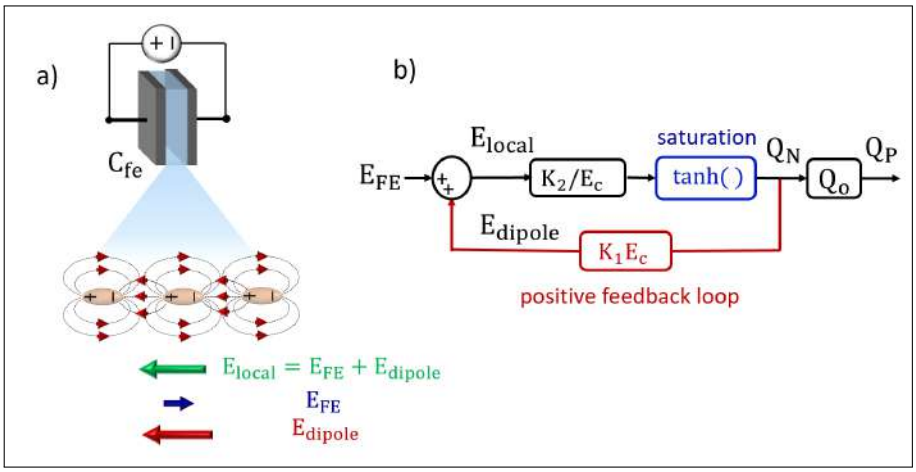


Figure 2.2. Arise of positive feedback loop from microscopic mechanism of domain switching. (a) Toy model of the electric field interactions at dipole level in ferroelectrics and (b) Representation of positive feedback loop that arises from the electric field interaction.

To account for the nucleation latency, we introduce a delay (τ_{fe}) related to the frictional inertia (ρ) of the ferroelectric and, for simplicity, we transform the electric field to voltages as shown in Fig. 2.3 (a). This new system is described by equations: (2.7), (2.8) and (2.9). Combining those three equations we can get a description of the dynamics of the system given by equation (2.10) [1]. The Charge-Voltage (Q − V) relation of equation (2.10) is shown in Fig. 2.3 (b).

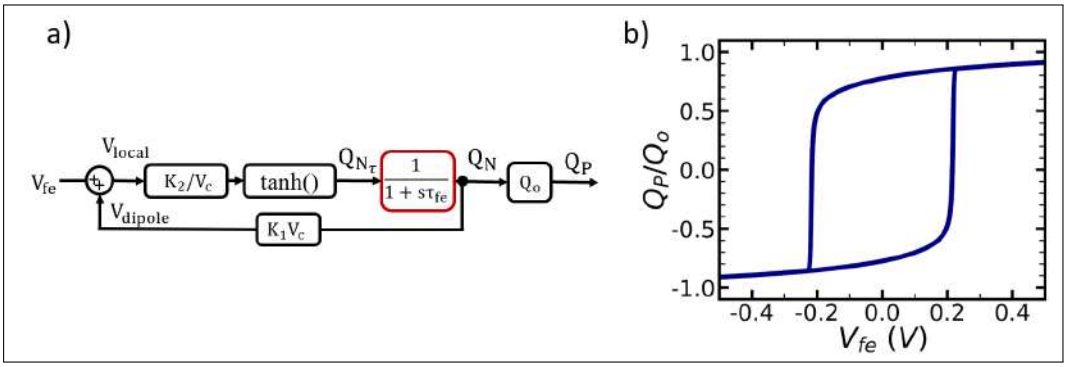


Figure 2.3. (a) Positive Feedback loop with incorporated nucleation latency and (b) Charge-Voltage relation of the positive feedback-loop.

$$V_{local} = V_{fe} + V_{dipole} \tag{2.7}$$

$$tanh^{-1}\left(Q_{N_\tau}\right) = \frac{K_2}{V_c}\left(V_{fe} + K_1 V_c Q_N\right) \tag{2.8}$$

[1] Note that Q_N is unit-less and can only take values between -1 and 1.

$$\frac{dQ_N}{dt} = \frac{Q_{N_\tau} - Q_N}{\tau_{fe}} \tag{2.9}$$

$$\frac{dQ_N}{dt} = \frac{1}{\tau_{fe} K_1 V_c} \left(V_{fe} - \frac{V_c}{K_2} tanh^{-1}\left(Q_{N_\tau}\right) + K_1 V_c Q_N \right) \tag{2.10}$$

In the steady-state scenario, i.e. $\frac{\mathrm{dQ_N}}{\mathrm{dt}} = 0$ and $\mathrm{Q_{N_\tau}} = \mathrm{Q_N}$, equation (2.10) becomes (2.11).

$$V_{fe} = \frac{V_c}{K_2} tanh^{-1}\left(Q_N\right) - K_1 V_c Q_N \tag{2.11}$$

Using Maclaurin expansion over $\tanh^{-1}$ in equation (2.11), we can demonstrate that the Landau theory is equivalent to the a positive feedback loop that arises from the microscopic mechanism of domain switching (see Appendix C). Fig. 2.4 (a) shows a direct comparison between both theories. Each theory has its own parameters; however, those parameters are equivalent and can be easily related. Fig. 2.4 (b) shows the $\mathrm{Q_N} - \mathrm{V}$ relation for both theories [2].

The biggest advantage of understanding the Landau theory as a positive feedback loop is that it opens the path to a simple an intuitive compact circuital implementation. The idea of understanding the Landau Theory as a positive feedback is not new, in fact, it was mentioned on the abstract of the seminal paper of 2008 [37]; however, here, we are showing how we derive the positive feedback from the microscopic mechanism, and develop an intuitive compact-model that takes direct advantage of the positive feedback loop.

[2]The plot done in Fig. 2.4 (b) presents an S shape but it does not mean that it is stable. To do this plot $\mathrm{Q_N}$ is the independent variable. We will go deeper into stability in the next section.

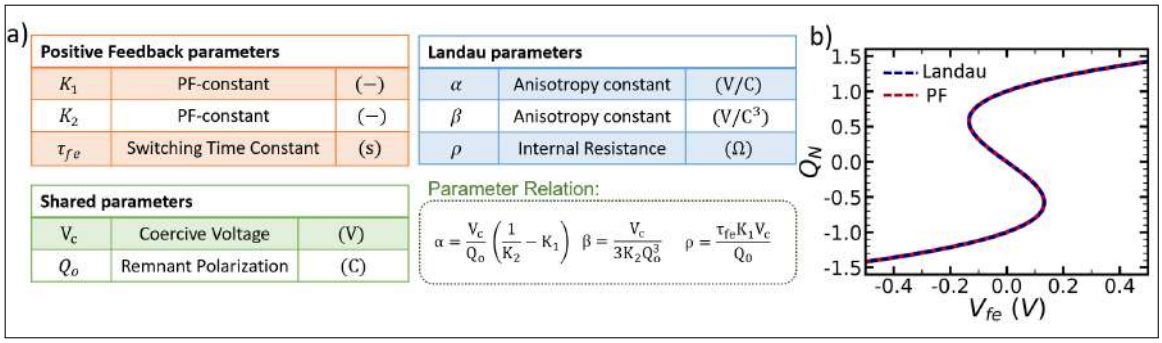


Figure 2.4. (a) Summary of parameters used in the Positive Feedback (PF) theory of domain switching and their relation with the anistropy parameters of Landau theory. (b) Charge-Voltage relation for Landau Theory and PF Theory.

2.3 Circuit Implementation of Ferroelectric Switching Dynamics

The ferroelectric capacitor (C_{fe}) (Figure 2.5 (a)) is modeled as a linear capacitor (C_D) in parallel with a current source (I_P) that depends on the total polarization charge from multiple domains (Figure 2.5 (b) and (c)). Each domain is modeled using a positive feedback loop with a characteristic τ_{fe} delay that accounts for the nucleation latency. The netlist for a single domain implementation can be seen in Appendix F.

We consider a Gaussian distribution of coercive voltages, polarization charges and nucleation times in the multiple domains scenario, and we calibrate our model to experimental data as is shown in Fig. 2.6.

The first applications of this model were for neuromorphic computing as is shown in Appendix G; however, in this work we will focus on the NC phenomena.

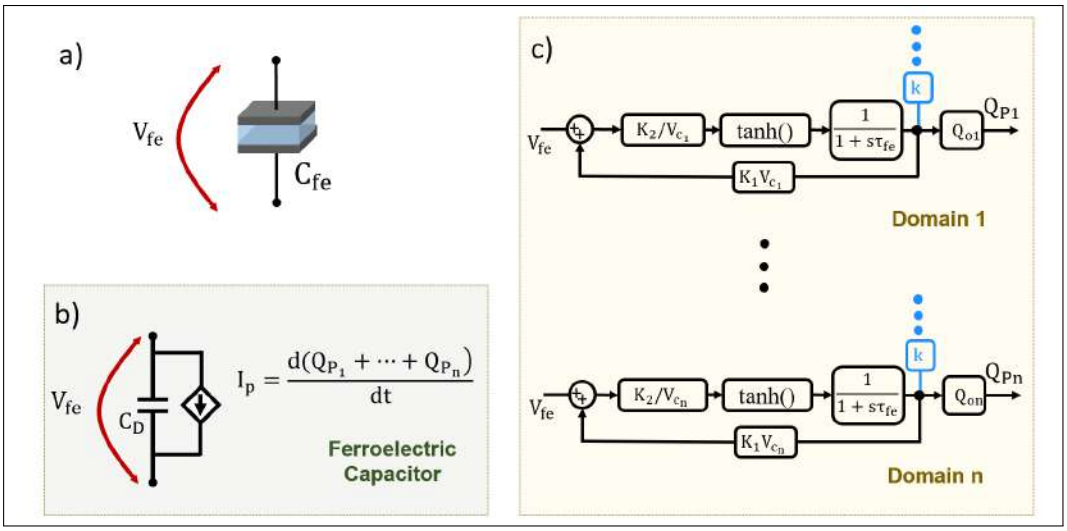


Figure 2.5. (a) Ferroelectric capacitor, (b) Circuit equivalent for a ferroelectric capacitor with two components in parallel: polarization current source and linear capacitor, and (c) positive feedback loops that define the switching dynamics of each one of the domains. k is a domain coupling factor that define the interaction between the domains.

2.4 Negative Capacitance

As we explained in chapter 1, the goal of NC is to improve the electrostatics of the transistor, which will show up as an improved SS, allowing to further reduce the supply voltage of transistors and therefore reducing the power dissipation of electronics. Also, better electrostatics enable to continue the scaling down of the transitor dimensions.

First, we will introduce the traditional explanation of the NC phenomenon based on the energy landscape of ferroelectric (see for example [17]). Capacitance is defined as the rate of increase of the charge (Q) with the voltage (V) and can also be derived from the free energy (U), as shown in equation (2.12). For a positive capacitor, Q increases linearly with V and the related energy landscape is a parabola (Fig. 2.7 (a)). In a negative capacitor Q decrease while V increases; thus, the energy landscape

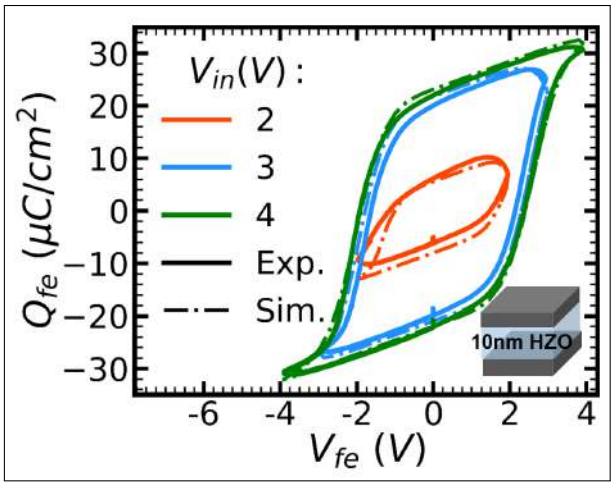


Figure 2.6. SPICE multi-domain model calibrated with a 10 nm HZO capacitor. By keeping the parameters constant, the model can capture multiple sub-loops.

is an inverted parabola (Fig. 2.7 (b)). According to Landau's ferroelectric model, the energy landscape of a ferrorlectric material (Fig. 2.7 (c)) has two energy minima and an invert curvature around Q = 0, which indicates a negative capacitance region; however, unstable.

$$C = \left[\frac{d^2U}{dQ^2}\right]^{-1} \tag{2.12}$$

Although unstable negative capacitance in a ferrolectric will show a steeper slope (lower SS), thanks to the polarization charge boost, the hysteresis, that we see in charge-voltage characteristics of the ferroelectric, will show up in the $I_d - V_g$ of the transistor, providing a device with two different (V_T) (Fig. 2.8 (b)). Although this device may be useful for memory applications, it can not be used for logic. If we can stabilize the NC region, the Charge-Voltage relation of the ferroelctric will present a hysteresis-free S shape; thus, providing a hysteresis-free steep-slope $I_d - V_g$ (Fig. 2.8 (c)), this device is known as NCFET.

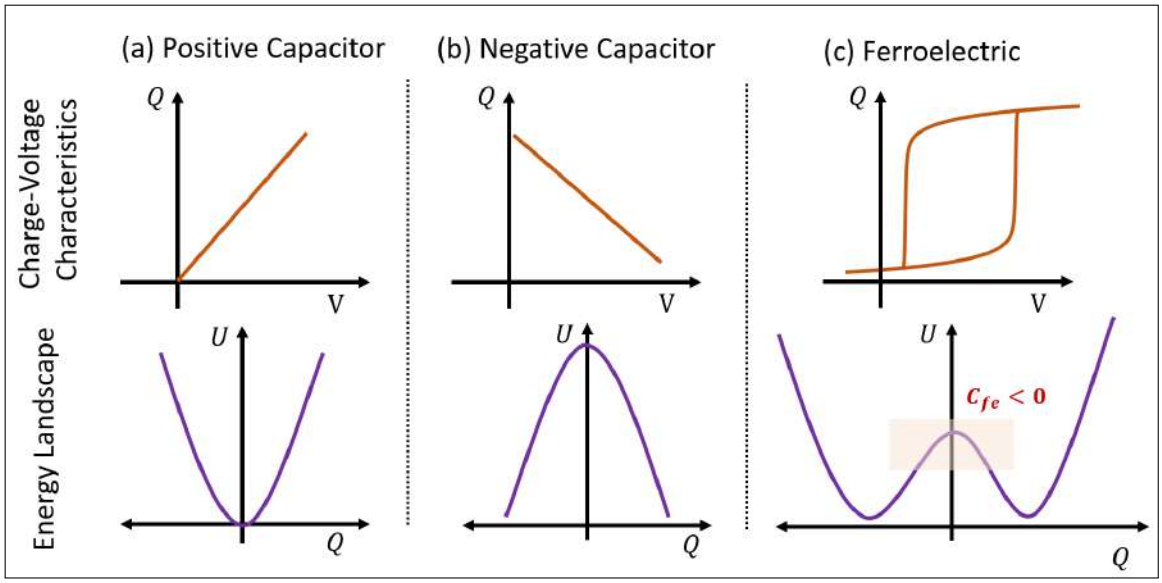


Figure 2.7. Charge- Voltage relation and Free Energy-Charge relation for: (a)positive capacitor, (b) negative capacitor, and (c) ferroelectric material.

So, the question is the following: is it possible to stabilize the negative capacitance region? By adding a capacitor in series with a ferroelectric, NC theory tells us that this should be possible. The free energy of a positive linear capacitor is given by equation (2.13); thus, a higher capacitance (C) will show up as a flattened parabola in the energy landscape (Fig. 2.9 (a)). By adding a ferroelectric in series (Fig. 2.9 (b)) and a positive linear capacitor (Fig. 2.9 (c)), the overall result will be a positive linear capacitor with a bigger capacitance than that of the ferroelctric and the positive linear capacitor. Although, this is not how normally series capacitors works, in this way the ferroelctric capacitance is negative and stable.

$$C = \frac{Q^2}{2C} \tag{2.13}$$

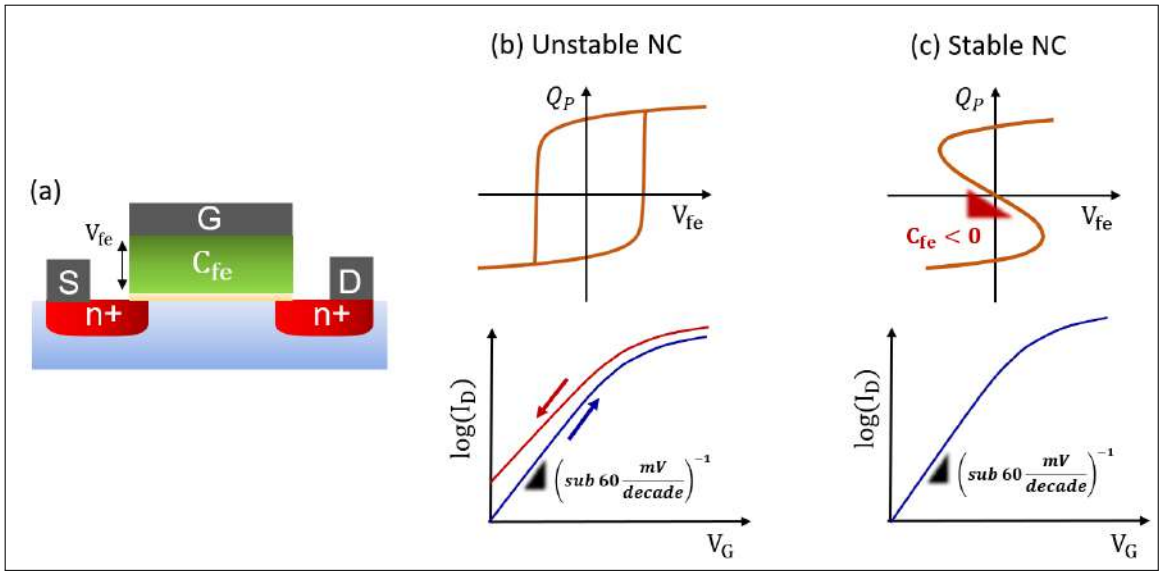


Figure 2.8. (a) MOSFET with a ferrolectric in the gate stack. Ferroelectric Charge-Voltage relation and consequent $I_d - V_g$ for: (b) Unstable ferroelctric fet (FeFET)and (c) Stabilized ferroelectric negative capacitance fet (NCFET).

2.4.1 Theory of Negative Capacitance from a Circuits Perspective

In this section, based on our circuital-compact model, we will present the theory of NC in a more intuitive circuit perspective; however, mathematically equivalent ([47], [48] and [49]).

NC arises when a linear capacitor (C_S) is in series with the ferroelectric capacitor (Fig. 2.10 (a)). When polarization switches, the internal node (V_{int}) charges up, which decreases the voltage across the ferroelectric (V_{fe}) providing a negative feedback. Then, the scenario is the following: we have an intrinsic positive feedback loop given by the dipolar interaction, and a negative feedback loop given by the series capacitor as shown in Fig. 2.10 (b). Using the dynamics of the system presented in

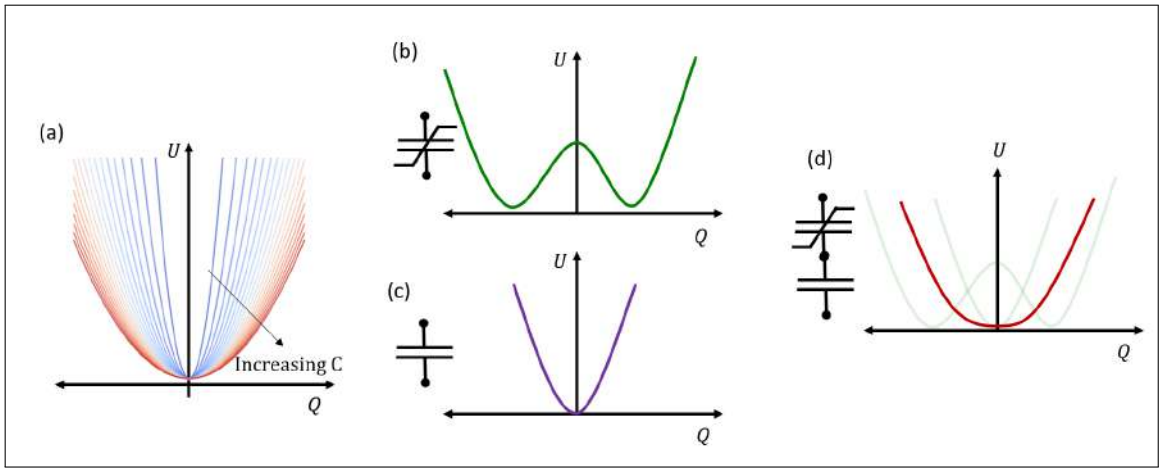


Figure 2.9. (a) Energy landscape of positive linear capacitor with different capacitance values, (b) circuital symbol and energy landscape of a ferroelectric, (c) circuital symbol and energy landscape of a linear capacitor, and (d) result of placing in series a ferroelectric and a linear capacitor.

equation (2.14) we can do a linear stability analysis[3] and derive the critical capacitor value to ensure stability, as shown in equation (2.15) (Appendices D and E). Figure 2.10 (c) shows the polarization switching dynamics: (i) when the the stability condition is not met (blue), and (ii) when the stability condition is met (red). When the stability condition is met, there is no hysteresis in the $Q_N - V_{fe}$ relation. This is important because, as we showed previously, the hysteresis that we see in the $Q_N - V_{fe}$ relation, shows up in the $I_D - V_G$ when we add the ferroelectric material to the gate stack of a transistor [49].

$$\frac{dQ_P}{dt} = \frac{Q_o}{\tau_f e K_1 V_c}\left(V_{in} + K_1 V_c Q_N - \frac{V_c}{K_2} tanh^{-1}(Q_N) - \frac{Q_o Q_N}{C_s}\right) \tag{2.14}$$

[3]The linear stability analysis is based on taking the eigenvalues of the Jacobian matrix and imposing that the real parts have to be less than zero.

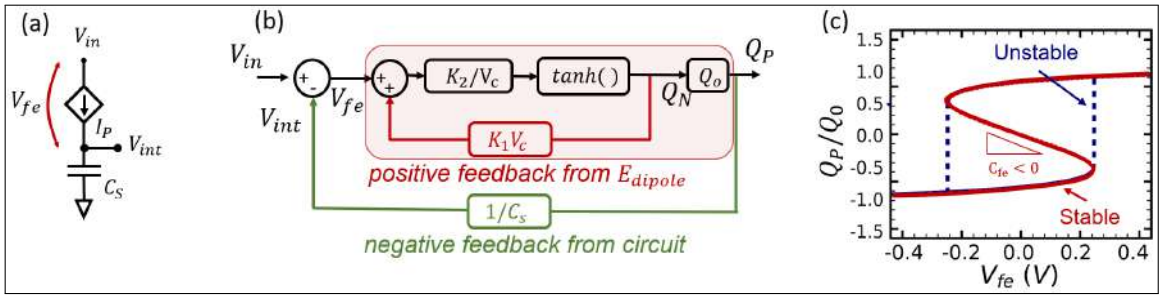


Figure 2.10. (a) Circuit model of an ideal ferroelectric with a capacitor in series, (b) Schematic of the interaction between the intrinsic positive feedback loop given by the dipolar interaction and a negative feedback loop given by the series capacitor, and (c) polarization switching dynamics: (i) when the the stability condition is not met (blue) and (ii) when the stability condition is met (red).

$$C_s < \frac{Q_o}{V_c}\left(\frac{K_2}{K_1K_2 - 1}\right) \tag{2.15}$$

The analysis presented above, works only for the ideal case; however, in real devices there are several non ideal conditions that can play a role on the stability of NC: leakage, ferroelectric dielectric contribution, muti-domain decomposition of ferroelectric and non linear semiconductor capacitance. Most of these non ideal conditions can be overcome if incorporated to the design of the system; however, muti-domain decomposition can play an important role in NC stability and we will take a closer look into it [47].

2.4.2 Multi-domain Switching Dynamics

Ferroelectrics are known to decompose into multi-domain structures when combined with dielectric (DE) layers. According to the current understanding ferroelectric NC cannot be stabilized under such conditions, and the operation of the device

would be hysteretic. For this analysis we distinguish between two different scenarios depending on whether or not a metal inter-layer is present between the ferroelectric and the dielectric.

When no metal inter-layer is present, as shown in a toy model with only 2 domains in Fig. 2.11 (a), the dynamics of each domain is independent of one another, and the negative feedback is localized to each domain ($1/Cs_1$ and $1/Cs_2$ in Fig. 2.11 (b)), and each domain has to meet separately the single domain stability condition. The FE-DE stack dynamics can then be described as the additive contribution of each one of the domains.

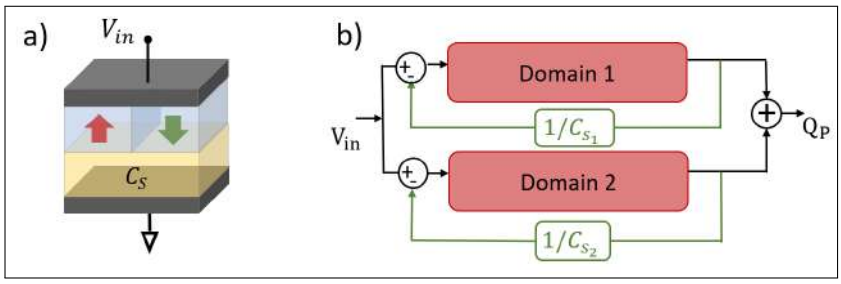


Figure 2.11. (a) Two domains in a MFIM capacitor without electrode metal layer and (b) different feedback mechanisms that arise inside the MFIM dielectric stack. Note that the negative feedback is local to each domain.

We then study the FE-DE stack dynamics, using a sample set of 100 domains (Figure 2.12). The distribution of the parameters are in-themselves Gaussian distributions used to fit experimental data. We then analyze three cases relative to the NC stability. (i) If the average value of the series capacitor is much larger than the stability threshold, none of the the domains is stable (Figure 2.12 (a)). (ii) If the

average value of the series capacitor is close to the stability threshold, then $\sim 50\%$ of the ferroelectric domains are stable; however, the cumulative FE-DE stack response cannot be considered stable (Figure 2.12 (b)). The stable domains are highly charged domains which is consistent with the stability condition (equation 2.15). When the negative feedback capacitor is much smaller than the stability threshold, we can achieve $\sim 95\%$ domain stabilization, and the cumulative FE-DE stack dynamics can be considered stable, even though a small portion of the domains remains unstable (Figure 2.12 (c)).

