Compact Device Technologies for Compact Integrated Systems

Lars Tatum
Tsu-Jae King Liu, Ed.

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-35
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-35.html

May 1, 2025

Copyright © 2025, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Compact Device Technologies for Compact Integrated Systems

By

Lars Prospero Tatum

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Tsu-Jae King Liu, Chair

Professor Ali Javey

Professor Junqiao Wu

Summer 2024

Compact Device Technologies for Compact Integrated Systems

Copyright ©2024

by

Lars Prospero Tatum

Abstract

Compact Device Technologies for Compact Integrated Systems

by

Lars Prospero Tatum

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Science

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

The rapid technological advancement over the past century, is a manifestation of Kurzweil's Law of Accelerating Returns. This theory, proposed by Ray Kurzweil, posits that technological progress accelerates exponentially, with each innovation spurring further advancements. This exponential growth has profound implications across various domains, driving innovation and fueling economic growth. A notable outcome of this progress is Generative AI, which is transforming communication, work, and learning, further accelerating technological advancement.

Since the 1970s, solid-state integrated circuit (IC) technology has been crucial to this exponential growth. Moore's Law, which observes the doubling of transistors on a chip approximately every two years, has enabled continuous enhancements in computational power and cost efficiency. However, around 2005, Dennard Scaling, a method for improving IC performance by scaling down MOSFETs, became impractical due to physical constraints. The future of IC technology faces significant challenges in maintaining improvements in Performance, Power, Area, and Cost (PPAC). Advancements in each of these aspects are increasingly interdependent, with improvements in one area often resulting in trade-offs in another. For instance, enhancing standby power consumption is constrained by the Boltzmann limit of the subthreshold slope, while performance is limited by parasitic resistances and capacitances. The semiconductor industry's continuous improvement is vital to sustaining technological progress, as predicted by Kurzweil's Law. Maintaining this trajectory is crucial to avoiding stagnation in technological capabilities. The dissertation aims to explore novel approaches to advancing CMOS technology platforms without significant trade-offs in PPAC, leveraging "Device and Circuit Cleverness" as noted by Gordon Moore.

A vertically oriented, nonvolatile Back End of Line Nanoelectromechanical Switch (NV BEOL NEMS) is introduced as an emerging nonvolatile memory device using compact-

finite-element-method simulations. BEOL NEMS can be integrated with CMOS processes, requiring only air-gap technology. A new differential half-select scheme enables dense NEMS arrays with low transistor overhead.

A CMOS-compatible, high PVCR negative differential resistance (NDR) device based on an optimized ferroelectric field-effect-transistor (FeFET) is proposed. TCAD studies explain its operation and optimization techniques to achieve peak currents over 400 μA/μm and PVCR over 10^6.

Finally, SRAM bit-cells based on the NDR FeFET are discussed and benchmarked against 6T CMOS FinFET SRAM using mixed-mode TCAD simulations. NDR FeFET SRAM offers significantly lower standby power, requires fewer devices, and leverages its unique hysteresis to reduce V_{DD}, achieve low retention V_{min}, and enable nonvolatile operation.

To my family, and particularly my late grandfather, for their inspiration, love, and support.

Contents

List of Figures

List of Tables

Acknowledgements

子曰。學而時習之、不亦說乎。 有朋自遠方來、不亦樂乎。人不知而不慍、不亦君子乎。

This opening line from *The Analects* captures my five years of graduate school quite well. The support of friends, and the camaraderie I built with the fellow students in Cory Hall, and subsequently Sutardja Dai Hall, have allowed me to stay sane, while diving deep into the challenging topics I took on for my thesis.

The greatest acknowledgment of this thesis, unsurprisingly, goes to Tsu-Jae. She has consistently gone beyond her academic duties as an advisor. Beyond daily technical challenges, she has served as a guide for how to be a leader—not only in leading others, but in leading your own life. Always willing to entertain my hare-brained ideas and push forward, her persistence has been paramount. She is also rare in that she cares deeply not just about pursuing technical work, but about the community and society at large. She is one of the few people I've met who truly lives like Christ, and she is an inspiration to us all. I am excited about the future work we will pursue and the collaborations for years to come.

A hearty note of gratitude is also extended to Professors Ali Javey and Junqiao Wu for serving on my dissertation committee, and additionally to Professor Bora Nikolic for serving on my qualifying committee. Bora provided key pointers to help me think "beyond the device" and consider whether device innovations could make a system-level impact, encapsulated in his advice: "read the news more." Thanks to Ali for your feedback and encouragement during my first year at Berkeley, especially after my first prelim experience.

It was also a fun and enriching experience to serve as a GSI for EE143 and EE130, the two most important semiconductor device courses offered when I was a graduate student. I wouldn't have developed my skills to the extent they are without having to explain these fundamental concepts to junior students. Thanks to Jeff Bokor and Sayeef Salahuddin for these opportunities!

Although none of the work made it into this thesis, I had a productive time working in the Marvell Nanofabrication Laboratory. The following staff were particularly helpful in assisting me with successfully fabricating various structures: Sam, Allison, Greg, Daniel, Richelieu, Ryan, Tariq, and Bill.

I also benefited from mentorship from industry members, which helped me build a good historical understanding and some foresight regarding the Semiconductor Industry. Paramount among these mentors have been Tahir Ghani, Sandy Liao, Charles Chu, Jamil Kawa, Hideki Takeuchi, Daniel Connelly, Hei Kam, and Reinaldo Vega. Hiu Yung Wong (now at SJSU) provided great troubleshooting advice for the TCAD work in this dissertation.

I have seen multiple groups of students come and go over the years. From the original King group students, who greeted me with "Wow, Tsu-Jae hasn't mentioned taking new students!", I am so thankful for Thomas, Fei, Robin, Tsegereda, Alice, Urmita, Yi-Ting, Ben, and my Bro, Xiaoer, for helping me settle into research life. Urmita, you have been a wonderful friend and coworker since the beginning, and I am very happy to have gone through Berkeley with you the whole way! To the other 373 gang with whom I have spent countless hours working and having candid discussions, particularly Ricky, Kevin, Qiutong, Xintian, Eric, Mutasem, Hanbin, and Qianyi, I am grateful. A special shoutout to the BSACers who took me in as a BSAC imposter: Daniel T., Daniel K., Lydia, and Jasmine; I definitely wouldn't have been able to get all those free lunches without you.

When the group moved to Sutardja Dai Hall, it was a blessing to get to know the residents of the 5th floor, particularly my friends Jason, Saavan, Adi, Pratik, Chirag, Nirmaan, Yejin, Chia-Chun, Junhao, Chien-Ting, Jianheng, Meshal, Li-Chen, Ava, Joe, Sean, Zeying, Ross, Min, and Tracy. One day I hope to repay Steve for my extensive bar tab. The recently joined junior students of the King group, Collin and Dasom, have been a pleasant surprise that I have treasured during my later years of graduate school. I especially look forward to seeing Nephew Roy grow up into a future Berkeley student! I also had the wonderful opportunity to mentor undergraduate students during my time here. It was a blast working with Isaac, Nhi, and Xiangwei, and seeing them become confident researchers. I am confident their futures are bright.

The BWRC folks have also been a significant part of my grad school experience: Erik, Avi, Cem, Daniel G., Chun-Yen, Yu-Chi, Daniel K., Sarika, Ryan, and Arya. Thanks for all the Monkey Heads and fun times!

Outside of my daily EECS gang, there are a few other professors and students who made grad school special. Naeem Zafar's course and mentorship on "Hard Tech" startups significantly shaped how I approach R&D problems and evaluate trends in the semiconductor space. I also greatly appreciate Lin Laoshi's introductory Mandarin language course and ongoing friendship; I wouldn't have been able to survive during my time in Taiwan without it! I also appreciate the friendship and Taiwan advice from Clara. Good luck in your future grad school endeavors!

Outside of Berkeley, I was lucky to have a small number of Florida friends in the Bay Area who consistently made good memories. Thanks to Cross, Herman, and Anthony for all the fun times, whether ripping on Cañada, raving at DNA Lounge, or impulsively driving across the country for a solar eclipse. Thanks to Mark and Bobby for keeping in touch from afar!

In 2023, I had the opportunity to go to Taiwan for an internship at the Semiconductor Disney World, TSMC. Going to Taiwan with limited mandarin skills might have gone quite disastrously had it not been for the friends I made along the way. Fortunately, the Taiwanese and other interns I met were incredibly welcoming. A special note of appreciation to the fellow interns, Tingyu, Vincent, Chongyock, Tanyu, Rutu, Sudhanshu, and Deborah, for the exceptional motivation to go for dinner and travel around Taiwan at every opportunity. I also made some great friends among the local Taiwanese at TSMC. I am particularly thankful for Yen-Ying, who acted as a Big Brother for me in Taiwan; he made sure that I was already a "Popular Guy" by the time I first set foot in TSMC! To my colleagues YC, KC, FC, Mickey, Kuen-Yi, Tzuyu, Yi-Hsuan, Yu-Shan, Hung-Ju, Wan-Chen, Yi-Han, Jeanie, and Allen: thank you for accepting me as one of your own and helping me do great work during my short time there. I was very lucky to meet Margaret while an intern. Margaret, your tenacity and drive inspire me immensely. It has been great to learn about semiconductor policy and strategy from you; I hope you become president (or CEO) one day!

None of this would have been possible without my care team at Sutter Health. Shortly before starting grad school, I was diagnosed with epilepsy, a chronic neurological disorder. This posed some heavy existential questions to me. I got very lucky to be paired with Dr. Mariel Velez, whose guidance and encouragement helped me to confidently push forward. My ongoing care team, Drs. Lewis Leng and Katherine Werbaneth, have done a great job of keeping my brain functioning. Thanks!

Finally, my deepest gratitude goes to my family. Without your emotional support and getting me on the right path raising me, I surely would not have made it to this point. Thank you!

Chapter 1: Sustaining the AI Revolution

The technological capabilities of humanity have advanced at an unprecedented pace over the past century, now approaching the potential to emulate human behavior [1.1]. This remarkable trend has been documented by Ray Kurzweil and is known as Kurzweil's Law of Accelerating Returns [1.2]. Kurzweil's Law posits that the rate of technological progress accelerates exponentially, with each innovation acting as a catalyst for further advancements. According to this theory, technological growth is not linear but instead exponential in speed and impact over time. This principle highlights the transformative potential of technological progress across various domains. Continuous technological evolution drives innovation, revolutionizes industries, and fuels economic growth. It enables the development of sophisticated solutions to complex problems and opens new possibilities in fields ranging from healthcare to communication and beyond. The most recent field to blossom out of technological evolution is Generative AI, which is actively changing the way we work, communicate, and learn, further fueling the exponential acceleration of technological advancement. Fig. 1.1 below illustrates Kurzweil's Law by plotting "Calculations per second per dollar", a proxy for "Technological Progress", since 1900.

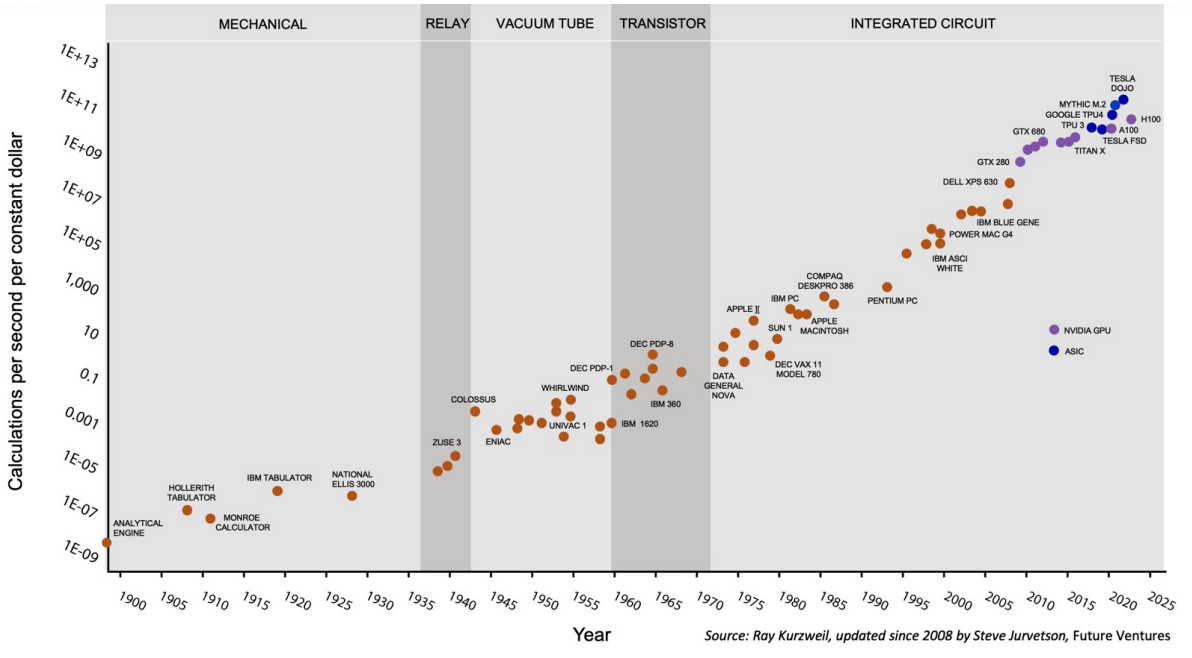


Figure 1.1: Tabulation of Kurzweil's Law since 1900. Adapted from [1.3].

Since the 1970s, solid-state integrated circuit (IC) technology has been the cornerstone of Kurzweil's Law. The exponentially increasing number of transistors integrated onto a chip, known as "Moore's Law" [1.4] (Fig. 1.2), has continuously enhanced the functionality and reduced the cost of computation. These enhancements lead to novel technological platforms that proliferate throughout society, expanding our capabilities and facilitating the design of the next generation of integrated circuit technology, thus perpetuating the exponential growth cycle. Consequently, these fundamental technology improvements drive advancements across the entire technology pipeline, enabling new software paradigms (e.g., machine learning, data analysis) and end-user applications (e.g., ChatGPT, TikTok). Therefore, it is essential to continue advancing our computational capabilities at the foundational level.

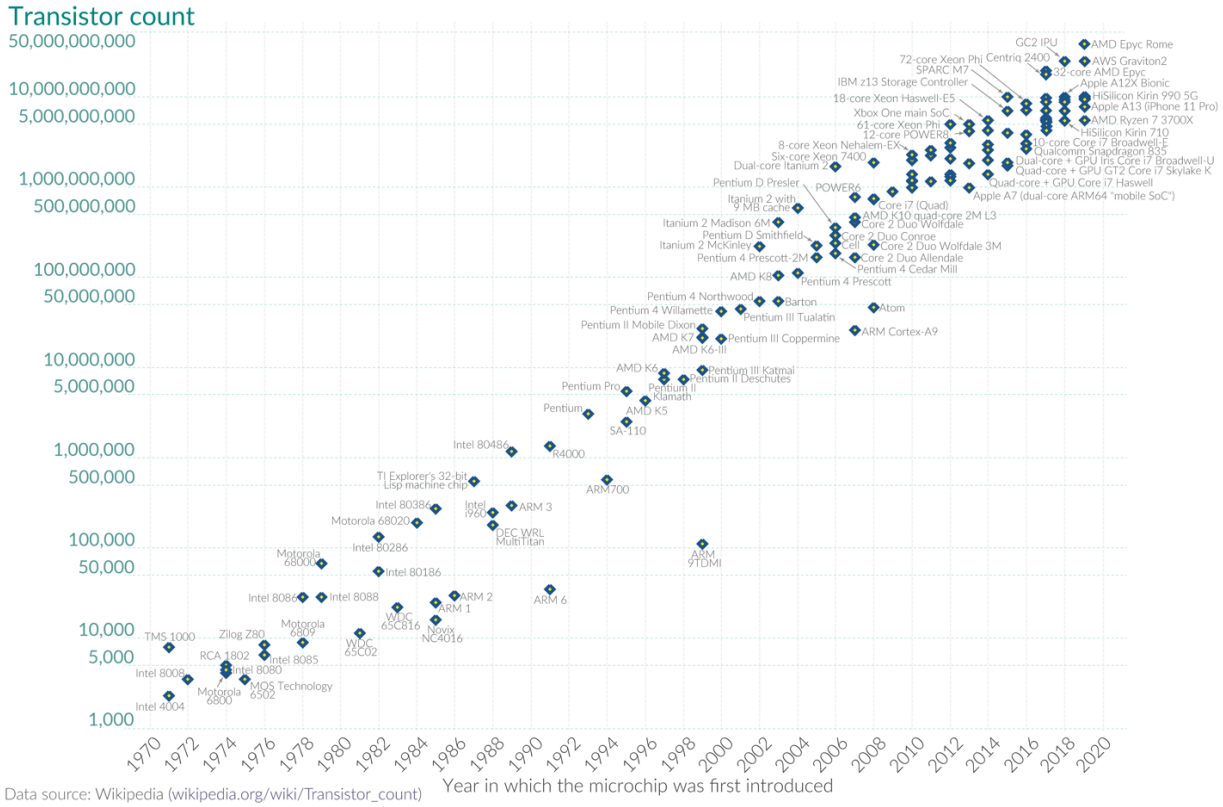


Figure 1.2: The number of transistors per (packaged) microchip has increased exponentially since 1970, following the prediction of Gordon Moore.

Until approximately 2005, the roadmap for CMOS technology advancement, the dominant IC platform based on Metal Oxide Semiconductor Field Effect Transistors (MOSFETs), could follow a formulaic approach known as "Dennard Scaling" [1.5]. The

constant-field scaling rule, proposed by Robert Dennard in 1974, provides a method for scaling MOSFETs which keeps the peak electric field roughly constant by scaling certain design parameters by a constant factor α. By adhering to Dennard Scaling rules, each new node could bring steady and predictable improvements in IC technology: circuit speed could be increased by α while transistor area decreased by $1/\alpha^2$. To enable and go beyond Dennard Scaling, new process innovations were adopted to further accelerate the pace of IC technology advancement; empirically, each time the number of transistors on a chip increased by a factor of 10.

Around 2005, Dennard Scaling became impractical as non-idealities became increasingly difficult to control. For instance, Dennard's scaling law mandates scaling the body doping concentration by α. However, higher doping concentrations reduce carrier mobility, limiting the device on-state current, and also lead to significant device-to-device variation [1.6], making it difficult for circuit designers to continuously improve the functionality and performance of state-of-the-art integrated circuits. Additionally, short-channel effects (SCE) led to increased leakage currents, limiting tradeoffs between high-performance and low-power device technology. Therefore, semiconductor engineers have become increasingly creative to continue improving the Performance, Power, Area, and Cost (PPAC) of integrated circuit technology. Beyond lithographic pitch scaling, new materials, processes, and structures have been successfully research, developed, and ramped to high volume manufacturing, in search of exponential technology advancement that lasts “forever” [1.7], [1.8]. Fig. 1.3 illustrates some of the key innovations deployed to facilitate continued advancements in CMOS logic technology in the past ~25 years.

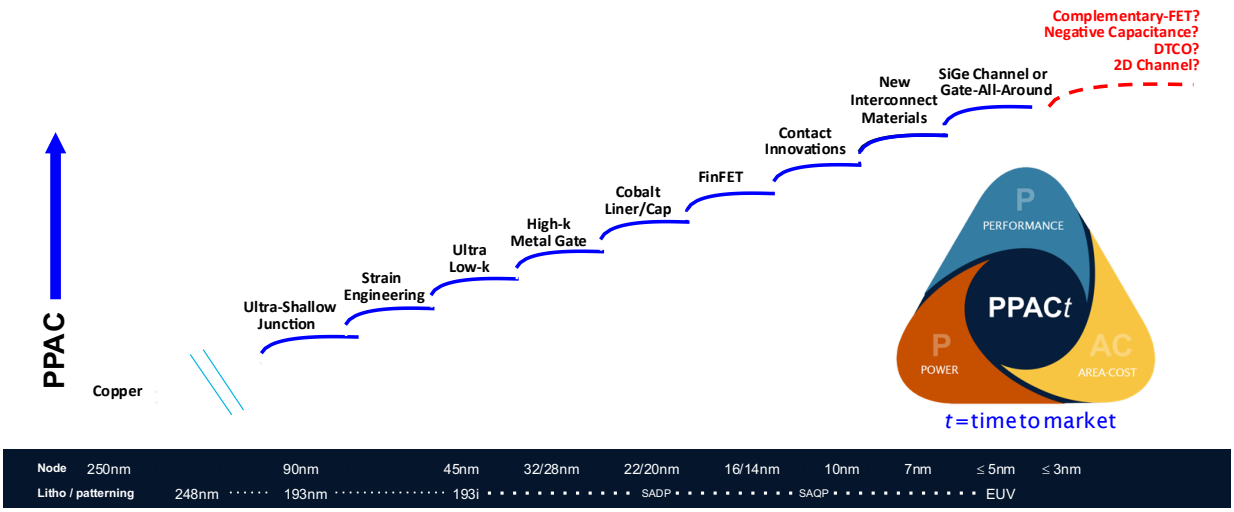


Figure 1.3: A showcase of some of the innovations that have continued CMOS logic technology advancement. Adapted from [1.9].

In the near future, improving or even maintaining each individual component of PPAC for new technology nodes will become increasingly challenging, as these aspects are interrelated. Improvements in standby leakage power consumption are constrained by the electrostatic integrity of the chosen transistor structure and ultimately by the Boltzmann limit of the subthreshold slope [1.10]. Conversely, performance is limited by worsening parasitic resistance and capacitance components, which further increase active power consumption. Area scaling has slowed significantly due to increasing sensitivity to process-induced variations at nanoscale geometries. Additionally, the cost per wafer is substantially increasing due to the increasingly complex manufacturing processes required to achieve these PPA improvements. Unfortunately, most of the proposed future innovations on the roadmap can only provide for tradeoffs between the components of PPAC. Complementary-FET technology [1.11] stacks N-type nanosheet transistors directly over P-type nanosheet transistors, reducing area footprint significantly, at the cost of substantially increased manufacturing cost. 2D-material channel transistors [1.12] can provide enhanced electrostatic integrity at reduced gate lengths to reduce area footprint, but may come at the cost of increased parasitic resistances, reduced on-state current, and increased manufacturing cost.

As a result of these trends, the historical reduction in cost per transistor, the original driver of Moore's Law, may be coming to an end (Fig. 1.4). If the cost per transistor

does begin increasing, the economic value added by each new process node (e.g., from performance enhancements, or reduced power consumption) must substantially outpace the increased cost to justify technology adoption by fabless semiconductor system companies. In other words, *someone* has to be willing to pay for it!

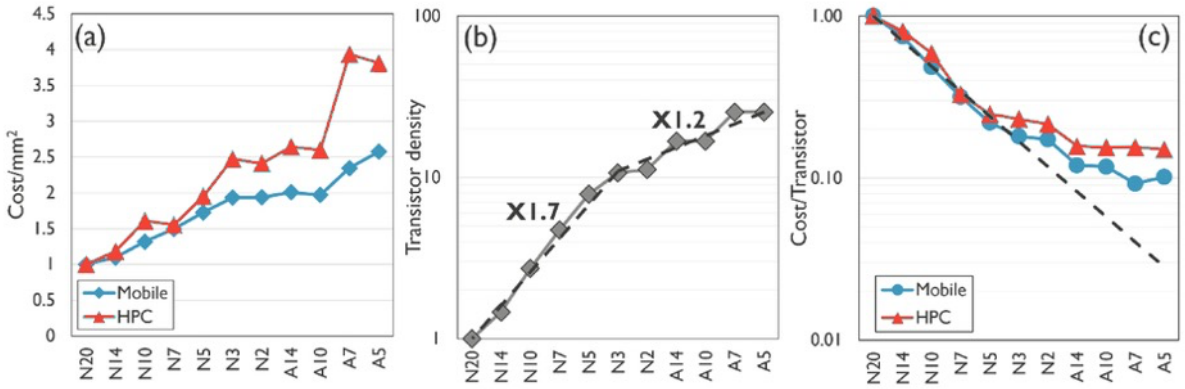


Figure 1.4: (a) Rising Die cost per unit area combined with (b) the slowing of transistor density scaling may lead to (c) a stagnating Cost/Transistor improvement trend, which threatens the current economic model of the semiconductor industry. Adapted from [1.13].

Since semiconductor R&D is primarily supported by the high revenues from the most advanced "state-of-the-art" technology nodes, facilitating the adoption of advanced nodes by continuously improving customer value is essential to continue pushing the semiconductor industry, and by extension, humanity's technology capabilities, forward. Maintaining this continuous improvement is existentially necessary— if it halts, this could have devastating long-term effects on our general technological capabilities, as predicted by Kurzweil's Law. Although exciting advances in computer architecture [1.14] have offered unprecedented boosts in performance or energy-efficiency for specific applications, the loss of an exponentially improving technology base would almost certainly be felt.

Therefore, the goal of this dissertation is to identify approaches to advancing CMOS technology platforms which do not incur a significant tradeoff between the components of PPAC, by broadening the scope outside of the contemporary CMOS device and circuit toolset. This is in the spirit of the observations made by Gordon Moore, who noted in 1975 [1.15] (Fig. 1.5) that "Device and Circuit Cleverness" contributed a significant factor to the increase in components per chip.

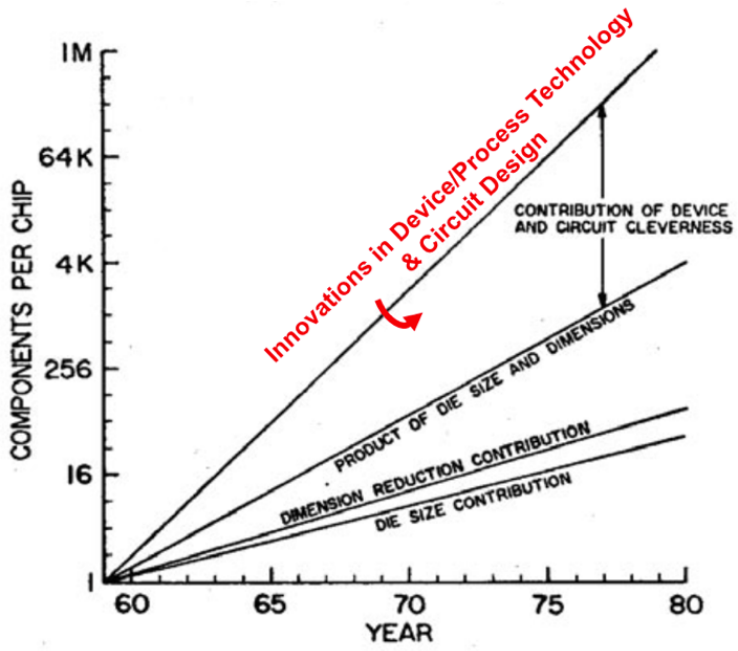


Figure 1.5: Moore attributed a substantial increase in components per chip to "Device and Circuit Cleverness". Adapted from [1.15].

In Chapter 2, vertically oriented, nonvolatile Back End of Line Nanoelectromechanical Switches (NV BEOL NEMS) are described and benchmarked as an emerging nonvolatile memory device via compact-finite-element-method simulations. BEOL NEMS technology can be readily integrated with a typical CMOS process, and at the fundamental level only requires the introduction of air-gap technology to a standard BEOL process. Simulations predict that a 5 nm node BEOL NV NEM switch memory cell can be programmed within 40 ns and less than 35 aJ of energy. A differential half-select scheme for programming single-pole, double-throw NV BEOL NEMS is introduced for the first time to enable the implementation of dense NEMS arrays with low transistor-peripheral overhead.

In Chapter 3, a CMOS process compatible, high peak-to-valley current ratio (PVCR) negative differential resistance (NDR) device technology is proposed, based on an optimized ferroelectric field-effect-transistor (FeFET). The NDR FeFET's method of operation is elucidated through Technology-Computer-Aided-Design (TCAD) studies, and design optimization techniques are described to obtain an NDR device with the highest peak current (> 400 μA/μm) and PVCR (> 10^6) ever proposed in a semiconducting material system.

In Chapter 4, static random access memory (SRAM) based on the NDR FeFET is described and benchmarked with 6T CMOS FinFET SRAM using mixed-mode TCAD simulations. Since SRAM scaling has nearly halted due to increased sensitivity to process-induced variations, new approaches to achieve compact SRAM cells are highly desirable, to meet the growing data processing demands of artificial intelligence. NDR FeFET-based SRAM is shown to have up to an order of magnitude lower static leakage current compared to conventional CMOS SRAM, and requires half of the device count. Compared to other NDR device approaches, the unique hysteresis characteristic of the NDR FeFET can be leveraged to reduce V_{DD} to less than 0.45 V, achieve a retention V_{min} of 0.125 V, and in certain design cases, nonvolatile operation.

In Chapter 5, this dissertation is summarized, and suggestions for future work are given.

Chapter 2: Design Technology Co-Optimization for Back-End-of-Line Non-Volatile NEM Switch Arrays

In recent years, nano-electro-mechanical (NEM) switches have attracted interest for implementing ultra-low-power digital logic and non-volatile (NV) memory integrated circuits (ICs) [2.1]–[2.8]. This is because NEM switches can achieve zero off-state leakage current, abrupt on/off switching characteristics enabling ultra-low-voltage operation, non-volatility, and relatively low on-state resistance. Furthermore, monolithic three-dimensional (3-D) integration of NEM switches with CMOS transistors can enable compact computing architectures for improved energy efficiency [2.9],[2.10], which is especially important for wireless devices and Internet of Things (IoT) applications. To minimize incremental manufacturing cost, the metallic layers in a conventional CMOS back-end-of-line (BEOL) process can be leveraged to implement NEM switches [2.11]–[2.16]. Recently, a compact, reprogrammable look-up table (LUT) IC comprising gated CMOS buffers and an array of vertically oriented NV-NEM switches implemented using multiple BEOL layers was successfully demonstrated using a conventional 65 nm CMOS process [2.17], and new compute-in-memory (CIM) architectures have been introduced to increase the energy efficiency of machine learning systems, in which NV-NEM switches may serve as a key element. In this chapter, design tradeoffs for vertically oriented non-volatile nano-electro-mechanical switches implemented using multiple interconnect layers in a 5 nm-generation CMOS back-end-of-line (BEOL) process are investigated via three-dimensional device simulation. Programming pulse voltage and width operating windows are identified for avoiding catastrophic pull-in. Simulation results indicate that sub-20 ns programming delay is possible with programming voltages compatible with standard input/output (I/O) CMOS circuitry, and that the write energy of a NV-NEM bit-cell will be less than 5 aJ. A crossbar array architecture operated with a half-select row/column bit-cell programming scheme is found to be effective for avoiding the issue of write disturbance.

2.1 NV-NEM Switch Structure

The NV-NEM switches studied in this work are of the single-pole/double-throw design comprising five terminals as illustrated in Fig. 2.1: a movable beam electrode anchored at its base, two fixed actuation electrodes on either side of the movable beam, and two corresponding fixed contact electrodes used to conduct current in their contacting state.

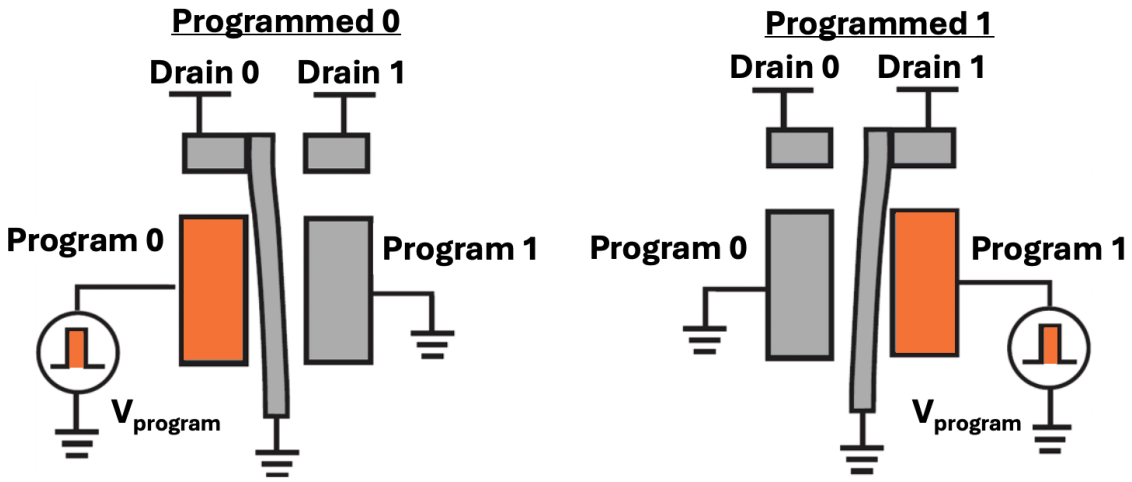


Figure 2.1: Schematic illustrating programming operation of a vertically oriented NV-NEM switch to State "0" and to State "1".

To program the NV-NEM switch into the "0" state, a programming voltage (V_{Prog}) pulse is applied between the "Prog0" (Program 0) actuation electrode and the grounded beam to induce attractive electrostatic force (F_{elec}) between these two electrodes and thereby actuate the beam into contact with the contact electrode D0. To program the switch into the "1" state, a programming voltage pulse is applied to the "Prog1" (Program 1) actuation electrode instead, to actuate the beam into contact with electrode D1. The programming voltage must be at least equal to the beam's pull-in voltage (V_{PI}), and can be larger to decrease the switching time. If this voltage is too large and/or the pulse width is too long, however, the beam will be pulled into contact with the actuation electrode; this is an undesirable situation referred to as catastrophic pull-in (CPI) [2.13],[2.14],[2.18].

In practice, vertically oriented NV-NEM switches are implemented using multiple BEOL metallic layers with the tightest possible pitch, which provides for the smallest

possible footprint and actuation gap (which in turn provides for the smallest possible programming voltage) as illustrated by the cross-sectional schematics in Fig. 2.2. The lower layers are used to implement the actuation (program) electrodes whereas the top layer is used to implement the contact (data) electrodes; the beam electrode is formed from all of these layers with vias in-between. After completion of the conventional CMOS fabrication process, an etch process is used to remove the inter-layer dielectric material surrounding the beam electrode, to "release" it for physical movement [2.17]. Note that the upper BEOL layers are patterned to serve as a mask for this etch process, so that no additional lithography step is needed.

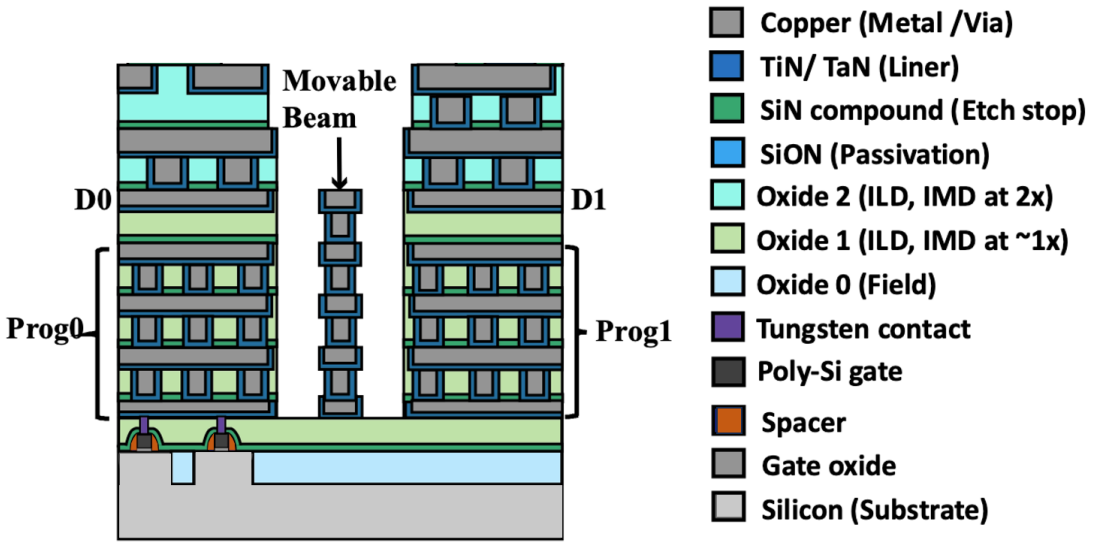


Figure 2.2: Schematic cross-sections illustrating the BEOL layers in a conventional CMOS process after the release etch process.

Fig. 2.3 shows scanning electron micrographs of a BEOL NV- NEM switch fabricated using a standard 65-nm CMOS process. As can be seen from Fig. 2.3a, this switch is formed from the five lowermost metal interconnect layers and intermediary via layers. The dielectric material surrounding the beam was selectively removed using a fluorine-based plasma etch process to allow its physical movement. Fig. 2.3a shows the BEOL NV-NEM switch as fabricated, in a neutral (non-contacting) state. After applying a voltage pulse between the Prog0 electrode and beam, the beam is actuated into physical contact with the D0 electrode, as shown in Fig. 2.3b.

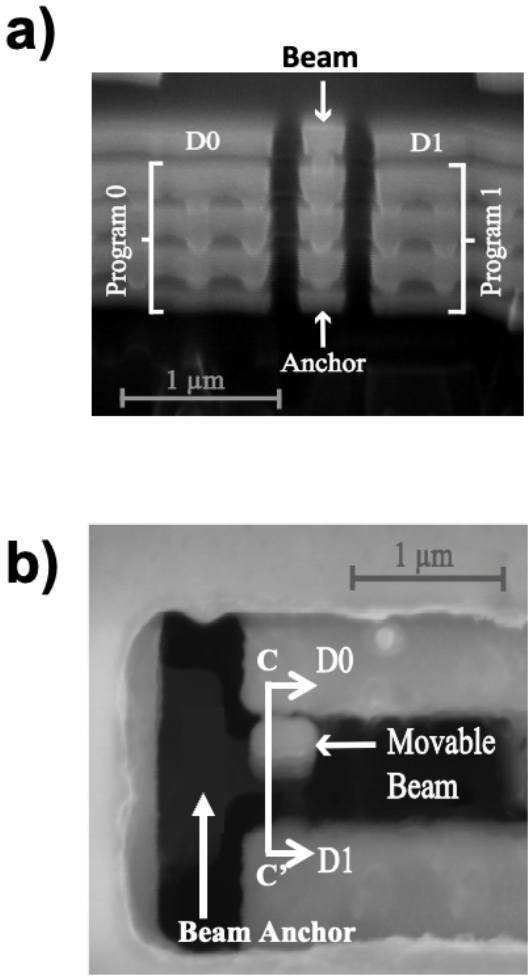


Figure 2.3: Scanning Electron Microscope images of fabricated vertically oriented NV-NEM switch: (a) cross-sectional view along cutline C-C' and (b) plan view after being programmed into state 0. Adapted from [2.17]

2.2 NEM Switch Design Optimization

In this chapter, Coventor MEMS+ finite-element method simulation software [2.19] is used to study the behavior of NV-NEM switches and explore design trade-offs. Previous works [2.17],[2.18],[2.20] have shown good agreement between NV-NEM physical models, MEMS+ simulations, and experimental results. Fig. 2.4 shows a simulated three-metal-layer (3ML) NV-NEM structure labeled to indicate the critical geometrical design parameters, the contact length and the actuator length. For a given contact metal layer thickness (defined by the process technology), the contact length determines the contact area and hence the contact adhesive force (F_{adh}) and on-state resistance. F_{adh} must be greater than the mechanical spring restoring force (F_{spring}) of the deformed beam in the contacting state, in order for the switch to be non-volatile; at the same time, F_{adh} should not be much higher than needed to achieve non-volatility, to avoid an unnecessarily large programming voltage to switch states. (F_{elec} + F_{spring} must be larger than F_{adh} to pull the beam out of contact.)

For given actuation metal layer thicknesses (defined by the process technology), the actuator length determines the effective actuation electrode area. Longer actuator length provides for greater actuation area, which is desirable for reducing the actuation (*i.e.*, programming) voltage. However, this comes at the trade-off of larger switch footprint, which can either increase the manufacturing cost or limit the size (number of rows and columns) of a NV-NEM array. 5-nm CMOS technology node process parameters were assumed, as shown in Table I.

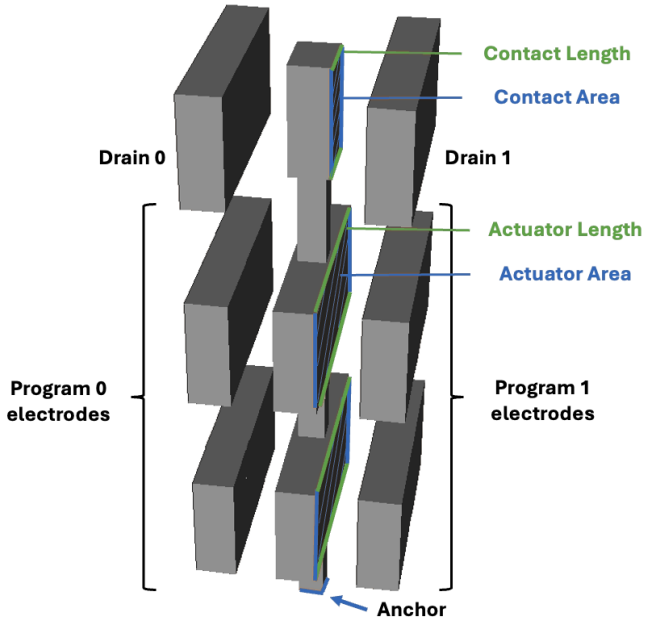


Figure 2.4: Simulated three-layer NV-NEM switch structure. The stacked program electrodes are assumed to be electrically connected (cf. Fig. 2.3(a)).

