12-2017

Design and Evaluation of a Sub-1-Volt Read Flash Memory in a Standard 130 Nanometer CMOS Process

David Andrew Basford
University of Tennessee, Knoxville, dbasford@vols.utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by David Andrew Basford entitled "Design and Evaluation of a Sub-1-Volt Read Flash Memory in a Standard 130 Nanometer CMOS Process." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

Jeremy Holleman, Major Professor

We have read this thesis and recommend its acceptance:

Benjamin Blalock, Syed Islam

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Design and Evaluation of a Sub-1-Volt Read Flash Memory in a Standard 130 Nanometer CMOS Process

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

David Andrew Basford

December 2017

Acknowledgements

I would like to thank the National Science Foundation for its partial funding of this project. The flash memory design presented here was intended to be part of a complete SoC solution for recording bat sounds in flight. The complete SoC was not finished, but the research on the flash memory design continued so that it could be applied to future projects.

Thanks also goes to Peixing Liu and Mohsen Judy for their contribution to the original work on the flash memory design. Their work was part of a class project that brought together a variety of scripting and automation tools to synthesize the complete flash memory for fabrication. The design has undergone several revisions since their involvement in the project. The majority of the circuit blocks have subsequently been replaced, but the sense amplifiers and phase generator remain intact.

I would like to give appreciation to Junjie Lu for his analog memory cell design. It turned out to be exceptionally useful for characterization of the CMOS process used in this project. Ultimately a condensed version of the circuit was used for the final flash memory cell design in this project.

I would also like to thank Jeremy Langford for his help in my initial foray into Cypress PSoC programming as a means of testing our team's ASIC designs. We spent many hours discussing our respective projects. He helped me learn ways to improve my C programming skills for embedded applications, and I helped him with the design of the analog front end for his bat board project.

Special thanks goes to Brian Coghlan for his help with the PSoC/MATLAB data connection as well as for performing extensive data retention tests on the analog memory

cells. The serial connection between the PSoC and MATLAB proved to be difficult to implement properly, but he was able to get it working reliably. Being able to download my test data directly from the PSoC to MATLAB meant that I could perform tests more efficiently. Additionally Brian's testing of the retention characteristics of the analog memory array was important for narrowing down the possible sources of leakage in the cells.

Finally, I would like to thank the members of my thesis committee. As the chairperson, Dr. Jeremy Holleman was the person I consulted most often on a variety of aspects related to this project. He was instrumental in my understanding of the physics behind nonvolatile memory as well as methods for improving digital circuit design. Dr. Syed Islam was an excellent source of knowledge in all areas of semiconductor physics and VLSI design. Finally, Dr. Benjamin Blalock has greatly enhanced my knowledge of analog circuit design and evaluation. As a group, the committee members proved to be an invaluable resource during my undergraduate and graduate career.

Abstract

Nonvolatile memory design is a discipline that employs digital and analog circuit design techniques and requires knowledge of semiconductor physics and quantum mechanics. Methods for programming and erasing memory are discussed here, and simulation models are provided for Impact Hot Electron Injection (IHEI), Fowler-Nordheim (FN) tunneling, and direct tunneling. Extensive testing of analog memory cells was used to derive a set of equations that describe the floating-gate characteristics. Measurements of charge retention also revealed several leakage mechanisms, and methods for mitigating leakage are presented.

Fabrication of flash memory in a standard CMOS process presents significant design challenges. The absence of multiple polysilicon layers requires that additional devices be used to control the floating-gate voltage. Furthermore high-voltage devices are often required to isolate the selected memory cells during write and erase cycles. However, a single-poly design allows portability to another standard process provided that the floating-gate characteristics are known.

A flash memory system is presented here that has been fabricated in a standard 130 nanometer CMOS process. The design utilizes capacitive feedback to maintain desired injection current during programming. It also includes a sense amplifier design which features auto-zeroing of inherent offsets. Comparisons to existing memory designs show that a significant improvement in areal density was achieved through the elimination of on-die high-voltage charge pumps and switches. Measurements were performed over a range of clock frequencies and supply voltages. Results show that this memory system is capable of a read access time of 3.5 microseconds with a 1 megahertz clock while consuming less than

25 microwatts from a 1 volt supply. Operation down to 650 millivolts was confirmed where power consumption was reduced to only 3.4 microwatts. The low power consumption and high density of this flash memory make it an excellent choice for on-die firmware storage in battery-powered embedded applications.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Flash memory is widely used for a variety of applications where nonvolatile storage is required. One such application is for the program memory on a microcontroller. This application requires fast random access read times, low power consumption, and high density. For application-specific integrated circuit (ASIC) designs it may be desired that the microcontroller and flash memory be fabricated together on the same die. This integration may limit the designer to processes that are specifically designed for nonvolatile memory. If the microcontroller has already been synthesized in a standard CMOS process, then changing the process to suit the flash memory may not be feasible. Instead it has been shown that nonvolatile memory can be fabricated without special structures [1]. The flash memory design presented here has been fabricated in a standard 130 nm CMOS process.

1.1 Statement of Problem

The goal of this project is to design a flash memory system that is intended to be used as firmware storage for an on-die microcontroller unit (MCU). Integration of flash memory onto the same chip as the MCU presents several design challenges. According to [2], the selection of flash topology depends on factors such as density requirements, total storage capacity, and the availability of special structures and layers.

For the flash memory presented here, write and erase operations will be performed before deployment. This leads to significant area savings since on-die high voltage switches and charge pumps are not required. Additionally it is desired that the MCU be able to execute single-cycle instructions, so the memory must have fast random read access times.

1.2 Single-Polysilicon Limitations

Typically nonvolatile memory systems are fabricated in processes with special structures such as multiple polysilicon layers. As seen in Figure 1.1, this allows transistors to have both a floating gate and a control gate. There are also designs in which the conductive poly floating-gate is replaced by a silicon-nitride insulating layer to store trap charge. This structure is called silicon-oxide-nitride-oxide-silicon (SONOS), and its application replaces dual-poly in smaller process nodes where tunneling leakage becomes significant. One such design is provided by [3] in which the structure is similar to Figure 1.1. Another design adds an address gate to the SONOS structure [4]. In this design the control gate and nitride layers partially overlap the address gate. This dual-gate structure allows for row and column selectivity within the cell, and it is ideal where high areal density is required. A further refinement shown by [5] demonstrates that local charge trapping allows the nitride layer to be shared between two adjacent cells.

The CMOS process used for this project is a standard single-polysilicon type. Additional devices are necessary in order to maintain control of the floating gate voltage. One possible configuration for a single-poly flash cell is shown in Figure 1.2 [1]. Although this configuration requires additional die area compared to the dual-poly design, it does allow flexibility in the relative sizes of the devices. For example if the gate area of the PMOS device is several times as large as that of the NMOS device, then capacitive voltage division would cause the bulk of the applied voltage to appear across the oxide of the NMOS.

This particular configuration was demonstrated by [6] where it was fabricated in a standard 0.5 µm CMOS process with a C_{GP}/C_{GN} ratio of 5. Programming was performed by applying 10.5 V pulses to the source and drain of the NMOS while all PMOS terminals were

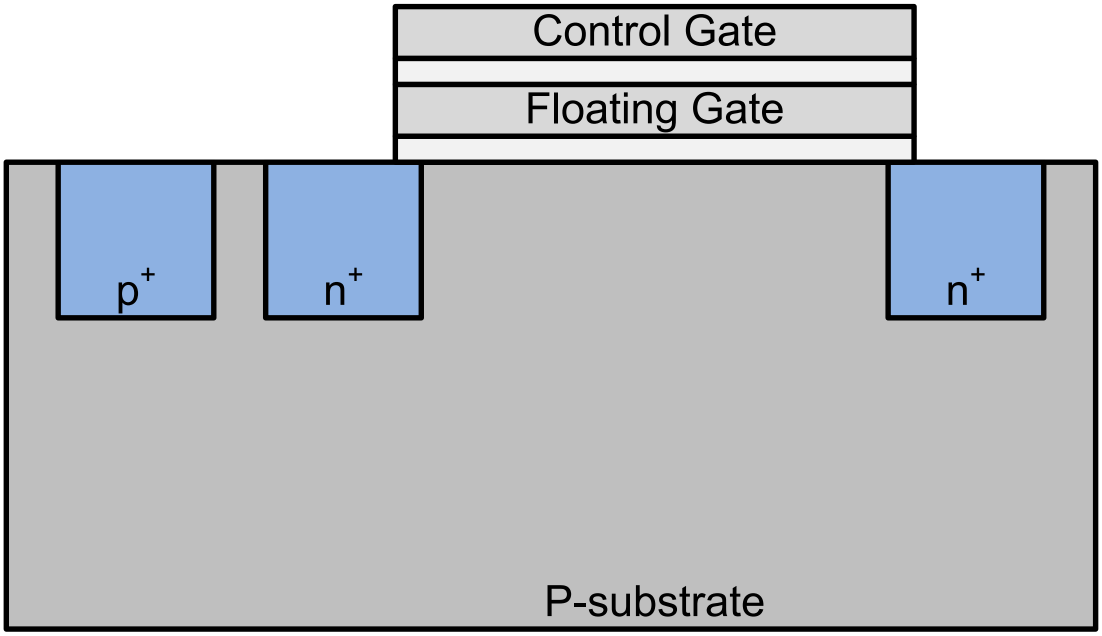

Figure 1.1: Dual-poly floating-gate structure

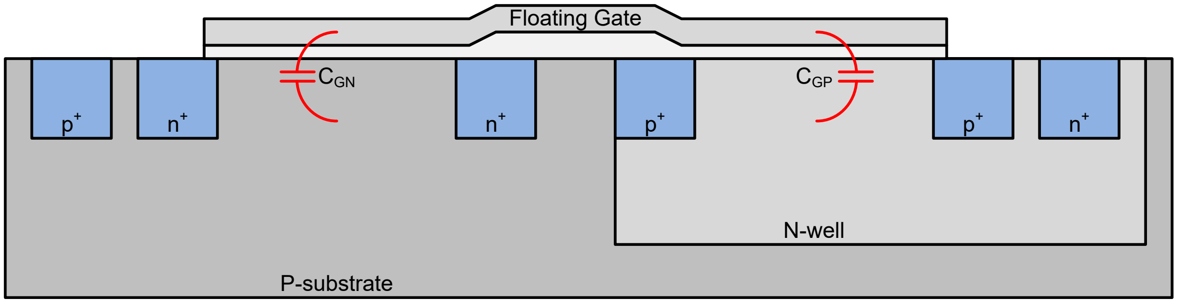

Figure 1.2: Single-poly floating-gate structure [1]

grounded. Likewise erasing was possible by applying that same high voltage to all PMOS terminals while grounding the NMOS source and drain.

1.3 Low-Power, Low-Voltage Requirements

The flash memory will be powered by a small, lightweight battery, so power consumption of the system while in read mode should be on the order of 10 μW. The typical supply voltage during read operations will be 1 V, though operation down to 0.5 V should be possible at lower clock speeds.

Chapter 2

Fundamentals of Operation

Designing a flash memory system requires an understanding of the physics involved in the operation of nonvolatile memory. The rates at which charge is added or removed from a floating gate depend on several circuit voltages and currents. The mechanics are also highly dependent on process parameters such as oxide thickness and dopant concentrations. The 130 nm process used in this paper was not intended for nonvolatile memory applications, so the foundry does not supply floating-gate characteristics in the design kit. For this reason it was necessary to measure the characteristics to develop a set of empirical models.

2.1 Methods for Writing and Erasing

There are several methods often employed for programming and erasing nonvolatile memory. The design presented in this paper utilizes Impact Hot Electron Injection (IHEI) for programming and Fowler-Nordheim (FN) tunneling for erasing.

2.1.1 Impact Hot Electron Injection

In IHEI, holes which are accelerated by a large electric field through the channel collide with the lattice near the drain to generate free electron-hole pairs. Given a strong enough electrical field across the oxide, the free electrons can overcome the oxide barrier and move from the channel to the floating gate. The decrease in stored charge lowers the floating-gate voltage, and during read operation this increases V_{sg} of the PMOS read transistor. If the charge is sufficiently depleted, then the cell will be interpreted as a logic 1 during read operation.

2.1.2 Tunneling

Tunneling is described by a quantum mechanical phenomenon in which charge may overcome an energy barrier in spite of having insufficient energy. There are several mechanisms in which tunneling may occur, and they are not mutually exclusive. The dominant mechanism depends on several factors including oxide thickness, the electric field across the oxide, and stress due to exposure to charges colliding with the oxide structure.

At lower oxide voltages, the oxide's conduction band is completely above that of the silicon, so the barrier is trapezoidal. Figure 2.1 shows the energy band diagram for direct tunneling. This tunneling mechanism is becoming increasingly significant as oxide thicknesses continue to decrease with newer CMOS processes. This has a significant effect on the retention of stored charge on a floating gate, a subject that will be discussed later in this chapter.

At higher oxide voltages FN tunneling is the dominant mechanism. Figure 2.2 shows the energy band diagram for FN tunneling. Under this condition the large electric field causes the oxide's energy bands to tilt to such a degree that a portion of its conduction band is at or below the conduction band of the silicon. This reduces the effective oxide thickness and increases the probability that electrons may tunnel through to the oxide's conduction band. Once this occurs the electrons are free to move to the gate's conduction band.

Another significant tunneling mechanism is caused by traps within the oxide. Figure 2.3 shows the energy band diagram for trap-assisted tunneling. The density of these traps is usually low after fabrication, but repeated exposure to high electric fields across the oxide can substantially increase the number of traps. Rosenbaum and Register tested devices with oxide thicknesses of 5.5 and 7.5 nm, and they found that there was a significant increase