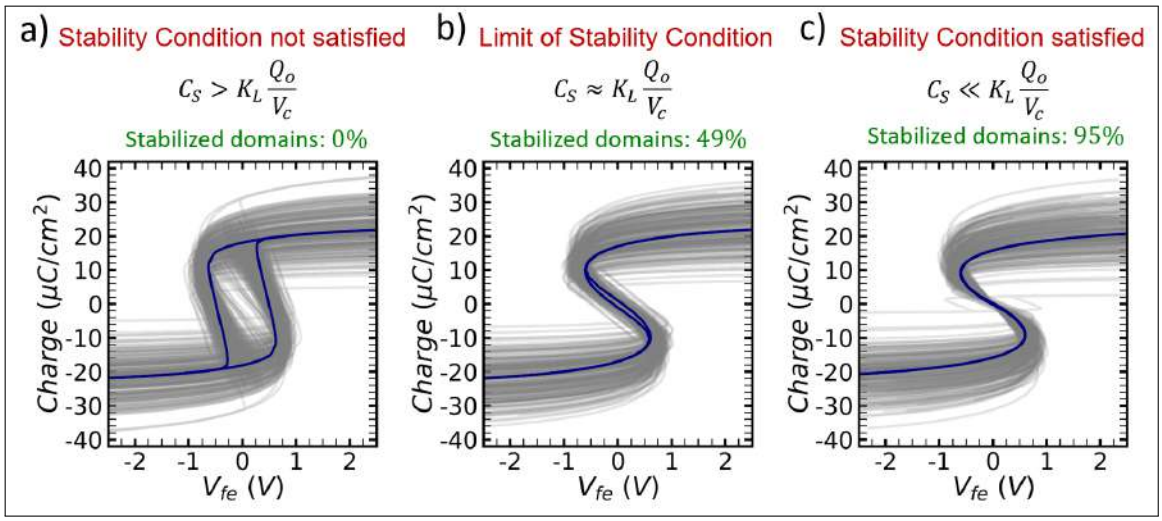


Figure 2.12. Study of the interaction of many domain polarization dynamics when no metal layer is present. (a) If C_S is much larger than the stability threshold none of the the domains is stable, (b) if C_S is close to the stability threshold and (c) if C_S is much smaller than the stability threshold.

A metal layer causes homogeneity in the interface between the FE and the DE in

the FE-DE stack (Fig. 2.13 (a)). This forces that all the domains share the negative feedback provided by the series capacitor (Fig. 2.13 (b)). Sharing of the negative feedback loop has two consequences: (i) there is a dynamic interaction between the domains while stabilizing, (ii) the cumulative series capacitance has to meet the stability criteria for all of the domains individually, which is the result obtained while deriving the stability criteria in a two-domain case.

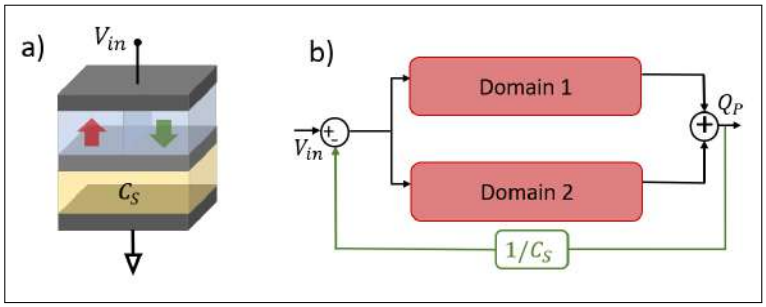


Figure 2.13. (a) Two domains MFIM capacitor with electrode metal layer, and (b) Different feedback mechanisms that arise inside MFIM dielectric stack. Note that the negative feedback is global.

Using the same conditions presented previously, in Figure 2.14 we study a sample of 100 domains at different stability conditions. In this case under any condition (Fig. 2.14 (a), (b) and (c)) the domains are unstable due to the sharing of the negative feedback loop as explained previously.

In conclusion, it is only possible to achieve the stability in the multi-domain scenario if no metal interlayer is present between the FE-DE interface.

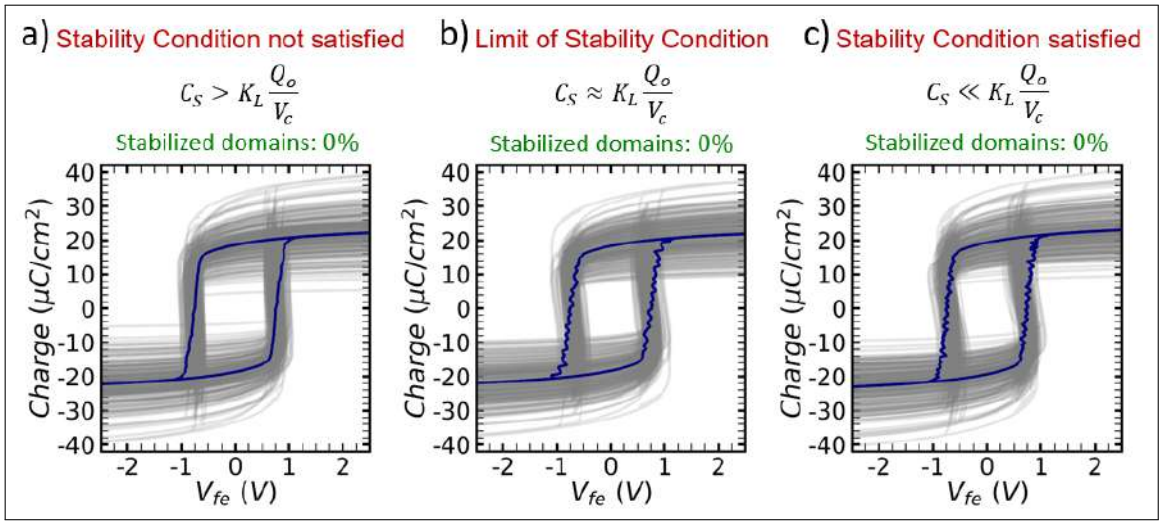


Figure 2.14. Study of the interaction of many domains polarization dynamics when a metal layer is present. In any case, none of the domains is stable.

2.5 Conclusions

In this chapter we introduced Landau theory of polarization switching. We also provided an equivalent framework to understand ferroelectric switching dynamics based on internal and external feedback mechanisms. Based on this framework, we developed a SPICE model that reproduces faithfully the experimental data, and we used this model to explore and understand the NC effect when a ferroelectric is introduced in the gate stack of a transistor. Finally, we expanded our analysis to the multi-domain scenario, and we concluded that it is only possible to achieve stability in the multi-domain scenario if no metal interlayer is present between the FE-DE interface.

CHAPTER 3

EXPERIMENTAL RESULTS OF FERROELECTRIC NEGATIVE CAPACITANCE FOR HIGH PERFORMANCE TRANSISTORS

> *"Since, however, the negative capacitance found in these systems hasn't been clearly demonstrated to extend to zero frequency, we shall not base our case solely on this, but will supply a more conceptually transparent discussion, though perhaps not an entirely conclusive one."*
>
> — Rolf Landauer (1976)

After the 2008 seminal paper on negative capacitance [37] and the proof of CMOS compatible ferroelectric field effect transistors in 2011 [43], the experimental demonstration of stable NC has become an object of major interest in industry and academia. However, although many groups have shown the presence of negative capacitance and the consequent sub-threshold slope reduction, the stability, and therefore hysteresis-free negative capacitance is still under discussion.

Fig. 3.1 shows an example of externally connecting a ferroelectric to the gate stack of a transistor. Fig. 3.1 (a) and (b), show an improvement in SS of the FeFET (blue) over the baseline (black). Also, when the charge-voltage relation is plotted, transient NC snap-backs can be seen (Fig. 3.1 (d)). However, as we explained previously, for logic applications it is not useful to have an hysteretic device because it is not possible to design any digital circuit with a varying threshold voltage.

In this chapter we will introduce some of the most relevant results of NC measurements and we will provide an explanation of these results using our simulation framework. First, we will briefly introduce some transient results on NC, then, we

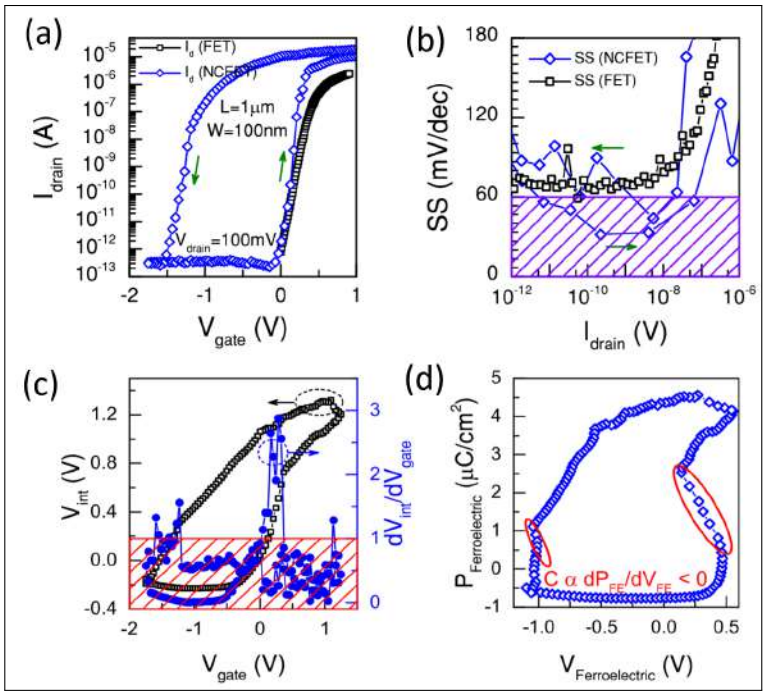


Figure 3.1. Common result presented in the literature of measurements of unstable NC (Fig. reproduced from [18]) based on an externally connected lead zirconate titanate (PZT) ferroelectric to a transistor. (a) Shows $I_{DS} - V_{GS}$ comparison with and without ferroelectric connected in series, (b) SS sub 60 mV/dec. in the presence of a ferroelectric, (c) internal voltage amplification due to the charge boost provided by the ferroelectric and (d) Charge-voltage relation with unstable NC snap-backs.

will show our own results in Metal-Ferroelectric-Insulator-Metal (MFIM) structures and, finally, we will show our NCFETs in which we show improved EOT.

3.1 Transient Measurements of NC

Once the concept of Negative Capacitance was on the table, the first step was to prove that NC can be shown experimentally. The first experimental setup was a series combination of a ferroelectric capacitor and a resistor [19] (Fig. 3.2 (a)). In this case the series resistor provides the negative feedback required for the NC phenomenon to arise (Fig. 3.2 (b)). Although NC can be observed using this experimental setup,

as shown in Fig. 3.2 (c) reproduced from [19] , the voltage drop across the resistor decreases over time which is a transient phenomena. Furthermore, by analyzing the dynamics of the system, the series resistor end up being in series with the internal resistance (equation 3.1); thus, causing a delay on the ferroelectric switching time.

$$\frac{dQ_P}{dt} = \frac{1}{\rho + R_S}\left(V_{in} - \frac{V_c}{K_2} tanh^{-1}\left(Q_P/Q_0\right) + K_1 V_c \frac{Q_P}{Q_0}\right) \tag{3.1}$$

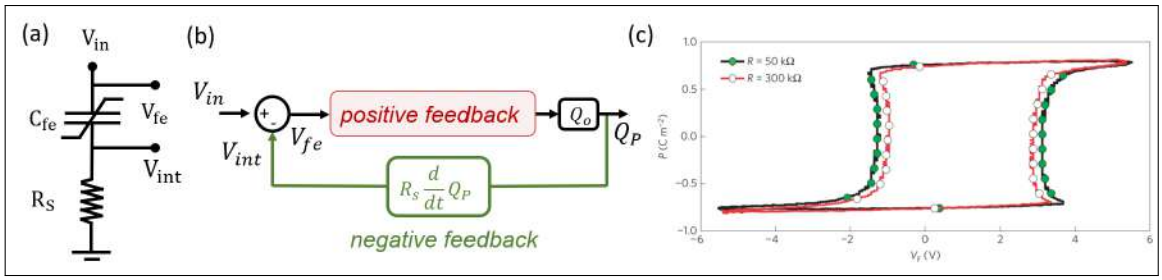


Figure 3.2. (a) Circuital topology used to measure transient NC, (b) equivalent feedback schematic and, (c) Charge-voltage relation with negative capacitance regions, reproduced from [19].

3.2 Stable NC in MFIM Structures

Fig. 3.2 (b) shows that the negative feedback depends on the derivative of the charge; thus, it has a transient component. The obvious choice is to replace the resistor with a circuital element that can integrate the polarization current, i.e. a capacitor. However, when experiments were done with a capacitor in series (Fig. 3.3 (a)), in Metal-Ferroelectric-Insulator-Metal (MFIM) structures, hysteresis still

showed up, as shown in Fig. 3.3 (c)

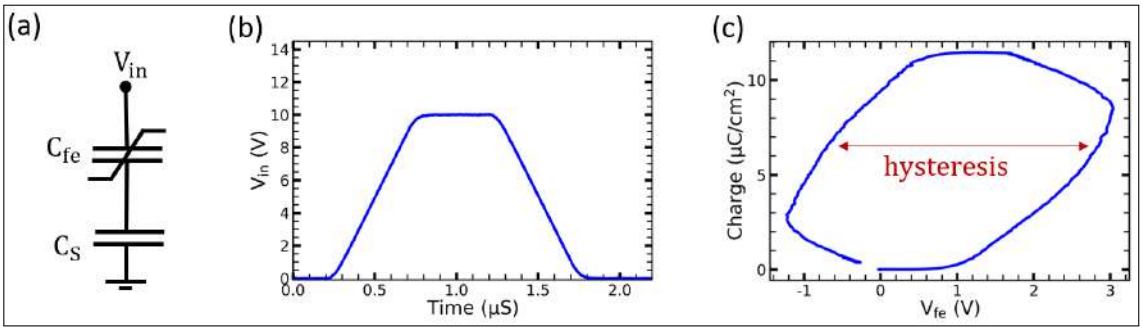


Figure 3.3. (a) Circuital topology used to measure NC, (b) input voltage applied to the MFIM structure and, (c) Charge-voltage relation with negative capacitance regions, but with hysteresis

By confronting the theory of NC with these strong results (repeated by multiple labs across the world, see for example [18], [50], and [51]), three main opinions appeared. The first one argued that NC may be an artifact of Landau phenomenological theory [51], i.e. Landau can reproduce experimental results, however it can not be considered the actual function that describes the system dynamics, so, it doesn't make any sense to do an stability analysis using Landau theory to describe the dynamics. The second group, slightly more optimistic, says that it is possible for NC to appear, however there are too many non ideal conditions in Ferroelectric compared to the system that Landau represents [44], ergo we need to eliminate conditions, such as: leakage, multi-domain decomposition, domain interaction, etc. Finally, the third group, which highly overlaps with the second group, argues that it is better to just go to scaled transistors and let the data tell us.

We think that the mismatch between the simulation results and experiments is

due to the ohmic losses of the measurements. Most of the NC simulation frameworks (not our positive feedback model) eliminate the internal resistance by imposing the condition $dU/dQ_p = 0$ (see for example [52]), and then, impose charge neutrality to solve the system state. In real systems, we can not eliminate the internal resistance, so we need a measurement methodology to de-embed it.

In 2018 a group from NaMLab proposed a way to measure hysteresis-free negative capacitance ([22], [20]). The method consists on applying a train of pulses with increasing amplitude. The charge is extracted at the max Q_D point which is defined as the max charge minus the residual charge. Using this method it is possible to extract the intrinsic S-shape as shown in Fig. 3.4 (a), (b) and (c), however, no clear explanation why this method works was provided; furthermore, by extracting the data presented in [22] and plotting the charge versus voltage across the ferroelectric we realized that each one of the loops presents hysteresis, as shown in Fig.3.4 (d).

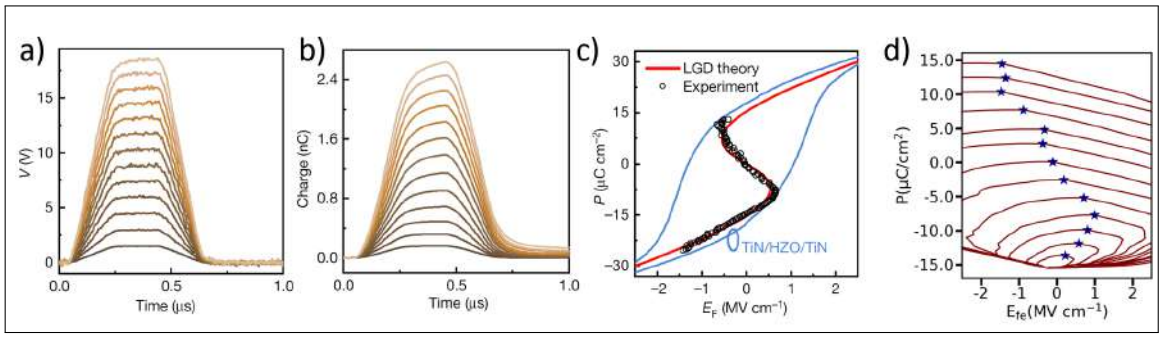


Figure 3.4. NC extraction method used by [20]. (a) Voltage-Time relation, (b) Charge-Time relation , and (c) Charge extracted-Electric field across the ferroelectric. (a), (b) and (c) are reproduced from [20]. (d) Extraction of complete Charge-Voltage relation for the entire waveform using data extracted from [20], hysteresis can be seen on each one of the loops.

TABLE 3.1

PARAMETERS OF MFIM STRUCTURE

Parameter	Magnitude
MFIM area	$100(\mu\text{m}) \times 100(\mu\text{m})$
HZO composition	5 : 5
HZO thickness	10(nm)
HZO Dielectric Constant	19
HZO Polarization Charge	$21(\mu\text{C/cm}^2)$
La_2O_3 thickness	10(nm)
La_2O_3 Dielectric Constant	20
Stability threshold	$3.9(\mu\text{F/cm}^2)$
Total Series Capacitor	$1.8(\mu\text{F/cm}^2)$

The first step towards understanding these experimental results is to replicate them with our own, designed and fabricated, MFIM structures, based on Hafnium Zirconium Oxide (HZO) as ferroelectric, and Lanthanum Oxide (La_2O_3) as insulator. All the design parameters are found in Table 3.1.

We need to decrease (C_s) so it is below the stability threshold. However, it can not be so small that it forces us to apply large voltages to effectively overcome the coercive voltage across the ferroelectric. Also, the thickness plays a fundamental role reducing the leakage which impacts the stabilization of NC [44]. We decided to use La_2O_3 which has a high dielectric constant allowing us to keep a thick insulator layer without excessively reducing the value of (C_s). This structure accomplishes all the requirements for presenting stable NC (Fig. 3.5 (a)). We then applied a series of pulses as shown in Fig. 3.5 (b), and we measured the total charge by integrating

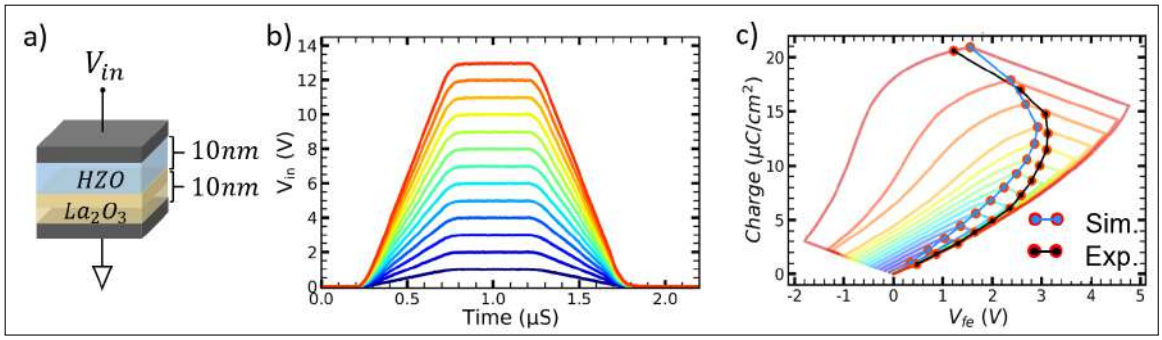


Figure 3.5. (a) HZO – La_2O_3 MFIM structure used for measurements, (b) Input voltage waveforms applied to the structure and (c)Q – V relation and extraction at max charge point for the experiments and simulations.

the current. Using the measured charge and the known capacitance of La_2O_3 we were able to get the voltage drop across the ferroelectric (V_{fe}) and plot the relation between the charge and V_{fe} (Fig. 3.5 (c)). In parallel we performed simulations with the same structure as shown also in Fig. 3.5 (c). Although we were able to obtain regions of NC, the charge-voltage relation presents hysteresis similar to the hysteresis extracted from [20]; furthermore, when we extract at the point of maximum charge, we also see how the S-shape arises. We observed the same results in the experiments and in the simulations.

We did our experiments at a rise time such that the leakage did not have a significant impact. In Fig. 3.6 we show the experimental results when the rise time is changed (×4). In Fig. 3.6 (a) we show the Voltage-Time relation with two different rise times and in Fig. 3.6 (b) the extraction for each different rise time.

In 1976 Rolf Landauer argued that the hysteresis on negative capacitance measurements can arise from ohmic losses given by internal damping mechanisms or external series resistivities [21]. Fig. 3.7 shows a schematic of the charge-voltage relation given by the ohmic losses, which was presented in his original paper on this

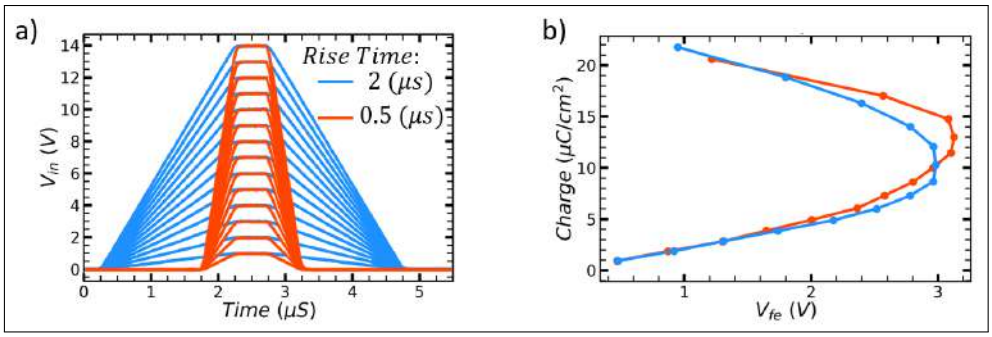


Figure 3.6. Experimental results of varying the rise time. (a) Shows the input Voltage-Time waveform and (b) shows the extracted Charge-Voltage relation for each one of the wave-forms.

topic.

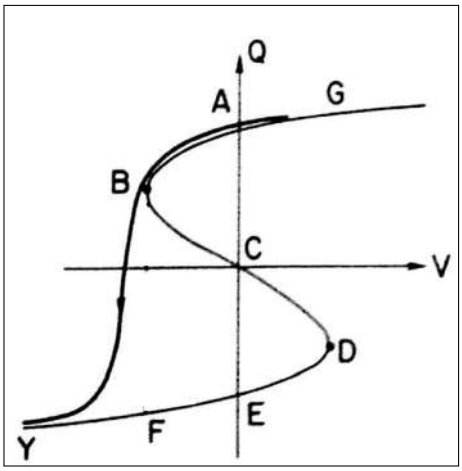


Figure 3.7. Landauer's original figure explaining ohmic losses [21]. The heavy line to the left shows the measured Q − V relation. The horizontal difference between the S-shape and the measured value is dropped in ohmic losses, or in soft model damping.

Starting from Landauer's explanation we realize that ohmic losses have not been considered while calculating the voltage across the ferrelectric (V_{fe}). We will refer as $V_{fe_{int}}$ when the voltage across the ferroelectric does not include the ohmic losses and as $V_{fe_{ext}}$ when it does (Fig. 3.8 (a)). The ohmic losses (R_S) have two components: the resistance given by the experimental setup and the internal resistance of the device (ρ) as shown in Fig. 3.8 (a). Applying a train of pulses we see that the relation $Q - V_{fe_{int}}$ has always the same S-shape, however the relation $Q - V_{fe_{ext}}$ has a dependence on the value of the external resistance in which $R_{S1} > R_{S2} > R_{S3}$ (Fig. 3.8 (b)).

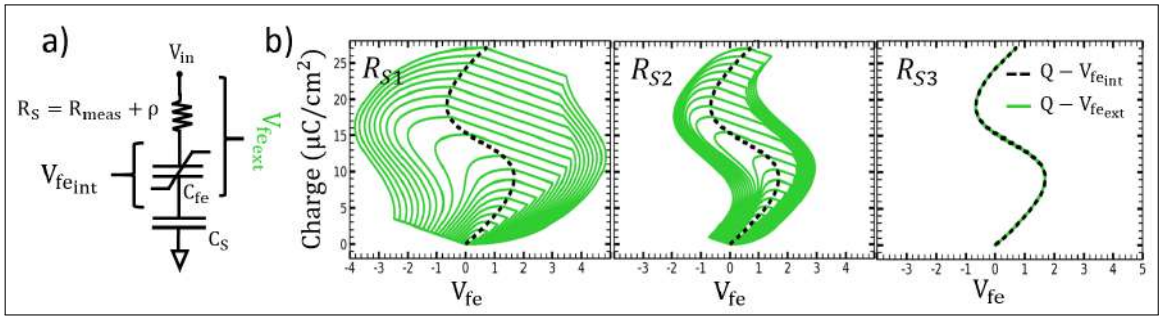


Figure 3.8. (a) Circuit level abstraction of MFIM structure with explicit ohmic losses in series with an ideal ferroelectric capacitor, (b) Impact of varying the ohmic loses with $R_{S1} > R_{S2} > R_{S3}$. $V_{fe_{int}} = V_{fe_{ext}}$ when $R_S \approx 0$.

The relation between the internal and the external voltage is given by equation (3.2). The point at which $V_{fe_{int}} = V_{fe_{ext}}$ is when $dQ/dt = 0$, which is the exactly at the proposed point of extraction in [22].

$$V_{fe_{int}} = V_{fe_{ext}} - \frac{dQ}{dt} R_S \tag{3.2}$$

Using the simulation framework, we can prove the functionality of the extraction methodology, as shown in Fig. 3.9. We apply the same train of pulses (Fig. 3.9(a)) and we extract at the point of maximum charge (Fig. 3.9(b)). The Charge-Voltage relation can be seen in Fig. 3.9 (c) in which, the extracted Ext. $Q - V_{fe}$ using the method proposed by [22], is the same that $Q - V_{fe_{int}}$, while $Q - V_{fe_{ext}}$ presents hysteresis, following the same trend that we saw in our experiments (Fig. 3.5 (c)).

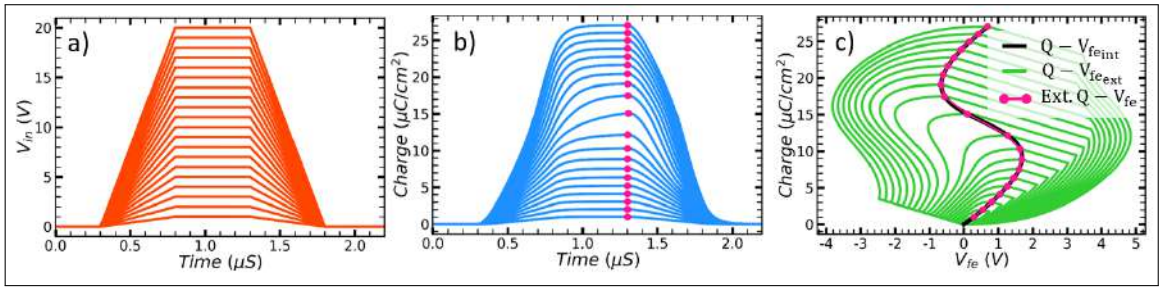


Figure 3.9. Simulation conditions to validate extraction method proposed in [22]. (a) Input voltage waveform applied to the simulated structure, (b) Charge-Time relation. In pink the point of maximum charge extraction is marked, and (c) $Q - V_{fe_{ext}}$, $Q - V_{fe_{int}}$ and extracted $Q - V_{fe}$ which overlays perfectly with $Q - V_{fe_{int}}$.

