UNIVERSITY OF CALIFORNIA
Los Angeles

An Embedded Nonvolatile SRAM in Logic CMOS Process

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Sepideh Nouri

2023

ABSTRACT OF THE DISSERTATION

An Embedded Nonvolatile SRAM in Logic CMOS Process

by

Sepideh Nouri
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2023
Professor Subramanian S. Iyer, Chair

There is an increasing demand for nonvolatile memory (NVM) due to the rapid growth of IoT devices that need to operate with a tight power budget. These devices often have low activity rates and most of the time are in standby mode, resulting in a high leakage energy consumption. To reduce the leakage energy, the supply power can be turned off, but before doing so, data need to be stored in an NVM. Also, many IoT devices are battery-operated or powered wirelessly and can see fluctuations in their supply power. Therefore their critical data need to be stored before a power-loss event. However, the conventional NVMs such as EEPROM and FLASH are not logic CMOS compatible in terms of technology node and operating voltages. They also require extra fabrication steps, which increases the cost for IoT devices that use small memories. Additionally, most NVMs are not embedded with logic. Embedding memory with the logic improves the memory access time and energy. It also increases data security by eliminating interface ports that are vulnerable to side-channel attacks and malicious activities. Furthermore, embedding nonvolatile memories (eNVM) with logic enables novel architectures, such as in-SRAM nonvolatile weight storage for compute-in-memory (nvCIM). In the past, researchers have proposed various eNVMs, such as RRAM, PCM, and STT-RAM [1]. However, these solutions require significant changes to the manufacturing process making it challenging to integrate them into modern SoCs. Therefore, there is a need for an eNVM that can be fabricated in a standard logic CMOS process.

This dissertation presents the first on-chip demonstration of an embedded Nonvolatile SRAM (eNVSRAM) in the standard CMOS logic process. We propose an 8T eNVSRAM architecture and present experimental results of a chip that was taped-out in GlobalFoundries 22FDX technology. In addition to eNVSRAM, I present a few other projects that I have worked on during my PhD studies. These include a nonvolatile circuit tuning technique used to modify the frequency of a ring oscillator, a supply-fluctuation resilient SRAM, and a low-power ASIC chip for detecting heart-rate and missing beats.

The dissertation of Sepideh Nouri is approved.

Joseph R. Cavallaro

C. K. Ken Yang

M. C. Frank Chang

Subramanian S. Iyer, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I take this opportunity to express my sincere gratitude to my advisor Prof. Subramanian Iyer, who has been the main source of encouragement and inspiration during my PhD studies.

I would like to thank my committee members, Prof. Ken Yang, Prof. Frank Chang, and Prof. Joseph Cavallaro for providing critical feedback on my PhD project and for serving as members of my dissertation committee. In particular, I am grateful to Prof. Cavallaro who advised me on my research at Rice University before moving to UCLA.

Many thanks to John Barth from Synopsys for his technical guidance and for taking time to meet with me despite his busy schedule. I am also very grateful to Prof. Sudhakar Pamarti for providing valuable advice during the design and testing. Additionally, thanks to GlobalFoundries for the fabrication support.

It has been a great experience to be a member of the UCLA CHIPS lab and I would like to thank all CHIPS members, especially Siva Jangam, Krutikesh Sahu, Faraz Khan, Jonathan Cox, Zhe Wan, Siyun Qiao, and Dhruv Srinivas .

At the end, I would like to thank my husband, my parents, and my sisters for their endless love and support.

VITA

Publications:

- **S. Nouri**, S. S. Iyer, "A Fully Logic-CMOS 8T eNVSRAM Macro in 22nm FDSOI with Simultaneous Full Array Data Restore for Secure IoT Devices", in IEEE ISSCC, Feb. 2023.
- **S. Nouri**, S. S. Iyer, "Non-Volatile Wideband Frequency Tuning of a Ring-Oscillator by Charge Trapping in High-k Gate Dielectric in 22nm CMOS", in IEEE Electron Device Letters, Nov. 2020.
- **S. Nouri**, J. R. Cavallaro, "A Supply Fluctuation Resilient SRAM", in Asilomar Conference on Signals, Systems, and Computers, 2018.
- **S. Nouri**, B. Aazhang, M. Razavi, J. R. Cavallaro, "A Low-Power Digital ASIC for Detecting Heart-rate and Missing-Beat", in Asilomar Conference on Signals, Systems, and Computers, 2017.

Patents:

- **S. Nouri**, S. Iyer, "Nonvolatile Memory Devices with Charge Trap Transistor Structures and Methods of Operation Thereof", PCT/US2022/014440.
- **S. Nouri**, S. Iyer, "Apparatus and Method for Changing the Functionality of an Integrated Circuit Using Charge Trap Transistors", PCT/US2021/024952.

Employment:

- **Meta-** Research Scientist at Meta Reality Labs, Silicon Research (Intern), Summer 2022.
- **Apple-** Engineer at Silicon Engineering Group (Intern), Summer 2021.
- **Qualcomm-** Engineer at Memory IP Group (Intern), Summer 2020.
- **Micron-** Technology, Engineer at Emerging Memory (Intern), Summer 2019.

CHAPTER 1

Introduction

1.1 Motivation and Objectives

Recently, there has been a growing interest in ultra-low power IoT devices. Many of these devices have a low activity rate and a large portion of the energy is consumed during idle time. It is possible to eliminate the stand-by power by turning off the supply, but the critical data needs to be saved in a non-volatile memory (NVM). Unfortunately, most NVM solutions require extra fabrication steps, which increases the cost for IoT devices that need only a small NVM. Ideally, this small NVM needs to be multi-time programmable, and also embedded with other SoC components to improve power, performance, and area (PPA). Integration of the multi-time programmable embedded non-volatile memories (eNVM) in advanced CMOS nodes remains challenging due to the scaling and voltage incompatibilities, and the need for extra fabrication steps and masks. For example, eFLASH requires a significant amount of additional masks and high voltages (up to $\sim 10V$) to operate. Also, eFLASH does not have a clear roadmap to sub-28nm nodes. Emerging memory technologies such as RRAM [2] , MRAM [3, 4, 5], and PCM [6, 7], while they can operate at logic compatible voltages, require additional complex fabrication processes [8]. Recently, a new multi-time programmable eNVM solution for advanced high-k/metal-gate (HKMG) CMOS technologies is introduced, which turns as-fabricated standard logic transistors into eNVM elements. These logic transistors, when employed as eNVM elements, are called “Charge Trap Transistors” (CTTs) [8]. CTT is the only eNVM technology that is scalable, can operate at logic-compatible voltages, and does not require any additional processes or fabrication steps. The existing NVM and eNVM technologies, including the CTT-based eNVM [9], require external interface ports or a bus to transfer data to/from SRAM. The interface port/bus add energy and area overhead as well as latency in data movement. Additionally, the interface ports are known to be vulnerable to the malicious activities and side-channel attacks. To address these issues, ideally, the CTT-based eNVM needs to be integrated with the SRAM at the bitcell level. This approach enables a one-shot whole array data transfer between the eNVM and SRAM without requiring any interface bus.

The objective of this work is to address the above challenges and demonstrate an embedded Non-Volatile SRAM (eNVSRAM), where the content of a 6T-SRAM core is stored in two additional logic NFETs in a non-volatile fashion. The non-volatile storage mechanism is based on charge trapping in the high-k gate dielectric of the two NFETs that are used as CTTs. This approach eliminates the need for using interface ports for transferring data between NVM and SRAM. eNVSRAM can operate as a regular SRAM with fast read and write speeds, when a nonvolatile data storage is not required. In this work, a $1kb$ CTT-based eNVSRAM macro has been designed and taped-out in $22nm$ CMOS technology. All modes of operations of the macro have been experimentally verified. In addition to using CTTs to build non-volatile memories, CTTs can also be used in other applications such as post-fabrication non-volatile calibration and mismatch compensation for analog circuits such as amplifiers, analog-to-digital converters (ADCs), etc. For example in ADCs, the input comparators are typically built with large transistors to reduce the mismatch between transistors of differential pairs. This limitation increases the input parasitic capacitance of the comparators, reduces the speed of the ADC, and increases its power consumption. CTTs can be potentially used to reduce the mismatch of the transistors in any differential pair by modifying their V_{TH}, hence, allowing to reduce their sizes. In another application, CTTs can be used to increase the fabrication yield by enabling post-fabrication tuning of circuits to bring their specs to an acceptable range before shipping them to customers. For example, some of the fabricated oscillators may not work at an intended frequency due to the process variation and mismatch. This issue can be resolved by tuning the V_{TH} of their transistors to adjust the oscillation frequency. To demonstrate this concept, we designed a ring oscillator in $22nm$ CMOS and performed wideband frequency tuning in a non-volatile fashion. This work serves as a proof-of-concept and the reported technique can be used in other circuits. Other CTT applications include in-SRAM nonvolatile weight storage for compute-in-memory (CIM) and analog memory for machine learning and artificial intelligence [10]. Additionally, since the CTT data is stored as trapped charges in a very thin dielectric ($< 1nm$), which is very secure from reverse engineering, CTT can be used in hardware security applications, such as on-chip reconfigurable encryption and Physically Unclonable Functions (PUF). Other eNVM solutions such as eFUSE and anti-fuse memories can be reverse engineered using scanning electron microscope (SEM) technique. In the next section, we review the fundamentals of the charge trapping in CTT.

1.2 Overview of Charge Trap Transistors (CTT)

High-k gate dielectrics are known to have oxygen vacancies (Figure 1.1) that can trap electrons from the channel, under appropriate bias conditions [11]. Since charge trapping is an electrostatically driven phenomenon, a sufficiently large positive gate-to-channel voltage, that is higher than nominal yet logic compatible, should be present to attract electrons to the gate dielectric. This process is significantly enhanced by heating the device through passing a high current from drain to source (self-heating) [11]. The self-heating assisted charge trapping produces much more stable trapped charges that are not possible to achieve with a conventional Positive Bias Temperature Instability (PBTI) effect. This is because the elevated temperature traps electrons into deeper energy states in the dielectric [11].

It has been shown empirically that the amount of shift in V_{TH} can be calculated based on equation 1.1 [12]:

$$\Delta V_{TH} = \Delta V_{TH_{max}}(1 - e^{(-t/\tau_0)^{\beta}}) \tag{1.1}$$

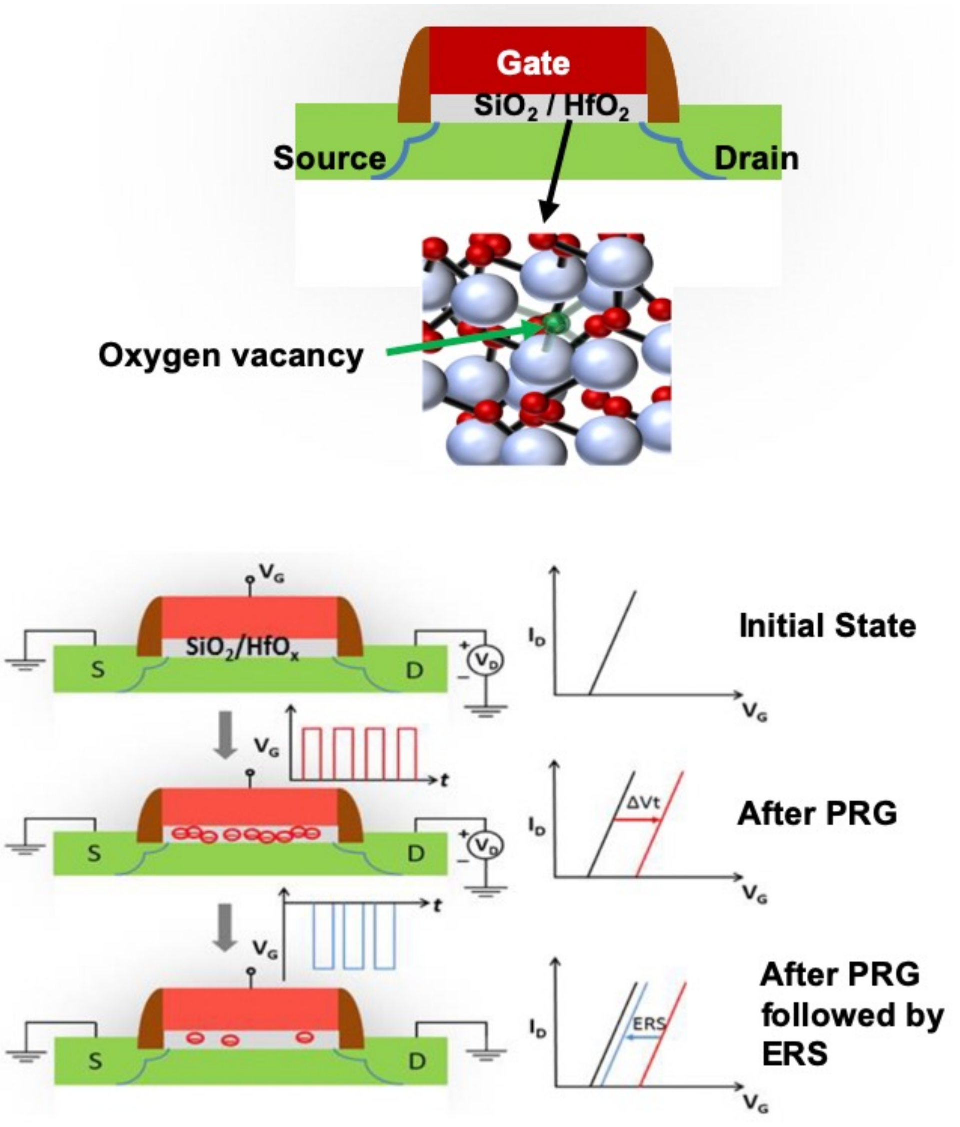


Figure 1.1: Oxygen vacancies in high-k gate dielectric (top), charge trapping and de-trapping and their impact on V_{TH} (bottom).

In equation 1.1, $\Delta V_{TH_{max}}$ represents the maximum amount of shift in V_{TH} that can be achieved through charge trapping process. Since, for a fixed device temperature, there are a finite number of traps in the gate dielectric, the amount of shift in V_{TH} saturates after these traps are filled. In equation 1.1, τ_0 and β are fitting parameters used to model the exponential saturation behavior of the trap filling process as a function of time t. It is shown that β ranges from 0.25 to 0.5 by increasing the temperature of the device and τ_0 decreases from $10s$ to $20ms$ as a logarithmic function of temperature. $\Delta V_{TH_{max}}$, itself is empirically shown to depend on the temperature T and gate bias voltage V_G according to equation 1.2:

$$\Delta V_{TH_{max}} = de^{gT} \left(\frac{V_G}{V_0} \right)^m \tag{1.2}$$

where, $d \sim 100nV$, $g \sim 0.02K^{-1}$, $m \sim 7$, and $V_0 = 1V$. Equation 1.2 indicates that $\Delta V_{TH_{max}}$ increases exponentially as a function of temperature and it also increases with the gate voltage [12].

Electron trapping increases V_{TH} of the CTT (called Programming or PRG), while detrapping reduces its V_{TH} (called Erase or ERS) as shown in Figure 1.1. CTT can be programmed and erased multiple times (>10,000), retain the shift in V_{TH} (ΔV_{TH}) for over 10 years at $125°C$, and is scalable down to $7nm$ nodes [12, 13].

1.3 Dissertation Organization

Chapter 2 discusses the design of an embedded Nonvolatile SRAM (eNVSRAM), along with the experimental results of a chip that was taped-out in GlobalFoundries 22FDSOI CMOS technology. In addition to eNVSRAM, I present a few other projects that I have worked on during my PhD studies. Chapter 3 introduces a nonvolatile circuit tuning technique in the post-fabrication stage by changing V_{TH} of NMOS transistors. To demonstrate this concept, a ring-oscillator was taped-out and nonvolatile tunability of its frequency has been tested and reported. Chapter 4 reports design of a supply-fluctuation resilient SRAM that retains the data up to several seconds after losing the power. This was done by adding a capacitor and 4 additional transistors to a 6T-SRAM core. Chapter 4 also presents a low-power ASIC chip for detecting heart-rate and missing beats. In this work to minimize the power consumption, several algorithmic and architectural techniques have been implemented. For example, to extract information from the sensed heart signal, a low-power method has been developed that does not rely on computationally intensive operations such as FFT, or even a division or multiplication. This design was taped-out in $0.5\mu m$ AMI technology.

CHAPTER 2

Embedded Nonvolatile SRAM (eNVSRAM)

2.1 Introduction

In this chapter, we report a $1kb$ embedded non-volatile SRAM macro (eNVSRAM), where the content of the 6T-SRAM core is stored in two additional logic NFETs in a non-volatile fashion. The non-volatile storage mechanism relies on charge trapping in the high-k gate dielectric of the two NFETs that are used as CTTs. The operation principle is based on adding an intentional mismatch to the bitcells, so they power up to their last state. This design consists of analog and digital blocks that will be covered in the next sections. The experimental results of the eNVSRAM chip that is fabricated in $22nm$ FDSOI process is presented in this chapter as well.

2.2 Analog Blocks

This design contains a few analog components that are presented in the following subsections. The analog blocks are designed and laid out from transistor level using RVT (regular V_{TH}) transistors with a regular oxide thickness, except for a level-shifter, where thick-oxide LVT (low V_{TH}) transistors have been used. In this section, the design and operation of different analog blocks are explained.

2.2.1 Bitcell Design and Operation

The bitcell schematic and different modes of operation are shown in Figures 2.1 and 2.2, respectively. eNVSRAM can operate as both regular SRAM and eNVM and has five modes of operation: Read, Write, Program (PRG), RECALL, and Erase (ERS).

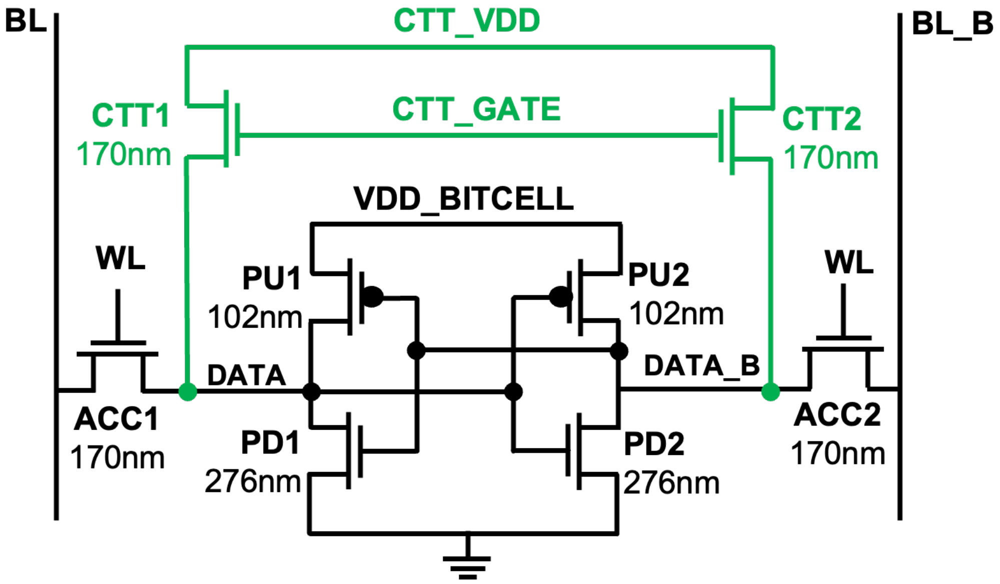


Figure 2.1: Schematics of eNVSRAM bitcell.

In this design, two CTTs are connected to DATA and DATA_B. To explain how the volatile data is saved in a non-volatile fashion, assume initially VDD_BITCELL=$0.8V$, CTT_GATE=CTT_VDD=$0V$, DATA=1, and DATA_B=0. Before turning off VDD_BITCELL, DATA needs to be stored as a V_{TH}-shift in CTTs via programming. To program a CTT, its V_{GS} and V_{DS} need to be in a specific range. By setting CTT_GATE=$2.7V$, CTT_VDD=$1.7V$, CTT2 will have the appropriate bias to be programmed, while CTT1 will not be programmed, due to having a lower V_{GS} and V_{DS} when DATA=1. Since only CTT2 is being programmed, a mismatch is created between the two CTTs, such that V_{TH}-CTT2 $>$ V_{TH}-CTT1. During PRG, DATA_B node raises to 0.4V since most of the PRG current is drained through the right pull-down NFET (PD2). Therefore, the effective V_{GS} and V_{DS} of CTT2 during

PRG are 2.3V and 1.3V, respectively. Transistor sizes for the bitcell have been optimized to ensure data stability and proper operation for all modes, across all PVT corners. Also, VDD_BITCELL is raised to $1.0V$ during PRG, to prevent any possible data disturb. PRG is done two rows at a time to minimize the surge of total PRG current, which is about $16mA$/row.

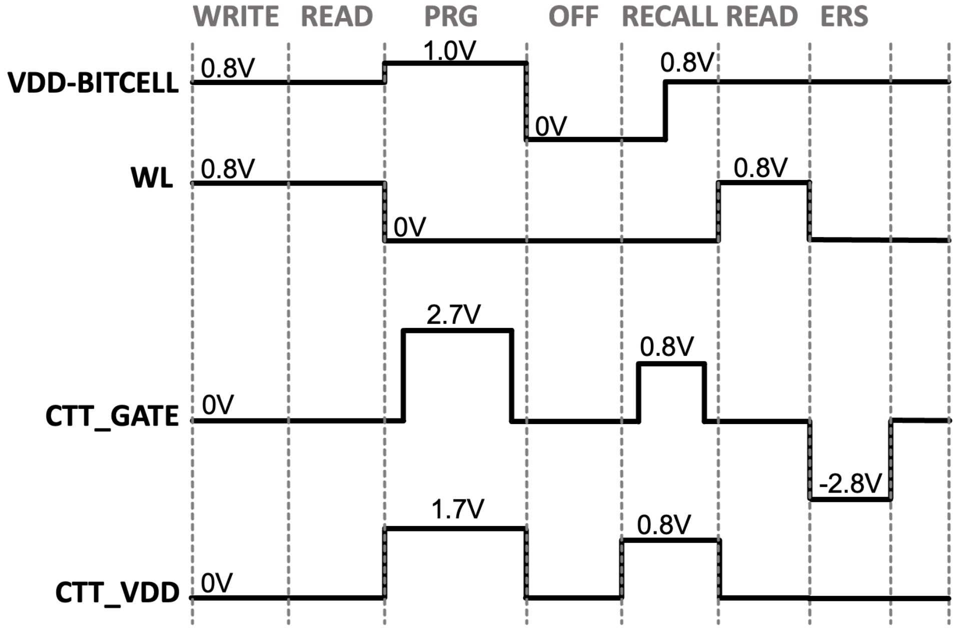


Figure 2.2: Different modes of operation and the corresponding voltages of eNVSRAM.

After PRG, the memory can be turned off. When VDD_BITCELL is turned back ON, CTT1, which has a lower V_{TH} than CTT2, becomes a faster pull-up and restores “1” at node DATA. This operation is called RECALL. During RECALL, first CTT_GATE and CTT_VDD are raised to $0.8V$ and after a brief delay, VDD_BITCELL is raised to $0.8V$ from $0V$. This delay ensures that the bitcell boot-up state is determined by the V_{TH} mismatch in CTTs. Direction of this V_{TH} mismatch is dictated by the bitcell original data. RECALL is performed in one-shot for the whole array, enabling a fast bus-free data restoration, in contrast with conventional SOCs, where data movement between NVM/eNVM and SRAM is done sequentially using a bus. This bus-free approach makes the data transfer secure against malicious activities by removing the interface ports that are vulnerable to side-channel attacks. CTTs can be erased after RECALL to reuse them in the next PRG cycle (ERS mode). During ERS, a negative V_{GS} is applied to CTTs to repel the trapped electrons electrostatically and reduce their V_{TH} back to almost original levels. ERS operation does not consume any power since CTTs are off.

Figure 2.3 shows the bitcell layout, where no DRC waiver is used. The commercial memories use DRC waivers to reduce their size. The bitcell dimensions are $0.208\mu m$ by $4.126\mu m$, which is a long and skinny design to minimize the area. For further area reduction, one CTT_GATE wire is shared with two rows. This configuration makes it possible to fit the entire row (consisting of 32 cells) within only two poly pitches. Each poly pitch in this process is $104nm$. The bitcell size is about 5.6× larger than Global Foundries standard 6T-SRAM, which uses extensive DRC waivers to shrink the layout. These waivers were not available to us at the time of the design. It is estimated that, eNVSRAM bitcell area is only 20% larger than the foundry 6T-SRAM IP when no DRC waiver is used in the IP. The bitcell layout is designed such that a PRG current of about $500\mu A$ can be delivered to each bitcell without any issues. To reduce the wire length and minimize the IR-drop, the PRG current is delivered horizontally to the bitcells.

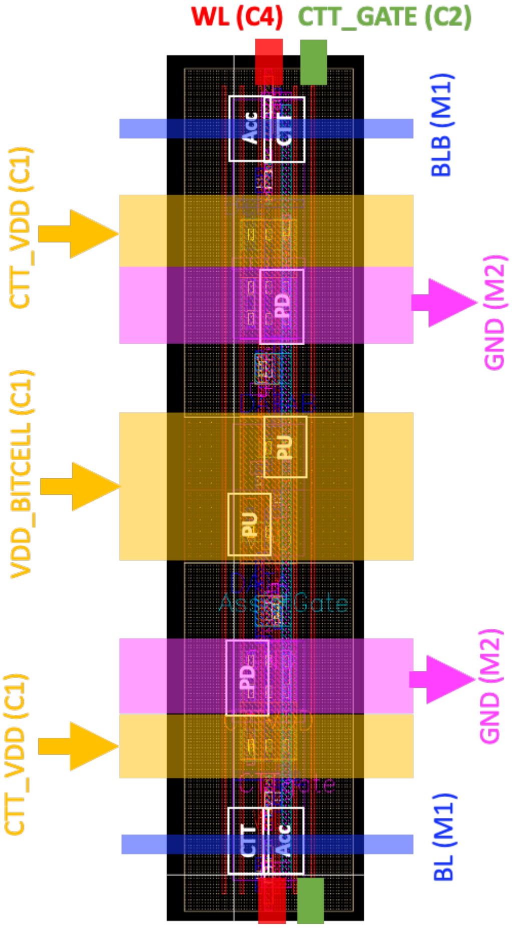


Figure 2.3: Layout of the Bitcell.

2.2.2 Precharge Circuit and Sense Amplifier

Figure 2.4 and 2.5 show the schematics and layout of precharge circuit respectively. This circuit consists of two PFETs, where PRECH_B signal is applied to their common gates. When PRECH_B is $0V$ (precharge phase) both BL and BL_B are pulled up to VDD. A shunt (PMOS-EQ) is added to equalize BL and BL_B and improve the speed of pre-charge operation.

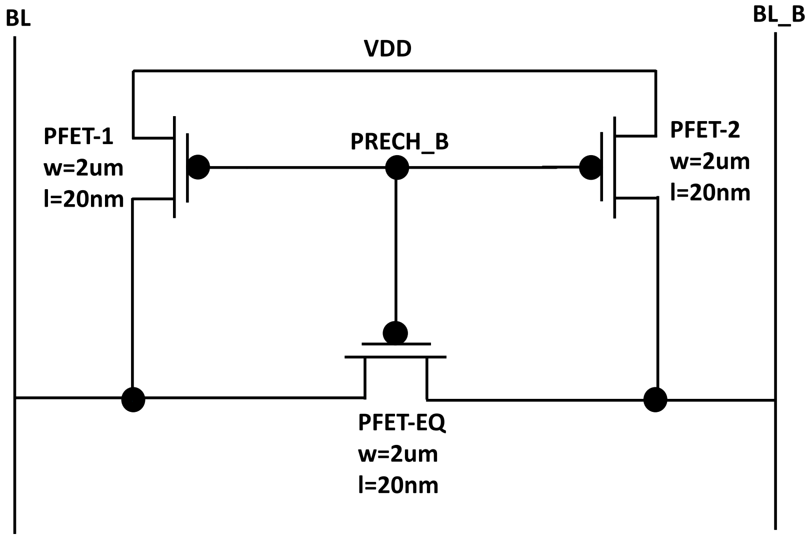


Figure 2.4: Schematics of precharge circuit.

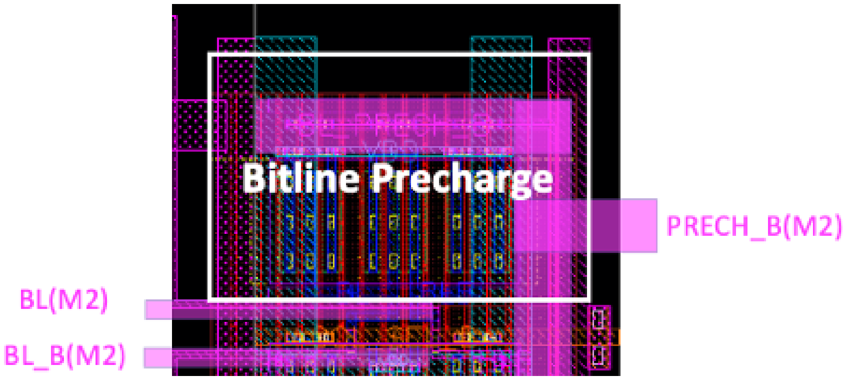


Figure 2.5: Layout of precharge circuit.

Figures 2.6 and 2.7 show the Sense Amplifier (SA) schematics and layout respectively. Every column has a dedicated SA, resulting in a total of 32 SAs in this design. A separate precharge circuit is used for the SA to pull up SENSE and SENSE_B to VDD. During the READ operation, the differential signal between BL and BL_B is transferred to SENSE and SENSE_B and amplified by SA. The sensed signal is first latched and then transferred to a shift register before sending it out of the chip, as shown in Figure 2.8. Since READ operation is performed one row (32 bits) at a time, to save on the number of pads, the read data of one row is stored in a 32-bit shift register and then sent out serially using one pad.