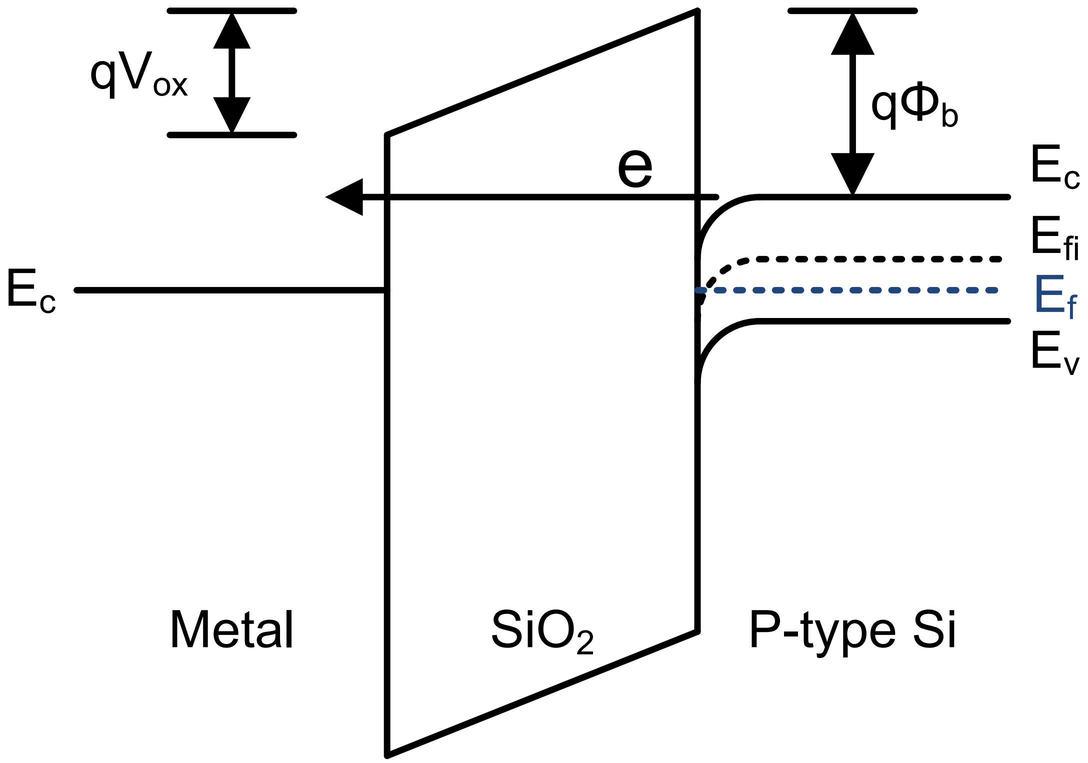

Figure 2.1: Direct tunneling

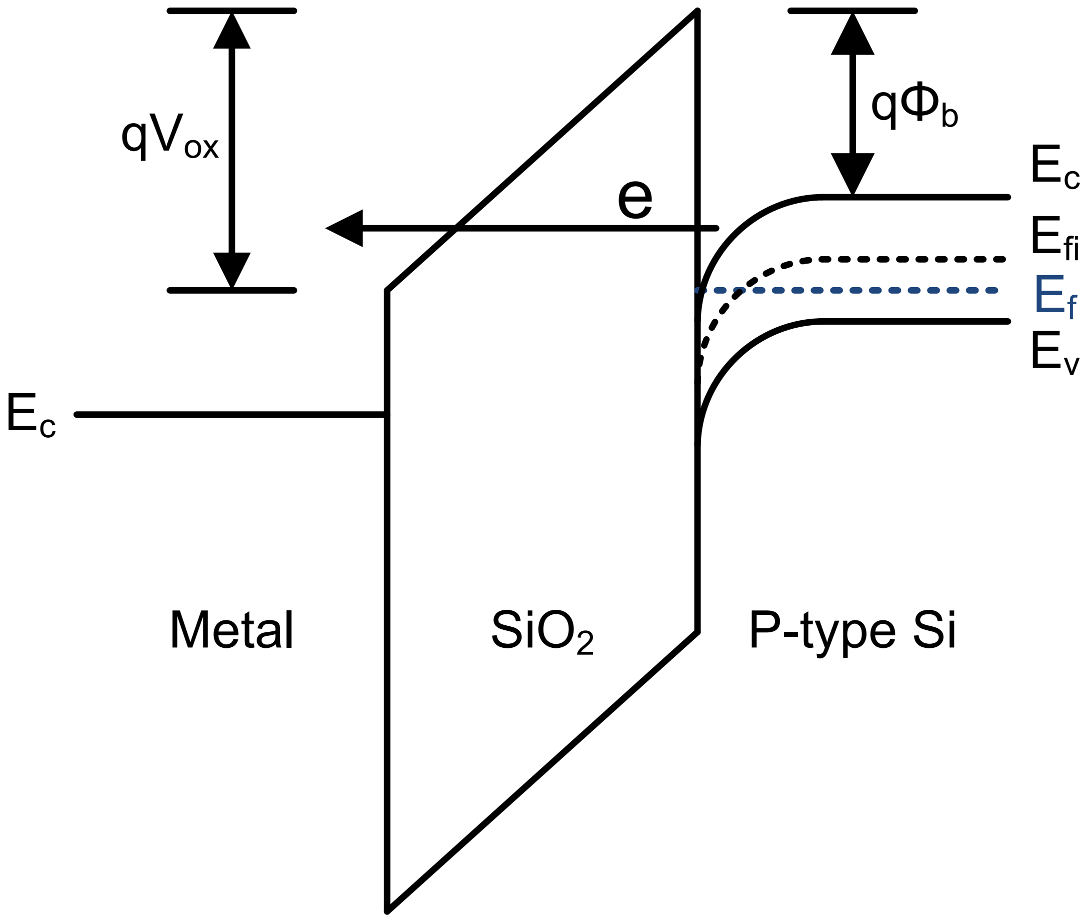

Figure 2.2: Fowler-Nordheim tunneling

in gate current at low oxide voltages after being subjected to stress [7]. Figure 2.4 shows the gate current versus oxide voltage for the two oxide thicknesses [7]. The dashed lines represents the gate current before stress, and solid lines show the current after 0.1 A/cm^2 was applied for 2 seconds. This degradation is important for nonvolatile memory, especially for architectures that utilize tunneling for both writing and erasing. The increase in traps within the oxide can significantly impact the charge retention and reliability of the memory.

2.2 Characterization and Derivation of a Simulation Model

In order to successfully design a nonvolatile memory system, a complete simulation model of the floating-gate characteristics is required. Characteristics vary significantly from one process to the next, so a new simulation model must be derived for the specific CMOS process in use.

2.2.1 Existing Publications

Many publications have been made on the modeling of gate currents in floating-gate devices. Some are aimed at providing approximate models for the purpose of simulation under typical operating conditions. Others provide extensive coverage of secondary effects such as charge quantization, image-force barrier lowering, and temperature. Ultimately the purpose of characterizing the CMOS process for this paper was to formulate a practical set of simulation models to predict the behavior of the final design.

Injection

Rahimi et al. [8] proposed a simulation model for the IHEI phenomenon in a PMOS injector which is described using terminal voltages and currents in (2.1). I_s is the source current, and V_{gd} and V_{sd} are the gate-drain and source-drain voltages respectively. Fit constants α, β, δ, and λ are empirically derived and depend on fabrication process parameters. From

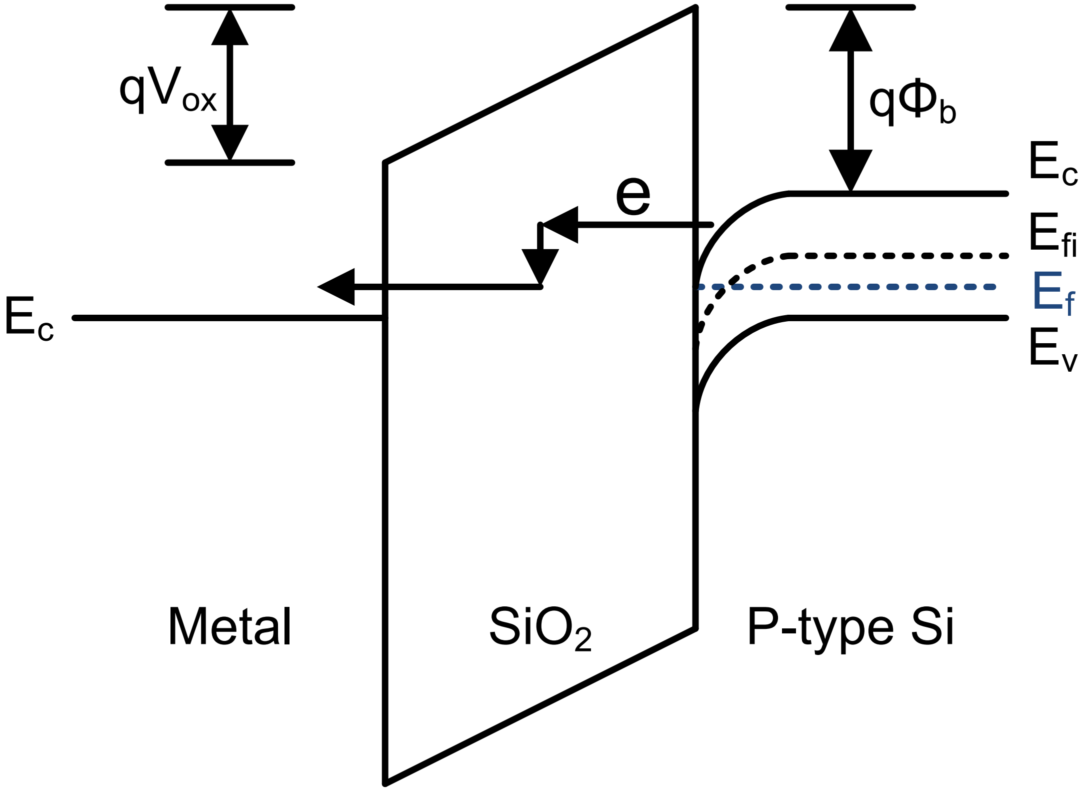

Figure 2.3: Trap-assisted tunneling

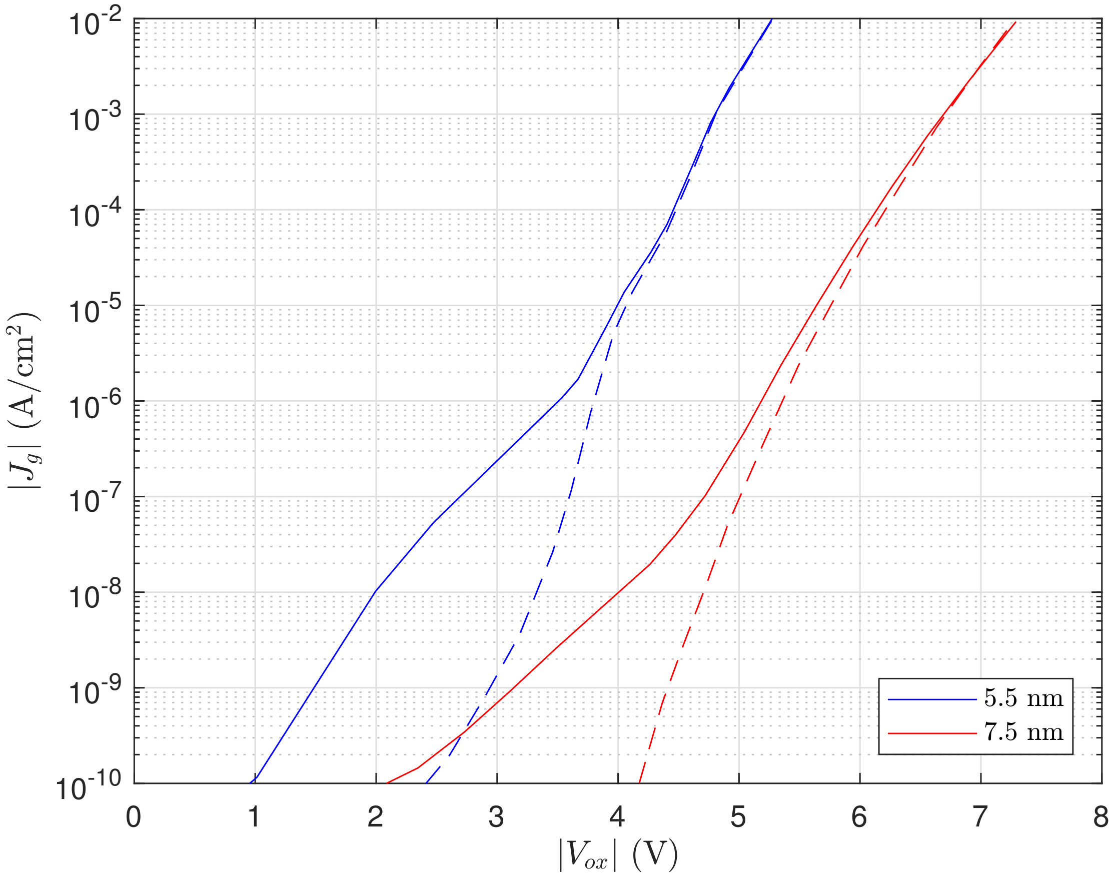

Figure 2.4: Gate current vs oxide voltage before and after stress [7]

this equation we can conclude that there are large dependencies on V_{gd} and V_{sd}. In order to achieve desired injection speed and efficiency, these voltages must be precisely controlled.

$$I_{inj} = \alpha I_s exp\left(\frac{-\beta}{(V_{gd}+\delta)^2} + \lambda V_{sd}\right) \tag{2.1}$$

Lu proposed an equation based on terminal voltages and currents as shown in (2.2) [9]. It is similar to (2.1), but the λV_{sd} term was omitted and V_{gd} was changed to V_{sd}. This equation was originally used in the early design stages of the flash memory system in this paper, however the omission of the V_{gd} dependence was found to be incorrect.

$$I_{inj} = \alpha I_s exp\left(\frac{-\beta}{(V_{sd}+\delta)^2}\right) \tag{2.2}$$

The electron-hole pairs generated by impact ionization leave behind holes after the hot electrons have been swept into the floating gate. The holes are collected by the drain, and this can be modeled as a body current as shown in (2.3) [8] where γ, κ, and λ are fit constants.

$$I_b = \eta I_s(\gamma V_{sd} - \kappa V_{sg} + V_t) exp\left(\frac{-\lambda}{\gamma V_{sd} - \kappa V_{sg} + V_t}\right) \tag{2.3}$$

To illustrate the strong dependence on V_{gd}, Figure 2.5 shows the effect of V_g on I_g and I_{sub} [10]. As V_g decreases, V_{sg} increases leading to an increase in I_s. This increases the generation of electron-hole pairs. While V_{sg} is low V_{gd} is high, so there is a strong electric field across the oxide near the drain that attracts the hot electrons. As V_g decreases further, the decrease in V_{gd} overshadows the increase in I_s, so I_g decreases. However, I_{sub} continues to increase implying that the generation of electron-hole pairs is still increasing. As V_g decreases further, the depletion region between the channel and the drain collapses, and the electron-hole pair generation declines.

Tunneling

A model for a single-poly, triple-device EEPROM cell was provided by Li et al. [11]. The cell uses tunneling for both programming and erasing. The equation (2.4) describes the charges

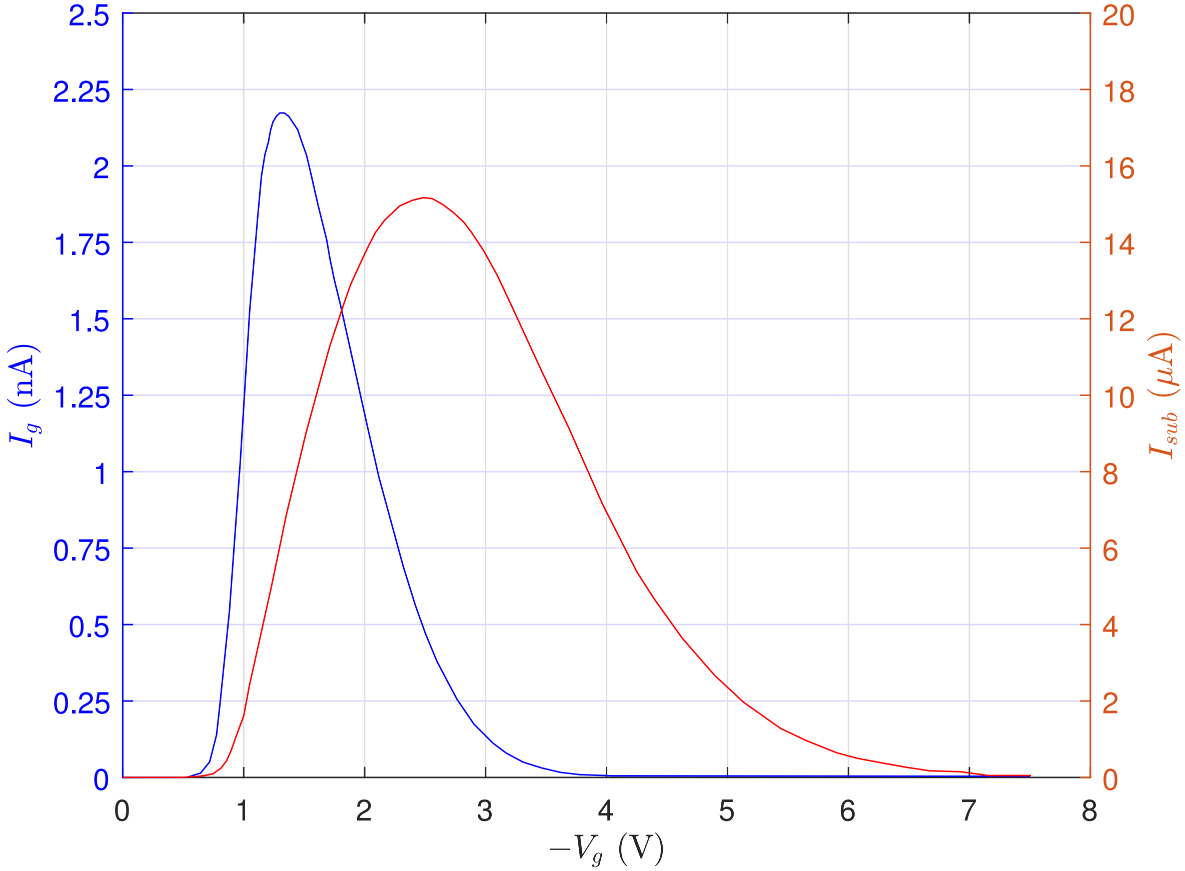

Figure 2.5: PMOS gate and substrate currents vs gate voltage [10]

and voltage-dependent capacitances of the circuit. The subscripts T, C, and are R refer to the tunneling, control, and read devices respectively. For cases where the inclusion of capacitances between individual terminals is necessary, (2.5) may be substituted into (2.4) for each device as required. Li postulated that ΔQ_{fg} can be determined from simulation, however this assumes that the tunneling current I_g is known. The actual equation describing the tunneling current was not provided.

$$Q_{fg} - Q_{g\_T,0} - Q_{g\_C,0} - Q_{g\_R,0} = \int_0^{V_{gt}} C_{gg\_T} dV_{gt} + \int_0^{V_{gc}} C_{gg\_C} dV_{gc} + \int_0^{V_{gr}} C_{gg\_R} dV_{gr} \tag{2.4}$$

$$Q_g - Q_{g,0} = \int_0^{V_{gb}} C_{gb} dV_{gb} + \int_0^{V_{gs}} C_{gs} dV_{gs} + \int_0^{V_{gr}} C_{gr} dV_{gr} \tag{2.5}$$

Rahimi et al. provided (2.6) as a means of simulating FN tunneling currents [8]. In this equation I_{tun0} is the pre-exponential gate current, and V_f is an fit constant derived from measured data. There is an exponential dependence on V_{ox}, and Rahimi showed that the model fit the measured data exceptionally well over the V_{ox} range. However, the process used was 350 nm, so the oxide thickness is likely significantly higher than that of the 130 nm process used in this paper. This could mean Rahimi would not have observed the transition between direct and FN tunneling due to the limited voltage range that was tested.

$$I_{tun} = -I_{tun0} W L \exp\left(\frac{-V_f}{V_{ox}}\right) \tag{2.6}$$

Another commonly quoted equation for FN tunneling is (2.7) [12]. In this equation q it the charge of an electron, $\hbar$ is Planck's reduced constant, Φ_b is the barrier height, and E_{ox} is the electric field across the oxide. Some publications add or omit terms depending on the degree of accuracy required. Several publications include a mass term in the pre-exponential in the denominator, but Ranuárez omitted it [12]. Lenzlinger and Snow included compensations for image-force barrier lowering and temperature dependence [13]. Weinberg included terms for quantization of the electrons into subbands at the potential wall [14]. Regardless of these minor differences, it can be seen that (2.6) is missing an important dependence on V_{ox}^2 before the exponential. For this reason, (2.7) will be used in the characterization discussion in the next section.

$$J_{FN} = \frac{q^3}{16\pi^2\hbar\Phi_b m_{ox}^*} E_{ox}^2 \exp\left[-\frac{4}{3} \frac{\sqrt{2m_{ox}^*}\Phi_b^{3/2}}{\hbar q} \frac{1}{E_{ox}} \right] \tag{2.7}$$

FN tunneling is rather well understood, but the discussion of direct tunneling often leads to complex equations while attempting to model the gate current behavior at low oxide voltages. Ranuárez cited several equations that model direct tunneling [12]. An equation similar to one from [15] is shown as (2.8). It is quite nearly identical to (2.7), where the addition of the $1 - \left(1 - \frac{V_{ox}}{\Phi_b}\right)^{3/2}$ term provides the correction for low oxide voltage behavior. Lee and Hu proposed an even more precise equation [16], but it is quite complex and beyond the scope of this paper.

$$J_{direct} = \frac{q^3}{16\pi^2\hbar\Phi_b m_{ox}^*} E_{ox}^2 \exp\left[-\frac{4}{3} \frac{\sqrt{2m_{ox}^*}\Phi_b^{3/2}}{\hbar q} \frac{\left[1 - \left(1 - \frac{V_{ox}}{\Phi_b}\right)^{3/2}\right]}{E_{ox}} \right] \tag{2.8}$$

Ranuárez provided an excellent coverage of tunneling mechanisms [12]. Most publications on tunneling are in reference to NMOS devices. The devices used in the flash memory system presented in this paper are PMOS, so it is important to discuss the differences. Tunneling current falls into one of three different component mechanisms. In electron conduction-band tunneling (ECB) the electron moves from the conduction band on one side of the oxide to the conduction band on the other. Electron valence-band tunneling (EVB) describes the movement from the valence band across the oxide to the conduction band. Hole valence-band tunneling (HVB) involves the movement of holes from the valence band on one side to the valence band on the other. Table 2.1 shows the dominant current component for both NMOS and PMOS devices in each region of operation [12].