Table 2.1: 5-nm node BEOL metal layer specifications

Parameter	Specification
Minimum metal pitch	30 nm
Minimum metal width	15 nm
Minimum via width	10 nm
Metal thickness (height)	40 nm
Via thickness (height)	35 nm
Metallic material and Young's Modulus	Copper 128 GPa

Five-metal-layer (5ML) and four-metal-layer (4ML) NV-NEM switch designs were investigated and found to be unviable for a 5-nm process technology. This is because the mechanical stiffness of the movable beam decreases with its length, making it more susceptible to CPI; at the same time, lower beam stiffness results in lower F_{spring} so that larger F_{elec} (*i.e.*, larger V_{Prog}) is needed to switch state – which increases the likelihood of CPI. In short, CPI was found to occur for a smaller programming voltage or pulse width than the minimum programming voltage and pulse width required to switch states, for the 5ML and 4ML NV-NEM switch designs. Fig. 2.5 shows how a 5ML NV-NEM switch is catastrophically pulled in to its Prog 1 electrode. Therefore, the remainder of this chapter focuses on the design and programming of 3ML NV-NEM switches.

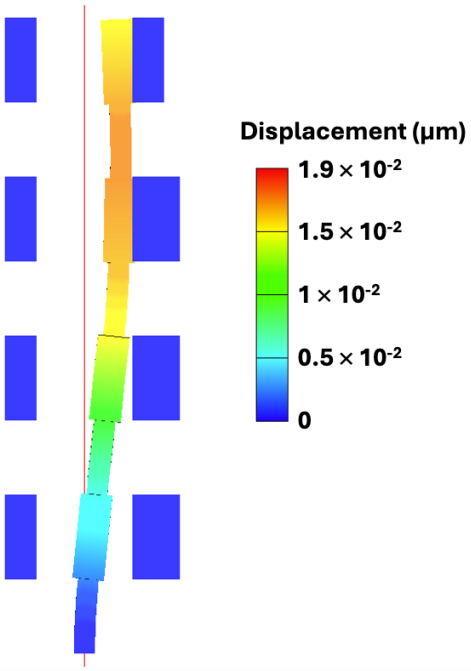


Figure 2.5: Simulated 4ML BEOL NV-NEM switch showing catastrophic pull-in after a voltage pulse was applied to the Prog1 electrode to switch from State 0 to State 1.

Contact Length Optimization

The contact length is minimized within the constraint of ensuring non-volatile operation ($F_{adh} > F_{spring}$), to provide for the lowest possible programming voltage and/or pulse width. For 3ML switches, the minimum required contact length was found to be 19 nm,

assuming a contact adhesive pressure of 130kN/cm^2 [2.18]. This contact length was assumed for the remainder of this study.

Effect of Actuator length

The stiffness of the beam electrode is largely determined by its via portions, *i.e.*, the beam's stiffness is not significantly dependent on the length of its actuation portions. Fig. 2.6 shows how the minimum programming voltage needed to switch states decreases with increasing actuator length. It should be noted that the NV-NEM switch programming voltage range is within that which is typical for input/output (I/O) CMOS circuitry; also, it is notable that the footprint of a NV-NEM bit-cell (replicated across a two-dimensional array of bit-cells arranged into rows and columns, as described below) is significantly smaller than that of a conventional static memory (SRAM) bit-cell which comprises six transistors within a layout area > 125F^2, where F is the minimum half-pitch. Below we consider in more detail NV-NEM switch designs with actuation electrode lengths of 90 nm, 140 nm, 190 nm and 240 nm, which all can be programmed with sub-2.5 V voltage pulses.

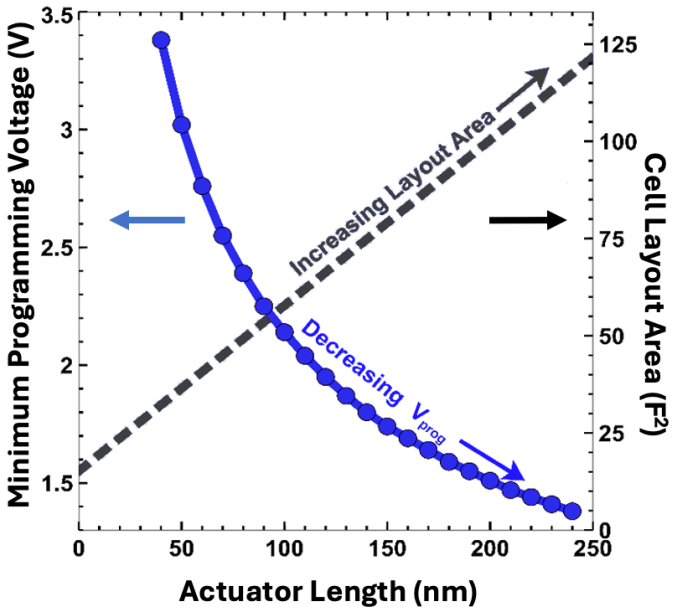


Figure 2.6: Minimum programming voltage and bit-cell layout area vs. actuator length, for a 3ML NV-NEM switch.

Programming Voltage Pulse Optimization

For a given programming voltage, there is a minimum programming voltage pulse width required to successfully change the state of the NV-NEM switch, and there is a

maximum programming voltage pulse width required to avoid the undesirable phenomenon of CPI. Fig. 2.7 plots curves for minimum and maximum pulse widths as functions of programming voltage, for 3ML NV-NEM switch designs of different actuation electrode lengths.

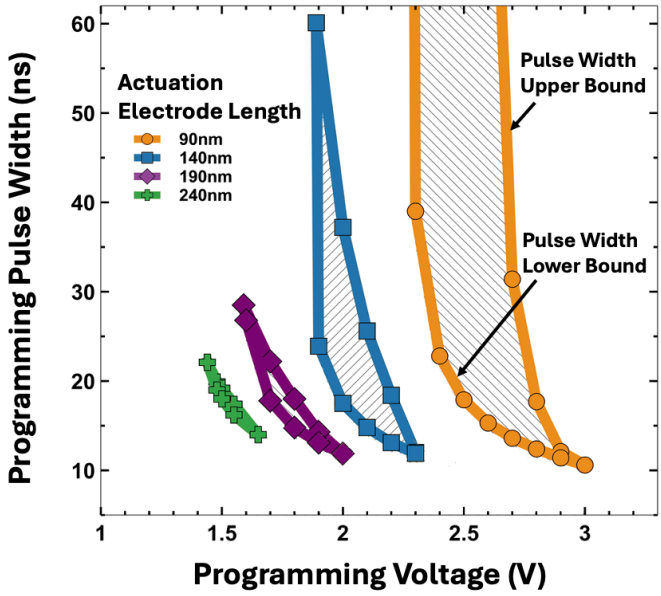


Figure 2.7: Minimum programming pulse width as a function of the programming voltage, for 3ML NV-NEM switches.

As expected, the programming pulse width decreases with increasing programming voltage; each of the 3ML NV-NEM switch designs can be operated with sub-20 ns programming pulse width. Interestingly, the operating voltage window shrinks with increasing actuator length, suggesting a maximum practical limit of 190 nm.

Programming Energy

Emerging NV memory devices have been proposed for IoT applications because they require far less energy to program than traditional floating-gate flash memory devices [2.21],[2.22]. Therefore, it is of interest to benchmark NV-NEM switches against other NV memory devices with regard to programming energy. Fig. 2.8 plots the minimum energy required to program 3ML NV-NEM switches *vs*. programming pulse width.

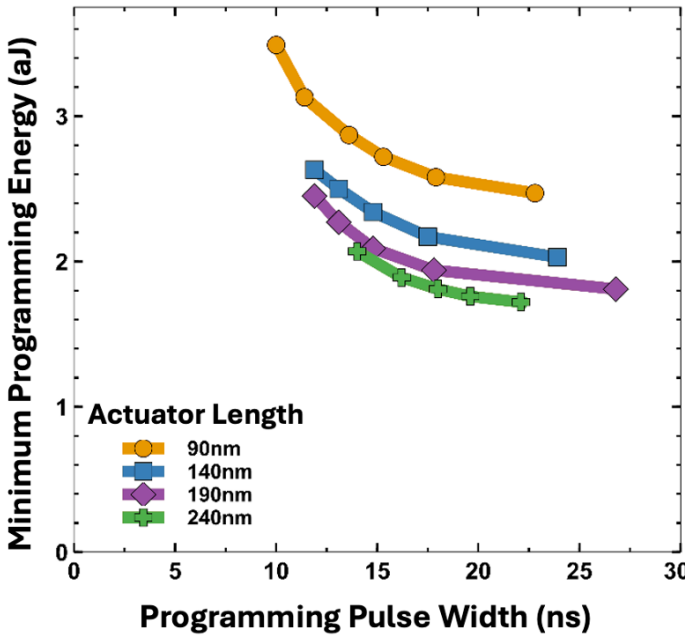


Figure 2.8: Minimum energy required to program 3ML NV-NEM switches, as a function of the programming pulse width.

It can be seen from Fig. 2.8 that NV-NEM switches are projected to have an intrinsic programming energy of less than 5 aJ, which is approximately two orders of magnitude lower than for RRAM and STT-RAM devices, and eight orders of magnitude lower than for NOR Flash devices, as shown in Table II. The extremely low write energy of a NV-NEM switch is a result of the fact that no direct current flows during the programming process; the program electrodes only need to be capacitively charged by displacement current, and the data (D0 and D1) electrodes can be either electrically floating or biased at ground potential (same as the beam potential) during the programming operation to ensure zero direct current flow. With a Ruthenium interconnect barrier material as the contact electrode material, the programming endurance of NV-NEM switches is expected to be $>10^{15}$ cycles [2.23], assuming the degradation of the contacting surface is the ultimate limiter for NV-NEM switch endurance, rather than plastic deformation of the switch.

The read time for a NV-NEM switch corresponds to the amount of time required to charge or discharge a data-line and in practice likely would be limited by the resistance of the beam-to-data electrode contact. The effective cell capacitance was found in simulation to be ~10aF. Conservatively assuming a contact resistance value of 250 kΩ (corresponding to the minimum contact area of ~0.00076 μm^2) [2.20] and a data-line length of 100 μm for 1024 switches (bringing the total capacitance to ~25 fF), the NV-NEM bit-cell read time for a LUT architecture [2.17] is less than 10 ns, which compares very favorably

against that of other emerging NV memory (NVM) devices. The minimum read voltage is limited by the array sense amplification circuitry. Assuming a read voltage of 100mV, a read energy of 250aJ per bit is possible.

Table 2.2: Benchmarking Emerging NVM Devices [2.21],[2.22]

Memory type	NOR Flash	PCRAM	RRAM	STT-MRAM	**NEMory (this work)**
Write Time	1000 ns	20-100 ns	<10 ns	<20 ns	**10-40 ns**
Read Time	10 ns	20-50 ns	<10 ns	<20 ns	**<10 ns**
Write Energy (per bit)	1nJ	>10 pJ	~fJ	~pJ	**15-35 aJ**
Read Energy (per bit)	~pJ	<<0.1 nJ	~ fJ	~pJ	**~250 aJ**
Lifetime	~10^5	~10^9	>10^{12}	~10^{15}	**>10^{15}**

The energy-delay product (EDP) is a metric that gauges the trade-off between performance and energy efficiency, and is most often used to benchmark digital logic technologies. It is worthwhile to note here that the EDP for each of the 3ML NV-NEM switch designs studied herein is less 10^{-24} J·s, which compares well against CMOS technology at >10^{-22} J·s [2.24].

2.3 NV-NEM Switch Array Programming Scheme

If each Prog0 electrode and Prog1 electrode of a NV-NEM switch bit-cell in a two-dimensional array of M rows and N columns were to be individually addressed, the number of program signal lines required would be 2×M×N. In [2.17] we introduced a novel crossbar array architecture that reduces the number of program signal (address) lines to M+N, together with a novel half-select programming scheme for setting the state of an individual bit-cell without disturbing the state of other bit-cells along the same row or along the same column, building upon prior works [2.6],[2.25],[2.26]. As illustrated in Fig. 2.9, the Prog0 electrodes for all bit-cells located in the same column are connected together as a single address line that runs continuously across the array, while the Prog1 electrodes for all bit-cells located in the same row are connected (through vias) to the same row address line that runs across the array and is formed in a higher metal layer. From a layout perspective, there must be sufficient spacing between each row address line and its adjacent beam electrodes, to allow the aforementioned release etch to remove the dielectric material surrounding the beam electrodes. (In other words, the row address lines should not overlap any beam electrodes, by some margin.) The data-lines run in the direction of the columns, on the same metal layer as the contact electrodes. The bit-lines, located on the M0 layer, run perpendicular to the data-lines, and anchor the switches. For maximum array density, the bit-cells are laid out in pairs sharing Prog1 electrodes.

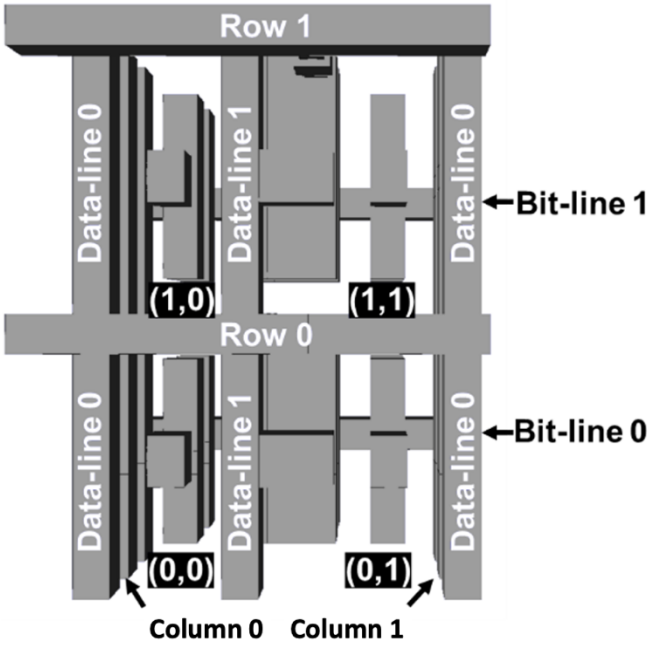


Figure 2.9: Layout view a 2×2 array of NV-NEM switch bit-cells.

As shown in Fig. 2.1, when an NV-NEM switch is being programmed into State 0, the Prog1 electrode is held at ground potential (same as the beam potential) so that there is no counteracting electrostatic force. The Prog1 electrode could be biased at a voltage greater than 0, however, to induce a counteracting electrostatic force that would prevent switching from State 1 to State 0. This "hold voltage" (V_{hold}) should not be too large so as to cause catastrophic pull-in of the beam to the Prog1 electrode; neither should it be large enough to cause switching from State 0 to State 1.

To "write" the state of a bit-cell in an array, a programming voltage pulse is applied to the appropriate actuation electrode while the opposite actuation electrode is held at ground potential. To avoid disturbing the states of the bit-cells located in the same row (with the same Prog1 electrode potential as the bit-cell being programmed), all of their Prog0 electrodes should be biased at V_{hold}. Likewise, to avoid disturbing the states of the bit-cells located in the same column (with the same Prog0 electrode potential as the bit-cell being programmed), all of their Prog1 electrodes should be biased at V_{hold}.

In practice, each of the row and address lines for the NV-NEM bit-cell array should be biased at V_{hold} except when they are needed to address the bit-cell being programmed, as illustrated in Fig. 2.10: to program a bit-cell, the voltage on one of its address lines is increased from V_{hold} to V_{prog} while the voltage on its other address line is decreased from V_{hold} to ground potential. This programming scheme is analogous to the half-select programming scheme used for crossbar RRAM arrays [2.27]. Since no direct current flows during the programming operation (as explained above), there is no half-select leakage current for NV-NEM bit-cell arrays. Also, in principle, multiple bit-cells sharing an address line could be programmed simultaneously into the same state.

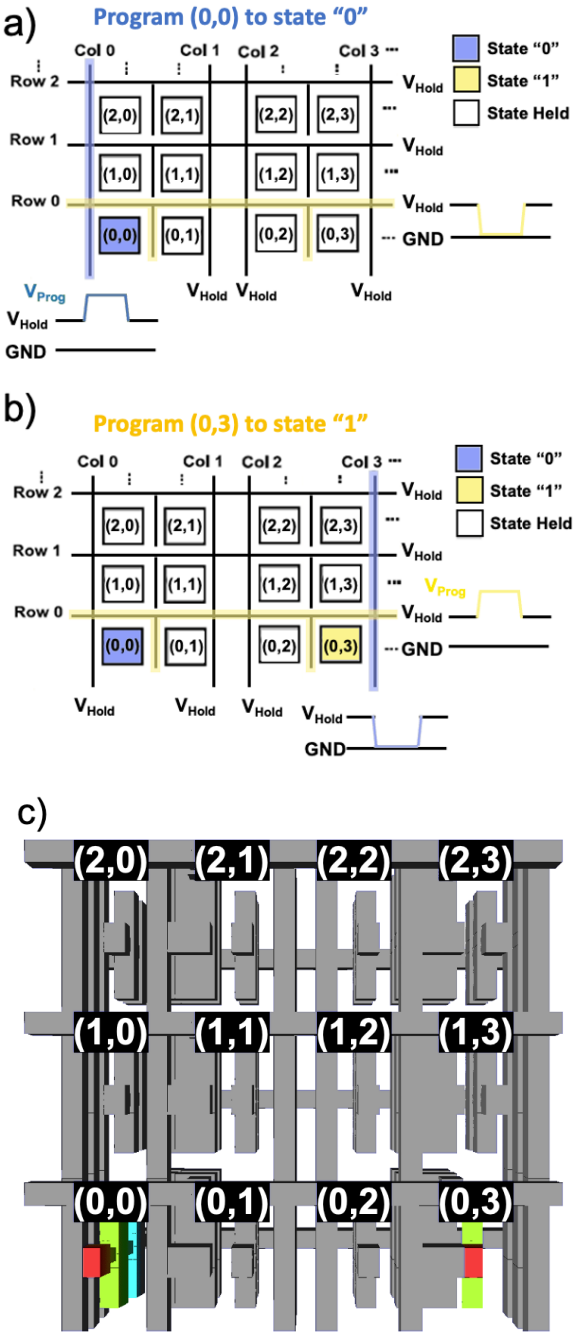


Figure 2.10: Circuit schematics showing the voltage pulses applied to the row and column address lines to program bit-cell a) (0,0) to State "0" and b) (0,3) to State "1". c) Layout view of the corresponding 3×4 array of NV-NEM bit-cells.

If the programming voltage is increased to lower the programming delay, the hold voltage also must be increased to avoid write disturb issues, as shown in Fig. 2.11. The hold voltage is very sensitive to the programming voltage since the beam is drawn physically

closer to the programming electrode than the holding electrode during programming. Similarly as the minimum programming voltage, the minimum hold voltage decreases with increasing actuator length.

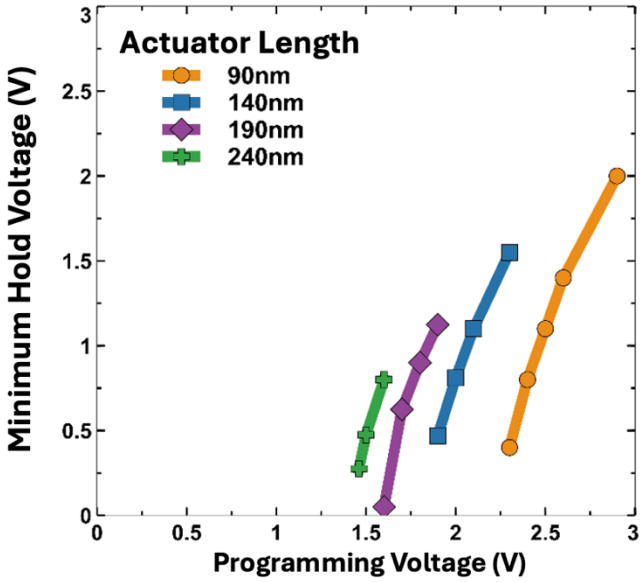


Figure 2.11: Minimum hold voltage vs. programming voltage for 3ML NV-NEM switches.

2.4 NEM Switch Crossbar Array Applications

The vertical orientation of the BEOL NEM switch makes it particularly well suited for dense, nonvolatile memory. A basic way to use the NEM switches for NEMory is to directly readout the state of the switches, by detecting if the drain electrode (data-line) is contacting the switch. For example, in order to read the state of the switches in column 0 below, the data-line 0 voltage is pulsed, which is the metal layer corresponding to the drain electrodes of the NEM switches. Then, the bit-line voltages on the lowest metal layer are monitored. If the voltage rises, and is sustained (i.e., not just to capacitive displacement current), then bit encoded by the switch's state is read as "1".

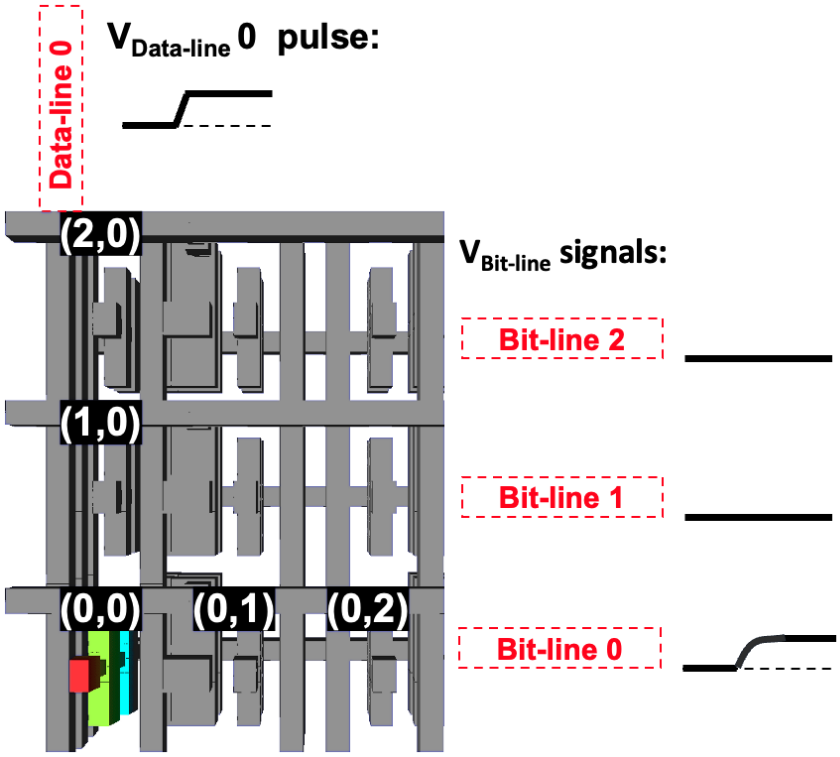


Figure 2.12: Basic NEMory readout. When data-line 0 is pulsed, bit-line 0 follows this pulse since switch (0,0) is contacting data-line 0. Bit-lines 1 and 2, however, does not follow the pulse, because switches (1,0) and (2,0) are not contacting data-line 0.

But this use of NEM switches does not take advantage of their single-pole double throw structure. More complex architectures could be realized to take advantage of their full functionality. A reconfigurable look-up table (LUT) [2.17] can be implemented like below, where the input/output values of the truth table could be programmed into the state (i.e., programmed 0 or programmed 1) of the NEM switches in the array.

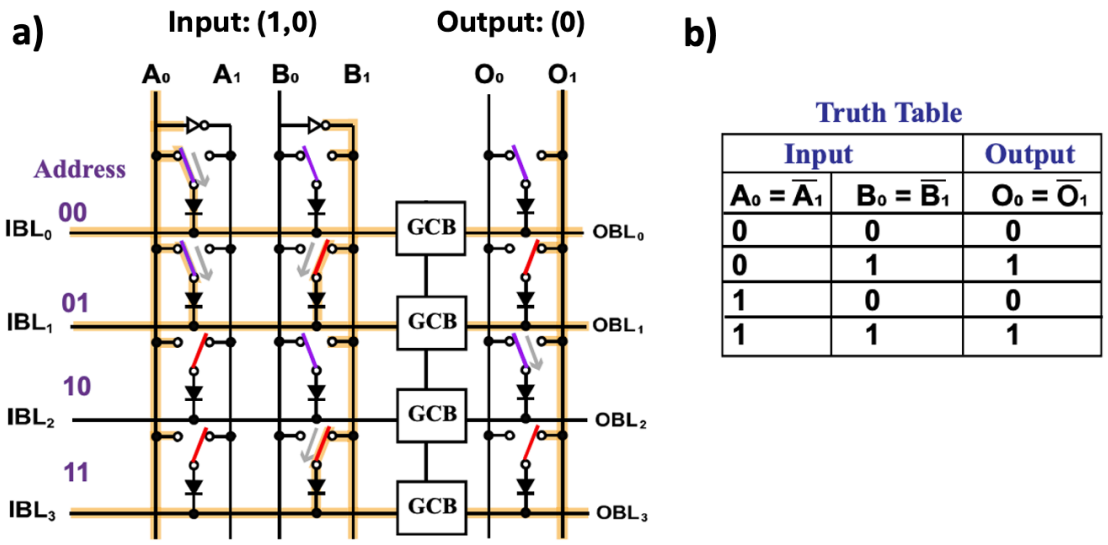


Figure 2.13: Reconfigurable NV-NEM based LUT [2.17].

Figure 2.13a depicts a schematic of a 2-bit input/1-bit output LUT, implemented using an array of NV-NEM switches and gated CMOS buffers (GCB). It features a cross-point array with the input section on the left and the output section on the right. The NV-NEM switches in a single column share contact electrodes D0 and D1, as well as Prog0 and Prog1 electrodes, which program each switch to make contact with either the D0 or D1 electrodes. For simplicity, the program electrodes are not shown.

The array's number of rows matches the number of possible input combinations (2^N address words). Each input combination and its corresponding truth table output (Fig. 2.13b) is programmed by grounding one input bit line (IBL) or output bit line (OBL) at a time, then using the half-select programming scheme so that the programmed states of the switches reflect the truth table. Each IBL connects to the corresponding OBL via a gated CMOS buffer. The IBL for the selected address row remains low while the others are pulled high. For example, input "10" raises all input bit lines except IBL2, which corresponds to the address word 10, as shown in the voltage waveforms in Fig. 2.6. Before reading, all OBLs are pre-charged high. The read enable signal is set high to transfer the IBL states to OBLs; only one OBL is pulled low to discharge an output line and read the stored. A diode in series with each movable beam prevents reverse leakage paths in the cross-point array, which could cause readout errors by discharging bit lines through an input NV-NEM switch.

While the previous two NEM switch architectures rely on passing direct current through the contact electrode surfaces, an alternative approach is to use them as nonvolatile, programmable capacitors.

One of the primary tasks for neural network-based inference systems is performing Multiply-Accumulate (MAC) operations, in which some pre-stored weights are multiplied with incoming data and summed to achieve an accumulated total. The most prominent compute-in-memory approach to achieve the MAC operation is with resistive crossbar arrays in which the current through many resistors is summed to achieve the sum of (Voltage × Conductance) from each resistor. Rather than using resistive elements, it was recently proposed to implement the MAC operation by the summation of the displacement current in a capacitive crossbar array, which avoids certain downsides of resistive crossbar arrays such as static and sneak-path currents.

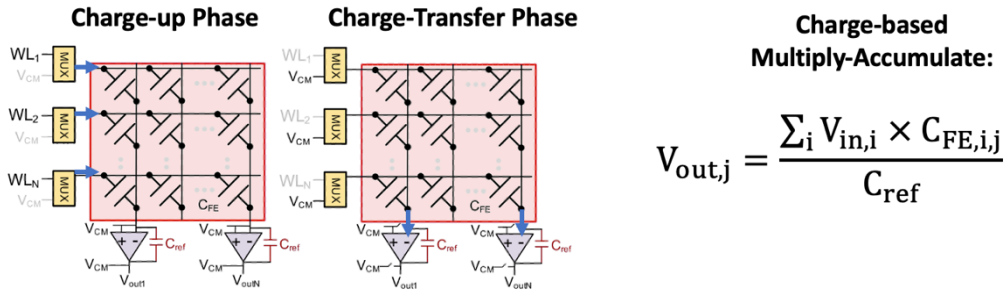


Figure 2.14: Capacitive Crossbar array for multiply-accumulate calculation, adapted from [2.28]. The left hand schematic shows the two-phase calculation and readout operation, and the right hand side shows a mathematical description of the charge-based MAC operation.

Most implementations of capacitive crossbar arrays use polycrystalline ferroelectric capacitors, which can achieve a programmable capacitance due to ferroelectric polarization switching which occurs near to the ferroelectrics' coercive voltage. However, most ferroelectric based capacitors can only achieve an on/off ratio of ~2 or less which limits their functionality, accuracy, and scalability [2.29]. Also, ferroelectric based capacitors for crossbar arrays have shown a reduction in capacitance over time due a gradual shift in the ferroelectric polarization state, which can limit the operating lifetime.

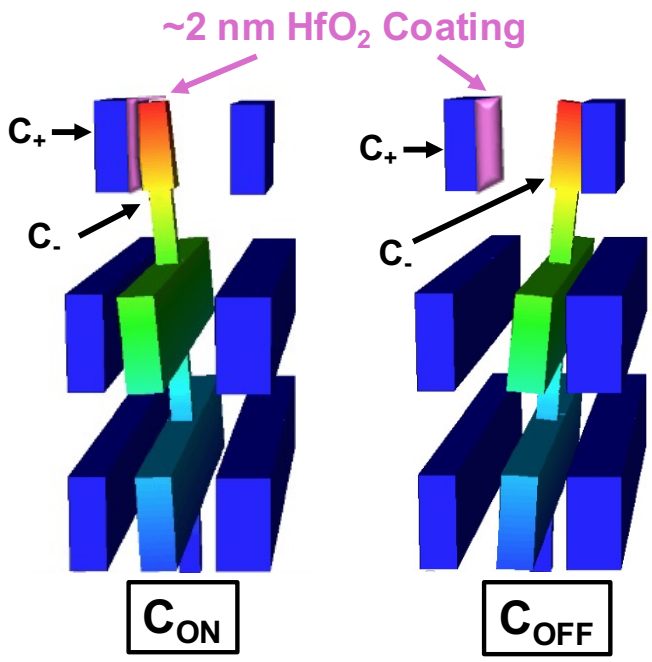


Figure 2.15: Programmable BEOL capacitor. For clarity, hafnia is only shown on the drain contacting surface, but would conformally coat all surfaces in an ALD-based fabrication process. This would have minimal effect on the capacitive ON/OFF ratio.

Since the NEM switch traverses the area between two metal lines when it switches, the gap distance between a drain electrode and the source changes significantly between the 0 and 1 states. To operate the NV-NEM switch as a programmable capacitor, the sidewalls should be coated with a high-K oxide, such as hafnia. Effectively, in the on state, the capacitance is dominated by the hafnia, and in the off state, the capacitance is dominated by the ~ 20 nm air gap between the switch and the electrode surface. For this 5 nm node switch, an ON/OFF capacitance ratio on the order of 200 could be achieved for 2 nm of hafnia coating, with limited drift in the capacitance over time.

2.5 Summary

Optimally designed NV-NEM switches implemented with three BEOL layers in a 5 nm CMOS technology can be programmed with CMOS-compatible voltages and are expected to be more compact than SRAM cells and much more energy efficient (with sub-5 aJ write energy) than any other type of non-volatile memory device. A crossbar array architecture provides for high bit-cell density, and a half-select programming scheme effectively eliminates the need for access transistors without causing write disturb issues. The crossbar array architecture additionally lends itself to the implementation of compact compute-in-memory architectures suitable for the implementation of machine learning systems.

Chapter 3: A Pathway to Giant Negative Differential Resistance in Nanoscale Ferroelectric Field-Effect Transistors

Semiconductor devices that exhibit a negative differential resistance (NDR) characteristic have long been sought after due to their promise of enabling more compact and/or more efficient integrated circuits compared to implementations using only complementary metal-oxide-semiconductor (CMOS) field-effect transistors (FETs). A significant challenge for the development of high-performance NDR devices is the need for them to be compatible with established integrated circuit (IC) manufacturing processes. Ferroelectric FETs (FeFETs) based on CMOS-compatible, hafnia-based ferroelectric gate stack materials have been investigated broadly in the past decade for potential uses in nonvolatile memory, steeply switching logic devices, and neuromorphic computing. This chapter presents a novel FET design that can achieve giant NDR, with a peak-to-valley current ratio (PVCR) exceeding 10^6 via drain-induced polarization switching.

3.1 Negative Differential Resistance Devices

As the pace of CMOS IC technology advancement has slowed, particularly for static memory (SRAM) area scaling in recent technology generations [3.1], the need has grown for new devices that can be easily integrated into a CMOS IC manufacturing process to enable more compact memory bit-cells and digital logic circuits than pure CMOS implementations. NDR devices, which are characterized by an operating region in which current decreases with increasing applied voltage (Fig. 3.1) [3.2], previously have been proposed to implement a wide array of digital logic and memory circuits in a more compact fashion than CMOS approaches (discussed further in the next chapter).

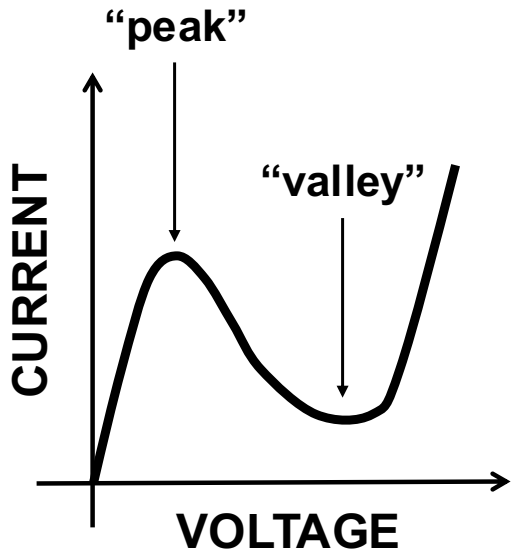


Figure 3.1: Tunnel Diode IV Characteristic. A region of negative differential resistance exists between a “Peak” voltage and a “Valley” voltage.

Electronic devices with a negative differential resistance region of operation have been reported since 1918 with the development of the Dynatron vacuum tube [3.3], originally intended to amplify signals for radio applications. Later in 1958, Leo Esaki serendipitously discovered a NDR IV characteristic in a heavily doped PN junction, and later a NDR characteristic in a resonant tunneling heterostructure [3.4]. In 1960, Goto developed circuits which made use of NDR diodes, and IBM continued development of these for ultra-high speed logic applications, though they were never successfully integrated into a very-large-scale-integration (VLSI) compatible process flow.

Later in the 1990s, RTDs were investigated for >10GHz VLSI systems, but proved challenging to integrate with CMOS transistors and they did not achieve high enough PVCR ($> 10^5$) for low standby power operation, and thus were inadequate for VLSI applications.

A novel concept for achieving NDR in the output characteristic of a MOSFET was proposed in 2000 [3.5] in which charge-carrier trapping and de-trapping dynamically modulate the transistor threshold voltage with changes in the applied drain voltage. However, this concept was also never successfully experimentally demonstrated to achieve very high PVCR.

The generic tunnel diode current-*vs*.-voltage (IV) characteristic [3.2] is compared with an NDR transistor characteristic in Fig. 3.2; note that the tunnel diode's IV characteristic (in blue) has an "N" shape, *i.e*., as the voltage increases, the current reaches a peak before entering the NDR region of operation, and it eventually increases again at higher voltage values. In contrast, the I_D-V_{DS} characteristic of an NDR transistor with fixed applied gate voltage (also illustrated in Fig. 3.2 in red) has a "Λ" shape, with current monotonically decreasing beyond the peak voltage [3.5], [3.6]. NDR transistors are generally based on a dynamically variable threshold voltage. Since the drain current of a transistor varies exponentially with threshold voltage in the subthreshold operating regime, if the threshold voltage could dynamically increase with increasing drain-to-source voltage (V_{DS}), the drain current could exponentially decrease. This is the basis for the high PVCR, "Λ" IV characteristic of NDR transistors.

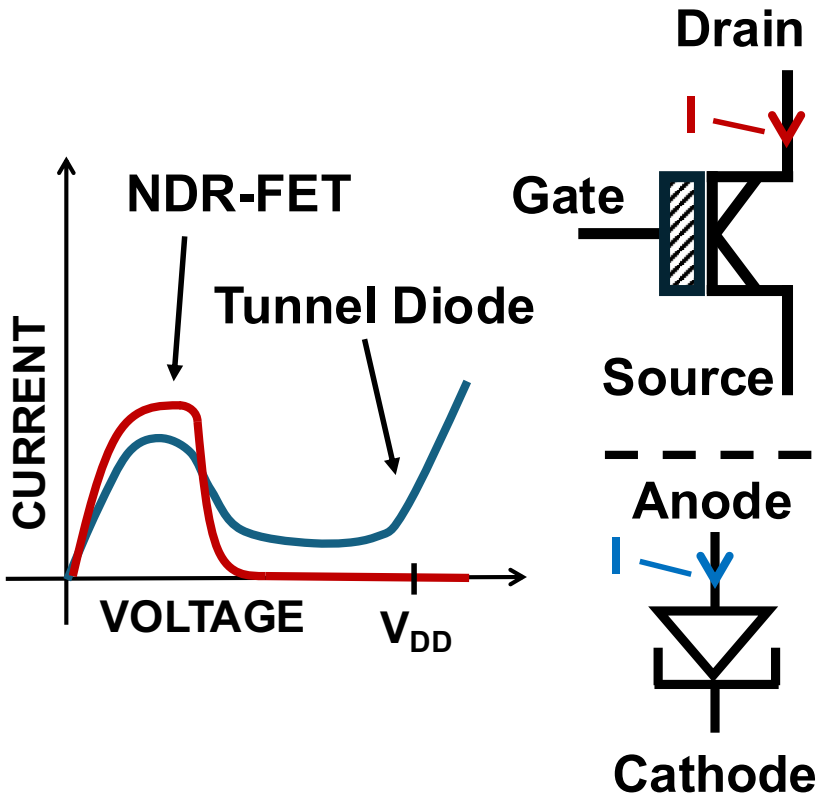


Figure 3.2: Generic IV characteristics of a tunneling diode (blue) and NDR transistor (red) on a linear scale.

3.2 Ferroelectric FETs

Ferroelectrics

Ferroelectric materials are materials which exhibit a spontaneous net dipole moment (polarization), in contrast to dielectric materials which are only polarized in response to an applied electric field. As Fig. 3.3 indicates, the charge-voltage relationship of a ferroelectric has a bistable region in which the material's polarization dictates the built-up charge in response to an applied voltage.

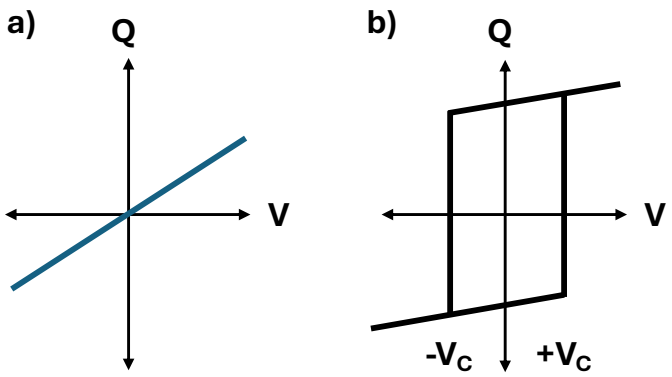


Figure 3.3: Charge vs. Voltage relationship for a) dielectric film and b) ferroelectric film.

The crystal lattice structure of ferroelectrics has a broken space inversion symmetry which allows for a spontaneous polarization; in their equilibrium state, some ionic charge is perturbed from the axis of symmetry, and this leads to a net dipole moment even in the absence of applied electric field. Fig. 3.4 below shows the free-energy diagrams for a dielectric (DE) and (FE) film under no external bias. For the DE, a minimum at zero charge exists, whereas for the FE two minima exist at +/-P_R charge.