An alternative method to measure the S-shape directly would be to take a slow or DC measurement; however, this is not possible due to the leakage through the ferroelectric and the dielectric [44]. Under this perspective, we will never have a DC negative capacitance, however if the leakage time constant is much larger than the

switching time constant, we could theoretically get a hysteresis-free S-shape. Thus, the frequency trade-off is the following: if the frequency is very high, the ohmic losses will dominate and hysteresis will arise. On the other hand, if the frequency is low, leakage will dominate and hysteresis will arise.

3.2.1 Impact of Metal Inter-Layer in NC Measurements

To experimentally explore NC in the multi-domain scenario, we explore two different structures: one structure without the internal electrode (3.10) (a)), and a second structure with the internal electrode (3.10) (b)). For these experiments we used the same devices presented previously ($HZO - La_2O_3$ MFIM structures, Fig. 3.5)

We show experimentally that when no internal electrode is present, we can extract an S-shape from the voltage-charge measurements; however, when a metal electrode is present, no region can be seen as our simulation framework predicted.

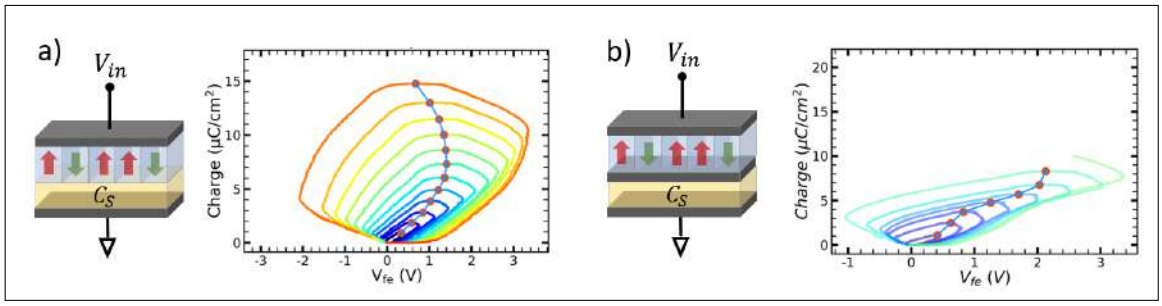


Figure 3.10. Experimental results to measure the impact of a metal electrode. (a) When no metal electrode is present the NC snap-back starts to appear, and (b) when the metal electrode is present no NC snap-back can be seen.

3.2.2 NC Measurements in MOS Capacitors

Direct DC measurements of NC remain elusive; however, previous works in MOS capacitor structures have shown that doping HfO_2 with Zr (3.11 (a)) can increase the dielectric constant of the high-κ beyond that of HfO_2 alone (see for example [53]). These results can be explained through the NC phenomenon. Figure 3.11 (b) shows the resulting C-V measurements for different compositions of $Hf_{1-x}Zr_xO_2$. The accumulation capacitance increases initially with Zr concentration up to 70% and then falls down (Table 3.2). This verifies that the NC charge boost is taking place.

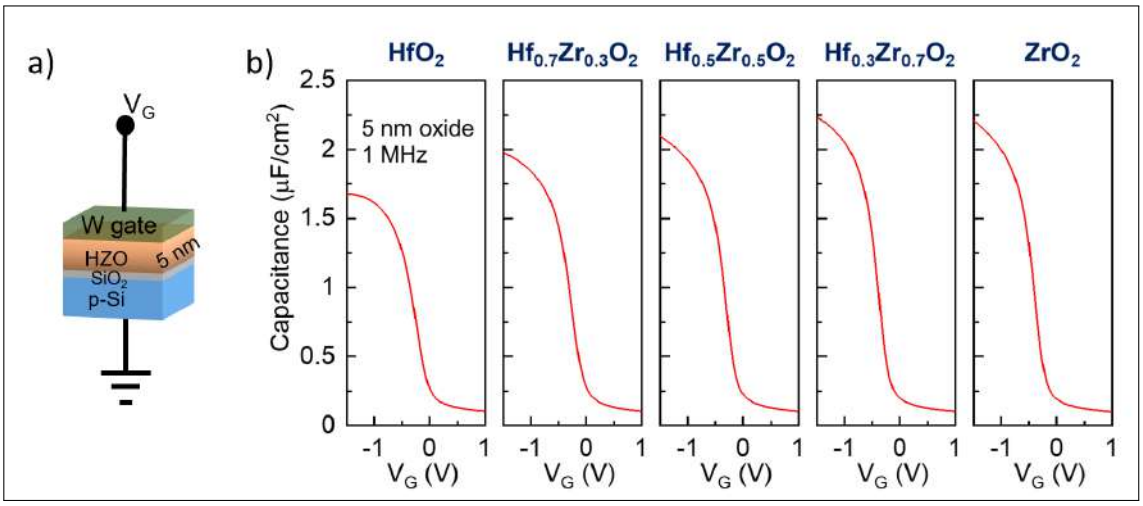


Figure 3.11. (a) MOS capacitor structure with HZO dielectric, and (b) CV curves at 1MHz for each one of five concentrations, measured from $50\mu m \times 50\mu m$ MOSCAPs.

Based on these results we decided to further explore the impact of HZO in transistors by building NCFETs and compare it with baseline high-κ HfO_2 devices.

TABLE 3.2

ACCUMULATION CAPACITANCES FOR HZO MOSCAPS

Parameter	HfO_2	HZO 7 : 3	HZO 5 : 5	HZO 3 : 7	ZrO_2
$C_{max}(\mu F/cm^2)$	1.64	1.88	2.00	2.16	2.12

3.3 HZO NCFETs Performance Evaluation

By allowing the equivalent oxide thickness to be reduced, high-κ dielectrics has allowed transistor scaling to continue past the limits of SiO_2-based gate stacks. However, band-gap of dielectric materials follow an inverse relation with κ as $E_g \approx \kappa^{-0.65}$ [54], limiting the choice of κ values to $20 \sim 30$, in order to avoid excessive direct tunneling gate-leakage, as shown in Fig. 3.12. High-κ dielectrics in their current form are reaching their own scaling limitations, such that further reducing the EOT will increase leakage to excessive levels, similar to the current densities ($> 100A/cm^2$) seen at the scaling limit of SiO_2-based dielectrics [55]. For 3 nm and beyond logic nodes, it is imperative to resume EOT scaling to meet the high performance, low-gate leakage (I_G) and improved reliability requirements of advanced CMOS.

An alternative choice would be reducing the EOT contribution from interlayer (IL) by IL scavenging, as shown in Fig. 3.13 (a). IL scavenging can potentially scale down the EOT [25] and increase the gate capacitance (Fig. 3.13 (b)). However, due to the closer proximity of the channel to the high-κ gate-dielectric, severe degradation in reliability (Fig. 3.13 (c)) and mobility (Fig. 3.13 (d)) is typically observed. Another option is to replace the SiO_2 interlayer by a high-κ dielectric; however, it has been shown that it also reduces mobility and shifts V_T [25].

The boost given by the NC phenomenon provides a path towards continue scaling

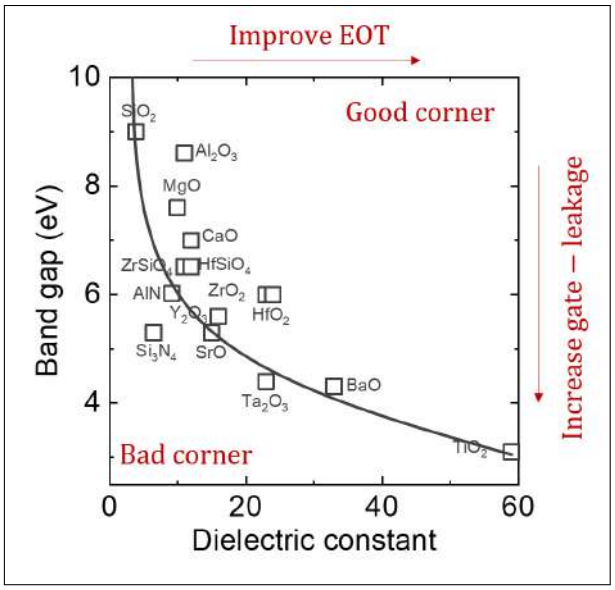


Figure 3.12. Band-gap of dielectric materials vs. dielectric constant. Figure adapted from [23].

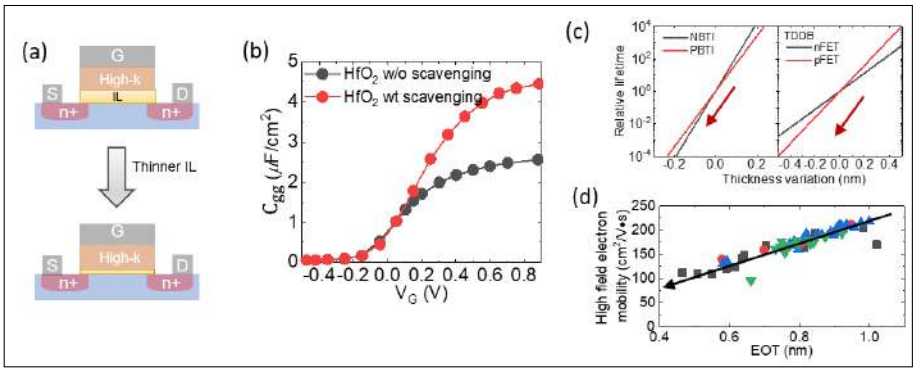


Figure 3.13. (a) IL scavenging by reducing SiO_2 interlayer, (b) capacitance boost by IL-scavenging (Figure reproduced from [24]), (c) degradation in reliability (TDDB: Time-dependent gate oxide breakdown, NBTI: Negative-bias temperature instability and PBTI: Positive-bias temperature instability) (Figure reproduced from [25]) and, (d) mobility degradation (Figure reproduced from [26]).

EOT without degradation in transistor performance. This approach can avoid the shortcomings of mobility and reliability degradation incurred in IL scaling approach and gate leakage increase from band gap reduction in traditional high-κ materials. First, we evaluate the impact in gate leakage using MOSCAPs. Using the optimal Zr concentration found previously (HZO 3 : 7), we fabricated 17 Å thickness MOSCAPs and we compare the result vs baseline (HfO_2). We see an improvement in gate capacitance (Fig. 3.14 (a)), thus an improvement in EOT, without increasing the gate leakage (Fig. 3.14 (b)).

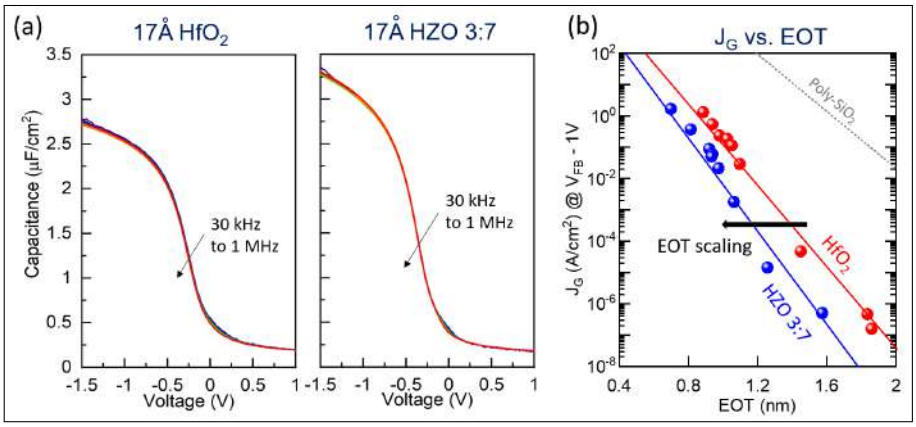


Figure 3.14. (a) Gate capacitance boost of HZO 3 : 7 over HfO_2, and (b) EOT boost at iso-gate leakage.

To further evaluate the impact of HZO in the gate stack we build NCFETs with different gate lengths (L_G) and compare it a with baseline high-κ HfO_2 devices with the same (L_G). Fig. 3.15 (a) shows an schematic of our own Trigate-on-SOI platform fabricated to investigate the impact of NC. The minimum channel length is 300nm. The devices are comprised of multiple parallel trigate channels, seen in the top-down

SEM of Fig. 3.15 (b), with line-widths of 75 nm as seen in the high-resolution TEM of Fig. 3.15 (c). The formed gate stack consists of a 25 Å of HZO thermal ALD film grown on $7-8$ Å SiO_2, with a 100A TiN PEALD film as the top metal gate.

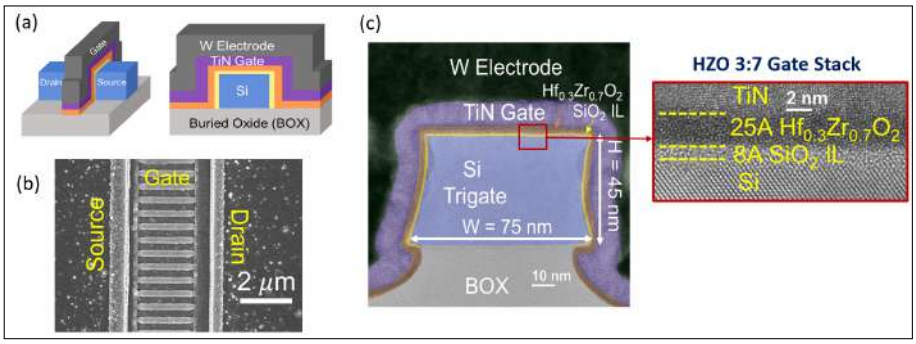


Figure 3.15. Trigate-on-SOI platform is fabricated to investigate HZO 3:7 as a MOSFET gate dielectric. (a) Trigate platform schematic, (b) Top-down SEM of Trigate channels, and (c) Trigate Cross-Section.

To evaluate the boost provided by HZO, we first perform an electrical characterization of the device, then, we measure the frequency dependent S-parameters, and, for an specific bias condition, we fit the measured S-parameters to a small signal circuit model of the transistor.

3.3.1 Electrical Characterization

Fig. 3.16 (a) shows the transfer characteristics of 300 nm gate-length trigate FETs with high-κ HfO_2 and higher-κ HZO as the gate-stack, showing better gate-control in HZO FETs with lower V_{TH}. $I_{DS}-V_{DS}$ characteristics for both HfO_2 and HZO trigate FETs (Fig. 3.16(b)) show excellent saturation with 20% higher V_{DS} at

1 V overdrive and $V_{DD} = 1$ V in HZO FET. 18% boost in g_m was also observed at 1 V V_{DS} in HZO (Fig. 3.16(c)). All these improvements can be understood thanks to a boost in gate capacitance (Fig. 4.3(a)).

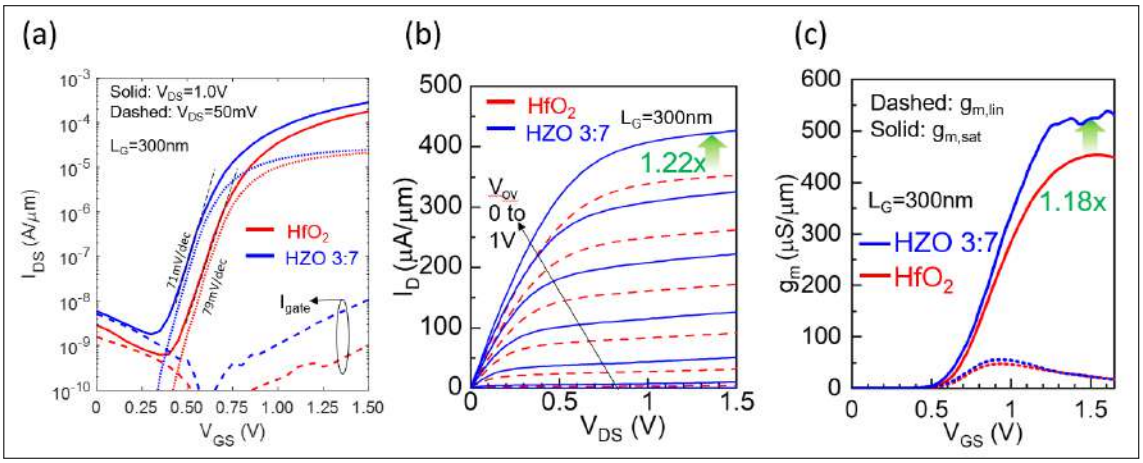


Figure 3.16. Comparison between HfO_2 and HZO devices: (a) $I_{DS} - V_{GS}$, (b) $I_{DS} - V_{DS}$ and (c) $g_m - V_{GS}$.

Fig. 4.3 (a) shows the CV measurements of our devices that we used to extract mobility. We conclude tha EOT scaling through HZO gate-dielectric has no adverse effect on carrier mobility (Fig. 4.3(b)), making HZO gate-dielectrics an excellent choice for high performance logic transistors.

Time kinetics of ΔV_{TH} under Positive Bias-Temperature-Instability (PBTI) stress (+1.3 V overdrive) at T = 85°C (Fig. 3.18(a)) showed 1.5× improvement in ΔV_{TH} at 1 ks stress time for HZO FET, indicating less electron trapping in bulk-oxide for HZO. Investigation of field dependence of ΔV_{TH} show 100× improved Time-to-Failure lifetime projections for HZO FET, which in turn allow 60 mV higher overdrive

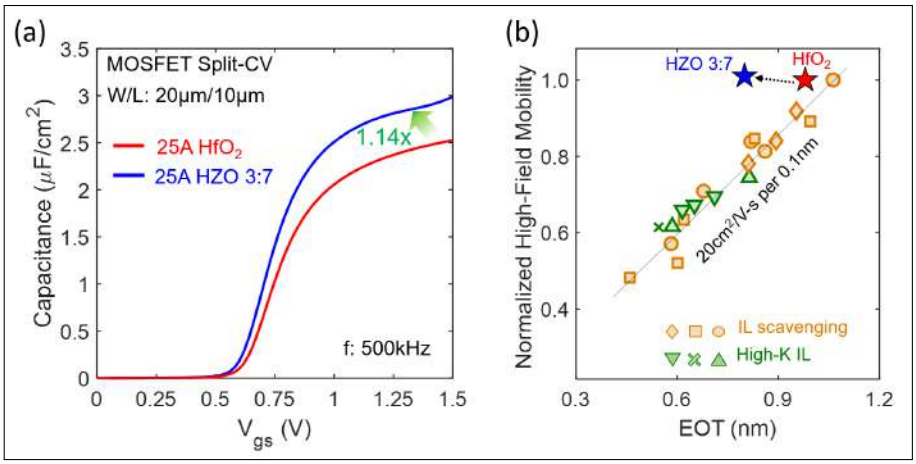


Figure 3.17. (a) 14% boost in C_{gg} at iso-overdrive obtained in higher-κ HZO FETs and, (b) EOT scaling through higher-κ is achieved w/o mobility degradation. Data obtained from [25].

operation for iso-reliability in 10 years.

The performance improvements of HZO FET over HfO_2 control device are summarized in Table 3.3.

3.3.2 RF Characterization

To investigate the device performance improvement in HZO FET at high-frequency range, S-parameters are measured on 300 nm L_G trigate FETs under different bias and frequency conditions ($V_{DS} = 1.0$ V, $V_{GS} = 0 \rightarrow 2$ V, $f = 0.01 \rightarrow 50$ GHz). To measure the S-parameters of the device we included Ground-Signal-Ground (GSG) structures in the sample (Fig. 3.19), and open and a short structures for pad-strip de-embedding purposes.

Three levels of de-embedding are necessary to access the intrinsic parameters of the device, a detailed explanation of the de-embedding process is shown in Appendix H. Once the probes and the pads have been de-embedded, and the transistor is in saturation, a linear small signal model can be used to capture the transistor behavior

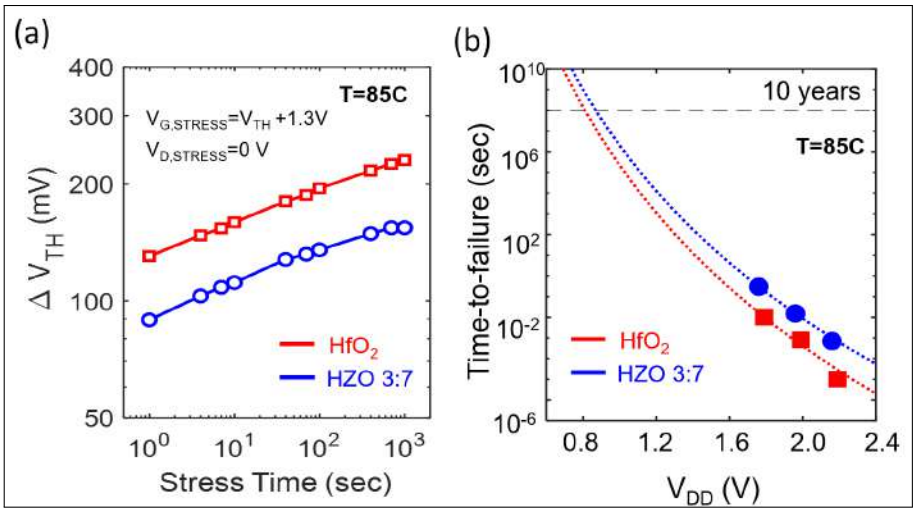


Figure 3.18. (a) Time-kinetics of ΔV_{TH} under PBTI stress (1.3 V overdrive) at T = 85°C show ∼ 80 mV lower V_{TH} degradation in HZO after 1 ks stress, and (b) Field activation of V_{TH} shift under PBTI stress at T = 85°C indicate ∼ 60 mV higher V_{max} can be achieved under iso-reliability (∼ 10 years) with HZO FET.

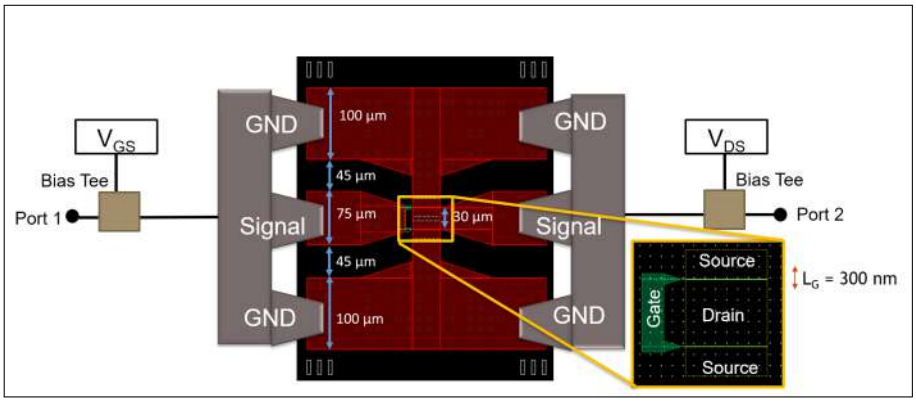


Figure 3.19. Ground-signal-ground (GSG) layout used to measure S-parameters of the devices.

for a specific bias condition (Fig.3.20 (a)). This model includes lumped extrinsic and intrinsic component. Using ADS Keysight we can optimize the component values to

TABLE 3.3

ELECTRICAL CHARACTERIZATION

	HfO_2	HZO3 : 7	Improvement
EOT(nm)	0.98	0.8	22%
SS_{min}(mV/dec.)	79	71	11%
$I_{D,sat}$@1.0V_{OV}(μA/μm)	338	407	20%
Max $g_{m,sat}$(μS/μm)	588	710	19%
μ_{eff}@1MV/cm($cm^2/V-s$)	200	205	1%
ΔV_{TH}@1ks(mV)	230	150	54%

fit the measured S-paramters at a specific bias, as shown in Fig. 3.20 (b).

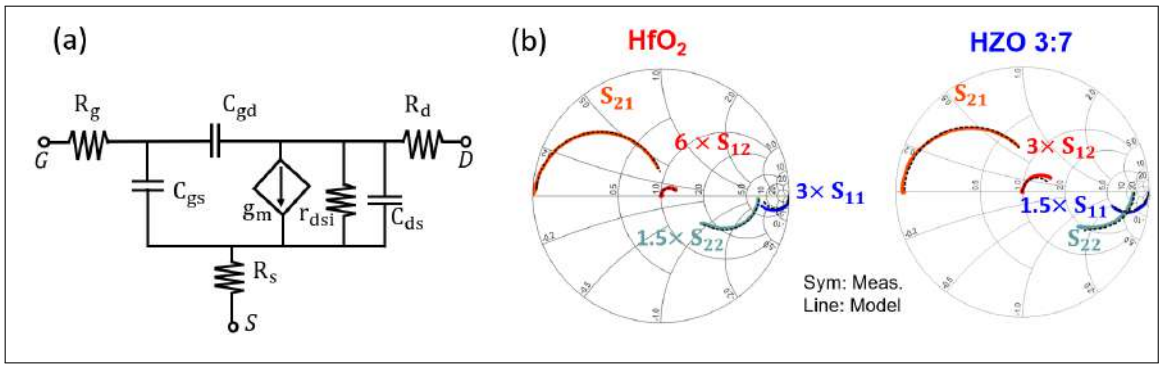


Figure 3.20. (a) Small-Signal Equivalent circuit model and, (b) measured and modeled S-parameters of HfO_2 and HZO gate-dielectric.

Also, using the linear model, we can derive expressions to calculate g_m (eq. (3.3)), C_{gs} (eq. (3.4)), C_{gd} (eq. (3.5)) and C_{gg} (eq. (3.6)). These extractions assume that $\omega^2 C_{gg}^2 R_g << 1$ and that R_g, R_s and R_d are small.

$$g_{m,RF} = Re(Y_{21})|_{\omega^2=0} \tag{3.3}$$

$$C_{gs} = \frac{Im(Y_{11}) + Im(Y_{12})}{\omega} \tag{3.4}$$

$$C_{gd} = -\frac{Im(Y_{12})}{\omega} \tag{3.5}$$

$$C_{gg} = C_{gs} + C_{gd} \tag{3.6}$$

Fig. 3.21 (a) shows the extraction methodology for g_m using the equation (3.3). We plot $Re(Y_{21})$ and then we extrapolate to $\omega^2 = 0$. Fig. 3.21 (b) shows the comparison of RF g_m for HZO and HfO_2 gate-dielectrics.

After de-embedding the effect of gate resistance, the improvement in C_{gg} and g_m is preserved up to the GHz range (Fig. 3.22), indicating the potential of HZO FETs for high-frequency applications.

f_T is determined from -20dB/dec extrapolation of current gain (h21) to 0 dB, as shown in Fig. 3.23 (a) and (b) for HfO_2 and HZO respectively. f_T is boosted by 11% in HZO FET compared to the baseline (Fig. 3.23(c)).

In order to understand the improvement in f_T we have to take a closer look on its expression shown in equation (3.7). Higher carrier concentration in HZO FET, due to higher-κ, contributes to a higher injection velocity (υ), also, there is an increased ratio between the parasitic capacitance C_{para} and oxide capacitance C_{ox} in higher-κ HZO FET, which shows up as boosted f_T.