Table 2.1: Dominant tunneling current components [12]

Current component	I_{gc}	I_{gb}		I_{gs},I_{gd}
Region of operation	Inversion	$V_{ox} > 0$	$V_{ox} < 0$	All
NMOS	ECB	EVB	ECB	ECB
PMOS	HVB	ECB	EVB	HVB

2.2.2 Testing Methods

As seen in the equations above, the injection and tunneling rates are dependent on multiple variables. In order to determine the fit constants empirically, a method of isolating each dependence is necessary. For this purpose an array of 16 analog memory cells was fabricated using a design based on a circuit provided by Lu [9]. The circuit in Figure 2.6 is the initial design. It was previously used to store bias voltages on-die to be supplied to other circuits. The diodes were used to shift the output voltage down to a range around 300 to 900 mV.

Tests on the array of cells proved to be beneficial to the understanding of the floating-gate characteristics of the 130 nm process used in the design. However, there were issues with the design of the cell that affected the range of voltages that could be applied as well as the accuracy of the test results. The voltage drop due to V_{gs} of T5 and the forward drop of the diodes prevented V_{DD} from being reduced below about 2.5 V, so low voltage testing could not be performed. Devices T1 and T3 shared the same N-well, and during testing it was discovered that the body effect in T1 caused the V_{sd} dependence to be significantly different from the expected behavior described by (2.1).

To combat these problems a new array of 32 analog memory cells was fabricated. The new cell design shown in Figure 2.7 allows independent control of V_{BIAS}, V_{DDI}, and V_{INJ} for injection and V_{TUN} for tunneling. The device specifications are provided in Table 2.2. Compared to Figure 2.6, the source-follower T5 was changed to a zero-V_t device to reduce the V_{gs} drop, and the diodes were also removed. This allowed V_{DD} to be reduced to less 1 V without V_{OUT} collapsing. The injection device T1 was moved to its own N-well, and a level shifter was added to the injection control multiplexer. These changes allowed the operation of T1 to be fully independent of V_{DD}, and the body effect was eliminated.

In order to determine the actual injection and tunneling currents from the measured data, simulations were run to estimate the unknown values of V_g and C_g. These values cannot be measured directly because no external connections can me made to the floating gate. The terminal voltages and currents were estimated at the same operating points that were used in the measurements. Then the floating-gate capacitance was determined from ΔV_{out} as the

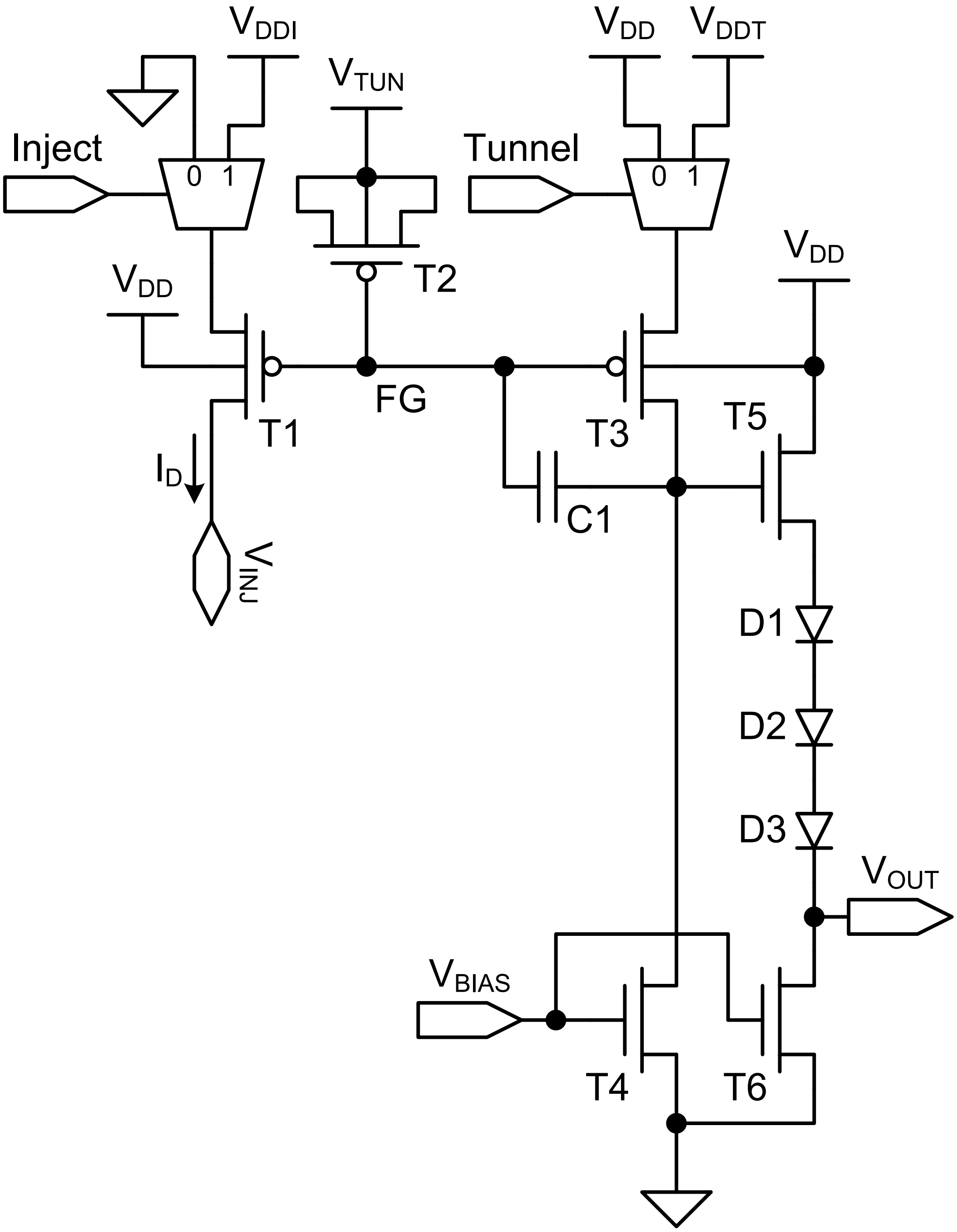

Figure 2.6: Initial analog memory cell schematic

Table 2.2: Analog cell device specifications

Device	**Type**	**W/L (μm)**	t_{ox} **(Å)**	**Value**
T1	PMOS	0.36/0.24	52	-
T2	PMOS	0.36/0.36	52	-
T3	PMOS	0.36/2.0	52	-
T4	NMOS	0.36/1.0	52	-
T5	NMOS (ZVT)	3.0/1.0	52	-
T6	NMOS	0.36/1.0	52	-
C1	dgncap	5.0/2.0	52	58.9 fF @ 3.3 V

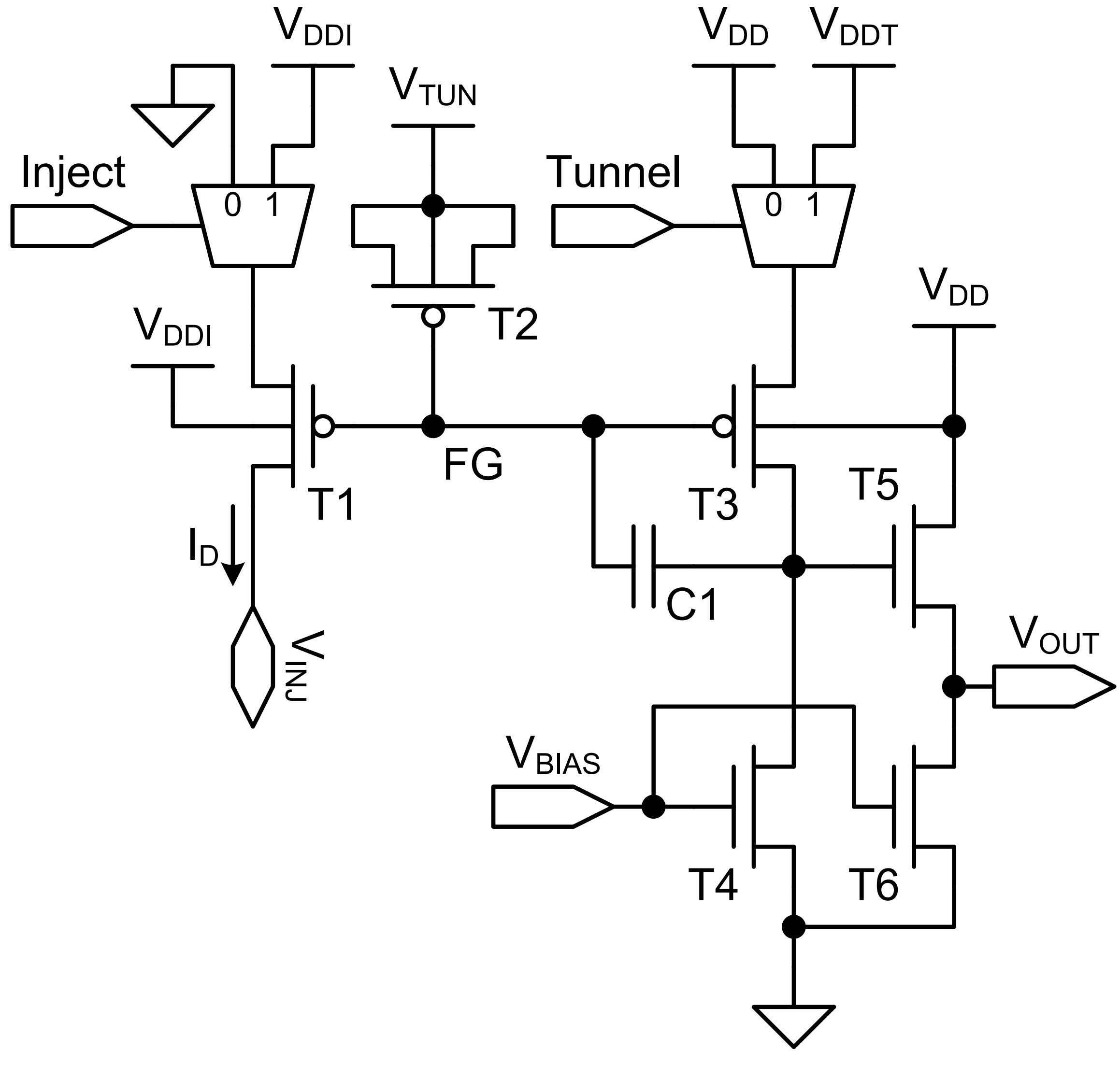

Figure 2.7: Analog memory cell schematic

result of injecting a fixed charge onto the gate. This simulation data was compiled into a lookup table so that V_{sd}, V_{gd}, V_{ox}, I_s, I_{inj}, and I_{tun} could be estimated from the measured data.

Injection

The injection equation (2.1) contains three dependent variables. Unfortunately it is not a trivial matter to isolate them when taking measurements. Changing V_d will affect both V_{sd} and V_{gd}. Secondly changing V_g affects both I_s and V_{gd}. Finally changing V_s affects both I_s and V_{sd}.

The first set of tests focused on the dependence of injection rate on V_{BIAS}. In Figure 2.7, V_{BIAS} affects both I_s and V_{gd} of T1. V_{DDI} is fixed at 3 V, and V_{INJ} is connected to ground. The dependence on V_{BIAS} is shown in Figure 2.8, where the update rate is defined as the change in V_{OUT} with respect to injection time.

If the update rate is divided by I_s, then this results in the curves shown in Figure 2.9. This effectively isolates the dependence on V_{gd} of the injection transistor. Based on (2.1) it is expected that the curves should follow an exponential of a quadratic function.

The simulation lookup table was then used to estimate the unknown values. Figure 2.10 shows the average I_{inj} as a function of V_{BIAS}. Dividing by I_s and plotting versus V_{gd} results in Figure 2.11. Both figures also show the derived model for the dependence on V_{gd} and I_s.

The next set of tests measured update rate as a function of V_{DDI}. During this measurement V_{BIAS} is set to 400 mV, and V_{INJ} is grounded. In this case varying V_{DDI} affects both I_d and V_{sd}, while V_{gd} is held constant. The update rate as a function of V_{DDI} is shown in Figure 2.12.

Dividing the update rate by I_s isolates the dependence on V_{sd}. This results in the curves shown in Figure 2.13. From (2.1) it would be expected that there should be an exponential relationship to V_{sd}. However, the slopes are surprisingly shallow.

The lookup table was again used to determine I_{inj} and V_{sd}. Figure 2.14 shows average injection current as a function of V_{DDI}. Dividing by I_s and plotting versus V_{sd} results in Figure 2.15. The two plots also contain the derived model of the dependence of I_{inj} on V_{sd}.