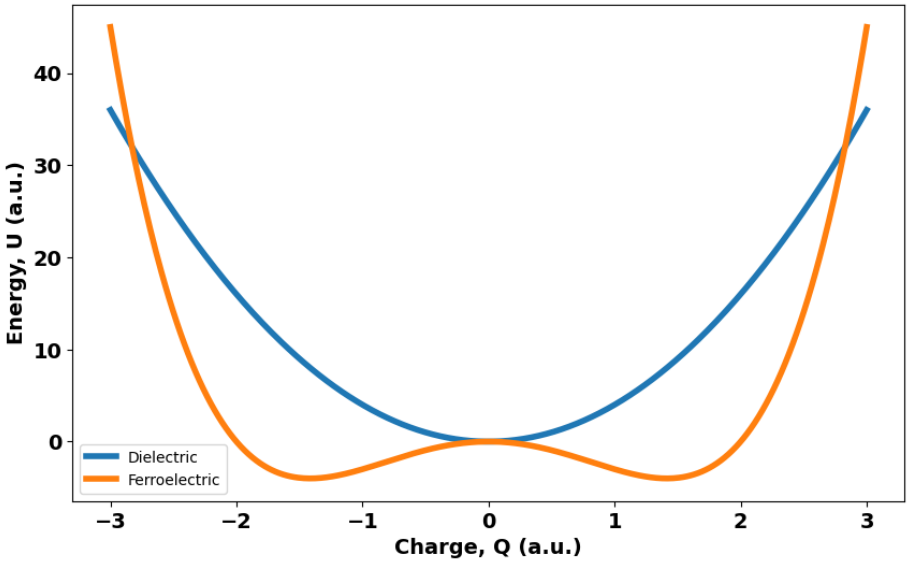


Figure 3.4: A generic Energy vs Charge diagram for a DE and FE film. The DE has a minimum energy for 0 charge buildup, whereas the FE has two local minima.

With an applied bias, the free energy curve shifts so that once the coercive field is reached, only one minimum exists near either $+P_R$ or $-P_R$, and the FE switches polarization states, as illustrated below (Fig. 3.5).

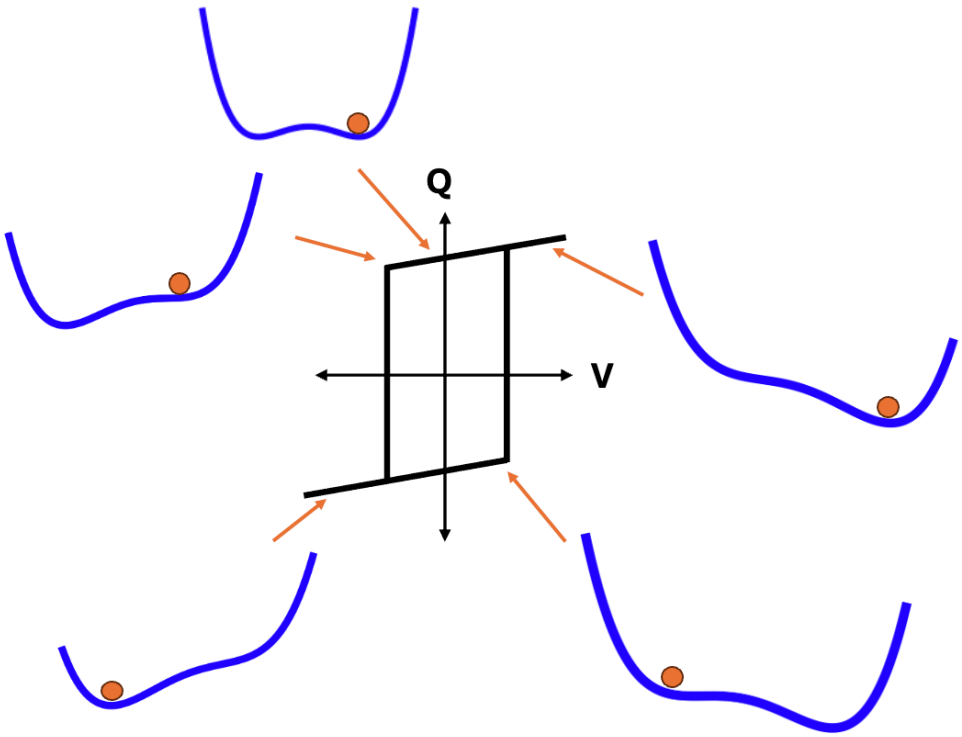


Figure 3.5: Application of an electric field to the ferroelectric shifts the energy vs charge curve (shown in blue, with the operating point indicated in orange) so that when the coercive field is reached, a single energy minimum exists and the FE switches polarization states.

Previously, materials such as strontium bismuth tantalite (SBT) and lead zirconium titanate (PZT) had been investigated for semiconductor memory applications. These exhibit ferroelectric properties due to the bistable states of its atoms about the symmetric crystal structure; a generalized schematic illustrates this for a perovskite structure in Fig. 3.6. However, perovskite-based ferroelectrics begin losing their ferroelectricity dramatically as their thickness is scaled below ~100 nm, which makes them difficult to integrate into most semiconductor process technologies.

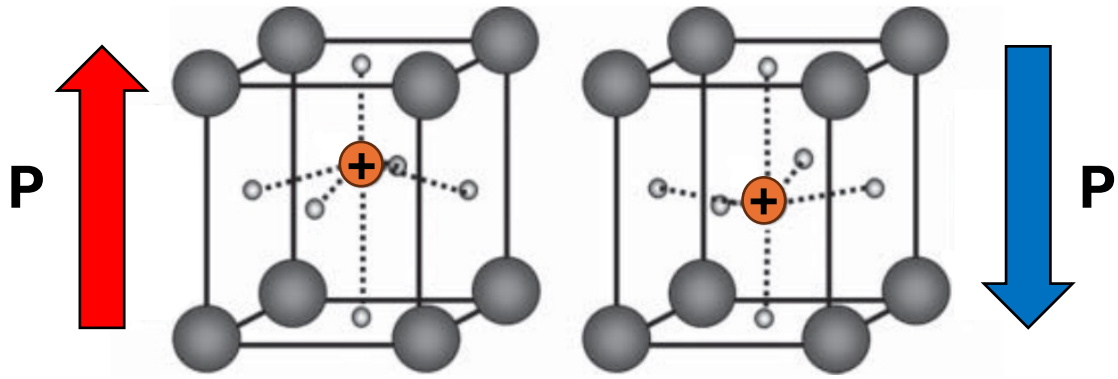


Figure 3.6: Cartoon illustration of ferroelectricity in a perovskite structure, $BaTiO_3$, adapted from [3.7]. The off-center equilibrium positions of Ti below the material's Curie temperature leads to a spontaneous polarization charge.

In 2011 [3.8], it was first publicly reported that HfO_2 can exhibit ferroelectricity in its crystalline orthorhombic phase. Orthorhombic HfO_2 (O-HfO_2) was found to be ferroelectric due to the off-center equilibrium position(s) of its oxygen ions. As shown below in Fig. 3.7, O-HfO_2 has alternating layers of oxygen ions which are polar (i.e., have two equilibrium off-center positions) and nonpolar (i.e., fixed position). Movement of the oxygen ions along the polar layer between their two stable states results in polarization charge switching.

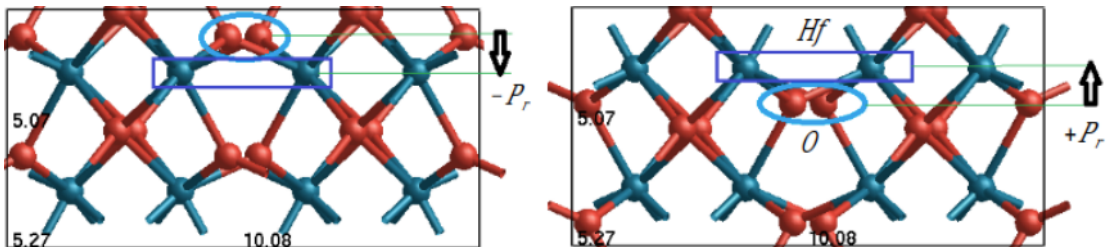


Figure 3.7: Atomistic model of the negative and positive polarized states in orthorhombic hafnia, adapted from [3.9]. The two stable positions of the oxygen ions about the hafnium ions lead to a spontaneous polarization.

Since HfO_2 is already used in advanced CMOS process technology, this ignited a renewed interest in ferroelectrics and their semiconductor applications, with an exponentially increasing number of published works since the first publicized report (Fig. 3.8). It was found that doping HfO_2 could make the orthorhombic phase thermodynamically favorable upon crystallization. The most popular and manufacturable approach is using a hafnia-zirconia

alloy, $Hf_{0.5}Zr_{0.5}O_2$ (referred to herein as "HZO"). Furthermore, HZO was shown to retain its ferroelectricity to nanometer thicknesses due to preferential textured crystal growth at thin thicknesses [3.10]. Ferroelectricity at the nanometer limit has motivated the investigation of ferroelectric based tunnel junctions, capacitors, and transistors for a variety of applications.

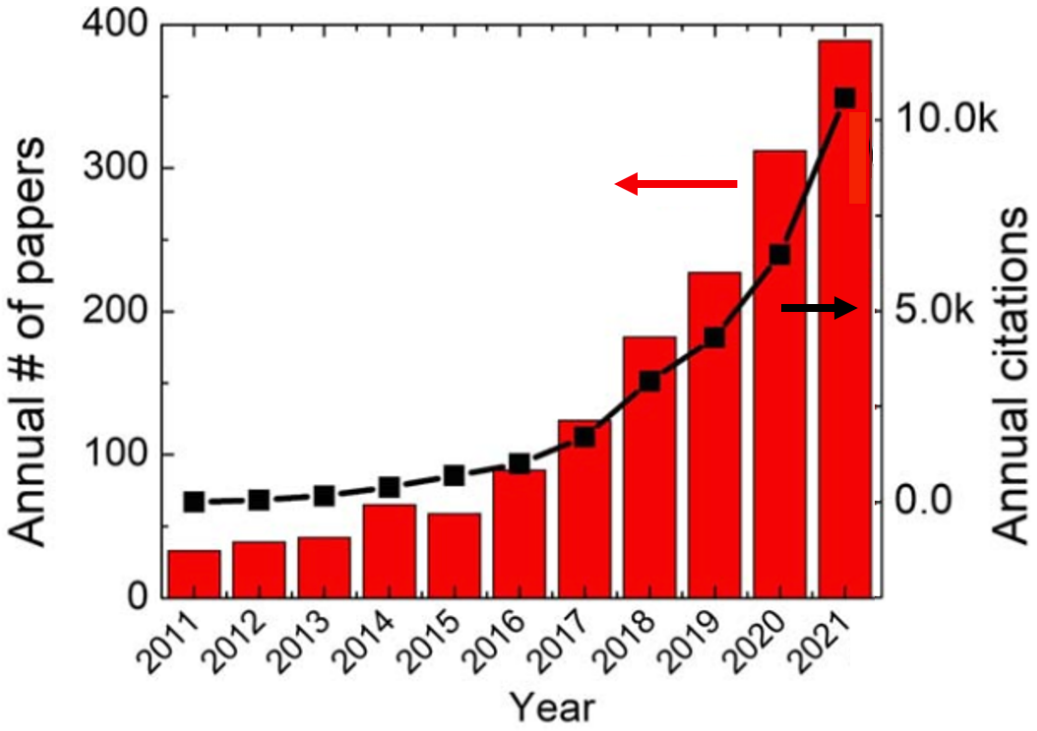


Figure 3.8: The discovery of ferroelectricity in hafnia-based films re-ignited interest in ferroelectrics research, with an exponentially growing body of work since 2011. Adapted from [3.11].

The FeFET

A ferroelectric field effect transistor (FeFET) is a MOSFET which includes a ferroelectric insulator in the gate. Like a classical MOSFET, the polarization charge of the ferroelectric shifts the threshold voltage of the MOSFET by screening the gate charge. An electric field is used to physically shift the ionic polarization charge within the gate stack such that the threshold voltage is electrically programmable. Fig. 3.9 below shows a generic FeFET structure and energy band diagrams for the programmed and erased states. In the "programmed" state, the FE's polarization is positive, and an inversion layer is at the reference gate voltage; this is the low V_T state. In the "erased" state, the FE's polarization is

negative, and the channel is depleted (or even in accumulation) at the reference voltage; this is the high V_T state.

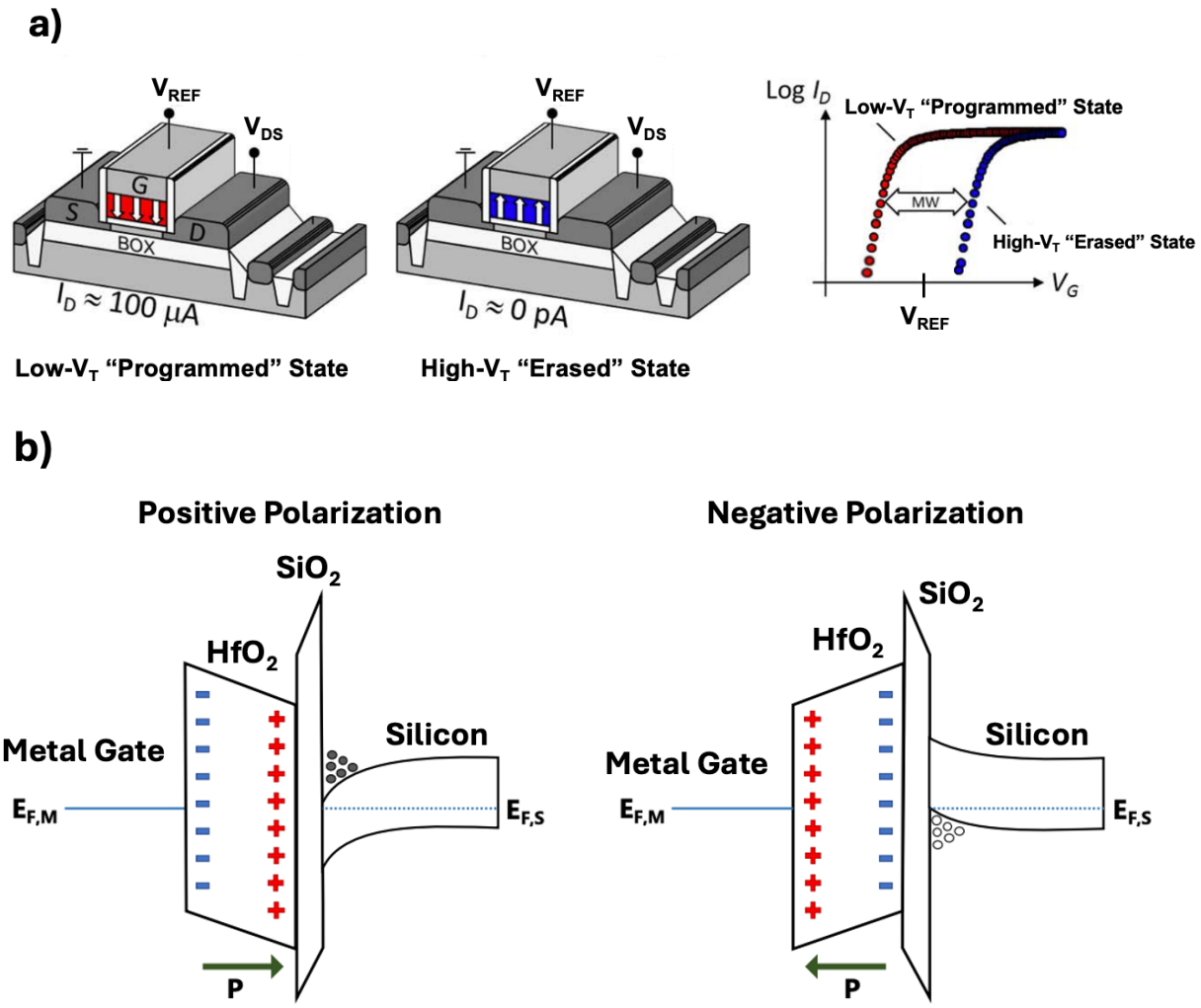


Figure 3.9a): An FeFET in the programmed (positive polarization) and erased (negative polarization) states. b) The energy band diagrams corresponding to these programmed and erased states. Adapted from [3.12], [3.13].

Wide-area FeFETs have different switching dynamics than nanoscale FeFETs. When a FE film is made up of many domains, each with slightly different resistance to switching, the net switching characteristic is quite gradual, like that shown in Fig. 3.10a. The switching of an individual domain is known to result in a sharp transition. This box-like characteristic is known as a "hysteron", shown in Fig. 3.10b. Variations in individual domains' coercive

voltage may be either due to the grain it is a part of (orientation, stress, etc.), or localized electrostatics about the individual domains. When a film has only a few domains, its polarization switching characteristic might exhibit a stepped form. Furthermore, this gradual vs. abrupt switching characteristic is also observed in FeFETs (Fig. 3.11), depending on the number of domains in its FE.

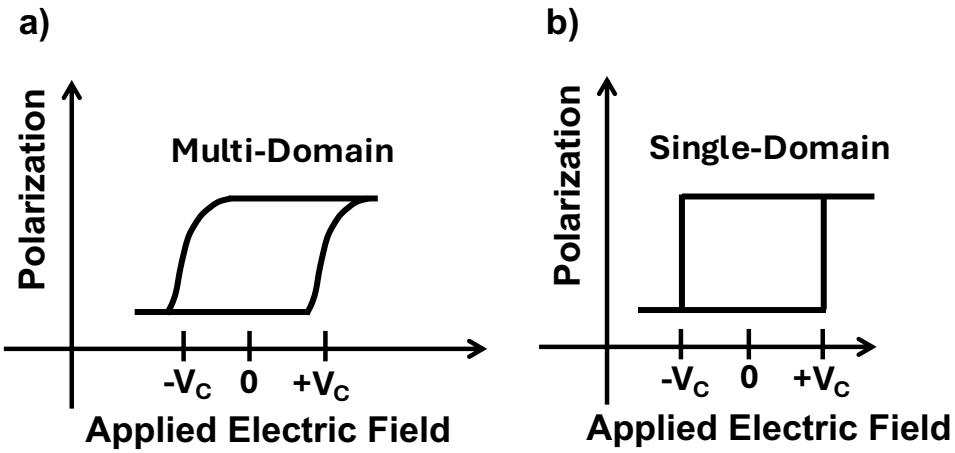


Figure 3.10: A multi-domain ferroelectric (a) switches polarization states gradually as the applied electric field is varied, while a multi-domain switches polarization states abruptly as the electric field is varied.

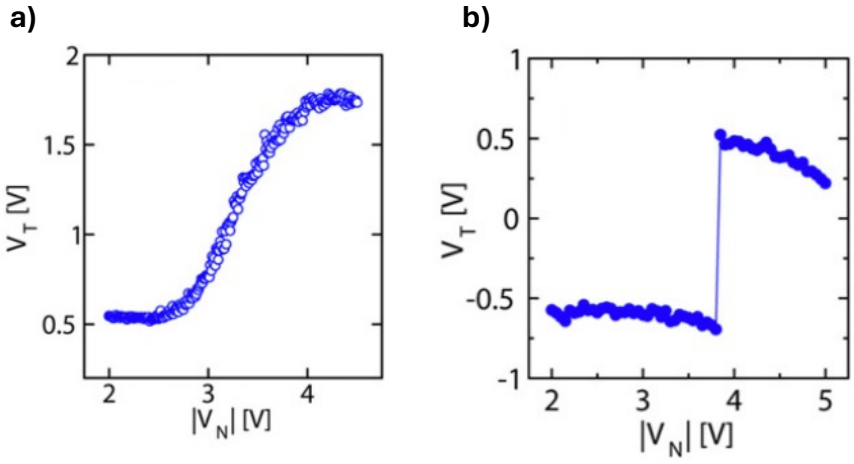


Figure 3.11: Comparing the pulsed-erase operation for a micron-scale area, many domain FeFET (a) with a nanoscale-area, few-domain FeFET (b). The V_T increases gradually as the

pulse voltage increases for the many domain FeFET, vs abruptly for the few-domain FeFET. Adapted from [3.14].

Since the FE polarization is modulated using the applied electric field, an FeFET can be configured for both gate-dependent and drain-dependent behavior. Since the gate and drain are at opposite ends of the ferroelectric, a positive gate voltage tends to polarize the FE in the positive direction, while a positive drain voltage tends to polarize the FE in the negative direction [3.15]. This drain voltage-induced negative polarization causes a positive V_T shift.

Mild NDR has been observed in simulated [3.16], [3.17] and measured [3.18], [3.19] output characteristics of Negative Capacitance FETs and FeFETs. Both gradual NDR (Fig. 3.12a) and abrupt (Fig. 3.12b) NDR have been observed. This has been attributed to drain-induced partial polarization switching of the ferroelectric layer, and is also described as "Reverse DIBL" stemming from Transient Negative Capacitance (TNC) [3.18]. Note that in both highlight cases, the "reverse" sweep (V_{DS} swept from V_{DD} to 0 V) does not show any NDR behavior, and the device does not return to a "high current" state. A "reset" pulse is required to re-polarize the ferroelectric to the positive direction, so this NDR behavior is not dynamic. While this behavior degrades transistor on-state current and hence is not desirable for conventional CMOS digital logic circuits, it could be leveraged for implementation of NDR-based memory and logic circuits. If the drain-induced V_T shift could be enhanced across the entirety of the channel, it could be harnessed to achieve a giant NDR characteristic.

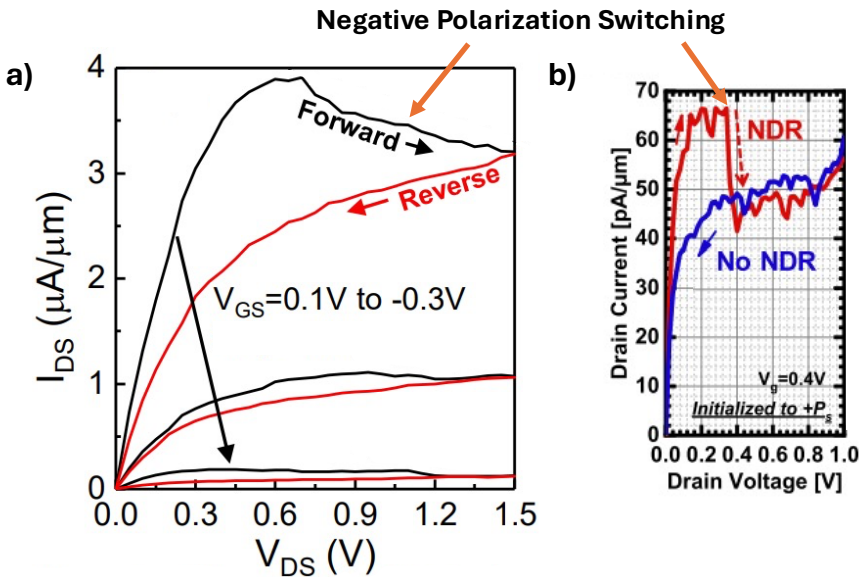


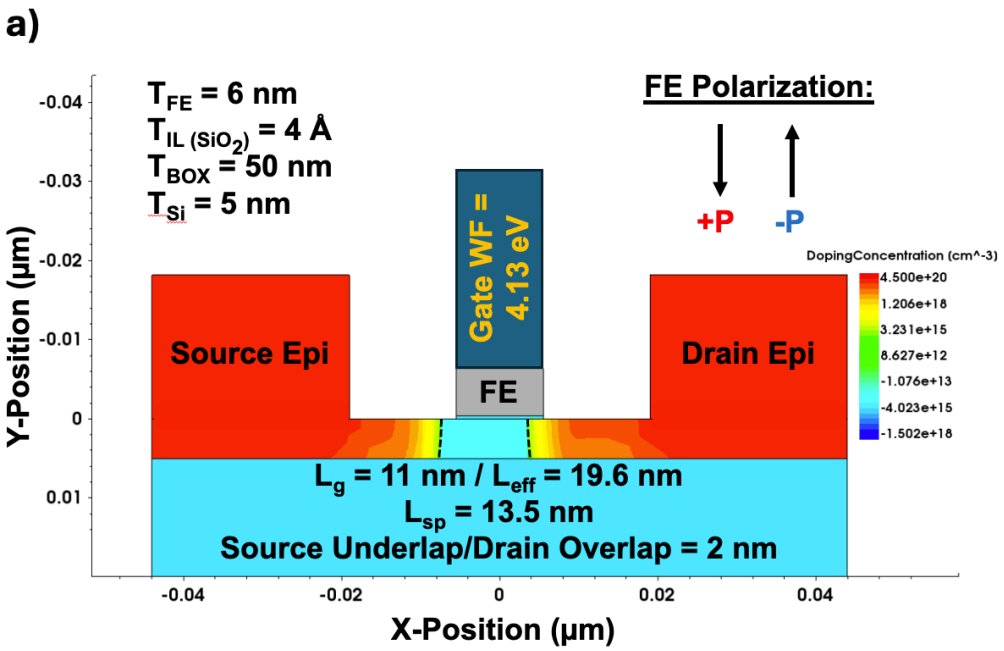
Figure 3.12: Previous reports of NDR in FeFETs include a) "gradual" NDR behavior and b) "abrupt" NDR behavior. In b) the current is low, due to subthreshold operation ($V_{GS} < V_T$). Adapted from [3.18], [3.20].

3.3 The NDR FeFET Device

Our proposed NDR FeFET design, schematically illustrated and detailed in Fig. 3.13a, uses polarization reversal induced by the applied drain voltage to dynamically modulate the transistor threshold voltage. In this work, 2D device simulations in Sentaurus TCAD software [3.21] were used to investigate the design of a fully depleted silicon-on-insulator (FDSOI) NDR transistor. Drift-diffusion, thin-layer mobility, velocity saturation, and band-to-band tunneling models were used to model carrier transport. Strain-enhanced mobility and ballistic transport were not considered. The third-order Ginzburg-Landau-Khalatnikov (G-L-K) model of polarization (below) was used to model the behavior of the HZO ferroelectric (FE) layer in the gate-insulator stack. Due to the scaled lateral dimensions (< 15 nm), the electric polarization within the HZO layer switches with a sharp, single-domain behavior (Fig. 3.13b) rather than a gradual multi-domain behavior.

$$-\Gamma\frac{\partial P_y}{\partial t} = \alpha P_y + \beta {P_y}^3 - g\nabla^2 P_y - E_y$$

(Ginzburg-Landau-Khalatnikov Model for FE Polarization)



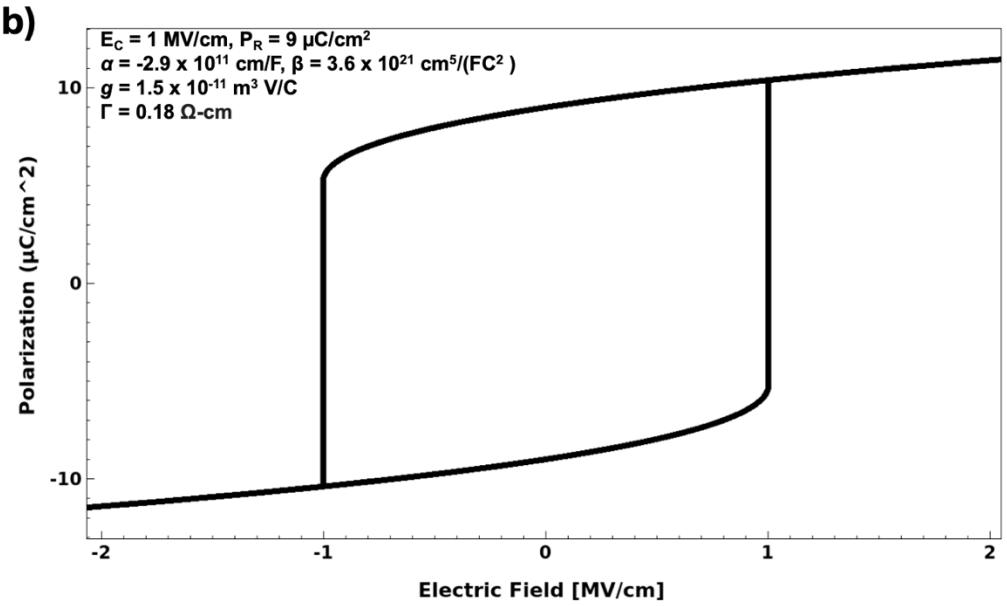


Figure 3.13a): Proposed NDR FeFET structure simulated in this chapter. The FDSOI device features air-gap gate-sidewall spacers and an asymmetric doping profile. The source/drain junctions are indicated by the dotted lines. b) Polarization vs. Electric Field characteristic of the baseline HZO layer used in this work using the Landau-Ginzburg model in Sentaurus Device.

Fig. 3.14 plots simulated quasi-DC (10 ms sweep time) I_D-*vs*.-V_{DS} characteristics for a depletion-mode NDR FeFET, for three values of applied gate-to-source voltage (V_{GS}). Note that the native threshold voltage in the positive polarization state ($V_{T0} \approx$ -0.35 V) and positive polarization switching voltage (-15 mV) of this device is designed to be negative (via gate work function tuning) so that it is in the on state for V_{GS} = 0 V, which enables the device to function as a 2-terminal NDR diode when the gate and source electrodes are tied together. When the drain-to-source voltage (V_{DS}) is small, the FE is polarized positively. As V_{DS} is increased to the peak-current voltage (V_{PEAK}) and beyond, the electric field across the FE layer at the drain end is reduced to such an extent that the FE layer switches to the negative polarization state, effecting an increase in transistor threshold voltage (V_T) and thereby an exponential decrease in transistor current. The greater the shift of charge within the FE layer, the greater the change in V_T and the lower the valley current [3.22]. As V_{DS} is reduced back towards 0 V, eventually the average electric field across the FE layer becomes sufficiently positive to switch back the FE polarization state, lowering V_T so that an inversion layer is once again formed in the channel region. Due to the hysteretic polarization switching behavior of the FE layer, the device exhibits hysteretic NDR behavior, but no reset operation is required. Fig. 3.14a shows that as V_{GS} is increased, PVCR is

dramatically reduced and the NDR region shifts to higher drain voltages, eventually resulting in no NDR behavior is for large V_{GS}. In other words, this device can switch between NDR and non-NDR operating modes by modulating the gate voltage. Fig. 3.15 compares the conduction band energy level for a low drain bias and a high drain bias. Remarkably, the source-side energy barrier for electron injection in the channel is massively raised at high drain bias, due to the ferroelectric switching to the negative polarization state!

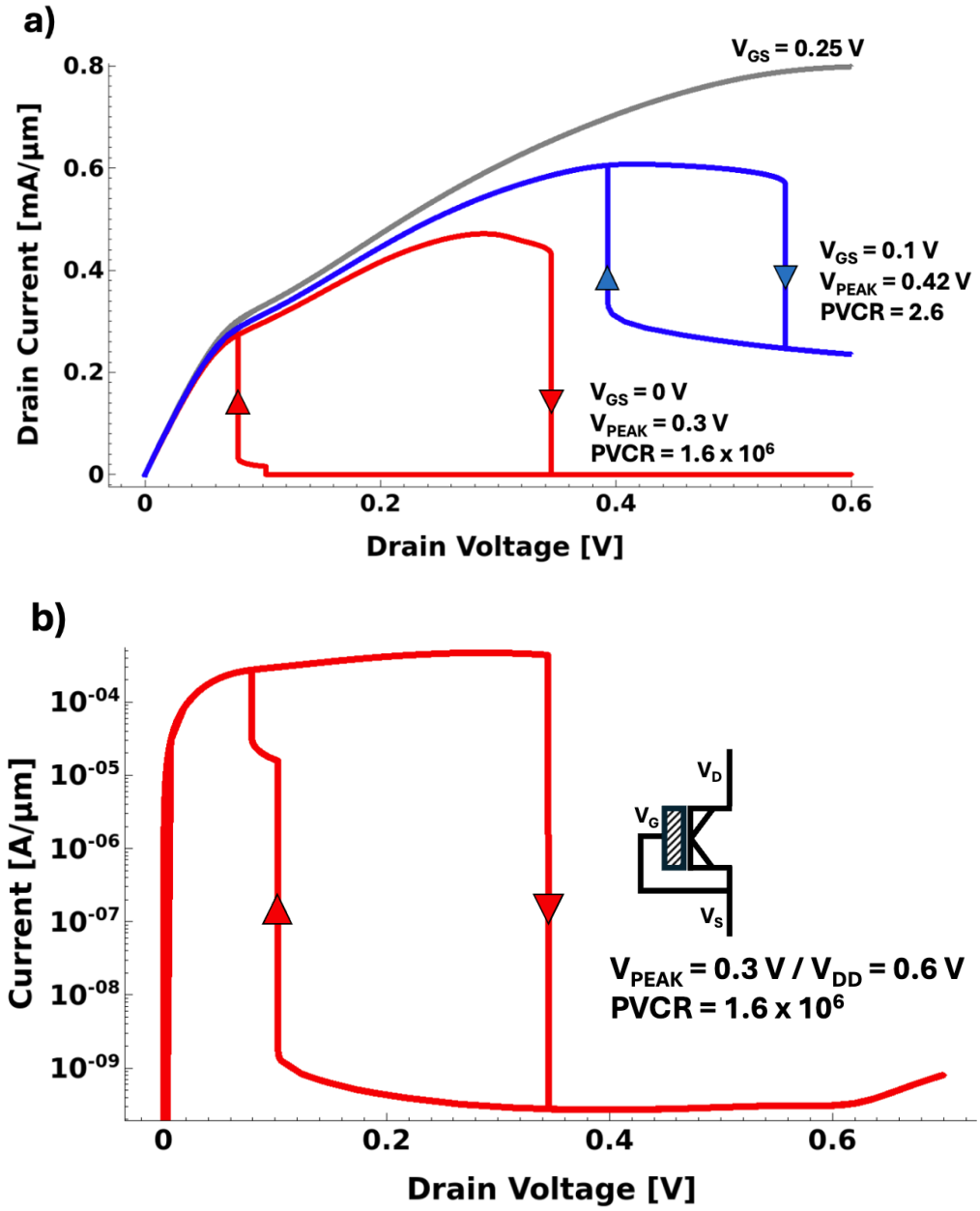


Figure 3.14: Simulated I_D-V_{DS} curves for a depletion-mode NDR FeFET for various values of applied gate voltage. The NDR region shifts to higher drain voltage ranges with increasing

V_{GS}, so that no NDR behavior is seen for large V_{GS}. b) Logarithmic scale I_D-V_{DS} curve for the depletion-mode NDR FeFET for V_{GS} = 0 V. Most of the NDR occurs in 1 or more discrete steps (c.f. Fig. 3.12b) as the FE polarization abruptly switches.

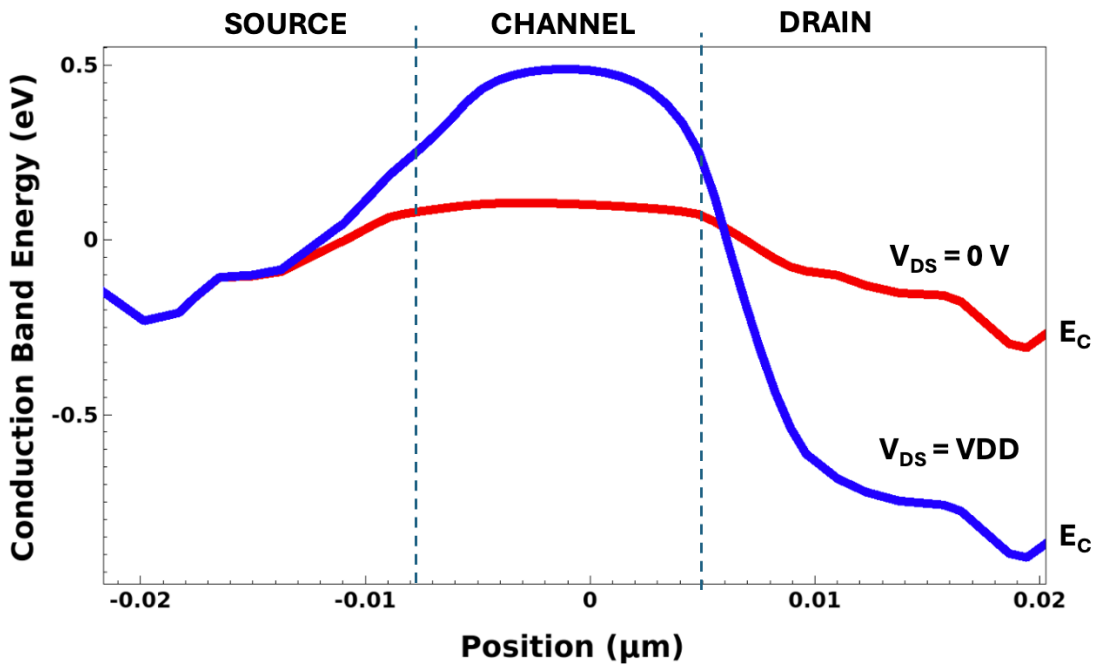


Figure 3.15: Conduction band edge profile of the NDR FeFET for low drain bias (red) and high drain bias (blue, V_{DD} = 0.6 V). The barrier to conduction is raised from ~0.15 eV to ~0.58 eV.

Fig. 3.16 plots simulated quasi-DC I_D-*vs*.-V_{GS} characteristics for the depletion-mode NDR FeFET, for two values of V_{DS}. A range of gate voltages exists (from -0.05 V to 0.15 V) for which the drain current is lower for the higher value of V_{DS} than for the lower value of V_{DS}, indicative of NDR behavior. Unlike most FeFETs reported in the literature, nanoscale NDR FeFET has an abrupt switching characteristic with low voltage hysteresis. The steep, hysteretic turn-on/turn-off switching behavior is typically reported for multi-domain devices as TNC [3.23], although it is seen herein for a single-domain FE layer. It also should be noted that the hysteresis voltage is smaller than that for the FE layer alone (*i.e.*, $2 \times E_C \times T_{FE}$) due to the "voltage snapback" effect wherein the voltage dropped across the FE layer reduces (or reverses polarity) when polarization switching occurs; this can be optimized to further reduce the hysteresis voltage.

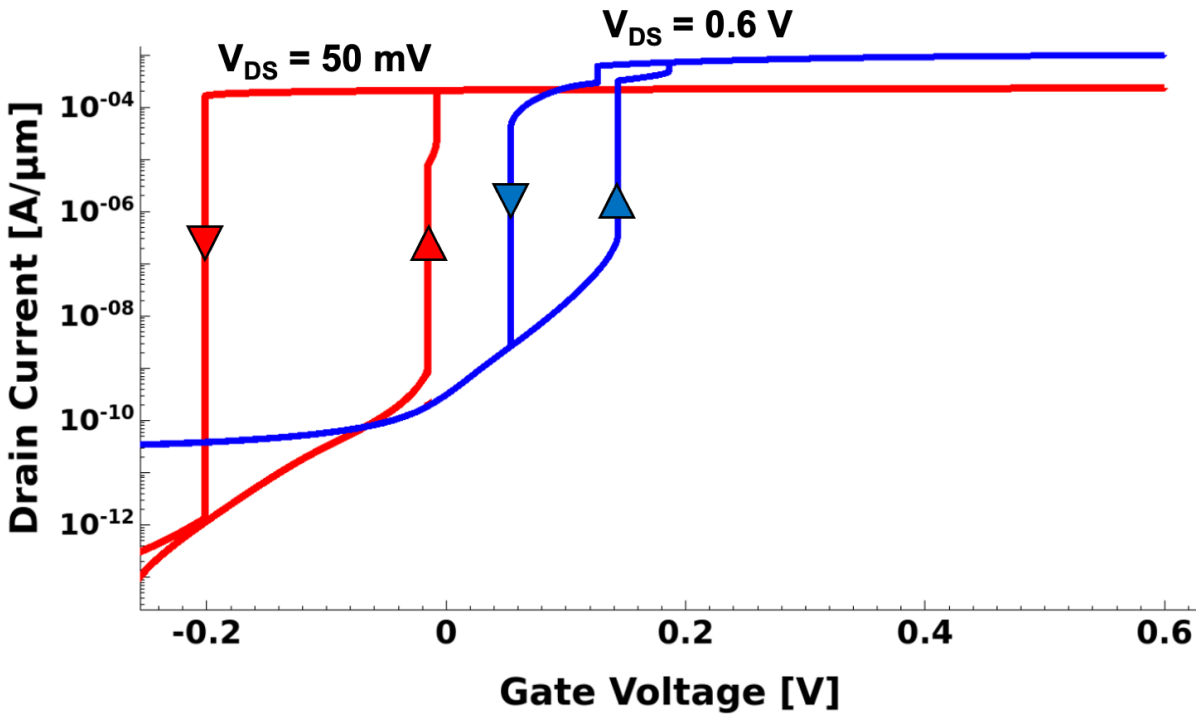


Figure 3.16: Simulated I_D-V_{GS} curves for a depletion-mode NDR FeFET for low V_{DS} (50 mV) and high V_{DS} (0.6 V). There exists a gate-voltage range in which the current is lower for high V_{DS}.

It should be noted that FeFETs used for nonvolatile memory are evaluated by their "Memory Window", which describes the threshold voltage shift between their "programmed" (positive polarization) and "erased" (negative polarization) state. For these multi-domain FeFETs with large coercive voltage gate stacks, the shift in threshold voltage is evaluated via a constant current method. It is typical to compare the V_{GS} at which 1 μA/μm drain current is obtained for the positive polarization and negative polarization states to evaluate the Memory Window. This type of FeFET's memory window is typically approximated as the shift in flat band voltage (when the net polarization of a multi-domain FeFET is equal to 0), $2 \times E_C \times T_{FE}$ [3.24]. For FeFETs with large coercive voltage, the point at which the drain current equals 1μA/μm does not correspond to the complete switching of the FE film. Fig. 3.17 illustrates that as the coercive voltage (and corresponding hysteresis) becomes even lower, the voltage difference between the high-V_T and low-V_T states for this constant current point becomes even lower even though the switched polarization charge is not reduced. This definition of memory window diverges from the V_T shift that the NDR FeFET experiences; therefore, a different metric must be used.