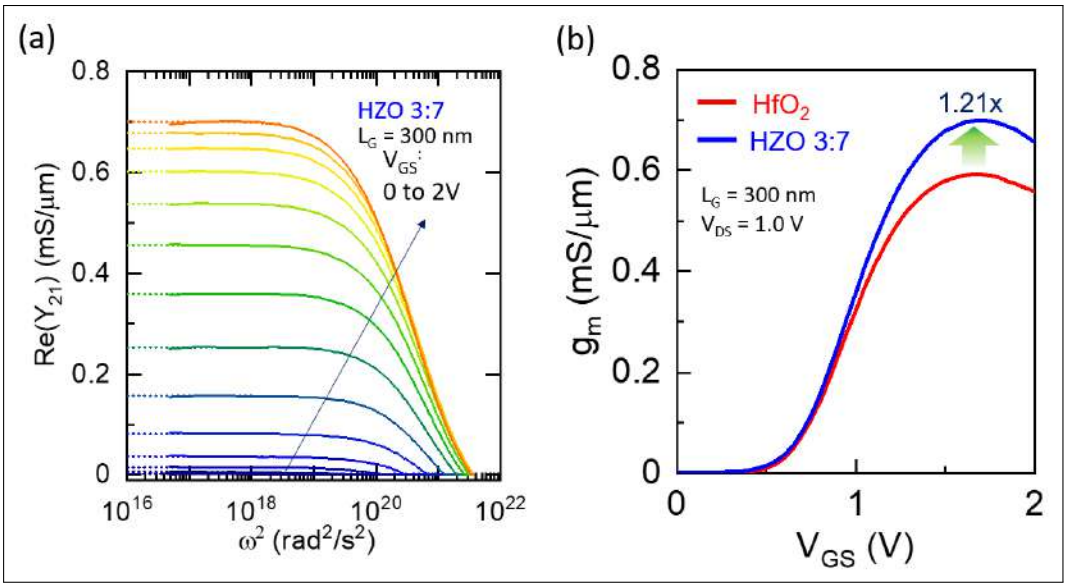


Figure 3.21. (a) Method for extracting $g_{m,RF}$ from $Re(Y_{21})$ in HfO_2 and (b) comparison between $g_{m,int}$ with $g_{m,RF}$

.

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_{ox}+C_{para}} = \frac{1}{2\pi}\frac{\upsilon}{L}\frac{1}{1+C_{para}/C_{ox}} \tag{3.7}$$

Using cold-fet de-embedding methodology (Appendix H) we can separate the intrinsic and extrinsic components of the FET (Fig. 3.24 (a)), and obtain the intrinsic transit frequency of the devices ($f_{T,i}$)(Fig. 3.24 (b)), which are: 26 GHz for HZO, and 24 GHz for HfO_2 gate dielectric. Using equation (3.8) we can obtain the injections velocities which are 0.33 cm/s for HZO, and 0.31 cm/s for HfO_2 (6% improvement).

$$\upsilon_{inj} = 2\pi \times L_{eff} \times f_{T,i} \tag{3.8}$$

Table 3.4 summarizes the intrinsic de-embedded parameters for HZO and HfO_2 gate dielctric.

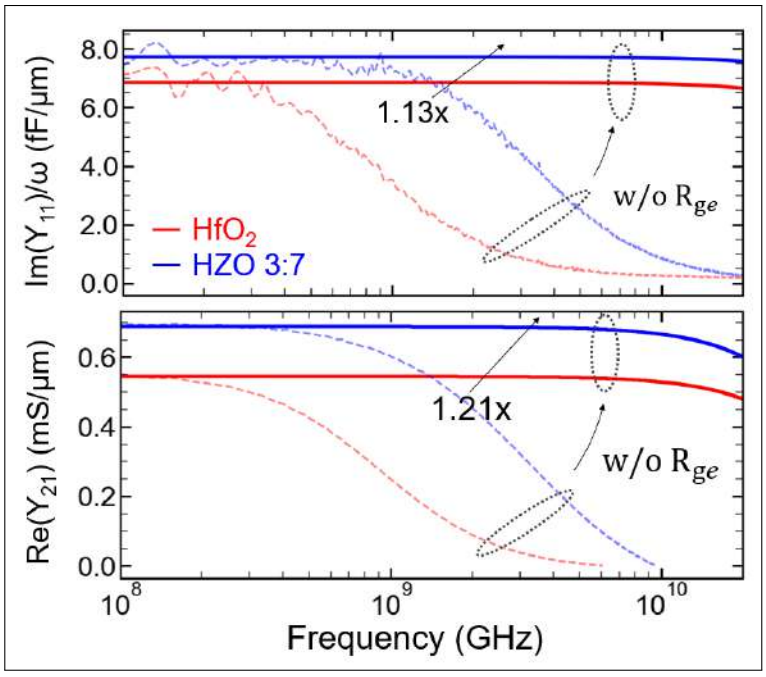


Figure 3.22. Measured and simulated Y-parameters shows that the improvement in C_{gg} (Imag(Y_{11})/ω) and RF g_m (Real(Y_{21})) is preserved even at GHz frequency range

3.4 Conclusions

In this chapter we experimentally explored the NC phenomenon. First we experimentally studied the arise of NC in a FE-DE stack and the methodology to de-embed ohmic losses from the measurements. We concluded that negative capacitance may be hysteresis-free in the multi-domain scenario if: 1) No internal electrode is present, 2) the series capacitance has been designed to meet the stability condition, 3) small ohmic losses on the ferroelectric switching dynamics, and 4) low leakage across the FE-DE stack. We also study the impact of introducing a ferroelectric layer to the gate stack of a MOSFET and discovered that it provides: 1) a higher-κ response compared to baseline (HfO_2), due to the NC phenomenon; 2) a 22 % EOT reduction in HZO FET over control HfO_2, without any mobility degradation, resulting in

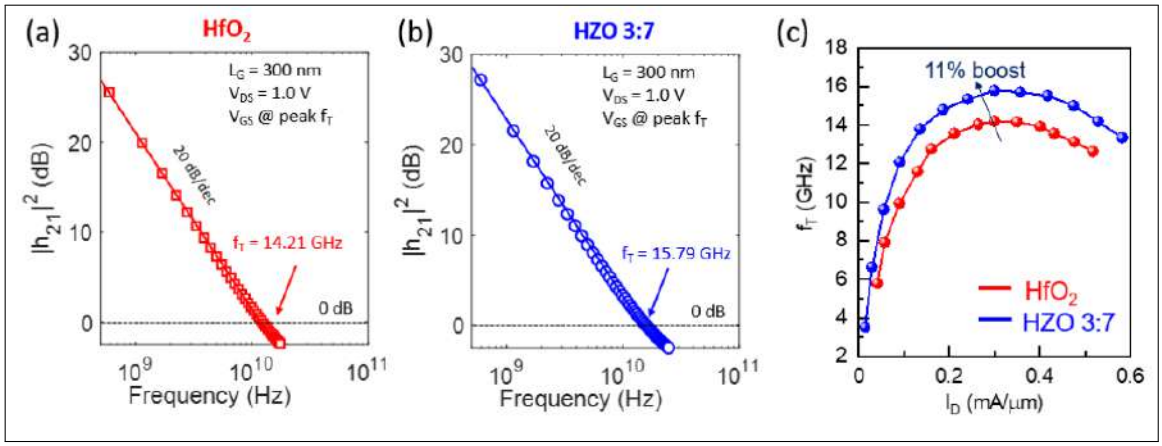


Figure 3.23. (a) Extrapolation of current gain for HfO_2 gate dielectric, (b) extrapolation of current gain for HZO gate dielectric, and (c) f_T of HfO_2 and HZO.

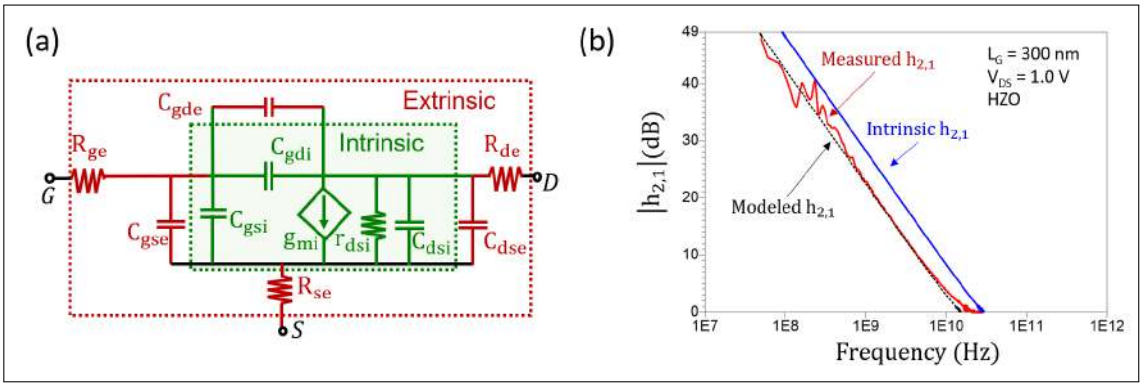


Figure 3.24. (a) Small-Signal Equivalent circuit model separating intrinsic and extrinsic components, and (b) difference in current gain (h_{21}) between intrinsic transistor and extrinsic+intrinsic transistor.

20% and 18% boost in drive-current ($I_{D,sat}$) and transconductance (g_m) respectively; 3) improved device reliability resulting in $\sim$ 60 mV higher V_{max} under iso-reliability than baseline; and 4) consistent enhancement in g_m arising from thinner EOT that

TABLE 3.4

EXTRACTED

	HfO_2	HZO3 : 7	Improvement
$g_{m,i}$ (mS/μm)	0.74	0.95	28%
$C_{gg,i}$ (μF/cm^2)	2.35	2.91	23%
υ_{inj} (cm/s)	0.31	0.33	6%

persists in the gigahertz frequency domain.

CHAPTER 4

DOUBLE-GATE W-DOPED AMORPHOUS INDIUM OXIDE TRANSISTORS FOR MONOLITHIC 3D CAPACITORLESS GAIN CELL eDRAM

> *"The architectural design of DNN accelerators tends to be more of an art than science due to the large design space and the lack of a systematic approach to explore it."*
>
> — Angshuman Parashar (NVIDIA)
>
> (2019)

In Deep Neural Networks (DNNs) the processing bottleneck is in the memory access. At the core of each convolution we find the multiply-accumulate operation which requires three memory reads and one memory write. S. Rabii in the plenary talk of VLSI 2019 [27] showed that for several relevant algorithms, most of the energy is lost on the memory and data movement. Fig. 4.1 refers two examples: the Hand Tracking algorithm and the Simultaneous Localization and Mapping (SLAM) algorithm. In both cases the energy consumed by the memory is more than 50% of the total energy.

Different hardware architectures and the many different strategies of scheduling operations and staging data on the same architecture (mapping), results in a huge variability of performance and energy efficiency. Fig. 4.2 (a) shows the histogram of the energy efficiency of various mapping of VGG convolutional network on a specific architecture. It can be seen that the energy efficiency can be improved by more than a factor of 15×, by efficiently deploying the workload into the hardware.

Multiple levels of memory hierarchy allows local low-energy/latency data accesses, at a cost of low capacity (Fig. 4.2 (b)). Managing the trade-offs of correctly sizing

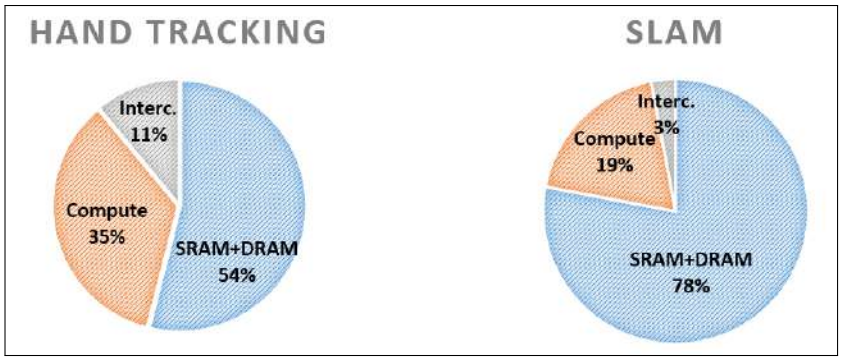


Figure 4.1. Popular AI algorithms energy breakdown. Figure modified from [27].

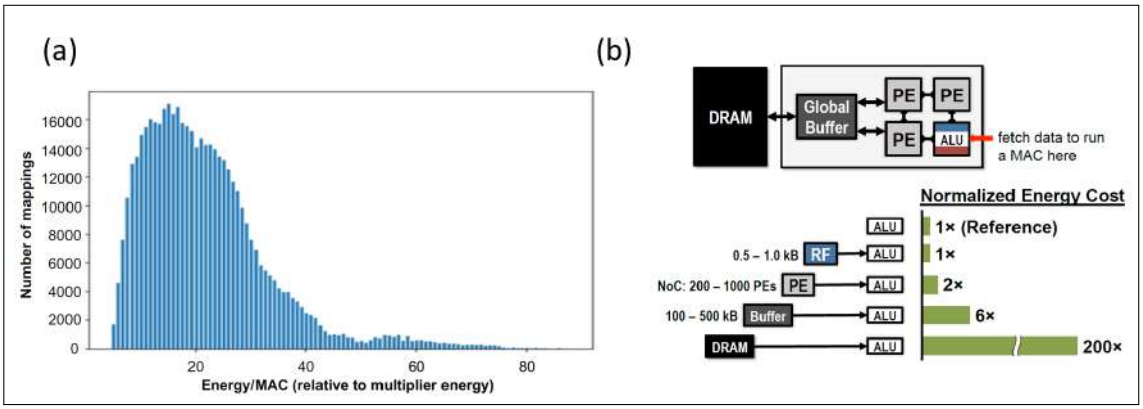


Figure 4.2. (a) Histogram of the energy efficiency of various mappings of VGG convolutional network on a specific architecture, and (b) memory hierarchy and different data movement energy. Figures reproduced from from [28] and [7] respectively.

the different memory hierarchy levels to optimize cost, energy, latency, and capacity, tends to be extremely challenging due to the large design space and the lack of a systematic approach to explore it.

Capacitor-less two-transistor gain cell embedded DRAM is a promising candidate to alleviate this trade-off, due to its potential for higher density, lower power con-

sumption, low latency and higher endurance than eNVRAM. Gain cells are dynamic memory cells made of two logic transistors, where the second transistor is used not only to increase the in-cell storage capacitance but also to amplify the readout charge via the transconductance gain resulting in a non-destructive read. In this section, we describe our Back-End-Of-Line (BEOL) compatible W-doped Indium Oxide (In_2O_3) (IWO) FETs with ultra-low leakage of $\sim$ 1fA/μm to demonstrate a monolithic 3D (M3D) capacitor-less 2T gain cell eDRAM.

Based on the on the M3D IWO-based gain cell we also developed a ternary content addressable memory (TCAM), specially suited for update-frequent associative search applications (e.g., clustering). Existing TCAM designs have a fundamental gap between low-density/high write performance SRAM, and high-density/poor write performance nonvolatile memories. In this work, we demonstrate: i) M3D TCAM designs based on IWO FETs that can simultaneously achieve high density and excellent write performance, thus bridging the performance gap among existing TCAM designs for update-frequent search applications; ii) the necessity of 6T IWO TCAM design to restore the desired independence among the TCAM words, which is destroyed in the 4T TCAM design; iii) excellent write performance with logic-compatible write voltage (< 1.5V), $<$ 20ns write latency, $> 10^{10}$ endurance; iv) up to 14 $\times$ /35$\times$ improvement in speed/energy over GPU in executing the K-Means clustering algorithm.

4.1 Channel Selection for n-Type BEOL Transistor

The selection of the TFT channel will depend on the intended function of the BEOL transistor. For an access transistor of a M3D DRAM array we target a low temperature (< 400°C) *in situ* synthesis of high-mobility n-type FET. Indium oxide (InOx)-based amorphous oxide semiconductor thin film transistor enables a low thermal budget process (< 400°C), high field-effect electron mobility (due to the unoccupied s-orbital of heavy transition metal In^{3+}) and low leakage (due to wide

bandgap), as shown in Fig. 4.3.

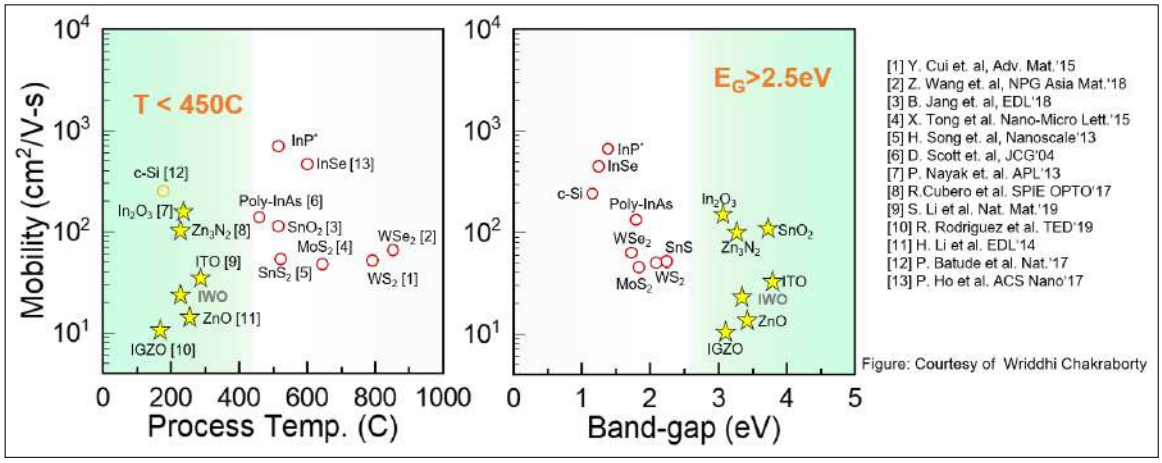


Figure 4.3. Indium, tin and zinc based semiconducting oxides are potential candidates for channel material in BEOL transistors (Field-effect mobility) (Figure:Courtesy of Wriddhi Chakraborty).

For BEOL compatibility, amorphous semiconductors are preferred over crystalline semiconductors from the viewpoints of processing temperature. In covalent semiconductors (for example silicon) the mobility is degraded in three order of magnitude by going from crystalline to amorphous phase. The low mobility in amorphous phase is associated with the intrinsic nature of the strong directivity of the sp^3 orbital bonding. The bond angle fluctuation significantly alters the mobility in covalent semiconductors (Fig. 4.4 (a)). High mobility is possible in amorphous semiconductor oxides due to the conduction band formed by spatially spread isotropic metal s-orbitals denoted by spheres, as shown in Fig. 4.4(b) [29].

To increase device stability we introduce tungsten (W)-doping that increases de-

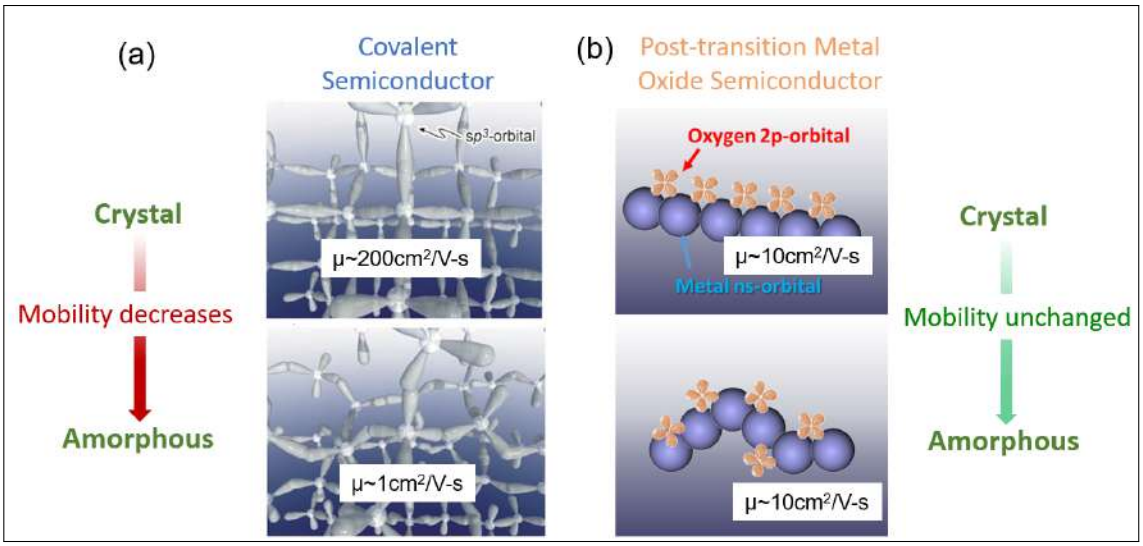


Figure 4.4. (a) In covalent semiconductors (for example silicone) mobility degrades in amorphous phase, and (b) high mobility possible in amorphous oxide semiconductors due to conduction bands formed by spatially spread isotropic metal s-orbitals. Figure reproduced from: [29]

vice stability as high W-O bond dissociation energy (720kJ/mol) suppresses oxygen deficiencies, as shown in Fig. 4.5.

All these characteristics make W-doped amorphous In_2O_3 FET a perfect candidate for n-channel transistors for DRAM application that requires:

1. High Density → BEOL compatible → Low fabrication tempertaure
2. Low access time → High I_{on} → High mobility
3. Long retention → Low I_{OFF} → Wide bandgap

4.2 Device Fabrication and Characterization

The low-thermal budget process flow and schematic structure of the fabricated tungsten (W)-doped amorphous In_2O_3 (IWO) FET with 5nm ALD HfO_2 gate oxide is shown in Fig. 4.6. First, 20 nm thick palladium (Pd) metal gate was deposited via e-beam evaporation and liftoff as back-gate followed by 5 nm-HfO_2 back gate oxide

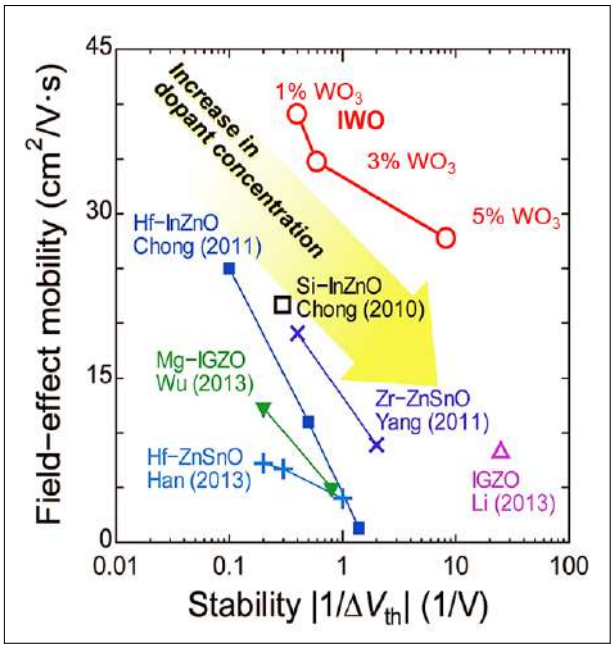


Figure 4.5. W-doping increase device stability as high W–O bond dissociation energy suppresses oxygen deficiencies. Figure reproduced from: [30]

deposition using thermal ALD at < 250°C. Next, 5 nm amorphous W-doped (1 wt. %) In_2O_3 channel was deposited by RF magnetron sputtering in the presence of 0.02 Pa excess O_2 at room temperature. 30 nm thick Ni was deposited as the source and drain electrode followed by 10-min anneal at 150°C under N_2 to improve the contact resistance. Finally, 5 nm thick HfO_2 top gate oxide was deposited using thermal ALD at 120°C, followed by deposition and patterning of Pd as the top gate.

Fig. 4.7 (a) shows the measured transfer characteristics of 50nm and 100nm channel length (L_G) IWO Double Gate (DG) FETs displaying SS_{AVG} of 73 mV/dec and 68 mV/dec respectively, under $V_{DS} = 1$ V. Output characteristics of the DG IWO FET with $L_G = 50$ nm, shown in Fig. 4.7 (b), exhibit a high saturation current of 550 μA/μm at $V_{GS} - V_{TH} = 2$ V and $V_{DS} = 1$ V, thanks to the low contact resistance between IWO film and Ni electrodes.

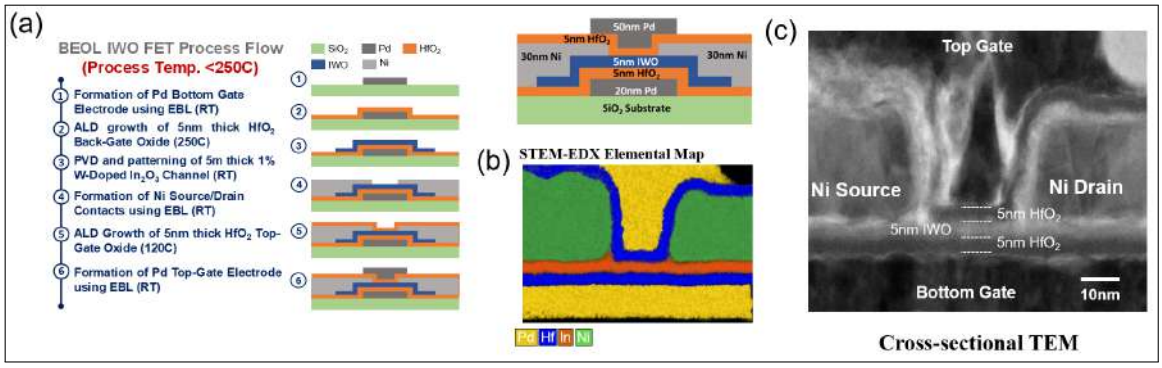


Figure 4.6. (a)Schematic device structure and process flow of Dual-Gate BEOL IWO FET,(b) STEM-EDX Elemental Map of IWO FET and (c) Cross Sectional TEM

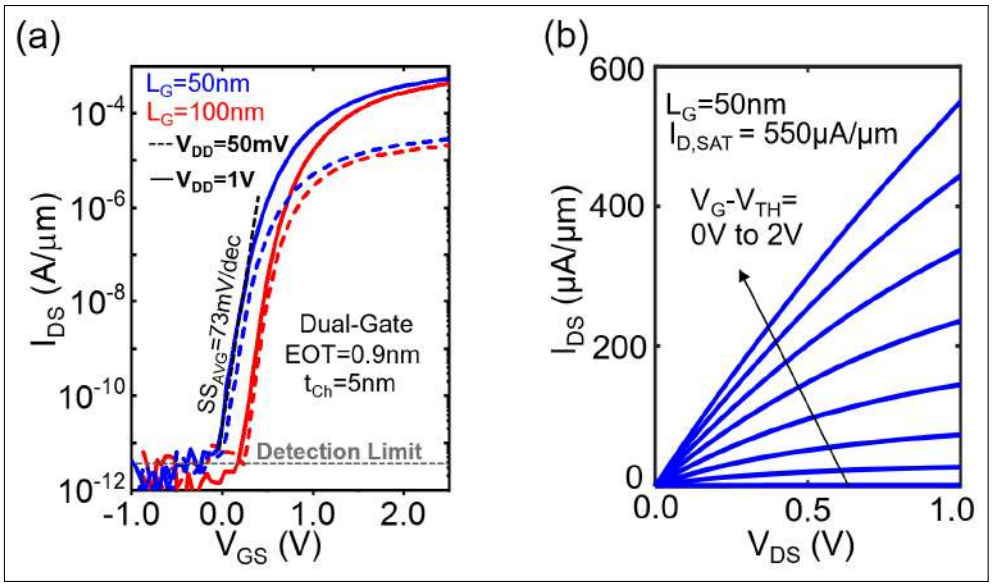


Figure 4.7. (a) Transfer characteristics of Dual-Gate (DG) IWO FET with L_G = 100 nm and 50 nm and (b) Output characteristics of DG IWO FET with L_G = 50 nm.