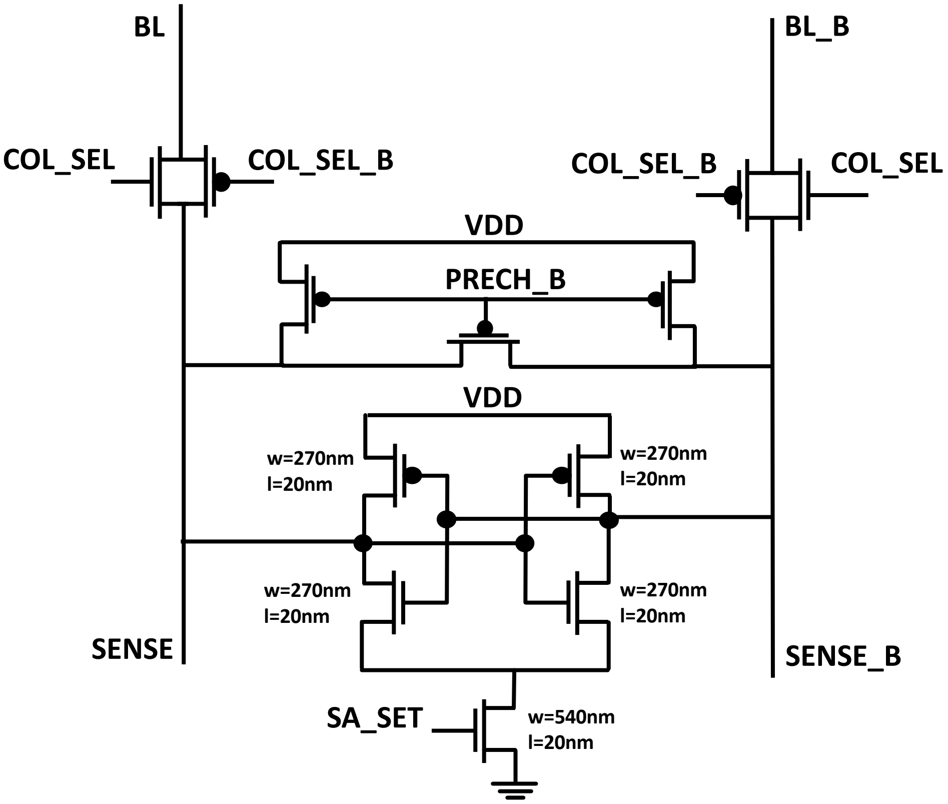


Figure 2.6: Schematics of the sense amplifier.

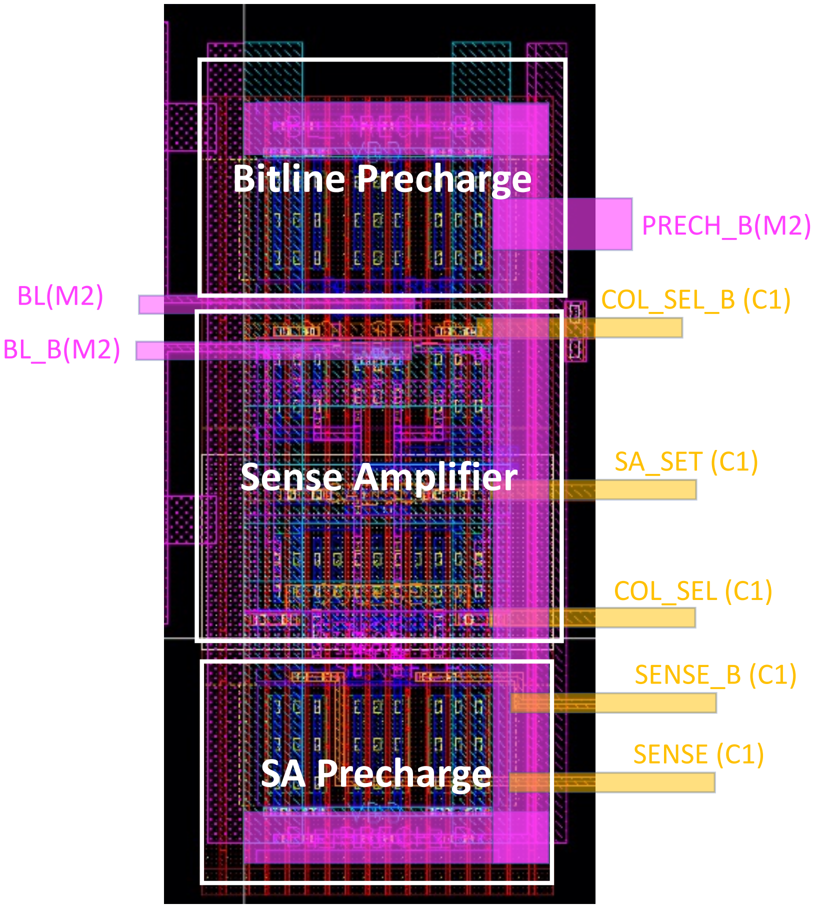


Figure 2.7: Layout of the sense amplifier.

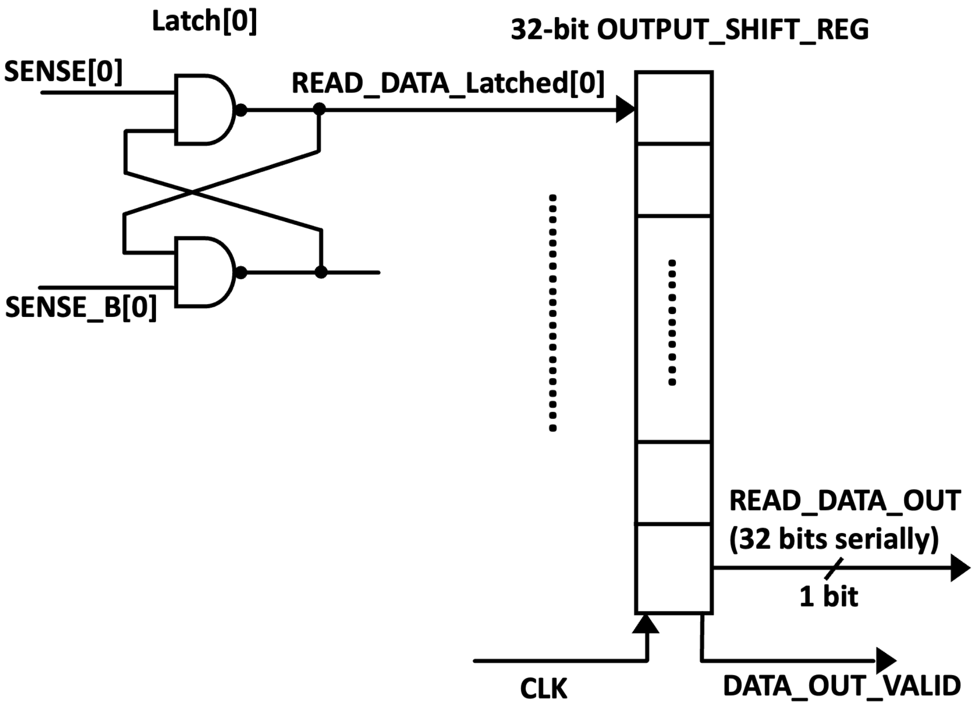


Figure 2.8: Schematics of the READ path.

2.2.3 Write Driver

Figure 2.9 shows the write driver schematics and Figure 2.10 shows its layout. To reduce the number of pads, only one pad is used for getting the external data into the chip. Since WRITE operation is performed 32 bits at a time, a 32-bit shift register is used to receive the external data serially and transfers them in parallel to 32 write drivers. Every write driver is connected to a BL/BL_B pair. Since WRITE_DATA is passed to BL/BL_B via two NFETs, cross-coupled PFETs have been used to maintain a full VDD level on the BL or BL_B that passes ”1” during WRITE operation.

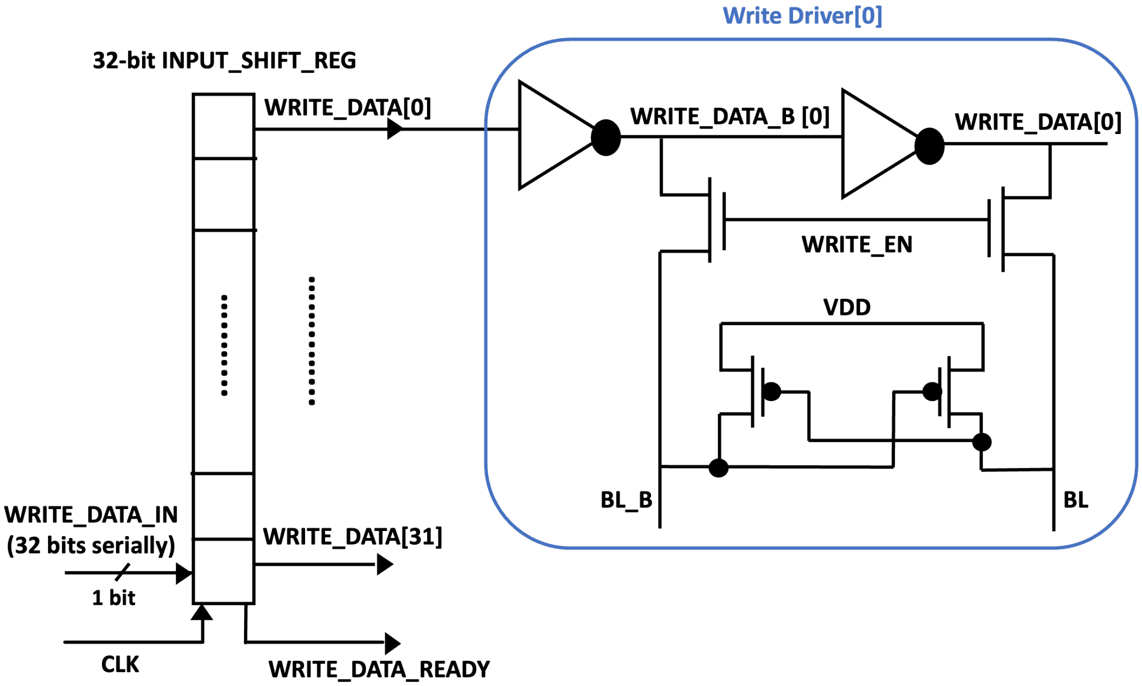


Figure 2.9: Schematics of the write driver.

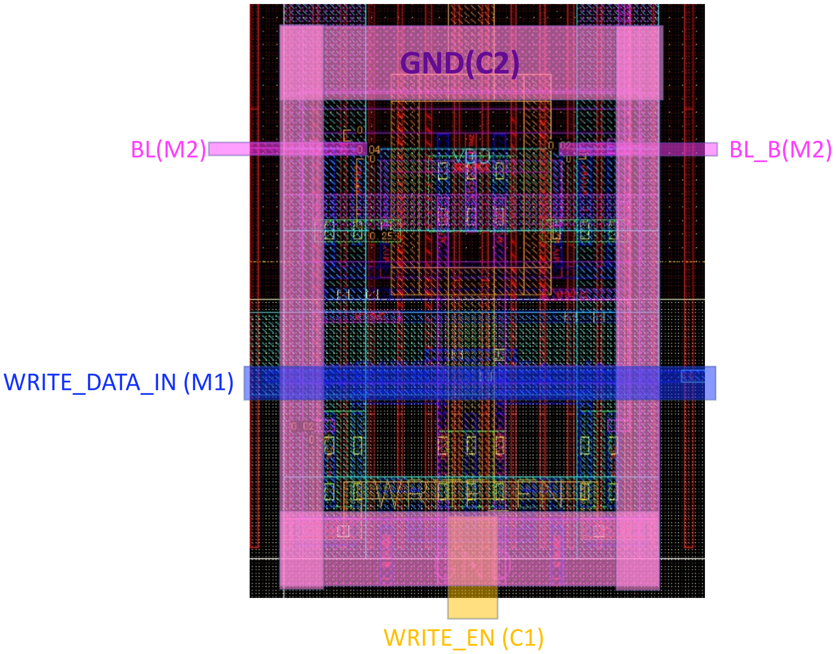


Figure 2.10: Layout of the write driver.

2.2.4 Level Shifter

A level shifter is designed to generate the appropriate voltages for CTT_GATE during PRG/RECALL/ERS. Figure 2.11 shows the level shifter's schematics and Figure 2.12 shows its layout. In this level shifter, VDD-PRG and VSS-ERS are supplied externally, and CTRL is generated internally by the digital control unit. Table 2.1 shows level shifter's input and output voltages in different modes of operation. This level-shifter uses thick-oxide I/O transistors that are rated up to $3.3V$.

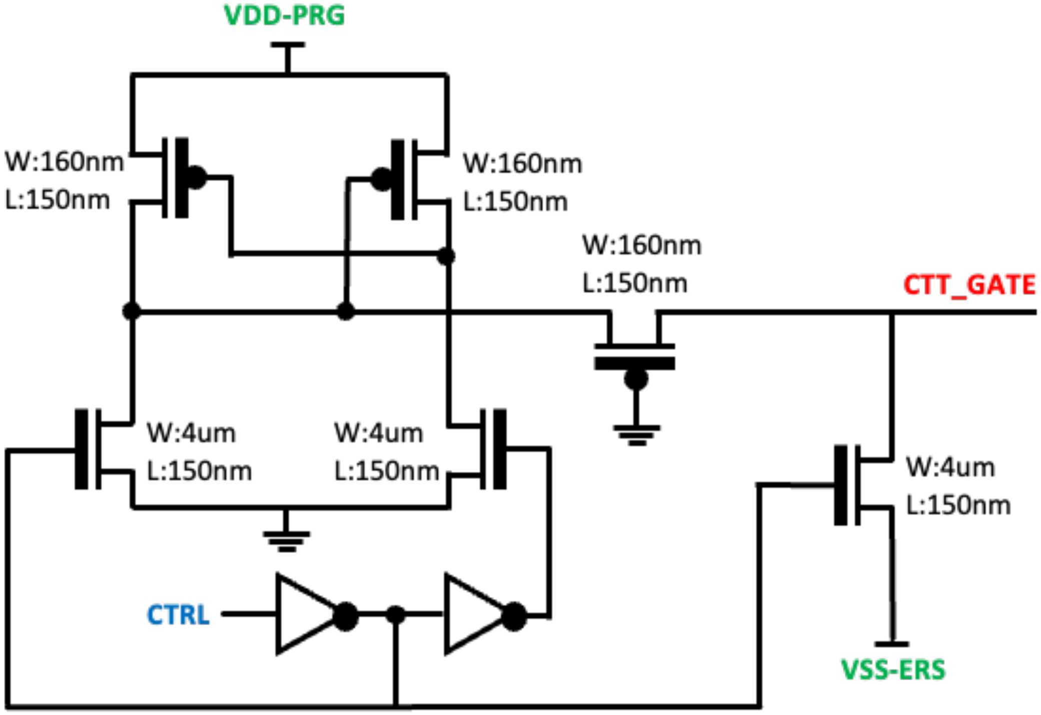


Figure 2.11: Schematics of the Level Shifter.

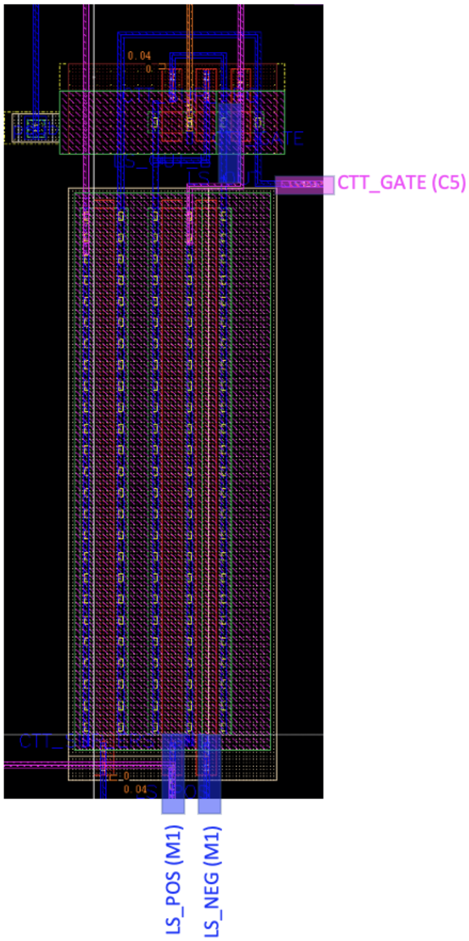


Figure 2.12: Layout of the Level Shifter.

	PRG	RECALL	ERS	OTHER (Read/Write/No-Op)
VDD-PRG (V)	2.7	0.8	0.8	0
VSS-ERS (V)	0	0	-2.8	0
CTRL (V)	0.8	0.8	0	0
CTT_GATE (V)	2.7	0.8	-2.8	0

Table 2.1: Level shifter's input and output voltages in different modes of operation.

2.3 Array Structure and Layout Considerations

Main components of the design are shown in Figure 2.13. Figure 2.14 shows the layout of array and periphery, which includes a 32 × 32 array, 32 write drivers, 16 level shifters, 32 sense amplifiers, 32 pairs of column switches, and 32 precharge circuits. The dimensions of the entire layout is $30\mu m$ by $132\mu m$, where the 1kb array occupies an area of $6.7\mu m$ by $132\mu m$.

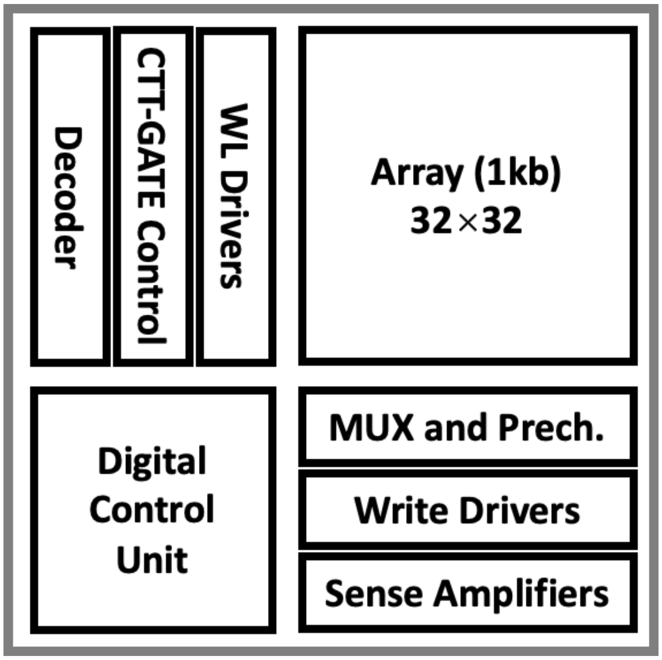


Figure 2.13: Block Diagram of the macro.

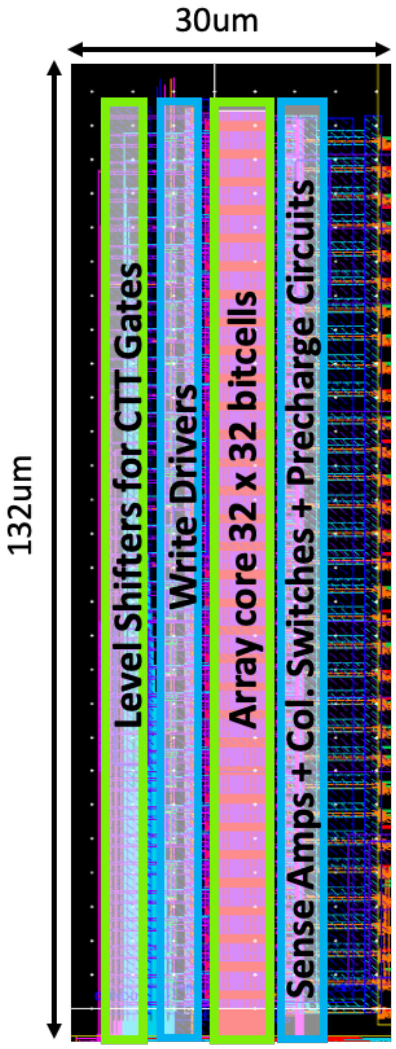

30um
132um
Level Shifters for CTT Gates
Write Drivers
Array core 32 x 32 bitcells
Sense Amps + Col. Switches + Precharge Circuits

Figure 2.14: Layout of the array.

In this layout, wordlines are vertical and bitlines are horizontal. Figure 2.15 shows the layout of a section of the array, where the top portion shows lower metal wirings and the bottom portion shows the higher metal wirings of the same section. In this design, a wide JA layer is used for wiring the GND, CTT_VDD, and VDD_BITCELL. At $100°C$, the DC current limit for a $5\mu m$-wide JA layer is $60mA$, while the peak current limit is $217mA$ for the same width. These limits are more relaxed at lower temperatures. The metal resistance per unit length of a $5\mu m$-wide JA layer is $2.7m\Omega/\mu m$. This results in a total resistance of

0.35Ω across the entire array ($132\mu m$ long). In this design, the maximum DC current is set by the PRG current of the bitcells. Since two rows are programmed at the time, the total PRG current does not exceed 32mA. The estimated IR drop on a $132\mu m$ long and $5\mu m$ wide JA is $11mV$, when it carries $32mA$.

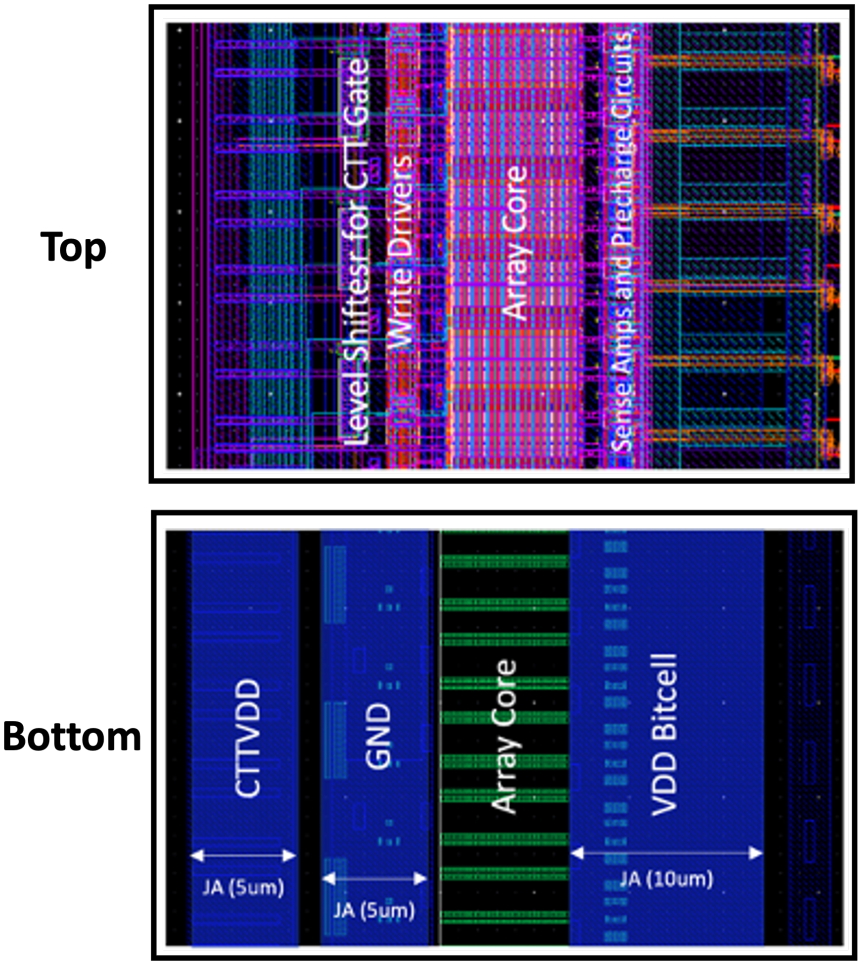


Figure 2.15: Layout of a section of the array.

In the layout of the topcell, extra attention is paid in the pad placement and routing them to the array. This is particularly important for GND, VDD_BITCELL, and CTT_VDD pads that draw large current. These pads are placed close to the array to minimize the IR drop. Figure 2.16 shows the placement of these pads as well as their routing to the array. Top metal layer (OI) wirings are used to connect GND, VDD_BITCELL, and CTT_DD to the array. Figure 2.16 also shows the placement of the digital blocks, de-caps, and other macros that are taped out.

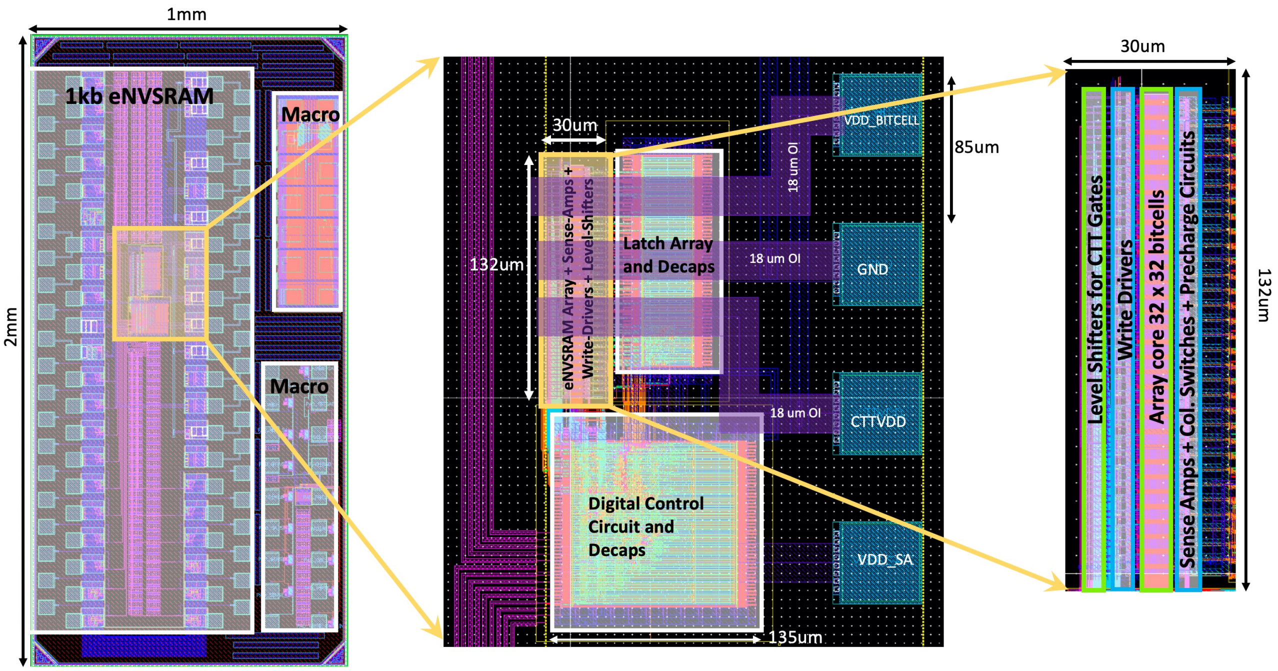
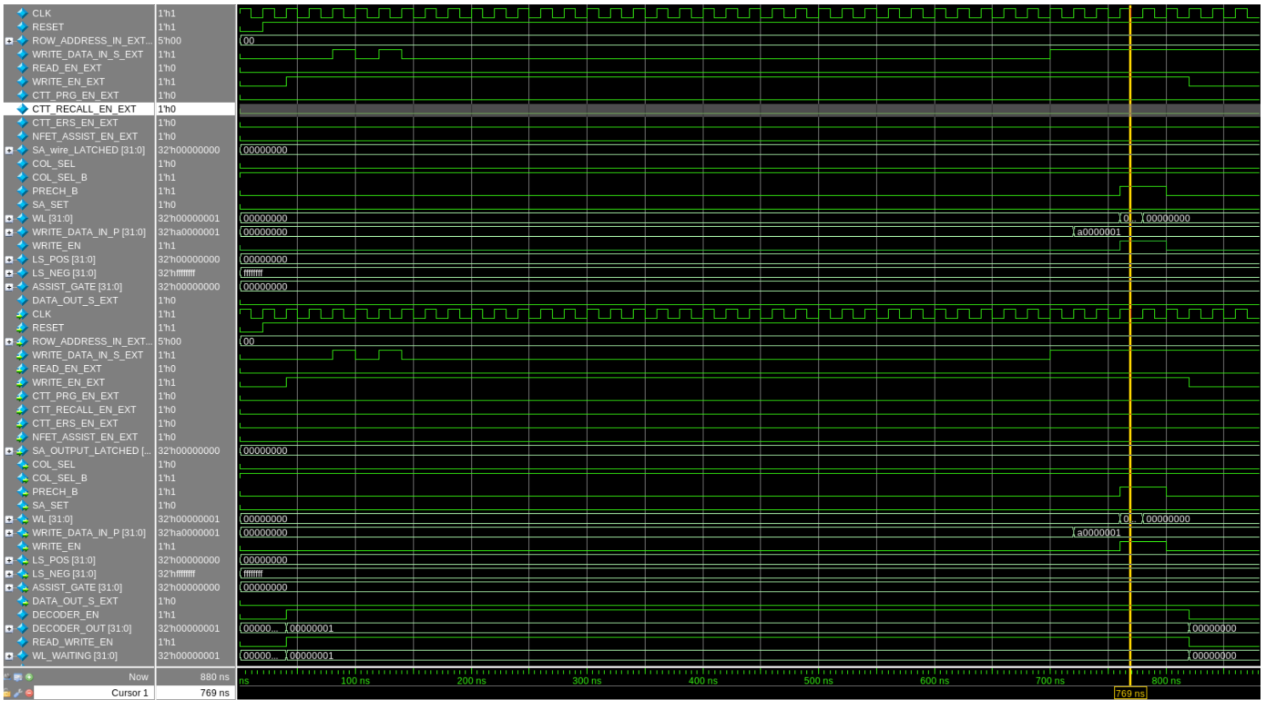

1mm
2mm
1kb eNVSRAM
Macro
Macro
30um
132um
eNVSRAM Array + Sense-Amps + Write-Drivers + Level-Shifters
Latch Array and Decaps
18 um OI
18 um OI
18 um OI
VDD_BITCELL
85um
GND
CTTVDD
VDD_SA
Digital Control Circuit and Decaps
135um
30um
132um
Level Shifters for CTT Gates
Write Drivers
Array core 32 x 32 bitcells
Sense Amps + Col. Switches + Precharge Circuits

Figure 2.16: Layout of the entire chip.