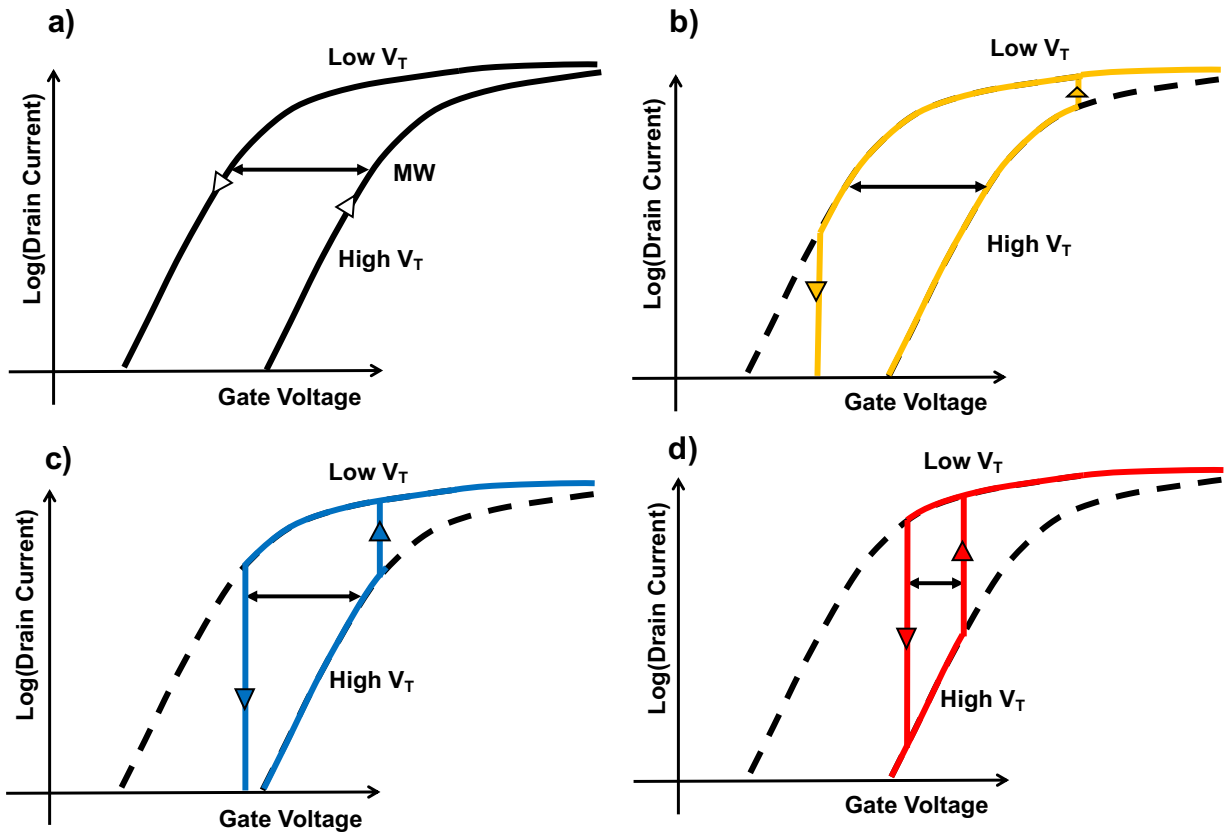


Figure 3.17: As coercive voltage of the FE is reduced from a) to d) (keeping P_R the same), the constant-current MW reduces by 60%, even though the ΔV_T is the same between all cases. The MW is equal to ΔV_T for a) and b) since chosen the constant-current value does not intersect the voltage at which polarization switching occurs.

Firstly, for the NDR FeFET device, the hysteresis and ΔV_T are not defined by $2 \times E_C \times T_{FE}$ since the FE dynamically switches during device operation. Instead, the hysteresis is reduced, due to the voltage snapback that occurs, to less than $2 \times E_C \times T_{FE}$, approximately:

$$V_{HYST} = 2 \times E_C \times T_{FE} - \frac{2P_R}{C_{FE} + C_{IL}}$$

Additionally, trapped electron charge tends to increase the hysteresis window since electrons are available for when an inversion layer is present; the negative polarization switching tends to occur at a lower voltage while the positive polarization switching voltage is unchanged.

The dynamic threshold voltage shift ΔV_T due to polarization switching can be approximated by the following equation [3.25]:

$$\Delta V_T = \frac{2P_R}{C_{FE}}$$

Furthermore, the theoretical maximum PVCR can then be estimated by:

$$PVCR_{max} = \frac{\Delta V_T - V_{OV}}{SS_{sat}}$$

...though this does not account for the impact of gate-induced drain leakage current. V_{OV} is the gate overdrive voltage, and SS_{sat} is the subthreshold slope at high V_{DS}. The abrupt polarization switching characteristic defining the hysteresis window sets a minimum overdrive voltage (as V_{GS} must be kept above the positive polarization switching gate voltage). Although a higher overdrive voltage effects a higher peak current, it can increase the peak voltage and limit the PVCR.

The NDR FeFET operating speed will be limited by the time required to switch FE layer polarization state. Studies have experimentally measured polarization switching time < 200 ps and indicate that sub-10 ps may be possible [3.26]. Chatterjee et al. experimentally measured the relaxation time of a thin HZO film and determined the inversion charge density of a MOSFET could be screened within 270 fs during FE switching, and the entire ferroelectric polarization could switch within 5 ps [3.27]. Further experimental work is needed to confirm that fast polarization switching can be achieved with electric field strength close to the FE layer's coercive field (E_C), for low voltage and high-speed operation.

3.4 NDR FeFET Design Considerations

For low-power operation, an NDR FeFET should have small V_{PEAK}, high PVCR, and small hysteresis voltage such that it returns to the low-V_T state when V_{DS} returns to 0 V. This requires a FE layer with low coercive voltage, adequate remanent polarization (P_R), and a high degree of drain voltage control over the polarization state of the FE layer. To achieve a low coercive voltage, a thin FE layer (< 10 nm thick) with low E_C (≤ 1 MV/cm) should be used. HZO and other fluorite-based ferroelectrics can maintain ferroelectricity for thicknesses below 2 nm [3.10], and E_C may be reduced below 1 MV/cm by the addition of atomically larger dopants such as La, Y, or Gd [3.28], or by seed-layer techniques [3.29]. In this work, 5 nm-thick and 6 nm-thick HZO films were considered, assuming modest values for P_R (9 μC/cm^2) and E_C (1 MV/cm), as a baseline. Currently, demonstrations of gate stacks with both a sharp switching characteristic and low voltage hysteresis in the literature are limited.

An alternative to a classical ferroelectric film is to use an anti-ferroelectric film, in which the polarization vs voltage hysteresis loop breaks into two half-loops:

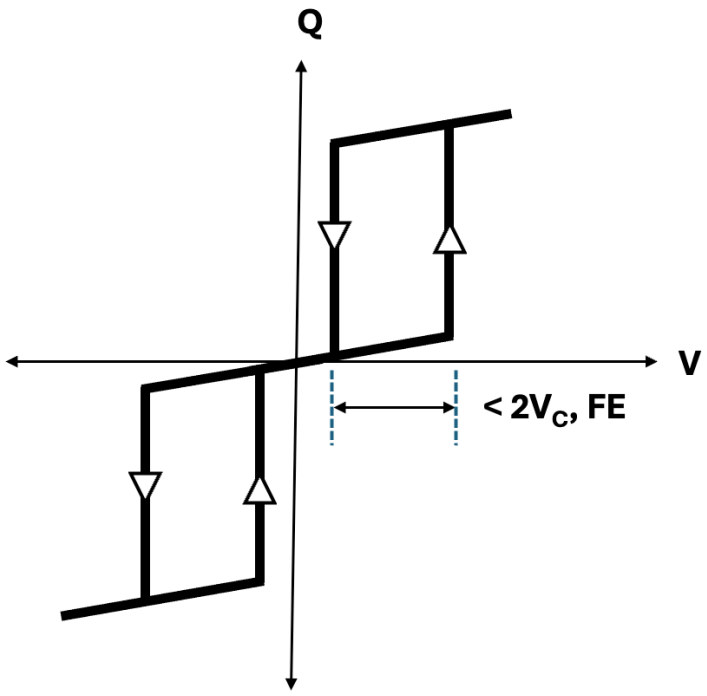


Figure 3.18: Charge-vs-Voltage characteristic for an antiferroelectric. The hysteresis window of the half loops individually is generally less than the hysteresis window of a ferroelectric made in the same material system.

This “half-loop” switches approximately half the amount of charge as its ferroelectric equivalent (reducing ΔV_T and PVCR), but also approximately half the voltage hysteresis. An anti-ferroelectric tetragonal-phase HZO or ZrO_2 film may be employed to achieve this behavior [3.30], [3.31]. In general, these anti-ferroelectrics also intrinsically exhibit up higher cycling endurance than their ferroelectric counterparts, which may be due to their reduced atomic distortion during polarization switching.

Key to achieving large PVCR in the NDR FeFET is polarization switching in the FE layer across the entire length of the channel. In other words, it is essential to maintain a single polarization domain in steady state (*i.e.*, for $V_{DS} = 0$ or $V_{DS} = V_{DD}$) rather than multiple domains. In addition to engineering the careful growth of a high quality, single-grain FE layer, classical device engineering techniques can promote single-domain behavior. Fig. 3.19 plots PVCR as a function of the NDR FeFET gate length (L_g), keeping the other transistor design parameter values fixed. It can be seen that PVCR increases dramatically when L_g is reduced below a critical length. The polarization contour plots show that for $L_g <$ 13 nm there exists a single polarization domain within the FE layer, whereas for $L_g \geq 13$ nm there are multiple domains. As evident in the contour plots, in this work the polarization is allowed to vary continuously within the FE layer due to the implementation of the G-L-K model in Sentaurus Device; DFT calculations suggest that polarization is likely more discrete with a small domain-wall transition region (a few lattice cells wide) [3.32].

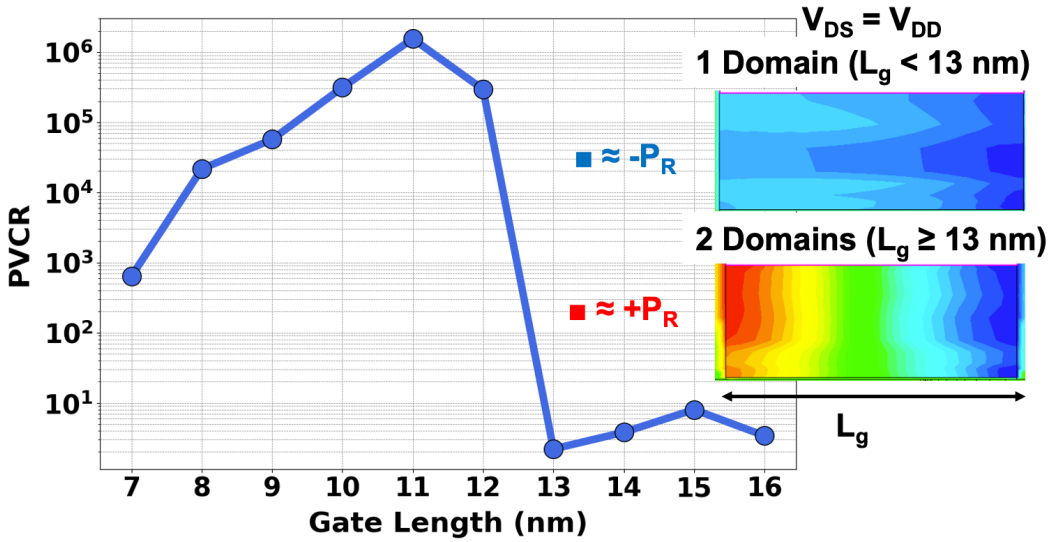


Figure 3.19: PVCR as a function of NDR FeFET gate length (T_{FE} = 6 nm, T_{IL}= 4 Å), showing a dramatic increase when the device transitions from multi-domain to single-domain FE. Inset: FE polarization profile calculated using the G-L-K model.

To ensure a high degree of drain voltage control, capacitive coupling between the FE layer and the source should be reduced by offsetting the source junction from the gate edge and using low-permittivity (low-k) gate-sidewall spacers. Accordingly, all NDR FeFET designs studied in this work have an asymmetric underlap/overlap source/drain doping profile (cf. Fig. 3.13a). Fig. 3.20 plots PVCR as a function of L_g for different values of gate-sidewall spacer relative permittivity. The use of low-k or air-gap spacers enables giant NDR at longer L_g, widening the design window. Alternatively, spacer thickness could be increased to reduce the capacitive coupling between the FE layer and the source. However, the increase in footprint area required makes this approach impractical: for the device with a 5 nm thick FE layer, to achieve a 1 nm (8%) increase in maximum gate length, the air-gap spacer thickness had to be increased by 3.5 nm (26%), for a total 7 nm increase in device length.

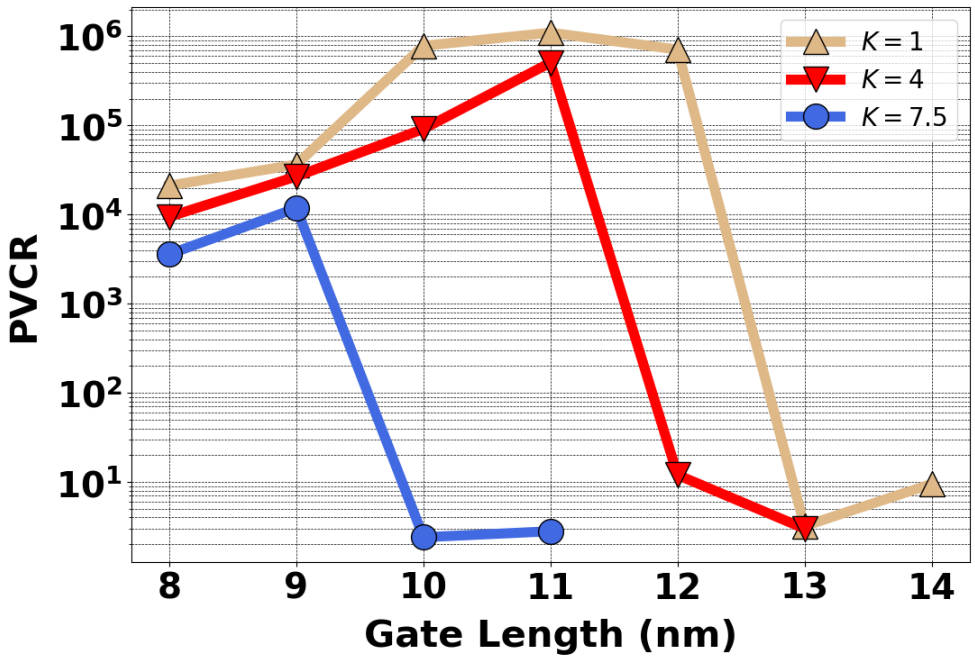


Figure 3.20: PVCR as a function of NDR FeFET gate length (T_{FE} = 5 nm T_{IL}= 1 Å) for various gate-sidewall spacer materials: silicon nitride (k = 7.5), low-k (k = 4) and airgap (k = 1).

Underscoring these device optimizations, the ferroelectric layer used in the gate stack must comprise a single grain to prevent domain-wall formation at the grain boundary (or boundaries). This may be achieved during the manufacturing process by choosing conditions which promote the formation of large orthorhombic (or tetragonal, for anti-ferroelectric-based devices) grains which can form over the entire channel area of a

minimum-sized NDR FeFET device (< 15 x 15 nm^2 for the studied FDSOI geometry) [3.10], [3.33], [3.34]. More recently, enhanced burn-in procedures such as a "thermal rewake-up" process [3.35] have been shown to obtain HZO grain sizes to at least 40 x 20 nm^2. Furthermore, the ferroelectric's domain density must also be reduced sufficiently to enable single-domain operation. In the G-L-K model, the polarization gradient energy term *g* describes how readily domain walls form to divide the FE into multiple domains. DFT calculations have suggested that applying compressive strain to the ferroelectric film along the "Continuous Polar Layer" direction (which is preferably aligned to the channel direction) can increase the size of the domains formed in the film by increasing the polarization gradient factor *g* [3.36]. Device simulations indicate that an unstrained (baseline g = 1.5 x 10^{-11} m^3 V/C) 5 nm HZO film with the FE properties studied in this work requires a metallurgical gate length below 12 nm to achieve single-domain operation and high PVCR. Consequently, a film under 1% compressive strain (g = 1 x 10^{-11} m^3 V/C) requires L_g below 9 nm, and 1% tensile strained film (g = 2 x 10^{-11} m^3 V/C) would require L_g below 16 nm to achieve high PVCR. Controlling the film orientation and polarization gradient factor may require new material growth innovations to seed preferential orientation growth, since the crystallization of HZO on an SiO_2 interfacial layer is typically a random nucleation-growth process. On the other hand, recent experimental reports suggest that the size of "elementary" nucleation-growth regions (setting the upper limit of domain-expansion) can be quite large; for example, in a 10 nm polycrystalline HZO film this elementary region was found to be on the order of 40 nm in diameter [3.37], which may allow for the use of gate lengths much greater than predicted by the G-L-K model here.

V_T-shift based NDR operation makes the PVCR sensitive to gate stack parameters. The SiO_2 interfacial layer (IL) between the FE and silicon channel region introduces a capacitive divider between the FE and IL for the switched polarization charge, dampening its effect on the channel potential. Oxygen scavenging can be used to reduce the IL thickness, and has been shown to reduce the coercive voltage of a FeFET gate stack and to enhance the ferroelectricity of thin HZO layers [3.38]. Fig. 3.21 plots the achievable PVCR for V_{DD} < 1 V as a function of IL thickness, showing that PVCR generally increases to 10^6 as the IL is scaled down. The PVCR for all cases reaches a limit somewhat above 10^6 due to gate-induced drain leakage (GIDL) current. A thinner FE layer increases the capacitance C_{FE} resulting in a smaller V_T shift, which generally reduces PVCR. However, Fig. 3.21 indicates that optimizing the gate stack to achieve greater switched polarization charge can retain a large V_T shift, therefore achieving high PVCR.

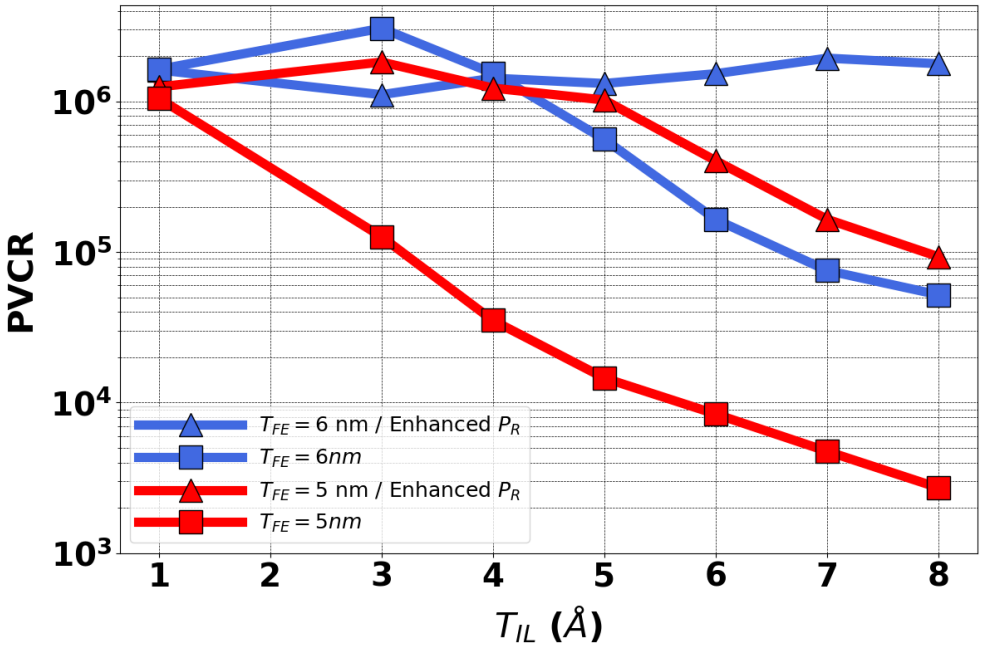
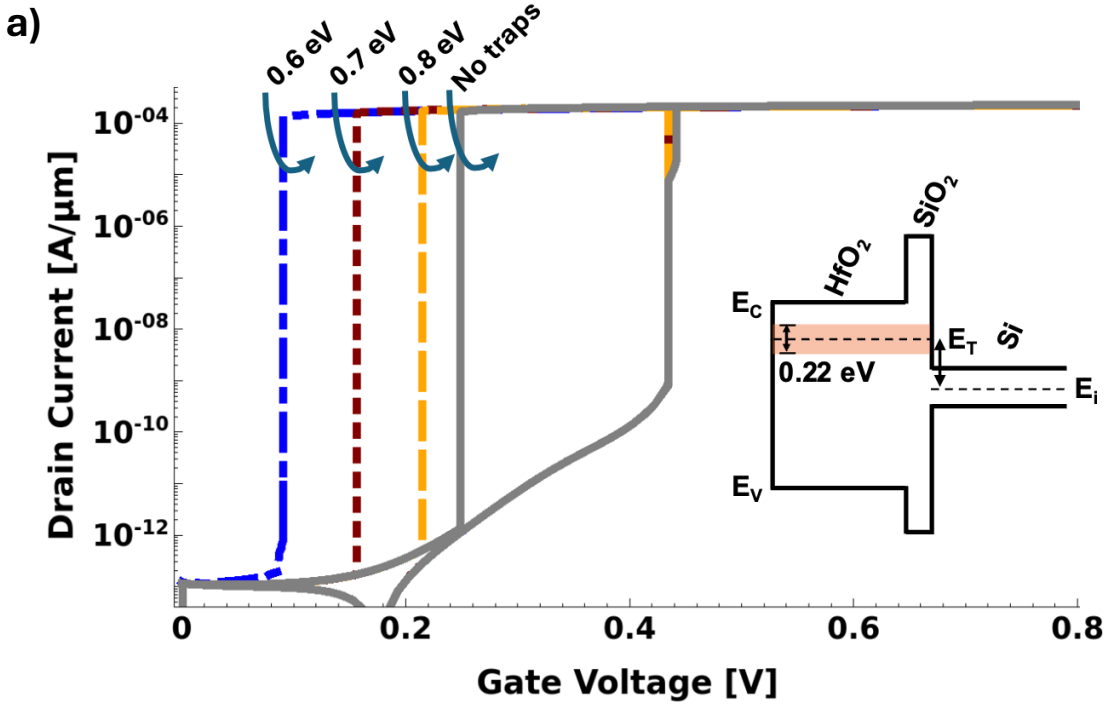


Figure 3.21: Effects of interfacial layer (IL) thickness scaling and enhancing ferroelectricity (P_R=13 μC/cm^2/E_C = 1.4 MV/cm) on PVCR for 11 nm gate length.

3.5 Impact of Trapped Charges

The effects of trapped charge are modeled herein using empirical parameters reported in [3.18], [3.32]. Tunneling between traps and the channel region was modeled using the Wentzel–Kramers–Brillouin model in Sentaurus Device. For dynamic electron charge traps ($2.6 \times 10^{20}/cm^3$ in the FE and IL), it was found that the I_D-V_{GS} hysteresis (Fig. 3.22a) increases as the traps become more energetically aligned with the silicon conduction band edge, resulting in undesirably larger V_{PEAK} (Fig. 3.22b). If the traps are distributed with a mean less than 0.7 eV above the Si midgap, a supply voltage > 1 V must be used to ensure proper NDR FeFET circuit operation.



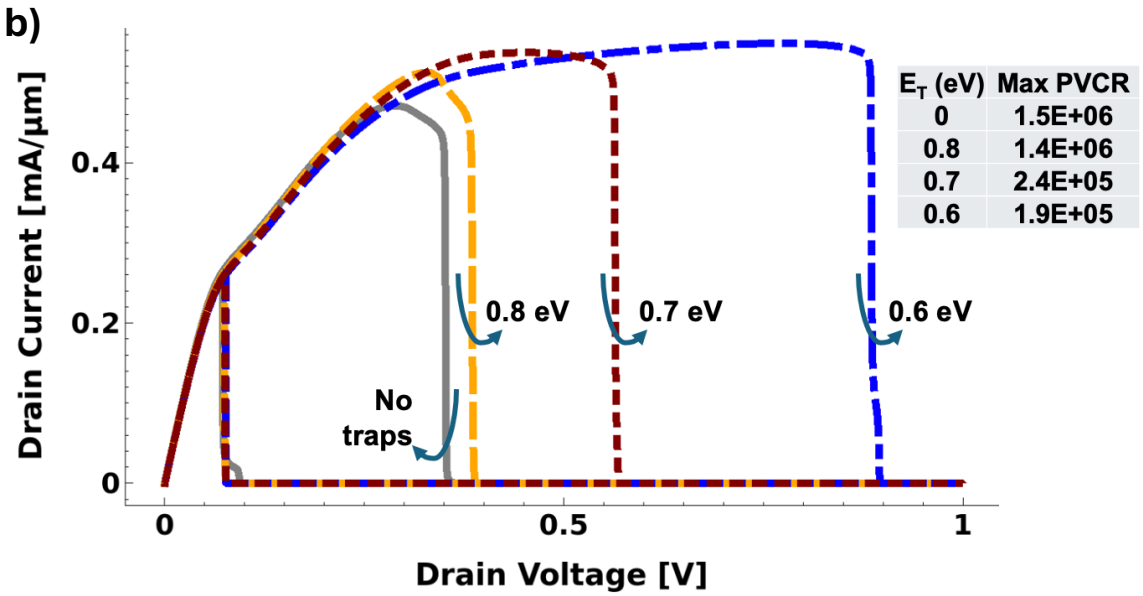


Figure 3.22: I_D-V_{GS} hysteresis increases due to dynamic electron trapping, as the electron trap energy level becomes aligned with the conduction band edge in the silicon channel region. The trap distribution bandwidth is 0.22 eV. b) Peak voltage increases due to dynamic electron trapping, as the electron trap energy level becomes aligned with the conduction band edge in the silicon channel region.

In addition to dynamic electron trapping, electrons may accumulate at the FE/IL interface after many FE switching cycles [3.32]. Fig. 3.23a shows how this buildup of negative charge (known as the imprint effect) reduces the gate voltage at which polarization switching occurs and introduces a pinched I_D-V_{GS} characteristic for $>2\times10^{12}/cm^2$ accumulated electron density. With increasing accumulation of electrons at the FE/IL interface, the peak (and FE switching) voltage becomes larger and PVCR is degraded (Fig. 3.23b), which will limit device endurance. Techniques to increase NDR FeFET cycling endurance beyond 10^{12} cycles (required for SRAM Last-Level-Cache write endurance) towards 10^{16} cycles based on IL optimization are needed to meet operating lifetime requirements and minimize the amount of trapped charge [3.39], [3.40]. FeFETs based on oxide semiconductors with no explicit IL have been shown to have superior endurance due to the substantial reduction in charge-trapping [3.23] and may be suitable for high-density, 3D-integrated NDR-based SRAM.

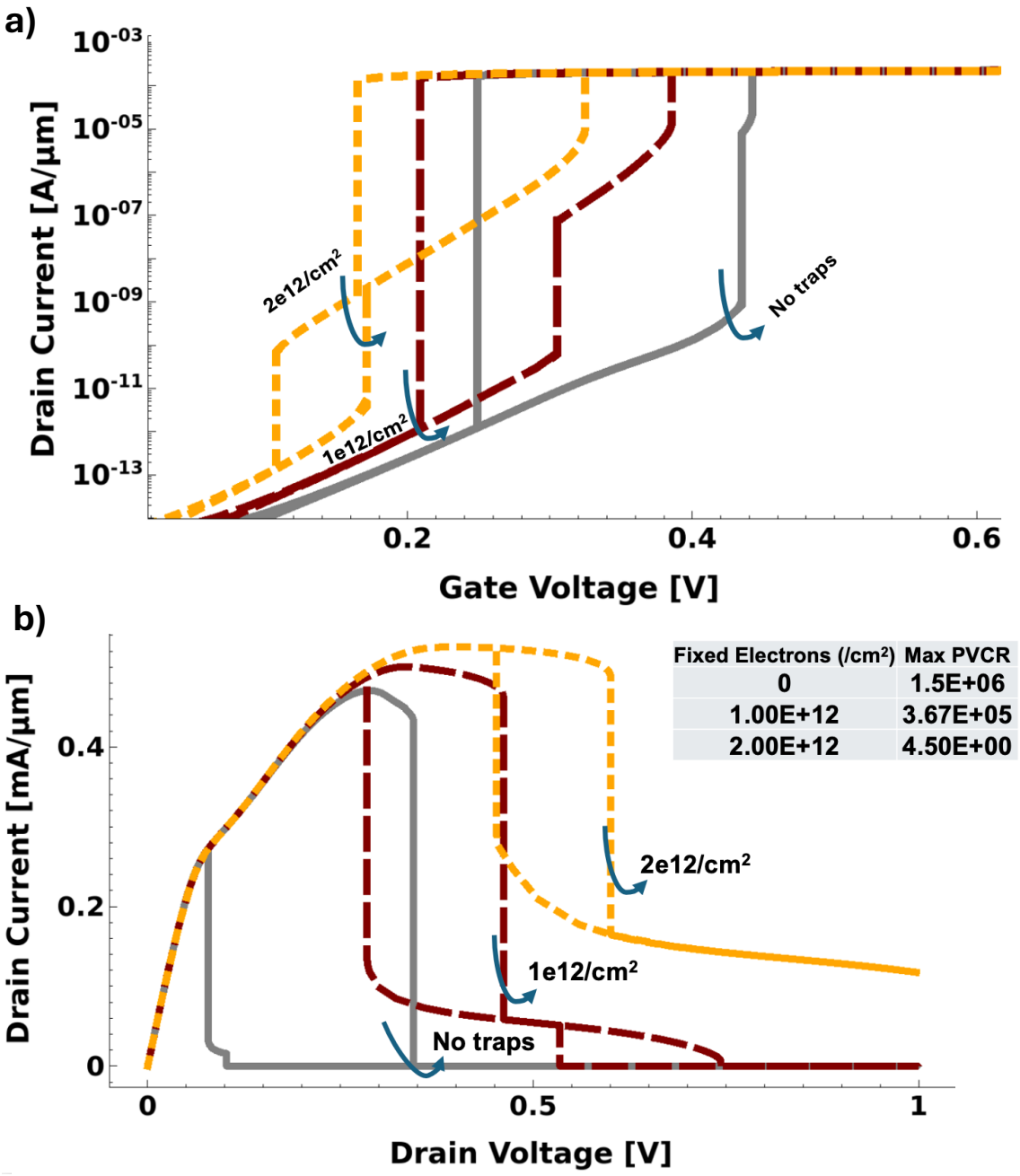


Figure 3.23: A high density of accumulated electrons at the FE/IL interface shifts the I_D-V_{GS} characteristic to the left and can introduce a double-loop characteristic. (V_{DS} = 50 mV) . b) Fixed electron charge at the FE/IL interface above 10^{12}/cm^2 can cause a rapid reduction in the PVCR by increasing the V_{DS} at which the ferroelectric layer becomes completely negatively polarized.

3.6 Summary

This study has investigated a new direction for nanoscale FeFET devices: an NDR mode of operation. A CMOS-compatible NDR FeFET was proposed and shown via TCAD simulations to be capable of achieving PVCR > 10^6 at low V_{DD} (< 1 V). Optimization of device design parameters and continued ferroelectric technology advancements are essential to fully realize the promise of FE-based NDR devices for more compact/efficient ICs. The table below compares the NDR FeFET with other proposed CMOS-compatible NDR devices having a peak voltage below 0.5 V. Compared to these ([3.6], [3.41], [3.42]), the NDR FeFET has an extremely high peak current density, PVCR, and is the most straightforward to integrate in advanced CMOS process technology.

Table 3.1: Benchmarking against proposed NDR devices

Reference/ organization	Description	Peak current	Peak voltage	PVCR
J. Plummer/ IEDM2005 [3.41] (Stanford)	SiGe gated diode	0.8×10^{-6} MA/cm^2	0.15 V	295
A. Lochtefeld/ IEDM2008 [3.42] (RIT/Notre Dame/AmberWave)	GaAs tunnel diode	10^{-3} MA/cm^2	0.16 V	43
J. Appenzeller/T-ED 2022 [3.6] (Purdue/Notre Dame)	Cross-coupled gated heterojunction diode	3.8 MA/cm^2 (23 μA/μm)	0.4 V	6.9 x 10^5
This work (UCB)	**Nanoscale FeFET**	**130 MA/cm^2 (0.5 mA/um)**	**0.3 V**	**3 x 10^6**

Chapter 4: Simulation-Based Study of Compact SRAM Bit-Cells Implemented using NDR FeFETs

Increased static-random-access memory (SRAM) storage capacity is necessary to achieve significant improvements in computational performance to meet today's artificial intelligence and machine learning demands. For many years, NDR-based SRAM bit-cell architectures have held the promise of greatly increasing the density and energy efficiency of embedded SRAM, but to date, a satisfactory NDR device technology has not been demonstrated which can fulfill the performance requirements. The NDR FeFET device concept described in the previous chapter could fill this need, with its CMOS-compatible manufacturing process and high PVCR greater than 10^6. This chapter describes NDR FeFET based SRAM bit-cells and benchmarks these with other emerging memory technologies. Unlike previously proposed NDR devices, the NDR FeFET's hysteretic output characteristic can be leveraged to enable low supply voltages (< 0.5 V) and nonvolatile operation. Static power consumption is projected through TCAD simulation to be > 10x less than state-of-the-art FinFET-based SRAM, with operating speeds suitable for L1 cache application (100 ps).

4.1 Data Abundance and the Demands of AI

In the late 2010s, the rise of "Big Data" led to the amount of information being stored and processed skyrocketing, much faster than embedded memory capabilities could keep up. However, the early 2020s shifted this focus towards another rapidly growing field: Artificial Intelligence (AI), and particularly Generative AI. Generative AI is, to date, the most successful approach towards emulating humans' behavior and has quickly found applications among a variety of fields. One of the key enablers of generative AI is the use of an exponentially growing number of model parameters which are calculated during model training and used to process data inputs during general operation (inference). Fig. 4.1 below shows a clear inflection point around 2020 when "Large Language Models", such as ChatGPT, were identified as a promising approach to emulate human communication. The most advanced models currently have over 1 trillion parameters.

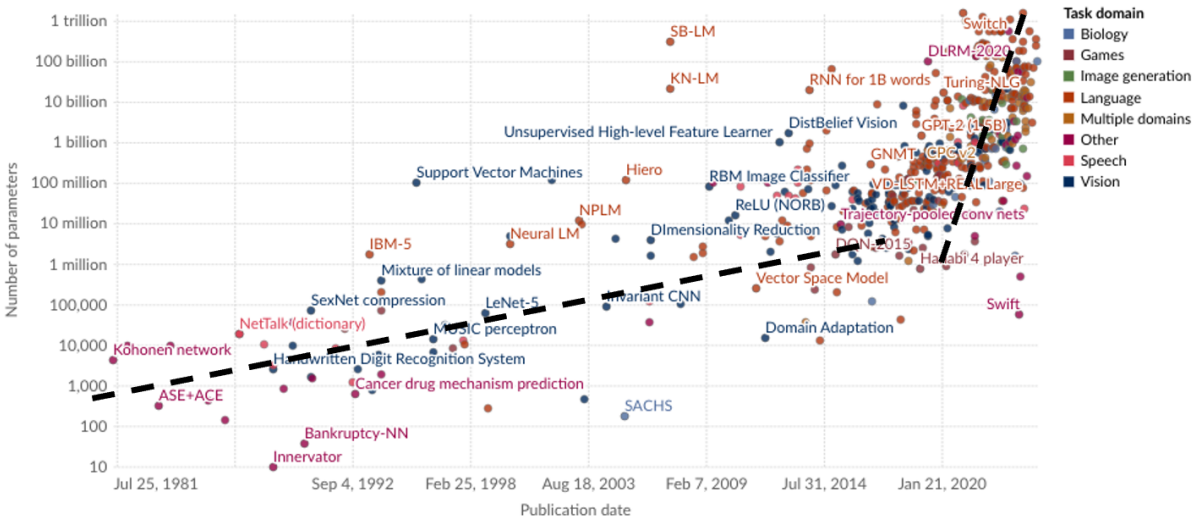


Figure 4.1: The number of parameters required for training AI models has increased exponentially over time, and hit a steep inflection point with the introduction of GPT models in the past few years. Most of the models with > 10 billion parameters are Language models. Adapted from [4.1].

Loading and processing this exponentially increasing number of parameters puts a significant strain on current state-of-the-art processors' resources. Limited on-chip SRAM capacity contributes a memory bottleneck, and can reduce the system's peak performance. To mitigate this issue, the amount of die area devoted to SRAM has steadily grown for recent chip designs. Companies such as Advanced Micro Devices have even

begun stacking discrete SRAM dies ("chiplets") in a configuration they call "V-Cache" in order to increase the amount of available cache memory and reduce the number of data transfer operations with off-chip high density memory devices. The diagram below (Fig. 4.2) is a logic layout analysis from TechInsights, a competitive analysis firm, which shows that for 7 nm and 5 nm node flagship chips, the amount of die area dedicated to static memory is about 30%, on average. It is expected that specialized AI chips will require an even greater portion of the chip dedicated to SRAM cache memory.

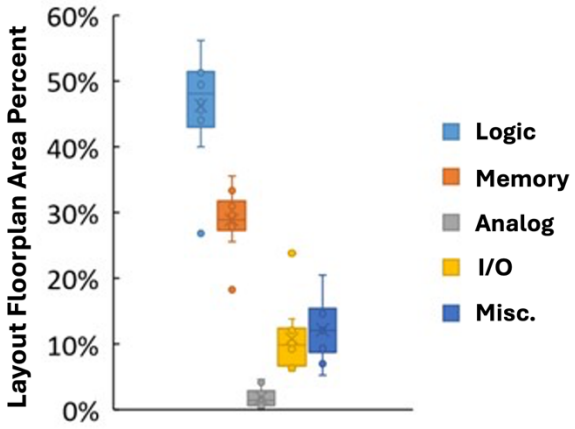


Figure 4.2: TechInsights survey of 7 nm and 5 nm node logic chip layouts indicate memory layout area accounts for, on average, ~30% of total area. Adapted from [4.2].

However, the pace of SRAM area scaling with each new generation of IC manufacturing technology has slowed dramatically in recent years, due to increasing sensitivity of transistor performance to process-induced variations with miniaturization. This presents a growing challenge for continued improvements in computing performance, energy-efficiency, and cost. The compilation of TSMC's SRAM "High Density" bit-cell area from 2015-2024 in Fig. 4.3 shows that SRAM bit-cell area has yet to significantly scale below 0.02 μm^2, staying flat since 2020.

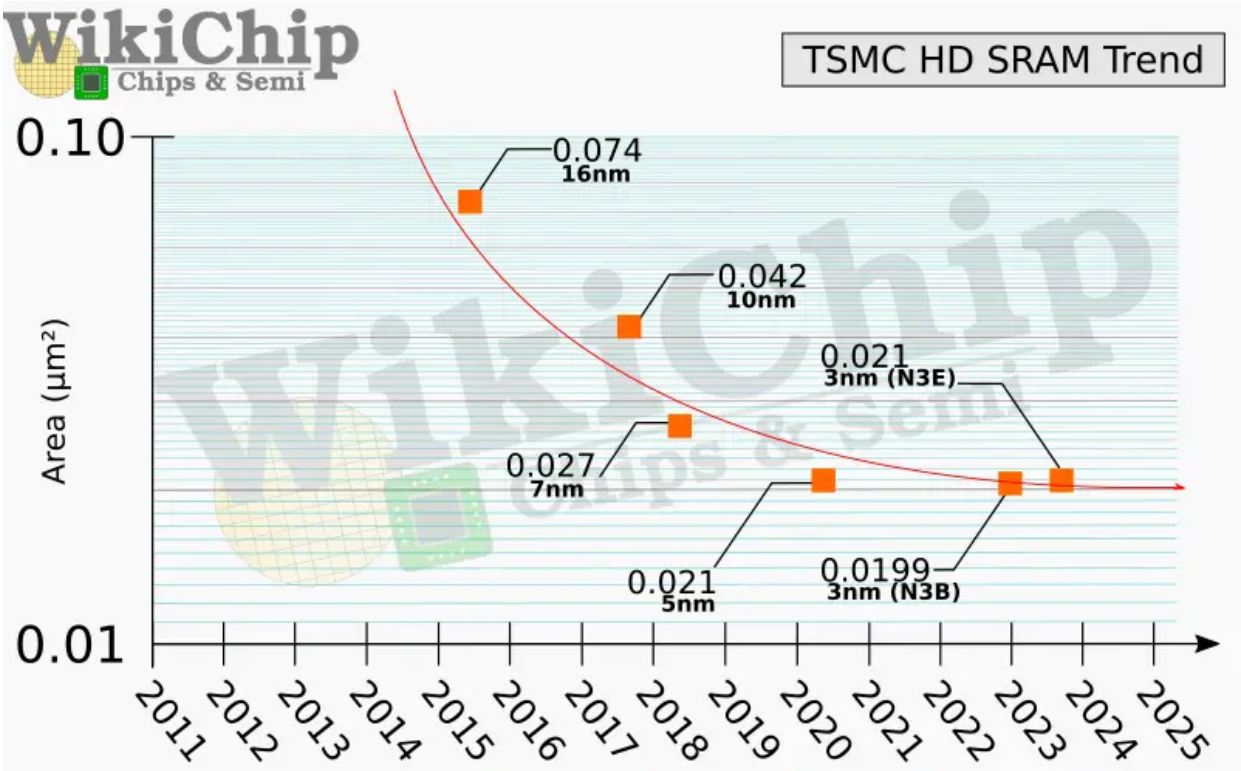


Figure 4.3: TSMC HD SRAM trend, from [4.3]. Between 2020-2024, the high density SRAM bit-cell area footprint trend has stayed roughly flat.