As the off-state leakage current of the IWO FET is extremely low [56], instrumentation with the current detection limit at approximately $\sim 10^{-13}$ A cannot directly measure the I_{OFF} leakage. Hence, an ultra-wide device (W = 100 μm) was used to directly measure the off-state leakage of $\sim 10^{-15}$ A in DG IWO FET, as shown in Fig. 4.8 (a), resulting in high I_{ON}/I_{OFF} ratio of 10^{12}. I_{OFF} was found to be limited by the gate-to-drain leakage and can be reduced by increasing the EOT. The performance of L_G = 50nm DG IWO FET was benchmarked against other BEOL compatible FETs (4.8 (b)), where I_{ON} is taken at V_{DS} = 1 V over a 1.8 V V_{GS} swing from the reported I_{MIN} point. DG IWO FETs in this work display the highest I_{ON} over a fixed swing, maintaining excellent I_{ON}/I_{OFF} ratio of 10^{12} compared to other BEOL compatible FETs. Hence, IWO DG FET is an excellent candidate for M3D capacitor-less 2T eDRAM application.

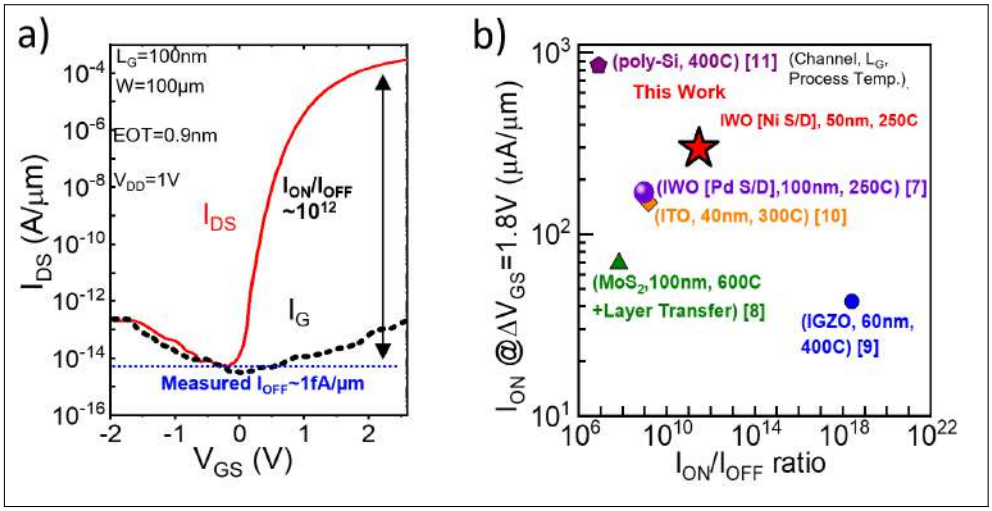


Figure 4.8. (a) Direct measurement of ultra-low ($\sim$ 1 fA/μm) OFF-state leakage in ultra-wide DG IWO FET showing I_{OFF} limited by gate-current (I_G) and (b) Benchmarking shows advantage of Dual Gate IWO (Ni S/D) FET with highest I_{ON} among oxide-semiconductor FETs.

4.3 Characterization of IWO Capacitor-less 2T eDRAM Cell

Fig. 4.9 (a) shows an optical image of the fabricated capacitor-less IWO 2T DRAM gain cell with four signal lines: read-bitline (RBL), write-bitline (WBL), write-wordline (WWL) and read-wordline (RWL), along with the corresponding circuit schematic shown in Fig. 4.9 (b). The state of the memory cell is given by the stored charge in the gate capacitance of the read transistor. Fig. 4.9 (c) illustrates the cell bias conditions used during the write, read, and hold modes. Write time and retention time are improved by holding WWL at V_{BOOST}, above V_{DD}, and V_{HOLD}, below V_{SS}, during write and hold phases respectively. 4.9 (d) shows how the cell is integrated into an eDRAM array.

We characterize the performance of the IWO eDRAM cell, particularly during the hold operation in which the storage node is discharged over time. A characteristic retention time (τ_r) is extracted when the node has discharged 80% of the total charge. We track the voltage of the storage node by continuously measuring the drain current of the read transistor. The storage node voltage ($V_{STORAGE}$) can be obtained from the $I_d - V_g$ of the read transistor. Fig. 4.10 (a.1) shows the measured discharge dynamics of the eDRAM cell for different hold voltages. By plotting the characteristic retention time for each hold voltage (4.10 (a.2)), an optimal hold voltage (V_{HOLD}) is extracted at which max retention is achieved before it is reduced by gate leakage. At the optimal V_{HOLD}, we study the impact of varying temperature on τ_r for three different temperatures: 25°C, 50°C and 85°C. The discharge dynamics are shown in Fig. 4.10 (b.1). Fig. 4.10 (b.2) shows the dependence of τ_r on temperature. Fig. 4.10 (c.1) shows the optical images of various eDRAM cells with different node capacitances that range from 960 fF to 7 fF. While the fabricated capacitor-less IWO eDRAM cell has the lowest node capacitance of 7 fF, we project it to be $\sim$ 1 fF at a further scaled geometry. Figs. 4.10 (c.2 and c.3) show the discharge dynamics for different in-cell node capacitances and the retention time as a function of node capacitance,

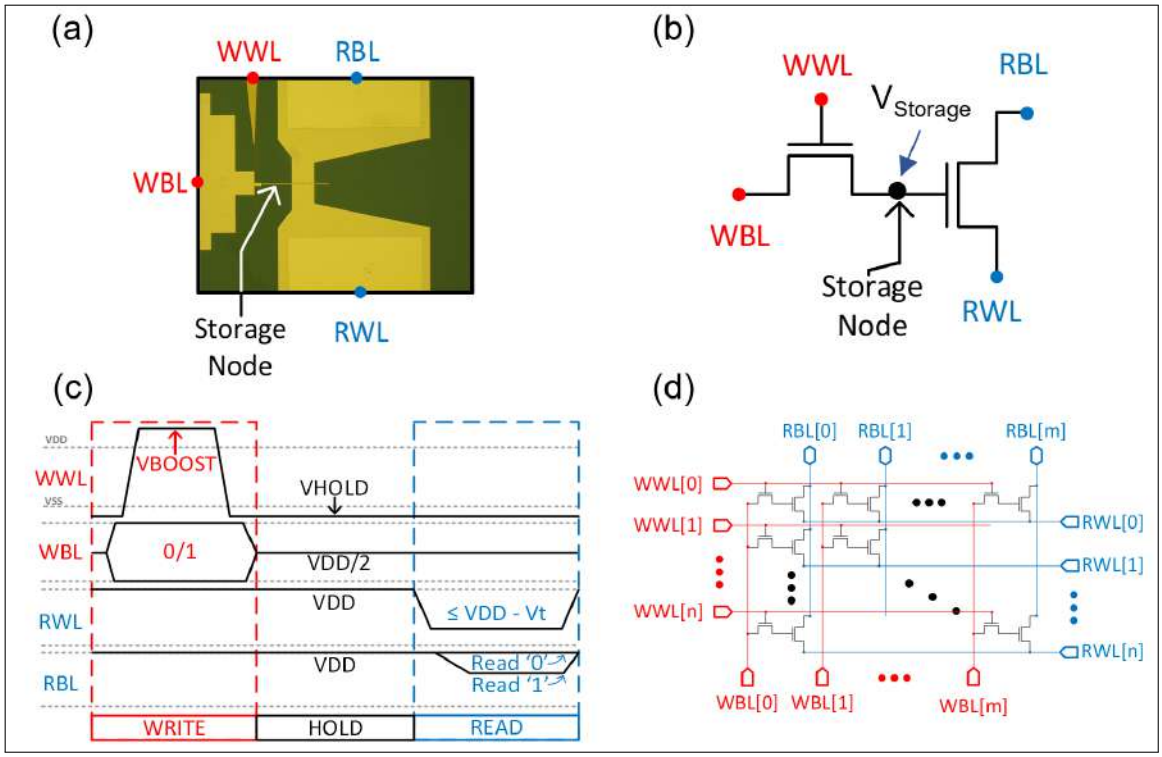


Figure 4.9. (a) Optical image and (b) corresponding circuit schematic of capacitor-less 2T-eDRAM, (c) timing diagram showing voltage waveforms for Write, Hold, and Read operations, including V_{BOOST} (above V_{DD}) and V_{HOLD}(below V_{SS}) and (d) schematic of the eDRAM array with individual cells.

respectively. Extrapolating these results, retention time for 1fF node capacitance at 25°C was found to be $\sim$ 10 s. Projected retention time at 50°C and 85°C are 2 s and 0.3 s , respectively.

4.4 Virtual Source Model for IWO FET

To perform 2T eDRAM Array simulations we extend the semi-empirical physics-based Virtual Source (VS) model for IWO transistors [57]. We incorporate gate-voltage (V_{GS}) modulated Schottky diodes in series with an intrinsic FET to capture

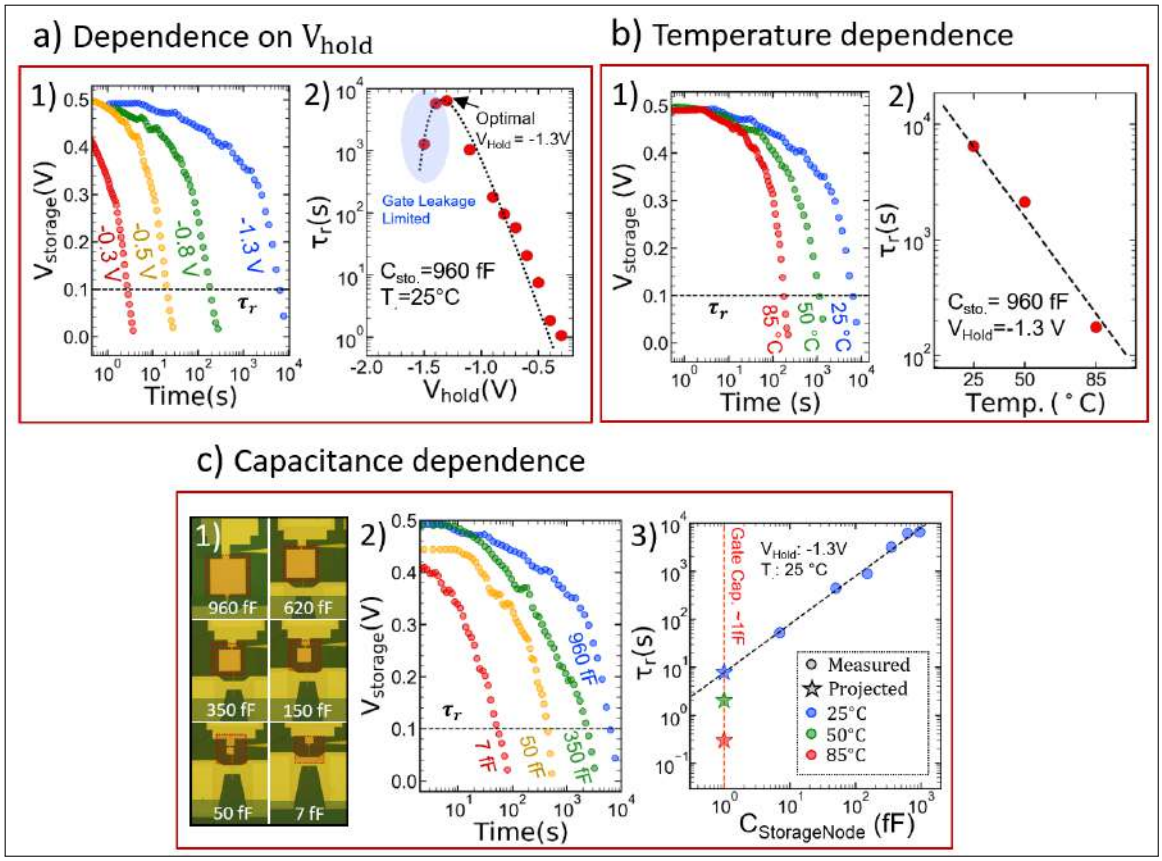


Figure 4.10. (a) 1) Discharge dynamics of storage node voltage ($V_{STORAGE}$) at different hold voltages (V_{HOLD}) and 2) dependence of retention time (τ_r) on different V_{HOLD}; (b) 1) Node voltage discharge characteristics at different temperatures and 2) dependence of retention time on operating temperature; and (c) 1) Array of fabricated node capacitance, 2) Node voltage discharge dynamics for different storage-node capacitances ($C_{STO.}$), and 3) dependence of retention time on C_{STO}. Projected retention times for $C_{STO} = 1$ fF at different operating temperatures show 300 ms retention time at 85°C.

the V_{GS} dependent source-drain contact resistance in IWO FET [58] (Fig. 4.11 (a)). According to the VS model, the intrinsic FET drain-to-source current (I_{DS}) is computed as the product of the mobile charge density (Q_{inv}) and injection velocity (v_{x0}) as shown in equations 4.1 and 4.2.

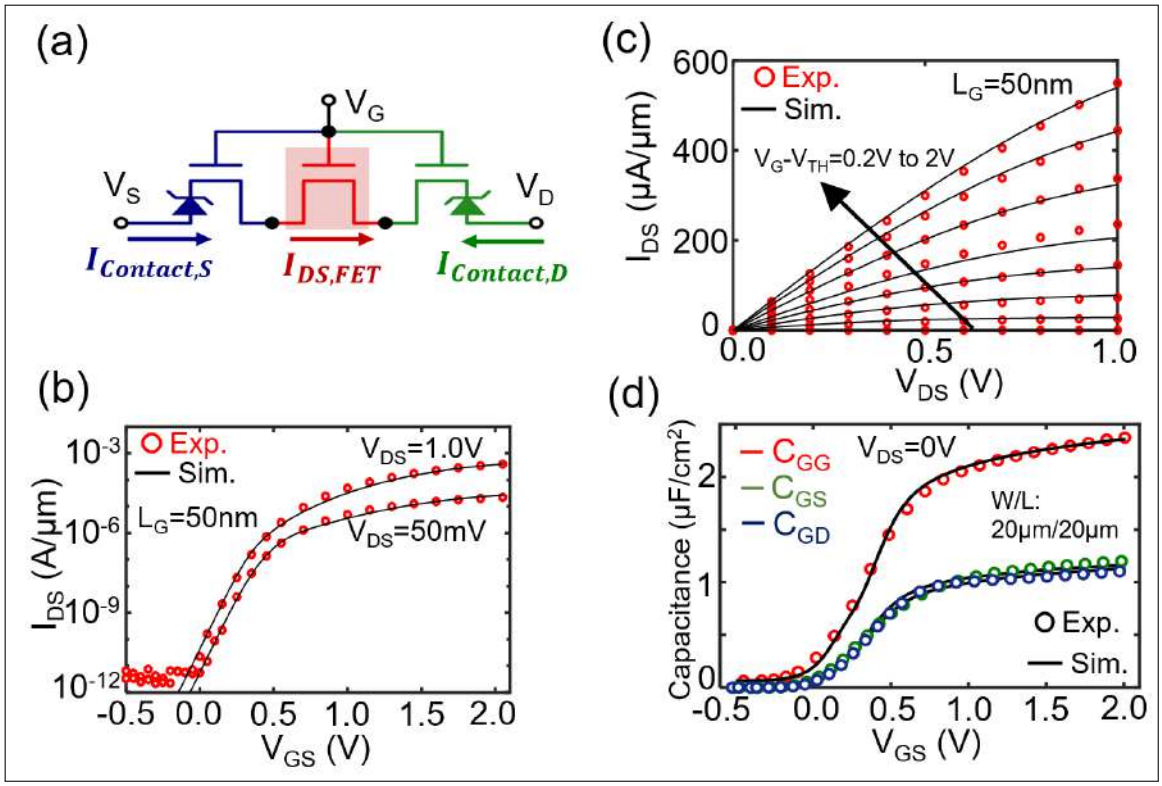


Figure 4.11. (a) IWO FET VS model. (b)/(c) Transfer/Output Characteristics of DG IWO FET and (d) C_{GG}, C_{GS} and C_{GD} vs V_{GS} at $V_{DS} = 0V$

$$I_{DS} = W \times Q_{inv} \times v_{x0} \tag{4.1}$$

$$v_{x0} = \left((\mu_0 V_{DS}/L_G)^{-1} + (v_{inj})^{-1}\right)^{-1} \tag{4.2}$$

where W is the geometric device width. The current through source-drain Schottky junction (I_{CS}, I_{CD} respectively) is empirically modeled as shown in equations 4.3 and 4.4.

TABLE 4.1

ELECTRICAL CHARACTERIZATION

Parameter	Value	Units
C_{ox}	2.1	(μF/cm^2)
I_{CS0}, I_{CD0}	1.2×10^{-4}	(A/μm)
η	10	-
L_G	50	(nm)
V_{inj}	1.5×10^6	(cm/s)

$$I_{Contact,S} = I_{CS0} \times \left(exp\left(\frac{-qV_{GS}}{\eta kT}\right) - 1 \right) \tag{4.3}$$

$$I_{Contact,D} = I_{CD0} \times \left(exp\left(\frac{-q|V_{GS}|}{\eta kT}\right) - 1 \right) \tag{4.4}$$

The model parameters are listed in table 4.1. Current continuity equation for the gated Schottky source-drain contacts and the intrinsic IWO FET are solved iteratively to match the experimental transfer (Fig. 4.11 (b)) and output characteristics (Fig. 4.11 (c)).

The voltage-dependent capacitances are estimated from the derivative of the terminal charges with respect to the terminal voltages, following channel-charge partitioning as shown in ref. [57]. Fig. 4.11 (d) shows that the VS charge model captures the V_{GS} dependence of gate-to-channel capacitance (C_{GG}), gate-to-source capacitance (C_{GS}) and gate-to-drain capacitance (C_{GD}) at $V_{DS} = 0$V.

4.5 IWO 2T eDRAM Array

We use the VS IWO FET model to simulated a 128×128 eDRAM array. Interconnect resistance, peripheral (BL/WL driver) output resistance and memory cell transistor channel resistance are treated separately. The total time-delay is the sum of the delay for the segments covering the voltage slew range and the limiting interconnect delay time in the case of read and write. The write and retention timing results across a range of V_{BOOST} and V_{HOLD}, as shown in Fig. 4.12 (a) and (b), show that ≤ 3 ns write time with ≥ 1 s retention are feasible within compact cell dimensions and moderate outside-the-rails voltages. Fig. 4.12 (c) shows that these timing results describe a class of memory that is much faster ($\sim 10\times$) than emerging non-volatile memories (eNVM) while requiring significantly less ($\sim 100\times$) standby power than conventional SRAM and eDRAM ([59], [60] and [61]).

Fig. 4.13 (a) shows an array density scaling path through 3D stacking of BEOL memory layers, where density eventually saturates with increasing number of stacking layers due to the peripherals area which can only be placed on the FEOL. At each layer, only a fraction of it can be used for bitcell, and the rest is used for interconnect, then, layer efficiency is defined as the ratio between the area used by the bitcells and the total area of the layer, as shown in equation (4.5). Fig. 4.13 (b) shows a 3D plot of how the layer efficiency and number of layer affects the memory density. A higher layer efficiency allows to reach the maximum density with a smaller number of layers. The memory density is limited at ~ 150 Mb/mm^2 by the peripheral area. For these plots, we conceive an equivalent peripheral area at 60% efficiency to be present in the scaled CMOS logic layer.

$$Layer\ Efficiency = \frac{A_{bitcells}}{A_{interconnect} + A_{bitcells}} \tag{4.5}$$

Considering all the previous results we benchmark our devices with other eDRAM

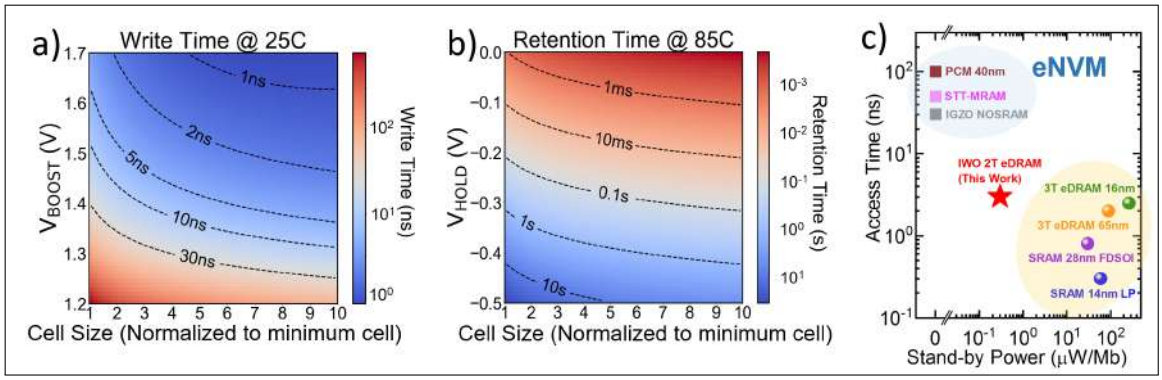


Figure 4.12. (a) Write time and (b) standby retention across access transistor width scaling and V_{BOOST} and V_{HOLD} respectively. ≥ 1 s retention and ≤ 3 ns write time are achievable with minimum access device width and moderate outside-the-rails voltages (~ 2 V). (c) Benchmarking shows that our memory is much faster ($\sim 10\times$) than emerging nonvolatile memories (eNVM) while requiring significantly less ($\sim 100\times$) standby power than conventional SRAM and eDRAM.

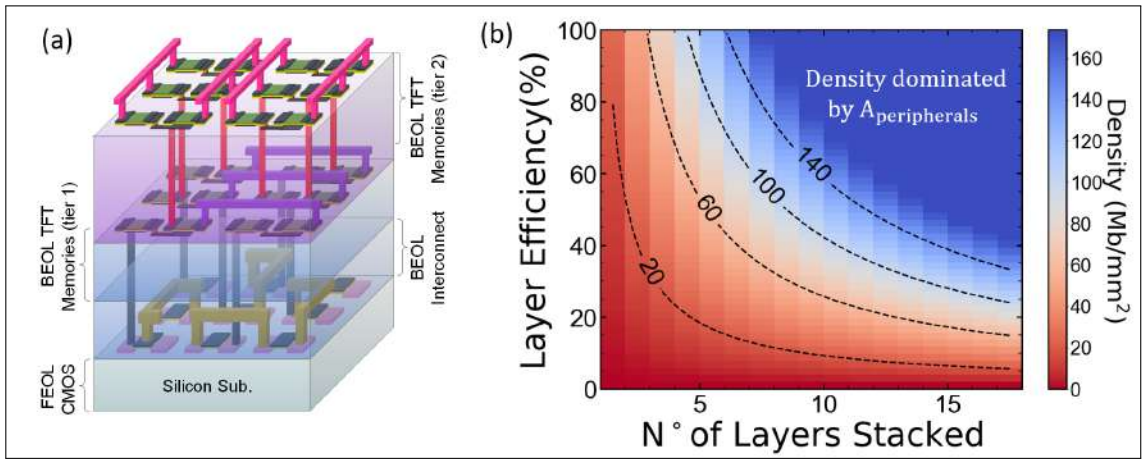


Figure 4.13. (a) Schematic of 3D stacking of BEOL memory layers and (b) memory density dependence on layer efficiency and number of layers.

TABLE 4.2

BENCHMARK OF MONOLITHIC-3D CAPACITORLESS 2T EDRAM BASED ON W-DOPED In_2O_3 FETS

eDRAM Type	2T [62]	1T-1C [62]	1T-1C [63]	2T [64]	This Work
V_{DD}(V)	1.2	1.2	1.05	1.2	1.0
Density (Mb/mm^2)	80	80	17.5	4	150
BEOL Compatible	Yes	Yes	No	No	Yes
Cell Cap. (fF/cell)	1.2	3.5	13.8	-	1
Retention (ms@85°C)	10^7	10^8	0.3	1	3
Access Time (ns)	30	30	5	1.6	3
Destructive Read	No	Yes	Yes	No	No

technologies, as shown in Table 4.2. Our devices present a good balance of memory density, access time and retention.

To visualize the performance improvement provided by our higher density M3D 2T-eDRAM, we estimate the impact in area and on-chip access frequency of deploying three different networks on chip: ResNet-110 [65], with a storage size of 54.4 Mb and Transformer in two flavors: base and big, with storage sizes of 2.1 Gb and 6.8 Gb respectively [66]. We use as baseline the Intel's 1T-1C 22 nm 2D eDRAM [67]. First, we estimate the area size that we would need to deploy the entire network on chip (Fig. 4.14 (a)). We can see that for transformer base, 8 layers IWO 2TeDRAM needs 7.3 times less area than baseline. Secondly, we consider On-chip frequency access at iso-area. Off-chip memory accesses have an energy cost of at least one order of magnitude higher than On-chip accesses, so, increasing the On-chip accesses, implies a direct energy improvement. Fig. 4.14 (b) shows that at a fix area of 10 mm^2, the on-chip access frequency of the 8 layers IWO 2TeDRAM is 6 times higher than the baseline.

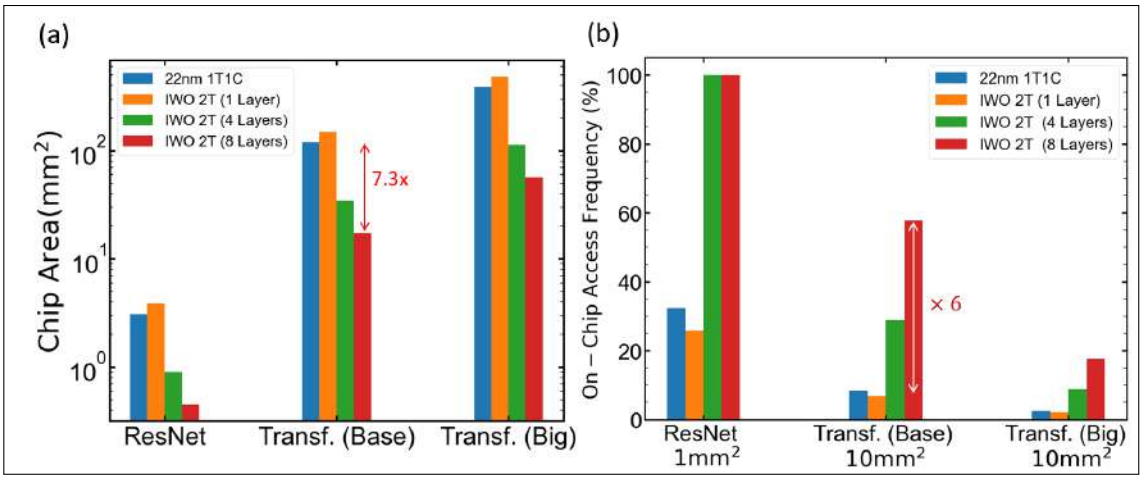


Figure 4.14. Impact of increasing on chip memory capacity:(a) chip area needed to deploy entire network on-chip, and (b) On-chip access frequency at fix area

4.6 Computational Associative Memory Based on 2T eDRAM

Computational associative memory, notably TCAM, is gaining popularity in accelerating pattern matching tasks because it can search across the whole memory in parallel for matched entries directly in-memory. In addition, its capability of computing the hamming distance (number of positions in which two symbols are different) enables its application as a distance kernel to accelerate various kinds of machine learning applications [68]. Among many of them, the TCAM works in the static inference mode, where it is mostly used as a distance kernel with only once or occasionally update. This opens up a large design space of TCAM by exploiting the dense nonvolatile memories, e.g., resistive random-access memory (ReRAM) and ferroelectric FET (FeFET), while avoiding their drawbacks of poor write performance (e.g., limited write endurance, high write voltage and latency). However, there are many other important applications, such as unsupervised clustering (Fig. 4.15(a) and (b)), where frequent update to the stored entries in the distance kernel is a must while the

long term retention can be relaxed [69]. As a rule of thumb, we estimated that about 0.13 update is required per search operation with the classical K-means clustering algorithm. Given a large dataset and many iterations required, the required updates could exceed the endurance limits of ReRAM or FeFET. SRAM based TCAM, with almost infinite endurance and excellent write performance, provides an alternative solution. However, each SRAM TCAM consumes 16 transistors, significantly limiting the TCAM capacity to accommodate big data.