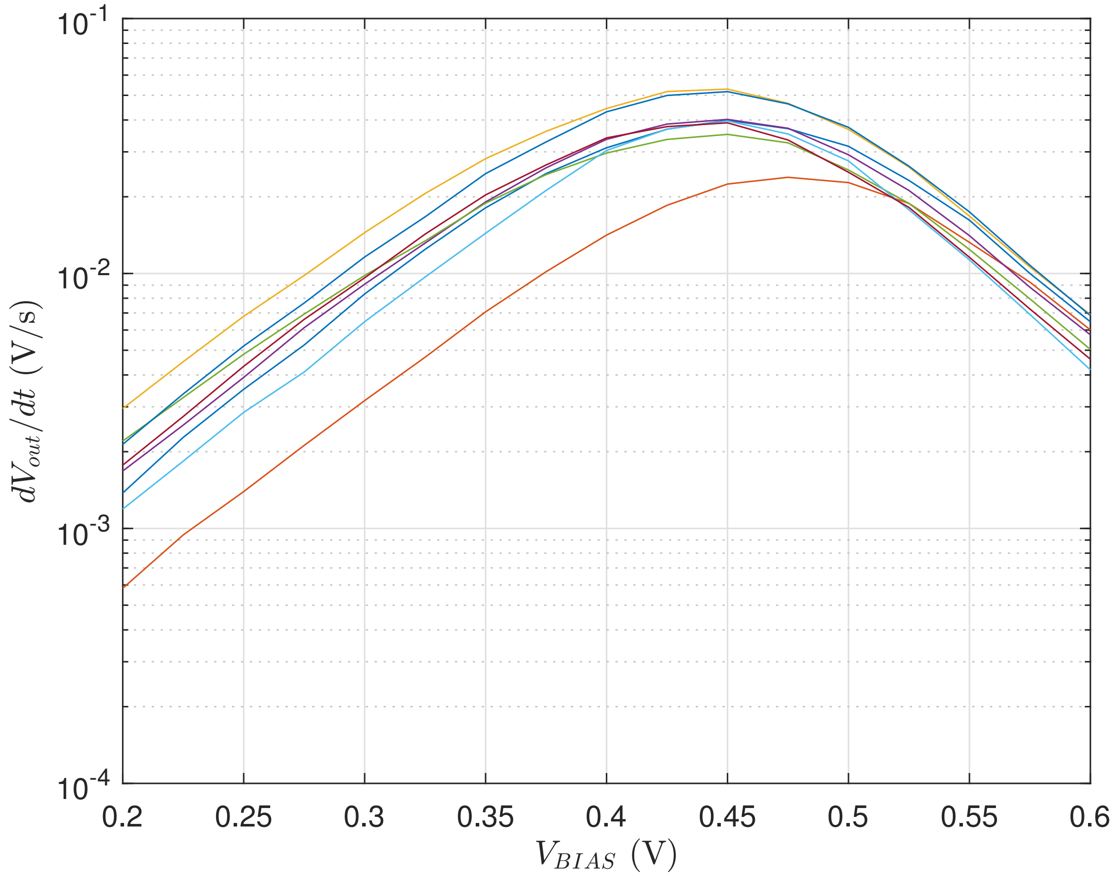

Figure 2.8: Injection update rate vs V_{BIAS}

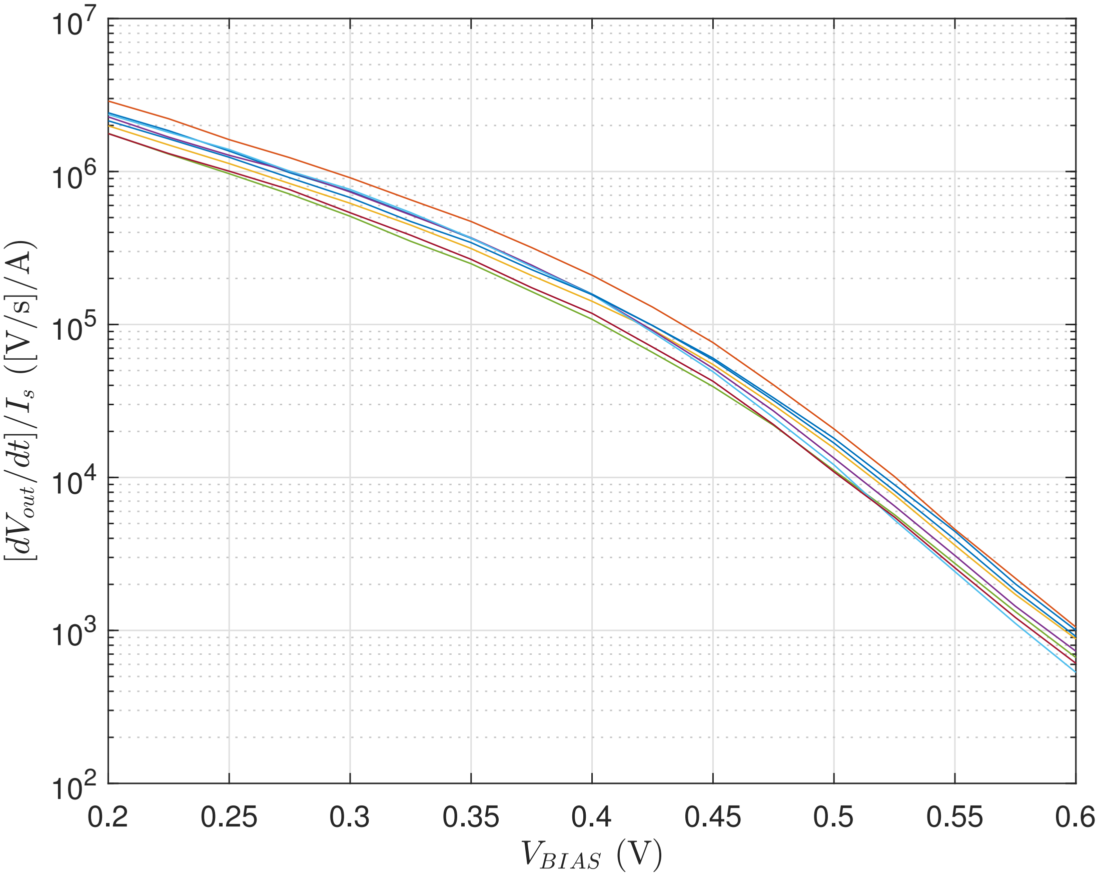

Figure 2.9: Normalized injection update rate vs V_{BIAS}

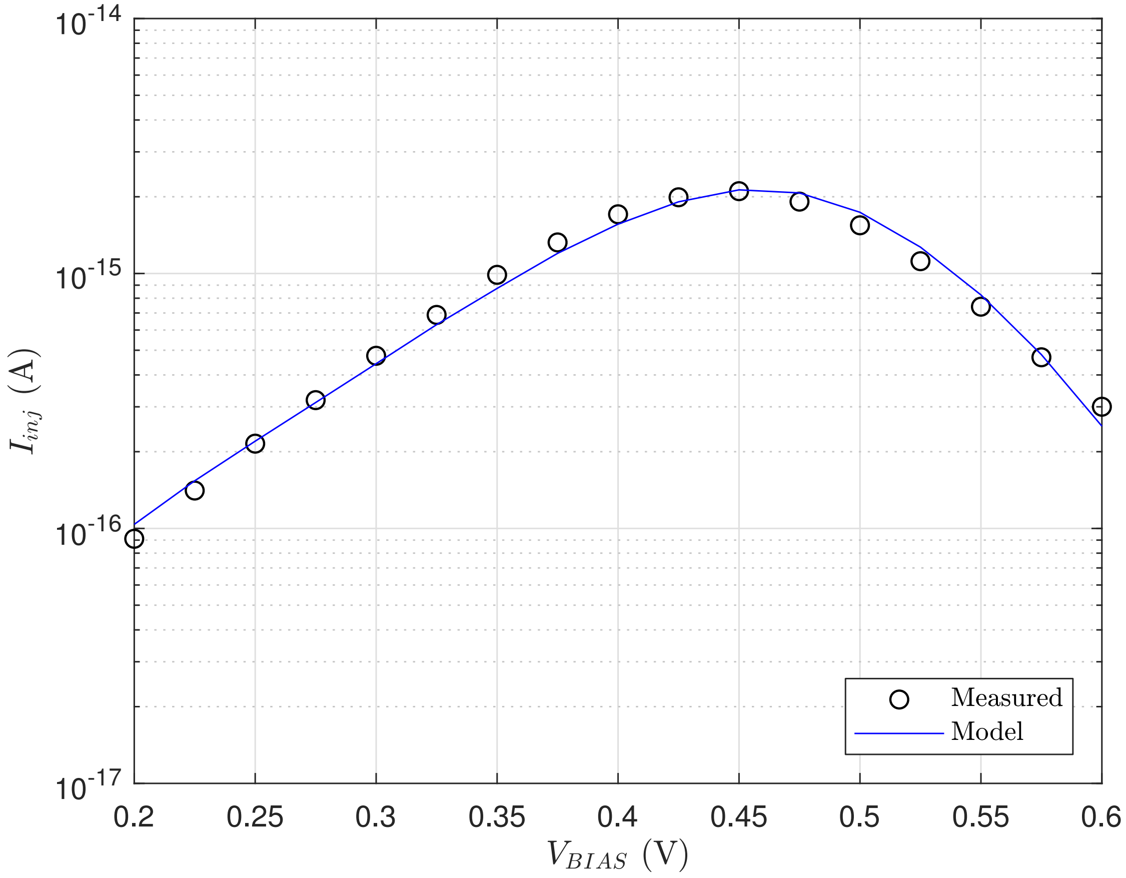

Figure 2.10: Average injection current vs V_{BIAS}

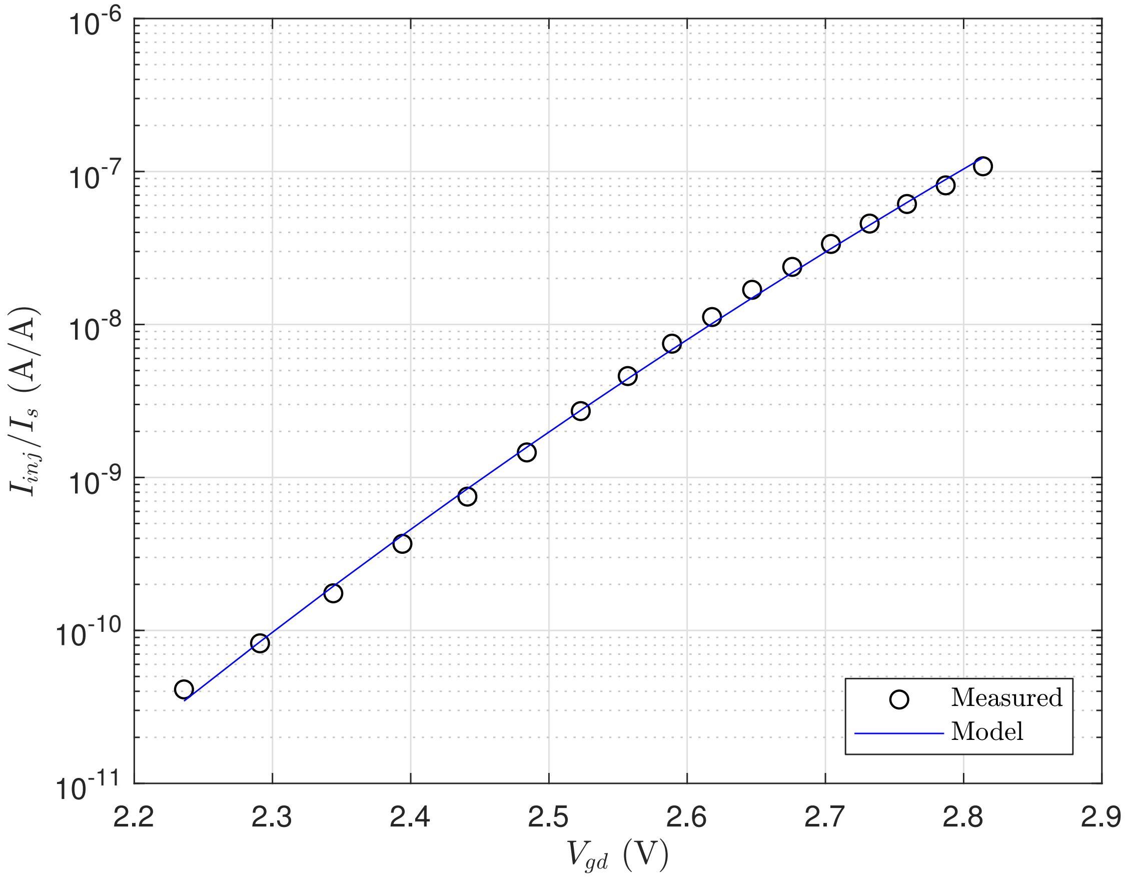

Figure 2.11: Normalized average injection current vs V_{gd}

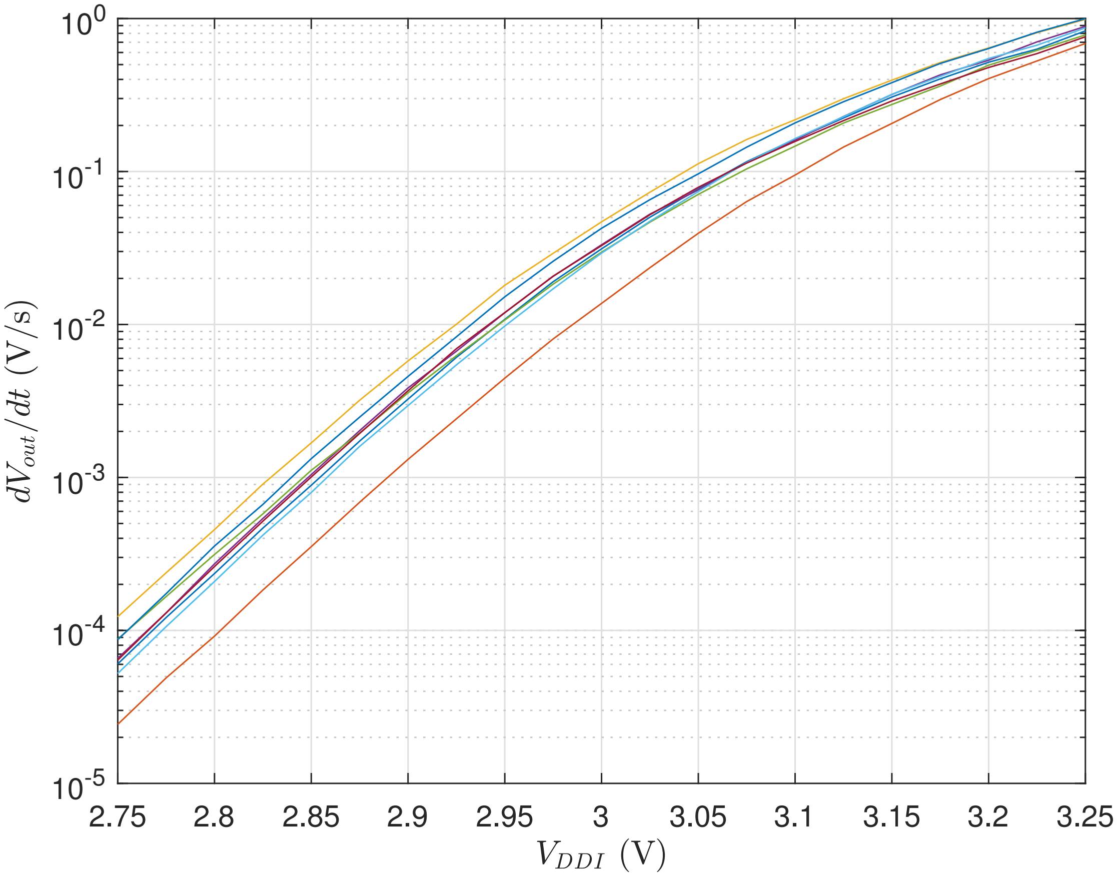

Figure 2.12: Injection update rate vs V_{DDI}

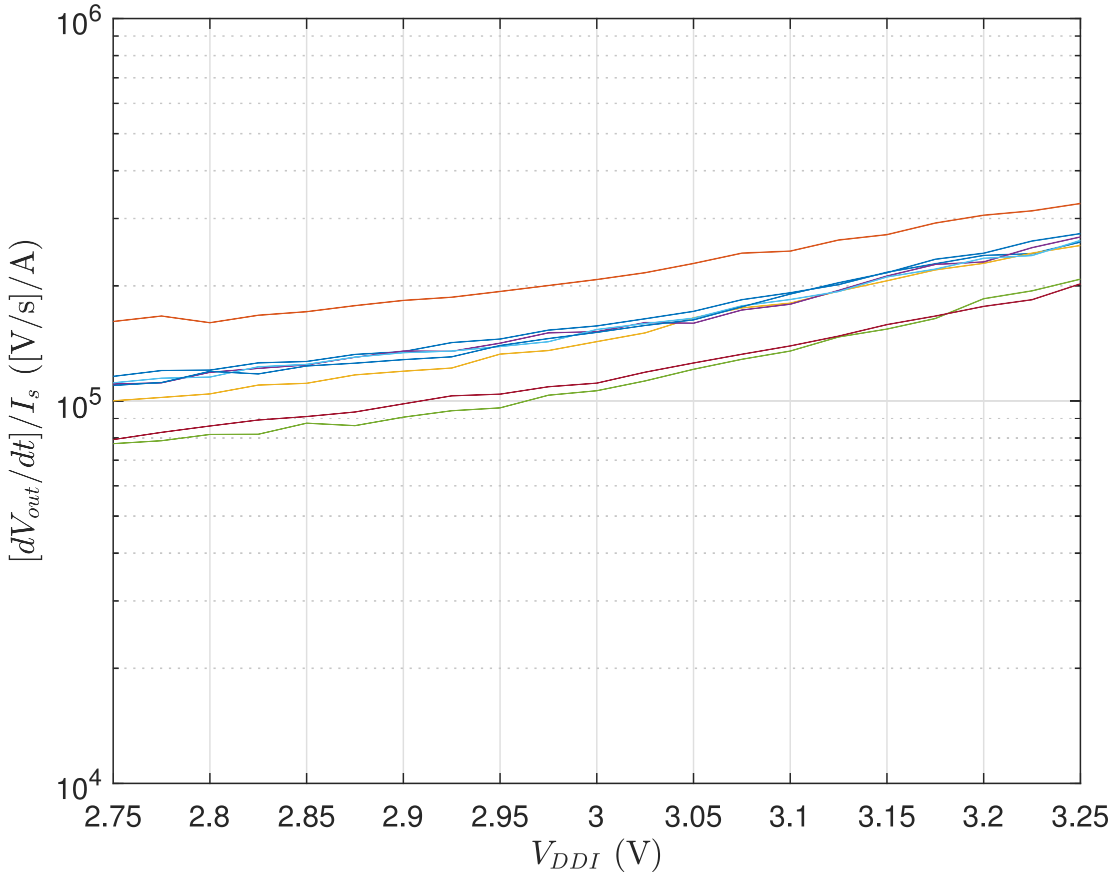

Figure 2.13: Normalized injection update rate vs V_{DDI}

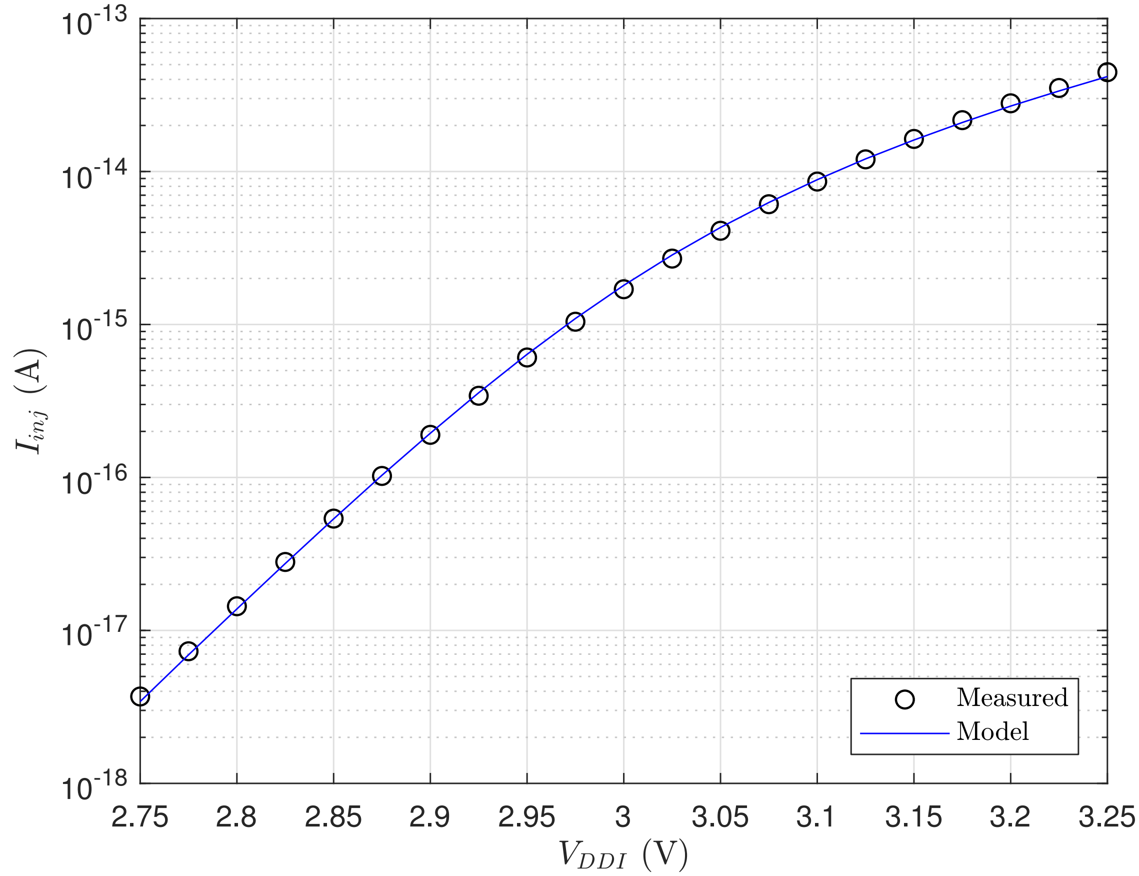

Figure 2.14: Average injection current vs V_{DDI}

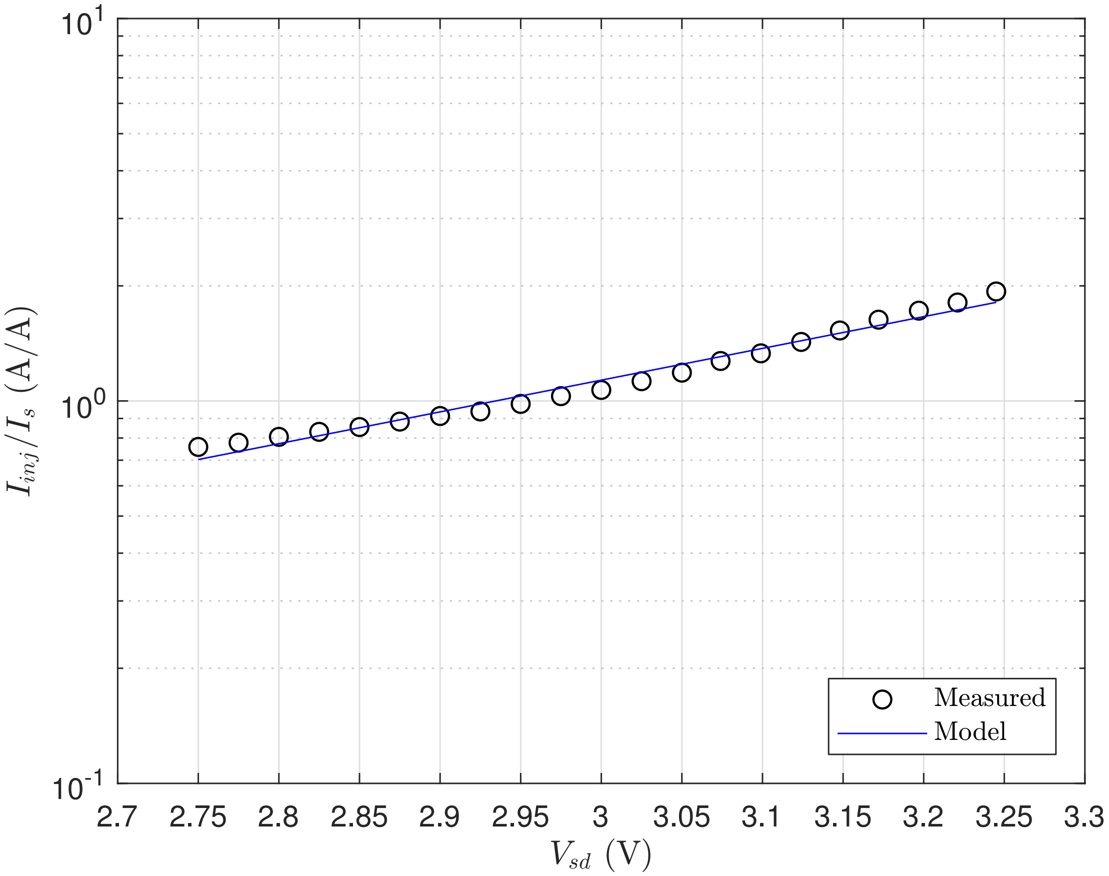

Figure 2.15: Normalized average injection current vs V_{sd}

Tunneling

There is only a single variable dependent on circuit operation in (2.7) and (2.8), so measurement of tunneling characteristics is simpler than for injection. The tunneling update rate was measured as a function of V_{TUN}. During these tests V_{DDT} is set to 1 V. The source terminal of T3 is connected to V_{DDT} for the selected cell, and all other cells remain at V_{DD}. This allows tunneling to occur only in the selected cell even though V_{TUN} is global for the array. The oxide voltage of T2 is roughly 2 V greater for the selected cell compared to the others, so isolation is possible due to the exponential dependence on V_{ox}.

The results of the V_{TUN} sweeps are shown in Figure 2.16. There is an interesting phenomenon that occurs around 6 V. Below this voltage the update rates decline less dramatically as V_{TUN} is decreased. This behavior exhibits the transition between direct and FN tunneling.

Simulations were again used to estimate I_{tun} and V_{ox} from the measurements. The average tunneling current as a function of oxide voltage is shown in Figure 2.17. Models for direct and FN tunneling are included to further illustrate the transition between them. Note that J_{tun} and V_{ox} are negative because the oxide voltage and tunneling current are defined as positive when the applied V_{gb} is positive. The measurements presented here were done with the opposite polarity. Also V_{ox} is offset from V_{gb} due to the flatband voltage.

2.2.3 Final Simulation Models

After the curves of each dependent variable had been established, it was finally possible to derive the fit constants for each equation. The final equations for IHEI, FN tunelling, and direct tunneling are shown in (2.9), (2.10), and (2.11). The injection current is in A, and the tunneling current density is in A/cm^2. The oxide voltage V_{ox} is defined in (2.12) where V_{gb} is the gate-to-body voltage.