2.4 Digital Blocks

The digital components of this design are generated with an standard digital flow, which starts by RTL scripting, followed by the synthesis and place and route. The RTL scripts have been verified with ModelSim for all modes of operation and different scenarios. A sample of ModelSim simulation is shown in Figure 2.17.



CLK	1'h1
RESET	1'h1
ROW_ADDRESS_IN_EXT...	5'h00
WRITE_DATA_IN_S_EXT	1'h1
READ_EN_EXT	1'h0
WRITE_EN_EXT	1'h1
CTT_PRG_EN_EXT	1'h0
CTT_RECALL_EN_EXT	1'h0
CTT_ERS_EN_EXT	1'h0
NFET_ASSIST_EN_EXT	1'h0
SA_wire_LATCHED [31:0]	32'h00000000
COL_SEL	1'h0
COL_SEL_B	1'h1
PRECH_B	1'h1
SA_SET	1'h0
WL [31:0]	32'h00000001
WRITE_DATA_IN_P [31:0]	32'ha0000001
WRITE_EN	1'h1
LS_POS [31:0]	32'h00000000
LS_NEG [31:0]	32'hffffffff
ASSIST_GATE [31:0]	32'h00000000
DATA_OUT_S_EXT	1'h0
CLK	1'h1
RESET	1'h1
ROW_ADDRESS_IN_EXT...	5'h00
WRITE_DATA_IN_S_EXT	1'h1
READ_EN_EXT	1'h0
WRITE_EN_EXT	1'h1
CTT_PRG_EN_EXT	1'h0
CTT_RECALL_EN_EXT	1'h0
CTT_ERS_EN_EXT	1'h0
NFET_ASSIST_EN_EXT	1'h0
SA_OUTPUT_LATCHED [...	32'h00000000
COL_SEL	1'h0
COL_SEL_B	1'h1
PRECH_B	1'h1
SA_SET	1'h0
WL [31:0]	32'h00000001
WRITE_DATA_IN_P [31:0]	32'ha0000001
WRITE_EN	1'h1
LS_POS [31:0]	32'h00000000
LS_NEG [31:0]	32'hffffffff
ASSIST_GATE [31:0]	32'h00000000
DATA_OUT_S_EXT	1'h0
DECODER_EN	1'h1
DECODER_OUT [31:0]	32'h00000001
READ_WRITE_EN	1'h1
WL_WAITING [31:0]	32'h00000001
Now	880 ns
Cursor 1	769 ns

Figure 2.17: A sample of ModelSim simulation for RTL verification.

For synthesis and place and route, Synopsys Design Compiler and Cadence Innovus are used, respectively. Post-layout simulations are performed at the top-level (including all digital and analog components) across PVT corners using Cadence Virtuoso for all modes of operations.

The digital portion of the design consists of combinatorial and sequential components. The sequential components are activated with a positive edge of the clock or reset. A Finite State Machine (FSM) is used to generate the necessary signals for controlling the operations. Figure 2.18 shows the FSM state transition diagram. When the reset signal, which is asynchronous, is activated, FSM goes to the Idle state at any point during the operation. During non-volatile modes (PRG, RECALL, ERS) FSM stays at the Idle state.

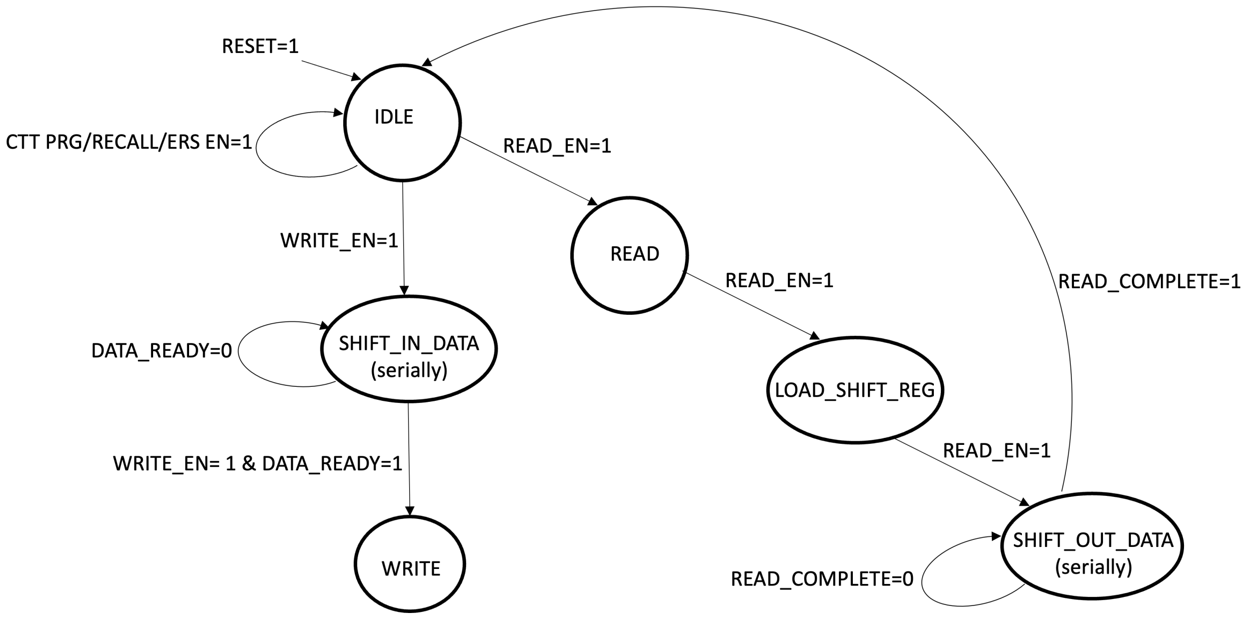


Figure 2.18: FSM state transition diagram.

Since the array has 32 rows, a 5-bit row decoder is used to activate the wordlines for READ or WRITE operations. The same decoder is also used during PRG to activate CTT_GATEs of two rows at a time.

To facilitate the chip testing and debugging with a small number of pads, two 8x1 multiplexers (MUXes) are added to monitor 16 important internal nodes, as shown in Figure 2.19 and Table 2.2.

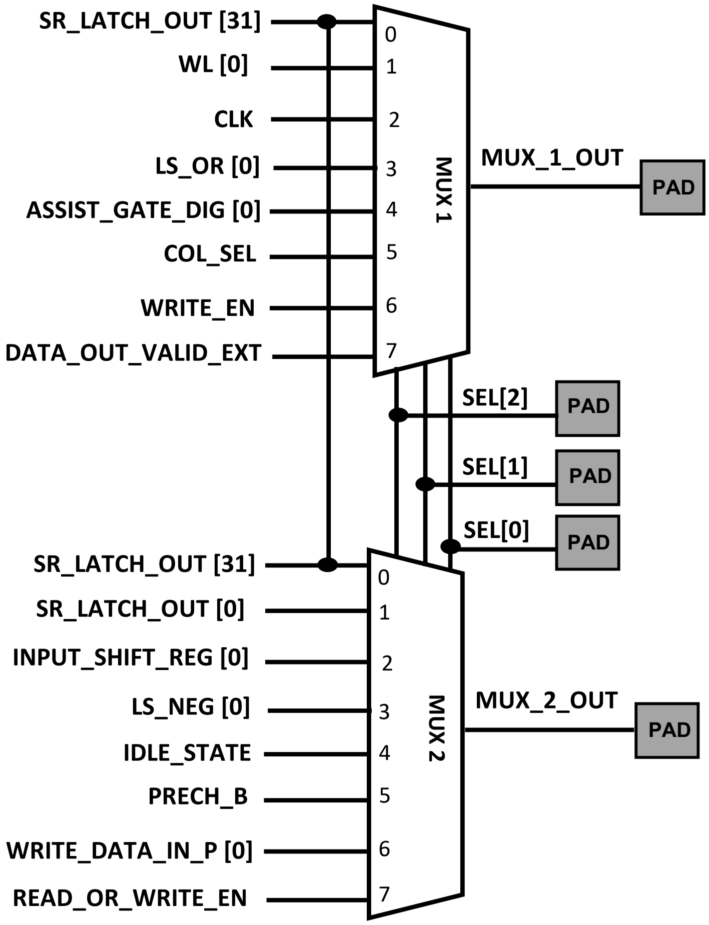
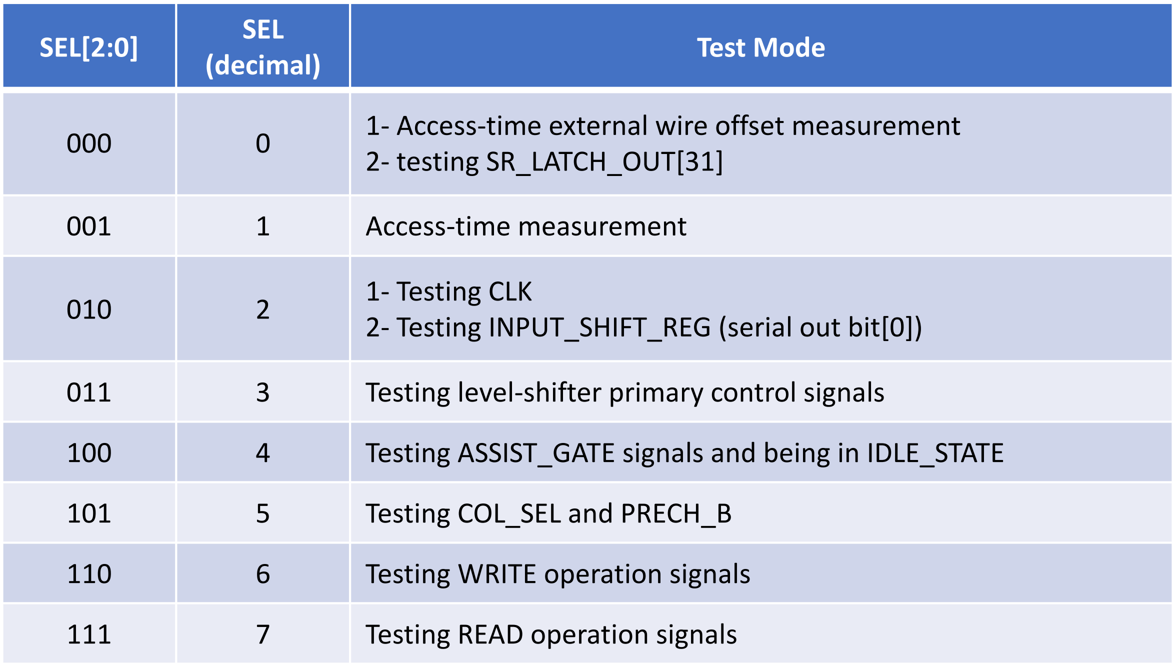


Figure 2.19: Two 8x1 multiplexers are used for monitoring internal nodes.



SEL[2:0]	SEL (decimal)	Test Mode
000	0	1- Access-time external wire offset measurement 2- testing SR_LATCH_OUT[31]
001	1	Access-time measurement
010	2	1- Testing CLK 2- Testing INPUT_SHIFT_REG (serial out bit[0])
011	3	Testing level-shifter primary control signals
100	4	Testing ASSIST_GATE signals and being in IDLE_STATE
101	5	Testing COL_SEL and PRECH_B
110	6	Testing WRITE operation signals
111	7	Testing READ operation signals

Table 2.2: SEL values vs. multiplexers' outputs.

In one case, when SEL: 000, both MUXes send the same signal to their outputs, so the delay offset between the two MUX outputs, caused by external wirings, can be measured. By knowing this offset, small delays between the two signals coming out of two MUXes can be measured. This method helps to measure the READ speed using differential delay technique, which is $550ps$ at VDD$= 0.8V$. In this case, after measuring the offset by setting SEL: 000, the time between the rising edges of WL[0] and SR_LATCH_OUT[0] is measured by setting SEL: 001. This is the worst-case READ speed that is calculated from the earliest WL[0], since it is tapped at the output of the WL driver, to the latest data coming out of the READ path. SR_LATCH_OUT[0] is the latest data, since it is coming from a bitcell that is farthest from WL driver.

2.5 Post-layout Simulation Results

All digital and analog blocks have been verified individually and also after integrating them for all modes of operations. The design top-level has been simulated in cadence virtuoso at the transistor level. Since CTTs operate in voltage ranges that do not have a foundry-provided model, for verification of the design, a custom model is developed based on experimental characterizations.

To minimize the risk of the failure, the entire chip is simulated after post-layout RC extraction. This is done for all modes of operations across all process, voltage and temperature corners. This is specially important when the RC delay of a wire is large. For example, the wordline wire for a single row has a parasitic resistance of $1.1K\Omega$ and capacitance of $35fF$. Also the total parasitic capacitance of access transistors in one row is 240fF. These parasitic capacitances and resistances correspond to an RC delay of about $300ps$. This value matches with the post-layout simulation result shown in Figure 2.20. In this simulation, the transient voltages at different points on a single WL wire are shown. The fastest waveform corresponds to the point that is closest to the WL driver, and the slowest waveform corresponds to the point that is farthest from the WL driver.

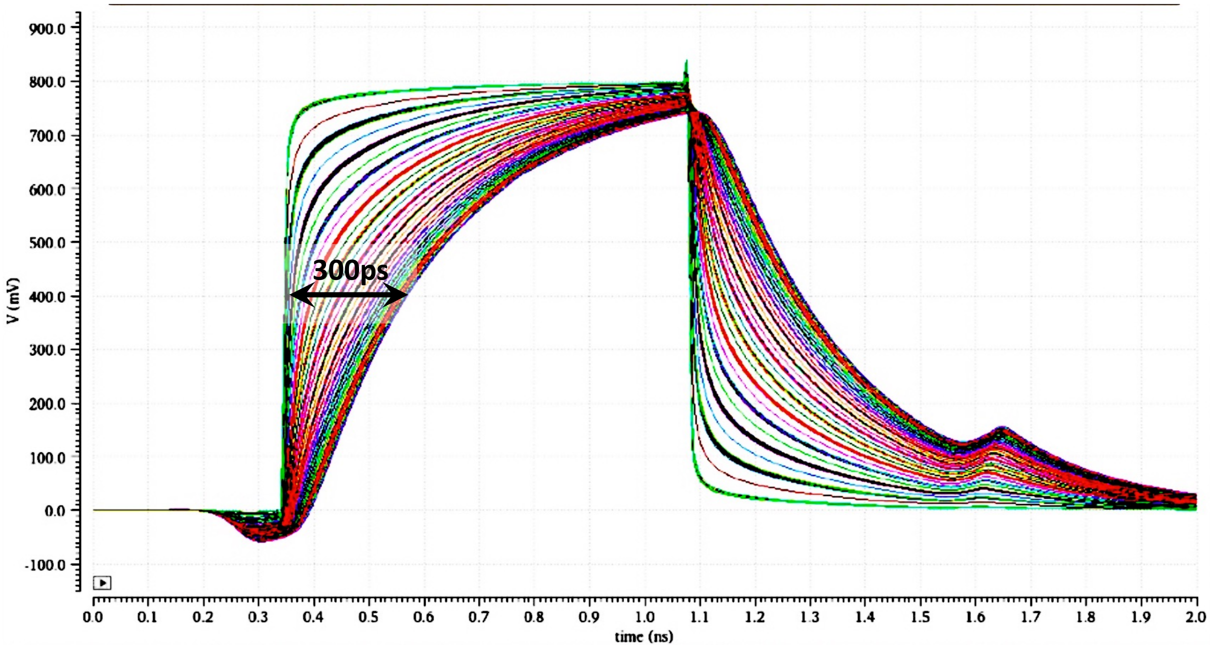


Figure 2.20: Post-layout simulation of single WL at different points from the beginning to the end of a row.

Figure 2.21 shows the post-layout simulation of a single row during READ operation, where the propagation delay of the WL is passed to the bitlines. In this design, there are 32 output latches (SR_LATCH_OUTPUTs) that are loaded with the data of 32 bitcells on a single row during READ. Due to the parasitic delay of the wirings and transistors, the data on SR_LATCH_OUTPUTs become available at different times. The output of the first SR_LATCH_OUTPUT becomes available 188ps before the output of the last SR_LATCH_OUTPUT. The post-layout transient results reported in Figure 2.21 are generated for nominal supply voltage 800mV and temperature ($25°C$).

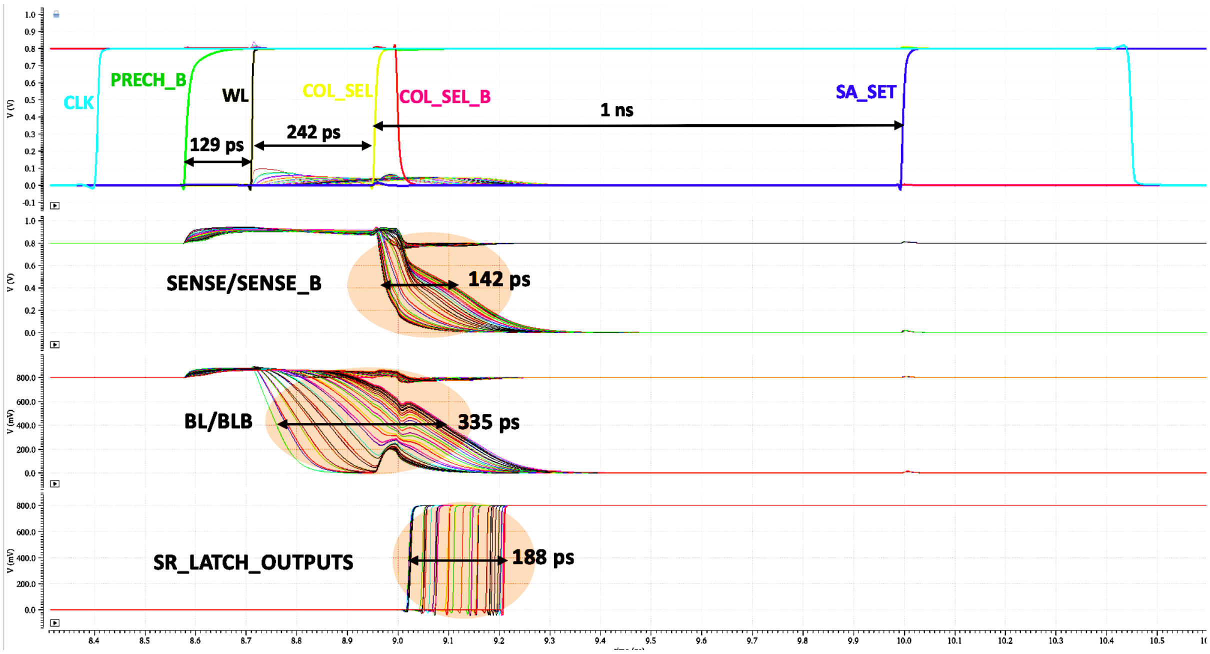


Figure 2.21: Full array post-layout simulation for READ operation at nominal supply voltage $800mV$ and temperature $(25°C)$.

In addition to the post-layout simulation at the typical corner (TT), nominal voltage $(800mV)$, and temperature $(25°C)$, we also verified the operation at Temp: -40$°C$, 25$°C$, 125$°C$, VDD: 720mV, 800mV, 880mV, across all corners (TT, FF, FS, SF, SS). Figure 2.22 shows the impact of temperature and supply voltage variation on important signals during READ operation. This plot shows that the timing of the WL signal, measured at the output of the WL driver, may vary by 210ps due to the PVT variation. Also, the timing of the Column Select signal (COL_SEL) may vary by 295ps.

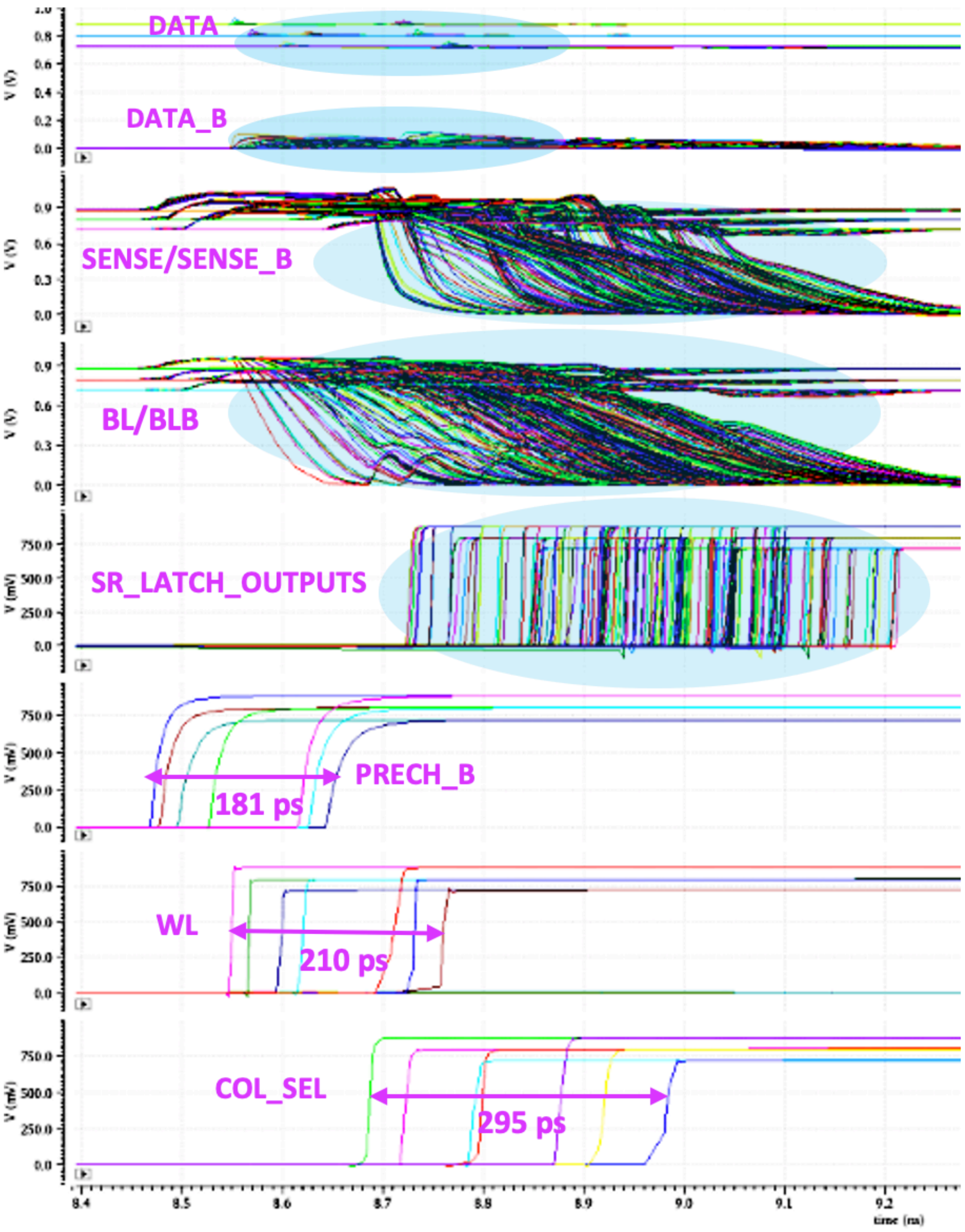


Figure 2.22: Full array post-layout simulation for READ operation showing the impact of temperature and supply voltage variation on important signals. (Temp: -40°C, 125°C and VDD: 720mV, 800mV, 880mV.)

During the READ operation, it is important to make sure that data are not disturbed.

READ disturb can happen when the voltage of a low storage node (e.g., DATA=0) rises too much during READ when the access transistor shorts DATA node to a high BL, pre-charged to VDD (0.8V in this design). When access transistor is too strong or pull-down NFET is too weak, READ disturb may happen. In this design, the size of these transistors are optimized to prevent a READ disturb across all PVT corners. Figure 2.23 shows that DATA and DATA_B remain stable at supply voltages VDD: 500mV, 600mV, 700mV, 800mV and corners: FF, FS, SF, SS, and TT.

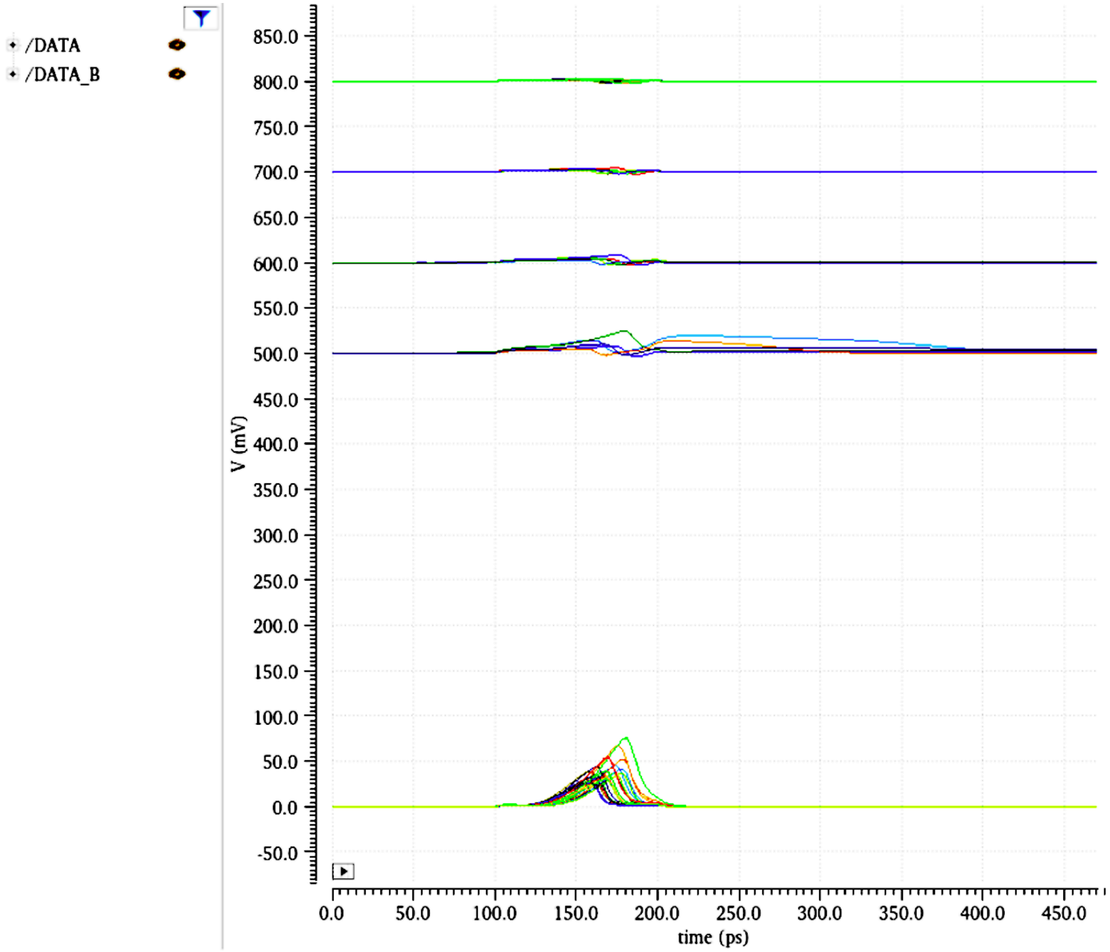


Figure 2.23: READ disturb simulations at VDD: 500mV, 600mV, 700mV, 800mV and corner: FF, FS, SF, SS, and TT.

Figure 2.24 shows the post-layout simulation for WRITE operation for the top-cell, where

WRITE_EN is asserted 190ps after PRECH_B is deactivated. Based on this simulation, the data is written in about 330ps after WL is activated.

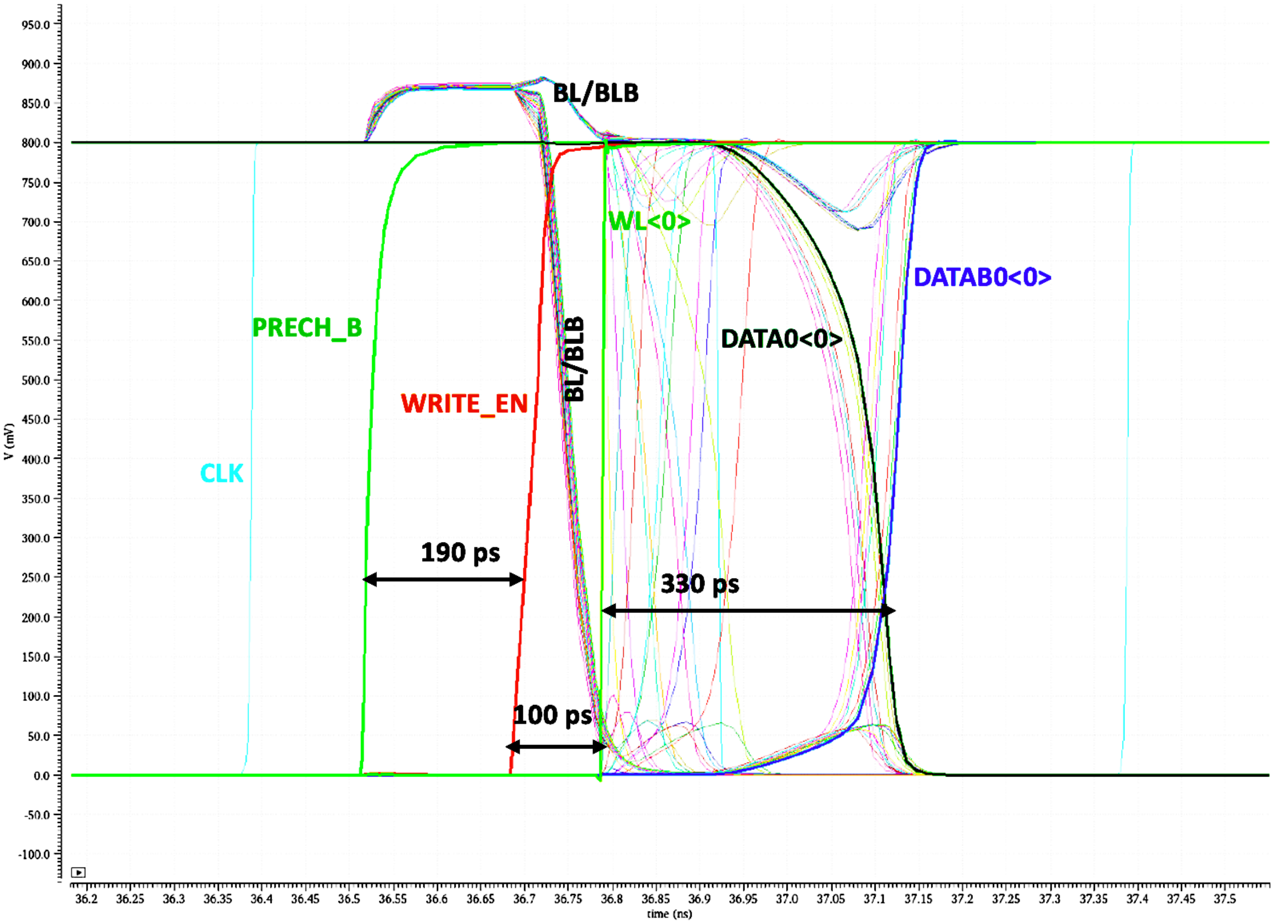


Figure 2.24: Top-cell post-layout simulation for WRITE operation.