4.2 3T NDR FeFET SRAM Bitcell

Negative differential resistance (NDR) devices previously have been proposed for more compact implementation of SRAM bit-cells [4.4], [4.5]. The most common architecture, first proposed by Goto [4.6] consists of two tunneling diodes (TDs) and a pass transistor. For a variety of NDR devices, including most TDs, the current first increases with increasing applied voltage, reaching a peak value, then decreases with increasing applied voltage, exhibiting negative differential resistance and reaching a minimum ("valley") value. A key figure of merit for NDR devices is the ratio of the peak current to the valley current (PVCR). The higher the value of the PVCR, the more useful the NDR device is for variety of circuit applications by either achieving a higher speed or lower static power dissipation. Fig. 4.4 shows the basic NDR-based SRAM cell in which one NDR device (blue) serves to keep the storage node V_{SN} pulled down to V_{SS} and the other (red) serves to keep the storage node pulled up to V_{DD}.

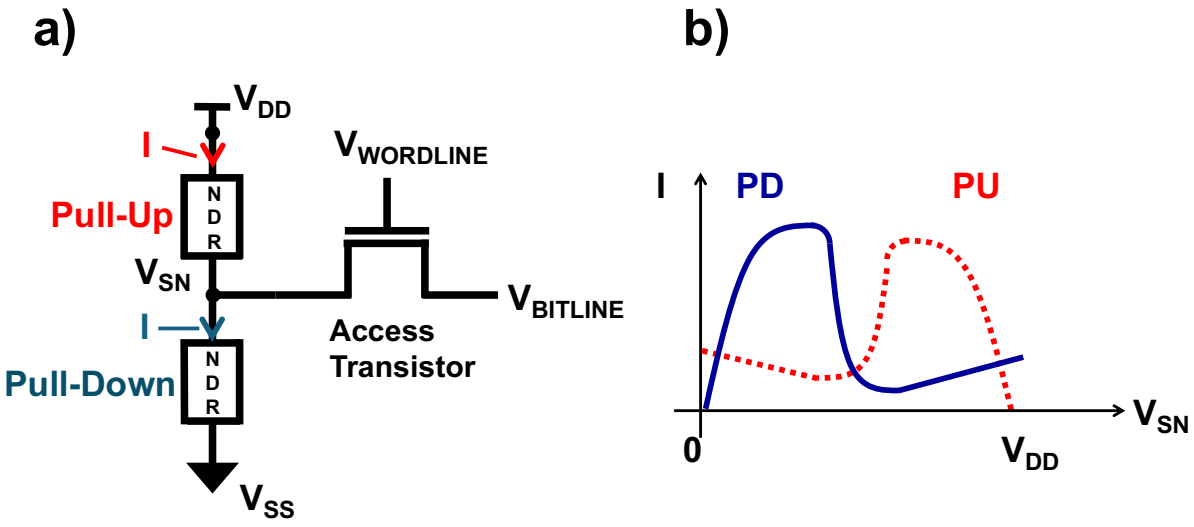


Figure 4.4a): 2NDR-1T NDR SRAM bit-cell and b) load-line diagram for the bit-cell, with the access transistor turned off.

The PVCR of TDs is generally not sufficiently large to make them practical for low-power compact SRAM application, because in order for the TDs in a Goto bit-cell to have sufficient current drive, the valley current is too large, resulting in large static power dissipation. Even if TDs could achieve PVCR large enough to implement low-power SRAM [4.5], [4.4.7], they would require specialized fabrication process sequences to integrate them monolithically with CMOS transistors used to other blocks of circuitry in a microprocessor; the increased complexity of an integrated TD-CMOS manufacturing

process would result in higher cost, offsetting the area reduction benefit of TD-based SRAM bit-cells. The NDR FeFET device proposed in chapter 3 relies on an optimized FeFET structure to achieve high PVCR NDR behavior, and can be readily integrated into a standard CMOS IC process flow.

NDR FeFET based SRAM is based on two depletion-mode NDR FeFET devices and a standard NMOS access transistor, schematically represented in Fig. 4.5. Using depletion-mode NDR-FeFETs with a native $V_{T0} \approx$ - 0.35 V and positive polarization switching voltage < 0 V (- 0.15 V) allows them to be used in a conventional TD-based memory bitcell in which the only external lines required are V_{DD}, V_{SS}, V_{BL}, and V_{WL}. Synopsys Sentaurus Device Mixed-Mode simulations are used to elucidate NDR FeFET memory cell behavior; the baseline ultra-thin-body FDSOI device studied in Chapter 3 (see chapter 3, Fig. 3.13) is used for this study. The pull-up (PU) and pull-down (PD) device are minimum-sized, with W/L = 11 nm/11 nm.

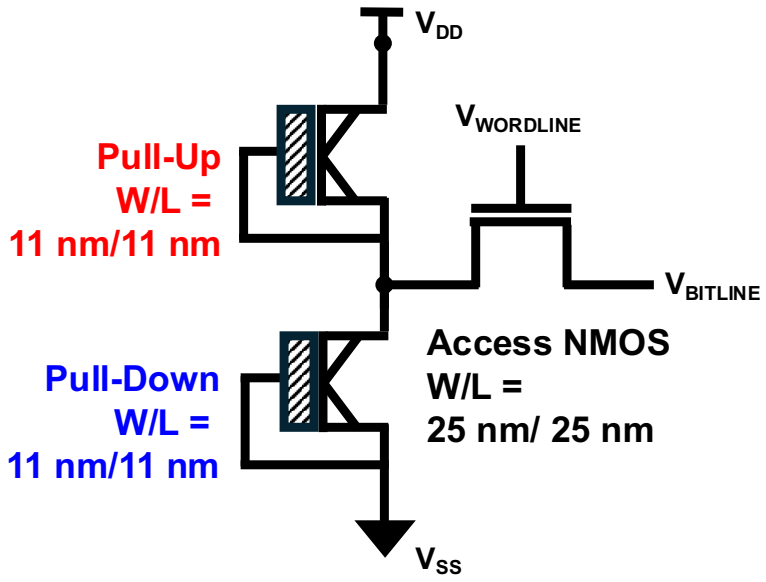


Figure 4.5: 3T NDR FeFET-based memory bit-cell.

The access transistor is an NMOSFET with the design parameters shown in Fig. 4.6. Alternatively, an enhancement-mode NDR FeFET could be used for this purpose, which might further increase the bitcell density by eliminating any required additional spacing between NDR FeFETs and NMOSFETs.

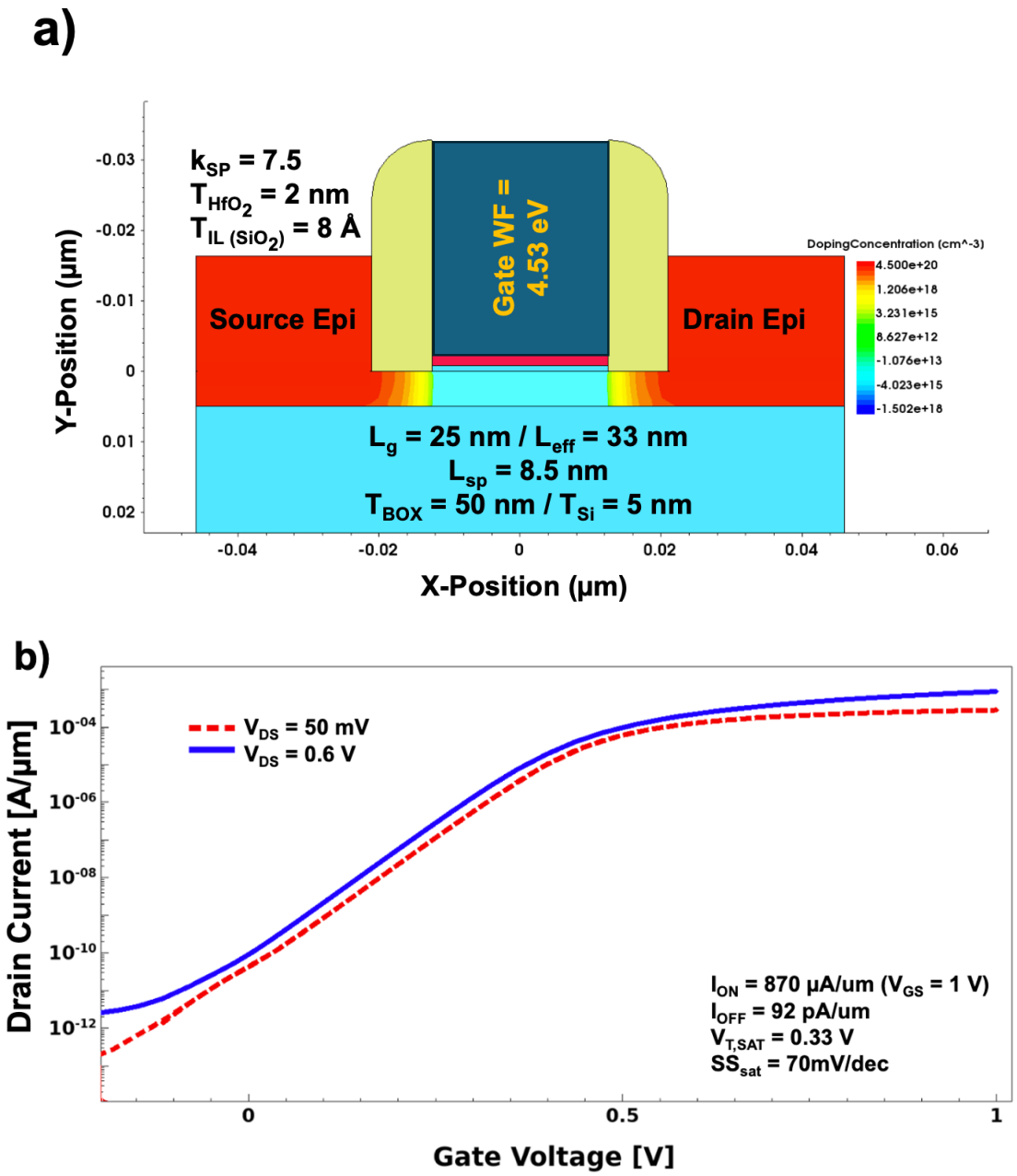


Figure 4.6: NMOS Access Transistor (a) structure and (b) transfer characteristic.

The annotated load-line below illustrates the two stable states of the NDR latch near 0 V and V_{DD}, for a V_{DD} of 0.7 V. The PU (red) and PD (blue) devices, while identical in structure, have slightly asymmetrical IV characteristics. Since they share the same N+ body well which is held at 0 V, the PU device is reverse body biased when the storage node is greater than 0 V, which slightly reduces its peak current. The peak voltage in the forward sweep direction is annotated as $V_{PEAK,\ PD\ (PU)}$, whereas the peak voltage in the reverse sweep

is annotated as $V_{PEAK,REV,PD\ (PU)}$. The FE polarization abruptly switches at the sharp jumps in current, indicated with the upward (positive switch) and downward (negative switch) arrows. As discussed in the previous chapter, the NDR FeFET devices have a high current path when the voltage is swept from 0 V towards V_{DD}, and a low current path when the voltage is swept from V_{DD} towards 0 V.

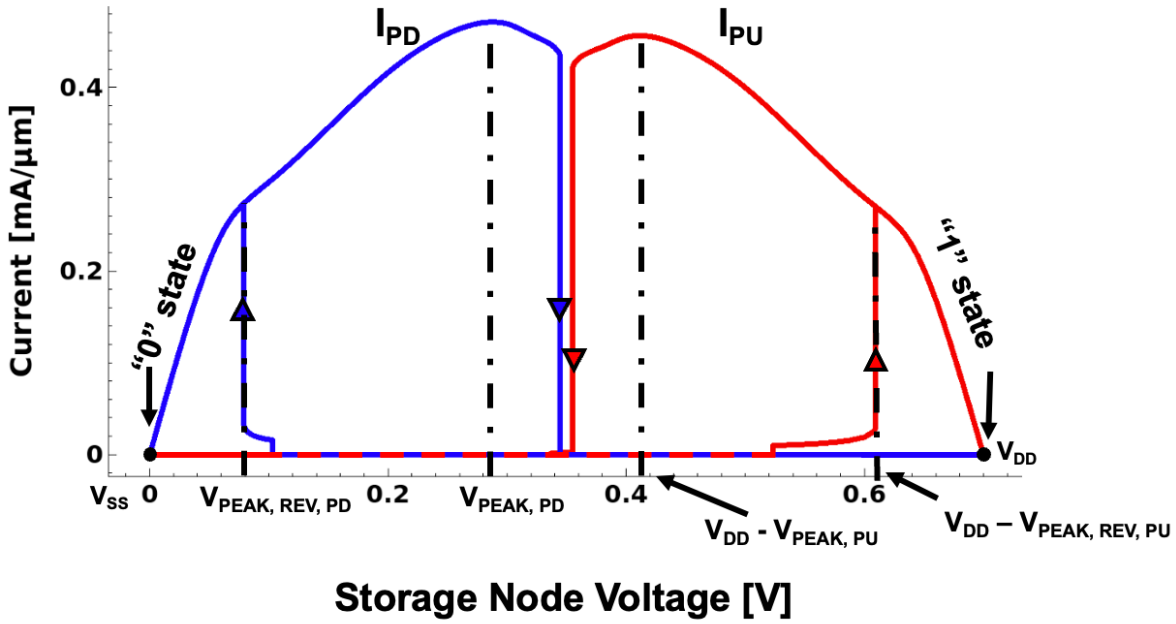


Figure 4.7: Load-line for a 2 NDR FeFET-based latch. A V_{DD} of 0.7 V is used for clarity.

For an NDR based latch to achieve bistability (i.e., one device operating in the valley region, the other in the peak region), V_{DD} must be greater than approximately $2 \times V_{PEAK}$ [4.8]. For this discussion in section 4.2, V_{DD} = 0.7 V is used for illustrative purposes, but for the rest of the chapter, V_{DD} = 0.6 V is considered in order to reduce the valley current and static power dissipation. Additionally, it will be shown that, after cell initialization, the V_{DD} for NDR FeFET based bitcells can be lowered even further to reduce power dissipation, due to the NDR FeFET's hysteretic characteristic.

Chiefly different than the NDR SRAM load-line of Fig. 4.4b, the valley currents are not visible on the linear scale; this is a "Λ" IV characteristic, characteristic of NDR transistors [4.9] and gated tunnel diodes [4.5]. Unique to the NDR FeFET, the IV characteristics are hysteretic; the peak voltage (and current) is significantly larger for the forward-sweep (V_{DS} 0 V → V_{DD}) than the return-sweep (V_{DS} from V_{DD} → 0 V). To analyze NDR

FeFET latch operation, it is necessary to consider whether the devices' V_{DS} voltages are being swept in the forward or reverse directions.

When the cell is first initialized (powered-on), the voltage supply rail is swept up to V_{DD}. In this case, the V_{DS} of both devices is increasing from zero, so both devices traverse the high current path, illustrated in Fig. 4.8.

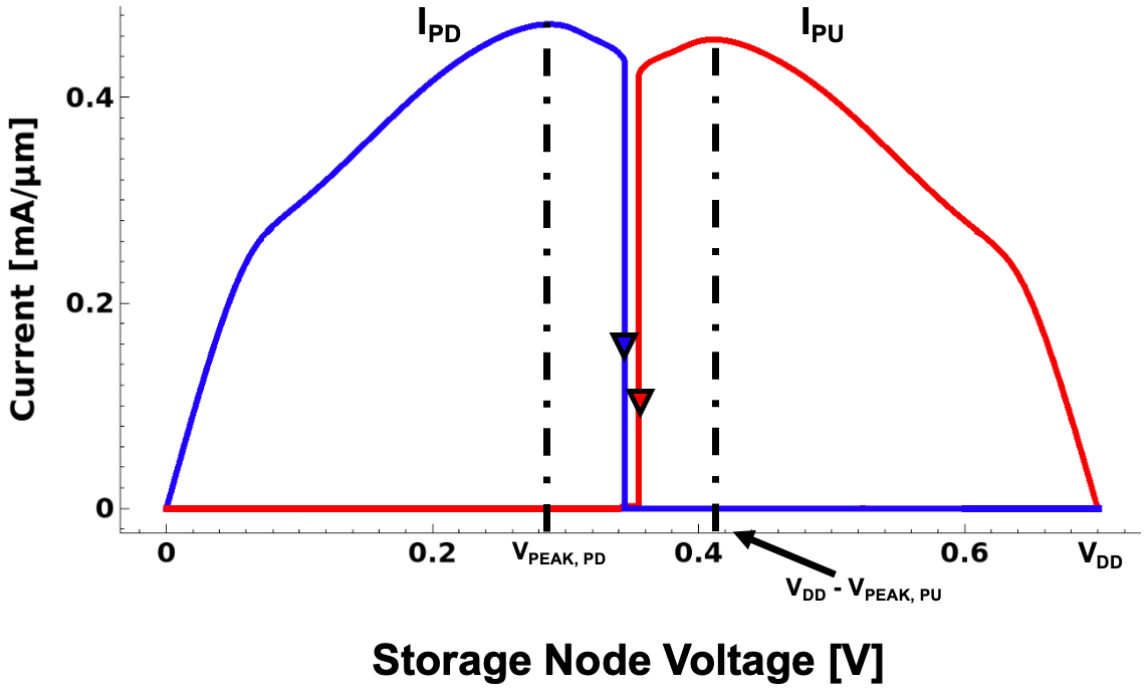


Figure 4.8: Effective Load-line during cell Power-up.

When '0' is latched, the effective loadline appears as below (Fig. 4.9). The PD device drops $V_{DS} = 0$ V, and PU device drops $V_{DS} = V_{DD}$. In other words, the PD device is in its high-current, positive polarization (peak) state, while the PU device is in its low-current, negative polarization (valley) state. So, during a read option, the PD device can pull down the bitline voltage with a high current up to its peak voltage (0.28 V) instead of its lower current reverse peak voltage (0.08 V), maintaining a large read static noise margin.

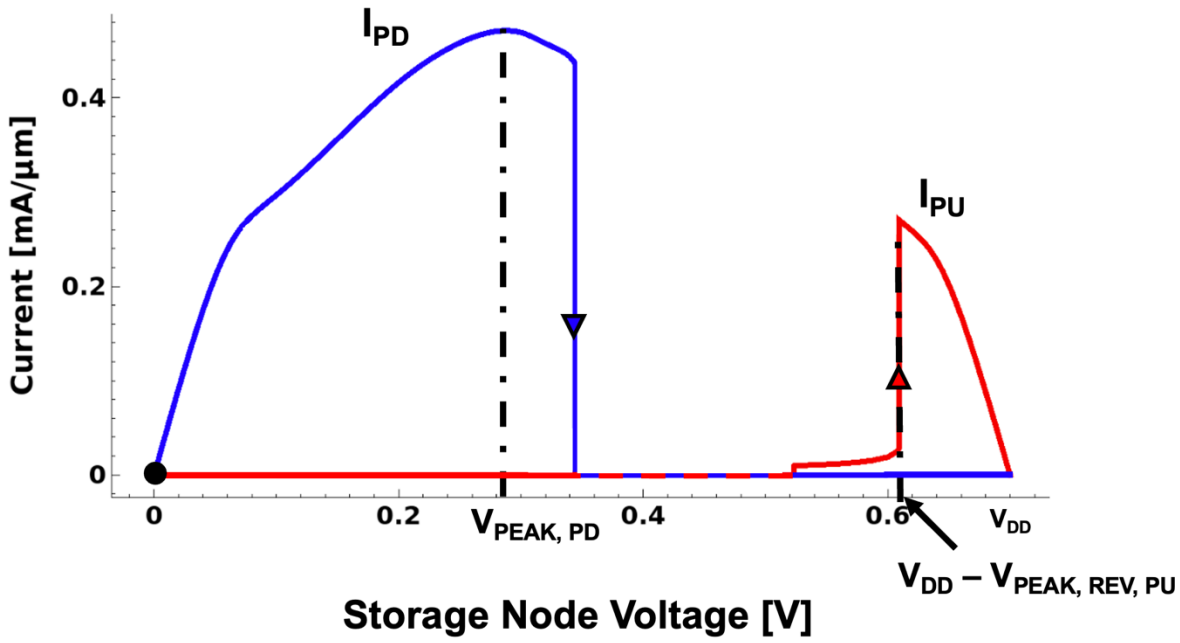


Figure 4.9: Load-line for the 2 NDR FeFET-based latch when the storage node is near 0 V.

The opposite is true when '1' is latched at the storage node (Fig. 4.10). In this case, the PU device is in the high current state. So, even though the NDR FeFET's hysteretic IV characteristic has a low-peak current return path, this hysteresis does not degrade the read noise margin for either the PD or PU devices.

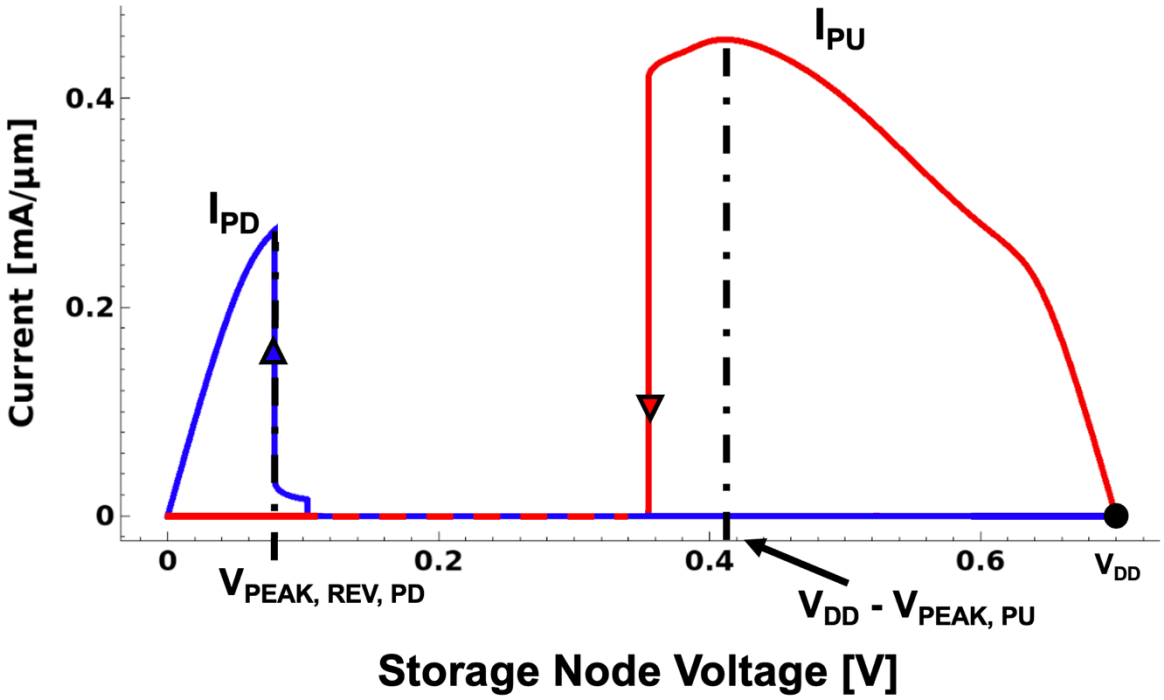


Figure 4.10: Load-line for the 2 NDR FeFET-based latch when the storage node is near V_{DD}.

4.3 NDR FeFET SRAM Bitcell Operation

In this section, standard NDR SRAM operations are described.

Dynamic Characteristics of the NDR FeFET

Unlike the previous chapter, transient simulations are now performed on short timescales rather than quasi-DC timescales to benchmark the operating speed of the NDR FeFET SRAM cell. To model the dynamic response of the NDR FeFET, the viscosity factor experimentally measured by Chatterjee et al. [4.10] was used. Fig. 4.11 plots the transient I_D-V_{DS} characteristic of the NDR FeFET for various sweep times. The sweep time was varied 6 decades from 100 ns to 1 ps. In all cases, the voltage was held at V_{DD} for 1 ns.

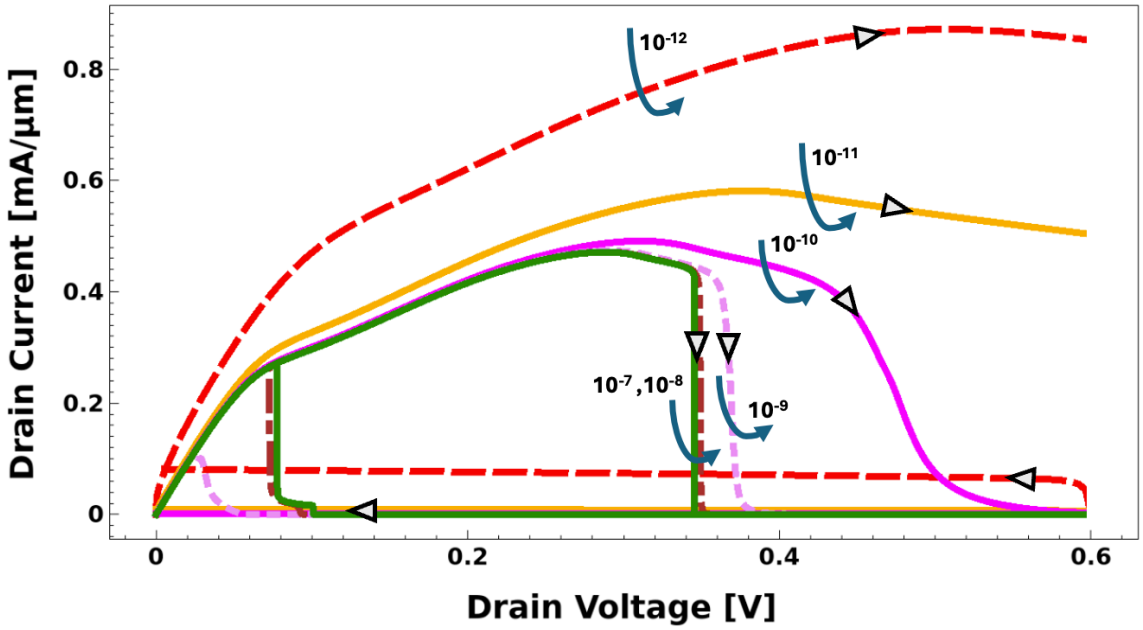


Figure 4.11: NDR FeFET I_D-V_{DS} characteristics as sweep speed is increased from 100 ns to 1 ps. The sweep direction is indicated with the grey arrowheads.

Below 10 ns sweep time, the IV characteristic becomes “smeared” in both the forward and reverse sweep directions due to the finite polarization switching time of HZO. Below 10 ps sweep time, the dynamic NDR behavior almost completely disappears. However, in all cases, the currents reduced to the same valley state current after 1 ns hold time (not shown). In the reverse sweep direction, only capacitive displacement current is visible below approximately 1 ns; however, the FE switches back to the positive polarization

state when V_{DS} is held at 0 V for 1 ns. Since the NDR FeFET does dynamically exhibit a giant NDR characteristic on the 100 ps timescale, it is reasonable to expect that an NDR FeFET SRAM cell will be competitive with the write speed requirements of state-of-the-art SRAM.

SRAM Initialization

An NDR based SRAM bitcell can initialize to the '0' or '1' state, depending on which device has a higher peak current, since a series NDR device pair is what's referred to as a MOnostable-BIstable transition Logic Element (MOBILE) [4.11]. When the supply rail of an NDR device pair begins increasing from 0 V ($V_{SN} < V_{PEAK}$), both devices are in the low resistance state (Fig. 4.12, leftmost panels). Once the supply rail reaches approximately 2 × V_{PEAK} the device with the lower peak current limits the current flow, then switches into the NDR region, so the opposite device pulls the storage node towards its rail (Fig. 4.12, middle panels). Above this voltage, once the positive and negative differential regions of the NDR devices intersect (Fig. 4.12, rightmost panels), the NDR pair is fundamentally bistable.

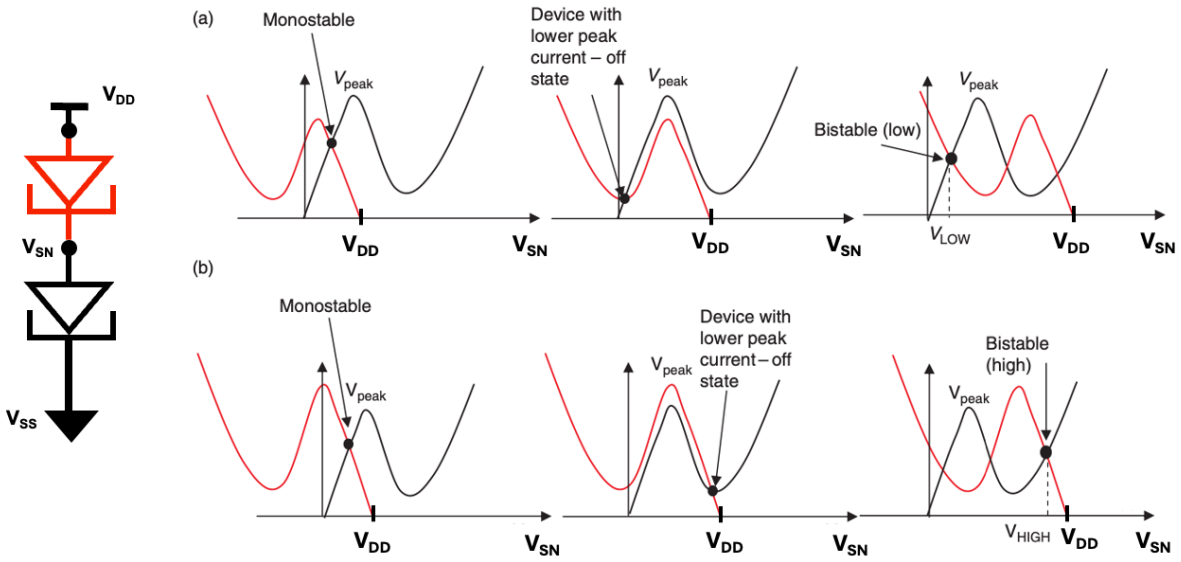


Figure 4.12: Load-Line illustration of NDR-based latch initialization with tunnel diodes. In case (a), the PD device has a larger peak current, so the cell latches to '0'. In (b), the PU device has a larger peak current, so the cell latches to '1'. Adapted from [4.8].

For the minimum sized NDR FeFET SRAM bitcell, the PU device has slightly lower peak current due to its reverse body bias (Fig. 4.8). Therefore, the cell initializes to the '0' state and quickly settles to its standby state, illustrated in Fig. 4.13:

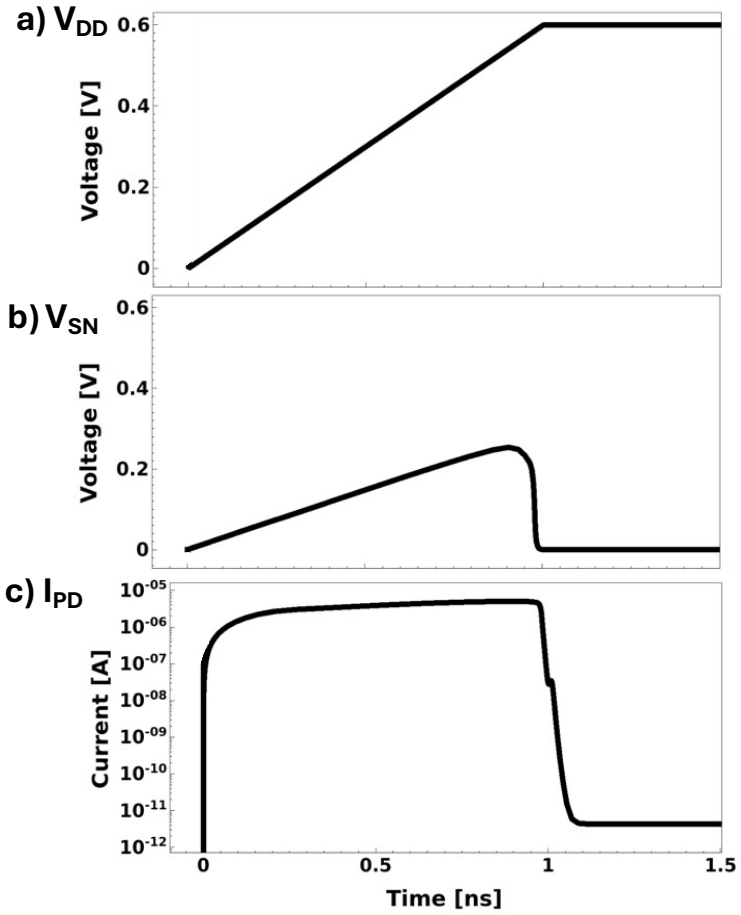


Figure 4.13: NDR FeFET SRAM initialization.

SRAM Write Operation

To write the bitcell, the bitline is biased at either V_{DD} or 0 V, and the word-line is pulsed to a positive voltage V_{WRITE}. The write '0' operation load-lines of a generic NDR MOSFET-based latch are graphically shown below in Fig. 4.14. The dot indicates that the value of the storage node is near V_{DD}; '1' is latched. In order to write a zero to the cell, the cell's bistability must be momentarily broken such that the only a stable state at V_{SN} = 0 V. During a '0' write, the bitline is held to 0 V, and word-line to V_{WRITE}. Now, the cell effectively has 1 NDR FeFET pulling up the storage node, but 1 NDR FeFET + 1 NMOSFET pulling down the storage node in parallel. Together, the load-line is transformed to have the characteristic of Fig. 4.14b, and the storage node discharges through the PD device and access FET to 0 V.

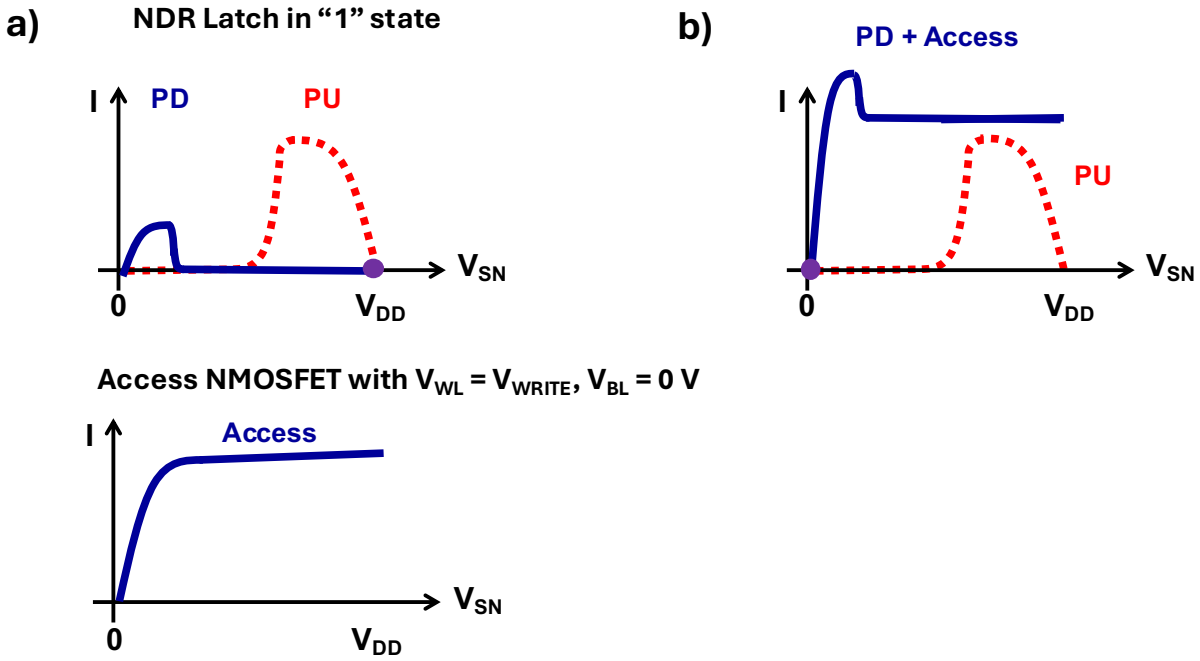


Figure 4.14: Graphic demonstrating the write '0' operation for NDR-based SRAM.

The NMOS access transistor passes a ‘0’ easily, but it is more difficult to pass ‘1’. The write ‘1’ operation is graphically represented in Fig. 4.15. As the storage node charges up from 0 towards V_{DD}, the gate-to-source voltage of the access transistor will drop from V_{WL} to V_{WL}-V_{SN}, reducing its current drive strength, so it becomes more difficult to pull-up the storage node as the storage node voltage rises.

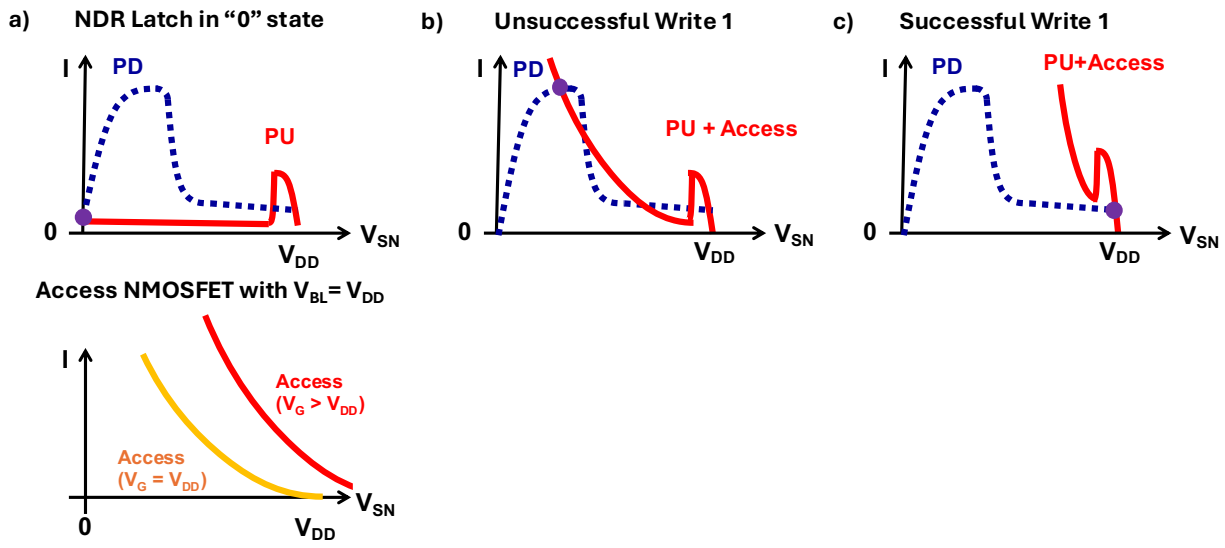


Figure 4.15: Graphic illustrating the write '1' operation for NDR-based SRAM. The valley currents of the NDR devices are exaggerated to display the operating point.

In the case of Fig. 4.15b, the write operation would not be successful since the peak current of the PD NDR FeFET is not exceeded. Since the NDR FeFET has a large peak current (> 400 μA/μm, similar to the ON-state current of the NMOS access transistor), the access transistor's area and/or gate overdrive must be increased to provide enough current to pull the storage node past the pull down device's peak voltage. Even if the drive current is increased such that only the stable state at V_{SN} = V_{DD} exists (Fig. 4.15c), it must be able to charge the storage node to that point quickly (< 100 ps). So, in general, the gate overdrive voltage of the access transistor must be somewhat greater than the V_{DD} voltage used for the NDR pair latch, as the overdrive will decrease towards 0 V as the storage node charges up. Here, the word-line voltage bias was increased to 1 V to achieve the sufficient V_{OV} required to overcome the peak current of the PD NDR FeFET.