$$I_{inj} = 1.68 \times 10^6 I_s \exp\left[\frac{-1256}{(V_{gd}+3.11)^2} + 1.91 V_{sd}\right] \tag{2.9}$$

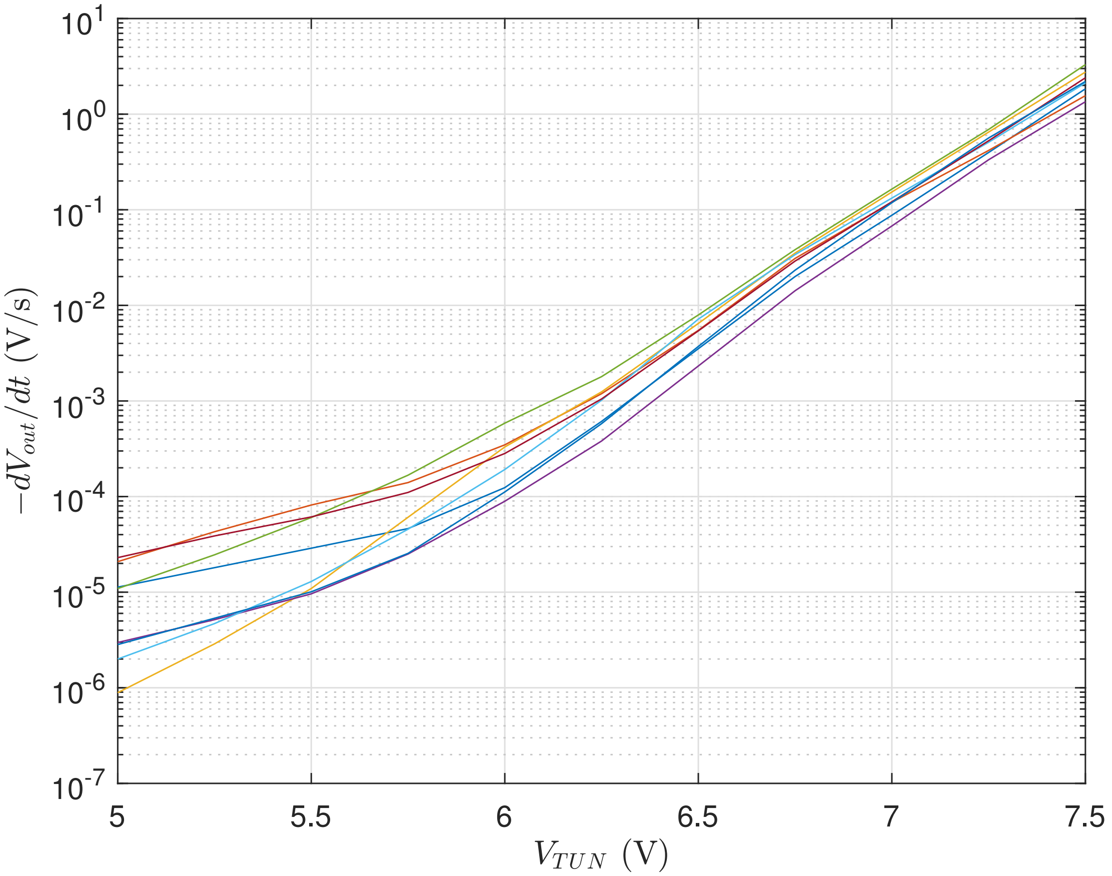

Figure 2.16: Tunneling update rate vs V_{TUN}

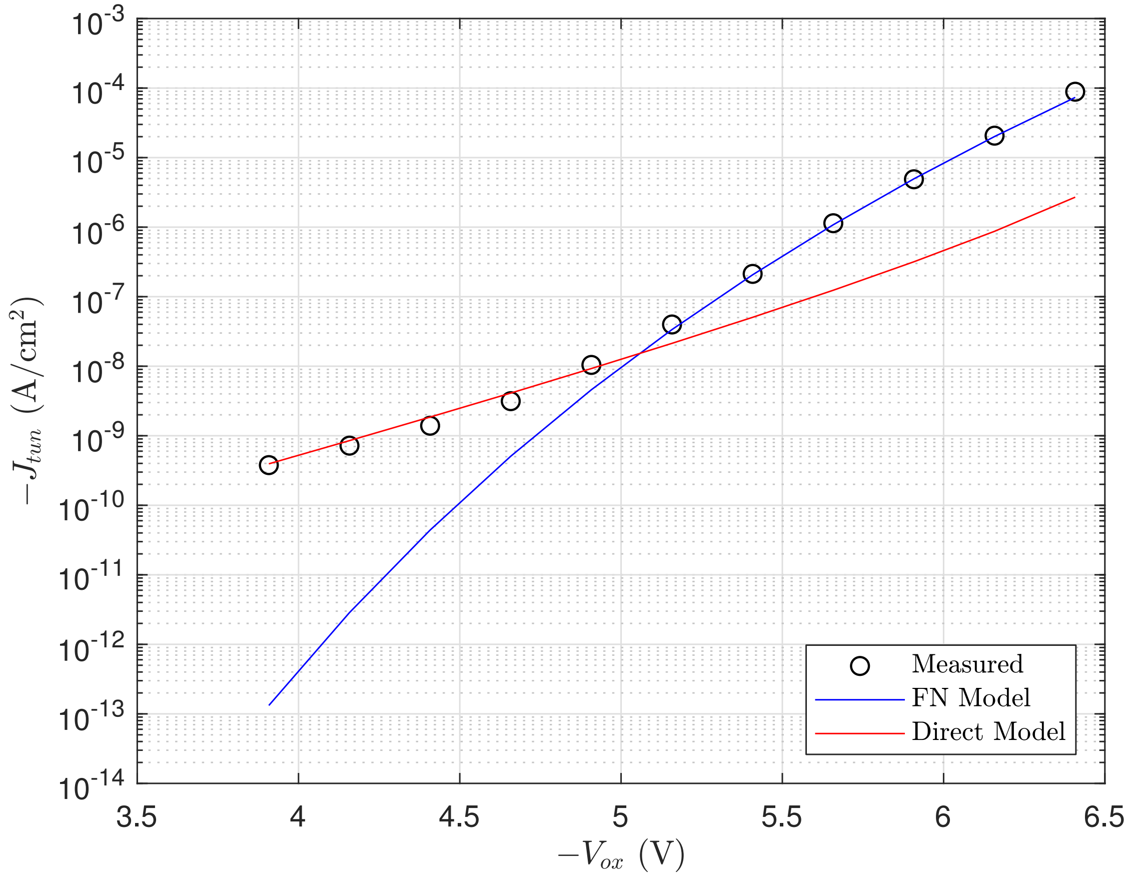

Figure 2.17: Average tunneling current density vs V_{ox}

$$J_{tun,FN} = -2.06 V_{ox}^2 \exp\left[\frac{-192}{V_{ox}}\right] \tag{2.10}$$

$$J_{tun,direct} = -1.69 V_{ox}^2 \exp\left[\frac{-212}{V_{ox}}\big(1 - (1 - 0.156 V_{ox})^{3/2}\big)\right] \tag{2.11}$$

$$V_{ox} = -V_{gb} - 0.5 \tag{2.12}$$

2.2.4 Retention Characteristics

Retention in nonvolatile memory is a serious concern for the reliability of the stored values. It is necessary to augment the program and erase characteristics with measurements of leakage. During all of the injection and tunneling measurements, it was observed that some cells' output voltages increased slowly even when no updates were being performed. The leakage was quite low even for cells with the worst retention, so the injection and tunneling measurements were not significantly affected.

Leakage Mechanisms

There are several possible mechanisms that could lead to leakage in the cell design. Since the feedback circuit is always active as long as power is supplied, it is possible that low level injection could occur in the feedback device T3 in Figure 2.7. However, V_{sd} and V_{gd} of this transistor are much lower than the voltages used in the injector T1. Additionally the length of T3 is so 8.3 times the length of T1, so the electric field through the channel is unlikely to be sufficient for the generation of hot carriers. Any leakage due to injection in T3 would be minuscule compared to the amount of leakage that was actually observed.

Another possible mechanism for leakage is due to the forming of silicide on the polysilicon floating gate. Silicide is formed to significantly reduce the sheet resistance of the poly as well as the contact resistance to metal interconnects, and its application is especially important in digital circuits to reduce rise time. However, deposition of the silicide can lead to the diffusion of metal into into the oxide. This could provide a resistive leakage path between the gate and the silicon below. Gambini and Cunningham showed that the annealing of $CoSi_2$ on As-doped polysilicon can cause metal to penetrate into the substrate of shallow-trench diodes [17]. They also noted that the resulting leakage was dependent on the thicknesses of both the poly and the Co deposits. It is reasonable to assume that the same penetration phenomenon can also occur at the interface between poly and oxide, and if it is severe enough it could provide a leakage path through the oxide barrier.