During the regular SRAM Read/Write operation, CTTs are kept OFF, hence the Static Noise Margin (SNM) is not impacted. Figure 2.25 shows SNM of eNVSRAM compared to a conventional 6T-SRAM, where they overlap each other.

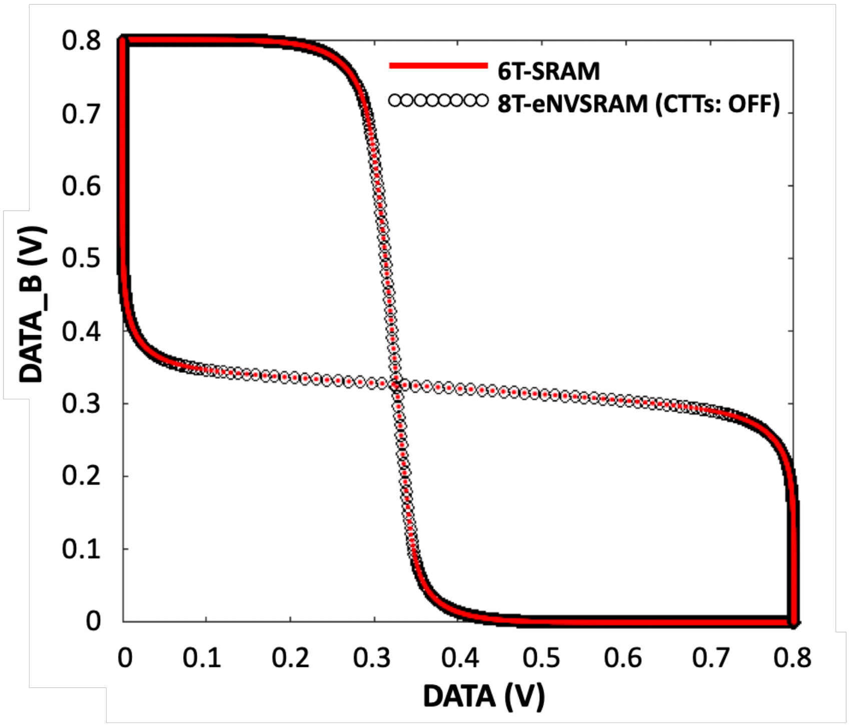


Figure 2.25: SNM comparison between eNVSRAM and 6T-SRAM.

2.6 Monte Carlo Simulation to Estimate RECALL Success-rate

In this section, the success-rate of RECALL is simulated to estimate the amount of V_{TH} shift (ΔV_{TH}) that needs to be created by PRG operation for a successful RECALL. To do so, a voltage source is placed at the gate of one of the CTTs to model the ΔV_{TH} created by PRG and RECALL is simulated for an eNVSRAM bitcell. To estimate the success-rate, Monte Carlo simulation is performed that includes both process variation and mismatch. The results of this simulation is reported in Figures 2.26 and 2.27. As shown in Figure 2.27, to achieve a lower error-rate, a larger ΔV_{TH} is required, as expected. It is also observed that the error rate goes down exponentially as a function of ΔV_{TH}, as shown in Figure 2.27.

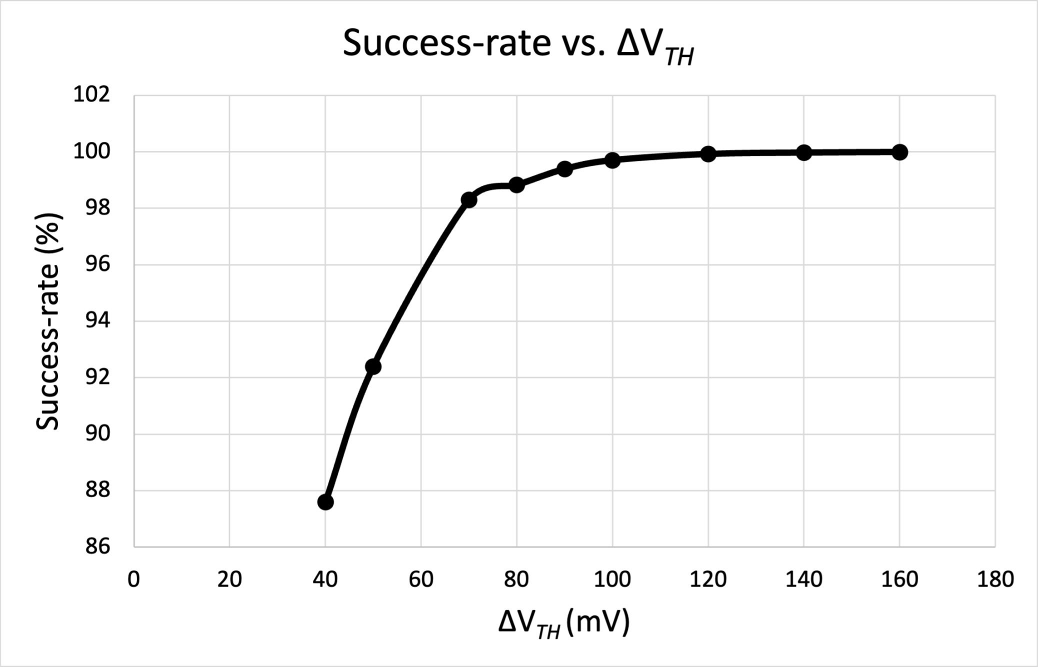


Figure 2.26: RECALL success-rate as a function of ΔV_{TH}.

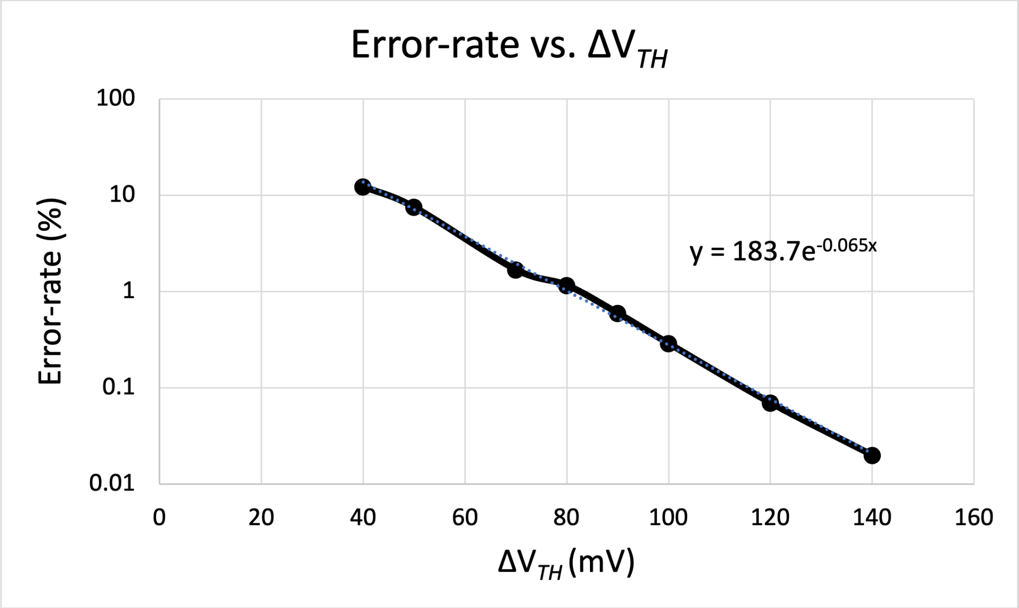


Figure 2.27: RECALL error-rate as a function of ΔV_{TH}.

2.7 CTT Characterization and Modeling

Since the foundry-provided models that are used for simulation purposes do not cover the voltage ranges needed for PRG and ERS, to verify these operations, a custom model is developed based on experimental device characterizations. In this process, a test structure consisting of regular NMOS transistors are characterized for different V_{GS} and V_{DS} voltages that are suitable for PRG and ERS operations. By comparing the I_{DS}- V_{GS} waveform of a transistor at its initial state (before PRG/ERS) with its I_{DS}- V_{GS} after PRG/ERS, we can calculate the ΔV_{TH} that is created by PRG/ERS. Figure 2.28 shows an example of such an experiment that is performed to calculate the post-PRG ΔV_{TH}. In this figure, I_{DS} vs. V_{GS} (with V_{DS}=760mV) is measured for a virgin device. Then, while V_{DS} is kept at 1V, PRG is performed with different V_{GS} values, ranging from 1.5V to 2.5V with steps of 0.1V. After each PRG, I_{DS} vs. V_{GS} is measured again (with V_{DS}=760mV) to see the impact of the PRG on I_{DS} (and therefore on V_{TH}). The measurements are repeated after 3$hours$ and it is confirmed that the results are very consistent (*i.e.* V_{TH}-shifts are stable). The maximum ΔV_{TH} is created by the largest V_{GS} (2.5V), which is 0.13V.

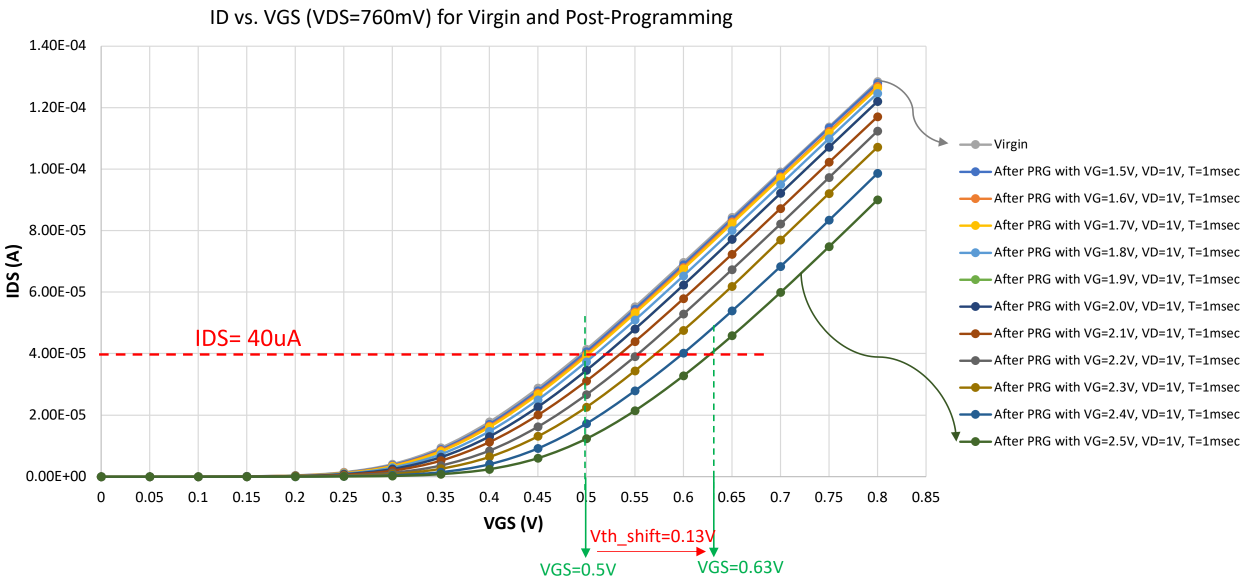


Figure 2.28: I_{DS} vs. V_{GS} is measured for virgin and programmed device to derive ΔV_{TH}.

Figure 2.29 shows characterization of I_{DS} vs. V_{DS} for high V_{GS} voltages up to $2.5V$ for a regular-V_{TH} (RVT) transistor with W=170nm and L=20nm, which is the same transistor used in the eNVSRAM design. Based on these device characterizations, a Verilog-A model is extracted for NMOS. This model is valid for V_{GS} and V_{DS} voltages that are used for PRG/ERS.

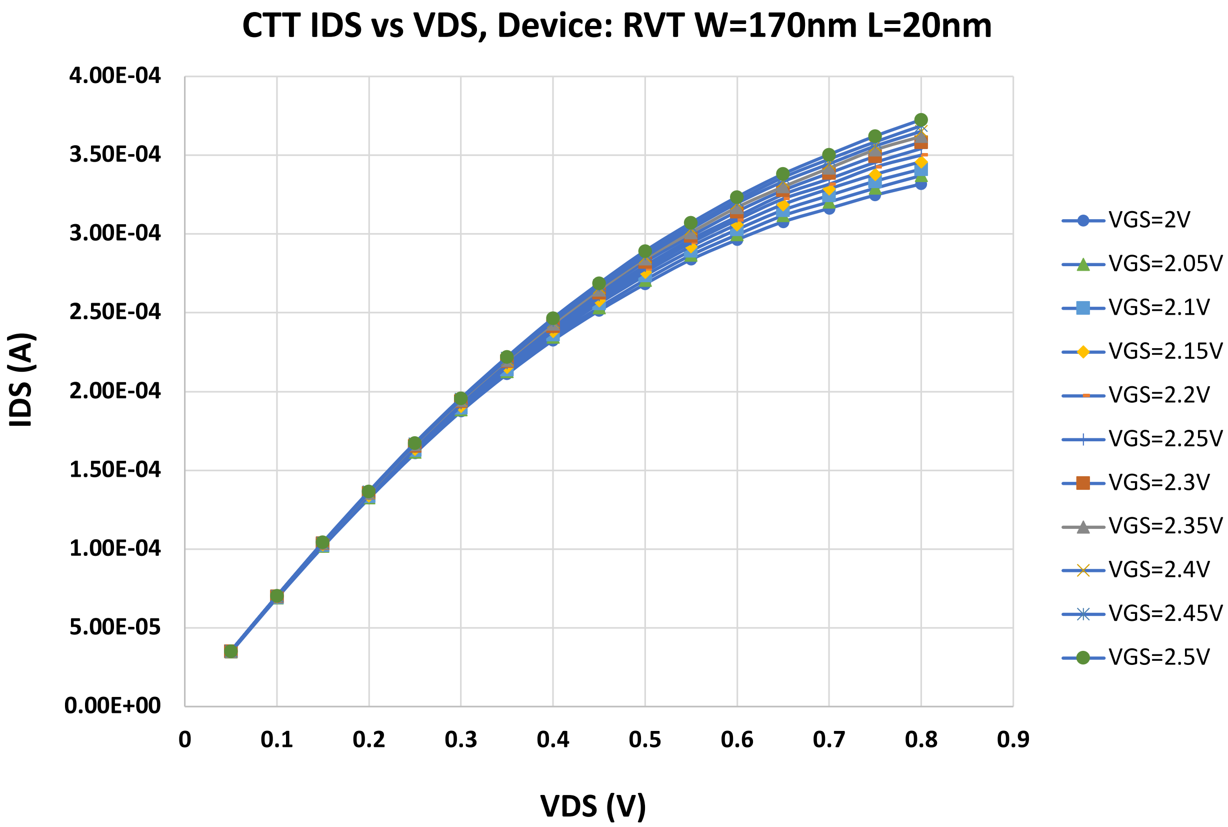


Figure 2.29: Experimental characterization of CTT transistors at high V_{GS} to derive a Verilog-A model.

The Verilog-A model is used to simulate the PRG operation and monitor DATA and DATA_B. The results are reported in Figure 2.30. In this simulation, during PRG, VDD_BITCELL is $1V$, CTT_VDD is kept at $1.6V$, and CTT_GATE is swept from $0V$ to $2.8V$. It is observed that, when CTT_GATE reaches $2.8V$, DATA node rises to $1.28V$ from $1V$, while DATA_B rises to $493mV$ from $0V$. Based on these results, at CTT_GATE=$2.8V$, the effective V_{GS} and V_{DS} of the CTT that is being programed is $2.3V$ and $1.1V$, respectively. Figure 2.31 shows the current drawn from different transistors and supplies during PRG. The maximum current drawn from CTT_VDD is $531\mu A$ for each bitcell. This value matches with the measured current drawn from CTT_VDD ($\sim 32mA$) during a single PRG attempt, which includes two rows (64 bitcells).

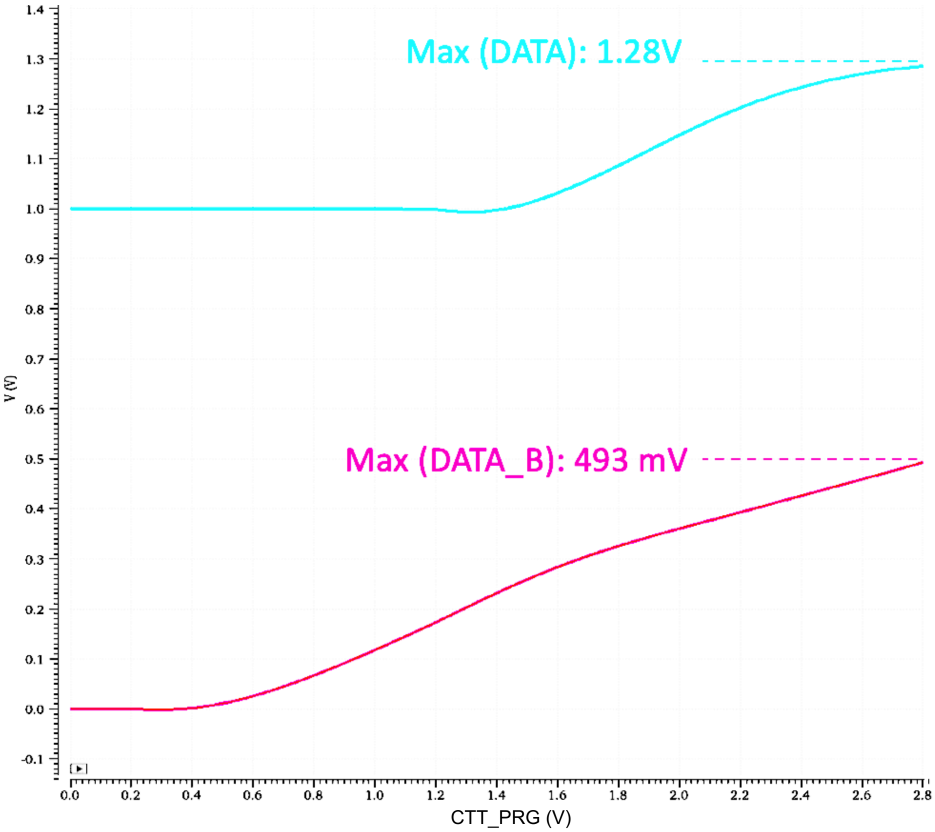


Figure 2.30: DATA and DATA_B during PRG.

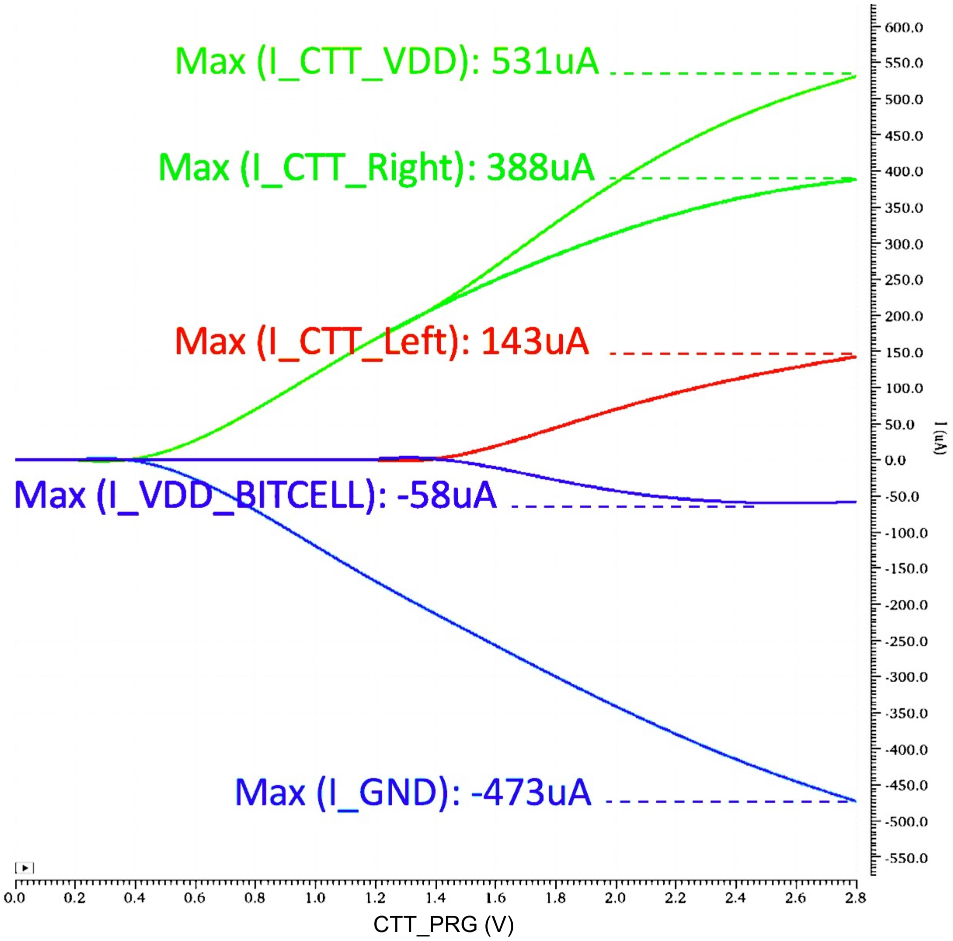


Figure 2.31: Currents of CTTs and supplies during PRG.

2.8 Measurement Results

To demonstrate the full functionality of the proposed eNVSRAM, a $1kb$ macro in $22nm$ FDSOI is designed, fabricated, and tested. Figure 2.32 shows the chip micrograph and Figure 2.33 shows the wirebonded die on a custom PCB. This macro includes a 32×32 array, write drivers, WL drivers, address decoder, precharge circuits, sense amplifiers, level shifters, and a digital control block.

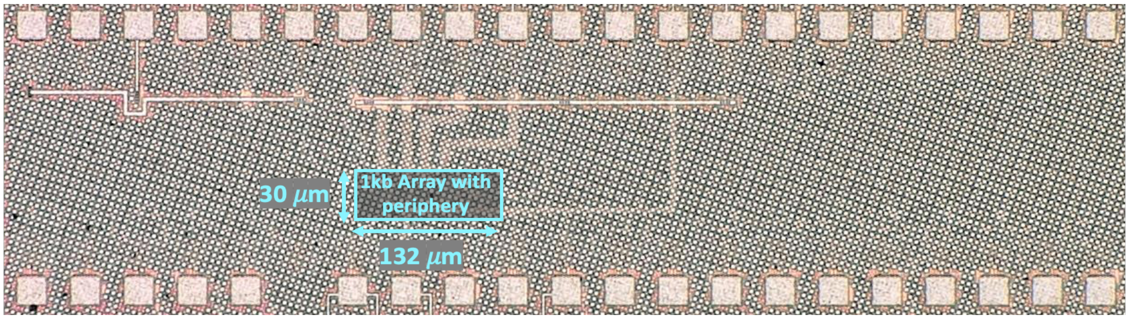


Figure 2.32: Chip micrograph.

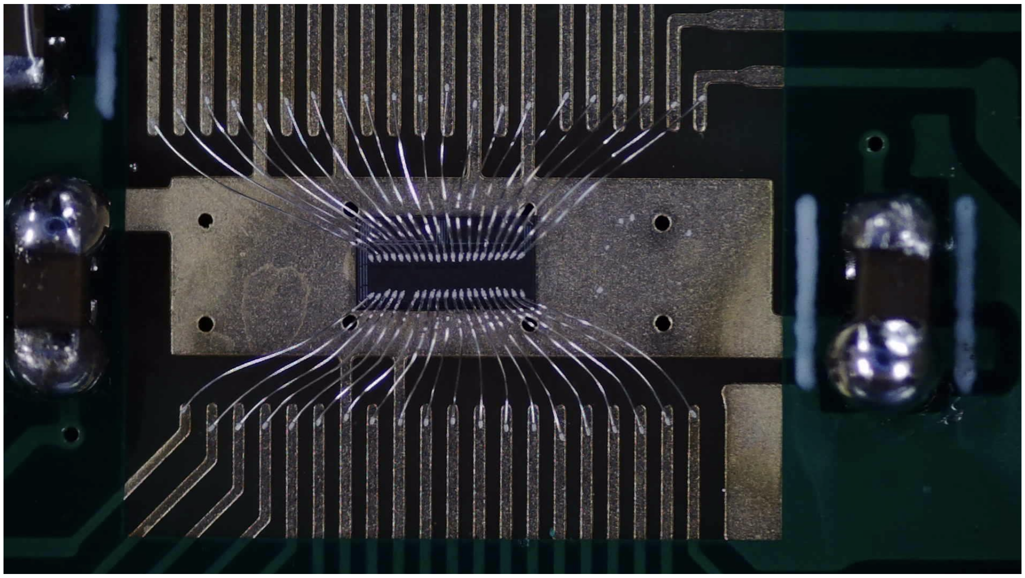


Figure 2.33: Image of a wirebonded die.

To characterize the eNVSRAM, a series of tests have been performed. These include verifying READ, WRITE, PRG, RECALL, ERS operations as well as measuring V_{min}, READ speed, and power consumption of the eNVSRAM. In the next subsections, the experimental results are reported.

2.8.1 Testing READ and WRITE Operations

To verify the successful READ and WRITE operations, various patterns are written to the memory and its content is read. Figure 2.34 shows a checkerboard board pattern that is written and read successfully. Figure 2.35 shows a custom pattern of UCLA CHIPS written and read. In all bitmap plots reported in this section, a white square represents DATA=1 and a black square represents DATA=0. After verifying proper READ and WRITE operations at the nominal supply voltage of $0.8V$, VDD_BITCELL is reduced and READ/WRITE operations are performed again. It is verified that READ/WRITE operations can be performed with VDD_BITCELL and VDD_SA as low as $0.55V$, as shown in Figure 2.36.

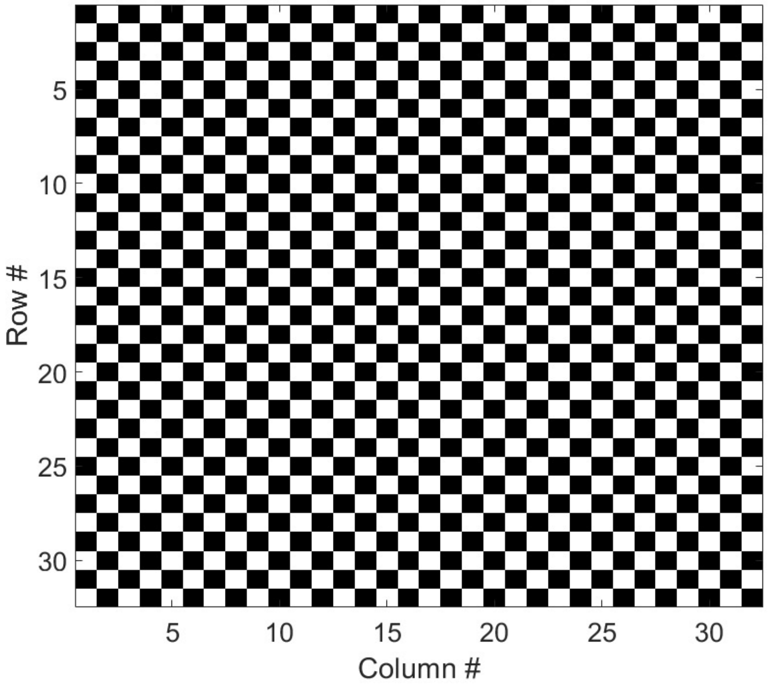


Figure 2.34: WRITE and READ of checkerboard pattern.

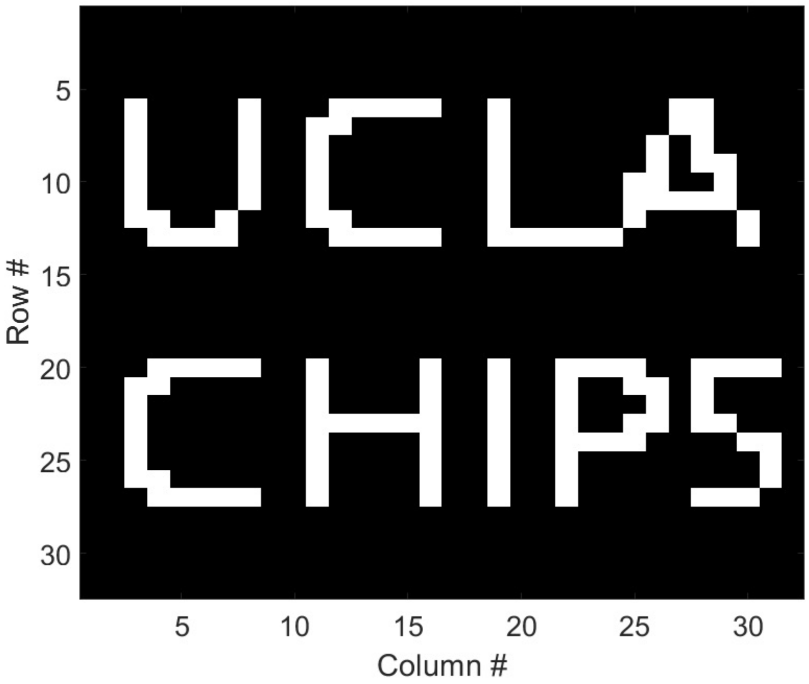


Figure 2.35: WRITE and READ of UCLA CHIPS pattern.