In the following sections, transient waveforms are shown to describe SRAM operations. A transient diagram showing a write '1' operation, followed by a write '0' operation, is shown in Fig. 4.16. It is clearly seen that the write '0' and write '1' operations are asymmetrical; the write '1' operation requires 10s of picoseconds to increase the storage voltage to V_{DD} while the write '0' operation seems to happen instantaneously (< 5 ps).

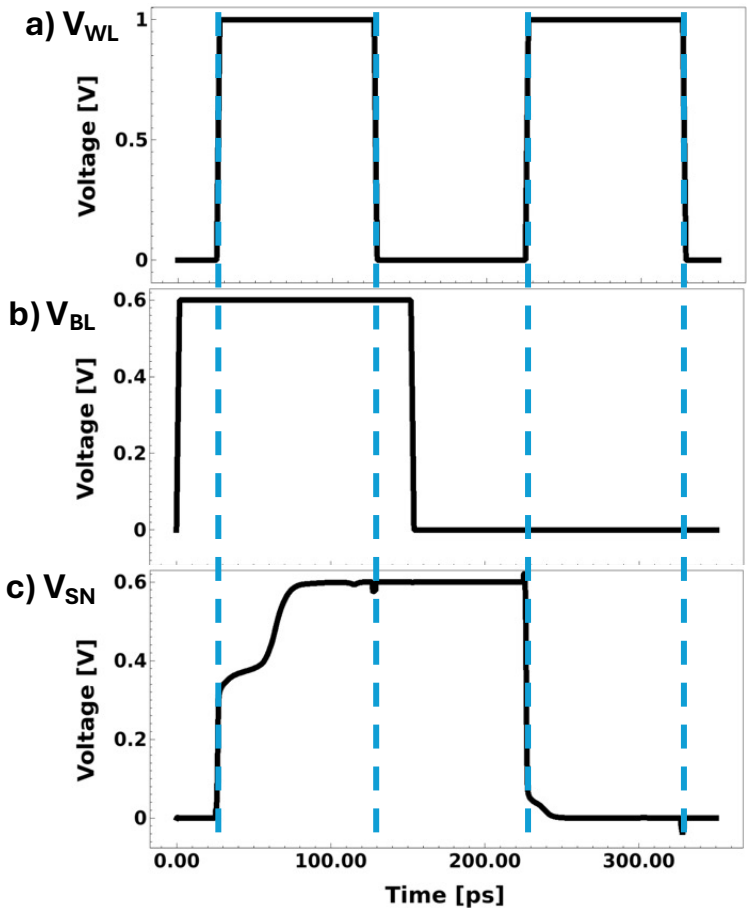


Figure 4.16: Write '1' and Write '0' operations for the NDR FeFET SRAM cell, for t_{write} = 100 ps.

The following compares the switching time for different levels of word-line bias. At lower word-line voltages, the switching time is limited by the electrical circuit (i.e., time to overcome the PD NDR FeFET's peak current and pull the storage node up to V_{DD}). Fig. 4.17a shows that as the word-line voltage is increased, the storage node voltage switching time seems to saturate near around 10 ps to reach 90% of V_{DD}. In contrast, Fig. 4.17b plots the net polarization of the FE layer in the PD NDR FeFET. The polarization switches significantly slower than the storage node potential (which is strongly driven by the access NMOSFET), saturating at around 35 ps to depolarize the FE layer (net polarization = 0), and about 60 ps to completely switch the FE polarization state. This polarization switching time similarly also applies to the write '0' operation; although the storage node voltage is pulled down to 0 V rapidly, it will still take around 35 picoseconds to switch the ferroelectric layer to depolarize the FE layer 0.

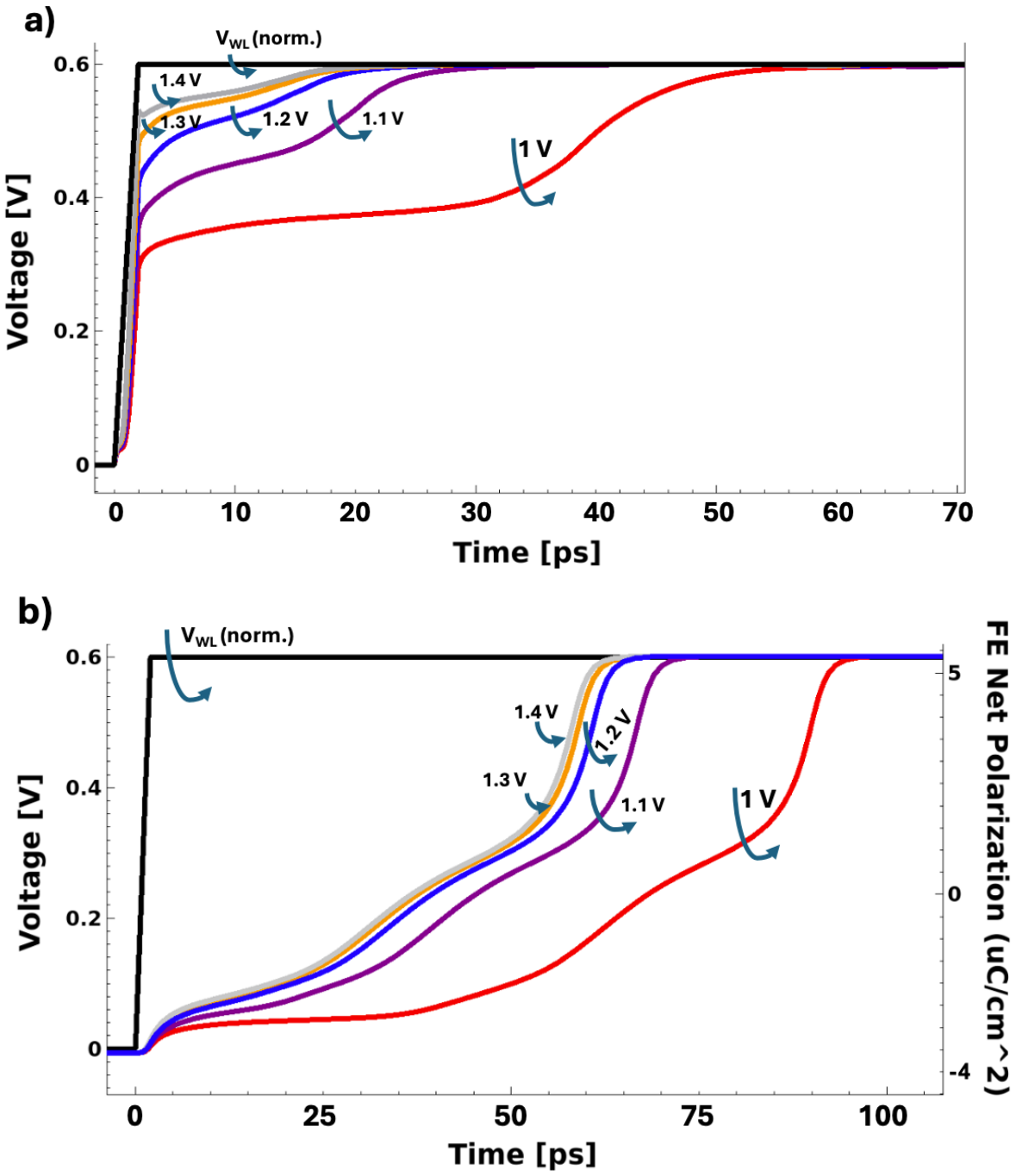


Figure 4.17a): Storage Node switching vs word-line voltage (b) FE Polarization Switching vs word-line Voltage.

For V_{WL} < 1 V, the write operation fails, as the peak current of the PD device is not overcome (as illustrated in Fig. 4.15b). The storage node potential can only rise to ~0.25 V, which is still in the high-current state of the PD device, just below the peak voltage. Key to

note is that this is, in fact, a stable operating point of the device for V_{WL} = 0.9 V, V_{BL} = 0.6 V, and will not change to another stable state unless a large perturbation pulls the storage node to a voltage greater than the negative polarization switching voltage of the PD device. This will be taken advantage of for the read operation.

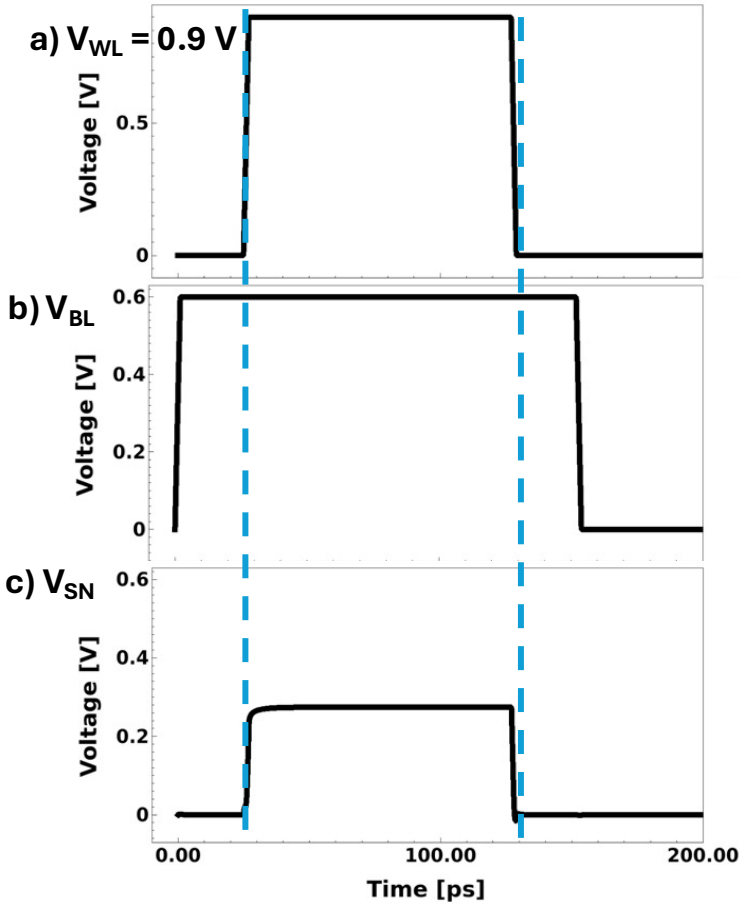


Figure 4.18: Failed write operation for V_{WL} = 0.9 V.

Based on the polarization speed in Fig. 4.17b, it should be possible to reduce the SRAM write '1' time (approx. the time taken to depolarize the FE layer) for V_{WL} = 1 V. A 60 ps write pulse width is illustrated in Fig. 4.19. Once the word-line voltage is reduced to 0 V, the storage node voltage is significantly disturbed, since the FE polarization of the NDR FeFETs is not fully switched; in other words, the NDR FeFETs are actively switching between the high and low current states. For this 60 ps write time, it takes approximately 50 more ps to strongly latch to '1', which is similar to the total switching time seen in Fig 17b.

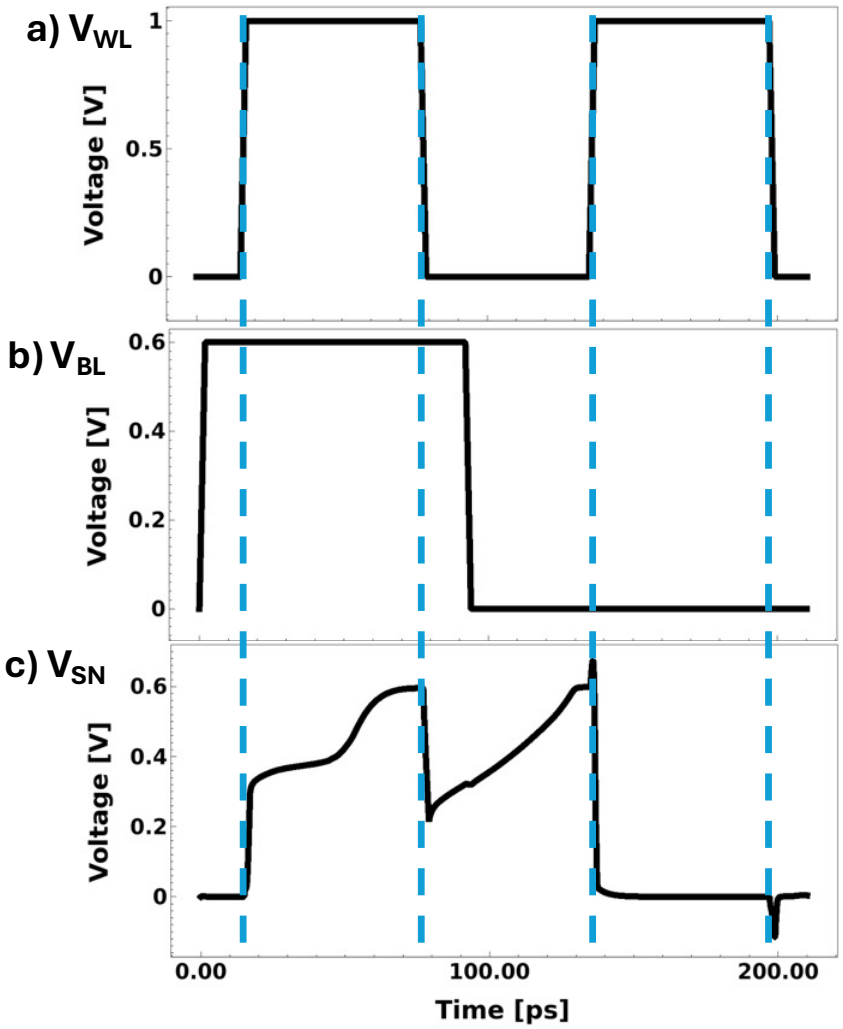


Figure 4.19: Write '1' and Write '0' operations for the NDR FeFET SRAM cell, for t_{write} = 60 ps.

Below 60 ps, the write operation fails (Fig. 4.20), as the FE polarization of the PU and PD devices cannot switch quickly enough to maintain a high enough storage node voltage required to complete polarization switching and successfully write the cell. However, the cell is not left in an indeterminate state, and fully recovers to the '0' state.

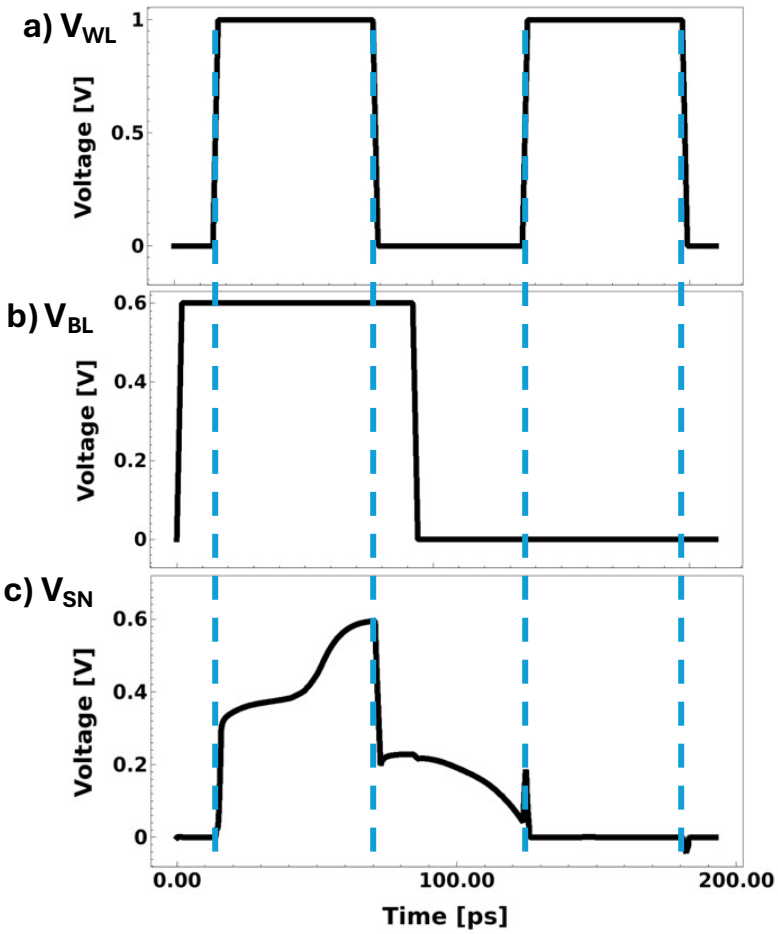


Figure 4.20: Failed write '1' operation for the NDR FeFET SRAM cell, for t_{write} = 55 ps.

SRAM Read Operation

To read the bitcell, the bitline is floated at V_{DD}, and the word-line is pulsed with a voltage V_{READ}. If the storage node is latched to '0', the PD NDR FeFET will pull the bitline down towards 0 V. In this way, the PD device can be optimized for fast read operations separately from the PU device. During the read operation, due to capacitive charge sharing between the bitline and storage node, V_{SN} will rise towards V_{DD} while the bitline is pulled down towards 0 V. Once the bitline voltage reaches a value which can be read out by the bitline sense amplifier (50 mV difference [4.12]), the read operation is completed. For the read operation, the ferroelectric does not need to switch polarization, which decouples the intrinsic time required for read and write operations; no write-after-read operation is required. This also implies that the endurance of an NDR FeFET based memory cell will not be limited by the number of read cycles. Instead of polarization switching speed, the read time is instead limited by the peak current of the NDR FeFET.

A minimum sized NDR FeFET SRAM cell has a very small cell capacitance. If the bitline capacitance is high, the NDR FeFET cell might be disturbed if the storage node voltage is pulled above the negative polarization switching voltage of the PD NDR FeFET. As previously discussed, for lower word-line voltages (< 1 V, in this case), the cell remains stable in the '0' state due to the PD device's large peak current. Therefore, a V_{READ} = 0.9 V can be used for the read operation, which guarantees no read disturb condition from capacitive charge sharing. Alternatively, the PD NDR FeFET could be sized larger to more effectively pull down the bitline voltage, or bitline capacitance (array size) may be reduced, to prevent read disturb. Figure 4.21 plots a read '0' operation for a bitline capacitance of 10.24 fF [4.12] and V_{READ} = 0.9 V. As before, V_{SN} rises to and is maintained at ~0.25 V, and the bitline is pulled down to 550 mV in 101 ps.

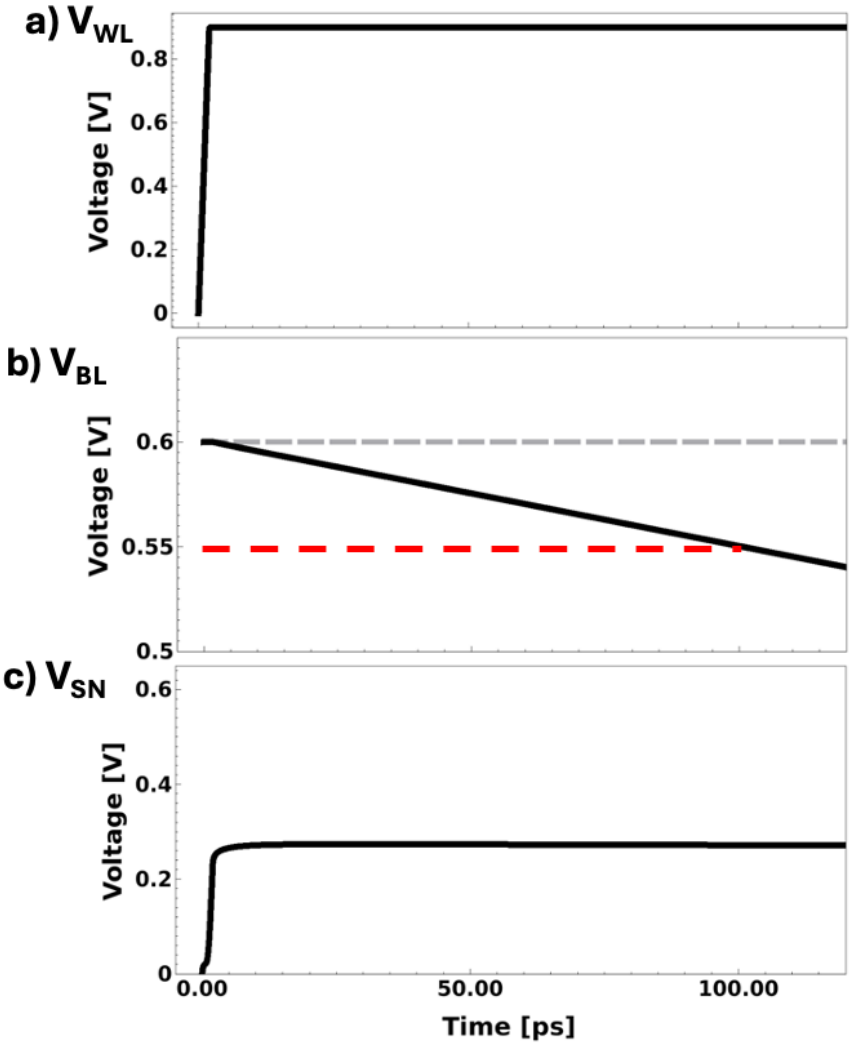
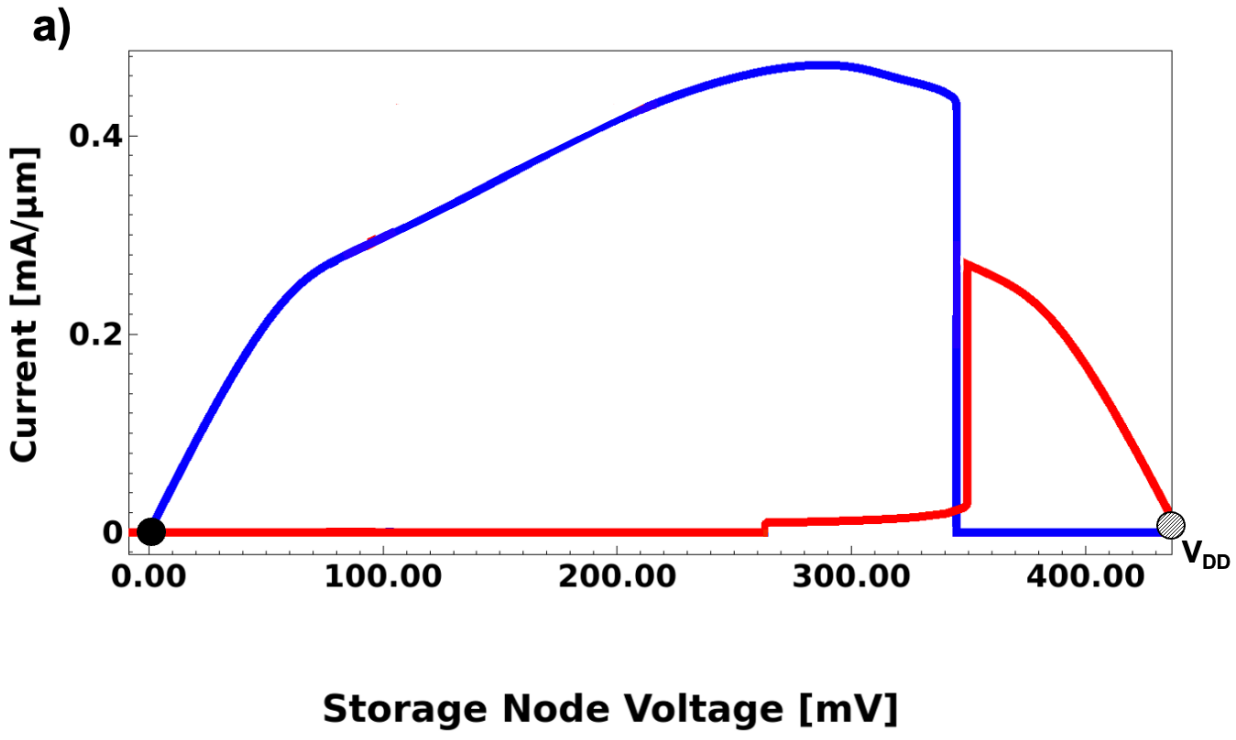


Figure 4.21: SRAM read '0' operation For V_{WL} = 0.9 V.

4.4 Low V_{min} of NDR FeFET based bit-cell

Reducing the minimum operating voltage "V_{min}" is of paramount importance in modern SRAM for reduced the power consumption. While the NDR FeFET SRAM must be initialized to a supply voltage of approximately $2 \times V_{PEAK}$, due to the hysteretic NDR FeFET IV characteristic, the cell operating voltage can then be reduced to significantly below $2 \times V_{PEAK}$. In the reverse sweep direction (V_{DS}: V_{DD} → 0 V), the peak voltage of an NDR FeFET is reduced to a value generally < 100 mV. Since, after cell initialization, one device in the NDR pair is always in the valley state (and will be swept in the reverse direction to change states), the bistability requirement of $V_{DD} > 2 \times V_{PEAK}$ becomes $V_{DD} > V_{PEAK} + V_{PEAK, REV}$. This is slightly increased to accommodate the sharp polarization switching NDR characteristic, which happens at a slightly larger voltage than V_{PEAK}, for the PU device. Fig. 4.22 below shows the load-lines for a V_{DD} of 440 mV. Since the peaks are non-overlapping in both the latched '0' and latched '1' state, this supply voltage supports bistability. Additionally, the large static noise margin is maintained (does not reduce with V_{DD}) for both the '0' and '1' states, even at the lower V_{DD}. This reduced V_{DD} mode of operation may be preferrable to the nominal voltage used in this work due to reduced GIDL current at low V_{DS}, and is left for more detailed future investigation.



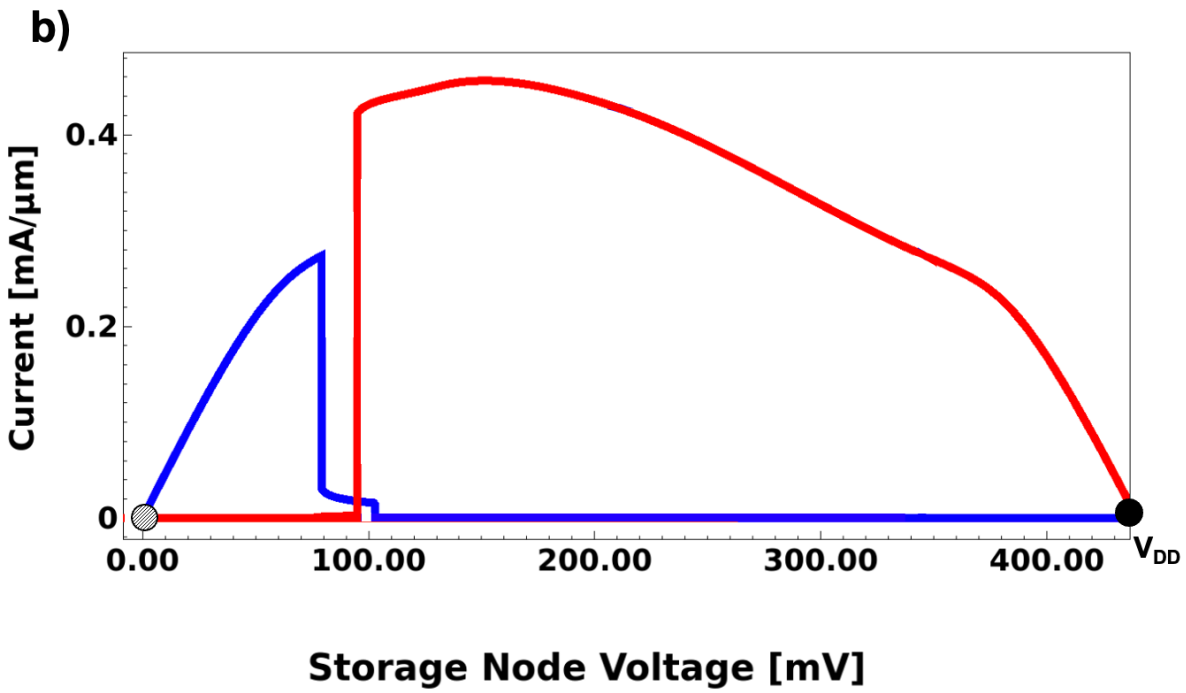


Figure 4.22 : Load-line for 440 mV operation of the NDR FeFET latch in the (a) '0' state (b) '1' state.

During retention, in order to prevent cell disturbance, the PD device cannot drive enough current to pull the storage node from V_{DD} to 0 V (or vice-versa). For conventional symmetric NDR devices, this means the supply voltage must be maintained above approximately $2 \times V_{PEAK}$. Due to the effective lower $V_{PEAK,REV}$ and $I_{PEAK,REV}$ as discussed, the supply voltage for an NDR FeFET can be reduced near to $2 \times V_{PEAK, REV}$ before the latch's bistability is broken. This enables a large supply-static noise margin. In fact, due to the asymmetry of the IV characteristic (gradual in the positive differential resistance state, abrupt in the negative differential resistance state), V_{DD} can be reduced to even less than $2 \times V_{PEAK, REV}$, as demonstrated in Fig. 4.23. Fig. 4.23 below is a load-line plot demonstrating a low standby V_{DD} retention for an NDR latch in the '1' state. A similar operation could be done for an NDR latch in the '0' state, with a similar lower-limit, since the PU and PD devices have similar $V_{PEAK, REV}$. Since the storage-node static noise margin is bound to V_{DD}-$V_{PEAK, REV}$, this technique trades-off cell static noise margin and static power dissipation as V_{DD} is reduced.

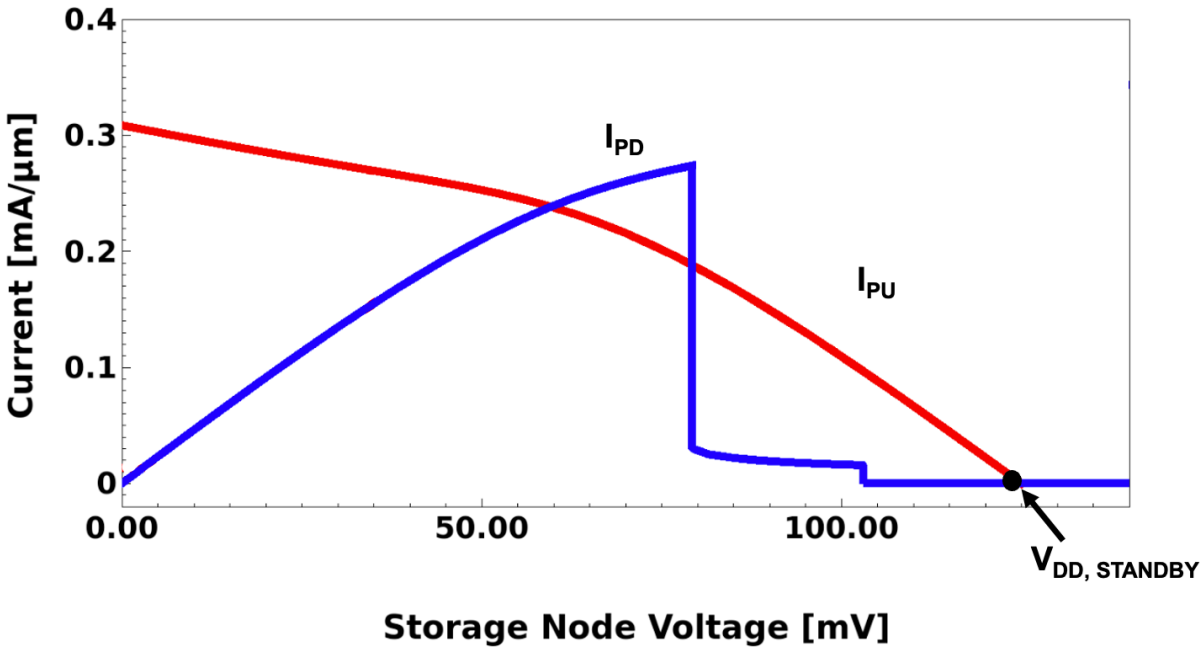


Figure 4.23: Load-Line plot for an NDR pair in the retention standby state.

Fig. 4.24 displays a transient plot of a standby and restore operation of the latched '1' state. Since this cell already initializes to the 0 state, demonstration of the standby operation for the '0' state is omitted here.

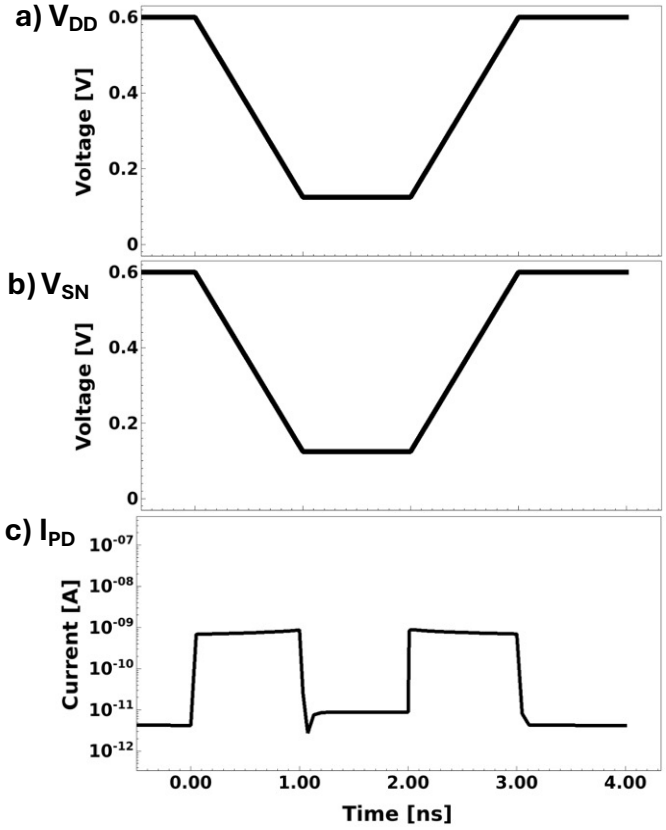


Figure 4.24: Transient waveforms for a standby/restore operation, for $V_{DD, standby}$ = 125 mV.

If the voltage is reduced to less than ~105 mV, the PD device begins to turn on (c.f. Fig. 4.23), so the leakage current is greatly increased. Fig. 4.25 plots transient waveforms for $V_{DD, standby}$ = 90 mV.

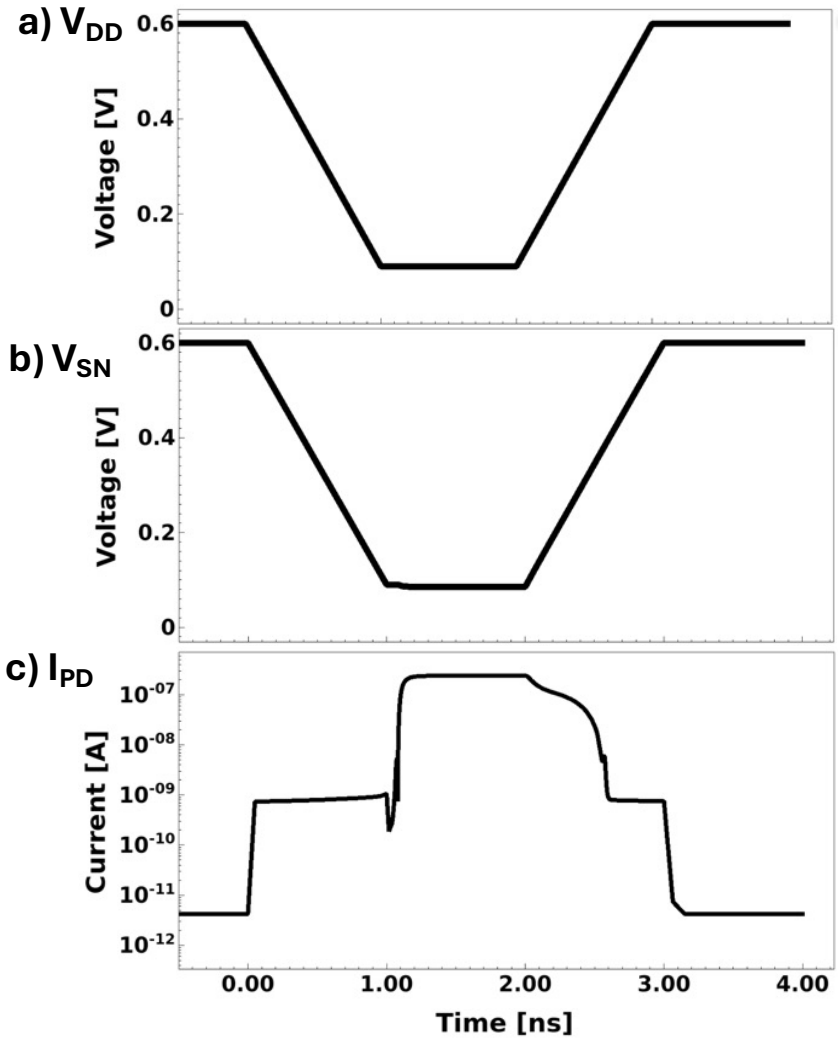


Figure 4.25: Transient waveforms for a standby/restore operation, for $V_{DD, standby}$ = 90 mV.

This increases the standby power dissipation by several orders of magnitude, although the state is still preserved. If the supply voltage is reduced even further, to 85 mV, the PD device fully turns on, and the state is lost: The cell re-initializes to the '0' state.

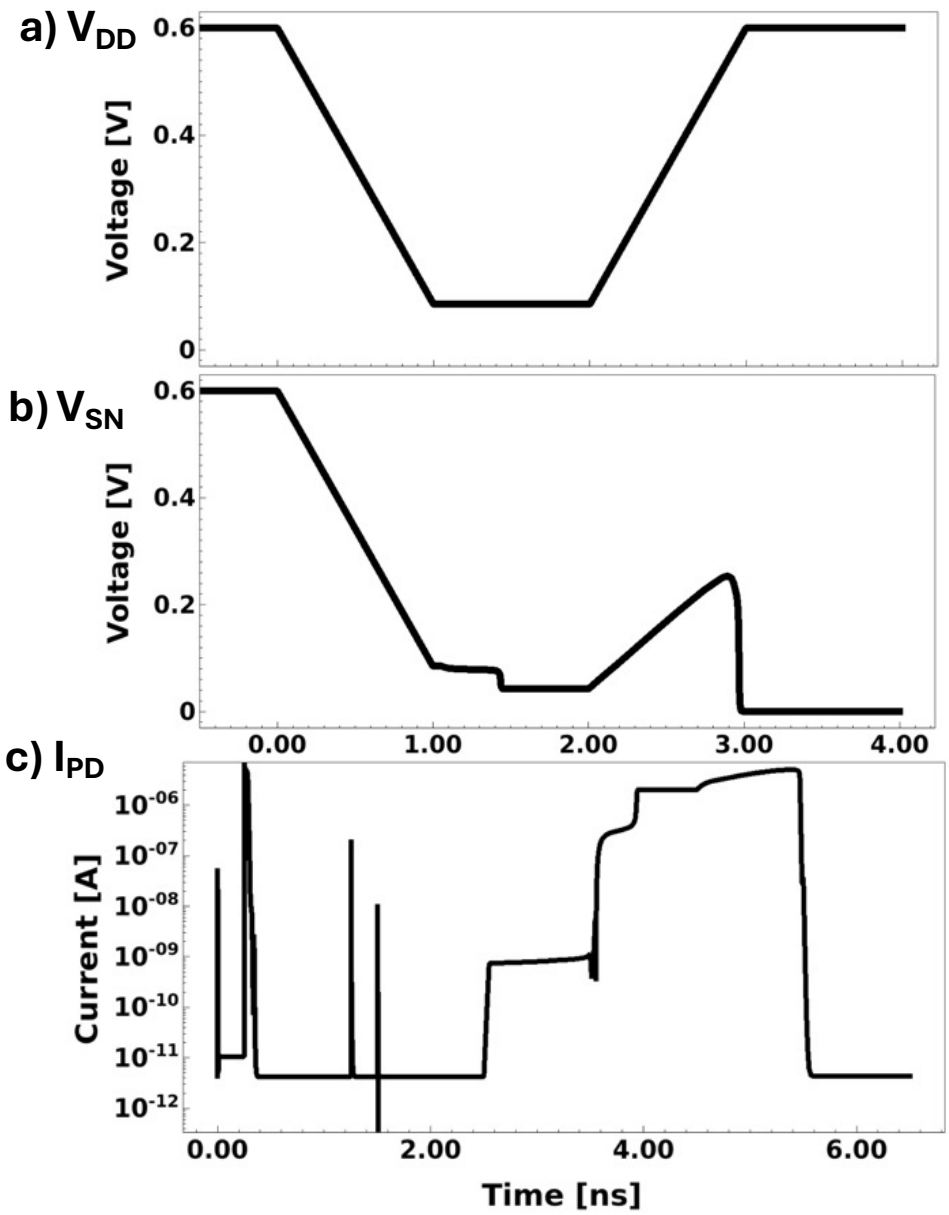


Figure 4.26: Transient waveforms for a standby/restore operation, for $V_{DD,\ standby}$ = 85 mV.