A third possible leakage path is via tunneling. There are four devices in the memory cell circuit that share the floating gate. While the circuit is in standby, the source and drain terminals of the injector are grounded. The source, drain, and body terminals of the tunneling transistor are also grounded. The voltage applied across the oxide of these two devices is around 2.6 V when V_{BIAS} is at 400 mV and V_{DD} is 3 V. As seen in the tunneling characterization tests, direct tunneling can be significant at this voltage.

Retention Characterization

In order to investigate the retention characteristics, a separate test setup was built to take measurements of V_{OUT} and ambient temperature over long time periods. A BeagleBone Black was used to acquire data from the PSoC, and this data was then sent to a local server to be stored in a CSV file. Several different measurements were made with varied starting conditions and supply voltages.

From the data obtained there were three major observations that could be made. The first was that there did not appear to be any correlation between the leakage rates and the injection and tunneling update rates. In other words cells with high leakage could have low update rates, and low leakage cells could have high update rates.

The second observation was that injection in the feedback circuit was unlikely to be the cause of the leakage. In fact turning off V_{BIAS} actually increased the leakage rate. With no bias voltage all drain currents are turned off, so there can be no injection.

The final observation was that there was an exponential dependence on supply voltage. Measurements of the leakage rates of 16 cells are shown in Figure 2.18. The exponential nature of the leakage implies that tunneling is the primary mechanism for the leakage. Note that there are discontinuities and unexpected peaks at certain voltages. These were due to large ambient temperature changes both during the test and between subsequent tests. Attempts were made to shield the test setup from temperature changes, but the circuit was still sensitive to the disturbances.

The initial design of the analog memory array did not take into consideration that retention testing would be performed, so the supply voltage could not be varied over a large range. A new design was fabricated with provisions for increased voltage range testing. The number of memory cells was doubled to 32 cells. The second half of those cells had the silicide blocked on the four devices that shared the floating gate so that the second leakage mechanism could be tested. The new set of retention tests were also performed in an environment where the ambient temperature was more stable. This led to more consistent results compared to the previous test setup.

The first retention test of the new analog memory array with V_{DD} at 3 V is shown in Figure 2.19. It is interesting to note that the cells without silicide appeared to have a higher leakage current on average. If the application of silicide was to be a significant source of leakage, it would have had the opposite effect on the results from this test. Also noteworthy are the cells whose leakage currents are significantly higher than the rest. These cells' leakage characteristics were compared to their measured tunneling characteristics, and again no correlation between high leakage current at low voltage and tunneling current at higher voltage could be found. It is possible that the dominant source of the leakage in these cells is through the injector's oxide instead of the tunneling device.

Tests at different supply voltages were then performed to observe the voltage-dependent nature of the leakage. The supply voltage was varied from 2.5 V to 3.25 V in 50 mV steps. The length of the tests ranged from six hours at higher voltages to over two days at low voltage. Across all cells there was a definite exponential relationship to the supply voltage. This is more apparent in Figure 2.21 where the average leakage current for each half of the cells is plotted.

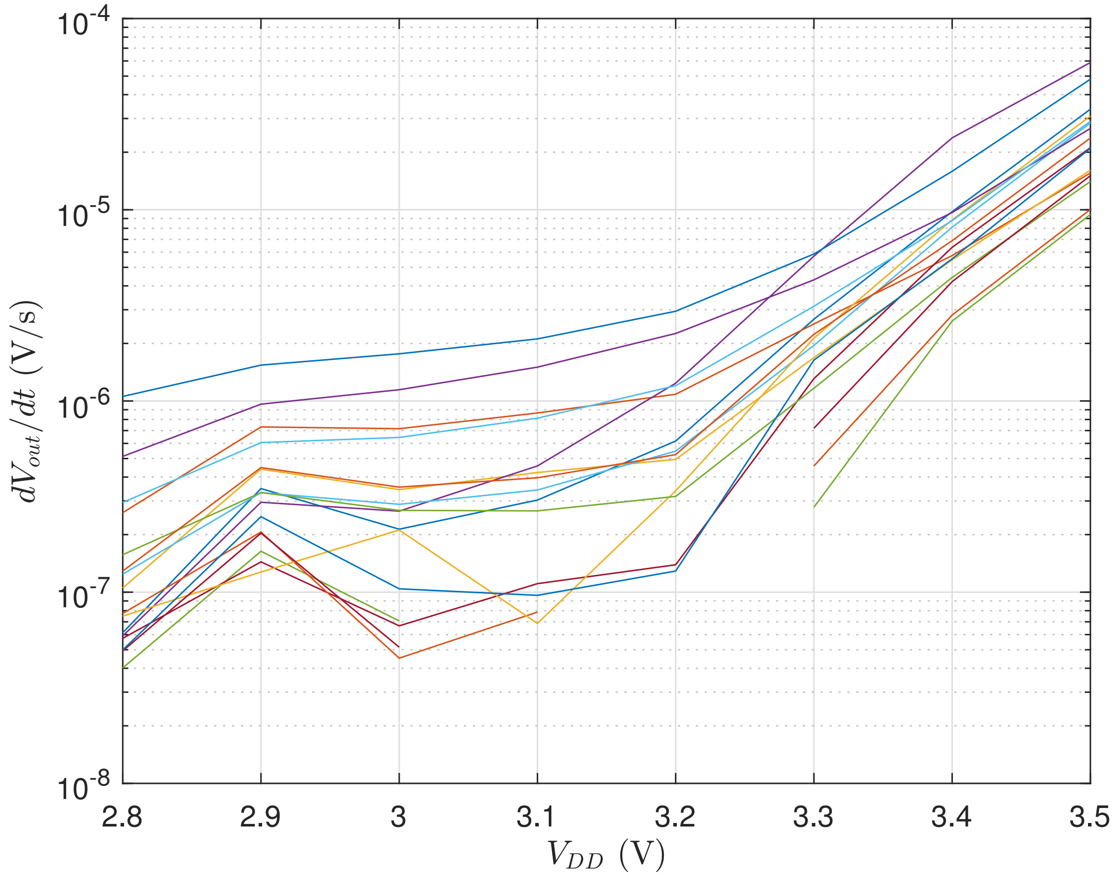

Figure 2.18: Leakage rate vs supply voltage for initial design

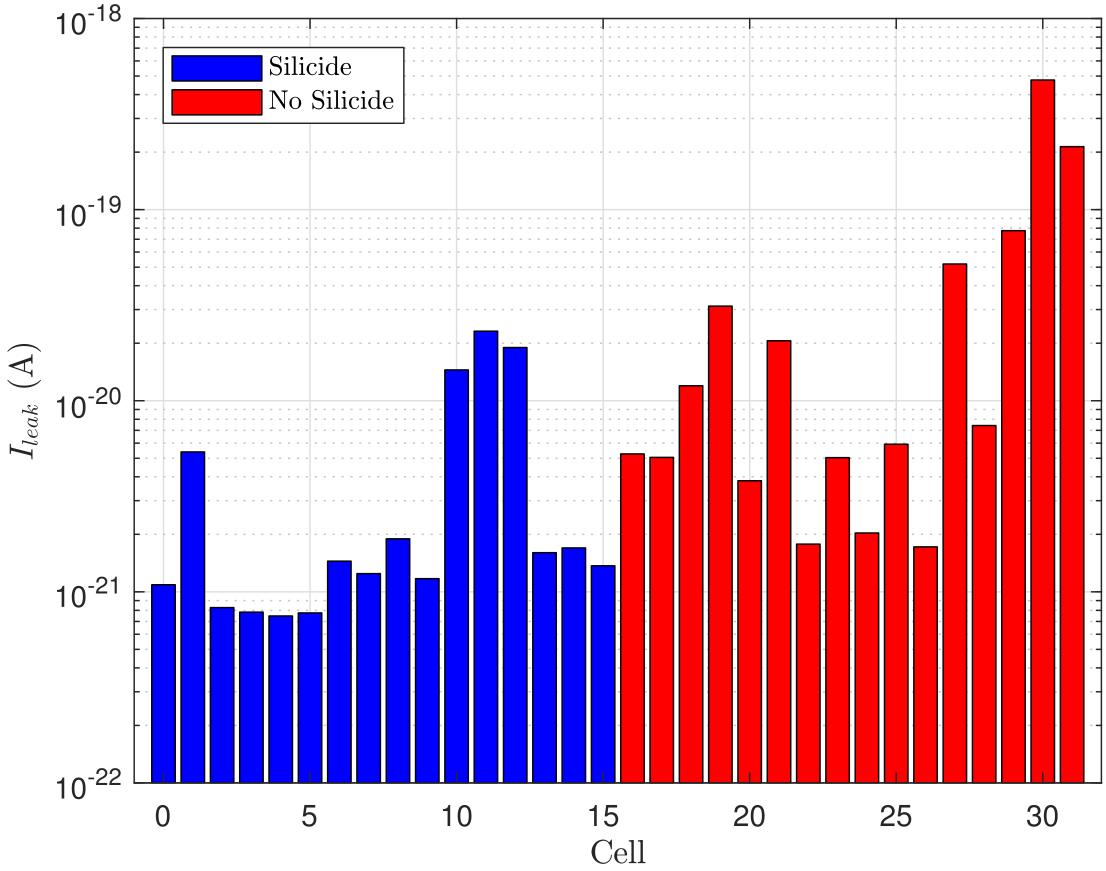

Figure 2.19: Leakage current at $V_{DD} = 3$ V

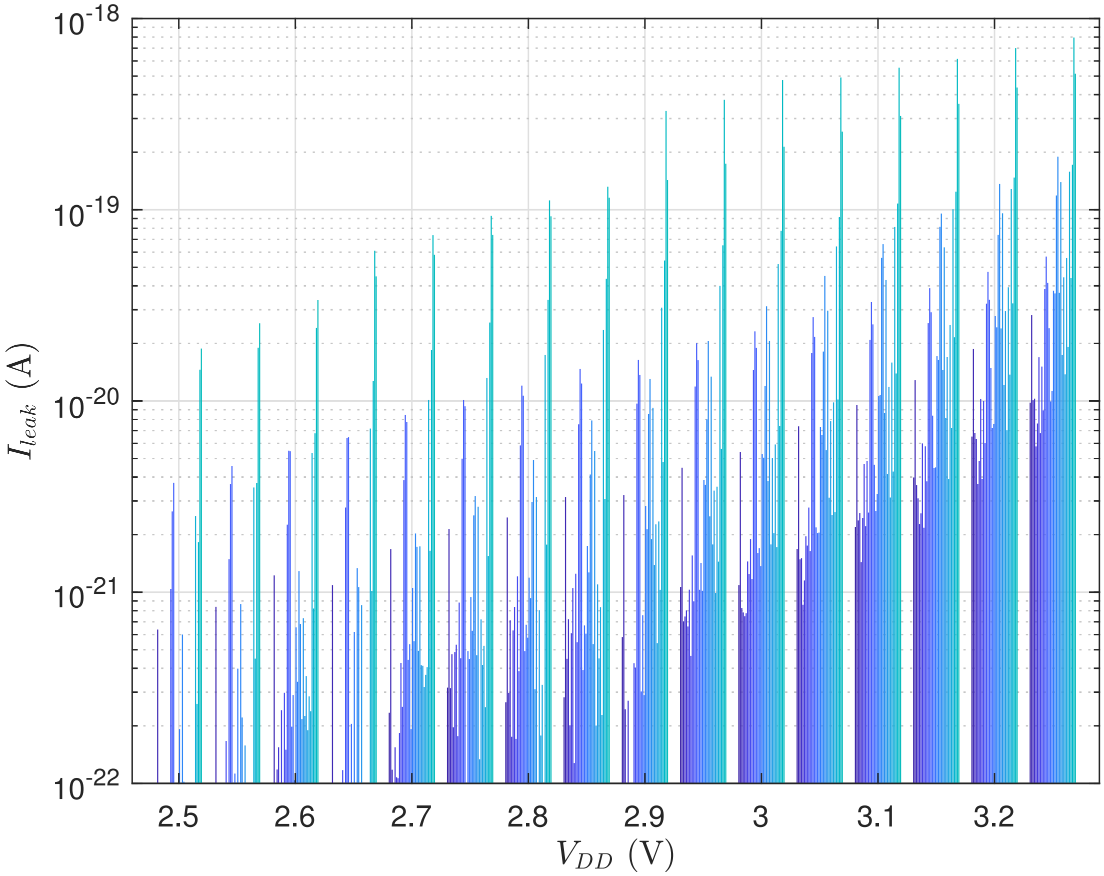

Figure 2.20: Leakage current vs supply voltage

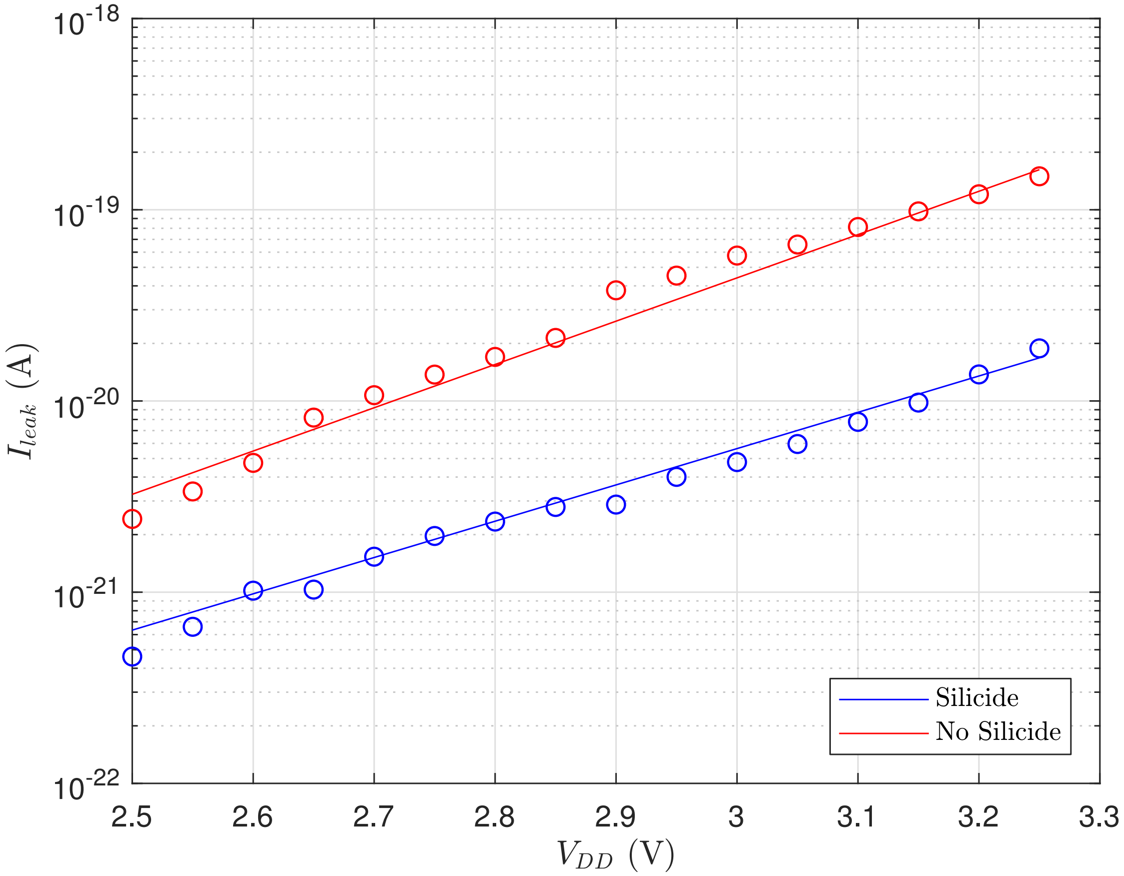

Figure 2.21: Average leakage current vs supply voltage

It appeared that an effective method to mitigate the leakage is to reduce or turn off V_{DD} when not performing measurements. In order to demonstrate this, a final retention test was run with V_{DD} at 1 V. Figure 2.22 shows the results of this test. After more than five days of runtime, there was no measurable leakage. The ambient temperature was included in the plot to show what caused the disturbance in output voltage near the 90 hour point.

Conclusions and Proposed Solution

It is important to note that the feedback circuits for all of the cells were active during the whole test. Varying the supply voltage had little effect on the bias currents and voltages within the feedback circuits, so any leakage due to injection should have been measurable at 1 V. Since there was no observable leakage over the five day period, it is clear that injection is not the source of leakage in the memory cell.