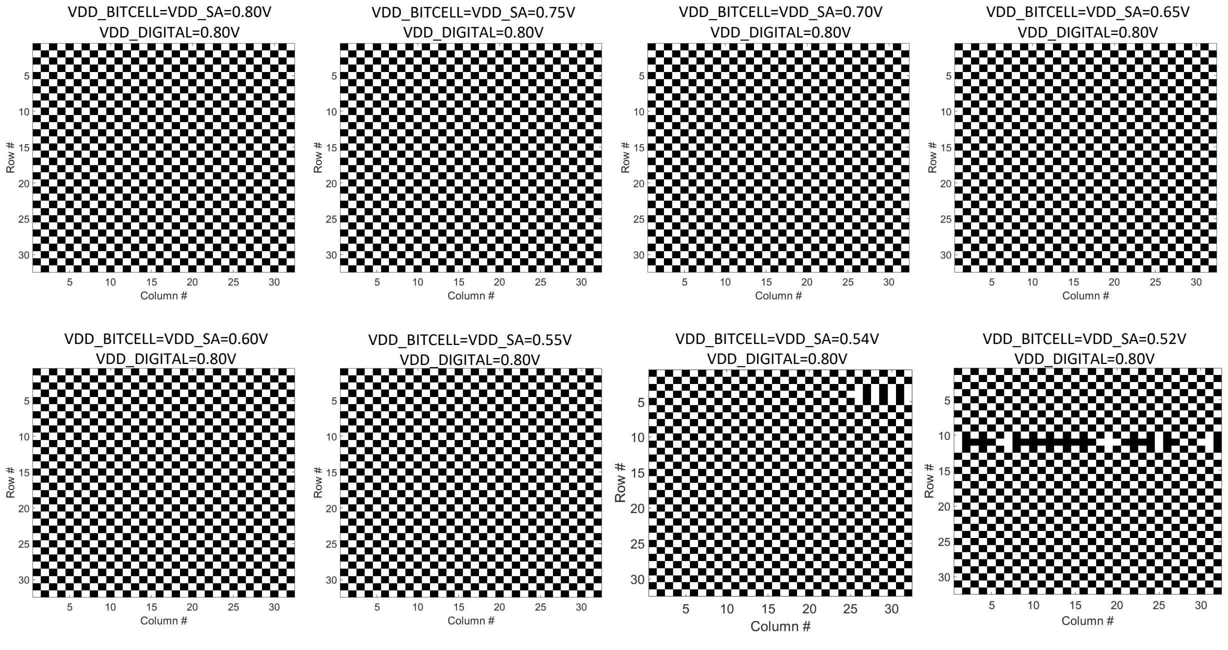


Figure 2.36: WRITE and READ of checkerboard pattern for different VDD_BITCELL.

After verifying WRITE and READ operations, speed of READ operation is measured. Figure 2.37 shows READ speed as a function of VDD, measured using differential delay technique. This is the worst case READ speed, which is measured by the time difference between the earliest WL and the latest READ DATA_OUT.

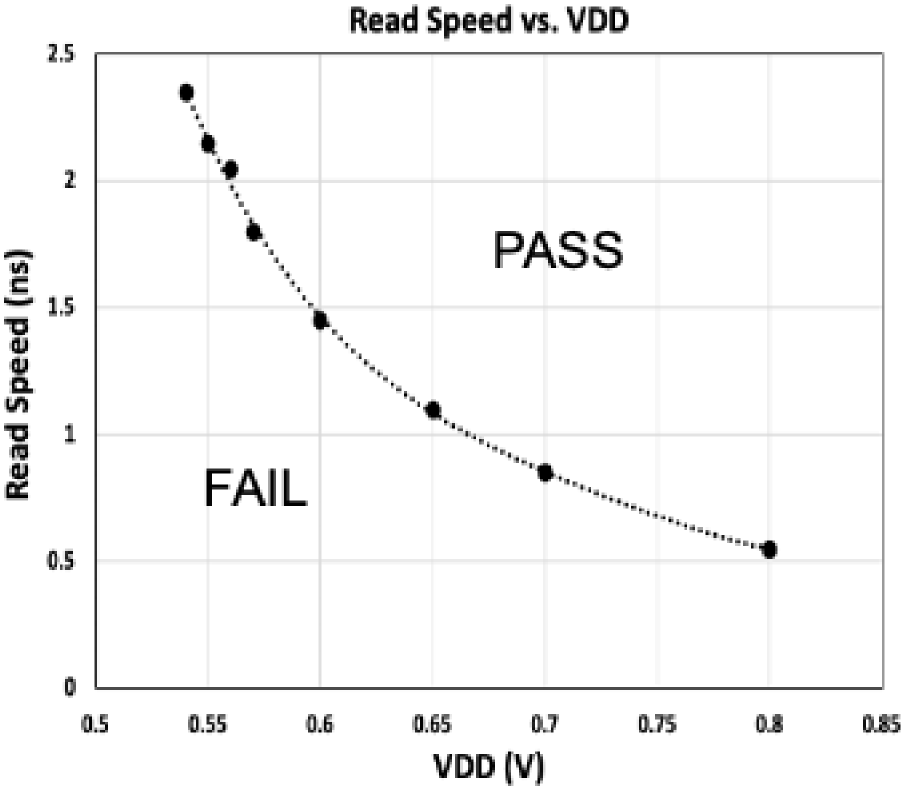


Figure 2.37: Measured READ speed versus VDD.

2.8.2 Testing PRG and RECALL Operations

Before performing PRG operation, the initial boot-up state of each die is read, to characterize the intrinsic process variation and mismatch of the bitcells. Figure 2.38 shows bitmap of the boot-up state of a single die after 10 experiments. Bitcells that their boot-up states are 0 in all 10 trials are shown in solid black, bitcells that are 1 in all 10 trials are shown in solid white, and bitcells that alternate between 1 and 0 are shown with different shades of gray, where a darker gray represents a higher number of 0 states. Figure 2.39 shows 75% of bitcells always boot-up to the same state, which means process variation impacts the boot-up state more than noise.

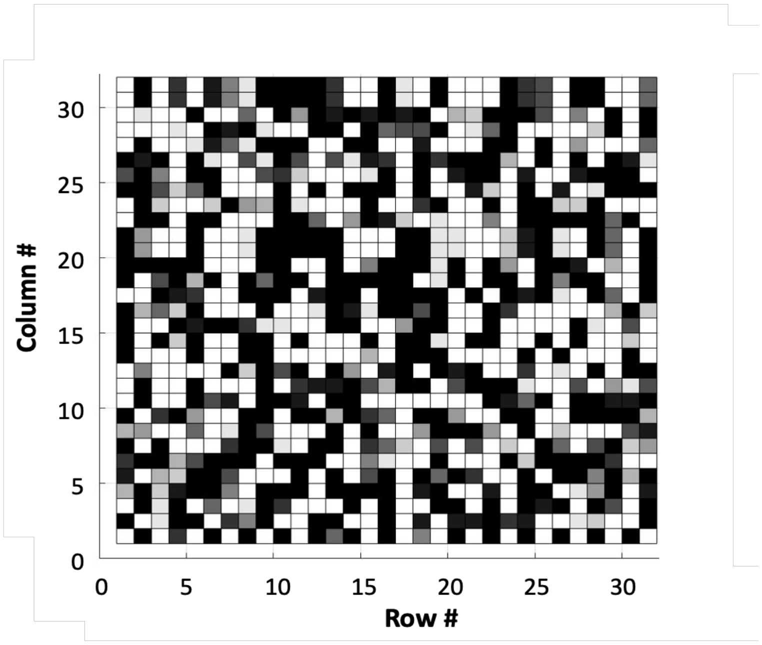


Figure 2.38: Average of boot-up state bitmaps of a single die after 10 experiments.

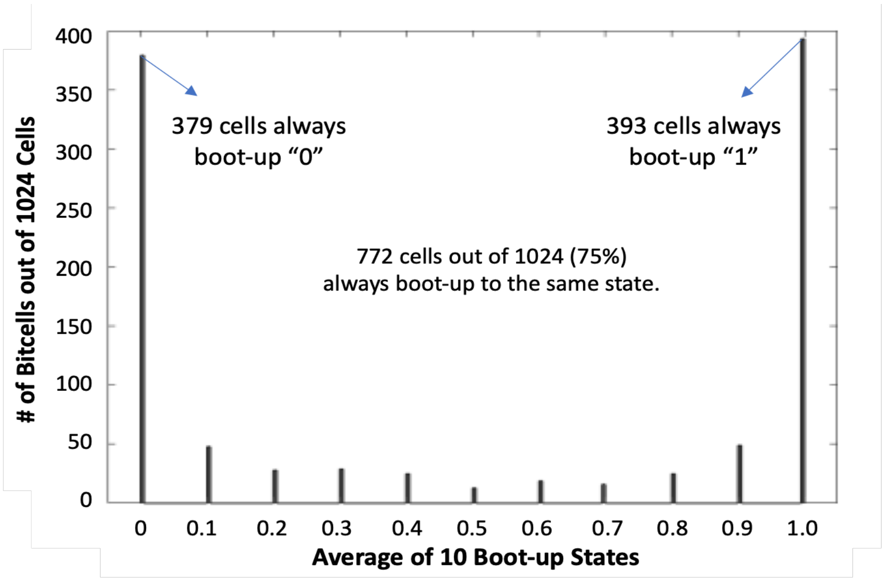


Figure 2.39: Statistics of 10 bootup states.

To optimize the programming conditions, PRG is performed with different voltages and durations across multiple dies. Each PRG operation is done on two rows at a time (64 bits). PRG success-rate is defined as % of those 64 bitcells that are recalled correctly. As shown in Figure 2.40, where X-axis is in log scale, at CTT_GATE=2.7V, CTT_VDD=1.7V, a shorter PRG time is required to achieve 100% success-rate for a single die. At this bias, the effective VGS and VDS of CTT2 is 2.3V and 1.3V, respectively due to 0.4V rise at zero storage node. The optimum PRG voltages (CTT_GATE=2.7V, CTT_VDD=1.7V) have been used to generate the statistics reported in Table 2.3, where over 100 pairs of rows across multiple dies have been tested.

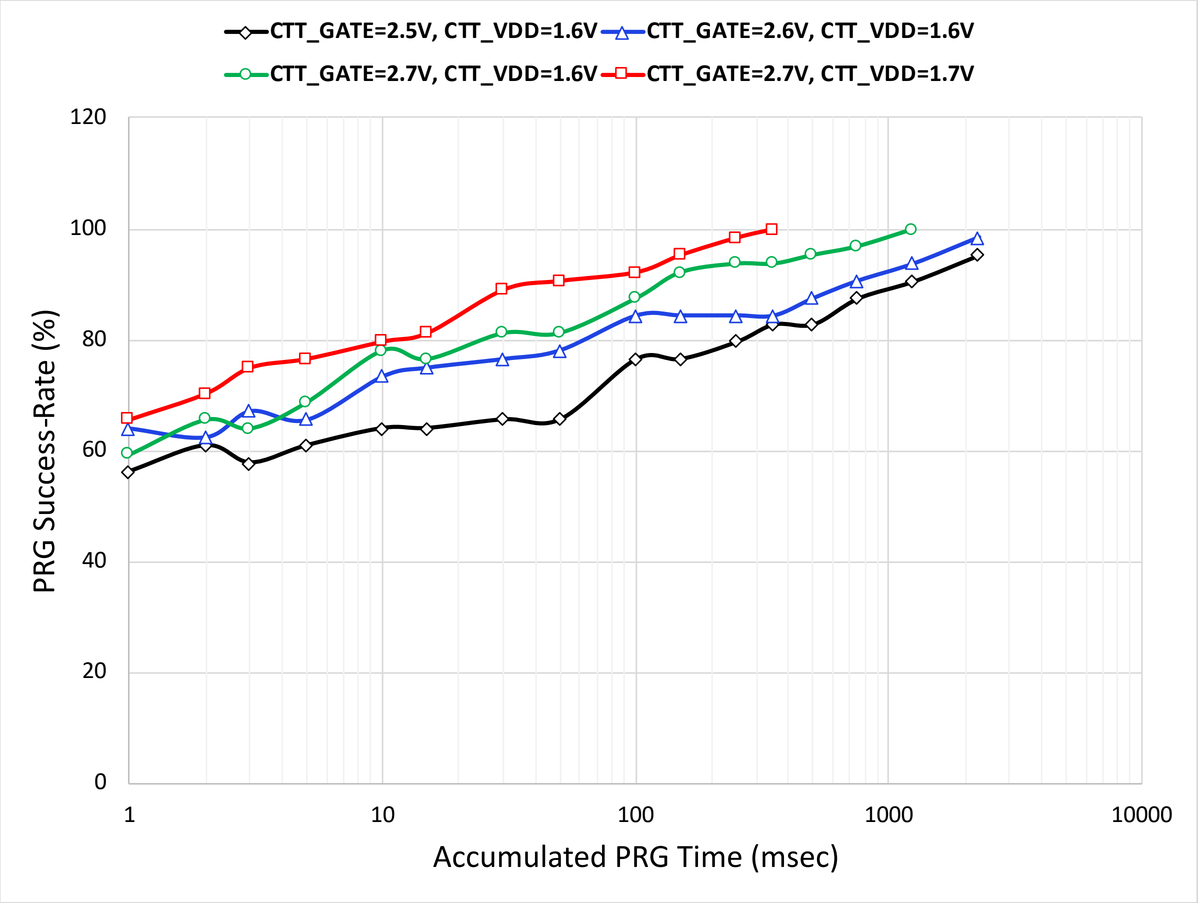


Figure 2.40: PRG success-rate for different programming time and bias conditions.

CTT_GATE=2.7V, CTT_VDD=1.7V		
PRG Time (msec)	PRG Success-Rate Mean (μ)	PRG Success-Rate STDEV (σ)
100	91.7%	1.9%
200	95.6%	1.8%
300	96.9%	2.2%
400	97.7%	1.3%

Table 2.3: The optimum PRG voltages (CTT_GATE=2.7V, CTT_VDD=1.7V) have been used to generate μ and σ of the success-rate in programming two rows.

Figure 2.41 shows the RECALL result after programming UCLA CHIPS pattern, which has a success-rate of 98.8%. This success-rate can be improved with further optimizations, such as increasing the size of CTTs, using redundancy, or utilizing error-correction methods.

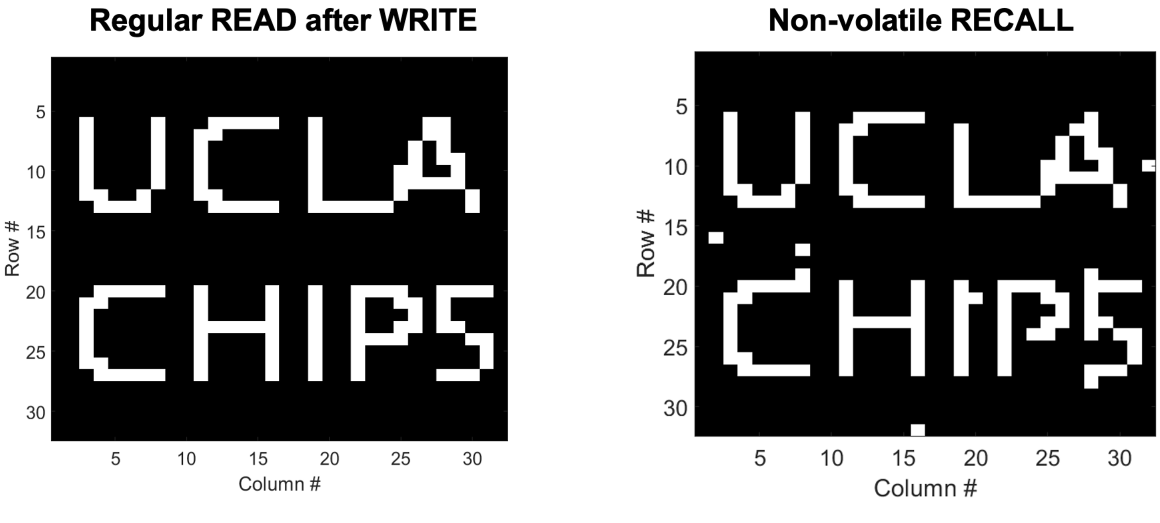


Figure 2.41: Ideal pattern (left), RECALL after PRG (right).

2.8.3 Testing Non-volatile Retention

To check the non-volatile retention, eNVSRAM is turned off for different durations after PRG, ranging from days to a few months, and then RECALL is performed. In all cases, the non-volatile state of the array remained unchanged. An example of such experiment is shown in Figure 2.42, where the non-volatile state of the array has not changed after 30 days of power-down.

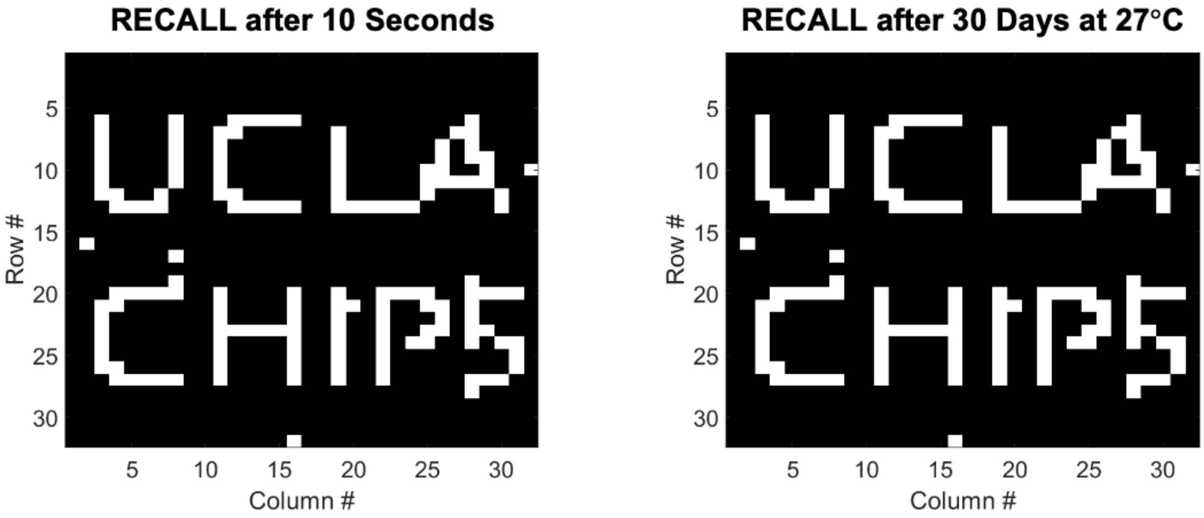
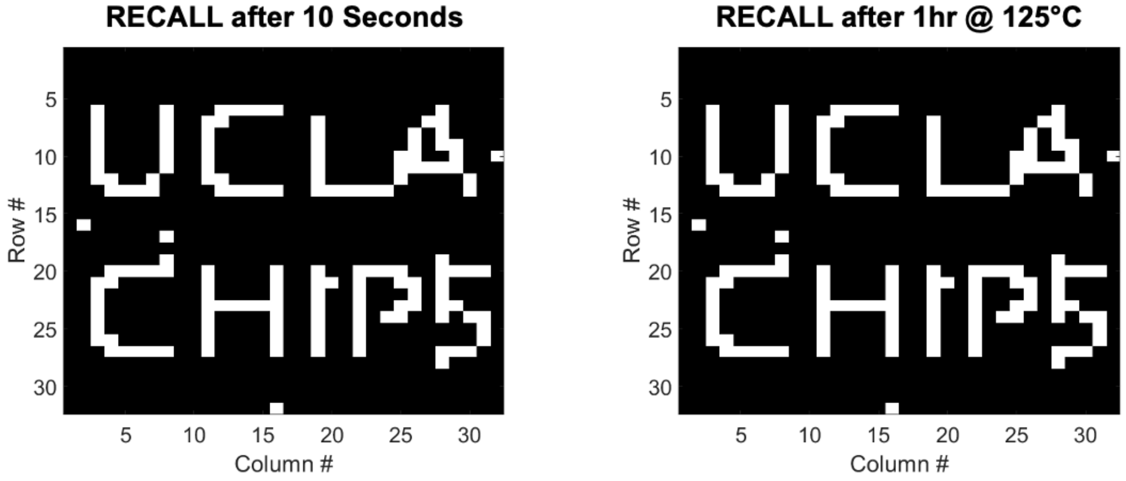


Figure 2.42: RECALL immediately after PRG (left), RECALL after 30 days of power-down (right).

Another important test, is to check the data retention in elevated temperatures. To study the impact of elevated temperatures on data retention, the chip was placed in an oven at a temperature of $125°C$ $(257°F)$ for 1 hour. After this step, RECALL is performed again and the results are shown in Figure 2.43. It is observed that the RECALLed patterns before and after heating up the chip are identical and none of the bitcells lost their non-volatile state.

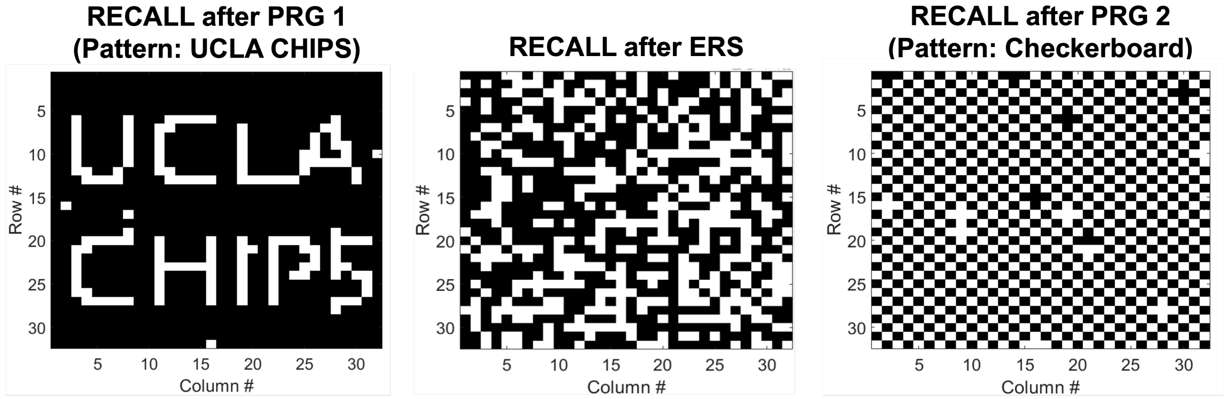

RECALL after 10 Seconds

RECALL after 1hr @ 125°C

Figure 2.43: RECALL immediately after PRG (left), RECALL after heating for 1-hour at 125°C (right).

2.8.4 Testing ERS Operation followed by re-PRG

To demonstrate the non-volatile reprogrammability of the chip with different patterns, first, UCLA CHIPS pattern is programmed. Then the memory is erased followed by programming a different pattern (checkerboard pattern). Results of this experiment are shown in Figure 2.44. As shown in this figure, a completely different pattern can be reprogrammed after performing ERS. Figure 2.44 (middle) shows the effectiveness of ERS operation, as no trace of UCLA CHIPS is left in this figure.



RECALL after PRG 1 (Pattern: UCLA CHIPS)

PRG Condition:
CTT_VDD= 1.7V, CTT_GATE= 2.7V
RECALL success-rate: 98.8%

RECALL after ERS

ERS Condition:
CTT_VDD=0V, CTT_GATE= -2.8V

RECALL after PRG 2 (Pattern: Checkerboard)

PRG Condition:
CTT_VDD= 1.7V, CTT_GATE= 2.7V
RECALL success-rate: 98.6%

Figure 2.44: RECALL after PRG with UCLA CHIPS pattern (left), RECALL after ERS (middle), RECALL after PRG with checkerboard pattern (right).

2.8.5 Studying the Impact of PRG on the 6T Core

To study the impact of PRG on the intrinsic mismatch of 6T-SRAM transistors at the core of eNVSRAM bitcell, the power-up pattern of an array that has never been programmed is read. The results are shown in Figure 2.45 (left). During a power-up operation, CTTs are kept off. Then, UCLA CHIPS pattern is programmed on the array and RECALL is performed, as shown on the Figure 2.45 (middle). After this step, the power-up operation (with CTTs off) is performed again and the results are shown in Figure 2.45 (right). This experiment shows that the power-up patterns (for before and after PRG) remain similar and UCLA CHIPS is not visible in the post-PRG power-up. Therefore, during PRG only CTTs are programmed and not the 6T-SRAM core transistors.

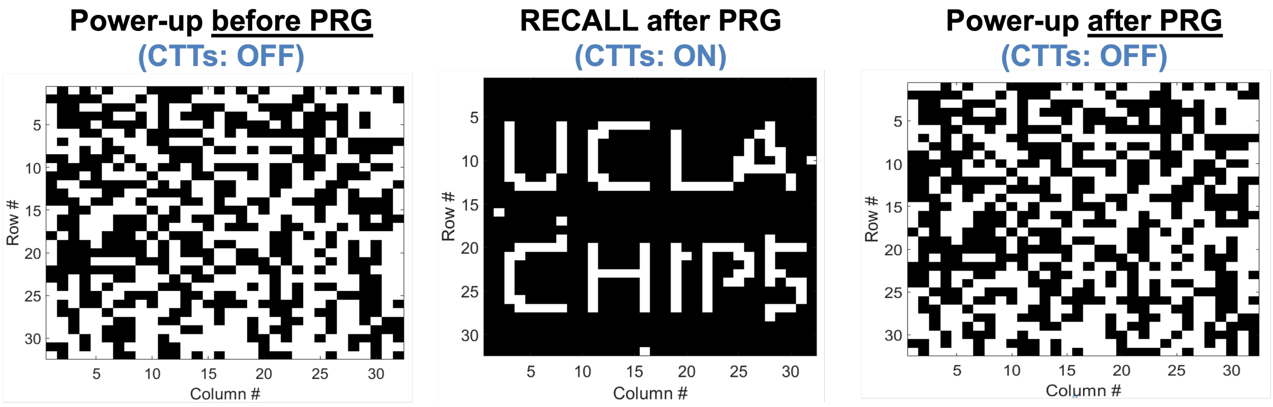
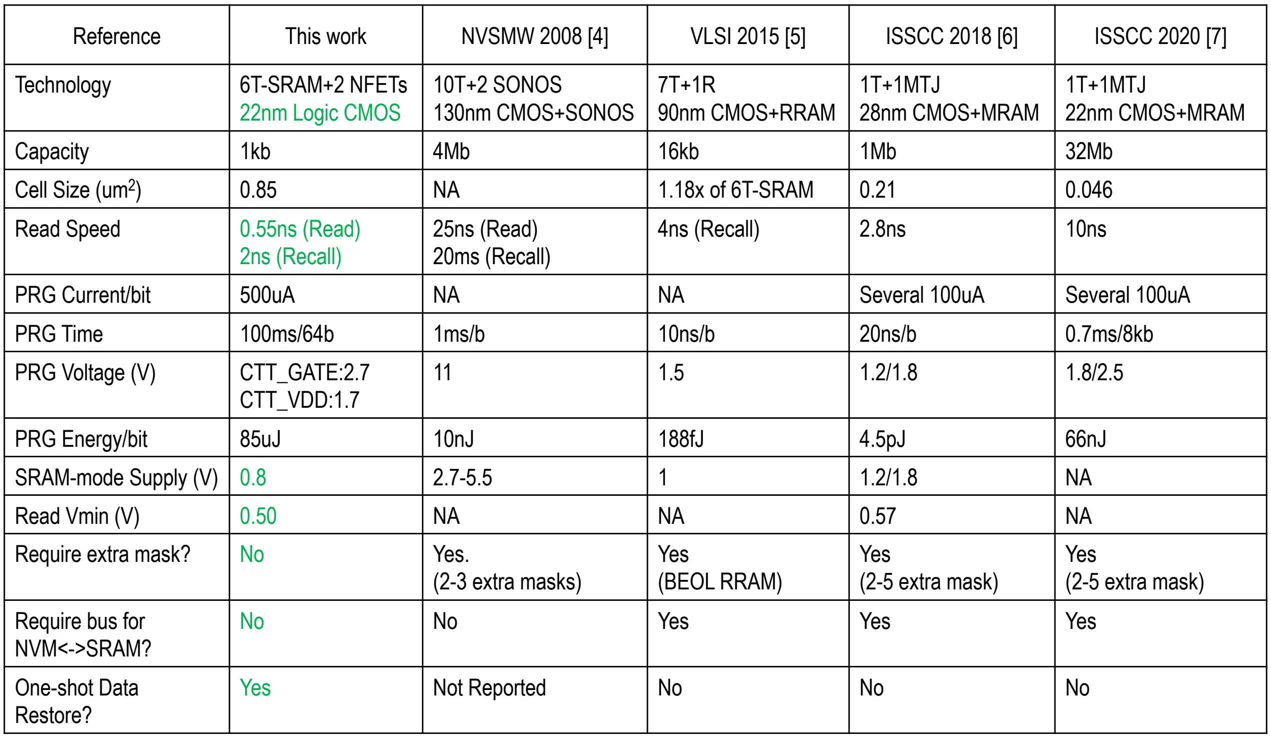


Figure 2.45: Power-up before PRG (left), RECALL after PRG (middle), power-up after PRG (right).