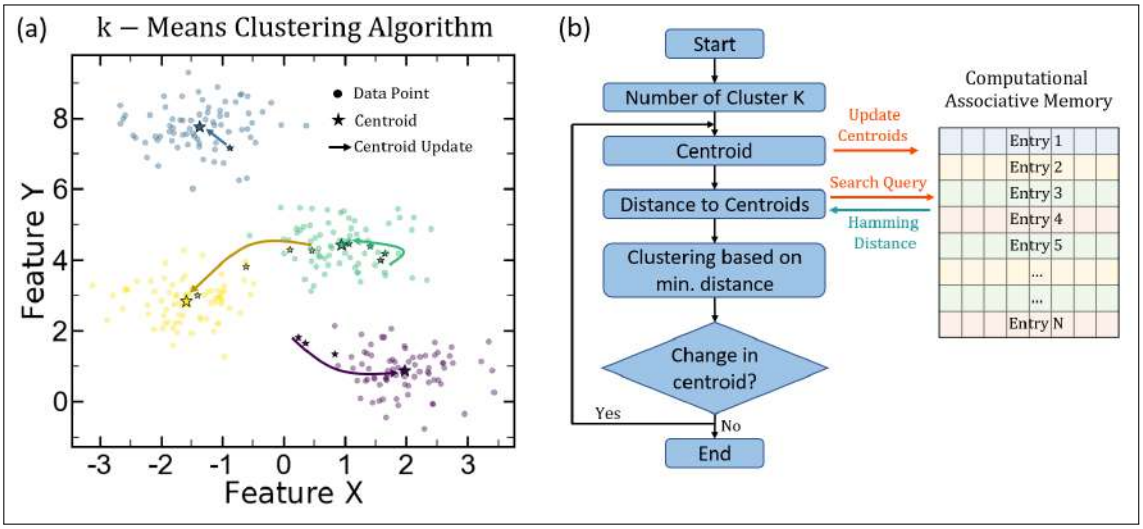
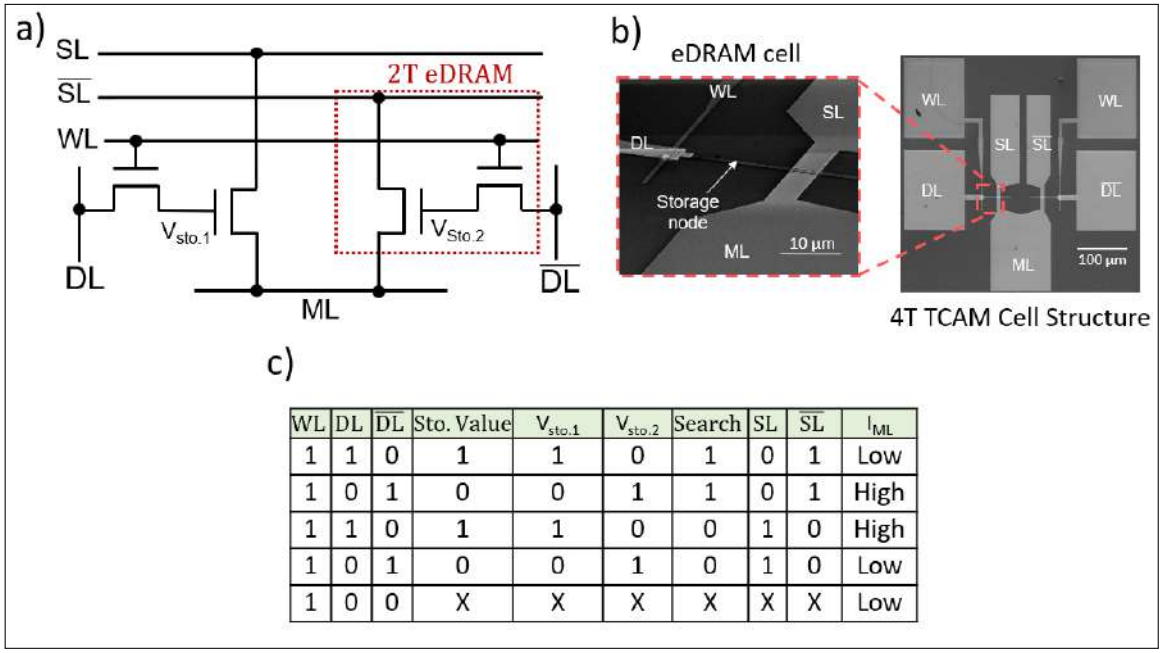


Figure 4.15. (a)Example of K-Means clustering algorithm with continuous update of the centroids and (b) K-Means clustering algorithm associative memory usage.

Our M3D TCAM can simultaneously achieve high density and excellent write performance, thus bridging the performance gap among existing TCAM designs for update-frequent search applications. M3D IWO-based TCAM has great write perfor-

mance and high memory density. For update-frequent search applications, retention is required only up to the next update. For the proposed M3D TCAM design, the 4T cell is first introduced, its critical issue of destroying the independence among TCAM words is identified, and then a 6T cell is proposed to address this issue.



WL	DL	$\overline{DL}$	Sto. Value	$V_{sto.1}$	$V_{sto.2}$	Search	SL	$\overline{SL}$	I_{ML}
1	1	0	1	1	0	1	0	1	Low
1	0	1	0	0	1	1	0	1	High
1	1	0	1	1	0	0	1	0	High
1	0	1	0	0	1	0	1	0	Low
1	0	0	X	X	X	X	X	X	Low

Figure 4.16. (a) M3D IWO-based TCAM topology, (b) SEM of the 4T TCAM cell and zoom-in of the eDRAM on one branch, and (c) 4T TCAM operation modes.

In the 4T TCAM design, the TCAM cells on the drain connected Search Line (SL), draw current from the SL along its propagation. Due to the presence of the interconnect metal wire resistance (R_i) and the SL driver launch resistance (RL) (Fig. 4.17 (a) and (b)) [70], [71], the SL voltage (V_{SL}) in the last TCAM word will fluctuate

between the best scenario, where all the cells on the SL match (Fig. 4.17 (c)), and the worst scenario, where all the cells mismatch (Fig. 4.17 (d)).

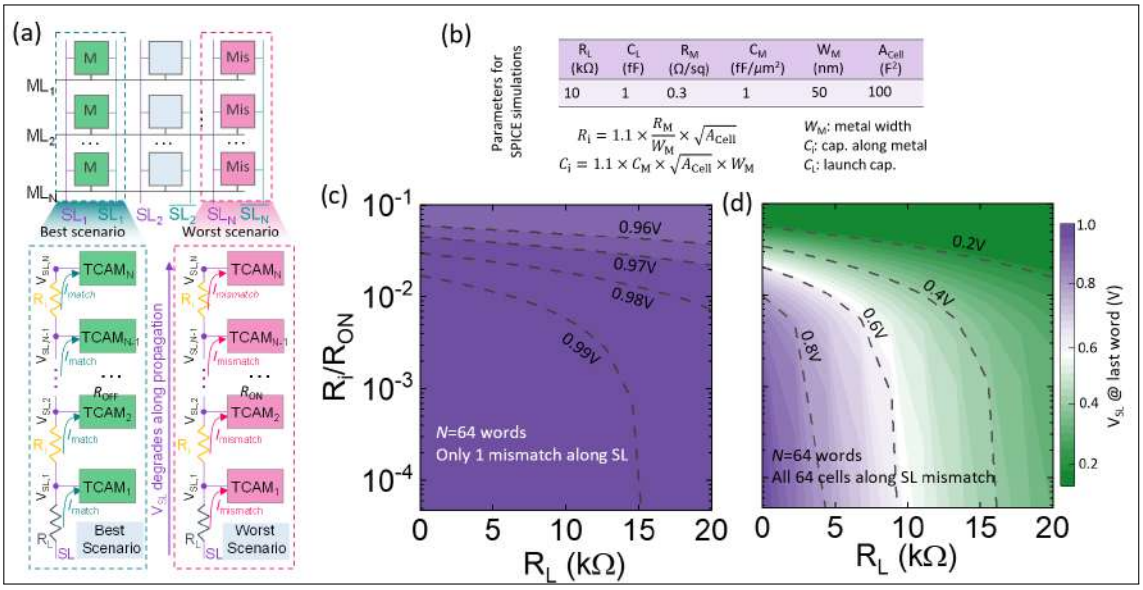


Figure 4.17. (a) By connecting the SL on the read FET drain we destroy the independence among TCAM words as the V_{SL} depends on the conditions of other cells on the SL. (b) Parameters used for the SPICE simulations. (c)/(d) simulated V_{SL} on the last word shows strong differences between best/worst scenarios.

6T TCAM (Fig. 4.18 (a) and (b)) enables the independence among TCAM words eliminating V_{SL} drop across the SL. In the 4T cell, the V_{SL} fluctuation on each column will cause hamming distance sensing almost impossible because each TCAM cell on the ML has a different current depending on the number of mismatches on the same word and in other words (Fig. 4.18 (c)). 6T TCAM allows to reduce this problem by applying the SL to the gate of the extra transistor at each branch of the

TCAM, allowing to sense the hamming distance (Fig. 4.18 (d)).

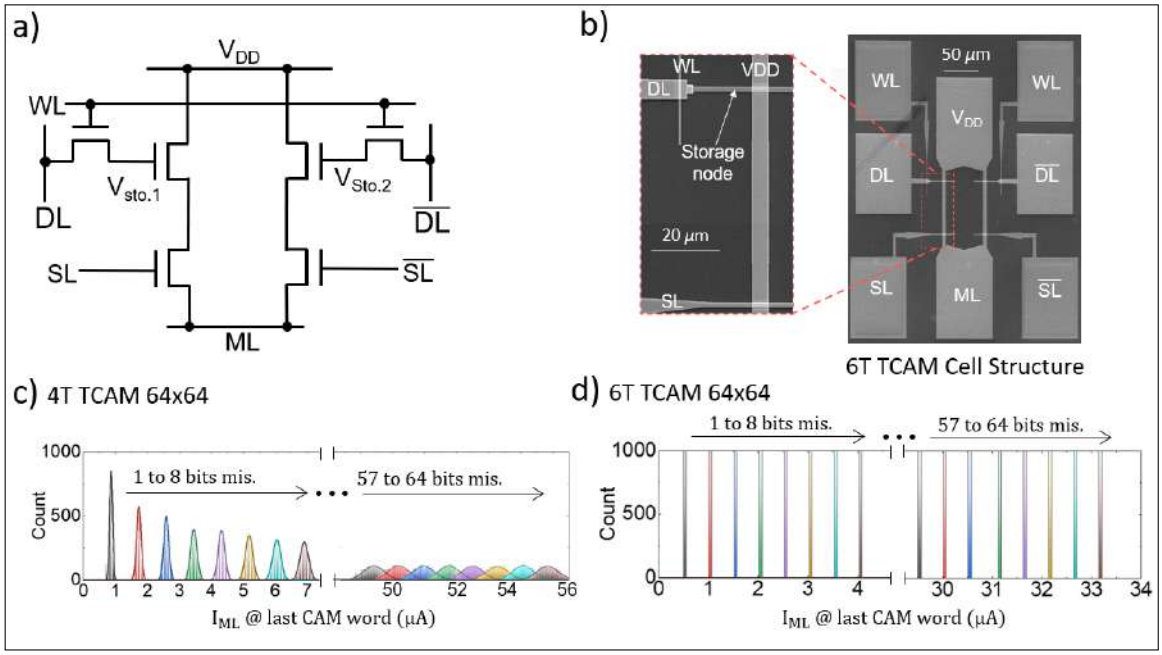


Figure 4.18. (a) M3D IWO-based 6T TCAM topology, (b) SEM of the 6T TCAM cell and zoom-in on one branch, (c) I_{ML} on the last TCAM word depends on the $N_{mismatch}$ cells along the SL. This creates V_{SL} fluctuation among cells of a word, resulting in variation in I_{ML} and (d) connecting the SL to the FET gate (6T configuration) restores the independence among words by delivering V_{SL} with no degradation and successful detection of hamming distance is achievable with independence on the number of transistors.

To evaluate the write performance of the TCAM, the write speed (Fig. 4.19 (a)) is experimentally characterized on one branch of the TCAM (Fig. 4.19 (d)). It shows that with logic-compatible 1.5V, 20ns (instrument limitation), a 5μm read current can be obtained, demonstrating great write performance. The retention of the intrinsic

eDRAM cell, i.e., no external storage capacitor, is measured as a function of hold voltage, V_{hold} and temperature (Fig. 4.19 (b)). Longer retention up to 1000s can be obtained by reducing V_{hold} to cut off the leakage current of the access transistor. Endurance of over 10^{10} is demonstrated (Fig. 4.19 (c)).

Based on the SPICE/Verilog-A simulation framework we can predict the performance of a 64x64 TCAM array (Fig. 4.20 (a)). The write performance is also predicted for scaled write transistor width (W_{TW}) (experimentally 5μm) and load capacitance on a 64x64 TCAM array. It shows around 10ns write latency and excellent write energy of 1-4 fJ/word/bit (Fig. 4.20 (b) (I) and (II)). The search performance is evaluated as a function of the R_L and the metal wire parasitic capacitance (C_i). The delay and energy (Fig. 4.20 (c) (I) and (II)) increase with the capacitance. Fig. 4.20 (d) presents a performance comparison of TCAM arrays using different technologies [72]. It is apparent that IWO based TCAM shows a logic-compatible write voltage (1.5V), high write speed (20ns), high endurance (10^{10} cycles), good search performance, making it a leading candidate for update-frequent search applications.

Finally we compare IWO TCAM in accelerating the K-Means algorithm to other technologies using eight different datasets summarized in table 4.3. The original real value data-points for clustering are first converted to binary high dimensional vectors, which allow to harness the Hamming distance calculation capability of TCAM. To implement K-Means algorithm, the clusters centers are stored in TCAM and then datapoints are searched through the cluster centers stored in the TCAM array. The minimum distance is identified, based on which of the cluster centers need to be updated in the TCAM array. System level benchmarking shows that IWO TCAM design provides on average the second highest speedup over GPU (14x) (Fig. 4.21 (a)), only second to SRAM TCAM and on average highest energy saving (35x) considering the overall search and update applications (Fig. 4.21 (b)).

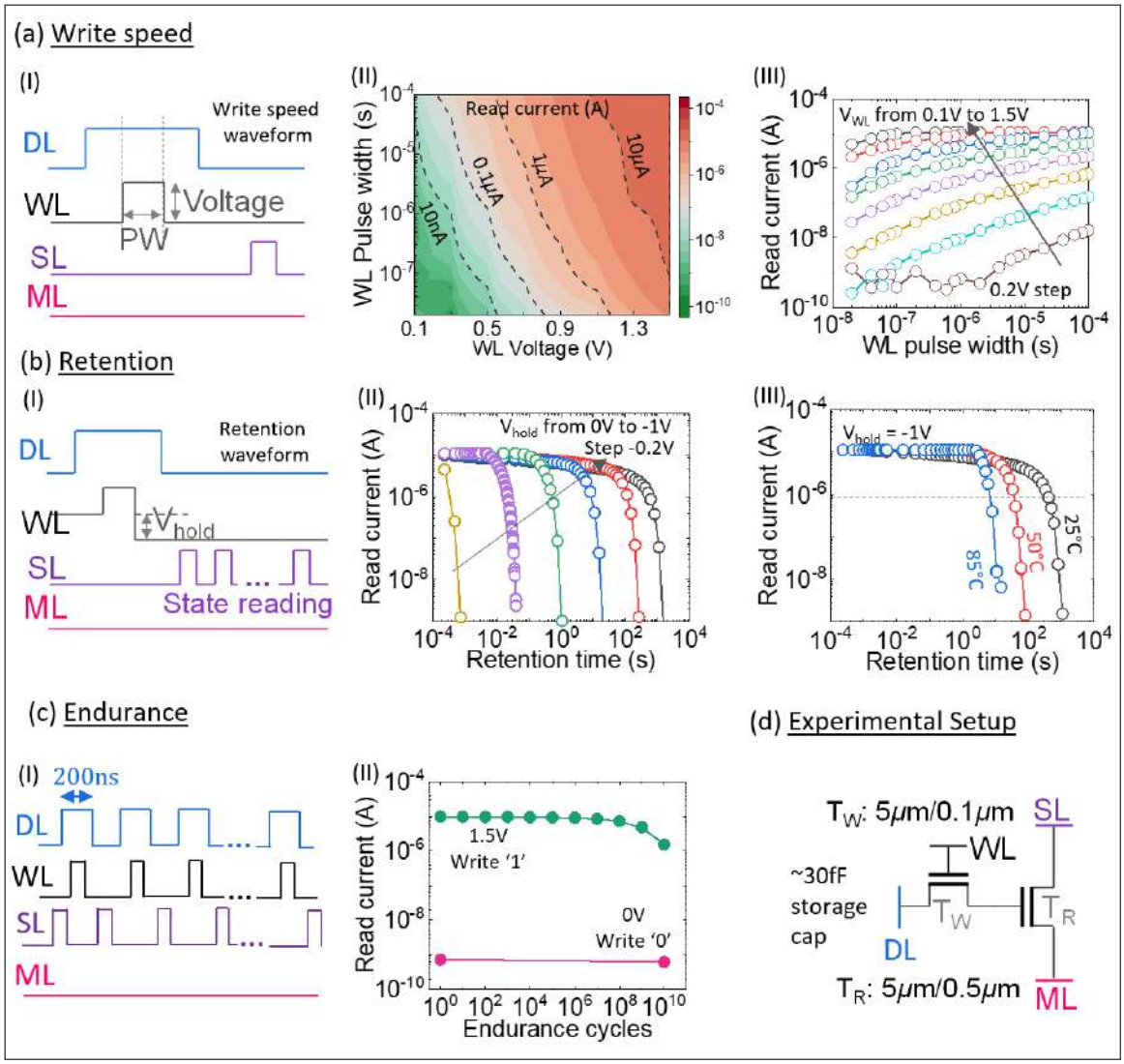
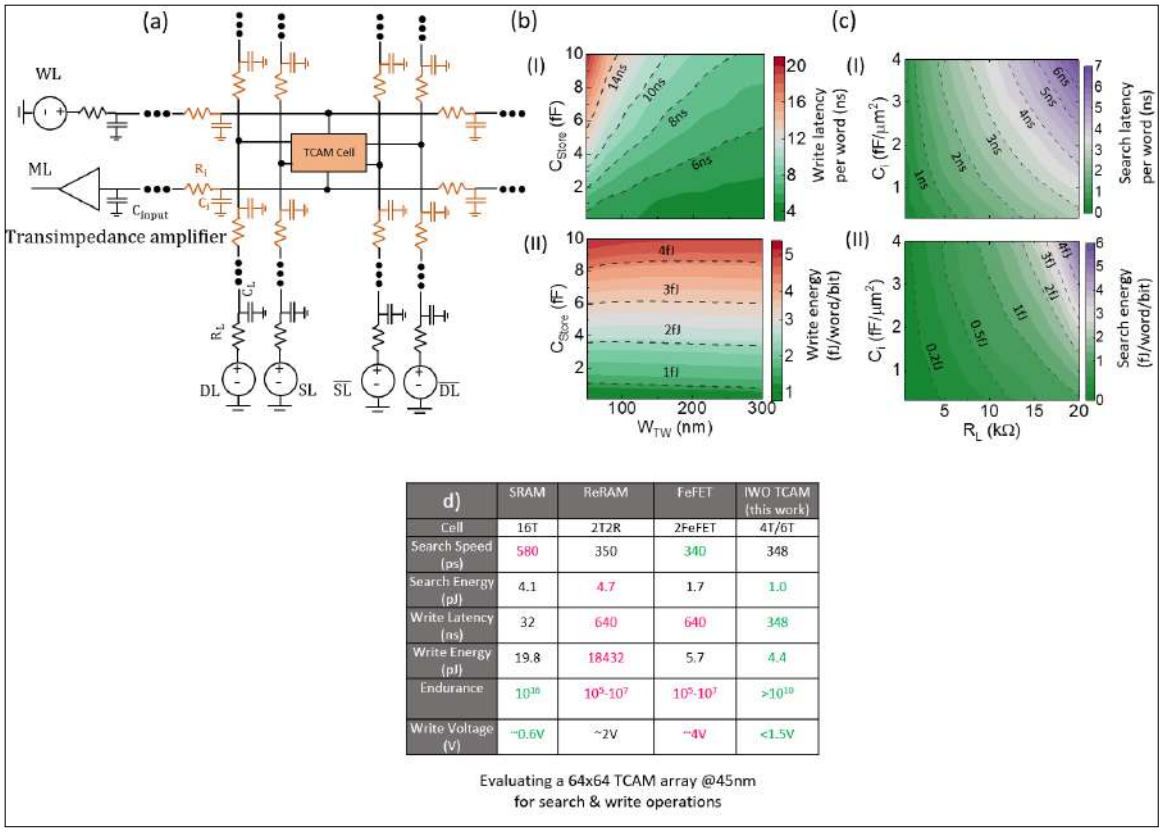


Figure 4.19. (a) Measured write speed:(I) Applied waveform, (II) and (III) Write delay of 20ns is demonstrated with < 1.5V .(b) Retention: (I) Applied waveform, (II) Retention up to 1000s is demonstrated with V_{hold} of -1V relying solely on the intrinsic storage capacitance, and (III) Retention degrades at high temperatures. (c) Endurance: (I) Applied waveform and (II) Negligible degradation up to 10^{10} cycles. (d) Experimental setup.



d)	SRAM	ReRAM	FeFET	IWO TCAM (this work)
Cell	16T	2T2R	2FeFET	4T/6T
Search Speed (ps)	580	350	340	348
Search Energy (pJ)	4.1	4.7	1.7	1.0
Write Latency (ns)	32	640	640	348
Write Energy (pJ)	19.8	18432	5.7	4.4
Endurance	10^{16}	10^{5}-10^{7}	10^{5}-10^{7}	$>10^{10}$
Write Voltage (V)	~0.6V	~2V	~4V	<1.5V

Figure 4.20. (a) Array Topology used in simulations. (b) Predicted performance of write latency (I) and energy (II) with the calibrated virtual source model. (c) Predicted performance of search latency (I) and energy (II). (d) Benchmarking of TCAM arrays shows that IWO TCAM has excellent write and search performance.

4.7 Conclusions

In this chapter we presented our fabricated BEOL compatible W-doped amorphous In_2O_3 (IWO) FETs with record I_{ON} (550 μA/μm) and ultra-low leakage (1 fA/μm). Using IWO FETs we experimentally demonstrated capacitor-less 2T-DRAM cell with high retention time and low refresh power. Experimental characteristics of the IWO transistor are used to calibrate the virtual source transistor model, which will be used for the eDRAM design exploration and performance evaluation.

TABLE 4.3

DATASETS

Dataset	# Data point	# Features	# Clusters	Description	Reference
MNIST	60000	784	10	Handwritten Digits	[73]
FACIAL	27965	300	2	Grammatical Facial Expressions	[74]
UCIHAR	7667	561	12	Human Activity Using Smartphones	[75]
SEIZURE	11500	178	5	Epileptic Seizure	[76]
SENSOR	139100	129	6	Gas Sensor Array Drift	[77]
GESTURE	9880	50	5	Gesture Phase Se	[78]
ISOLET	7797	617	26	Speech data	[79]

IWO 2T-eDRAM bridges the gap between eNVRAM and traditional SRAM and eDRAM by demonstrating a lower access time than eNVRAM and lower standby power than traditional SRAM and eDRAM. We also estimated the density of the IWO 2T-eDRAM memory and showed that a higher density allows to reduce the chip area, or to reduce the Off-chip access frequency, which has a direct impact on the energy/latency performance of the system. Finally, based on the eDRAM gain cell, 3D IWO TCAMs are proposed for update-frequent associative search applications. Together with its high density through 3D stacking, the IWO-based TCAM can fill the gap of existing TCAM designs for update-frequent search applications.

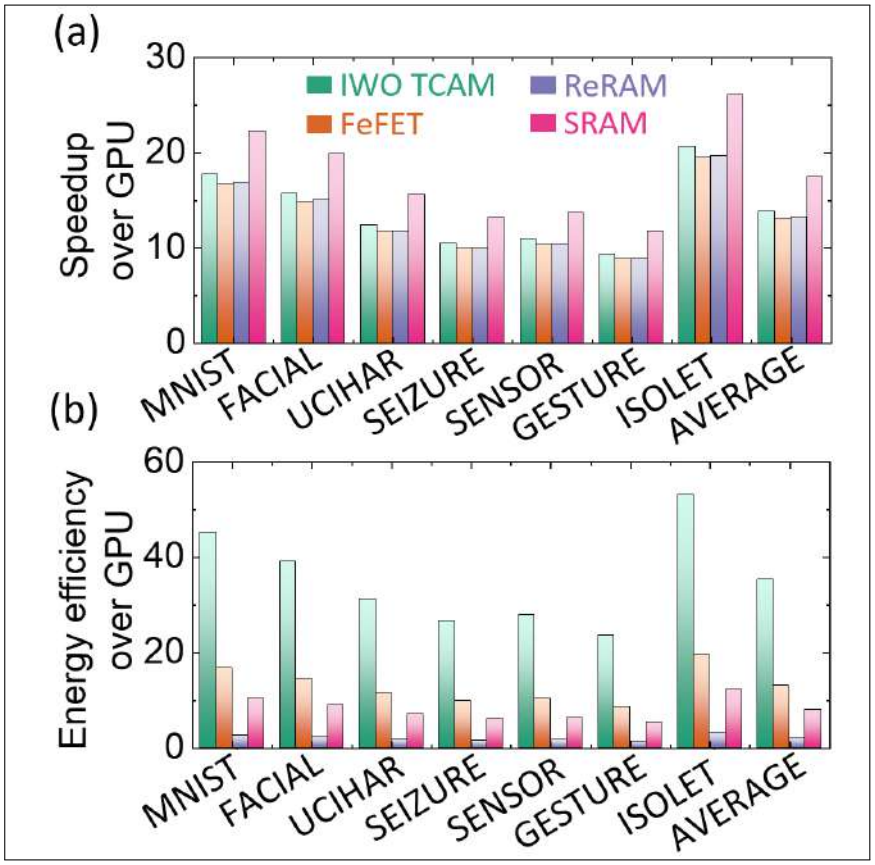


Figure 4.21. (a) Overall latency show 14x improvement on average of IWO TCAM over GPU and (b) Overall energy show 35x improvement on average of IWO TCAM over GPU

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

At the verge of the end of the second era of transistor scaling, there is a global search for devices capable of bringing in new functionality and to further lower the power dissipation in computer. When the baseline is 3 nm CMOS technology, this is not an easy game. In fact, we have observed a continuous reduction of players over the last years in the node scaling race, as shown in Fig. 5.1. Although it is a difficult race, it is worth running. Every technological component benefits directly or indirectly from transistor scaling, hence, it impacts every aspect of our daily life. If this is a difficult race for industry, for Academia it is even worse. Smaller teams, reduced budget and less experienced students, can make it seem even as an impossible task, however, ideas are not a monopoly of industry, and precisely during these transition times, new ideas are most needed, and academia may play major role on developing them.