On the subject of silicide being the cause of leakage, no definitive conclusion can be drawn from these tests. The plots would seem to indicate that leakage is actually worse without silicide, but an interesting phenomenon is visible in Figure 2.22. Prior to running each retention test, all cells were programmed to the same output voltage with V_{DD} at 3 V. The supply was then set to the desired voltage, and the test was started. It was observed that the output voltage shift is different for cells with and without silicide. At 1 V this difference was quite significant. The cells without silicide shifted about 100 mV lower than the other cells. The reason for this behavior is that the addition of silicide not only affects the resistance of the polysilicon, but it causes the work function to change. Heavily doped n^+ polysilicon has the work function Φ equal to χ_{si} at 4.05 eV. Adding transition metals to silicon causes the barrier height to change. For example, according to [18] the work function of $CoSi_2$ ranges between 4.62 and 4.77 eV. This shift in barrier height affects the flatband voltage of the MOS structure, and by extension the threshold voltage and capacitance characteristics are altered. So in these tests there is likely a significant difference in floating gate voltage between the two groups of cells even though the target output voltage at 3 V supply was the same. This is readily apparent in Figure 2.21 where the two groups of cells differ by a large vertical shift and a small change in slope.

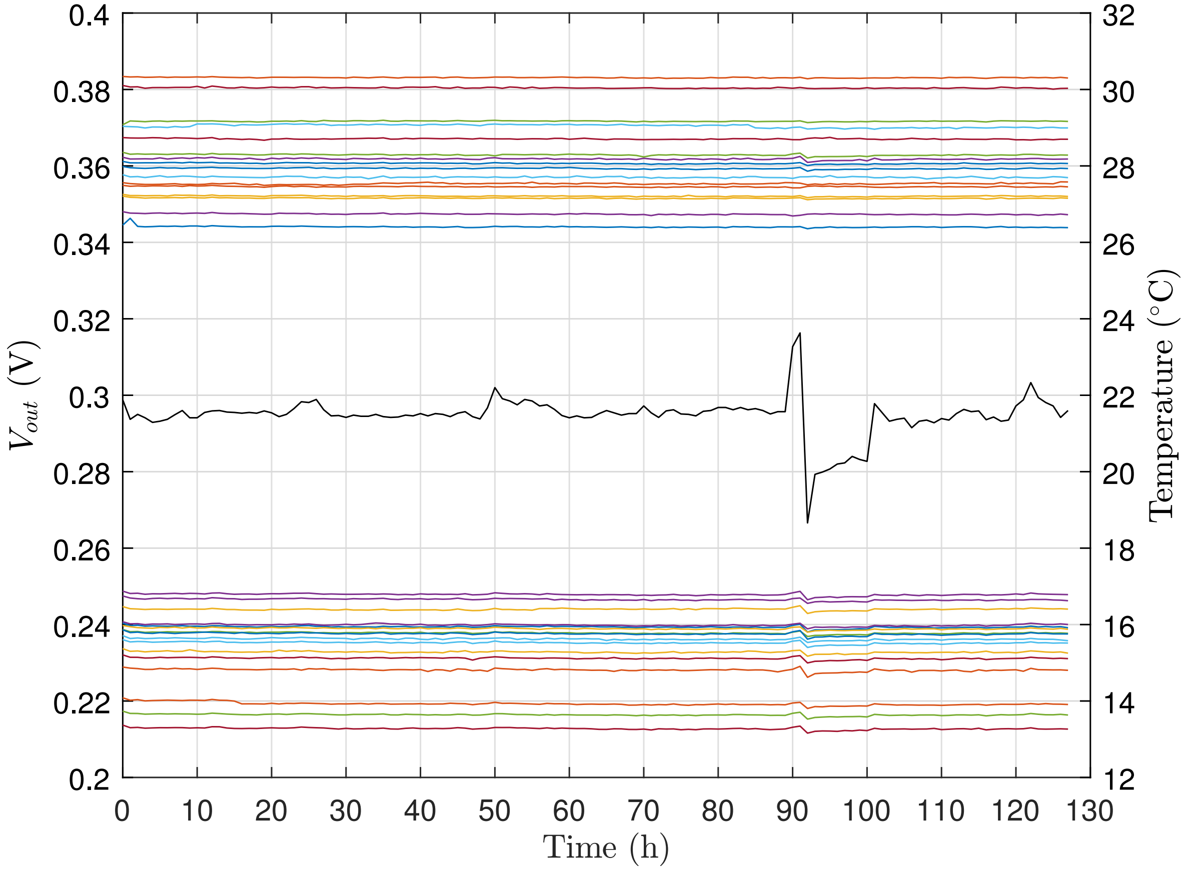

Figure 2.22: Output voltage and temperature vs time at $V_{DD} = 1$ V

The blocking of silicide to reduce leakage is common in some applications. For example CMOS image sensors benefit from reduced leakage by excluding silicide. Merrill notes in his patent on silicide blocking that few publications on imagers mention silicides and those that do specifically state that silicides should not be used [19]. A secondary consideration for blocking silicides in imagers is that the opacity reduces the available light to the sensor.

Regardless of the cause of the leakage observed in the analog memory array shown here, it is clear that the most effective method to minimize the leakage is to reduce the supply voltage. This does mean that if the output is to be used in a continuous fashion (e.g. as a bias voltage), then the difference between the output voltage in programming and operation modes needs to be accounted for. It is a trade off to consider if low leakage is to be obtained.

Chapter 3

Flash Memory System

The design of a complete flash memory system depends on several factors of the intended application including power consumption, speed, capacity, and die area. The design presented here is intended to be firmware storage for a microcontroller. Write and erase operations will be performed before deployment, so only the read operation is constrained by power requirements. For firmware storage it is required that the memory be able to fetch instructions quickly, so fast random read access is desired. The capacity of the memory does not need to be as large as for data storage, but sharing die space with the MCU still requires that the design be optimized for high density.

3.1 Existing Publications

Several single-poly memory cell designs have been previously published. Ohsaki et al. [1] proposed a cell design that consisted of two devices: an NMOS injector and a PMOS control gate. Their approach allows for multiple methods for writing and erasing. For writing, either IHEI or FN tunneling in the NMOS device is used. For erasing, FN tunneling is used in either device. The choice of writing and erasing methods determine the relative sizes of the two devices in order to achieve the desired voltage division between the gate capacitances.

Another design published by Raszka et al. [20] consists of three devices: a PMOS read transistor, a coupling capacitor, and a tunneling capacitor. FN tunneling is used for both writing and erasing the cell. For writing a high voltage is applied to the coupling cap while the tunneling cap is grounded and the PMOS is held at an intermediate voltage. The tunneling cap is small in comparison to the coupling cap, so the bulk of the high voltage develops across the tunneling cap. Electrons tunnel onto the floating gate causing the stored charge to decrease. For erasing the coupling cap is grounded while the tunneling cap has a high voltage applied. Again the majority of the voltage appears across the tunneling cap, though it is of the opposite polarity compared to the writing operation. This allows holes to tunnel onto the floating gate, thus increasing the stored charge.

A third design proposed by Wang et al. [21] uses a core cell with three PMOS devices. Like [20] it also uses FN tunneling for both writing an erasing. The devices are sized such that tunneling occurs across one of the smaller devices, and the direction is determined by which terminals have high voltage applied. Two of the core cells are combined with four additional transistors to form a complete storage cell.

Unfortunately the designs discussed above require significant additional circuitry on-die. This includes high-voltage switches and charge pumps to provide the high-voltage supply. In the case of [21], an additional intermediate voltage at half the high-voltage supply is required. The design in [1] requires that additional devices be used to block the high voltages from being applied to sensitive low-voltage devices in the peripheral circuitry.

The design presented here attempts to alleviate some of these issues. No on-die high-voltage switches, blocking devices, or supply generators are required. The core cell design is also minimal and compact to allow for dense arrays to be fabricated. Table 3.1 shows comparisons of overall dimensions between this design and two of those discussed above. Both [20] and [21] suffer from significant area penalties due to the additional circuitry to handle high voltage generation and switching. The differential topologies and additional devices within the complete memory cells also increase the area of each bit. Only the total system area for 1 kb was provided for [20], so the other designs were scaled up or down in order to compare the areas. The presented design is 94% smaller than [20] and 69% smaller than [21] for a 1024-bit system.

Table 3.1: System area comparison

Parameter	**This Design**	**Raszka** [20]	**Wang** [21]
Technology (nm)	130	130	130
Devices per core cell	4	3	3
Devices per complete cell	4	26	10
Rows	64	128	32
Columns	32	16	16
Total bits	2048	2048	512
Bit area (μm^2)	22.1	N/A	72
Array area for 1 kb (mm^2)	0.023	N/A	0.14
System area for 1 kb (mm^2)	0.051	0.9	0.165

3.2 System Overview

The array is a 2048-bit system arranged in 64 rows and 32 columns. The word length is 8 bits giving 256 bytes total storage. Bytes are individually addressable and writable, and the erase is performed globally. The schematic of the system is shown in Figure 3.1. When fabricated in a 130 nm process, the dimensions of the total system are 530 × 195 µm as seen in Figure 3.2. A summary of specifications is provided in Table 3.2.

3.3 Memory Cell

To overcome the control limitations of single-polysilicon devices available in standard CMOS processes, a feedback circuit was included to maintain floating gate voltage during programming. The cell design is based on an analog memory system provided by Lu [9]. In order to maximize array density, all control logic has been separated and placed in the row and column decoders.

The memory cell contains three PMOS transistors and one capacitor. The schematic is shown in Figure 3.3, and device specifications are provided in Table 3.3. T1 and C1 provide feedback during programming. T2 is used to erase the cell. T3 is used for both writing to and reading from the cell. The layout shown in Figure 3.4 is designed for mirrored abutment in both columns and rows for maximum density. The dimensions are 2.88 × 7.68 µm.

Terminals CL and SL are connected by row, CG and BL are shared per column, and GL is global to the array. The voltages applied to the terminals in each of the three operation modes are shown in Table 3.4.

In read mode CL, CG, and GL are grounded. BL is discharged to ground prior to a read operation. When the row is selected, 1 V is applied to SL and T3 will begin charging the bitline. The rate at which the voltage increases depends on the drain current set by V_{fg}. At the end of the read cycle, if the bitline voltage has increased sufficiently compared to a reference then it will be interpreted as a logic 1.

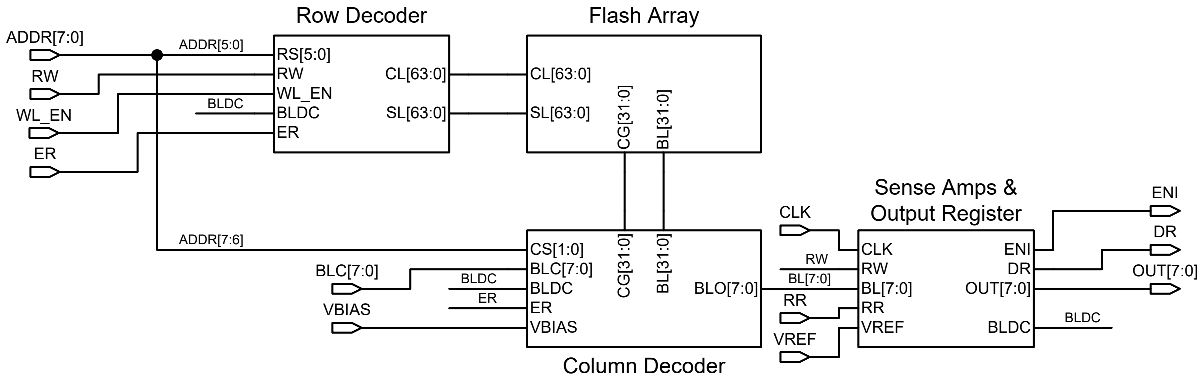

Figure 3.1: Schematic of complete flash memory system

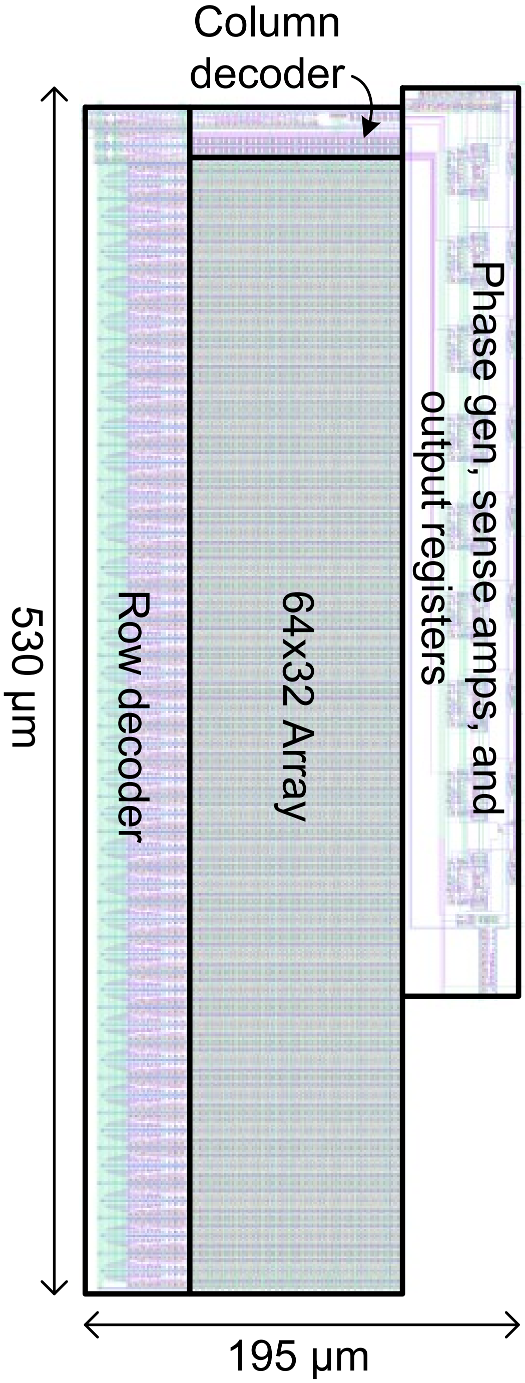

Figure 3.2: Layout of complete flash memory system

Table 3.2: System specification summary

Parameter	**Value**	**Unit**
Technology	130	nm
System dimensions	530 × 195	μm
Array size	64 × 32	bits
Word size	8	bits
Cell dimensions	2.88 × 7.68	μm
Read voltage	1	V
Write method	IHEI (byte)	-
Write voltage	3	V
Erase method	FN (global)	-
Erase voltage	7	V

Table 3.3: Flash cell device specifications

Device	**Type**	**W/L (μ m)**	t_{ox} **(Å)**	**Value**
T1	PMOS	0.36/0.24	52	-
T2	PMOS	0.36/0.24	52	-
T3	PMOS	0.36/0.24	52	-
C1	dgncap	1.16/1.56	52	10.7 fF @ 3.3 V