2.9 Table of Comparison

Table 2.4 compares this work with several state-of-the-art eNVMs. As we see, this is the only work that is implemented completely in a logic CMOS process and does not require any extra mask or fabrication step. This design can work as both regular SRAM and a non-volatile memory. Therefore, it benefits from a fast READ/WRITE in the SRAM mode. When used as an eNVM, the non-volatile RECALL is also fast, since it is performed simultaneously for the whole array. Another advantage of this design is the elimination of IO ports for transferring data between NVM and SRAM, since the nonvolatile devices are embedded inside the bitcells.



Reference	This work	NVSMW 2008 [4]	VLSI 2015 [5]	ISSCC 2018 [6]	ISSCC 2020 [7]
Technology	6T-SRAM+2 NFETs 22nm Logic CMOS	10T+2 SONOS 130nm CMOS+SONOS	7T+1R 90nm CMOS+RRAM	1T+1MTJ 28nm CMOS+MRAM	1T+1MTJ 22nm CMOS+MRAM
Capacity	1kb	4Mb	16kb	1Mb	32Mb
Cell Size (um^2)	0.85	NA	1.18x of 6T-SRAM	0.21	0.046
Read Speed	0.55ns (Read) 2ns (Recall)	25ns (Read) 20ms (Recall)	4ns (Recall)	2.8ns	10ns
PRG Current/bit	500uA	NA	NA	Several 100uA	Several 100uA
PRG Time	100ms/64b	1ms/b	10ns/b	20ns/b	0.7ms/8kb
PRG Voltage (V)	CTT_GATE:2.7 CTT_VDD:1.7	11	1.5	1.2/1.8	1.8/2.5
PRG Energy/bit	85uJ	10nJ	188fJ	4.5pJ	66nJ
SRAM-mode Supply (V)	0.8	2.7-5.5	1	1.2/1.8	NA
Read Vmin (V)	0.50	NA	NA	0.57	NA
Require extra mask?	No	Yes. (2-3 extra masks)	Yes (BEOL RRAM)	Yes (2-5 extra mask)	Yes (2-5 extra mask)
Require bus for NVM<->SRAM?	No	No	Yes	Yes	Yes
One-shot Data Restore?	Yes	Not Reported	No	No	No

Table 2.4: Table of comparison.

2.10 Conclusions

In conclusion, we reported a multi-time programmable eNVSRAM in a $22nm$ FDSOI standard CMOS process for low-cost IoT devices. The eNVSRAM can operate as a regular SRAM as well as a non-volatile memory. The non-volatile data storage is embedded inside the bitcell, which enables a bus-free simultaneous data transfer between SRAM and NVM. This approach eliminates the power, performance, and area overhead of interface ports, and enhances data security. Other potential applications of the reported design are Physically Unclonable Functions (SRAM-based PUF), secure key storage for authentication, non-volatile weight storage for ML/AL applications, and non-volatile in-memory computing (nvCIM) [14].

CHAPTER 3

Non-Volatile Wideband Frequency Tuning of a Ring-Oscillator by Charge Trapping in High-k Gate Dielectric in 22nm CMOS

3.1 Introduction

As CMOS technology scales down to nanometer regime, process variation increases, which negatively impacts the circuit performance [15]. For example, in a 22nm CMOS node, one-sigma V_{TH} variation is about 10% of the mean value for a minimum size transistor. Process variations may be compensated by employing different techniques such as increasing the device dimensions. However, increasing the device size increases the parasitic capacitance, reduces the switching speed, and increases the power consumption. This is particularly important in high-frequency circuits such as GHz oscillators or amplifiers. To compensate the effect of V_{TH} variation on the frequency of GHz oscillators and amplifiers, frequency tuning and calibration may be accomplished by means of switched-capacitor tuning [16], varactor tuning [17], or back-gate/body biasing [18, 19]. Unfortunately, none of these techniques are non-volatile and the calibration information is lost if the chip loses power. To address this issue, we report a technique that utilizes charge trapping to adjust V_{TH} and tune the frequency of an oscillator in a non-volatile fashion. To verify this technique, a frequency-tunable ring oscillator is designed, taped-out, and tested. This work serves as a proof-of-concept and the reported technique can be applied to other circuits such as amplifiers.

3.2 Schematics of a Frequency-Tunable Ring Oscillator Based on CTT

Schematics of the frequency-tunable ring oscillator circuit is shown in Figure 3.1, which includes an 11-stage ring oscillator, a 2-stage output buffer, two designated NMOS transistors as CTT devices inside the ring oscillator, and two CTT programing circuits. Each programming circuit includes a transmission gate (T-GATE) to control V_{GS} of the CTT and another T-GATE to control its V_{DS}. A 2-stage buffer is designed to provide sufficient drive to an external load. Width of the oscillator and buffers transistors are shown in Figure 3.1. These transistors have a length of 20nm. This design employs thin-oxide transistors for the ring oscillator and output buffers, while thick-oxide transistors are used for the transmission gates to handle a large external voltage during programming. The frequency of a ring oscillator depends on the delay of each stage (T_D) and number of stages (N), according to the following equation 3.1:

$$f = \frac{1}{2NT_D} \tag{3.1}$$

As V_{TH} of the ring oscillator transistor increas, T_D also increases. In this work, we show that by trapping charges in the high-k gate dielectric of the ring oscillator NMOS, we can reduce the oscillation frequency in a controllable and reliable manner.

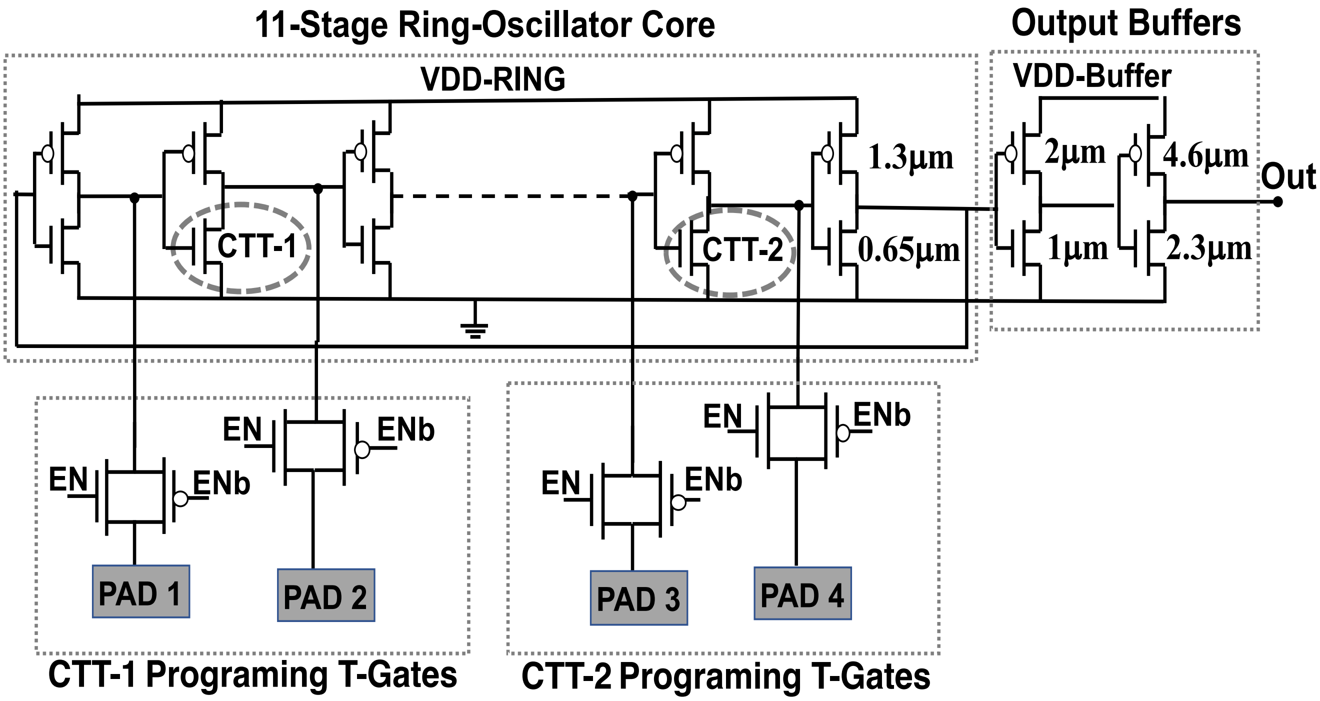


Figure 3.1: Schematics of an 11-stage CTT programmable ring oscillator.

3.3 Measurement Results

In this section, the measurement results of a 12-stage frequency-tunable ring oscillator that is designed and taped-out in 22 nm CMOS process are reported. The size this macro $57\mu m \times 36\mu m$. To reduce the area overhead of pads in this design, we have utilized an advanced packaging technique known as Silicon Interconnect Fabric (Si IF), which has been developed in our lab. Si IF technology uses Cu-Cu thermal compression bonding at $10\mu m$ pitch, as compared to the conventional C4 bump/pillar technology which uses a $100\mu m$ pitch. More information on Si IF technology can be found in [20]. Without using Si IF technology, the pad area would have been $30 \times 100 \times 100\mu m^2$ or 150 times the area of the macro.

The ring oscillator performance is characterized and the non-volatile frequency tuning is demonstrated. In the experimental setup, output of the final stage buffer is connected to a spectrum analyzer (Keysight PXA N9030A) to measure frequency of the chip. A probe card is used to probe the pads on the SiIF platform. Figure 3.2 shows a picture of the probes landed on the SiIF platform.

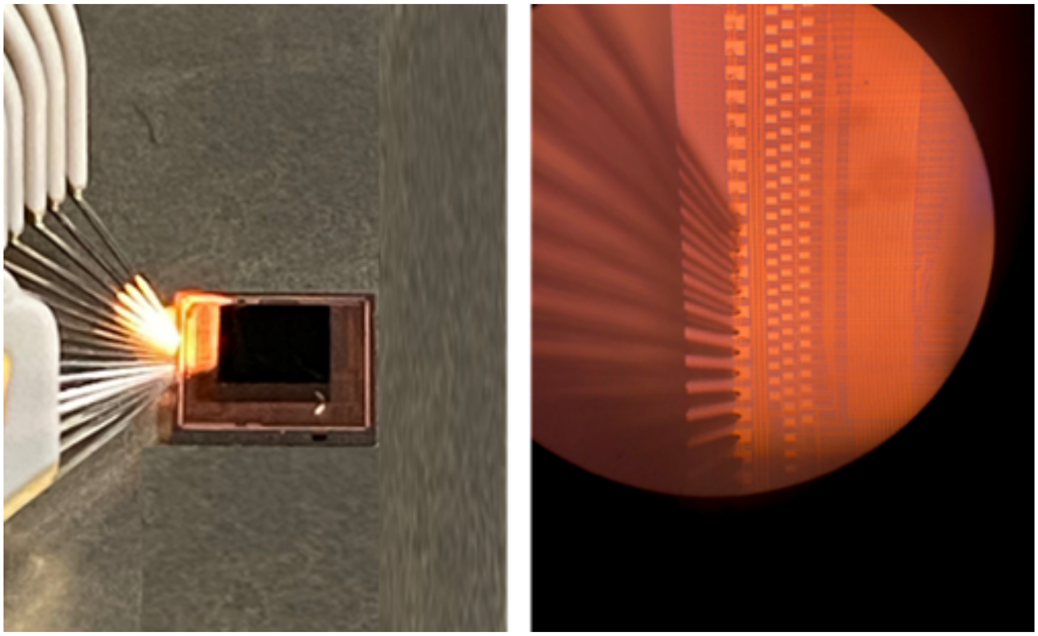


Figure 3.2: A photo of the die on the Si IF platform and the measurement probes.

Figure 3.3 shows frequency of the ring oscillator as a function of supply voltage in both simulation and measurement that match closely with each other. Based on the measurement results, the 11-stage ring oscillator starts to oscillate at VDD=0.29V. At this voltage, the oscillation frequency is 13.66MHz and the oscillator consumes 0.12μW. By increasing the supply voltage, the frequency increases to 2.165GHz at VDD=0.8V and the power consumption rises to 169.2μW.

The frequency of the ring oscillator is shown as a function of the supply voltage in Figure 3.3. The simulation and measurement results closely match each other. The 11-stage ring oscillator starts oscillation at VDD=0.29V as determined from the measurement results. At this voltage, the oscillator's frequency is 13.66$MHz$and power consumption is 0.12μW. Upon raising the supply voltage, the frequency increases, reaching 2.165GHz at VDD=0.8VV and the power consumption rises to 169.2μW.

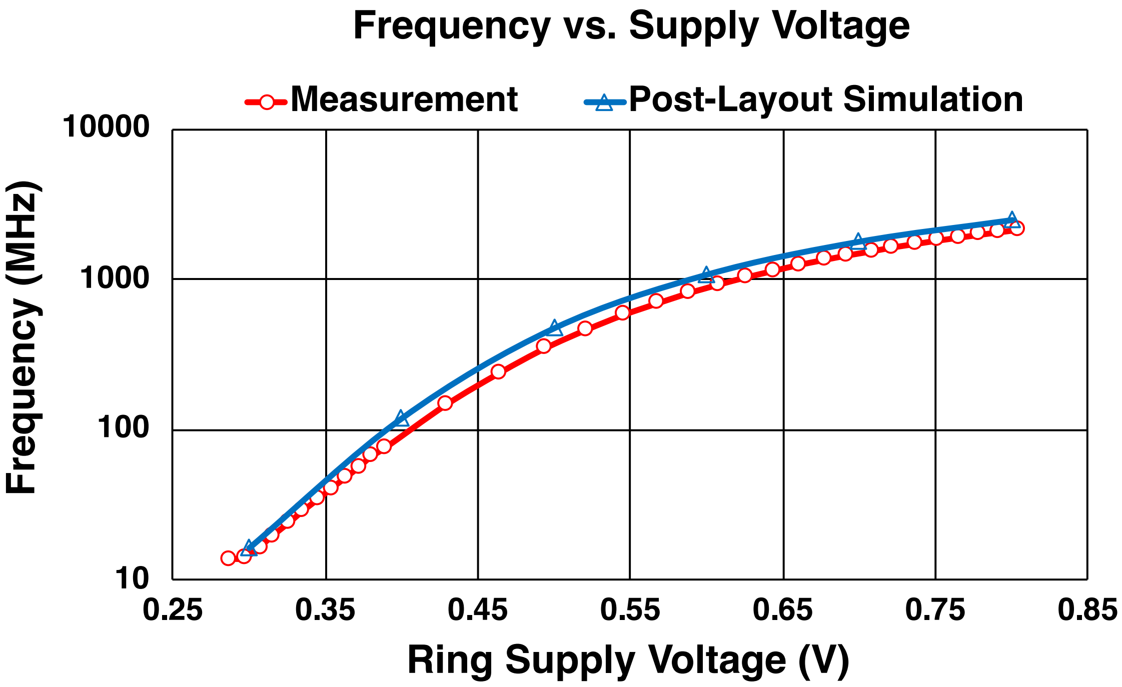


Figure 3.3: 11-stage ring oscillator frequency vs. supply voltage.

In the following section, we will discuss how CTTs can be used to change the frequency of the oscillator in a nonvolatile fashion. In order to demonstrate the impact of trapping electrons in the high-k gate dielectric of the CTT, we have applied voltage pulses at PAD1. As shown in Figure 3.1, PAD1 is connected to the gate of CTT-1 through a T-GATE. In each programming attempt, we have applied 300 pulses over a period of 30sec, with a pulse rate of 10Hz at PAD1. The duration of each pulse is about 100μsec. Figure 3.4 shows the impact of programming voltage levels (i.e. V_{GS} and V_{DS} of the CTT) on the frequency shift. In this figure, X-axis shows the voltage levels at PAD1, which varies from 1.3V to 3.9V. Each curve in Figure 3.4 corresponds to a different DC voltage level applied at PAD2. These values are 1V, 1.3V, 1.6V, 2.2V, and 2.5V. It is important to note that due to the voltage drop on the T-GATEs, the actual voltages reached to the CTT gate and drain are smaller than the voltages applied at PAD1 and PAD2, respectively.

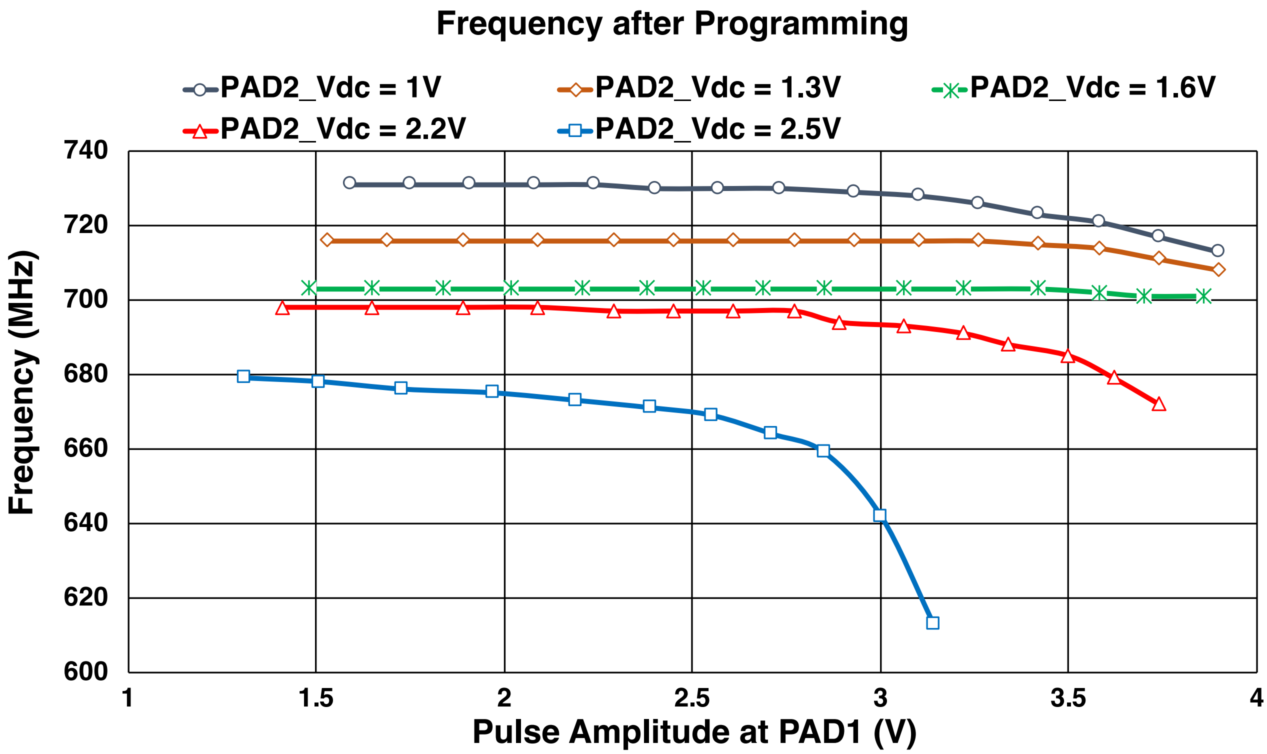


Figure 3.4: Effect of charge trapping on the oscillation frequency.

After every programming attempt, the programming circuitry is disabled, VDD-RING is set to $570mV$, and frequency of the ring oscillator is remeasured. We observed that by increasing the amplitude of the pulses applied at PAD1, more electrons are trapped in the high-k gate dielectric and threshold voltage of the NMOS is increased, as expected. Also, by increasing the voltage applied at PAD2 and passing higher I_{DS}, the self-heating is increased, which assists in the charge trapping process. The increased self-heating creates a larger V_{TH} shift and further reduces the oscillation frequency. In another programming method, we disabled the programming circuity and tried to program the ring oscillator by raising its supply voltage (VDD-RING). In this experiment, a baseline is produced by measuring the ring oscillator frequency as a function of VDD-RING (Figure 3.5). Then a large voltage of $2.3V$ is applied to VDD-RING for about $3sec$ to program the ring oscillator transistors. After programming with this method, frequency of the ring oscillator is measured as a function of VDD-RING again. It is observed that the stress induced by a high voltage of VDD-RING=2.3V caused electrons to be trapped in the gate dielectric, which increased V_{TH} of transistors and reduced frequency of the ring oscillator.

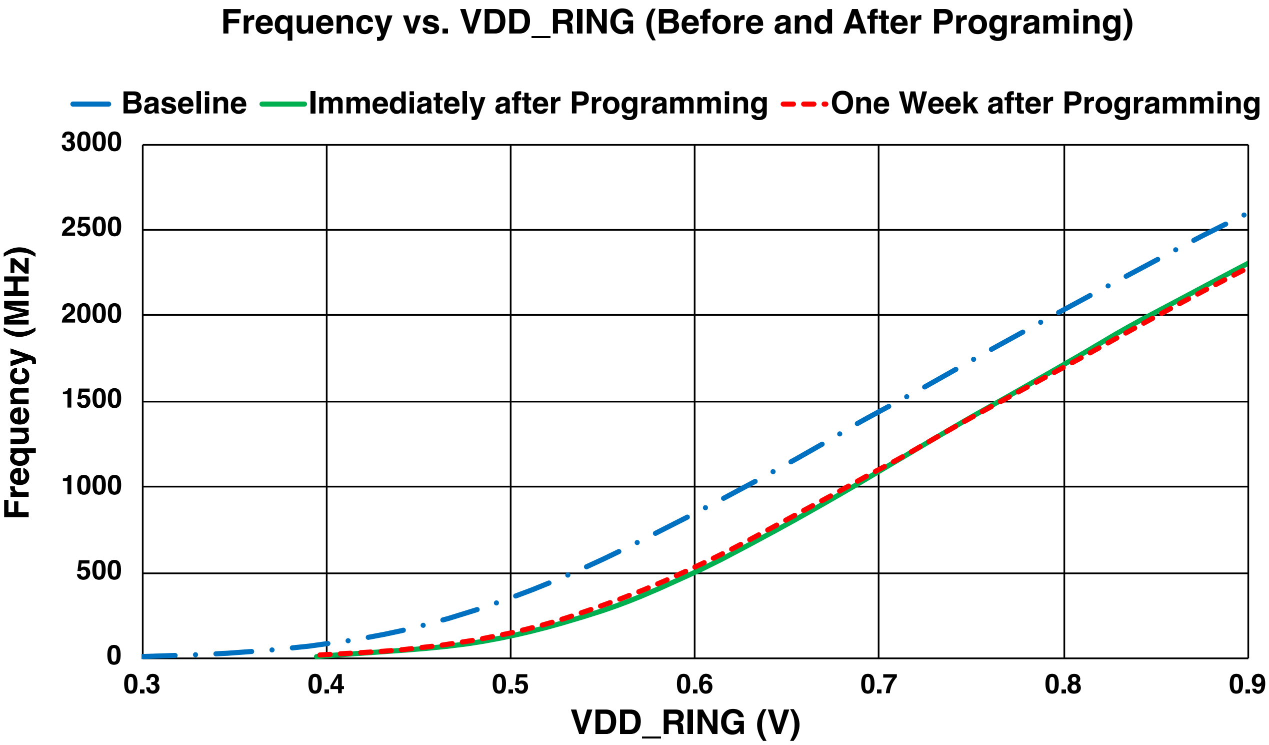


Figure 3.5: Oscillation frequency vs. VDD before and after programming.

As shown in Figure 3.5, at VDD-RING=0.8V, we were able to change the frequency from 2075GHz to 1746GHz (frequency-shift=329MHz). Based on Monte Carlo analysis, one-sigma frequency-shift due to the process variation and mismatch is 119MHz at VDD-RING=0.8V. Therefore, our programming technique allows us to cover a tuning range of 2.8σ at VDD-RING=0.8V. In this experiment, the oscillation frequency decreased by 16% at VDD-RING=0.8V. Based on the post-layout simulation in Cadence Virtuoso, to reduce the oscillation frequency by 16% at VDD-RING=0.8V, the V_{TH} of NMOS transistors should increase by 110mV. According to the transient simulation, at VDD-RING=2.3V, about 1% of the time in every cycle, both V_{DS} and V_{GS} are higher than 1.5V. This level of V_{DS} and V_{GS}

is necessary to create deep traps. This means the actual programming happens during only 1% of the time that the elevated VDD-RING is applied. This is in agreement with a previous work [21], where a $\Delta V_{TH_{max}}$ of $100mV$ was achieved with V_{DS}=$1.5V$ and V_{GS}=$1.5V$ applied for at least $10msec$. Finally, we have compared different V_{TH} tuning methods in Table 3.1

Table 3.1: Comparison of Different V_{TH}Tuning Methods.

Method	**Freq. Range (MHz)**	**Non-Volatile**	**Node**	**Hardware**	**Ref.**
CTT PRG	1746-2075	Yes	22nm FDSOI	Oscillator Tuning	This work
CTT PRG	N/A	Yes	22nm FDSOI	Analog Memory Cell	[22]
CTT PRG	N/A	Yes	22nm FDSOI	Digital Memory Cell	[11]
CTT PRG	N/A	Yes	22nm FDSOI	Digital Memory Cell	[23]
Back-gate/Body-Biasing	377-556	No	65nm CMOS	Oscillator Tuning	[18]
Back-gate/Body-Biasing	25300-26300	No	22nm FDSOI	Oscillator Tuning	[19]

3.4 Conclusions

In this chapter, we reported a non-volatile technique for tuning the frequency of a GHz ring oscillator using CTTs. This work servers as a proof-of-concept for non-volatile tuning of circuits using charge trapping in high-k gate dielectric of advanced CMOS nodes. We acknowledge GlobalFoundries for fabrication of the macro and UCLA CHIPS students especially Siva Jangam and Krutikesh Sahoo for SiIF fabrication.

CHAPTER 4

Other Works

4.1 A Supply Fluctuation Resilient SRAM

4.1.1 Background and Motivation

In the pervious chapter, we introduced a CTT-based non-volatile SRAM that can keep data when SRAM loses power for an extended duration (>10years). In this chapter, we are presenting an earlier work that is not CTT-based. Using this technique SRAM data can be retained after a short duration (few seconds) of power-loss. This is particularly important for the systems that rely on energy harvesting and may experience brief supply fluctuations, which can result in data loss in SRAM. [24, 25, 26, 27]. This work is designed and simulated in TSMC's 180nm CMOS process, which is described the next section.

4.1.2 Conventional SRAM vs. Proposed SRAM

Figure 4.1 shows the block diagram a conventional 6T-SRAM [28]. In a conventional SRAM, when the supply voltage drops below V_{min}, the value of DATA and DATA_B are lost. The time that it takes for DATA to decline to zero depends on the total capacitance at node DATA and the effective resistance to GND. Figure 4.2 shows that when VDD transitions to zero, DATA follows the decline at almost same time. This is because transistor M3 is mostly ON during this transition. In this case, DATA and DATA_B become zero in about 10$msec$.

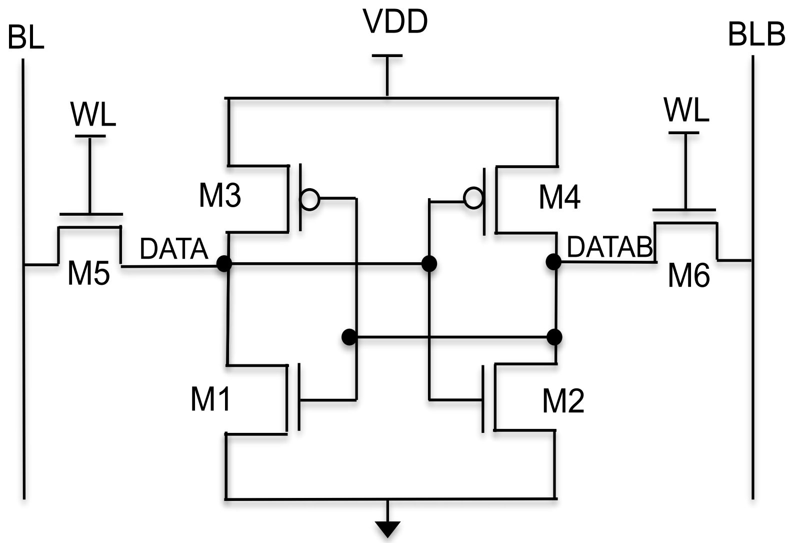


Figure 4.1: Schematic of a 6T-SRAM cell.

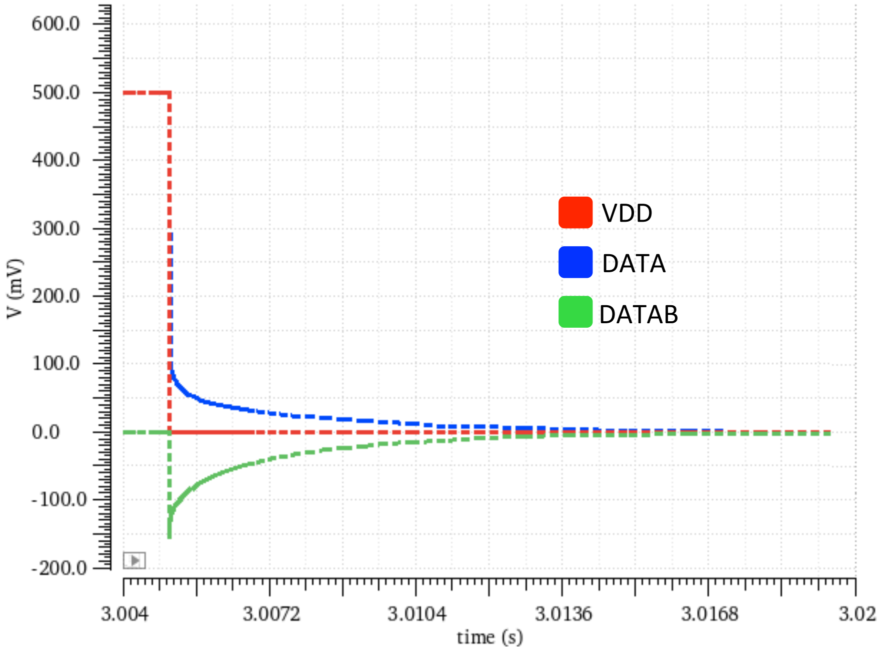


Figure 4.2: DATA discharges in less than 10$msec$ when VDD is lost.