Oxide-based devices can play a major role on enabling new functionality and improving current devices. Two of them -NCFET and IWO FET- have been the subject of discussion in this thesis.

5.1 Contribution of this Research

This work attempts to position itself as a timely effort at exploring the phenomenon of Negative Capacitance and its implications on transistors. It also explores Monolithic 3D integration using BEOL compatible IWO transistors. The hypothesis of this work was that functional oxide electronics can enable the third era of scaling:

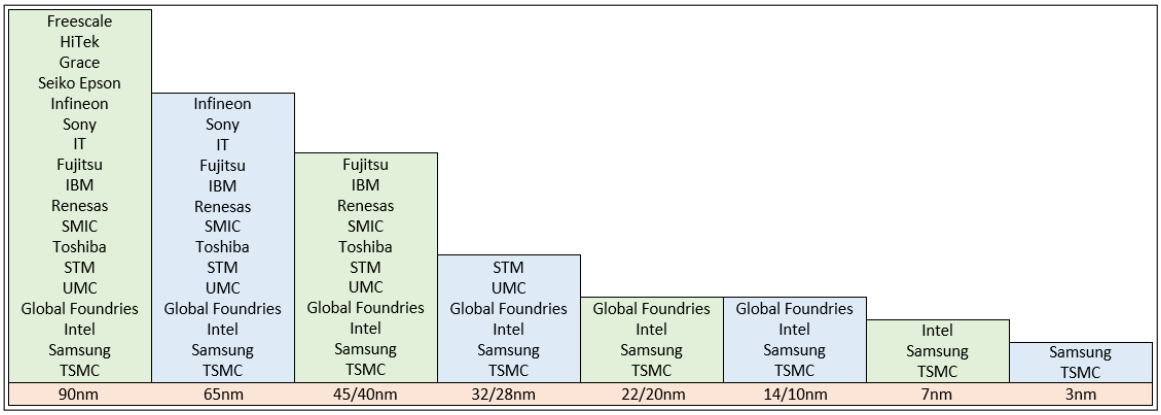


Figure 5.1. Players of major process nodes from 90nm to 3nm. Figure adapted from [31]

hyper-scaling, specifically by enabling two areas: beyond-Boltzmann transistors and monolithic 3D integration of memory. Analyzing the results of this work we think that, although there are still technical challenges to overcome, oxide-based devices are promising candidates to provide a path towards the hyper-scaling era due to the following specific contributions of this thesis:

5.1.1 Contribution on Negative Capacitance

The contributions of this work on the Theory of Negative Capacitance are the following:

1. Starting from the Microscopic Mechanism of domain switching, we developed an intuitive ferroelectric model based on a Positive Feedback loop. We demonstrate that our model is equivalent to the Landau Model, and can reproduce and explain experimental data. This model is easy to implement in circuit simulators and shows better results in capturing transient effects.

2. We developed a quantitative explanation of the theory of NC based on the interplay of a positive feedback loop with a negative feedback loop.

3. We expanded the analysis of NC to the multi-domain scenario and derived conditions for hysteresis-free negative capacitance.

4. Based on circuit theory, we explained the role of ohmic losses on NC measurements.

5. We explored the impact of introducing a ferroelectric layer to the gate stack of a transistor (NCFET) and, comparing it to a baseline high-κ transistor, we concluded that there are improvements in: EOT (22%), SS (11%), $I_{D,Sat}$ (20%), Max $g_{m,sat}$ (19%) and ΔV_{TH}@1ks (54%). All this without mobility degradation.

6. We performed a RF characterization of our NCFETs and we conclude that compared to baseline, our devices present an improvement in the following intrinsic parameters: $g_{m,i}$ (28%), $C_{gg,i}$ (23%), and υ_{inj} (6%).

5.1.2 Contribution on BEOL Compatible Indium Oxide Transistors

The contributions of this work on BEOL compatible Indium Oxide Transistors are the following:

1. We fabricated BEOL IWO transistors that present the highest I_{ON} among oxide-semiconductors keeping a high I_{ON}/I_{OFF} ratio of 10^{12}. Hence, IWO FET is an excellent candidate for M3D capacitor-less 2T eDRAM application.

2. We showed that IWO 2T-eDRAM has lower access time than eNVM and lower standby power than traditional SRAM and eDRAM. Hence, it bridges the gap between these two types of memories.

3. Based on the 2T eDRAM gain cell, we demonstrated and studied the performance of the 4T and 6T TCAMs designs, and studied its capability as a computational associative memory.

5.2 Suggestions for Future Work

5.2.1 Negative Capacitance

1. Although the phenomenon of negative has been studied through several years, there is still discussion on the fundamental physics behind the polarization switching dynamics, and how well a phenomenological model can reflect these physics. Using novel techniques, such as high-resolution transmission electron microscopy (HRTEM), we can improve our understanding of this phenomenon.

2. A future step towards direct observation of NC is to reduce the ohmic losses in MFIM. It is necessary to understand if the ohmic losses are dominated by the experimental setup resistance, internal resistance of the MFIM stack, or the internal time delay of the polarization switching dynamics.

3. Scaling is a fundamental requirement to any emerging technology. It is mandatory to demonstrate that all the benefits showed in this work of HZO over HfO_2 gate-dielectric at 300 nm, still stand when the transistor length is reduced to a single digit nanometer.

5.2.2 BEOL Compatible Indium Oxide Transistors

1. Some aspects of the transistor have to be improved to allow IWO transistor to achieve a commercial level, such as Bias-Temperature Instability.
2. PMOS BEOL compatible transistors to enable BEOL compatible logic.
3. It is possible to introduce a ferroelectric to the gate-stack of an IWO FET, thus getting the same improvement that we saw in NCFETs.
4. Demonstration of fully integrated BEOL memory with FEOl logic.

APPENDIX A

GETTING ANISOTROPY PARAMETERS FOR LK EQUATION

To obtain the anisotropy parameters we follow the procedure presented in [44]. We start with equation (2.4), shown as reminder in equation (A.1).

$$V_{fe} = \alpha Q_P + \beta Q_P^3 \tag{A.1}$$

At $V_{fe} = 0$:$Q_P = \pm Q_0$, and it is possible to obtain equation (A.2).

$$Q_0 = \pm\sqrt{-\frac{\alpha}{\beta}} \tag{A.2}$$

Now, differentiating both sides of equation (A.1) and considering at $V_{fe} = V_c$, $\frac{dV_{fe}}{dQ_P} = 0$ and $Q_P = -Q_c$, we can get equation (A.3).

$$Q_c = \sqrt{-\frac{\alpha}{3\beta}} \tag{A.3}$$

Replacing $V_{fe} = V_c$ and $Q_P = Q_c$ in equation (A.1), we get equation (A.4). By simplifying this expression, we get the relation between V_c and Q_0, as show in equation (A.5).

$$V_c = \alpha\sqrt{-\frac{\alpha}{3\beta}} + \beta\left(\sqrt{-\frac{\alpha}{3\beta}}\right)^3 \tag{A.4}$$

$$V_c = -\frac{2\alpha}{3\sqrt{3}}Q_0 \tag{A.5}$$

From equation (A.5) we can directly obtain β by replacing equation (A.5) in equation (A.2). The anisotropy parameters are shown in equation (A.6).

$$\alpha = -\frac{3\sqrt{3}}{2}\frac{V_c}{Q_0} \; and \; \beta = \frac{3\sqrt{3}}{2}\frac{V_c}{Q_0^3} \tag{A.6}$$

APPENDIX B

EXPRESSION OF POSITIVE FEEDBACK LOOP

From the positive feedback loop, shown in fig. B.1, we can directly derive the expressions that are shown in equations: (B.1), (B.2), (B.3) and (B.4).

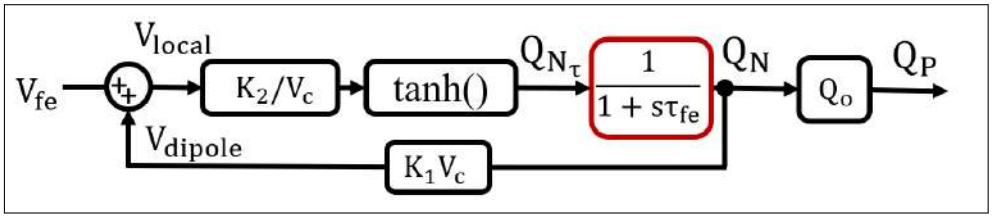


Figure B.1. Positive Feedback loop that arise from microscopic mechanism of domain switching

$$V_{local} = V_{fe} + V_{dipole} \tag{B.1}$$

$$V_{dipole} = K_1 V_c Q_N \tag{B.2}$$

$$tanh^{-1}\left(Q_{N_\tau}\right) = \frac{K_2}{V_c}\left(V_{local}\right) \tag{B.3}$$

$$\frac{dQ_N}{dt} = \frac{Q_{N_\tau} - Q_N}{\tau_{fe}} \tag{B.4}$$

Now by merging these expressions, we can derive equation (B.5).

$$\tanh^{-1}(Q_{N\tau}) = \frac{K_2}{V_c}\left(V_{fe} + K_1 V_c\left(Q_{N\tau} - \tau_{fe}\frac{dQ_N}{dt}\right)\right) \tag{B.5}$$

By simplifying equation (B.5) and assuming $Q_N = Q_{N\tau}$, we can derive the final dynamic expression used to describe the positive feedback loop that is shown in equation (B.6).

$$\frac{dQ_N}{dt} = \frac{1}{\tau_{fe}K_1V_c}\left(V_{fe} - \frac{V_c}{K_2}tanh^{-1}(Q_N) + K_1V_cQ_N\right) \tag{B.6}$$

We can move equation (B.6) to the charge domain by simply multiplying both sides by Q_0, the final expression of ferroelectric switching dynamics is shown in equation (B.7).

$$\frac{dQ_P}{dt} = \frac{Q_0}{\tau_{fe}K_1V_c}\left(V_{fe} - \frac{V_c}{K_2}tanh^{-1}(Q_N) + K_1V_cQ_N\right) \tag{B.7}$$

APPENDIX C

EQUIVALENCE OF POSITIVE-FEEDBACK WITH LK EQUATION

Starting from equation (2.11), we expand $\tanh^{-1}$ using Maclaurin expansion [1], and equation (2.11) becomes equation (C.1).

$$V_{fe} = \left(\frac{V_c}{K_2 Q_0} - \frac{K_1 V_c}{Q_0}\right) Q_P + \frac{V_c}{3K_2 Q_0^3} Q_P^3 \tag{C.1}$$

Imposing the conditions $Q_P = \pm Q_0$ at $V_{fe} = 0$ in equation (C.1), we can derive the product of the constants shown in equation (C.2).

$$K_1 K_2 = \frac{4}{3} \tag{C.2}$$

Differentiating equation C.1, we get the expression shown in equation (C.3).

$$\frac{dV_{fe}}{dQ_P} = \frac{V_c}{K_2 Q_0} - \frac{K_1 V_c}{Q_0} + \frac{V_c}{K_2 Q_0^3} Q_P^2 \tag{C.3}$$

Using the conditions $\frac{dV_{fe}}{dQ_P} = 0$ at $V_{fe} = V_c$, and, replacing in equation (C.3), we can get an expression for Q_c as is shown in equation (C.4).

$$Q_c = \pm\sqrt{\frac{1}{3}} Q_0 \tag{C.4}$$

Replacing V_c and Q_c in equation (C.1), we can get K_2, and, using equation (C.2), we can get K_1. The values are shown in equation (C.5).

---

[1] $tanh^{-}1(x) = x + \frac{x^3}{3} + \frac{x^5}{5}...$, for this work we use $tanh^{-}1(x) = x + \frac{x^3}{3}$

$$K_1 = 6\sqrt{3} \text{ and } K_2 = \frac{2}{9\sqrt{3}} \tag{C.5}$$

Finally, replacing in equation (C.1), the steady state relation between the charge and the voltage is given by equation (C.6), which is exactly the same equation derived from Landau Theory, shown in equation (2.6).

$$V_{fe} = \frac{-3\sqrt{3}}{2}\frac{V_c}{Q_0}Q_P + \frac{3\sqrt{3}}{2}\frac{V_c}{Q_0^3}Q_P^3 \tag{C.6}$$

APPENDIX D

STABILITY CONDITION USING LK EQUATION

To derive the stability condition for a ferroelectric in series with a capacitor we use a linear stability analysis. Equation (D.1) shows the differential equation that describes the dynamics of a ferroelectric in series with a capacitor.

$$\dot{Q}_P = f\left(Q_P\right) = \frac{1}{\rho}\left(V_{in} - \frac{Q_P}{C_S} - \alpha Q_P - \beta Q_P^3\right) \tag{D.1}$$

Now we take the derivative of $f\left(Q_P\right)$ and evaluate and the point of interest i.e. $Q_P = 0$. The result is shown in equation (D.2).

$$\left.\frac{df}{dQ_P}\right|_{Q_P=0} = \frac{1}{\rho}\left(-\frac{1}{C_s} - \alpha\right) \tag{D.2}$$

According to linear stability theory, if all the eigenvalues of the Jacobian matrix have real part less than zero, then the steady state is stable. Our case is trivial: the determinant of a 1×1 matrix is a single value, and we don't have have imaginary part. The stability condition is shown in equation (D.3).

$$C_s < -\frac{1}{\alpha} = \frac{2}{3\sqrt{3}}\frac{Q_0}{V_c} \tag{D.3}$$

APPENDIX E

STABILITY CONDITION USING POSITIVE FEEDBACK MODEL

We follow the same approach presented previously in appendix D. Starting from the positive feed-back loop, the differential equation that describes the dynamic of a ferroelectric in series with a capacitor is shown in equation (E.1).

$$\dot{Q}_P = f\left(Q_P\right) = \frac{1}{\rho}\left(V_{in} - \frac{Q_P}{C_s} - \frac{V_c}{K_2} tanh^{-1}\left(\frac{Q_P}{Q_0}\right) + \frac{K_1 V_c Q_P}{Q_0}\right) \tag{E.1}$$

We take the derivative of f (Q_P) and evaluate and the point of interest i.e. $Q_P = 0$. The result is shown in equation (E.2).

$$\left.\frac{df}{dQ_P}\right|_{Q_P=0} = \frac{1}{\rho}\left(-\frac{1}{C_s} - \frac{V_c}{Q_0 K_2} + \frac{K_1 V_c}{Q_0}\right) < 0 \tag{E.2}$$

Starting from equation equation (E.2), we can derive the stability condition shown in equation (E.3). This stability condition depends on the positive feedback gains: K_1 and K_2.

$$\frac{V_c}{Q_0}\left(K_1 - \frac{1}{K_2}\right) < \frac{1}{C_s} \tag{E.3}$$

Now, if we replace the gains by the values obtained in appendix B, we can derive the same stability condition.

$$C_s < \frac{2}{3\sqrt{3}} \frac{Q_0}{V_c} \tag{E.4}$$

APPENDIX F

SPICE NETLIST OF FERROLECTRIC MODEL

Spice model of single domain ferroelectric.

```
* Optimal SPICE implementation of LK-eqution
* University of Notre Dame, IN, USA
* Implemented by Jorge Gomez
* 9/25/2019
* Netlist Ferroelectric
.subckt Cfe_Landau in out s d PARAMS:
+Vc=2.2 Qo=3n K1=10.392 K2=0.128
* Initial Conditions
.ic V(d)= 0
.ic V(s)= 0
* Input
Bref Fe 0 V=V(in,out)
* Positive Feedback
Bs s 0 V=tanh(({K2}/{Vc})*(V(Fe)+{K1}*{Vc}*V(d))) ic = 0
* Switching Time Contant
Rd s d 100
Cd d 0 1e-12
* Ouput Current
Bi in out I=(ddt(({Qo}*V(d))))
.ends
```

APPENDIX G

FERROELECTRIC MODEL FOR NEUROMORPHIC COMPUTING

Our ferroelectric model has been also used to successfully model a FeFET base spiking neuron ([80], [81], [82] and [83]). Fig. G.1 (a) illustrates the circuit of the FeFET based spiking neuron. Fundamentally, it works as a relaxation oscillator that periodically charges and discharges capacitor C with I_D and I_M, which are the currents flowing through the two transistors. The pull-up FeFET which charges C, can be viewed as a ferroelectric layer connected to the gate of a normal N-type FET. A traditional pull-down MOSFET is connected to the capacitor and provides a discharging path. To understand the source of oscillation, let us assume the gate voltages V_{GF}, V_{GM} and V_{DD} are fixed. During the charging phase, the voltage of capacitor C, V_N, is relatively low and triggers the coercion of the ferroelectric layer, injects charge into the gate node V_g and abruptly switches the FeFET to an ON state. Consequently, I_D charges the capacitor throughout this phase. As V_N rises to a certain threshold that initiates the discharging phase, the ferroelectric layer reaches another coercive voltage, removes the charge from V_g and switches the FeFET to an OFF state. Since I_D is now almost zero, I_M is able to discharge the capacitor and V_N decreases. Then the whole process resumes and the two phases are repeated alternatively and V_N keeps oscillating between the two boundary voltages V_{t1} and V_{t2}. V_{GF} and V_{GM} can be used to control the oscillating frequency when one of these two gate voltages is fixed. V_{GF} and V_{GM} are defined through the integration of exicitatory and inhibitory incoming spikes. Fig. G.1 (b) shows the experimental demonstration of the neuron and Fig. G.1 (c) shows the results of the SPICE simulations. In this

case the model was used to simulate a Lead Zirconate Titanate (PZT) ferroelectric. PZT presents bigger and less domains than HfO_2 [84] making it a good alternative for oscillatory implementations.

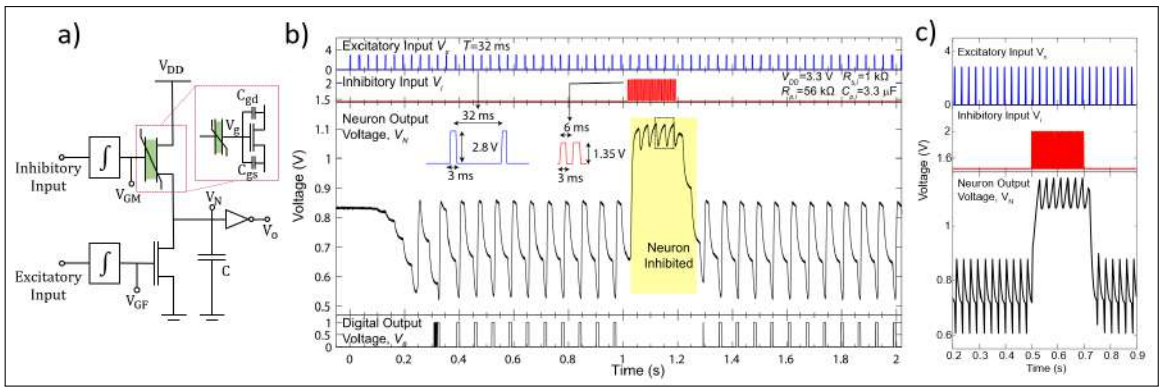


Figure G.1. (a)Circuit of the FeFET based spiking neuron, (b) experimental demonstration of the neuron and (c) results of the SPICE simulations using our developed model.

This model has been used to evaluate a fully connected network [80] and Swarm Intelligence - Spiking Neuronal Network (SI-SNN) [83].

APPENDIX H

DE-EMBEDDING SCHEME

The goal of this section is to obtain the S-parameters of the intrinsic device (S_{int}). We separate the process in three steps:

1. *Experimental Setup Calibration:* We de-embed the setup by doing an Off-waffer LRRM (line-reflect-reflect-match) calibration with a calibration standard (Cascade 104-783 W-band impedance standard).
2. *Pad Strip de-embedding:* We included two On-wafer test structures to de-embedd the pad strip: open and short, which will provide a model of the pad parasitic, so we can de-embed them.
3. *Extrinsic elements de-embedding:* By measuring Cold-Fet parameters of the transistor we obtain only the extrinsic parameters of the transistor and, hereafter, we can de-embed them from the intrinsic parameters.

Fig. H.1 shows the equivalent circuit model used to de-embed all on-wafer parasitics of the DUT. The experimental setup consists on a DC Voltage Source (HP4142B) and a vector network analyzer (VNA) (Agilent E8381C). For probing we used 150μm pitch Cascade Infinity i110-A probes. The experimental setup calibration is done automatically by WinCal XE program installed on the VNA, so we will not discuss it here.

H.1 Pad Strip De-Embedding

Fig. H.2 shows on-wafer test structures for pad strip de-emebedding: layout view and equivalent circuits. The pad strip s-parameters are used to subtract parasitcs from the DUT data, as shown in Fig. H.3. Pad capacitances are removed from

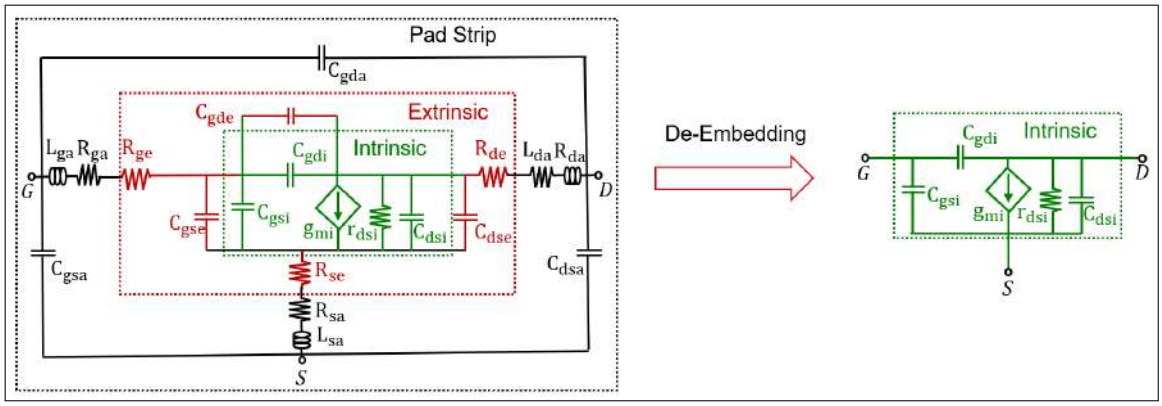


Figure H.1. Equivalent circuit model used to de-embed all on-wafer parasitics

both the DUT and Short structures by subtracting Y_{OPEN} from their respective Y-parameters. Then, they are converted to Z-parameters to facilitate removal of pad resistance and inductance. The corrected (pad-stripped) S-matrix of the DUT is acquired, which still includes Extrinsic elements.

H.2 Extrinsic Elements De-Embedding

Using keysight ADS software, we optimize the values of the extrisic circuital components to fit the S parameters of the transistor measured in Cold-Fet conditions ($V_g = 0,\ V_d = 0,\ V_s = 0$) (Fig. H.4 (a)). Then, we also optimized the S-parameter fitting at the point of max f_T (Fig. H.4 (b)). Once all the circuital elements have been properly calibrated, we can subtract the extrinsic components of the circuit and simulate the intrinsic performance of the device.

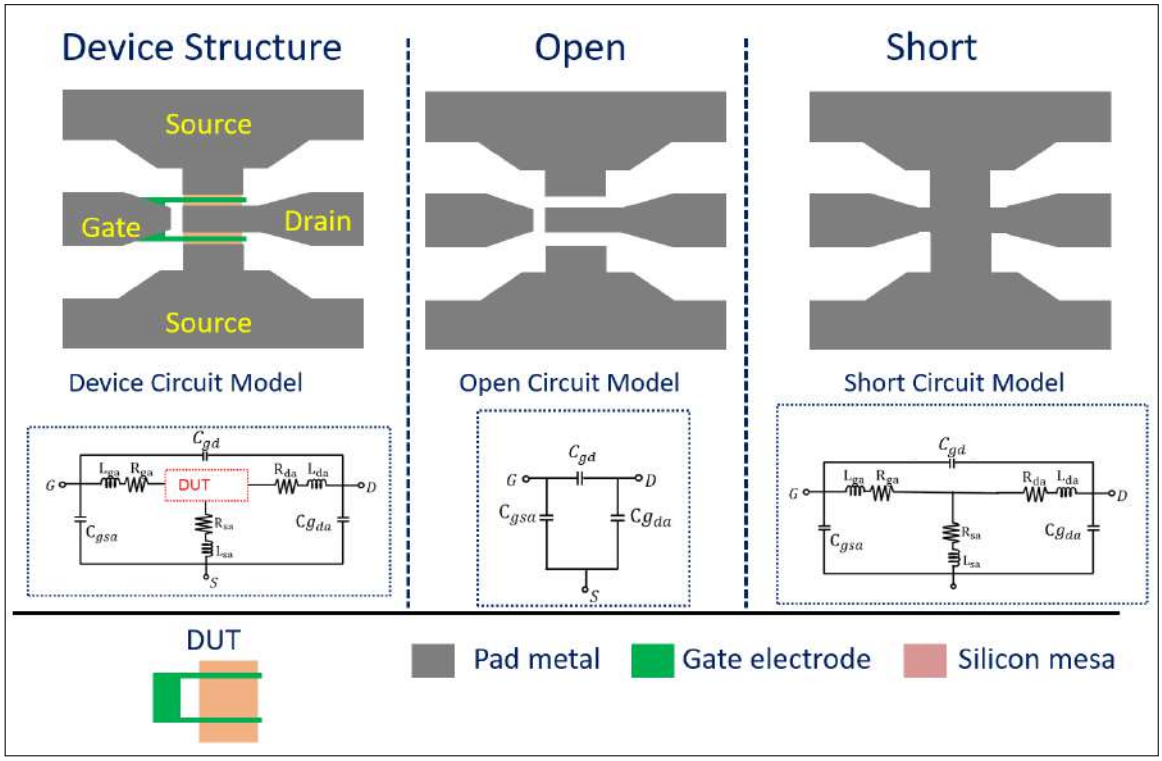


Figure H.2. On-wafer test structures for pad strip de-emebedding with equivalent circuit model

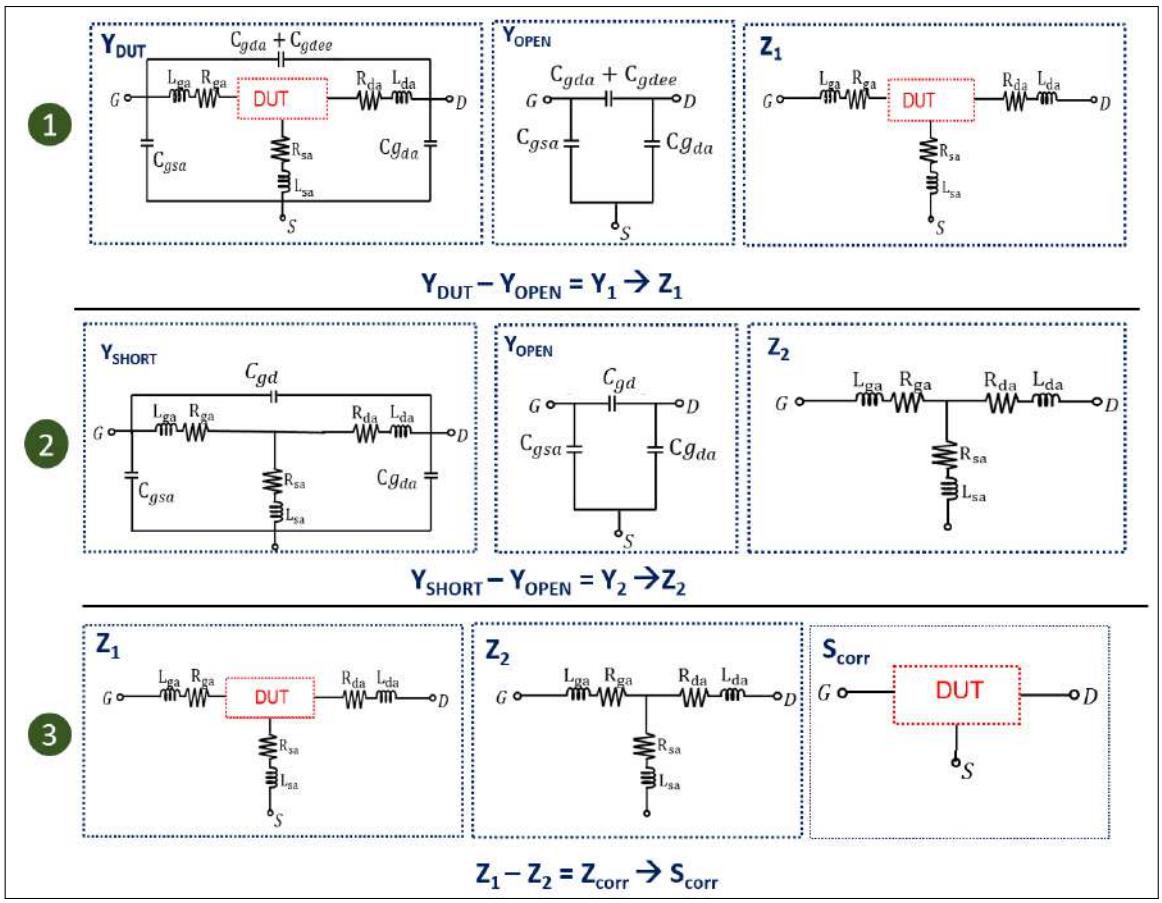


Figure H.3. Procedure to de-embed pad-strip parasitic from DUT.