4.5 NDR FeFET Nonvolatile SRAM Operation

Emerging memory technologies based on ferroelectrics [4.13], spintronics [4.14], resistive switching , and other material systems have been explored to with the goal of being a compelling option for either embedded nonvolatile memory or Last-Level-Cache memory. In addition to the volatile memory operation explored thus far in this chapter, the hysteresis window of an NDR FeFET can be leveraged to preserve the polarization state of one or both of the NDR FeFETs in a SRAM cell. By operating the NDR FeFET in a nonvolatile mode, a save-and-restore scheme can be used to retain the latched storage node value upon power down. Fig. 4.27 shows a schematic for an NDR FeFET SRAM which can operate in a nonvolatile mode. Depending on the saved polarization state of the PD NDR FeFET (indicated to be nonvolatile in the schematic), the cell initializes to either the '1' state or the '0' state.

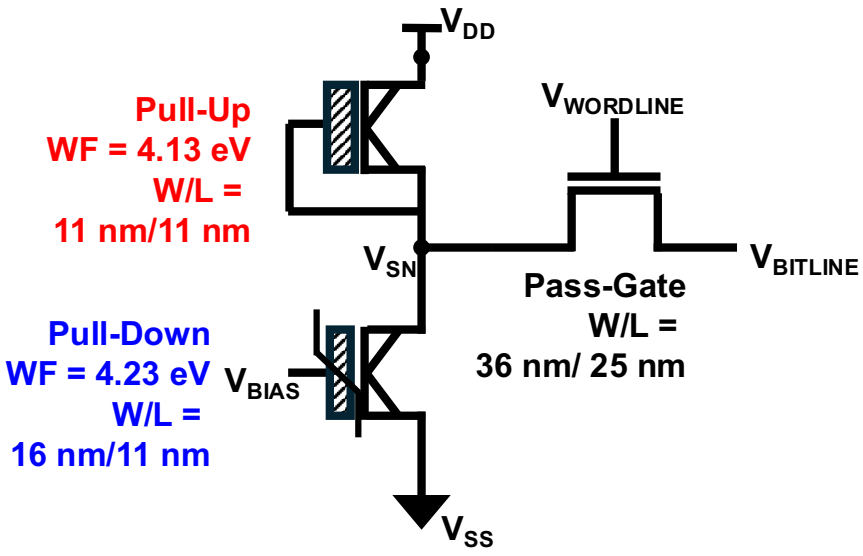


Figure 4.27: Schematic of a nonvolatile NDR SRAM cell.

By increasing the gate work function of the PD NDR FeFET, the hysteresis window could be centered about 0 gate-to-source voltage (Fig. 4.28), so that it operates in a nonvolatile enhancement-mode. A gate bias signal is used to control PD device during the save and restore operations. The sizing of devices in this cell is slightly larger than the minimum-sized NDR FeFET cell previously considered, in order to ensure polarization retention during the save and restore operations.

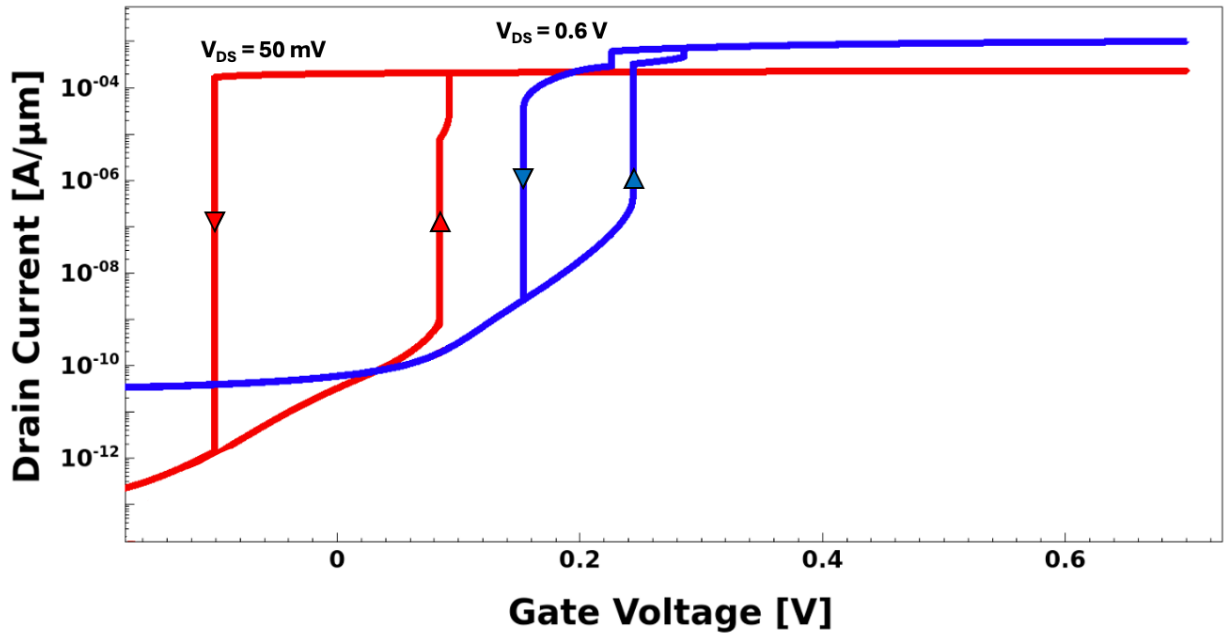


Figure 4.28: I_D-V_{GS} characteristic of a nonvolatile NDR FeFET.

During active operation, the gate voltage of the PD NDR FeFET is kept at a constant bias just above the positive polarization switching voltage, in this case, 100 mV. The shutdown operation is visually described in Fig. 4.29. To "save" the state of the cell, the gate bias is first lowered to 0 V (Fig. 4.29a), which is within the hysteresis window of the PD NDR FeFET. The storage node voltage is retained, since one NDR FeFET remains in the positive polarization state and strongly pulls the node towards its rail. Next, V_{DD} is reduced to 0 V (Fig. 4.29b). If V_{SN} = 0 V, then the voltage will be slightly raised when the PU device turns on at low V_{DS}; as long as the margin between V_{GS} = 0 V and negative polarization switching gate voltage is great enough, the PD NDR FeFET will be able to retain its polarization state. Now, the cell is powered down, and the PD NDR FeFET has retained its polarization state.

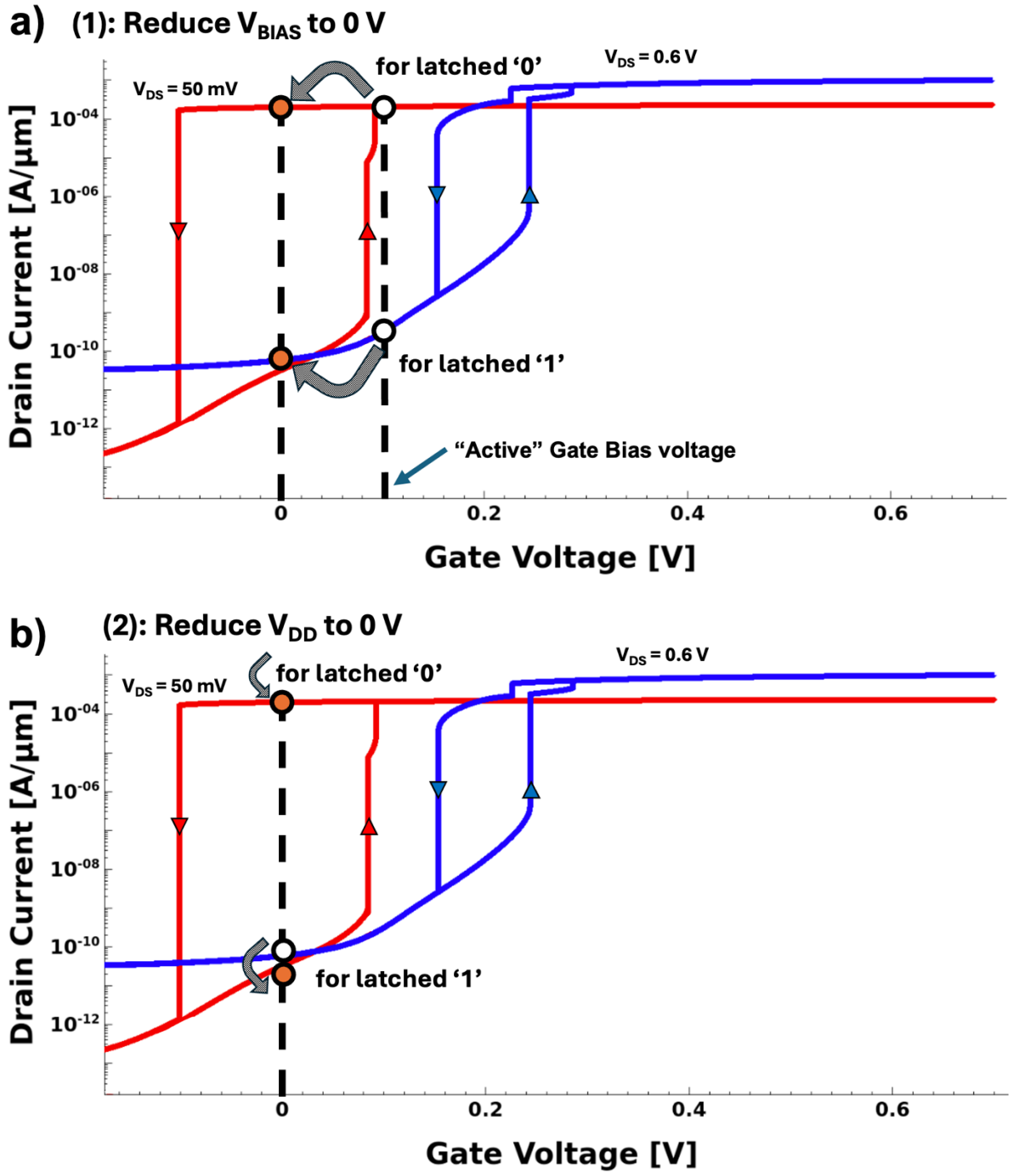
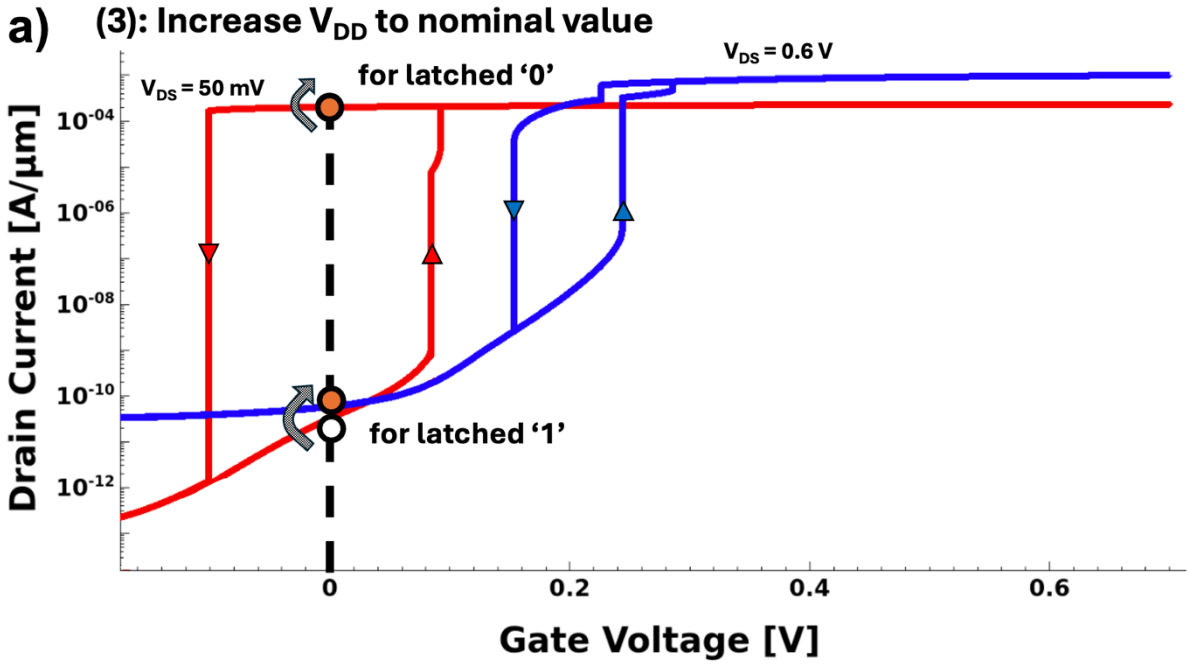


Figure 4.29: Illustrating the Nonvolatile NDR FeFET "save" operation on the I_D-V_{GS} transfer characteristics low (red) and high (blue) V_{DS}. In (a), the V_{GS} bias is reduced to 0 V for the PD device, to be within the hysteresis window for low V_{DS}. In (b) V_{DD} is reduced to 0 V, which reduces V_{DS} to 0 V. The low and high current states are preserved because the FE polarization does not switch.

To power up (“restore”) the cell, the opposite procedure is applied, visually described in Fig. 4.30. First, the supply rail is increased to V_{DD} (Fig. 4.30a). If the PD NDR FeFET is in the positive polarization state, then the PD device will be a similar conductivity to the PU device and keep the storage node pulled down towards 0 V; if it is in the negative polarization state, the PD NDR FeFET will be highly resistive, and PU NDR FeFET will pull the storage node strongly up to V_{DD}; the negative polarization state will be retained. To restore the device state for active operation, the gate bias signal is then returned to 100 mV (Fig. 4.30b).



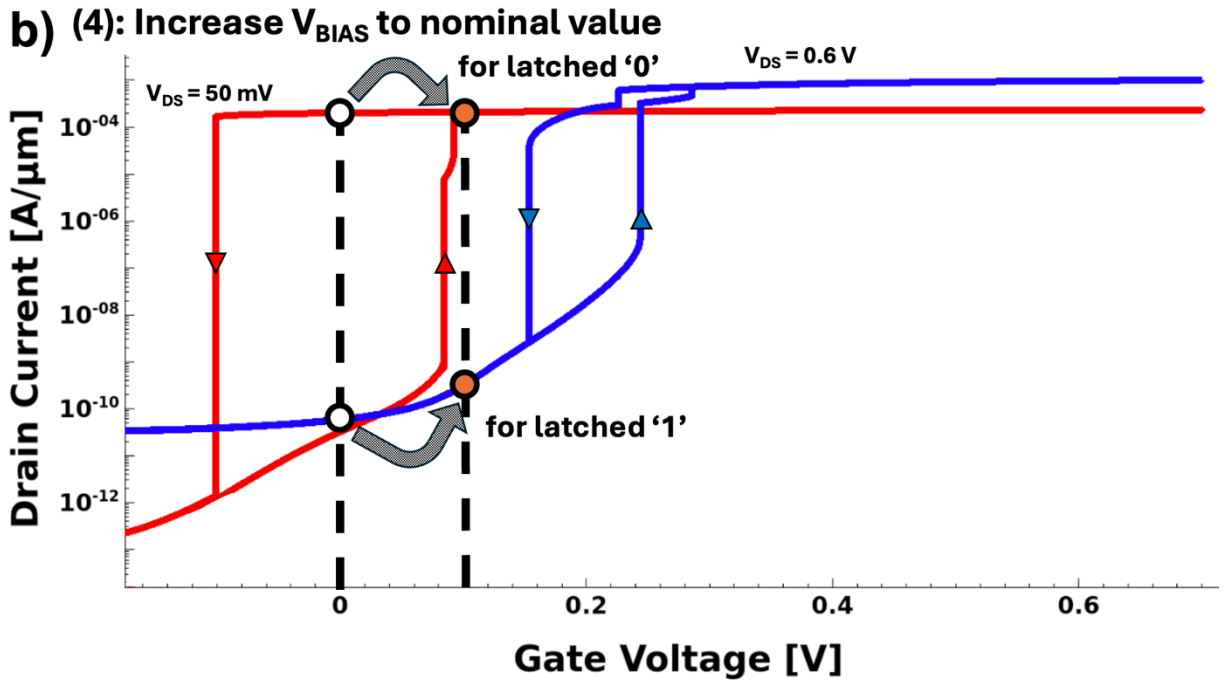


Figure 4.30: Illustrating the Nonvolatile NDR FeFET "restore" operation on the I_D-V_{GS} transfer characteristics low (red) and high (blue) V_{DS}. In (a), V_{DD} is increased to the nominal operating voltage. If the latched state was '0', the PD device's high current state is retained, whereas if the latched state was '1', the low current state is retained. In (b) the PD device's V_{GS} bias is increased to the nominal voltage, out of the hysteresis window of either V_{DS} bias. This enables successful read/write operations.

The NDR FeFET cell can restore itself in this way since the NDR pair is a MOBILE [4.11]. When the PD NDR FeFET is restored in its positive polarization state, it has a high peak current, as shown in Fig. 4.31a. When it is restored in its negative polarization state, it has very low peak current (on the order of the valley current), as shown in Fig. 4.31b.

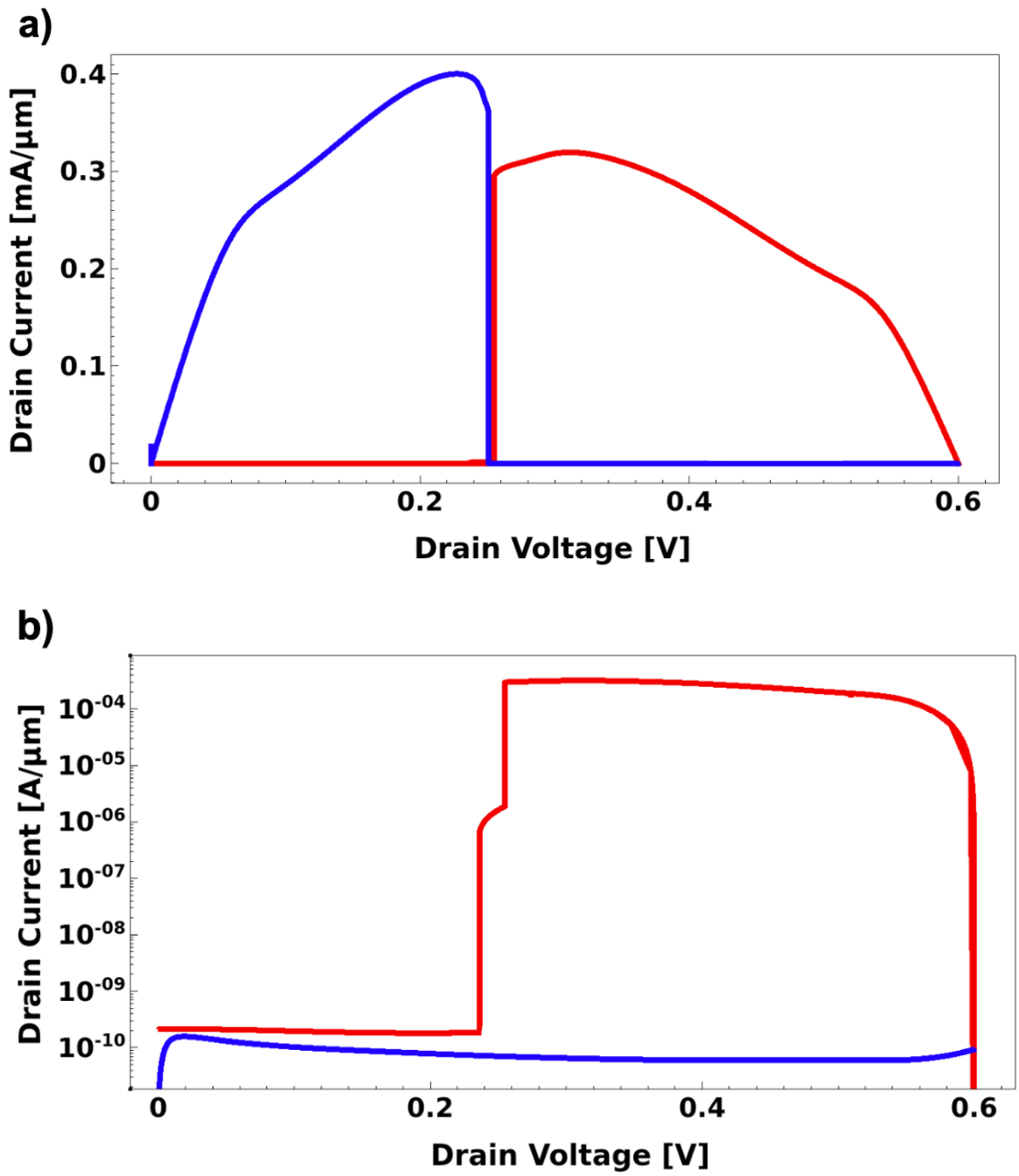


Figure 4.31: Load-line diagrams for a nonvolatile NDR FeFET cell (V_{BIAS}= 0 V) restored in the (a) positive polarization state and (b) negative polarization state.

Fig. 4.32 on the following page shows a timing diagram for the save and restore operation, for a saved '1' and saved '0'. By using the prescribed scheme, the pulldown device never switches polarization, and the state is retained. This may be used to achieve zero standby power similar to other emerging memory devices. In fact, this scheme could

substantially relax the typical requirement of high PVCR for NDR-based memory cells to less than 10 if the SRAM cell is powered-down in standby, as the array would only draw valley current when a read or write operation is performed.

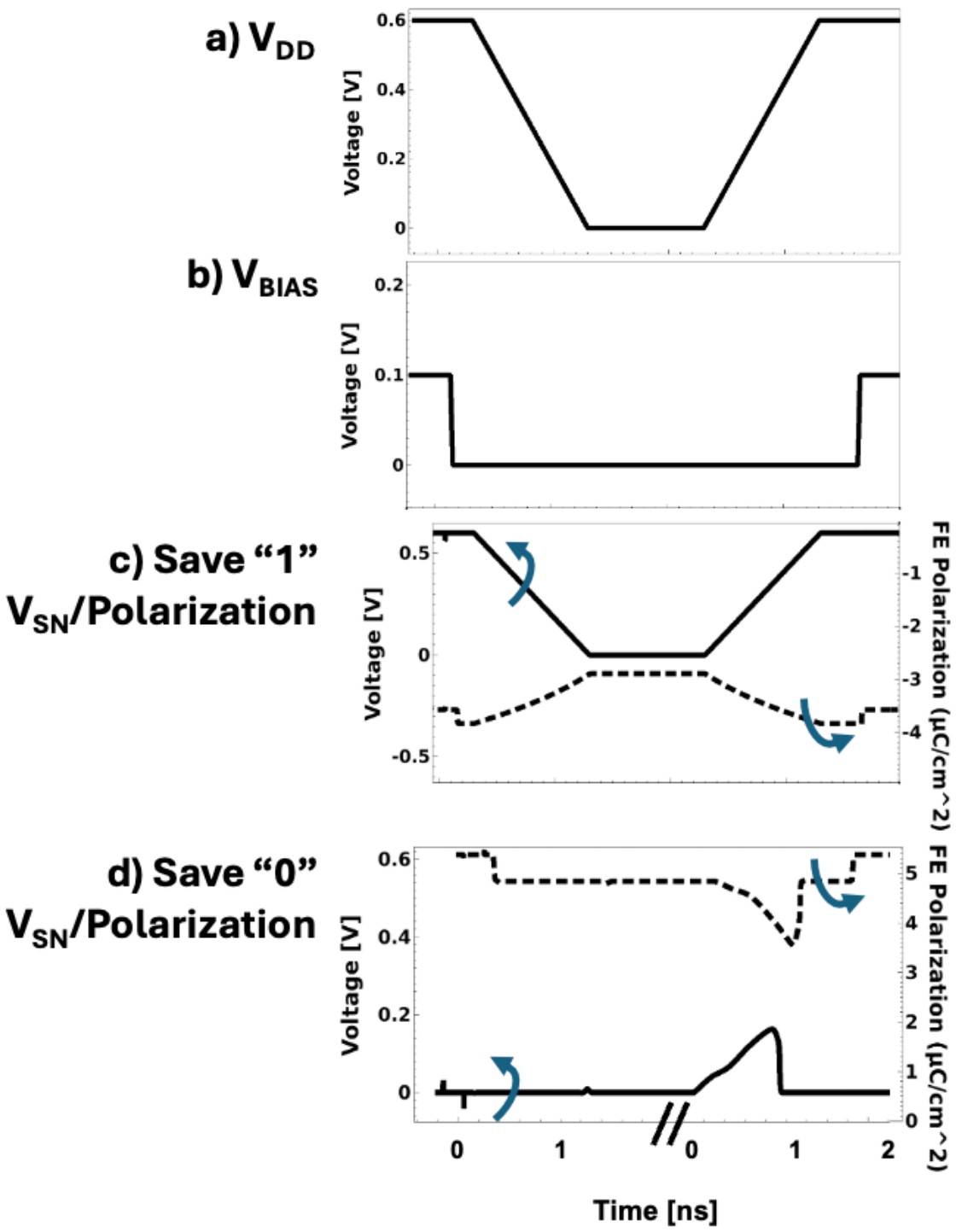


Figure 4.32: Transient waveforms for an NV NDR FeFET SRAM Save and Restore operation. Panels (c) and (d) show the V_{SN} and net ferroelectric polarization for the nonvolatile PD NDR FeFET. The orange break-line on the time-axis indicates the passage of 1 ms in the powered-down state.

4.6 Summary and Benchmarking

This chapter has presented a compact 3T SRAM bit-cell based on NDR FeFETs. When compared with the conventional CMOS-based 6T SRAM design, NDR FeFET based SRAM offers reduced device count and up to an order of magnitude lower standby power consumption, at compatible speeds. Table I summarizes the NDR FeFET SRAM performance characteristics. Note that unlike the symmetrical behavior of 6T CMOS SRAM, standby power is increased for the '1' state due to leakage through the access transistor.

Table 4.1: NDR FeFET SRAM Performance Benchmarking

	3T NDR SRAM	**6T FinFET SRAM** [4.12]
Nominal V_{DD}	600 mV	650 mV
Nominal Standby Power ('1')	6.18 pW	26 pW
Nominal Standby Power ('0')	2.52 pW	26 pW
Nominal Read Speed	101 ps	38.4 ps (access time)
Nominal Write Speed	60 ps (V_{write} = 1 V)	8.5 ps (access time)
Standby Static Noise Margin	345 mV	140 mV (read SNM)
Active V_{min}	440 mV	490 mV (read)/ 730 mV (write)
Active V_{min} power ('1')	2.1 pW	26 pW
Active V_{min} power ('0')	1.29 pW	26 pW
Retention V_{min}	125 mV	490 mV (read)
Retention Power ('1')	1.41 pW	26 pW
Retention Power ('0')	1.09 pW	26 pW

There are many similar embedded memory approaches currently under investigation, such as FE capacitors for high-density DRAM [4.15], FeFETs for nonvolatile [4.16], [4.17] memory, and high PVCR tunnel diodes [4.18]. NDR FeFET-based circuits, due to their versatility, provide a rich new set of options for IC designers to consider. Compared with ultimately-scaled FE memories, it does not suffer the same read/write disturb issues as the FE-based architectures [4.19], [4.20], though these can provide for higher density. It also does not require additional process complexity such as the integration of a BEOL capacitor [4.15]. However, like most FE-based devices, continuous advancements must be made to increase the write-endurance, which is limited by FE wear-out and trapped charge [4.21]. The first boundary to break is 10^{12} write cycles to meet the requirements for SRAM Last-Level-Cache [4.22], and towards 10^{16} for low level cache and general logic applications.

The NDR FeFET is qualitatively compared to other emerging memory technologies below.

Table 4.2: Comparison of novel FE and NDR memory technologies

	6T CMOS	FE eDRAM	X-Coupled TD	NDR FeFET
Speed	✓	X	~	✓
Area	X	✓	✓	✓
Power	~	✓	✓	✓
Process Complexity	✓	X	X	✓
Functionality	~	✓	~	✓
Endurance	✓	~	✓	~

Chapter 5: Conclusions

This dissertation has examined various new approaches to enhancing the Power, Performance, Area, and Cost efficiency (PPAC) of integrated circuit technology by exploring novel functionalities of familiar integrated circuit components. The intense computation and data-handling requirements of AI workloads can be met by the improvements in PPAC unlocked by the compact device technologies proposed in this work.

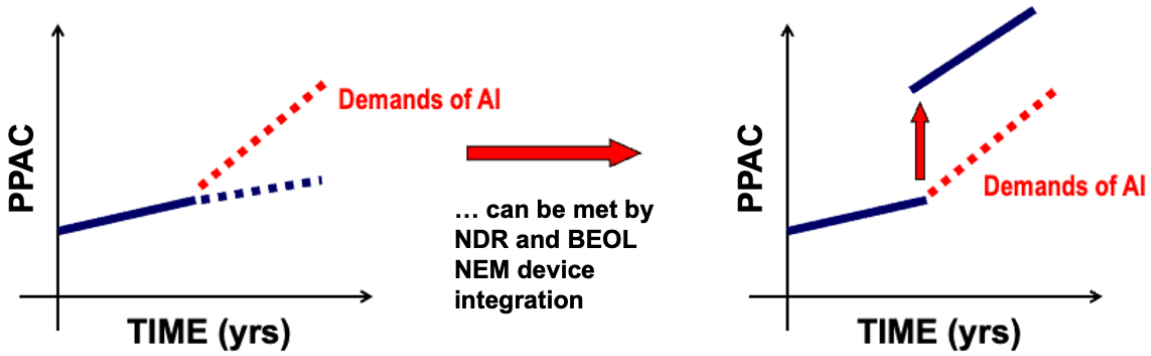


Figure 5.1: Integrating the technologies described in this dissertation can increase the achievable PPAC of advanced IC platforms to meet or exceed the demands of AI.

5.1 Contributions of This Work

Both the NEMS and NDR FeFET device approaches described in this thesis have their own unique advantages (and disadvantages) compared to contemporary and emerging technologies, and will require thoughtful Design Technology Co-Optimization when integrating them into systems.

In chapter 2, BEOL NV-NEMS were investigated. It was projected that optimally designed NV-NEM switches implemented with three BEOL layers in a 5 nm CMOS technology can be programmed with CMOS-compatible voltages and are expected to be more compact than SRAM bit-cells and more energy efficient other contemporary non-volatile memory schemes. The differential half-select programming scheme proposed effectively eliminates the need for access transistors without causing write disturb issues. With the compact crossbar array architecture and programming scheme, BEOL NV-NEM switches can not only implement traditional embedded nonvolatile memory, but can also be used for compute-in-memory architectures suitable for machine learning systems.

In chapter 3, a novel transistor-based NDR device was proposed and optimized. The NDR FeFET is highly compatible with state-of-the-art CMOS process technologies and was shown via TCAD simulations that it could achieve higher peak current (> 400 μA/μm) and PVCR (> 10^6) than any other proposed semiconductor-based NDR device.

In chapter 4, compact NDR FeFET-based memory bit-cells were investigated. NDR FeFET bit-cells display rich new behaviors unseen in previous NDR-based memory bit-cells due to the hysteretic NDR output characteristic and giant PVCR; namely,

- Minimum operating voltage below $2 \times V_{PEAK}$
- Minimum retention voltage far below $2 \times V_{PEAK}$
- Standby power dissipation far below that of state of the art CMOS SRAM
- Nonvolatile operation

5.2 Opportunities for Future Investigation

BEOL NV NEM Switches

In the course of this work, careful simulation work was pursued to benchmark the vertically oriented BEOL NEMS. However, experimental demonstration of the differential half-select scheme, which allows the NEMS to be programmed in a compact fashion similar to other crossbar-type memory, has yet to be completed. Additionally, the ultimate limits to BEOL NEMS reprogramming endurance should be experimentally studied; currently, it is unclear if the device will first fail from material fatigue in the bulk of the beam, or due to contact surface degradation. This will have implications for switch design optimization in future nodes in which hard metals like Ruthenium and Cobalt are being considered. New applications of BEOL NEMS beyond traditional nonvolatile memory, such as compute-in-memory schemes, should continue to be investigated (elaborated further in the final section of this chapter). Additionally, backside power delivery schemes- in which the backside of the silicon wafer is removed and built up to have wide metal lines for power delivery- have changed guidelines for the BEOL scaling roadmap. For Intel's 4 nm node (Fig. 5.2), the introduction of the backside power delivery ("PowerVia") scheme came with a simultaneous increase in Metal 0 pitch from 30 nm to 36 nm [5.1]; this increase in metal pitch increases the minimum programming voltage of BEOL NEMS by approximately 30%.

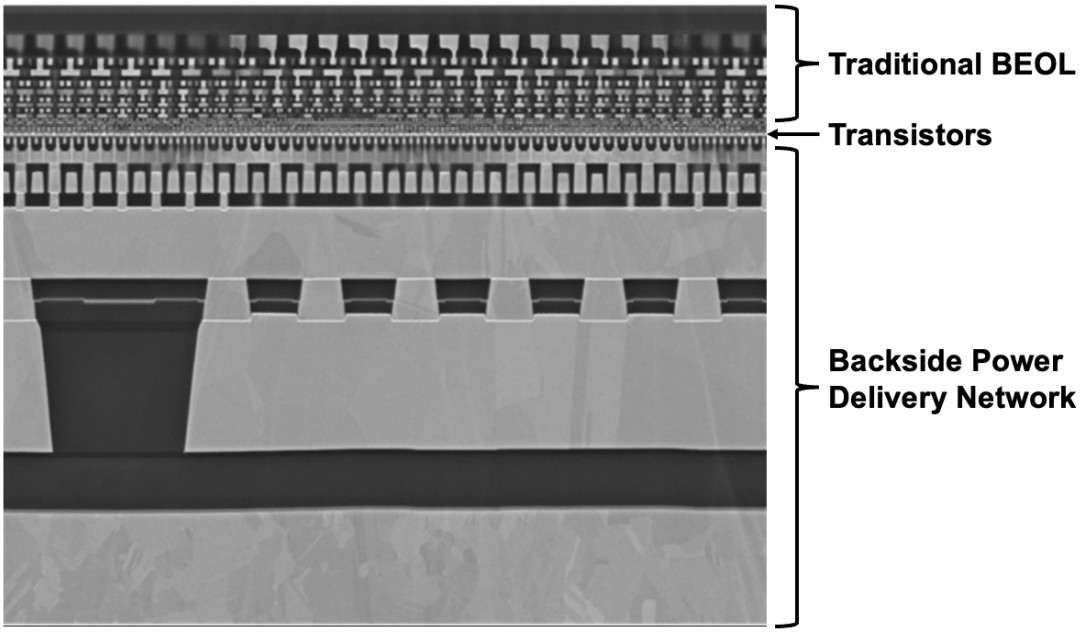


Figure 5.2: Cross-Sectional TEM of Intel 4 Backside Power Delivery scheme. Adapted from [5.1].

NDR FeFET

TCAD work done thus far has given valuable insight into the capabilities and key enabling process modules for the NDR FeFET. Further TCAD work should be performed to benchmark the NDR FeFET performance for 3D MOSFET structures, such as FinFET and Gate-All-Around-FET (GAA-FET).

Firstly, these structures, due to their geometry, have an effective channel width significantly longer than their channel length. Therefore, along the channel width, the ferroelectric may break up into multiple domains, which risks increasing the drain voltage required to switch the entire ferroelectric into the negative polarization state to achieve high PVCR. This should not be as significant as the FE breaking up into multiple domains along the channel length, since these positively polarized FE domains will maintain strong capacitive and low resistance coupling to the drain. On the other hand, the enhanced transverse electric field at the edges of these structure may act as seed sites to switch the FE to the negative polarization at a lower drain voltage. Furthermore, modern GAA-FET architectures have strict requirements on gate stack thickness to stack 3+ nanosheets and minimize gate fringing field capacitance. This imparts a maximum insulator thickness (FE + IL) of about 2.5 nm, which may limit the achievable PVCR as per Fig. 3.7. With an optimized, high-remanent-polarization gate stack and enhanced subthreshold slope present in the GAA-FET architecture, it is feasible that the GAA-FET could achieve a PVCR of at least 10^5 to enable VLSI applications.

The first step to experimentally demonstrate the NDR FeFET (and motivate further investigation) will be to develop a manufacturing process flow compatible with the facilities available in the Berkeley Marvell Nanofabrication lab. There are several key manufacturing module challenges which must be addressed:

- Manufacturing of an ultra-short-channel, small width transistor (< 12 x 40 nm^2)
- Preparation of an ultra-thin-body SOI substrate (T_{si} ~ 5 nm)
- Integrating low K spacers (or air-gaps) to the nanoscale FET, with the underlapped source/overlapped drain doping profiles.
- Manufacturing of a low-coercive voltage FE gate stack (Metal/FE/IL/Silicon) integrated into a (preferably, replacement metal gate) transistor flow[1]
- Low WF metal gate (< 4.15 eV) to enable diode-connected operation of a depletion mode device.

[1] With limited experimental demonstration of low coercive voltage, sharply switching FeFET gate stacks, this remains one of the largest scientific challenges to be resolved for the NDR FeFET.

If any one of these modules is not within specification, then the device will either display mild NDR behavior (PVCR < 10), or require large voltages (> 1 V) to achieve NDR behavior. Prof. Salahuddin's group has already developed a manufacturing process for short-channel, ultra-thin-body SOI NCFETs [5.2]; however, the device width would need to be reduced substantially below 40 nm [5.3] in order to prevent multiple domains from forming along the width of the device.

Future work must also seriously consider whether to attempt NDR FeFET realization with alternative channel materials, such as an amorphous oxide semiconductor. Amorphous metal oxide semiconductors, such as Indium Tungsten Oxide (IWO), do not have a native oxide layer, which can help enable a low voltage operation, high PVCR, and high endurance NDR FeFET device [5.4]. For these alternative channel materials, methods of reducing the source-to-FE capacitive coupling will have to be investigated, since the junctionless FETs developed for them cannot have a tunable source-doping profile. New characterization capabilities will have to be added to the Device Characterization Lab. Namely, the ability to measure the polarization charge displacement current from nanoscale-area films will have to be developed. This will enable the analysis of nanoscale ferroelectric films to determine coercive voltage, remanent polarization, and qualitative domain count. Since the absolute magnitude of the displacement current will be very small, a measurement setup with a low noise transimpedance amplifier will likely have to be developed.

In addition to working out these key module challenges, variability will have to be addressed, in order to demonstrate basic circuits using the NDR FeFET. The most important parameters which must be controlled for a proof-of-concept demonstration are voltage hysteresis and peak voltage value. If the hysteresis is too large, then the NDR behavior will not be dynamic (i.e., require a reset operation), and if V_{PEAK} is too large, then the NDR behavior will only occur for very high voltages. Even if the PVCR is low, device operation can still be demonstrated.

Pertinent tor NDR FeFET operation are variations in:

- FE, Oxide IL, and channel thickness
- Spacer Thickness and dielectric properties
- Source/drain profiles and under/overlap with the gate edge
- Defects and other charge traps in the gate stack

The ferroelectric brings its own serious variability concerns. Ferroelectric remanent polarization, coercive field, and domain density are dependent on (to varying degrees):

- Crystal orientation

- Crystal phase
- Oxygen vacancy concentration
- Presence of grain boundaries/grain size
- Presence of pinned domains or dead layers

Various sources in the literature have reported improving these parameters to various degrees. One of the core issues with Ferroelectric HZO is that its crystallization is a random nucleation-growth process. Due to this there is little consensus on how to best control the orientation of the regions that form during the crystallization process. Since the polarization gradient factor (domain wall energy) is thought to depend on the orientation of the FE film, which will significantly affect the domain density, this is a key point of concern which must be resolved with future FE work.

NDR FeFET Circuits

Future investigation should be pursued to ensure that the NDR FeFET based SRAM cell can be written with a word-line voltage under 1 V, to ensure lifetime operating reliability. This may be achieved by increasing the intrinsic drive strength (drain current) of the pass gate "access" NMOSFET, while keeping the off-state current low; alternatively, the peak current of the minimum-sized pull-down NDR FeFET could be intentionally reduced so that the access NMOSFET could more-easily drive the storage node up to V_{DD}. Similarly, techniques to increase the write speed of an NDR FeFET towards the theoretical polarization switching speed limit of < 10 ps should be investigated in order to achieve a similar read/write access time as CMOS SRAM [5.5]. Additionally, investigation of NDR FeFET circuit operation, for V_{DD} near to the "active" V_{min}, should be pursued, to verify the ultimate energy efficiency limits of an NDR FeFET technology platform. In addition, the NDR FeFET should be investigated for use in other types of NDR-device based digital logic, to implement more compact, energy efficient systems than contemporary CMOS [5.6]. While NDR FeFETs are intrinsically slower than standard MOSFETs due to the ferroelectric polarization switching time, circuit-level optimizations may amortize the polarization switching time and help make NDR FeFET-based logic a compelling approach for energy-efficient computing. The nonvolatile operation capability of the NDR FeFET, which opens the possibility of zero standby leakage NDR SRAM, should be further explored, as it may enable the use of high peak-current, low PVCR NDR FeFET devices for NDR SRAM. Finally, a compact model of the NDR FeFET should be developed, for the investigation of NDR FeFET-based VLSI systems.