Figure 4.3 shows the schematic of the proposed SRAM cell. In this design, one MIM capacitor (C1) and four transistors are added to a 6T-SRAM. Transistors M9 and M10 are used as a voltage divider to produce a voltage of VDD/2 at the gates of M7 and M8. During the normal operation, since VDD is high, M7 and M8 behave like large resistors and C1 is disconnected from the SRAM core. Therefore, C1 does not consume power during the normal operation. When VDD drops, the effective resistances of M7 and M8 are reduced. Therefore, C1 is charged briefly through M7 and M8. Polarity of the charge stored on C1 helps the SRAM to recover its original data when VDD becomes high again. V_{GS} of transistors M9 and M10 are intentionally kept at 0V to minimize the leakage current drawn from VDD.

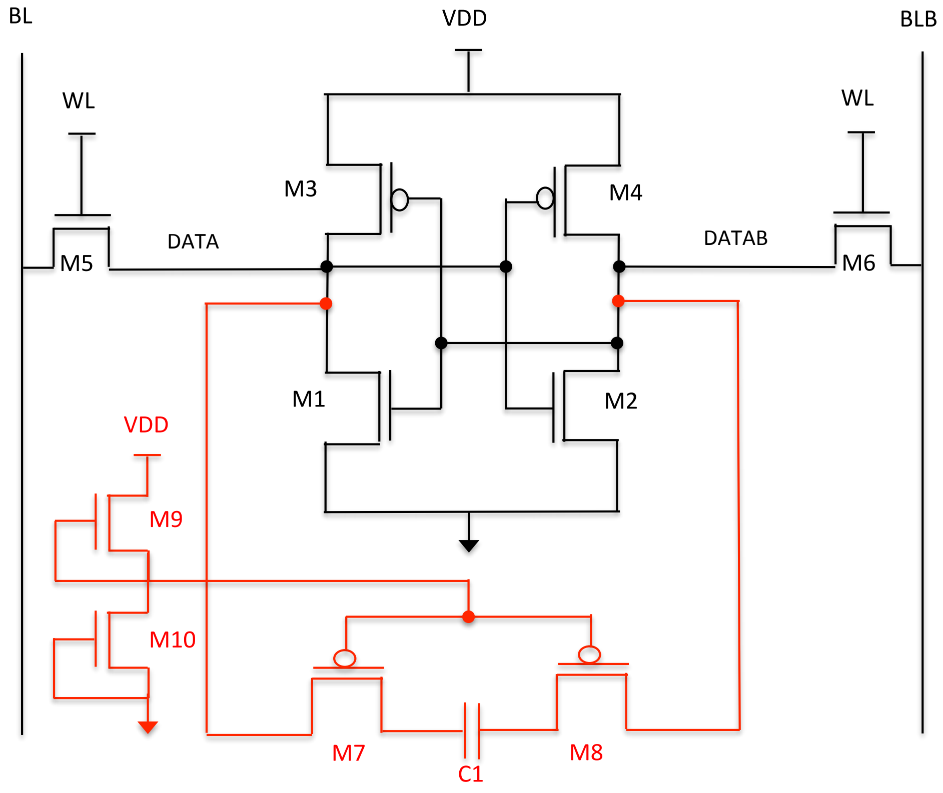


Figure 4.3: Architecture of the proposed SRAM.

To describe the operation, we first assume DATA is '1', DATA_B is '0', and there is no charge on C1 before VDD turns off. In this case, when VDD is transitioning to 0, M7 and M8 turn on briefly and since DATA is '1', left plate of C1 is charged positively. After VDD reaches 0, since M3 and M1 are OFF, they create a high resistance discharge path for C1. This helps to keep the charge on C1 for about 1*sec*. The cap polarity is always in a direction that the original data can be recovered once VDD is back ON.

As shown in Figure 4.4, DATA becomes '1' again after VDD turns back on. Figure 4.5 shows the charge on C1 as a function of time. In this figure, it takes about 1*sec* for C1 to lose its charge (C1=288fF). This simulation is performed for a sub-threshold regime with VDD=500mV. In this process, $V_{TH-NMOS}$=730mV and $V_{TH-PMOS}$=-640mV.

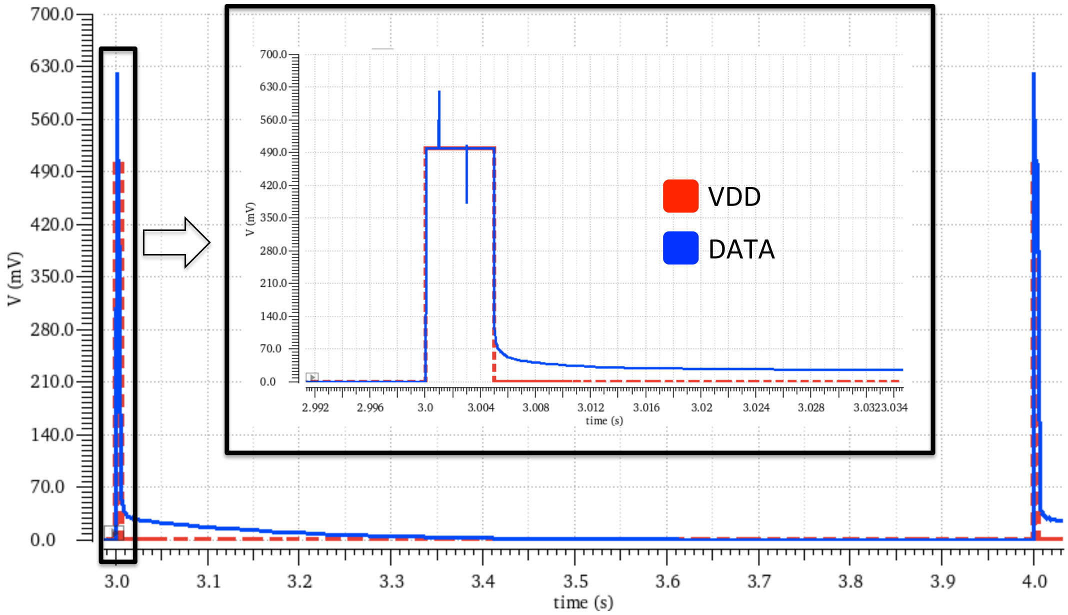


Figure 4.4: DATA and VDD vs. time. DATA regains its correct value once VDD turns back ON after being OFF for 1*sec*.

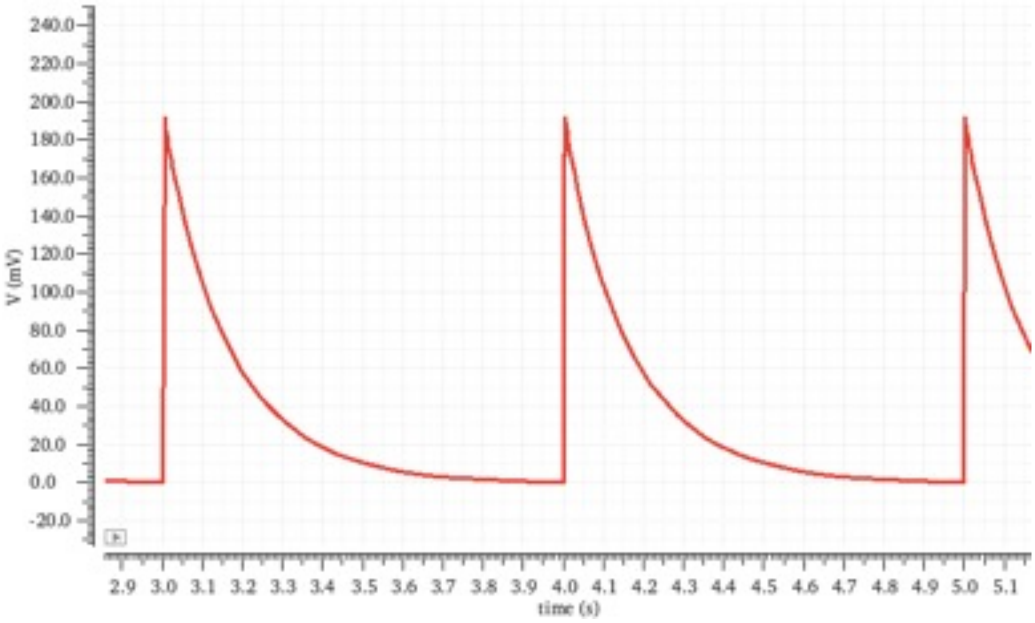


Figure 4.5: C1 voltage vs. time.

As shown in Figure 4.5, the initial voltage on C1 due to the charging process is only $190mV$. Although, this value decays after $1sec$, due to the large size of the capacitor (C1=$288fF$) compared to the parasitic capacitances of the SRAM transistors, the remaining charge on C1 creates sufficient imbalance to recover the correct data. To verify the proper operation of the proposed SRAM, yield analysis and PVT corner simulations have been performed. Next section focuses on the yield analysis.

4.1.3 Yield Analysis

In this section, yield has been analyzed as a function of several parameters. These parameters are VDD, size of C1, and the power-down duration (T_{OFF}). For example, by increasing the size of C1, yield of SRAM improves because it takes a longer time for C1 to lose its charge. However, a large capacitor is undesirable, due to the area overhead. Increasing VDD also improves the yield because the initial charge on C1 becomes larger and the discharge time becomes longer. However, a large VDD increases the power consumption of the SRAM in the normal operating mode. Hence, there is another trade-off between the power consumption and the maximum time that VDD can remain off. The last parameter that we study for the yield analysis is duration of the time that VDD remains off. By increasing this time, yield decreases due to C1 charge leakage .

Figure 4.6 shows the yield of a single bitcell as we increase the capacitance of C1 from $200fF$ to $570fF$. In this simulation, the OFF duration is kept at $1sec$ and VDD is $0.5V$. As shown in this plot, to achieve a yield of 99.7%, size of C1 should be greater than $288fF$. By increasing the capacitance value to $337fF$, the yield increases to 99.9%.

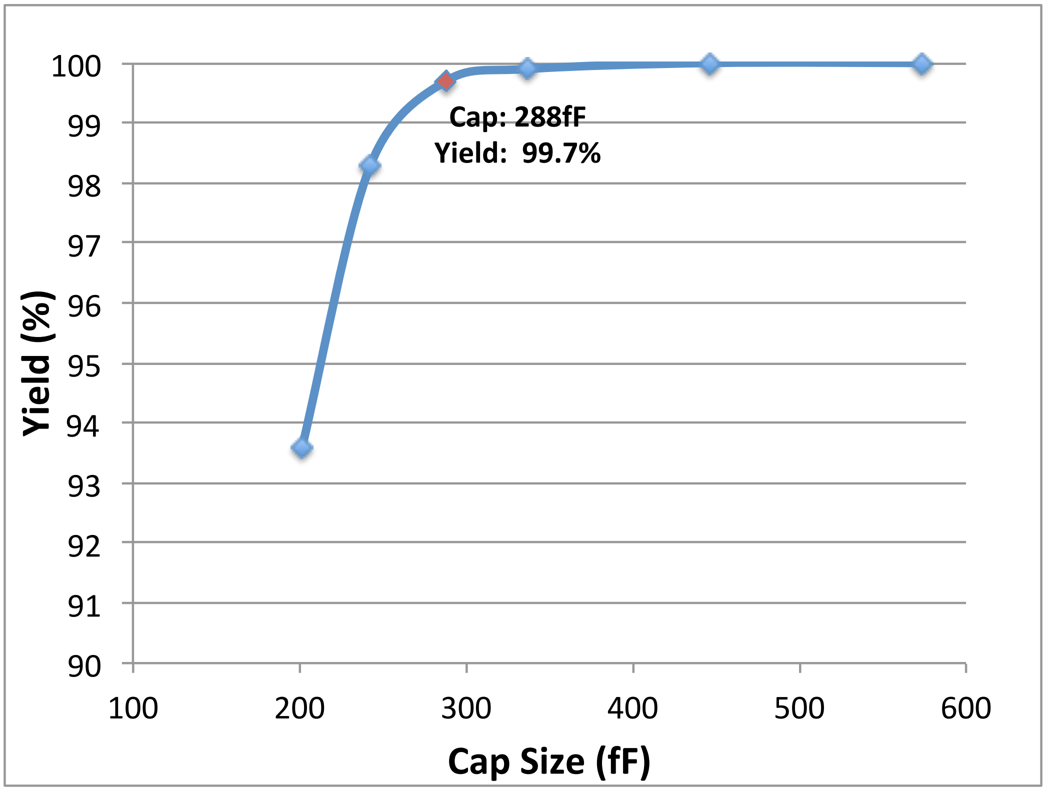


Figure 4.6: Yield as a function of cap size. VDD=$0.5V$, T_{OFF}=$1sec$.

Figure 4.7 shows the yield as we increase VDD from $0.35V$ to $0.5V$. In this simulation, the VDD OFF duration is kept at $1sec$ and C1=$288fF$. Based on this plot, to achieve a yields of 98.9% and 99.7%, VDD should be greater than $0.4V$ and $0.5V$, respectively.

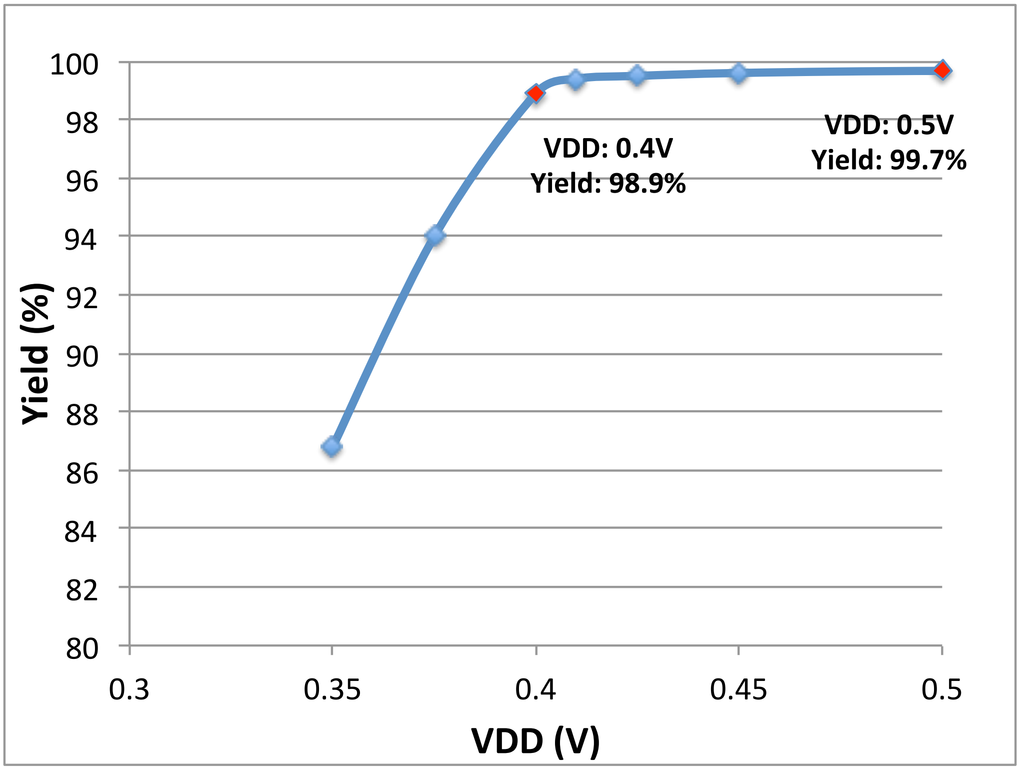


Figure 4.7: Yield as a function of VDD. C1=288fF, T_{OFF}=1sec.

Figure 4.8 shows yield as we increase the VDD OFF duration from 0.9sec to 1.4sec. In this simulation, VDD is kept at 0.5V and C1=288fF. As shown in this plot, to achieve a yield of 94.8% and 99.7%, the VDD OFF duration should be smaller than 1.3sec and 1sec, respectively.

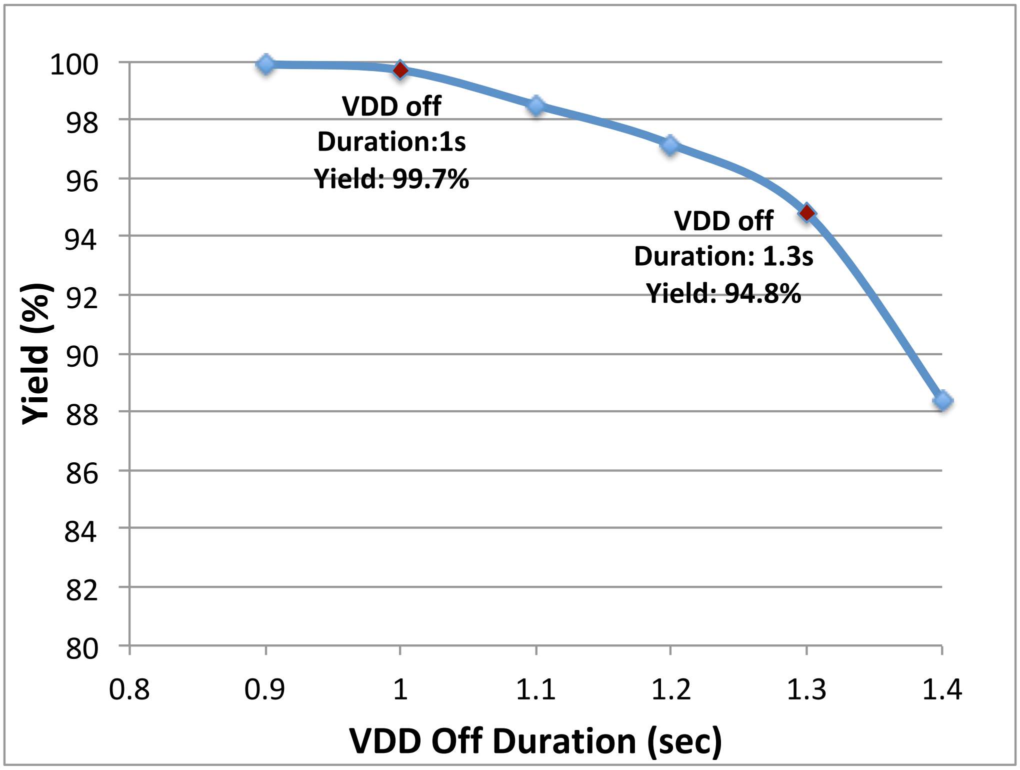


Figure 4.8: Yield as a function of VDD OFF duration. C1=288fF and VDD=0.5V.

4.2 A Low-Power Digital ASIC for Detecting Heart-rate and Missing Beats

4.2.1 Background and Motivation

To reduce the power consumption of an SOC, the algorithmic techniques that are implemented as a part of digital design, play an important role. These algorithms should not consume high power when they are implemented in the hardware. This is particularly important for medical implants, since they are either battery operated or rely on energy harvesting. Medical implants can be used to record body's electrical activities (e.g. neural activities or cardiac signals) or perform electrical stimulation in therapeutic applications [29, 30]. For example, they can be used to control seizures by monitoring neural activities and inducing brain stimulations to prevent the occurrence of the seizure. In heart applications, an irregular heart-rate or a missing-beat can be an indicator of an underlying disease, which can be detected with several methods such as sensing the heart's electrical activities [31, 32, 33].

Traditionally a pacemaker, which is a device about the size of a pocket watch, is used to monitor electrical activity of the heart and perform stimulation to bring the heart rhythm to the normal condition [32, 33]. The pacemaker is placed under the skin (in the chest or abdomen area) and uses leads to connect to the heart tissue.

During the last 50 years, the basic concept of pacemakers remained unchanged. In traditional pacemakers, an implantable pulse generator sends electrical signals to one or more leads. These leads must pass through the veins to connect to the heart tissue. This approach has caused significant complications including infection and lead failure. To mitigate many issues caused by the leads, lead-less pacemakers have been introduced [31, 27]. A lead-less pacemaker, can perform sensing and stimulation and eliminates the complications caused by leads. Unfortunately, the lead-less pacemakers (such as Micra by Medtronics [31]) suffer from two major issues: (a) They can only pace a single point, (b) they need a battery to operate and if the battery depletes, they cannot be removed from the heart [32, 33]. Multi-site pacing is particularly important in Cardiac Resynchronization Therapy (CRT) [32]. In CRT, a single-point pacing is not sufficient. To address the challenges caused by battery-powered lead-less devices, researchers have focused on implementing battery-less lead-less pacemakers that can be powered with various mechanisms such as the heart's mechanical movement [27], ultrasonic energy harvesting, or electromagnetic energy harvesting [26].

The wirelessly powered pacemakers impose a more stringent requirements on the power consumption. This is because the amount of harvested energy is very limited and the efficiency of the wireless energy transfer is low. In [25], A. Amar, et. al., provided a summary on the power consumption of various Implantable Medical Devices (IMDs) and the available technologies for providing power. Based on this summary, electromagnetic energy harvesting can provide power up to $100\mu W$. The goal of this work is to design an ultra-low power custom digital ASIC that consumes less than $1\mu W$ and occupies an area of less than $1mm$ by $1mm$. This level of power consumption and size makes it possible to integrate this ASIC on a single chip that can eventually host the complete system that includes: on-chip antennas for harvesting electromagnetic energy, analog amplifiers to amplify the electrical activities of the heart, an analog-to-digital converter (ADC), a digital ASIC to perform on-chip data processing, and pulsing circuitry to stimulate the heart. Figure 4.9 shows block diagram of the wirelessly-powered chip that is used in an implantable pacemakers. In the following sections we will discuss the design of an ASIC chip for low-power processing of cardiac signals. This design has been taped-out in AMI's $0.5\mu m$ process technology.

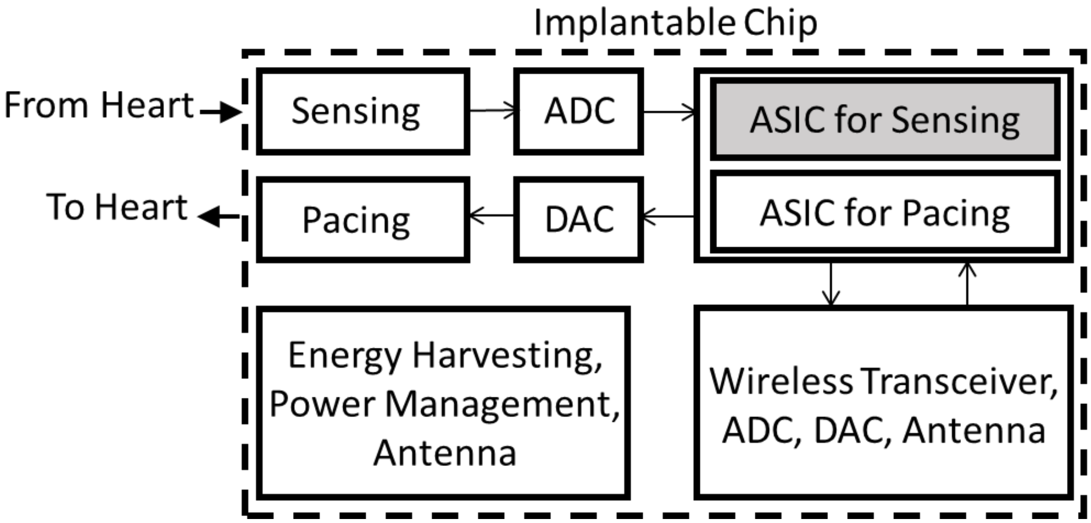


Figure 4.9: Block Diagram of a wirelessly-powered chip used in an implantable pacemaker.

4.2.2 System Description and ASIC Details

The implantable chips can be placed in different parts of the heart tissue to perform multi-site pacing. These chips can be delivered to heart using standard catheters through the vein with a minimally invasive procedure [31]. The chips receive electromagnetic energy from an external battery-powered control unit. The control unit, which stays outside of the body, provides power/commands to the chip and receives sensed signals from the chip wirelessly. Since this chip relies on the harvested energy, to minimize the energy used for sending data from the chip to the control unit, raw data goes through an initial processing on the implanted chip and only the important features are transmitted to the external unit. This work presents a power-efficient algorithm for detection of heart-rate and missing beats and its implementation on an ASIC chip.

Figure 4.10 shows the algorithm flowchart for detecting the heart-rate and missing beats and Figure 4.11 shows the schematic diagram of the heart-rate detection unit. To detect a heart-rate, first we detect a valid peak (i.e. a peak that is higher than a certain level) for every heat-beat cycle that is called QRS waveform. In this design, the sensed cardiac signals are fed to the circuit sequentially. Peak detection is performed as the following: First, the voltage level of a sensed signal is compared with a given peak threshold (P_T). If the sensed signal exceeds P_T, the value of this signal is stored in a 8-bit register as the temporary maximum value. Then, the new sensed data enters and if this data exceeds the value stored in this register, the register is updated with the new data. At the end of every heart-beat cycle this register stores the value of the largest sensed data during one cycle. To detect the real maximum and not the local maxima, a ten input AND gate with nine delay components is used. A peak detection signal is activated only if the value of the register is higher than the next 9 incoming data points. To calculate the heart-rate, the peak-to-peak time intervals are accumulated after successful detection of a valid peak. In addition to the time accumulation, a counter is used to count the number of detected peaks. When the accumulated time reaches 60 seconds, the value of the peak counter is reported as the heart-rate in terms of number of Beats per Minute (BPM). This is a novel ultra-low-power method for detecting a heart-rate that avoids using FFT and computationally intensive operations to save power.

To find a missing beat, the average peak-to-peak duration needs to be calculated from the sensed data. For this purpose, another time accumulator is used. This accumulator stores the total time for 17 beats and divides this value by 16 (2^4). We chose number 16, which is a power of 2, so the division can be implemented by four shift-right operations instead of a conventional division. A shift operation consumes less power than a conventional division when it is implemented in hardware. This division provides an approximate peak-to-peak average, without sacrificing the accuracy of the missing-beat detection. In this design whenever possible, we took advantage of the approximate computation without losing the accuracy to minimize the power consumption. A missing beat flag is raised if a future peak-to-peak interval exceeds a safe margin, where the safe margin is $\beta\times$ (average peak to peak interval). This multiplication is also performed with shift operations. β can be adjusted for different patients and can be chosen from the range of numbers between 1 to 2. β is set to be 1.75 in this work to avoid a false missing beat alarm if the heart-rate reduces naturally. The accuracy of the proposed heart-rate and missing beat detection algorithms have been tested and verified on real patients data.

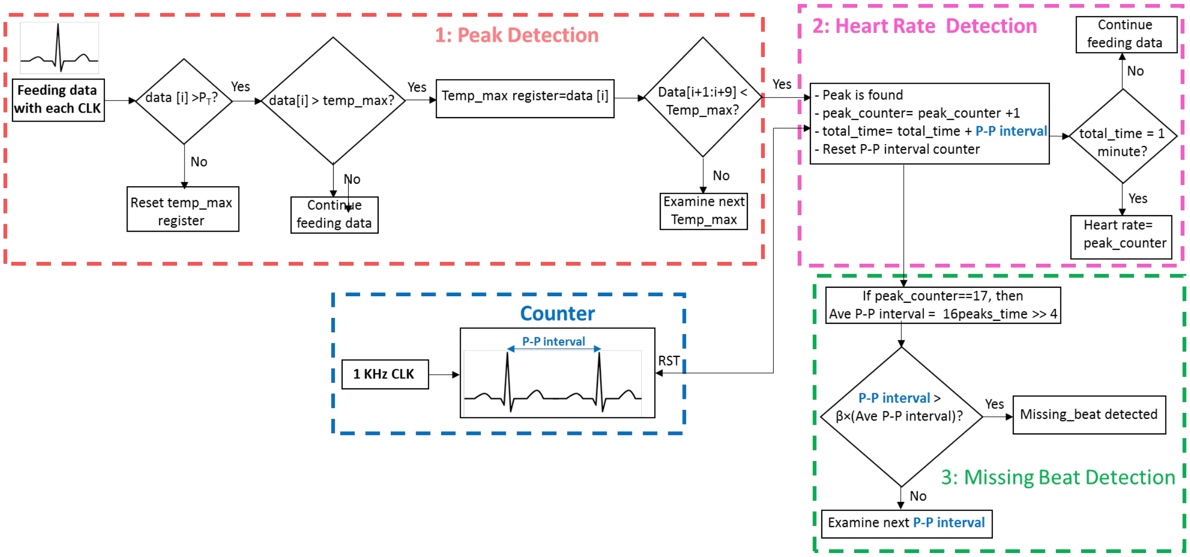

1: Peak Detection
Feeding data with each CLK
data [i] >P_T?
Yes
No
Reset temp_max register
data[i] > temp_max?
Yes
No
Continue feeding data
Temp_max register=data [i]
Data[i+1:i+9] < Temp_max?
Yes
No
Examine next Temp_max
2: Heart Rate Detection
- Peak is found
- peak_counter= peak_counter +1
- total_time= total_time + P-P interval
- Reset P-P interval counter
total_time = 1 minute?
No
Continue feeding data
Yes
Heart rate= peak_counter
Counter
1 KHz CLK
P-P interval
RST
If peak_counter==17, then
Ave P-P interval = 16peaks_time >> 4
P-P interval > β×(Ave P-P interval)?
Yes
Missing_beat detected
No
Examine next P-P interval
3: Missing Beat Detection

Figure 4.10: A high-level algorithm flowchart for detecting the heart-rate (in BPM) and missing beats.