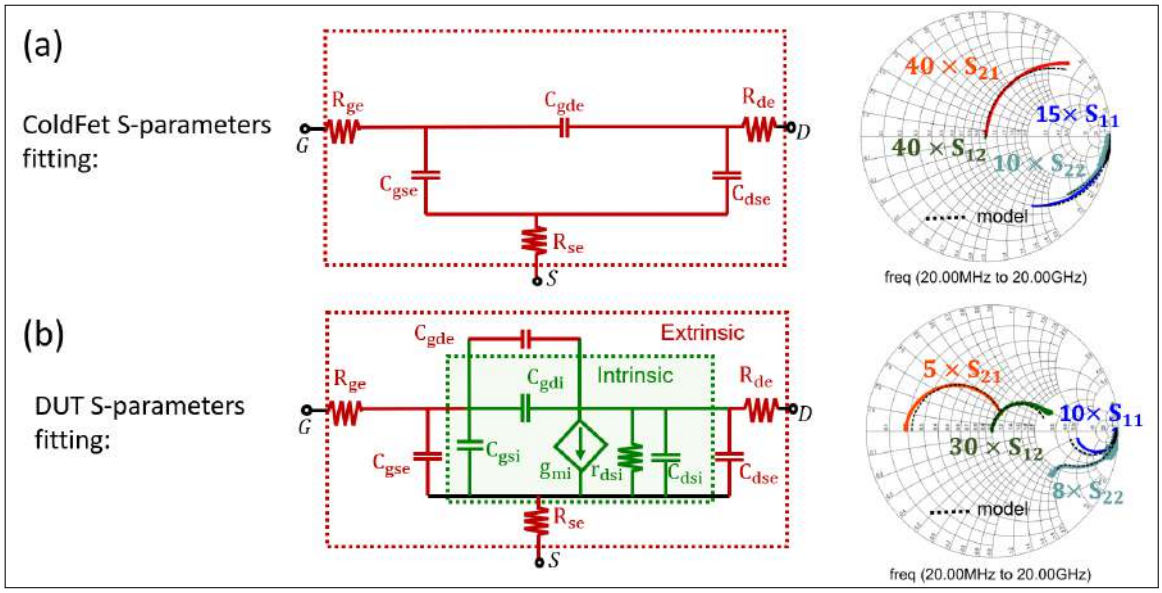


Figure H.4. (a) Equivalent circuit and S-paramaeters fitting of ColdFet, and (b) equivalent circuit and S-paramters fitting of NCFET.

APPENDIX I

ARTICLES PUBLISHED ABOUT THIS WORK

1. Z. Wang, B. Crafton, J. Gomez, R. Xu, A. Luo, Z. Krivokapic, L. Martin, S. Datta, A. Raychowdhury, and A. I. Khan, "Experimental demonstration of ferroelectric spiking neurons for unsupervised clustering," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 13.3.1–13.3.4, 2018

2. J. Gomez, S. Dutta, K. Ni, S. Joshi, and S. Datta, "Steep slope ferroelectric field effect transistor," in *2019 Electron Devices Technology and Manufacturing Conference (EDTM)*, pp. 59–61, 2019

3. Y. Fang, J. Gomez, Z. Wang, S. Datta, A. I. Khan, and A. Raychowdhury, "Neuro-mimetic dynamics of a ferroelectric fet-based spiking neuron," *IEEE Electron Device Letters*, vol. 40, no. 7, pp. 1213–1216, 2019

4. S. Dutta, A. Saha, P. Panda, W. Chakraborty, J. Gomez, A. Khanna, S. Gupta, K. Roy, and S. Datta, "Biologically plausible ferroelectric quasi-leaky integrate and fire neuron," in *2019 Symposium on VLSI Technology*, pp. T140–T141, 2019

5. S. Dutta, W. Chakraborty, J. Gomez, K. Ni, S. Joshi, and S. Datta, "Energy-efficient edge inference on multi-channel streaming data in 28nm hkmg fefet technology," in *2019 Symposium on VLSI Technology*, pp. T38–T39, 2019

6. J. Gomez, S. Dutta, K. Ni, B. Grisafe, J. Smith, A. Khan, and S. Datta, "Significance of multi and few domain ferroelectric switching dynamics for steep-slope non-hysteretic ferroelectric field effect transistor," in *2019 Device Research Conference (DRC)*, pp. 247–248, 2019

7. Y. Fang, Z. Wang, J. Gomez, S. Datta, A. I. Khan, and A. Raychowdhury, "A swarm optimization solver based on ferroelectric spiking neural networks," *Frontiers in Neuroscience*, vol. 13, p. 855, 2019

8. J. Gomez, S. Dutta, K. Ni, J. A. Smith, B. Grisafe, A. Khan, and S. Datta, "Hysteresis-free negative capacitance in the multi-domain scenario for logic applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 7.1.1–7.1.4, 2019

9. S. Dutta, C. Schafer, J. Gomez, K. Ni, S. Joshi, and S. Datta, "Supervised learning in all fefet-based spiking neural network: Opportunities and challenges," *Frontiers in Neuroscience*, vol. 14, p. 634, 2020

10. H. Ye, J. Gomez, W. Chakraborty, S. Spetalnick, S. Dutta, K. Ni, A. Raychowdhury, and S. Datta, "Double-gate w-doped amorphous indium oxide transistors for monolithic 3d capacitorless gain cell edram," in *2020 IEEE International Electron Devices Meeting (IEDM)*, pp. 28.3.1–28.3.4, 2020

11. N. Tasneem, P. V. Ravindran, Z. Wang, J. Gomez, J. Hur, S. Yu, S. Datta, and A. I. Khan, "Differential charge boost in hysteretic ferroelectric–dielectric heterostructure capacitors at steady state," *Applied Physics Letters*, vol. 118, no. 12, p. 122901, 2021

12. W. Chakraborty, M. S. Jose, J. Gomez, A. Saha, K. A. Aabrar, P. Fay, S. Gupta, and S. Datta, "Higher-k zirconium doped hafnium oxide (hzo) trigate transistors with higher dc and rf performance and improved reliability," in *2021 Symposium on VLSI Technology*, pp. 1–2, 2021

BIBLIOGRAPHY

1. C. 2020, "Cisco annual internet report (2018-2023)," *White Paper Cisco public*, 2020.
2. K. Rupp, "48 years of microprocessor trend data," *https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/*, Access: 06/03/2021.
3. A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of cmos device performance from 180nm to 7nm," *Integration*, vol. 58, pp. 74–81, 2017.
4. S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nat. Electron*, vol. q, pp. 442–450, 2018.
5. M. Hoffmann, S. Slesazeck, and T. Mikolajick, "Progress and future prospects of negative capacitance electronics: A materials perspective," *APL Materials*, vol. 9, no. 2, p. 020902, 2021.
6. S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018.
7. V. Sze, Y. Chen, T. 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.
8. "The memory hierarchy," *https://computationstructures.org/lectures/caches/caches.html*, Access: 08/08/2021.
9. "Pulp platform," *https://pulp-platform.org/*, Access: 08/08/2021.
10. J.-h. Sim, "Technology challenges and directions of sram, dram, and edram cell scaling in sub-20nm generations," *SK hynix*, 2013.
11. "Intel's embedded dram: New era of cache memory," *https://www.eetimes.com/intels-embedded-dram-new-era-of-cache-memory*, Access: 08/03/2021.
12. S. K. Kim, S. W. Lee, J. H. Han, B. Lee, S. Han, and C. S. Hwang, "Capacitors with an equivalent oxide thickness of ¡0.5 nm for nanoscale electronic semiconductor memory," *Advanced Functional Materials*, vol. 20, no. 18, pp. 2989–3003, 2010.

13. H. Koji, "Dram: Challeneges and opportunities," *VLSI Simposia Short Course*, 2021.

14. P. Sung-ki, "The future memory technologies," *Hynix semiconductor Inc.*, 2011.

15. "Micron's d1a dram products use arf-i based lithography without euvl photomask applied," *https://www.techinsights.com/blog/memory/micron-1a-dram-technology*, Access: 08/03/2021.

16. "Dram scaling," *https://semiwiki.com/semiconductor-services/297904-spie-2021-applied-materials-dram-scaling/*, Access: 08/03/2021.

17. A. Khan, "Negative capacitance for ultra-low power computing," *EECS Department University of California Berkeley*, Jul 2015.

18. A. Saeidi, F. Jazaeri, F. Bellando, I. Stolichnov, C. C. Enz, and A. M. Ionescu, "Negative capacitance field effect transistors; capacitance matching and non-hysteretic operation," in *2017 47th European Solid-State Device Research Conference (ESSDERC)*, pp. 78–81, 2017.

19. A. Khan, K. Chatterjee, and B. Wang, "Negative capacitance in a ferroelectric capacitor," *Nature Mater*, vol. 14, pp. 182–186, 2015.

20. M. Hoffmann, F. Fengler, M. Herzig, T. Mittmann, B. Max, U. Schroeder, R. Negrea, P. Lucian, S. Slesazeck, and T. Mikolajick, "Unveiling the double-well energy landscape in a ferroelectric layer," in *Nature*, 2019.

21. R. Landauer, "Can capacitance be negative?," in *Collective Phenomena*, vol. 2, pp. 167–170, 1976.

22. M. Hoffmann, B. Max, T. Mittmann, U. Schroeder, S. Slesazeck, and T. Mikolajick, "Demonstration of high-speed hysteresis-free negative capacitance in ferroelectric hf0.5zr0.5o2," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 31.6.1–31.6.4, 2018.

23. J. Robertson and R. M. Wallace, "High-k materials and metal gates for cmos applications," *Materials Science and Engineering: R: Reports*, vol. 88, pp. 1–41, 2015.

24. T. Ando, M. Copel, J. Bruley, M. M. Frank, H. Watanabe, and V. Narayanan, "Physical origins of mobility degradation in extremely scaled sio2/hfo2 gate stacks with la and al induced dipoles," *Applied Physics Letters*, vol. 96, no. 13, p. 132904, 2010.

25. T. Ando, "Ultimate scaling of high-k gate dielectrics: Higher-k or interfacial layer scavenging?," *Materials*, vol. 5, no. 3, pp. 478–500, 2012.

26. M. M. Frank, "High-k / metal gate innovations enabling continued cmos scaling," in *2011 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, pp. 25–33, 2011.

27. S. Rabii, "Plenary talk," in *Symposia on VLSI Technology and circuits (VLSI)*, 2019.

28. A. Parashar, P. Raina, Y. S. Shao, Y. Chen, V. A. Ying, A. Mukkara, R. Venkatesan, B. Khailany, S. W. Keckler, and J. Emer, "Timeloop: A systematic approach to dnn accelerator evaluation," in *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 304–315, 2019.

29. K. Nomura, H. Ohta, A. Takagi, T. Kamiya, M. Hirano, and H. Hosono, "Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors," *Nature*, vol. 432, Nov 2004.

30. T. Kizu, S. Aikawa, N. Mitoma, M. Shimizu, X. Gao, M.-F. Lin, T. Nabatame, and K. Tsukagoshi, "Low-temperature processable amorphous in-w-o thin-film transistors with high mobility and stability," *Applied Physics Letters*, vol. 104, no. 15, p. 152103, 2014.

31. "Intel and taiwan semiconductor: A tale of two cities revisisted," *https://seekingalpha.com/article/4297720-intel-and-taiwan-semiconductor-tale-of-two-cities-revisited*, Access: 06/24/2021.

32. E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020.

33. G. E. Moore, "Cramming more components onto integrated circuits," *Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. IEEE Solid-State Circuits Soc. Newsl.*, vol. 11, 2009.

34. K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, "Impact ionization mos (i-mos)-part i: device and circuit simulations," *IEEE Transactions on Electron Devices*, vol. 52, no. 1, pp. 69–76, 2005.

35. A. C. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond cmos logic," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095–2110, 2010.

36. C. Onal, "Improving the realtibility and performance of impact-ionization based transistores for low power logic applications," *Stanford University*, June 2010.

37. S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Letters*, vol. 8, no. 2, pp. 405–410, 2008.

38. Z. Jia, M. Maggioni, J. Smith, and D. P. Scarpazza, "Dissecting the nvidia turing t4 gpu via microbenchmarking," 2019.

39. S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "Eie: Efficient inference engine on compressed deep neural network," *SIGARCH Comput. Archit. News*, vol. 44, p. 243–254, June 2016.

40. L. Wei, O. Mysore, and D. Antoniadis, "Virtual-source-based self-consistent current and charge fet models: From ballistic to drift-diffusion velocity-saturation operation," *IEEE Transactions on Electron Devices*, vol. 59, no. 5, pp. 1263–1271, 2012.

41. H. Ye, J. Gomez, W. Chakraborty, S. Spetalnick, S. Dutta, K. Ni, A. Raychowdhury, and S. Datta, "Double-gate w-doped amorphous indium oxide transistors for monolithic 3d capacitorless gain cell edram," in *2020 IEEE International Electron Devices Meeting (IEDM)*, pp. 28.3.1–28.3.4, 2020.

42. J. Valasek, "Properties of rochelle salt related to the piezo-electric effect," *Phys. Rev.*, vol. 20, pp. 639–664, Dec 1922.

43. T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, "Ferroelectricity in hafnium oxide: Cmos compatible ferroelectric field effect transistors," in *2011 International Electron Devices Meeting*, pp. 24.5.1–24.5.4, 2011.

44. A. I. Khan, U. Radhakrishna, K. Chatterjee, S. Salahuddin, and D. A. Antoniadis, "Negative capacitance behavior in a leaky ferroelectric," *IEEE Transactions on Electron Devices*, vol. 63, pp. 4416–4422, Nov 2016.

45. P. Chandra and P. Littlewood, "A landau primer for ferroelectrics," *arXiv:cond-mat/0609347v1*, 2006.

46. A. I. Khan, "On the microscopic origin of negative capacitance in ferroelectric materials: A toy model," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 9.3.1–9.3.4, 2018.

47. J. Gomez, S. Dutta, K. Ni, J. A. Smith, B. Grisafe, A. Khan, and S. Datta, "Hysteresis-free negative capacitance in the multi-domain scenario for logic applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 7.1.1–7.1.4, 2019.

48. J. Gomez, S. Dutta, K. Ni, S. Joshi, and S. Datta, "Steep slope ferroelectric field effect transistor," in *2019 Electron Devices Technology and Manufacturing Conference (EDTM)*, pp. 59–61, 2019.

49. J. Gomez, S. Dutta, K. Ni, B. Grisafe, J. Smith, A. Khan, and S. Datta, "Significance of multi and few domain ferroelectric switching dynamics for steep-slope non-hysteretic ferroelectric field effect transistor," in *2019 Device Research Conference (DRC)*, pp. 247–248, 2019.

50. P. Sharma, J. Zhang, K. Ni, and S. Datta, "Time-resolved measurement of negative capacitance," *IEEE Electron Device Letters*, vol. 39, no. 2, pp. 272–275, 2018.

51. B. Obradovic, T. Rakshit, R. Hatcher, J. A. Kittl, and M. S. Rodder, "Modeling transient negative capacitance in steep-slope fefets," *IEEE Transactions on Electron Devices*, vol. 65, no. 11, pp. 5157–5164, 2018.

52. F. Zhang, Y. Peng, X. Deng, J. Huo, Y. Liu, G. Han, Z. Wu, H. Yin, and Y. Hao, "Theoretical study of negative capacitance finfet with quasi-antiferroelectric material," *IEEE Transactions on Electron Devices*, vol. 68, no. 6, pp. 3074–3079, 2021.

53. K. Ni, A. Saha, W. Chakraborty, H. Ye, B. Grisafe, J. Smith, G. B. Rayner, S. Gupta, and S. Datta, "Equivalent oxide thickness (eot) scaling with hafnium zirconium oxide high-k dielectric near morphotropic phase boundary," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 7.4.1–7.4.4, 2019.

54. S. Mohsenifar and S. M.H., "Gate stack high-k materials for si-based mosfets past, present, and futures," *Phys. Rev.*, vol. 1, pp. 12–24, 2015.

55. M. M. Frank, "High-k/metal gate innovations enabling continued cmos scaling," pp. 50–58, 2011.

56. W. Chakraborty, H. Ye, B. Grisafe, I. Lightcap, and S. Datta, "Low thermal budget (¡250c) dual-gate amorphous indium tungsten oxide (iwo) thin film transistor for monolithic 3d integration," in *2020 International Symposium on VLSI Technology*, pp. 1–1, 2020.

57. L. Wei, O. Mysore, and D. Antoniadis, "Virtual-source-based self-consistent current and charge fet models: From ballistic to drift-diffusion velocity-saturation operation," *IEEE Transactions on Electron Devices*, vol. 59, no. 5, pp. 1263–1271, 2012.

58. W. Chakraborty, B. Grisafe, H. Ye, I. Lightcap, K. Ni, and S. Datta, "Beol compatible dual-gate ultra thin-body w-doped indium-oxide transistor with ion = 370ua/um, ss = 73mv/dec and ion /ioff ratio," in *2020 IEEE Symposium on VLSI Technology*, pp. 1–2, 2020.

59. K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim, "A 3t gain cell embedded dram utilizing preferential boosting for high density and low power on-die caches," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1495–1505, 2011.

60. N. Planes, O. Weber, V. Barral, S. Haendler, D. Noblet, D. Croain, M. Bocat, P. . Sassoulas, X. Federspiel, A. Cros, A. Bajolet, E. Richard, B. Dumont, P. Perreau, D. Petit, D. Golanski, C. Fenouillet-Béranger, N. Guillot, M. Rafik, V. Huard, S. Puget, X. Montagner, M. . Jaud, O. Rozeau, O. Saxod, F. Wacquant, F. Monsieur, D. Barge, L. Pinzelli, M. Mellier, F. Boeuf, F. Arnaud, and M. Haond, "28nm fdsoi technology platform for high-speed low-voltage digital applications," in *2012 Symposium on VLSI Technology (VLSIT)*, pp. 133–134, 2012.

61. S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu,

Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang, and K. Zhang, "A 14nm logic technology featuring 2nd-generation finfet, air-gapped interconnects, self-aligned double patterning and a 0.0588 µm2 sram cell size," in *2014 IEEE International Electron Devices Meeting*, pp. 3.7.1–3.7.3, 2014.

62. M. Oota, Y. Ando, K. Tsuda, T. Koshida, S. Oshita, A. Suzuki, K. Fukushima, S. Nagatsuka, T. Onuki, R. Hodo, T. Ikeda, and S. Yamazaki, "3d-stacked caac-in-ga-zn oxide fets with gate length of 72nm," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 3.2.1–3.2.4, 2019.

63. M. Meterelliyoz, F. H. Al-amoody, U. Arslan, F. Hamzaoglu, L. Hood, M. Lal, J. L. Miller, A. Ramasundar, D. Soltman, I. Wan, Y. Wang, and K. Zhang, "2nd generation embedded dram with 4x lower self refresh power in 22nm tri-gate cmos technology," in *2014 Symposium on VLSI Circuits Digest of Technical Papers*, pp. 1–2, 2014.

64. K. C. Chun, P. Jain, T. Kim, and C. H. Kim, "A 1.1v, 667mhz random cycle, asymmetric 2t gain cell embedded dram with a 99.9 percentile retention time of 110µsec," in *2010 Symposium on VLSI Circuits*, pp. 191–192, 2010.

65. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

66. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

67. M. Meterelliyoz, F. H. Al-amoody, U. Arslan, F. Hamzaoglu, L. Hood, M. Lal, J. L. Miller, A. Ramasundar, D. Soltman, I. Wan, Y. Wang, and K. Zhang, "2nd generation embedded dram with 4x lower self refresh power in 22nm tri-gate cmos technology," in *2014 Symposium on VLSI Circuits Digest of Technical Papers*, pp. 1–2, 2014.

68. K. Ni, X. Yin, A. F. Laguna, S. Joshi, S. Dünkel, M. Trentzsch, J. Müller, S. Beyer, M. Niemier, X. S. Hu, and S. Datta, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, pp. 521–529, 2019.

69. D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *Journal of Big Data*, vol. 2, pp. 2196–1115, 2014.

70. H. Lee, A. Lee, F. Ebrahimi, P. K. Amiri, and K. L. Wang, "Array-level analysis of magneto-electric random-access memory for high-performance embedded applications," *IEEE Magnetics Letters*, vol. 8, pp. 1–5, 2017.

71. B. Gopireddy and J. Torrellas, "Designing vertical processors in monolithic 3d," in *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*, pp. 643–656, 2019.

72. X. Yin, K. Ni, D. Reis, S. Datta, M. Niemier, and X. S. Hu, "An ultra-dense 2fefet tcam design based on a multi-domain fefet model," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 9, pp. 1577–1581, 2019.

73. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

74. "Grammatical facial expressions data set," *https://archive.ics.uci.edu/ml/datasets/Grammatica* Access: 08/08/2021.

75. D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Ambient Assisted Living and Home Care* (J. Bravo, R. Hervás, and M. Rodríguez, eds.), (Berlin, Heidelberg), pp. 216–223, Springer Berlin Heidelberg, 2012.

76. R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Phys. Rev. E*, vol. 64, p. 061907, Nov 2001.

77. A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles," *Sensors and Actuators B: Chemical*, vol. 166-167, pp. 320–329, 2012.

78. R. C. B. Madeo, C. A. M. Lima, and S. M. Peres, "Gesture unit segmentation using support vector machines: Segmenting gestures from rest positions," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, (New York, NY, USA), p. 46–52, Association for Computing Machinery, 2013.

79. "Isolet data set," *http://archive.ics.uci.edu/ml/datasets/ISOLET*, Access: 08/08/2021.

80. Z. Wang, B. Crafton, J. Gomez, R. Xu, A. Luo, Z. Krivokapic, L. Martin, S. Datta, A. Raychowdhury, and A. I. Khan, "Experimental demonstration of ferroelectric spiking neurons for unsupervised clustering," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 13.3.1–13.3.4, 2018.

81. S. Dutta, A. Saha, P. Panda, W. Chakraborty, J. Gomez, A. Khanna, S. Gupta, K. Roy, and S. Datta, "Biologically plausible ferroelectric quasi-leaky integrate and fire neuron," in *2019 Symposium on VLSI Technology*, pp. T140–T141, 2019.

82. Y. Fang, J. Gomez, Z. Wang, S. Datta, A. I. Khan, and A. Raychowdhury, "Neuro-mimetic dynamics of a ferroelectric fet-based spiking neuron," *IEEE Electron Device Letters*, vol. 40, no. 7, pp. 1213–1216, 2019.

83. Y. Fang, Z. Wang, J. Gomez, S. Datta, A. I. Khan, and A. Raychowdhury, "A swarm optimization solver based on ferroelectric spiking neural networks," *Frontiers in Neuroscience*, vol. 13, p. 855, 2019.

84. F. P. G. Fengler, M. Pešić, S. Starschich, T. Schneller, C. Künneth, U. Böttger, H. Mulaosmanovic, T. Schenk, M. H. Park, R. Nigon, P. Muralt, T. Mikolajick, and U. Schroeder, "Domain pinning: Comparison of hafnia and pzt based ferroelectrics," *Advanced Electronic Materials*, vol. 3, no. 4, p. 1600505, 2017.

85. S. Dutta, W. Chakraborty, J. Gomez, K. Ni, S. Joshi, and S. Datta, "Energy-efficient edge inference on multi-channel streaming data in 28nm hkmg fefet technology," in *2019 Symposium on VLSI Technology*, pp. T38–T39, 2019.

86. S. Dutta, C. Schafer, J. Gomez, K. Ni, S. Joshi, and S. Datta, "Supervised learning in all fefet-based spiking neural network: Opportunities and challenges," *Frontiers in Neuroscience*, vol. 14, p. 634, 2020.

87. N. Tasneem, P. V. Ravindran, Z. Wang, J. Gomez, J. Hur, S. Yu, S. Datta, and A. I. Khan, "Differential charge boost in hysteretic ferroelectric–dielectric heterostructure capacitors at steady state," *Applied Physics Letters*, vol. 118, no. 12, p. 122901, 2021.

88. W. Chakraborty, M. S. Jose, J. Gomez, A. Saha, K. A. Aabrar, P. Fay, S. Gupta, and S. Datta, "Higher-k zirconium doped hafnium oxide (hzo) trigate transistors with higher dc and rf performance and improved reliability," in *2021 Symposium on VLSI Technology*, pp. 1–2, 2021.

This document was prepared & typeset with pdfLaTeX*, and formatted with* NDdiss2$_\varepsilon$ *classfile (v3.2017.2[2017/05/09])*