The additional functionality provided by NDR FeFETs, with their reconfigurable behavior, nonvolatility, and low V_{min} show that they not only can be drop-in replacement for 6T CMOS SRAM, but may offer circuit designers new flexibilities to address the age of generative AI. Although this chapter has focused on the memory applications of NDR FeFET devices, NDR devices have been previously proposed to implement a plethora of compact integrated circuits. Table I compares the number of components required for common logic circuits. Since the gate voltage of the NDR FeFET can be varied to switch the NDR FeFET from a NDR to non-NDR mode, this may lower the component count even further for certain circuit architectures.

Table 5.1: Device Count for Hybrid NDR+CMOS Circuits [5.7]

Circuit	CMOS	NDR + NMOS
XOR	**16**	**4**
NOR+flip-flop	**12**	**4**
SRAM	**6**	**3**

Moreover, due to the sharp switching characteristic of the ferroelectric (similar to "Negative Capacitance" transistors [5.8]), and beneficial hysteresis characteristic, NDR FeFET devices are not bound by the Boltzmann limit for conventional CMOS transistors. Therefore, NDR FeFET based circuits may be an appealing approach to move past the Boltzmann limit and reduce supply voltages below 0.4 V to continue enhancing the energy efficiency of CMOS-based electronics [5.9], if a scaled low-hysteresis voltage gate stack could be developed with low peak voltage NDR FeFETs.

5.3 Breaking the Memory Wall with Novel Devices

An alarming phenomenon faced in advanced computing systems is the so-called "Memory Wall". Even if the energy efficiency of individual compute operations is continually enhanced with computer architecture approaches, it takes a significant amount of energy and time to load the data itself into the CPU from external high-density memory and store it again off chip [5.10]. For example, in a modern technology node, a 32-bit add operation only takes ~20 fJ and 150 ps to complete. However, fetching the two 32-bit words from the cache memory takes 1.9pJ and 400ps, and fetching the words from off chip main memory takes an astounding 1.3nJ of energy! As a result, specialized domain-specific accelerators seldom achieve an energy efficiency better than 100fJ/operation, regardless of their peak power and performance [5.11].

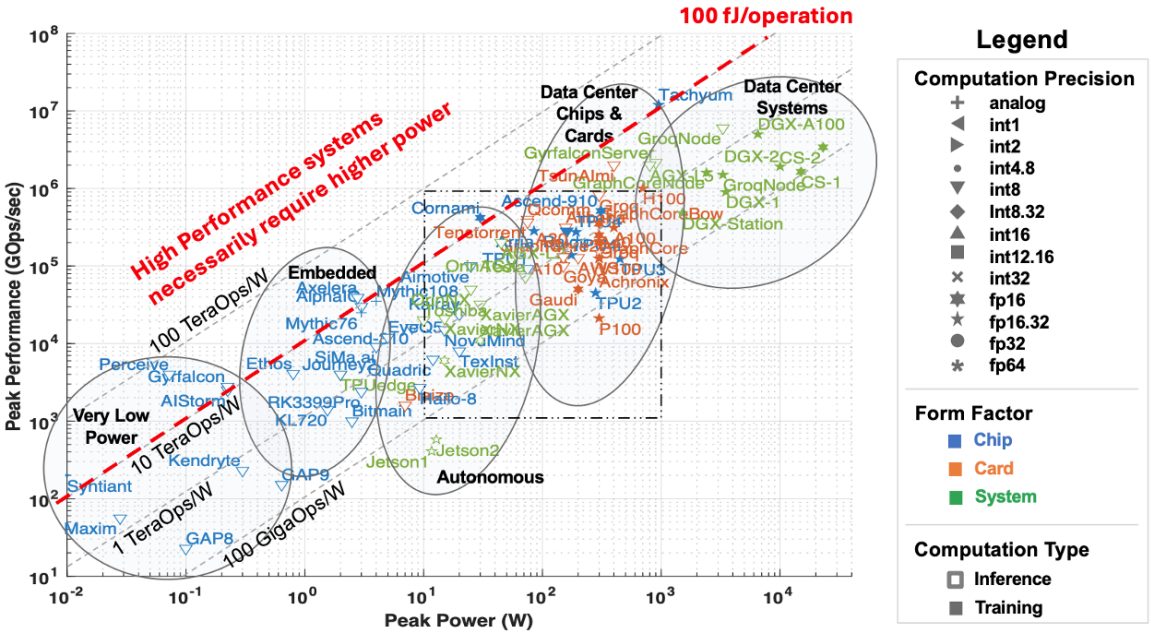


Figure 5.3: Peak Performance vs Peak Power consumption for various AI accelerators. While specialized architectures may massively achieve a high computing performance, they may only do so with a high power consumption. Adapted from [5.11].

Therefore, while familiar SRAM and crossbar technology architectures might reduce the area, standby power, or cost of IC technology, these alone will not substantially reduce the computational efficiency of CMOS-based systems when taking into account the energy required to transfer data to and from the main memory. To address this issue, "beyond-von Neumann" architectures, particularly "Compute In Memory" approaches [5.12] in which

the data is both operated on and stored physically on-chip, must continue to be aggressively pursued. In the course of this work, Compute In Memory circuits for BEOL NEM crossbar arrays were demonstrated (look-up table [5.13]) and proposed (capacitive Multiply-Accumulate crossbar array [5.14]). NDR devices have previously been proposed as a compact way to implement Cellular Neural Networks [5.15], [5.16], a type of neural network which is well-suited for image processing; however, new methods of using NDR devices for in-memory computation or neuromorphic computing should be pursued, to maintain their technological importance as new computing paradigms are explored [5.17].

Bibliography

[1.1] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, “Generative AI,” *Business & Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, Feb. 2024, doi: 10.1007/s12599-023-00834-7.

[1.2] R. Kurzweil, “The Singularity is Near,” in *Ethics and Emerging Technologies*, London: Palgrave Macmillan UK, 2014, pp. 393–406. doi: 10.1057/9781137349088_26.

[1.3] S. Jurvetson, “Carrying the mantle of Moore’s Law,” https://x.com/FutureJurvetson/status/1681783187792613376.

[1.4] G. E. Moore, “Cramming More Components onto Integrated Circuits,” *Electronics (Basel)*, vol. 38, no. 8, pp. 1–14, 1965.

[1.5] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, “Design of ion-implanted MOSFET’s with very small physical dimensions,” *IEEE J Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974, doi: 10.1109/JSSC.1974.1050511.

[1.6] A. Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET’s: A 3-D ‘atomistic’ simulation study,” *IEEE Trans Electron Devices*, vol. 45, no. 12, pp. 2505–2513, 1998, doi: 10.1109/16.735728.

[1.7] S. E. Thompson, R. S. Chau, T. Ghani, K. Mistry, S. Tyagi, and M. T. Bohr, “In Search of ‘Forever,’ Continued Transistor Scaling One New Material at a Time,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 18, no. 1, pp. 26–36, Feb. 2005, doi: 10.1109/TSM.2004.841816.

[1.8] T. K. Liu and L. P. Tatum, “Materials Innovation: Key to Past and Future Transistor Scaling,” in *75th Anniversary of the Transistor*, Wiley, 2023, pp. 403–414. doi: 10.1002/9781394202478.ch35.

[1.9] O. Nalamasu, “Technology Trends,” in *SEMI International Trade Partners Conference*, Waimea, HI, Nov. 2022.

[1.10] J. D. Meindl, Q. Chen, and J. A. Davis, “Limits on Silicon Nanoelectronics for Terascale Integration,” *Science (1979)*, vol. 293, no. 5537, pp. 2044–2049, Sep. 2001, doi: 10.1126/science.293.5537.2044.

[1.11] S. Liao *et al.*, “Complementary Field-Effect Transistor (CFET) Demonstration at 48nm Gate Pitch for Future Logic Technology Scaling,” in *2023 International Electron Devices Meeting (IEDM)*, IEEE, Dec. 2023, pp. 1–4. doi: 10.1109/IEDM45741.2023.10413672.

[1.12] Y. Liu, X. Duan, H.-J. Shin, S. Park, Y. Huang, and X. Duan, “Promises and prospects of two-dimensional transistors,” *Nature*, vol. 591, no. 7848, pp. 43–53, Mar. 2021, doi: 10.1038/s41586-021-03339-z.

[1.13] G. Mirabelli *et al.*, “Investigation of die-cost scaling scenarios in future technologies,” in *DTCO and Computational Patterning III*, N. V. Lafferty and H. Grunes, Eds., SPIE, Apr. 2024, p. 13. doi: 10.1117/12.3010149.

[1.14] J. L. Hennessy and D. A. Patterson, “A new golden age for computer architecture,” *Commun ACM*, vol. 62, no. 2, pp. 48–60, Jan. 2019, doi: 10.1145/3282307.

[1.15] G. E. Moore, “Progress in Digital Integrated Electronics,” in *Technical Digest 1975 International Electron Devices Meeting*, IEEE, 1975, pp. 11–13. doi: 10.1109/N-SSC.2006.4804410.

[2.1] W. Y. Choi, H. Kam, D. Lee, J. Lai and T.-J. K. Liu, “Compact Nano-Electro-Mechanical Non-Volatile Memory (NEMory) for 3D Integration,” in International Electron Devices Meeting, pp. 603–606, Dec. 2007. DOI: 10.1109/IEDM.2007.4419011

[2.2] F. Chen, H. Kam, D. Marković, T.-J. K. Liu, V. Stojanović and E. Alon, “Integrated circuit design with NEM relays,” in 2008 IEEE/ACM International Conference on Computer-Aided Design, pp. 750–757, Nov. 2008. DOI: 10.1109/ICCAD.2008.4681660

[2.3] W. Y. Choi, T. Osabe and T.-J. K. Liu, “Nano- Electro-Mechanical Nonvolatile Memory (NEMory) Cell Design and Scaling,” IEEE Trans. Electron Devices, vol. 55, no. 12, pp. 3482–3488, Nov. 2008. DOI: 10.1109/TED.2008.2006540

[2.4] F. Chen, M. Spencer, R. Nathanael, C. Wang, H. Fariborzi, A. Gupta, H. Kam, V. Pott, J. Jeon, T.-J. K. Liu, D. Marković, V. Stojanović and E. Alon, "Demonstration of integrated micro-electro- mechanical (MEM) switch circuits for VLSI applications," 2010 International Solid State Circuits Conference (San Francisco, California, USA), pp. 150-151, Feb. 2010. DOI: 10.1109/ISSCC.2010.5434010

[2.5] C. Chen, R. Parsa, N. Patil, S. Chong, K. Akarvardar, J Provine, D. Lewis, J. Watt, R. T. Howe, H.-S. P. Wong and S. Mitra, “Efficient FPGAs using nanoelectromechanical relays,” in Proceedings of the 18th annual ACM/SIGDA International Symposium on Field programmable gate arrays, pp. 273–282, Feb. 2010. DOI: 10.1145/1723112.1723158

[2.6] C. Chen, W. S. Lee, R. Parsa, S. Cong, J Provine, J. Watt, R. T. Howe, H.-S. P. Wong and S. Mitra, “Nano-Electro-Mechanical Relays for FPGA Routing: Experimental Demonstration and a Design Technique,” in Design, Automation and Test in Europe Conference and Exhibition (DATE), pp. 1361- 1366, March 2012. DOI: 10.1109/DATE.2012.6176703

[2.7] T. K. Liu, N. Xu, I. Chen, C. Qian and J. Fujiki, "NEM relay design for compact, ultra-low-power digital logic circuits," 2014 IEEE International Electron Devices Meeting, pp. 13.1.1-13.1.4, San Francisco, CA, Dec. 2014. DOI: 10.1109/IEDM.2014.7047042

[2.8] A. Peschot, C. Qian and T.-J. K. Liu, “Nanoelectromechanical Switches for Low-Power Digital Computing,” Micromachines, vol. 6, no. 8, pp. 1046–1065, Aug. 2015. DOI: 10.3390/mi6081046

[2.9] K. Kato, V. Stojanović and T.-J. K. Liu, “Non-volatile nano-electro-mechanical memory for energy-efficient data searching,” IEEE Electron Device Lett., vol. 37, no. 1, pp. 31-34, Dec. 2015. DOI: 10.1109/LED.2015.2504955

[2.10] K. Kato, V. Stojanović and T.-J. K. Liu, “Embedded Nano-Electro-Mechanical Memory for Energy- Efficient Reconfigurable Logic,” IEEE Electron Device Lett., vol. 37, no. 12, pp. 1563–1565, Oct. 2016. DOI: 10.1109/LED.2016.2621187

[2.11] N. Xu, J. Sun, I-R. Chen, L. Hutin, Y. Chen, J. Fujiki, C. Qian and T.-J. K. Liu, "Hybrid CMOS/BEOL- NEMS technology for ultra-low-power IC applications," 2014 IEEE International Electron Devices Meeting, San Francisco, CA, Dec. 2014, pp. 28.8.1-28.8.4. DOI: 10.1109/IEDM.2014.7047130

[2.12] T.-J. K. Liu, U. Sikder, K. Kato and V. Stojanović, “There’s plenty of room at the top,” in 2017 IEEE 30th International Conference on Micro Electro Mechanical Systems (MEMS), Jan. 2017, pp. 1–4. DOI: 10.1109/MEMSYS.2017.7863324

[2.13] U. Sikder, G. Usai, L. Hutin and T.-J. K. Liu, “Design Optimization Study of Reconfigurable Interconnects,” 2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), Kobe, March 2018, pp. 128-130. DOI: 10.1109/EDTM.2018.8421482

[2.14] U. Sikder and T.-J. K. Liu, “Design Optimization for NEM Relays Implemented in BEOL Layers,” 2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Burlingame, CA, Oct. 2017, pp. 1-3. DOI: 10.1109/S3S.2017.8309249

[2.15] H. S. Kwon, S. K. Kim and W. Y. Choi, “Monolithic Three-Dimensional Reconfigurable Logic for Sub-1.2- V Operation,” IEEE Electron Device Lett., vol. 38, no. 9, pp. 1317-1320, July 2017. DOI: 10.1109/LED.2017.2726685

[2.16] W. Y. Choi and Y. J. Kim, “Three-Dimensional Integration of Complementary Metal-Oxide- Semiconductor-Nanoelectromechanical Hybrid Reconfigurable Circuits,” IEEE Electron Device Lett., vol. 36, no. 9, pp. 887–889, July 2015. DOI: 10.1109/LED.2015.2455556

[2.17] U. Sikder, L. P. Tatum, T. Yen and T.-J. K. Liu, “Vertical NV-NEM Switches in CMOS Back-End-of- Line : First Experimental Demonstration and Array Programming Scheme,” in 2020 IEEE International Electron Devices Meeting (IEDM) Technical Digest, Dec. 2020, pp. 21.2

[2.18] G. Usai, L. Hutin, U. Sikder. J. L. Muñoz-Gamarra, T. Ernst, T.-J. K. Liu and M. Vinet, “Balancing Pull-In and Adhesion Stability Margins in Non-Volatile NEM switches,” in IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Oct. 2017, vol. 2, no. 2, pp. 1–2. DOI: 10.1109/S3S.2017.8309215

[2.19] Coventor, CoventorMP 1.2 Documentation. 2019. [2.20] U. Sikder, G. Usai, T. Yen, K. Horace-herron, L. Hutin, and T. K. Liu, “Back-End-of-Line Nano- Electro-Mechanical Switches for Reconfigurable Interconnects,” IEEE Electron Device Lett., vol. 41, no. 4, pp. 625–628, Feb. 2020. DOI: 10.1109/LED.2020.2974473

[2.21] K. L. Wang, J. G. Alzate, and P. K. Amiri, “Low- power non-volatile spintronic memory: STT-RAM and beyond,” J. Phys. D. Appl. Phys., pp 074003, vol. 46, no. 7, Jan. 2013. DOI: 10.1088/0022- 3727/46/7/074003

[2.22] F. Wang, "Energy efficient memory technologies," 2017 IEEE Green Energy and Smart Systems Conference (IGESSC), Long Beach, CA, Nov. 2017, pp. 1-5. DOI: 10.1109/IGESC.2017.8283454

[2.23] Y. Chen, "Reliability Studies of Micro-Relays for Digital Logic Applications," Ph.D dissertation, Dept. of EECS, UC Berkeley, Berkeley, CA, 2015

[2.24] M. T. Bohr and I. A. Young, "CMOS Scaling Trends and Beyond," in IEEE Micro, vol. 37, no. 6, pp. 20-29, Nov. 2017. DOI: 10.1109/MM.2017.4241347

[2.25] S. Braun, J. Oberhammer, and G. Stemme, “Row / Column Addressing Scheme for Large Electrostatic Actuator MEMS Switch Arrays and Optimization of the Operational Reliability by Statistical Analysis,” J. Microelectromechanical Syst., vol. 17, no. 5, pp. 1104–1113, Aug. 2008. DOI: 10.1109/JMEMS.2008.928710

[2.26] S. Braun, J. Oberhammer and G. Stemme, “Smart Individual Switch Addressing of 5×5 and 20×20 MEMS Double-Switch Arrays,” TRANSDUCERS 2007 - 2007 International Solid-State Sensors, Actuators and Microsystems Conference, Lyon, June 2007, pp. 153-156. DOI: 10.1109/SENSOR.2007.4300094

[2.27] Yi-Chou Chen, C.F. Chen, C.T. Chen, J.Y. Yu, S. Wu, S.L. Lung, R. Liu, and Chih-Yuan Lu, “An Access- Transistor-Free (OT/lR) Non-Volatile Resistance Random Access Memory (RRAM) Using a Novel Threshold Switching, Self-Rectifying Chalcogenide Device,” in International Technical Digest on Electron Devices Meeting, Dec. 2003, pp. 905–908. DOI: 10.1109/IEDM.2003.1269425

[2.28] Y. -C. Luo, A. Lu, J. Hur, S. Li and S. Yu, "Design and Optimization of Non-Volatile Capacitive Crossbar Array for In-Memory Computing," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 3, pp. 784-788, March 2022, doi: 10.1109/TCSII.2021.3108148.

[2.29] Yu, E., K, G., Saxena, U. *et al.* Ferroelectric capacitors and field-effect transistors as in-memory computing elements for machine learning workloads. *Sci Rep* **14**, 9426 (2024). https://doi.org/10.1038/s41598-024-59298-8

[3.1] S. Y. Wu *et al.*, “A 3nm CMOS FinFlex™ Platform Technology with Enhanced Power Efficiency and Performance for Mobile SoC and High Performance Computing Applications,” in *Technical Digest - International Electron Devices Meeting, IEDM*, IEEE, Dec. 2022, pp. 2751–2754. doi: 10.1109/IEDM45625.2022.10019498.

[3.2] L. Esaki, “New Phenomenon In Narrow Germanium p-n Junctions,” *Phys. Rev.*, vol. 109, no. 2, pp. 603–604, Jan. 1958, doi: 10.1103/PhysRev.109.603.

[3.3] A. W. Hull, “The Dynatron: A Vacuum Tube Possessing Negative Electric Resistance,” *Proceedings of the Institute of Radio Engineers*, vol. 6, no. 1, pp. 5–35, 1918, doi: 10.1109/JRPROC.1918.217353.

[3.4] L. Esaki and R. Tsu, “Superlattice and Negative Differential Conductivity in Semiconductors,” *IBM J Res Dev*, vol. 14, no. 1, pp. 61–65, 1970, doi: 10.1147/rd.141.0061.

[3.5] T.-J. King and D. Y. K. Liu, “CMOS COMPATIBLE PROCESS FOR MAKING A TUNABLE NEGATIVE DIFFERENTIAL RESISTANCE (NDR) DEVICE,” 6596617, Jul. 22, 2003

[3.6] P. Wu, M. Li, B. Zhou, X. S. Hu, and J. Appenzeller, “Cross-Coupled Gated Tunneling Diodes With Unprecedented PVCRs Enabling Compact SRAM Design - Part I: Device Concept,” *IEEE Trans Electron Devices*, vol. 69, no. 11, pp. 6078–6084, Nov. 2022, doi: 10.1109/TED.2022.3207139.

[3.7] J. Dai, “Introduction to Ferroelectrics,” in *Ferroic Materials for Smart Systems*, Wiley, 2020, pp. 15–46. doi: 10.1002/9783527815388.ch2.

[3.8] T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, “Ferroelectricity in hafnium oxide thin films,” *Appl Phys Lett*, vol. 99, no. 10, Sep. 2011, doi: 10.1063/1.3634052.

[3.9] M. N. K. Alam *et al.*, “Transition-state-theory-based interpretation of Landau double well potential for ferroelectrics,” Apr. 2024.

[3.10] H. Lee *et al.*, “Unveiling the Origin of Robust Ferroelectricity in Sub-2 nm Hafnium Zirconium Oxide Films,” *ACS Appl Mater Interfaces*, vol. 13, no. 30, pp. 36499–36506, Aug. 2021, doi: 10.1021/acsami.1c08718.

[3.11] J. Y. Park, D. H. Lee, G. H. Park, J. Lee, Y. Lee, and M. H. Park, “A perspective on the physical scaling down of hafnia-based ferroelectrics,” *Nanotechnology*, vol. 34, no. 20, p. 202001, May 2023, doi: 10.1088/1361-6528/acb945.

[3.12] S. Dunkel *et al.*, “A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond,” in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Dec. 2017, pp. 19.7.1-19.7.4. doi: 10.1109/IEDM.2017.8268425.

[3.13] A. Tan, “Towards High-Endurance, Nonvolatile, CMOS-Compatible Ferroelectric Memories for Next-Generation Computing,” University of California, Berkeley, Berkeley, 2022.

[3.14] E. Covi, H. Mulaosmanovic, B. Max, S. Slesazeck, and T. Mikolajick, “Ferroelectric-based synapses and neurons for neuromorphic computing,” *Neuromorphic Computing and Engineering*, vol. 2, no. 1, p. 012002, Mar. 2022, doi: 10.1088/2634-4386/ac4918.

[3.15] P. Wang *et al.*, “Drain–Erase Scheme in Ferroelectric Field-Effect Transistor—Part I: Device Characterization,” *IEEE Trans Electron Devices*, vol. 67, no. 3, pp. 955–961, Mar. 2020, doi: 10.1109/TED.2020.2969401.

[3.16] A. K. Saha, M. Si, K. Ni, S. Datta, P. D. Ye, and S. K. Gupta, “Ferroelectric thickness dependent domain interactions in FEFETs for memory and logic: A phase-field model based analysis,” in *Technical Digest - International Electron Devices Meeting, IEDM*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 4.3.1-4.3.4. doi: 10.1109/IEDM13553.2020.9372099.

[3.17] T. Yu, W. Lü, Z. Zhao, P. Si, and K. Zhang, “Negative drain-induced barrier lowering and negative differential resistance effects in negative-capacitance transistors,” *Microelectronics J*, vol. 108, Feb. 2021, doi: 10.1016/j.mejo.2020.104981.

[3.18] C. Jin, C. J. Su, Y. J. Lee, P. J. Sung, T. Hiramoto, and M. Kobayashi, “Study on the roles of charge trapping and fixed charge on subthreshold characteristics of FeFETs,” *IEEE Trans Electron Devices*, vol. 68, no. 3, pp. 1304–1312, Mar. 2021, doi: 10.1109/TED.2020.3048916.

[3.19] H. Zhou *et al.*, “Negative Capacitance, n-Channel, Si FinFETs: Bi-directional Sub-60 mV/dec, Negative DIBL, Negative Differential Resistance and Improved Short Channel Effect,” in *IEEE Symposium on VLSI Technology*, Honolulu, HI, USA: IEEE, Jun. 2018, pp. 53–54. doi: 10.1109/VLSIT.2018.8510691.

[3.20] M. Jerry, J. A. Smith, K. Ni, A. Saha, S. Gupta, and S. Datta, “Insights on the DC Characterization of Ferroelectric Field-Effect-Transistors,” in *76th Device Research Conference*, Santa Barbara, CA, USA, 2018, pp. 1–2. doi: 10.1109/DRC.2018.8442191.

[3.21] Synopsys Inc, *Sentaurus User’s Manual, Version T-2022.03*. Mountain View, CA, 2022.

[3.22] I. Myeong *et al.*, “Strategies for a wide Memory Window of Ferroelectric FET for multilevel Ferroelectric VNAND operation,” *IEEE Electron Device Letters*, 2024, doi: 10.1109/LED.2024.3400983.

[3.23] S. Dutta *et al.*, “Logic Compatible High-Performance Ferroelectric Transistor Memory,” *IEEE Electron Device Letters*, vol. 43, no. 3, pp. 382–385, Mar. 2022, doi: 10.1109/LED.2022.3148669.

[3.24] J. Mueller, S. Slesazeck, and T. Mikolajick, “Ferroelectric Field Effect Transistor,” in *Ferroelectricity in Doped Hafnium Oxide: Materials, Properties and Devices*, Elsevier, 2019, pp. 451–471. doi: 10.1016/B978-0-08-102430-0.00022-X.

[3.25] Y. Taur and T. H. Ning, “MOS Capacitors,” in *Fundamentals of Modern VLSI Devices*, 3rd ed., Cambridge University Press, 2021, pp. 99–170. doi: 10.1017/9781108847087.007.

[3.26] M. M. Dahan, H. Mulaosmanovic, O. Levit, S. Dünkel, S. Beyer, and E. Yalon, “Sub-Nanosecond Switching of Si:HfO2 Ferroelectric Field-Effect Transistor,” *Nano Lett*, vol. 23, no. 4, pp. 1395–1400, Feb. 2023, doi: 10.1021/acs.nanolett.2c04706.

[3.27] K. Chatterjee, A. J. Rosner, and S. Salahuddin, “Intrinsic speed limit of negative capacitance transistors,” *IEEE Electron Device Letters*, vol. 38, no. 9, pp. 1328–1330, Sep. 2017, doi: 10.1109/LED.2017.2731343.

[3.28] M. I. Popovici *et al.*, “High-Endurance Ferroelectric (La, Y) and (La, Gd) Co-Doped Hafnium Zirconate Grown by Atomic Layer Deposition,” *ACS Appl Electron Mater*, vol. 4, no. 4, pp. 1823–1831, Apr. 2022, doi: 10.1021/acsaelm.2c00063.

[3.29] T. Onaya *et al.*, “Improvement in ferroelectricity of HfxZr1-xO2 thin films using ZrO2 seed layer,” *Applied Physics Express*, vol. 10, no. 8, Aug. 2017, doi: 10.7567/APEX.10.081501.

[3.30] S. C. Chang *et al.*, “FeRAM using Anti-ferroelectric Capacitors for High-speed and High-density Embedded Memory,” in *International Electron Devices Meeting, IEDM*, San Francisco: IEEE, Dec. 2021, pp. 33.2.1-33.2.4. doi: 10.1109/IEDM19574.2021.9720510.

[3.31] M. Pešić, M. Hoffmann, C. Richter, T. Mikolajick, and U. Schroeder, “Nonvolatile Random Access Memory and Energy Storage Based on Antiferroelectric Like Hysteresis in ZrO2,” *Adv Funct Mater*, vol. 26, no. 41, pp. 7486–7494, Nov. 2016, doi: 10.1002/adfm.201603182.

[3.32] R. Ichihara *et al.*, “Accurate Picture of Cycling Degradation in HfO2-FeFET Based on Charge Trapping Dynamics Revealed by Fast Charge Centroid Analysis,” in *Technical Digest - International Electron Devices Meeting, IEDM*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6.3.1-6.3.4. doi: 10.1109/IEDM19574.2021.9720516.

[3.33] S. Lee *et al.*, “Design Guidelines of Hafnia Ferroelectrics and Gate-Stack for Multilevel-Cell FeFET,” *IEEE Trans Electron Devices*, vol. 71, no. 3, pp. 1865–1871, Mar. 2024, doi: 10.1109/TED.2024.3355873.

[3.34] Z. Zhang, Y. Yang, R. Su, T. Lin, X. Miao, and X. Wang, “Recorded Ferroelectric Polarization Switching of Hf0.5Zr0.5O Capacitors Achieved by Thermal Rewake-up,” in *IEEE Electron Devices Technology and Manufacturing Conference: Strengthening the Globalization in Semiconductors, EDTM 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/EDTM58488.2024.10511655.

[3.35] Z. Zhang, Y. Yang, R. Su, T. Lin, X. Miao, and X. Wang, “Recorded Ferroelectric Polarization Switching of Hf0.5Zr0.5O2 Capacitors Achieved by Thermal Rewake-up,” in *2024 8th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, IEEE, Mar. 2024, pp. 1–3. doi: 10.1109/EDTM58488.2024.10511655.

[3.36] T. K. Paul, A. K. Saha, and S. K. Gupta, “Direction-Dependent Lateral Domain Walls in Ferroelectric Hafnium Zirconium Oxide and their Gradient Energy Coefficients: A First-Principles Study,” *Adv Electron Mater*, vol. 10, no. 1, Jan. 2024, doi: 10.1002/aelm.202300400.

[3.37] Y. Shao *et al.*, “Discrete Domain Switching in Scaled Oxide-Channel Ferroelectric FETs,” in *Device Research Conference*, College Park, Maryland: IEEE, Jun. 2024, pp. 1–1.

[3.38] N. Tasneem *et al.*, “Remote Oxygen Scavenging of the Interfacial Oxide Layer in Ferroelectric Hafnium-Zirconium Oxide-Based Metal-Oxide-Semiconductor Structures,” *ACS Appl Mater Interfaces*, vol. 14, no. 38, pp. 43897–43906, Sep. 2022, doi: 10.1021/acsami.2c11736.

[3.39] N. Gong and T. P. Ma, “A Study of Endurance Issues in HfO2-Based Ferroelectric Field Effect Transistors: Charge Trapping and Trap Generation,” *IEEE Electron Device Letters*, vol. 39, no. 1, pp. 15–18, Jan. 2018, doi: 10.1109/LED.2017.2776263.

[3.40] A. J. Tan *et al.*, “Ferroelectric HfO2 Memory Transistors with High-κ Interfacial Layer and Write Endurance Exceeding 10^10 Cycles,” *IEEE Electron Device Letters*, vol. 42, no. 7, pp. 994–997, Jul. 2021, doi: 10.1109/LED.2021.3083219.

[3.41] Yue Liang, K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, “From DRAM to SRAM with a novel SiGe-based negative differential resistance (NDR) device,” in *IEEE International Electron Devices Meeting, 2005. IEDM Technical Digest.*, IEEE, 2005, pp. 959–962. doi: 10.1109/IEDM.2005.1609520.

[3.42] S. L. Rommel *et al.*, “Record PVCR GaAs-based tunnel diodes fabricated on Si substrates using aspect ratio trapping,” in *2008 IEEE International Electron Devices Meeting*, IEEE, Dec. 2008, pp. 1–4. doi: 10.1109/IEDM.2008.4796801.

[4.1] Epoch (2024) – with major processing by Our World in Data, “Parameter, Compute and Data Trends in Machine Learning,” https://ourworldindata.org/grapher/artificial-intelligence-parameter-count.

[4.2] S. Jones, “Logic 2034 – Technology, Economics, and Sustainability,” in *Industry Strategy Symposium 2024*, Half Moon Bay, California, Jan. 2024.

[4.3] David Schor, “IEDM 2022: Did We Just Witness The Death Of SRAM?,” https://fuse.wikichip.org/news/7343/iedm-2022-did-we-just-witness-the-death-of-sram/.

[4.4] J. P. A. Van Der Wagt, “Tunneling-Based SRAM,” *Proceedings of the IEEE*, vol. 87, no. 4, pp. 571–595, 1999, doi: 10.1109/5.752516.

[4.5] M. Li, P. Wu, B. Zhou, J. Appenzeller, and X. S. Hu, “Cross-Coupled Gated Tunneling Diodes With Unprecedented PVCRs Enabling Compact SRAM Design—Part II: SRAM Circuit,” *IEEE Trans Electron Devices*, vol. 69, no. 11, pp. 6085–6088, Nov. 2022, doi: 10.1109/TED.2022.3207122.

[4.6] E. Goto *et al.*, “Esaki Diode High-Speed Logical Circuits,” *IRE Transactions on Electronic Computers*, vol. EC-9, pp. 25–29, 1960.

[4.7] P. Wu, M. Li, B. Zhou, X. S. Hu, and J. Appenzeller, “Cross-Coupled Gated Tunneling Diodes With Unprecedented PVCRs Enabling Compact SRAM Design—Part I: Device Concept,” *IEEE Trans Electron Devices*, vol. 69, no. 11, pp. 6078–6084, Nov. 2022, doi: 10.1109/TED.2022.3207139.

[4.8] P. R. Berger and A. Ramesh, “Negative Differential Resistance Devices and Circuits,” in *Comprehensive Semiconductor Science and Technology*, Elsevier, 2011, pp. 176–241. doi: 10.1016/B978-0-44-453153-7.00013-4.

[4.9] T.-J. King and D. Y. K. Liu, “CMOS COMPATIBLE PROCESS FOR MAKING A TUNABLE NEGATIVE DIFFERENTIAL RESISTANCE (NDR) DEVICE,” 6596617, Jul. 22, 2003

[4.10] K. Chatterjee, A. J. Rosner, and S. Salahuddin, “Intrinsic speed limit of negative capacitance transistors,” *IEEE Electron Device Letters*, vol. 38, no. 9, pp. 1328–1330, Sep. 2017, doi: 10.1109/LED.2017.2731343.

[4.11] K. M. Koichi Maezawa and T. M. Takashi Mizutani, “A New Resonant Tunneling Logic Gate Employing Monostable-Bistable Transition,” *Jpn J Appl Phys*, vol. 32, no. 1A, p. L42, Jan. 1993, doi: 10.1143/JJAP.32.L42.

[4.12] Y.-T. Wu, F. Ding, M.-H. Chiang, J. F. Chen, and T.-J. K. Liu, “Simulation-Based Study of Low Minimum Operating Voltage SRAM With Inserted-Oxide FinFETs and Gate-All-Around Transistors,” *IEEE Trans Electron Devices*, vol. 69, no. 4, pp. 1823–1829, Apr. 2022, doi: 10.1109/TED.2022.3150645.

[4.13] J. Y. Park *et al.*, “Revival of Ferroelectric Memories Based on Emerging Fluorite-Structured Ferroelectrics,” *Advanced Materials*, vol. 35, no. 43, Oct. 2023, doi: 10.1002/adma.202204904.

[4.14] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. N. Piramanayagam, “Spintronics based random access memory: a review,” *Materials Today*, vol. 20, no. 9, pp. 530–548, Nov. 2017, doi: 10.1016/j.mattod.2017.07.007.

[4.15] S.-C. Chang *et al.*, “FeRAM using Anti-ferroelectric Capacitors for High-speed and High-density Embedded Memory,” in *2021 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Dec. 2021, pp. 33.2.1-33.2.4. doi: 10.1109/IEDM19574.2021.9720510.

[4.16] A. Sharma and K. Roy, “1T Non-Volatile Memory Design Using Sub-10nm Ferroelectric FETs,” *IEEE Electron Device Letters*, vol. 39, no. 3, pp. 359–362, Mar. 2018, doi: 10.1109/LED.2018.2797887.

[4.17] K.-T. Chen *et al.*, “Non-Volatile Ferroelectric FETs Using 5-nm $Hf_{0.5}Zr_{0.5}O_2$ With High Data Retention and Read Endurance for 1T Memory Applications,” *IEEE Electron Device Letters*, vol. 40, no. 3, pp. 399–402, Mar. 2019, doi: 10.1109/LED.2019.2896231.

[4.18] P. Wu, M. Li, B. Zhou, X. S. Hu, and J. Appenzeller, “Cross-Coupled Gated Tunneling Diodes With Unprecedented PVCRs Enabling Compact SRAM Design - Part I: Device Concept,” *IEEE Trans Electron Devices*, vol. 69, no. 11, pp. 6078–6084, Nov. 2022, doi: 10.1109/TED.2022.3207139.

[4.19] Z. Jiang *et al.*, “On the Feasibility of 1T Ferroelectric FET Memory Array,” *IEEE Trans Electron Devices*, vol. 69, no. 12, pp. 6722–6730, Dec. 2022, doi: 10.1109/TED.2022.3216819.

[4.20] K. Ni, X. Li, J. A. Smith, M. Jerry, and S. Datta, “Write Disturb in Ferroelectric FETs and Its Implication for 1T-FeFET AND Memory Arrays,” *IEEE Electron Device Letters*, vol. 39, no. 11, pp. 1656–1659, Nov. 2018, doi: 10.1109/LED.2018.2872347.

[4.21] N. Gong and T. P. Ma, “A Study of Endurance Issues in HfO2-Based Ferroelectric Field Effect Transistors: Charge Trapping and Trap Generation,” *IEEE Electron Device Letters*, vol. 39, no. 1, pp. 15–18, Jan. 2018, doi: 10.1109/LED.2017.2776263.

[4.22] Y. Shiokawa *et al.*, “High write endurance up to 10^12 cycles in a spin current-type magnetic memory array,” *AIP Adv*, vol. 9, no. 3, Mar. 2019, doi: 10.1063/1.5079917.

[5.1] W. Hafez *et al.*, “Intel PowerVia Technology: Backside Power Delivery for High Density and High-Performance Computing,” in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, IEEE, Jun. 2023, pp. 1–2. doi: 10.23919/VLSITechnologyandCir57934.2023.10185208.

[5.2] L.-C. Wang, “Design and Characterization of Advanced Negative Capacitance Field-Effect Transistors Featuring Atomic-Scale CMOS-Compatible Ferroic HfO2-ZrO2 Gate Stack for Next-Generation Computing Technologies,” Doctoral Thesis, University of California, Berkeley, Berkeley, CA, 2023.

[5.3] Y. Shao *et al.*, “Discrete Domain Switching in Scaled Oxide-Channel Ferroelectric FETs,” in *Device Research Conference*, College Park, Maryland: IEEE, Jun. 2024, pp. 1–1.

[5.4] S. Dutta *et al.*, “Logic Compatible High-Performance Ferroelectric Transistor Memory,” *IEEE Electron Device Letters*, vol. 43, no. 3, pp. 382–385, Mar. 2022, doi: 10.1109/LED.2022.3148669.

[5.5] Y.-T. Wu, F. Ding, M.-H. Chiang, J. F. Chen, and T.-J. K. Liu, “Simulation-Based Study of Low Minimum Operating Voltage SRAM With Inserted-Oxide FinFETs and Gate-All-Around Transistors,” *IEEE Trans Electron Devices*, vol. 69, no. 4, pp. 1823–1829, Apr. 2022, doi: 10.1109/TED.2022.3150645.

[5.6] P. R. Berger and A. Ramesh, “Negative Differential Resistance Devices and Circuits,” in *Comprehensive Semiconductor Science and Technology*, Elsevier, 2011, pp. 176–241. doi: 10.1016/B978-0-44-453153-7.00013-4.

[5.7] P. Mazumder, S. Kulkarni, M. Bhattacharya, Jian Ping Sun, and G. I. Haddad, “Digital circuit applications of resonant tunneling devices,” *Proceedings of the IEEE*, vol. 86, no. 4, pp. 664–686, Apr. 1998, doi: 10.1109/5.663544.

[5.8] S. Salahuddin and S. Datta, “Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices,” *Nano Lett*, vol. 8, no. 2, pp. 405–410, Feb. 2008, doi: 10.1021/nl071804g.

[5.9] E. Alon, “Energy efficiency limits of digital circuits based on CMOS transistors,” in *CMOS and Beyond*, Cambridge University Press, 2014, pp. 3–13. doi: 10.1017/CBO9781107337886.003.

[5.10] W. Dally, “On the model of computation,” *Commun ACM*, vol. 65, no. 9, pp. 30–32, Sep. 2022, doi: 10.1145/3548783.

[5.11] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, “AI and ML Accelerator Survey and Trends,” in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, IEEE, Sep. 2022, pp. 1–10. doi: 10.1109/HPEC55821.2022.9926331.

[5.12] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, “Compute-in-Memory Chips for Deep Learning: Recent Trends and Prospects,” *IEEE Circuits and Systems Magazine*, vol. 21, no. 3, pp. 31–56, Jul. 2021, doi: 10.1109/MCAS.2021.3092533.

[5.13] U. Sikder, L. P. Tatum, T.-T. Yen, and T.-J. K. Liu, “Vertical NV-NEM Switches in CMOS Back-End-of-Line: First Experimental Demonstration and Array Programming

Scheme," in *2020 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Dec. 2020, pp. 21.2.1-21.2.4. doi: 10.1109/IEDM13553.2020.9372116.

[5.14] E. Yu, G. K. K, U. Saxena, and K. Roy, "Ferroelectric capacitors and field-effect transistors as in-memory computing elements for machine learning workloads," *Sci Rep*, vol. 14, no. 1, p. 9426, Apr. 2024, doi: 10.1038/s41598-024-59298-8.

[5.15] L. O. Chua and L. Yang, "Cellular neural networks: applications," *IEEE Trans Circuits Syst*, vol. 35, no. 10, pp. 1273–1290, Oct. 1988, doi: 10.1109/31.7601.

[5.16] M. Hänggi and L. O. Chua, "Cellular neural networks based on resonant tunnelling diodes," *International Journal of Circuit Theory and Applications*, vol. 29, no. 5, pp. 487–504, Sep. 2001, doi: 10.1002/cta.172.

[5.17] J. Shalf, "The future of computing beyond Moore's Law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 378, no. 2166, p. 20190061, Mar. 2020, doi: 10.1098/rsta.2019.0061.