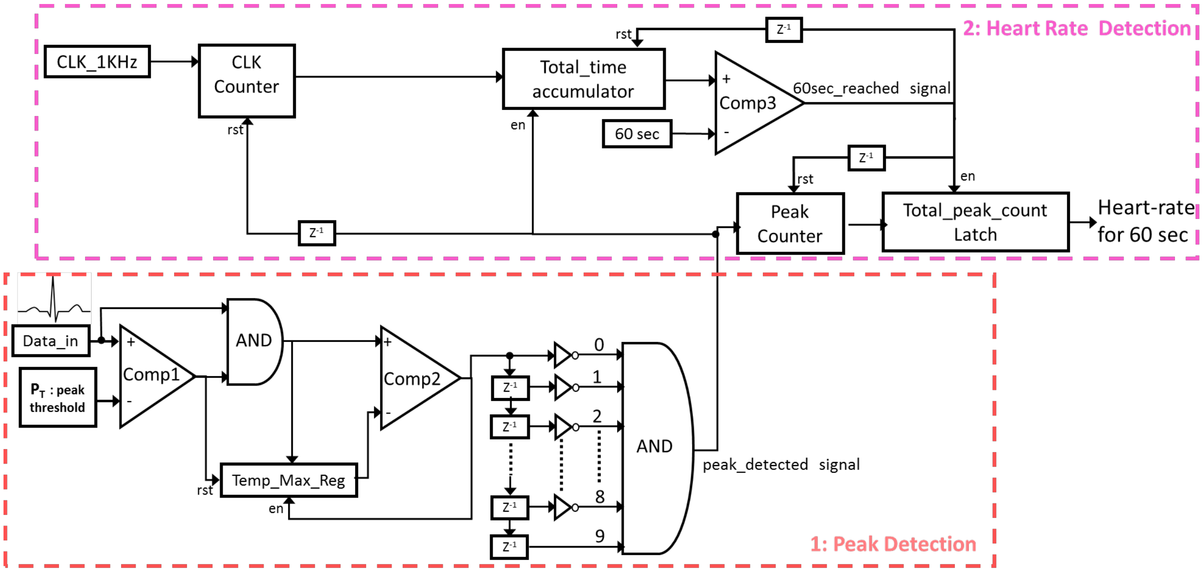


Figure 4.11: Schematic for heart-rate detection unit (in BPM).

4.2.3 Simulation Results and Hardware Implementation

In this work, a set of data provided by Texas Heart Institute (THI) is used to verify the algorithm. This data is generated by a BARD system [34] that amplifies the electrical activities of the heart and performs band-pass filtering with cutoff frequencies of $30Hz$ and $250Hz$. BARD system uses a sampling rate of $1KHz$ and it can sense up to $5mV$. The number of bits of the data recorded by BARD system is reduced to 8-bit signed to minimize the power consumption without losing significant information. The 8-bit data are fed to the ASIC chip that is reported in this section. Figure 4.12 shows one beat cycle for one of the 17 channels recorded by the BARD system.

The algorithm is tested and verified in MATLAB before it is implemented with RTL. Figure 4.13 shows the number of beats detected as a function of the threshold voltage simulated in MATLAB. Based on this simulation, The middle point of the flat region is used to find a threshold value (P_T) that later on is used for finding a valid peak. The reported algorithm for the ASIC chip is implemented with Verilog. Figure 4.14 shows the Modelsim logic level simulation results. This figure shows that the detected heart-rate is 99 BPM, which is verified by examining the patient cardiac data. Figure 4.15 shows that missing beats can be detected with. this algorithm. After inspecting the data, it is confirmed that there is one missing beat in the data because one of the peaks is smaller than P_T.

The ASIC chip is taped-out using AMI's $0.5\mu m$ process technology. To minimize the number of pads, only one pad is used to get data into the chip and another pad is used for sending data out of the chip sequentially. The entire design occupies a total area of $785\mu m$ by $744\mu m$. The layout of the ASIC chip is shown in Figure 4.16.

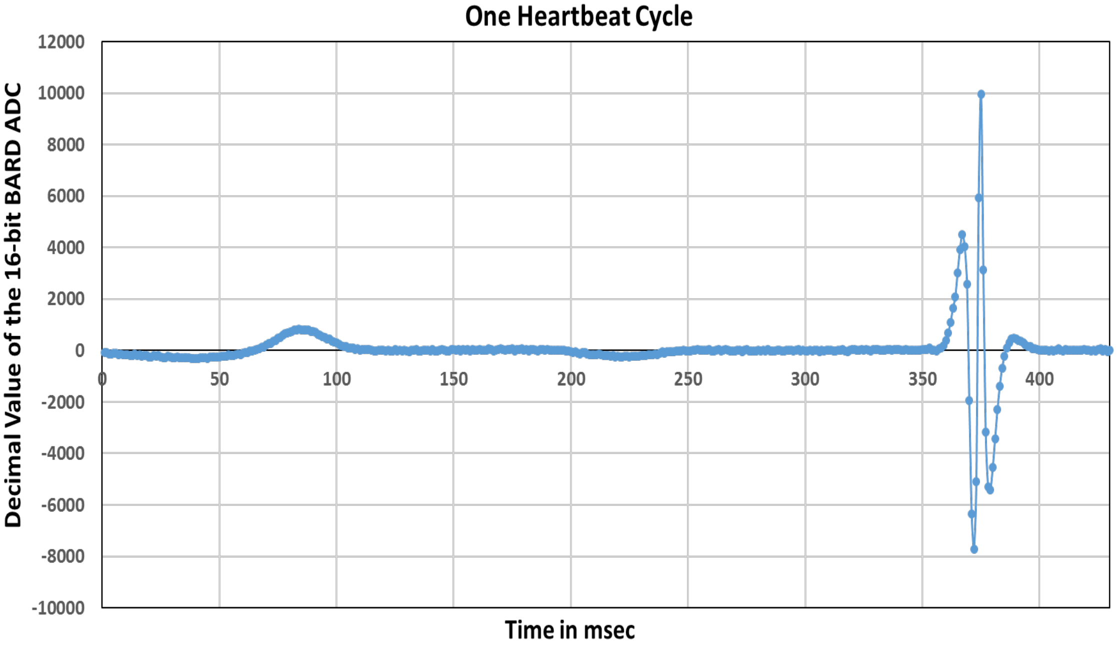


Figure 4.12: Time-domain waveform of a single beat (actual patient cardiac data).

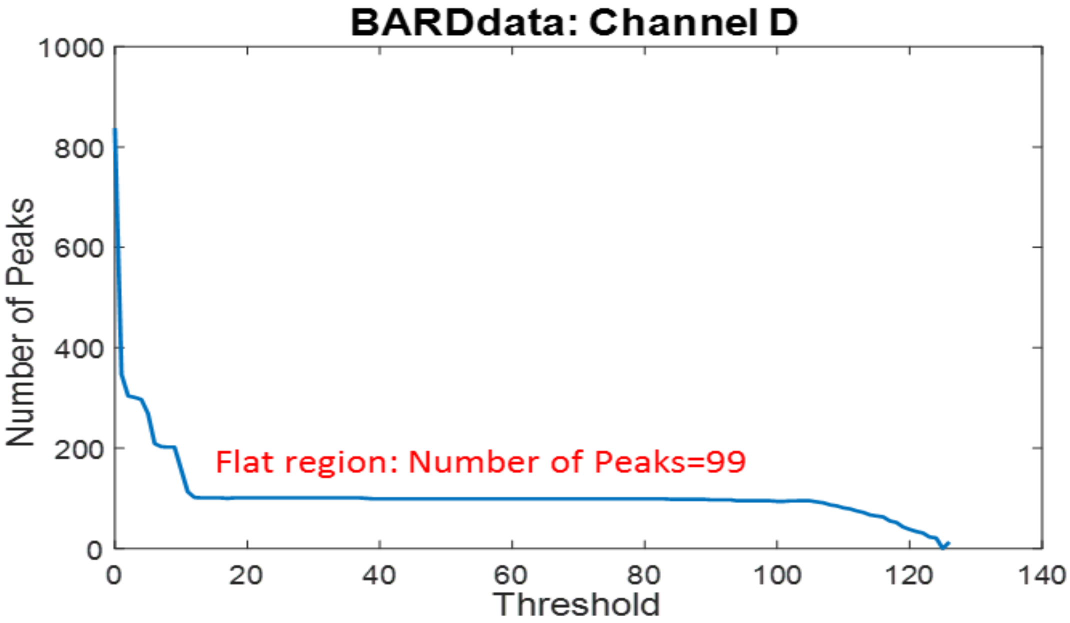
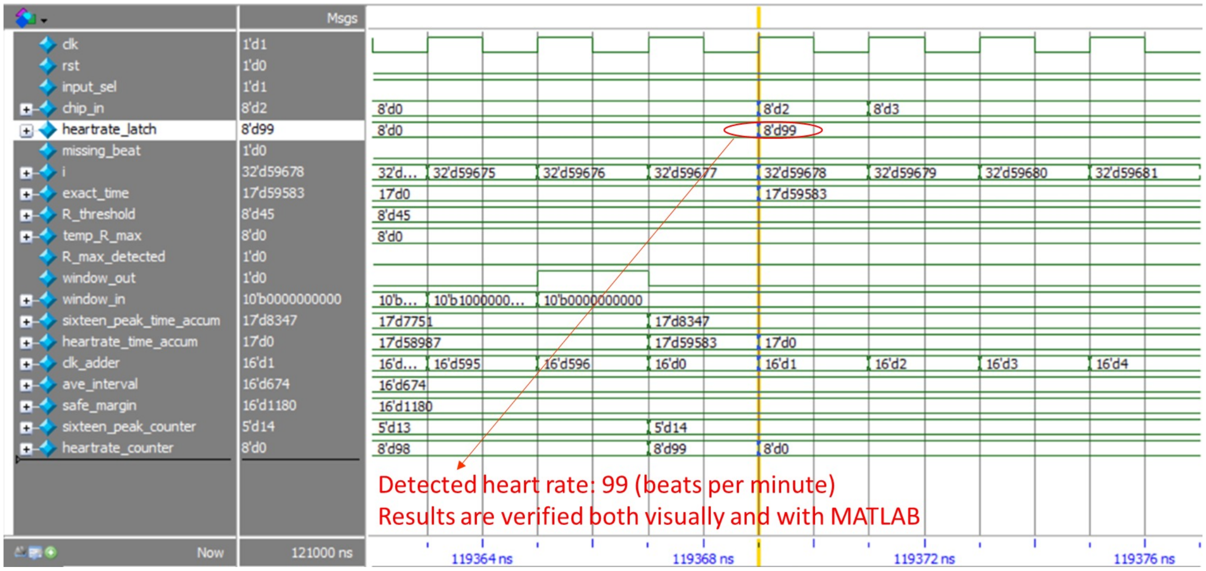


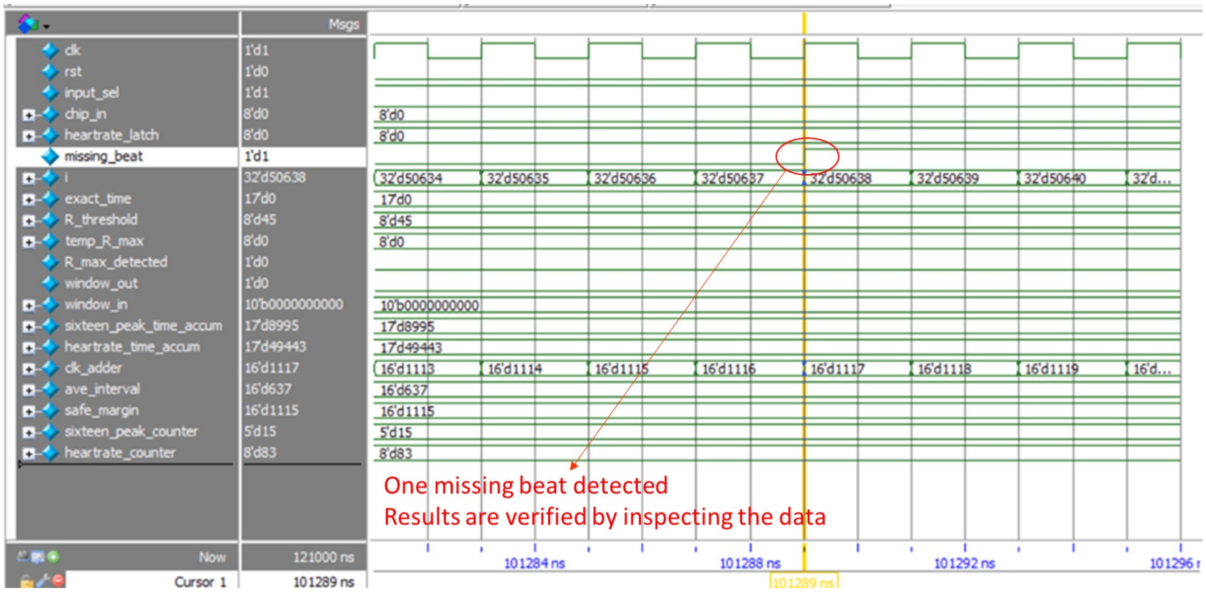
Figure 4.13: Number of peaks vs. peak threshold voltage simulated in MATLAB.



	Msgs
clk	1'd1
rst	1'd0
input_sel	1'd1
chip_in	8'd2
heartrate_latch	8'd99
missing_beat	1'd0
i	32'd59678
exact_time	17'd59583
R_threshold	8'd45
temp_R_max	8'd0
R_max_detected	1'd0
window_out	1'd0
window_in	10'b0000000000
sixteen_peak_time_accum	17'd8347
heartrate_time_accum	17'd0
clk_adder	16'd1
ave_interval	16'd674
safe_margin	16'd1180
sixteen_peak_counter	5'd14
heartrate_counter	8'd0
Now	121000 ns

Detected heart rate: 99 (beats per minute)
Results are verified both visually and with MATLAB

Figure 4.14: Modelsim simulations for detecting the heart-rate.



4.2.4 Power Analysis

The power consumption of the ASIC chip is composed of static and dynamic components. The static part is due to the leakage of the digital gates. The entire ASIC chip designed in this work, consumes a static power of $80nW$ and a dynamic power of $238nW$. The heart-rate detection unit consumes a static power of $48nW$ and a dynamic power of $164nW$. These values are are estimated using Synopsys Design Compiler based on the foundry provided library. The dynamic power is a linear function of the clock frequency and is due to the switching activities of the gates. Figure 4.17 shows the dynamic power as well as the total power as a function of clock frequency plotted on a logarithmic scale for the portion of the design that performs heart-rate detection. Equation 4.1 provides the total power of the whole design (heart-rate and missing beats detection units) as a function of the clock frequency at nominal VDD=$5V$. To further reduce the power consumption, the supply voltage (VDD) can be reduced. Since the model of the logic gates provided by the foundry are defined only at nominal VDD=$5V$, HSpice simulations are performed to demonstrate that the ASIC chip can still function correctly if VDD is lowered down to $1.7V$, as shown in Figure 4.18.

$$TotalPower = 80nW + (238nW \times \frac{f_{clk}}{1KHz}) \tag{4.1}$$

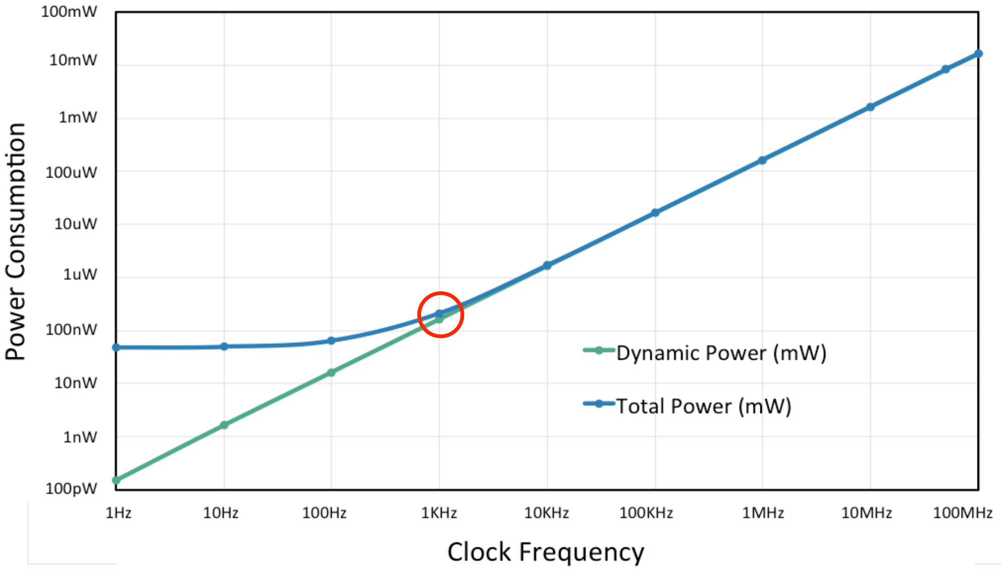


Figure 4.17: Dynamic and total power of heart-rate detection unit at nominal VDD=5V.

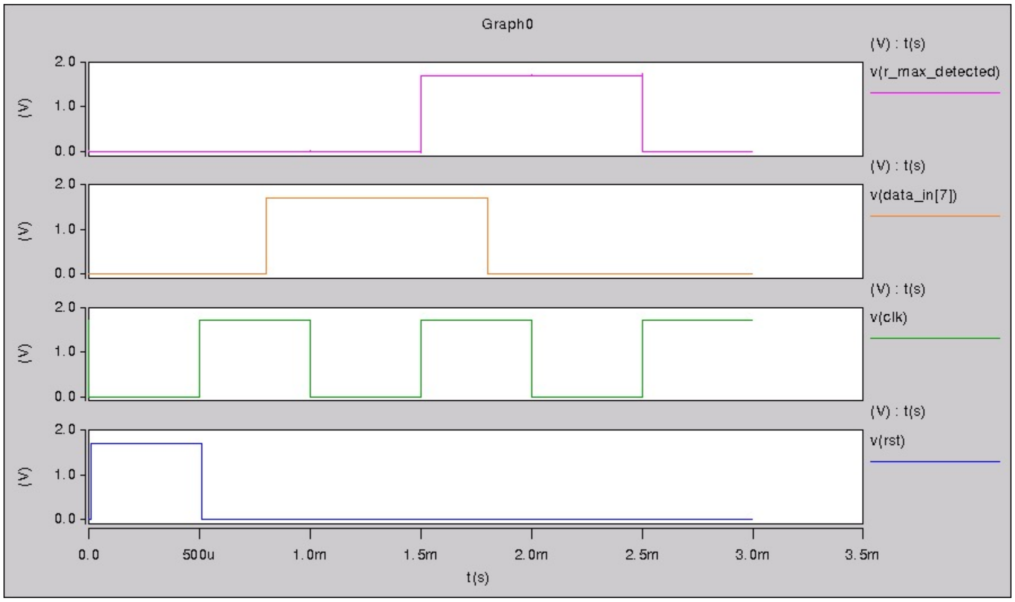


Figure 4.18: HSpice transient simulations at VDD=1.7V.

4.2.5 Conclusions

The reported ASIC chip calculates the number of beats per minute and detects missing beats. To meet the power consumption requirements, traditional power hungry signal processing techniques, that are used in the existing pacemakers, are avoided. The entire ASIC is implemented without using any computationally intensive operations such as FFT or even a division or multiplication. In this design whenever possible, we took advantage of the approximate computation without losing the accuracy to minimize the power consumption. The chip consumes a total power of $318nW$ at a clock frequency of $1KHz$ and supply voltage of $5V$. The power consumption of the reported ASIC chip is far below $100\mu W$ that can be harvested with an electromagnetic energy harvesting circuit [26, 25]. The low power consumption of the ASIC and its small size makes it possible to integrate this module with a wirelessly-powered implantable pacemaker chip that is shown in Figure 4.9.

REFERENCES

[1] K. Nii, Y. Taniguchi, and K. Okuyama, "A cost-effective embedded nonvolatile memory with scalable lee flash®-g2 sonos for secure iot and computing-in-memory (cim) applications," in *2020 International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*, 2020, pp. 1–4.

[2] O. Golonzka, U. Arslan, P. Bai, M. Bohr, O. Baykan, Y. Chang, A. Chaudhari, A. Chen, J. Clarke, C. Connor, N. Das, C. English, T. Ghani, F. Hamzaoglu, P. Hentges, P. Jain, C. Jezewski, I. Karpov, H. Kothari, R. Kotlyar, B. Lin, M. Metz, J. Odonnell, D. Ouellette, J. Park, A. Pirkle, P. Quintero, D. Seghete, M. Sekhar, A. S. Gupta, M. Seth, N. Strutt, C. Wiegand, H. J. Yoo, and K. Fischer, "Non-volatile rram embedded into 22ffl finfet technology," in *2019 Symposium on VLSI Technology*, 2019, pp. T230–T231.

[3] K. Lee, K. Yamane, S. Noh, V. B. Naik, H. Yang, S. H. Jang, J. Kwon, B. Behin-Aein, R. Chao, J. H. Lim, S. K., K. W. Gan, D. Zeng, N. Thiyagarajah, L. C. Goh, B. Liu, E. H. Toh, B. Jung, T. L. Wee, T. Ling, T. H. Chan, N. L. Chung, J. W. Ting, S. Lakshmipathi, J. S. Son, J. Hwang, L. Zhang, R. Low, R. Krishnan, T. Kitamura, Y. S. You, C. S. Seet, H. Cong, D. Shum, J. Wong, S. T. Woo, J. Lam, E. Quek, A. See, and S. Y. Siah, "22-nm fd-soi embedded mram with full solder reflow compatibility and enhanced magnetic immunity," in *2018 IEEE Symposium on VLSI Technology*, 2018, pp. 183–184.

[4] K. Nishioka, H. Honjo, S. Ikeda, T. Watanabe, S. Miura, H. Inoue, T. Tanigawa, Y. Noguchi, M. Yasuhira, H. Sato, and T. Endoh, "Novel quad interface mtj technology and its first demonstration with high thermal stability and switching efficiency for stt-mram beyond 2xnm," in *2019 Symposium on VLSI Technology*, 2019, pp. T120–T121.

[5] O. Golonzka, J. G. Alzate, U. Arslan, M. Bohr, P. Bai, J. Brockman, B. Buford, C. Connor, N. Das, B. Doyle, T. Ghani, F. Hamzaoglu, P. Heil, P. Hentges, R. Jahan, D. Kencke, B. Lin, M. Lu, M. Mainuddin, M. Meterelliyoz, P. Nguyen, D. Nikonov, K. O'brien, J. Donnell, K. Oguz, D. Ouellette, J. Park, J. Pellegren, C. Puls, P. Quintero, T. Rahman, A. Romang, M. Sekhar, A. Selarka, M. Seth, A. J. Smith, A. K. Smith, L. Wei, C. Wiegand, Z. Zhang, and K. Fischer, "Mram as embedded non-volatile memory solution for 22ffl finfet technology," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 18.1.1–18.1.4.

[6] F. Arnaud, P. Zuliani, J. Reynard, A. Gandolfo, F. Disegni, P. Mattavelli, E. Gomiero, G. Samanni, C. Jahan, R. Berthelon, O. Weber, E. Richard, V. Barral, A. Villaret, S. Kohler, J. Grenier, R. Ranica, C. Gallon, A. Souhaite, D. Ristoiu, L. Favennec, V. Caubet, S. Delmedico, N. Cherault, R. Beneyton, S. Chouteau, P. Sassoulas, A. Vernhet, Y. Le Friec, F. Domengie, L. Scotti, D. Pacelli, J. Ogier, F. Boucard, S. Lagrasta,

D. Benoit, L. Clement, P. Boivin, P. Ferreira, R. Annunziata, and P. Cappelletti, "Truly innovative 28nm fdsoi technology for automotive micro-controller applications embedding 16mb phase change memory," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 18.4.1–18.4.4.

[7] P. Zuliani, A. Conte, and P. Cappelletti, "The pcm way for embedded non volatile memories applications," in *2019 Symposium on VLSI Circuits*, 2019, pp. T192–T193.

[8] F. Khan, "Charge trap transistors (ctt): Turning logic transistors into embedded non-volatile memory for advanced high-k/metal gate cmos technologies," Ph.D. dissertation, UCLA, 2020.

[9] J. Viraraghavan, D. Leu, B. Jayaraman, A. Cestero, R. Kilker, M. Yin, J. Golz, R. R. Tummuru, R. Raghavan, D. Moy, T. Kempanna, F. Khan, T. Kirihata, and S. Iyer, "80kb 10ns read cycle logic embedded high-k charge trap multi-time-programmable memory scalable to 14nm fin with no added process complexity," in *2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits)*, 2016, pp. 1–2.

[10] X. Gu, "Charge-trap transistors for neuromorphic computing," Ph.D. dissertation, UCLA, 2018.

[11] F. Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. C. S. Woo, and S. S. Iyer, "The impact of self-heating on charge trapping in high- k -metal-gate nfets," *IEEE Electron Device Letters*, vol. 37, no. 1, pp. 88–91, 2016.

[12] F. Khan, M. S. Han, D. Moy, R. Katz, L. Jiang, E. Banghart, N. Robson, T. Kirihata, J. C. S. Woo, and S. S. Iyer, "Design optimization and modeling of charge trap transistors (ctts) in 14 nm finfet technologies," *IEEE Electron Device Letters*, vol. 40, no. 7, pp. 1100–1103, 2019.

[13] F. Khan, D. Moy, D. Anand, E. Schroeder, R. Katz, L. Jiang, E. Banghart, N. Robson, and T. Kirihata, "Turning logic transistors into secure, multi-time programmable, embedded non-volatile memory elements for 14 nm finfet technologies and beyond," in *2019 Symposium on VLSI Technology*, 2019, pp. T116–T117.

[14] S. Nouri and S. S. Iyer, "Non-volatile wideband frequency tuning of a ring-oscillator by charge trapping in high-k gate dielectric in 22nm cmos," *IEEE Electron Device Letters*, vol. 42, no. 1, pp. 110–113, 2021.

[15] K. J. Kuhn, M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, "Process technology variation," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2197–2208, 2011.

[16] A. Kral, F. Behbahani, and A. Abidi, "Rf-cmos oscillators with switched tuning," in *Proceedings of the IEEE 1998 Custom Integrated Circuits Conference (Cat. No.98CH36143)*, 1998, pp. 555–558.

[17] L. M. Dani, N. Mishra, S. K. Banchhor, S. Miryala, A. Doneria, and B. Anand, "Design and characterization of bulk driven mos varactor based vco at near threshold regime," in *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 2018, pp. 1–2.

[18] T. Yoshio, T. Kihara, and T. Yoshimura, "A 0.55 v back-gate controlled ring vco for adcs in 65 nm sotb cmos," in *2017 IEEE Asia Pacific Microwave Conference (APMC)*, 2017, pp. 946–948.

[19] F. Gerfers, N. Lotfi, E. Wittenhagen, H. Ghafarian, Y. Tian, and M. Runge, "Body-bias techniques in cmos 22fdx® for mixed-signal circuits and systems," in *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2019, pp. 466–469.

[20] A. Bajwa, S. Jangam, S. Pal, N. Marathe, T. Bai, and T. Fukushima, "Heterogeneous integration at fine pitch ($\leq$ 10 um) using thermal compression bonding," in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, pp. 1276–1284.

[21] C. Kothandaraman, X. Chen, D. Moy, D. Lea, S. Rosenblatt, F. Khan, D. Leu, T. Kirihata, D. Ioannou, G. LaRosa, J. B. Johnson, N. Robson, and S. S. Iyer, "Oxygen vacancy traps in hi-k/metal gate technologies and their potential for embedded memory applications," in *2015 IEEE International Reliability Physics Symposium*, 2015, pp. MY.2.1–MY.2.4.

[22] X. Gu, Z. Wan, and S. S. Iyer, "Charge-trap transistors for cmos-only analog memory," *IEEE Transactions on Electron Devices*, vol. 66, no. 10, pp. 4183–4187, 2019.

[23] F. Khan, E. Cartier, J. C. S. Woo, and S. S. Iyer, "Charge trap transistor (ctt): An embedded fully logic-compatible multiple-time programmable non-volatile memory element for high- k -metal-gate cmos technologies," *IEEE Electron Device Letters*, vol. 38, no. 1, pp. 44–47, 2017.

[24] S. Kim, S. J. Choi, K. Zhao, H. Yang, G. Gobbi, S. Zhang, and J. Li, "Electrochemically driven mechanical energy harvesting," *Nature communications*, vol. 7, no. 1, pp. 1–7, 2016.

[25] A. Ben Amar, A. B. Kouki, and H. Cao, "Power approaches for implantable medical devices," *sensors*, vol. 15, no. 11, pp. 28 889–28 914, 2015.

[26] Y. Sun, B. Greet, D. Burkland, M. John, M. Razavi, and A. Babakhani, "Wirelessly powered implantable pacemaker with on-chip antenna," in *2017 IEEE MTT-S International Microwave Symposium (IMS)*, 2017, pp. 1242–1244.

[27] A. Zurbuchen, A. Haeberlin, A. Pfenniger, L. Bereuter, J. Schaerer, F. Jutzi, C. Huber, J. Fuhrer, and R. Vogel, "Towards batteryless cardiac implantable electronic devices—the swiss way," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 1, pp. 78–86, 2017.

[28] N. H. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective.* Pearson Education India, 2015.

[29] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger, and F. Solzbacher, "A low-power integrated circuit for a wireless 100-electrode neural recording system," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 1, pp. 123–133, 2007.

[30] R. M. Walker, H. Gao, P. Nuyujukian, K. Makinwa, K. Shenoy, T. Meng, and B. Murmann, "A 96-channel full data rate direct neural interface in 0.13 μm cmos," in *2011 Symposium on VLSI Circuits-Digest of Technical Papers.* IEEE, 2011, pp. 144–145.

[31] P. Ritter, G. Z. Duray, C. Steinwender, K. Soejima, R. Omar, L. Mont, L. V. Boersma, R. E. Knops, L. Chinitz, S. Zhang *et al.*, "Early performance of a miniaturized leadless cardiac pacemaker: the micra transcatheter pacing study," *European heart journal*, vol. 36, no. 37, pp. 2510–2519, 2015.

[32] S. K. Mulpuru, M. Madhavan, C. J. McLeod, Y.-M. Cha, and P. A. Friedman, "Cardiac pacemakers: function, troubleshooting, and management: part 1 of a 2-part series," *Journal of the American College of Cardiology*, vol. 69, no. 2, pp. 189–210, 2017.

[33] M. Madhavan, S. K. Mulpuru, C. J. McLeod, Y.-M. Cha, and P. A. Friedman, "Advances and future directions in cardiac pacemakers: part 2 of a 2-part series," *Journal of the American College of Cardiology*, vol. 69, no. 2, pp. 211–235, 2017.

[34] www.bardaccess.com. Bard system.