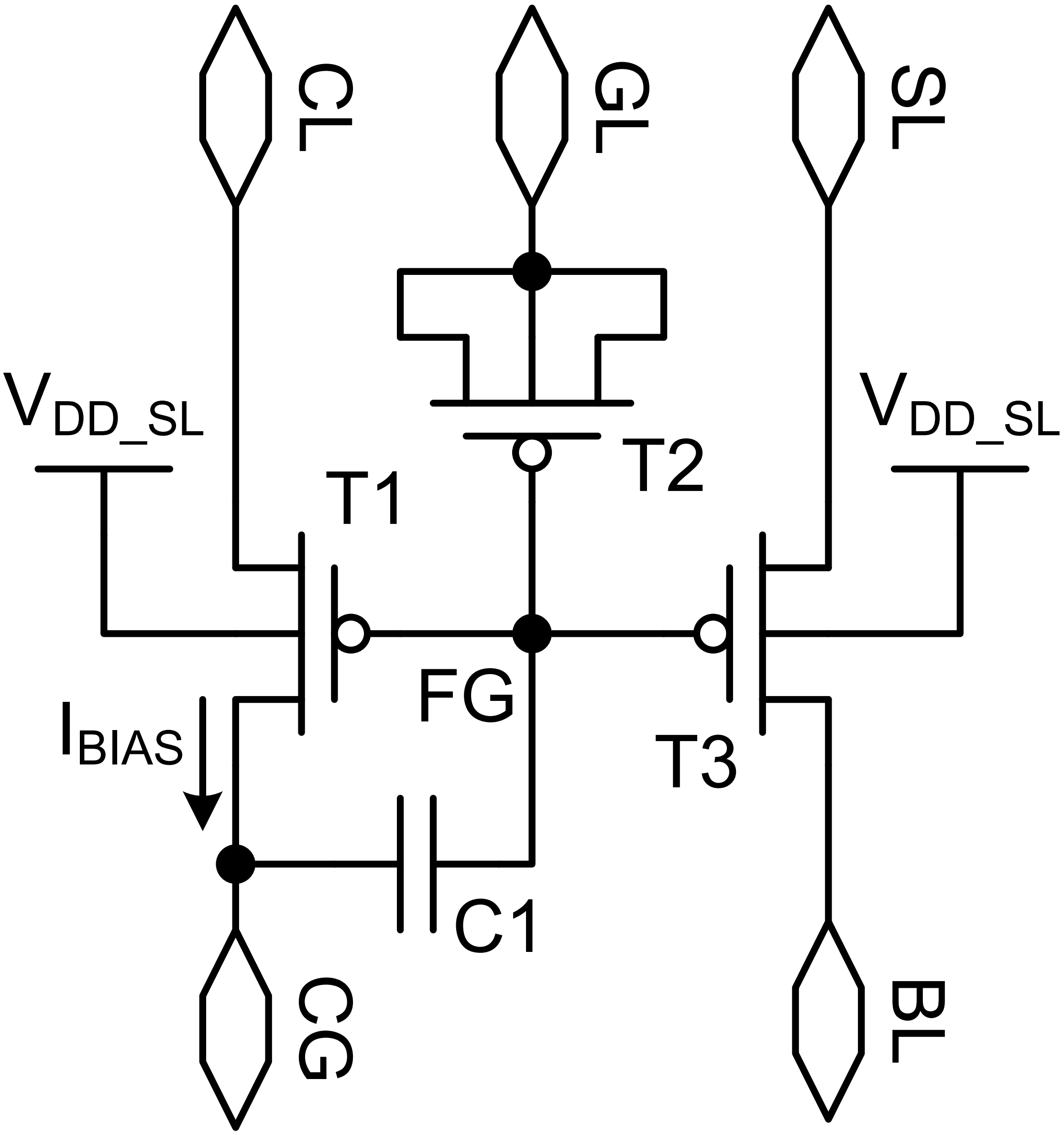

Figure 3.3: Schematic of flash cell

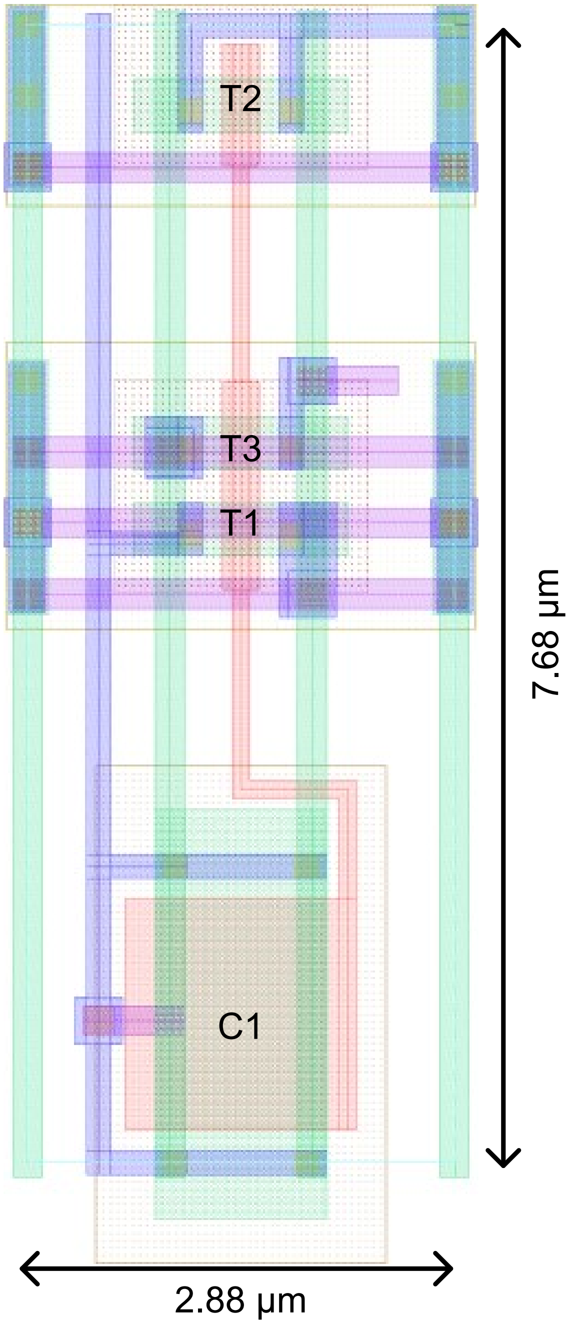

Figure 3.4: Layout of flash cell

Table 3.4: Flash cell operation

Terminal	**Read**	**Write**	**Erase**
CL	0 V	3 V	0 V
CG	0 V	$\mathrm{I_{BIAS}}$	0 V
SL	1 V	3 V	0 V
BL	High Z	0 V	0 V
GL	0 V	0 V	7 V

In write mode CL is connected to 3 V. If the column is selected, then CG is connected to an NMOS current sink to set I_{BIAS}. Capacitive feedback through C1 holds V_{fg} relatively constant. When the row is selected 3 V is applied to SL. If the cell is intended to be written to, then BL is grounded and I_{BIAS} is mirrored onto T3. Injecting electrons onto the floating gate will decrease the voltage across C1, which in turn increases the bitline drive current during read operations.

In erase mode all terminals except GL are grounded. The gate capacitance of T2 is small compared to the total floating gate capacitance. Applying 7 V to GL results in a large electric field across T2's oxide, so tunneling may occur. Holes tunnel from the channel of T2 to the floating gate causing the voltage across C1 to increase.

3.4 Row Decoder

The row decoder is responsible for multiple duties. Address lines 0 through 5 are decoded to select the desired row. During read operations the Word Line Enable (WL_EN) signal sets the source of the memory cells' injection transistors high to begin charging the bitlines. In write mode the row decoder turns on the feedback transistors. Setting WL_EN high causes current to flow in the injection transistors. In erase mode the row decoder grounds all all row connections of the memory array.

To allow scaling for larger memory array dimensions, the row decoder is split into predecoders that feed into an array of decoder cells. The current design uses three 2-to-4 predecoders shown in Figure 3.5 to decode six address lines to 64 rows.

The row decoder cell is shown in Figure 3.6. Inputs A through D are used to select the desired row from the predecoders' outputs. In this design input D is tied to V_{DD_SL} since there are only three predecoders. One more predecoder could be added to allow up to 256 rows.

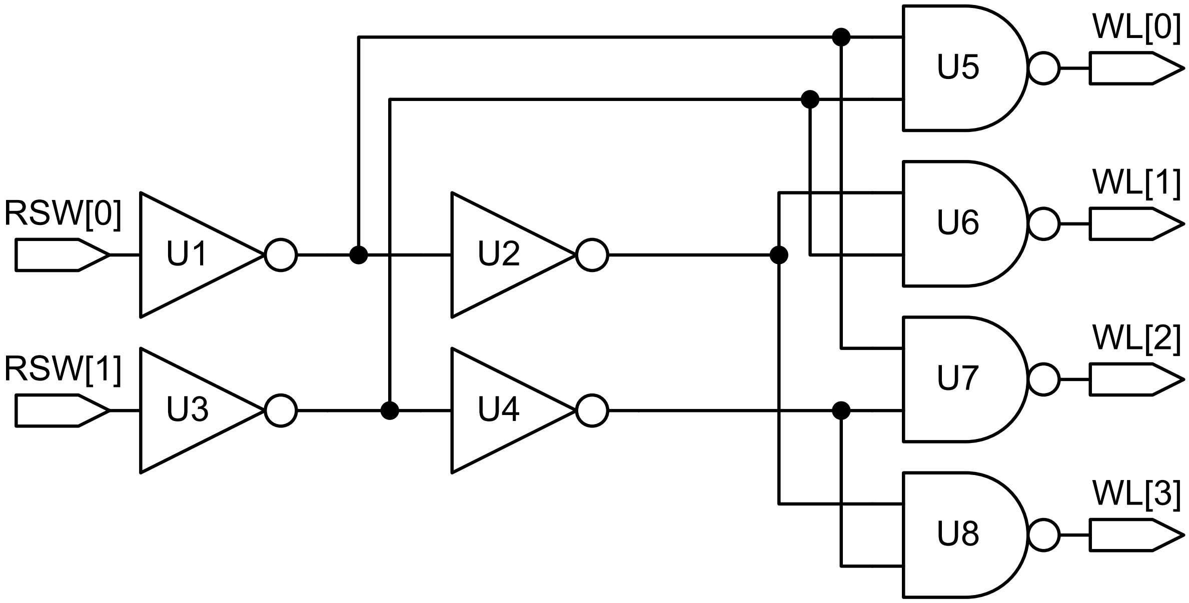

Figure 3.5: Row 2-to-4 predecoder schematic

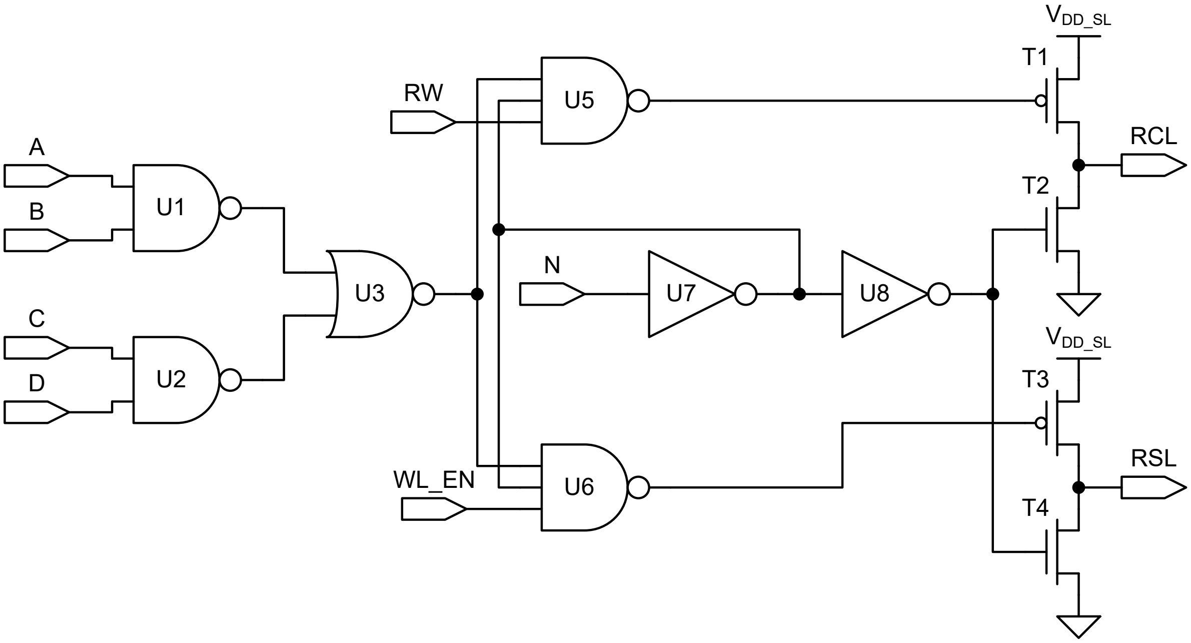

Figure 3.6: Row decoder cell schematic

3.5 Column Decoder

The column decoder is also multi-purpose. Address lines 6 and 7 are decoded to select the desired column. During read operations it is responsible for discharging the bitlines prior to energizing the array. It also sends the bitline outputs to the appropriate sense amplifiers. In write mode the column decoder sets the bias current of the feedback transistors though the CG lines. During write operation it determines which bits are written to by grounding the desired BL lines. In erase mode it connects all column lines to ground.

3.6 Sense Amplifier and Phase Generator

In order to convert the analog voltages of the bitlines to digital data, sense amplifiers are used. This circuit is essentially a comparator, and in the case of the system presented here, the sense amp compares the bitline voltage to a reference. If the voltage is above the reference, then the output is a logic 1. The sense amp in this design is clocked, so a phase generator is used to control the operation.

3.6.1 Phase Generator

The phase generator is comprised of two cascaded flip-flops whose outputs are decoded into each of the four phases using logic gates. The schematic is shown in Figure 3.7. In order to minimize skew between phases, the output of the first flip-flop is delayed using a string of inverters.

3.6.2 Sense Amplifier

Process variations lead to offsets in threshold voltages of differential comparators used in sense amplifiers. If it is desired to compare small differences in voltages, then these offsets can lead to comparison errors. In order to mitigate these errors, an automatic scheme for offset compensation may be used.

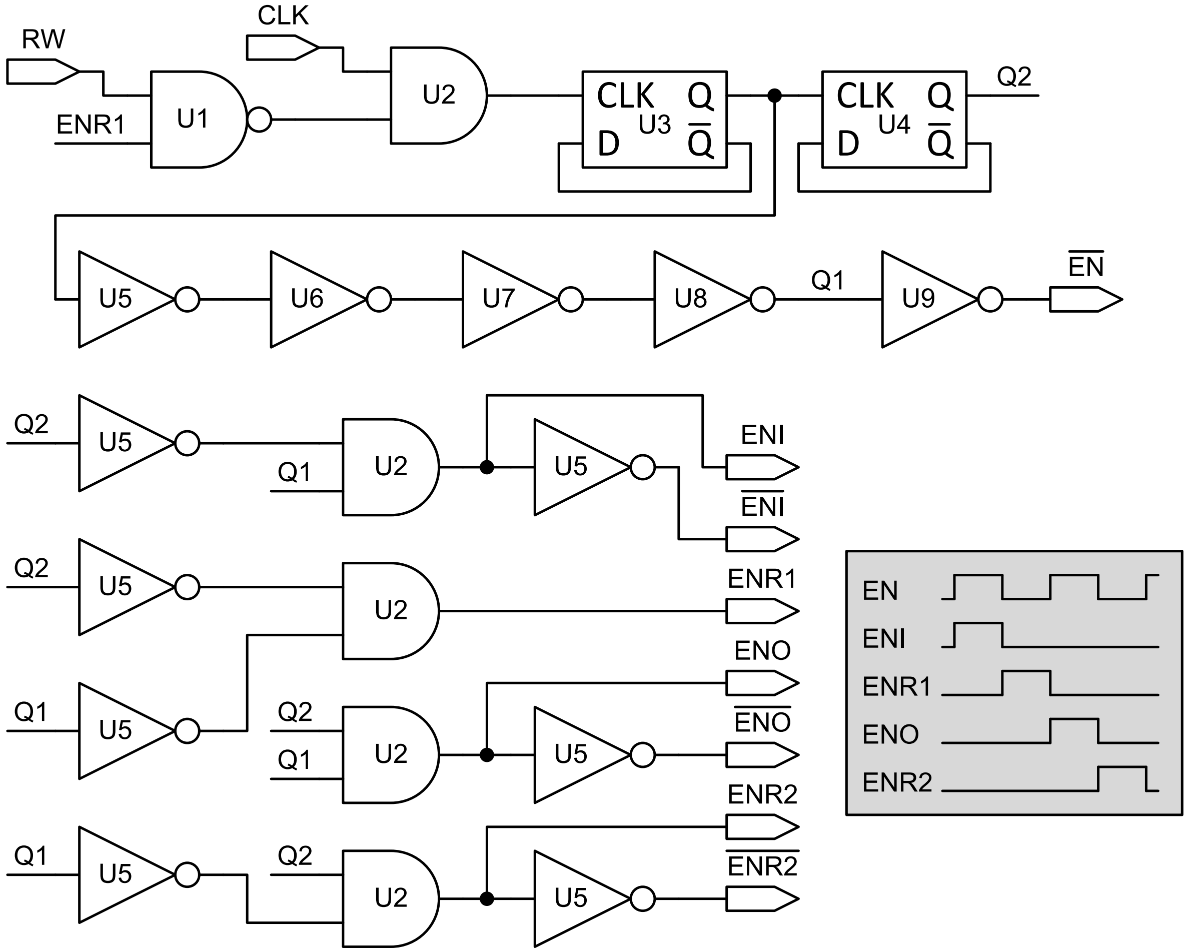

Figure 3.7: Schematic of phase generator

In [22] a Digital Auto-Zeroing (DAZ) scheme is proposed. The sense amplifier is a PMOS differential pair with cross coupled loads to provide positive feedback for fast latching. A charge pump adjusts voltage on a capacitor which increases or decreases additional current drive to one side of the sense amplifier through a pair of PMOS devices. This approach is functional, but the additional devices on one side of the amplifier present an unbalanced capacitive load.

Another approach is presented in [23]. Instead of adding devices in parallel with the sense amplifier, compensation is applied to the body terminals of the differential pair to directly adjust the threshold voltages. The sense amplifier itself is a two-stage design, and the phase detector and charge pump are separate entities.

The design presented here combines the two approaches described above. The single-stage sense amp core from [22] was used together with the bulk-driven scheme from [23]. A pair of charge pumps from [22] provide complementary phase detection and offset correction. The schematic for the fully differential design is shown in Figure 3.8. The charge pumps are shown in Figure 3.9. One charge pump has OUT as its input and BN as its output, and the other is connected to $\overline{\text{OUT}}$ and BP.

The phase diagram in Figure 3.7 illustrates the timed operation of the sense amplifier and charge pumps. During ENI the sense amplifier's inputs are connected to the bitlines, and the tail of the differential pair is activated. Positive feedback provided by the cross-coupled load causes the outputs to latch quickly. In ENR1 the bitlines are disconnected, and the inputs and outputs of the sense amp are reset to ground. In ENO phase the differential pair is activated again, and the outputs are sampled by the charge pumps. When entering ENR2 the charge stored at the output of the pump's inverter will switch on either T7 or T8 depending on the polarity of the sampled input. The charge stored in the selected transitor is pumped into C1, and its voltage will either increase or decrease. This voltage is applied to the body terminal of the respective device in the differential pair to adjust its threshold voltage.

Settling time of the DAZ is dependent on the capacitor values. A value of 200 fF was chosen for the design. Simulations have shown that the DAZ compensation is able to reduce

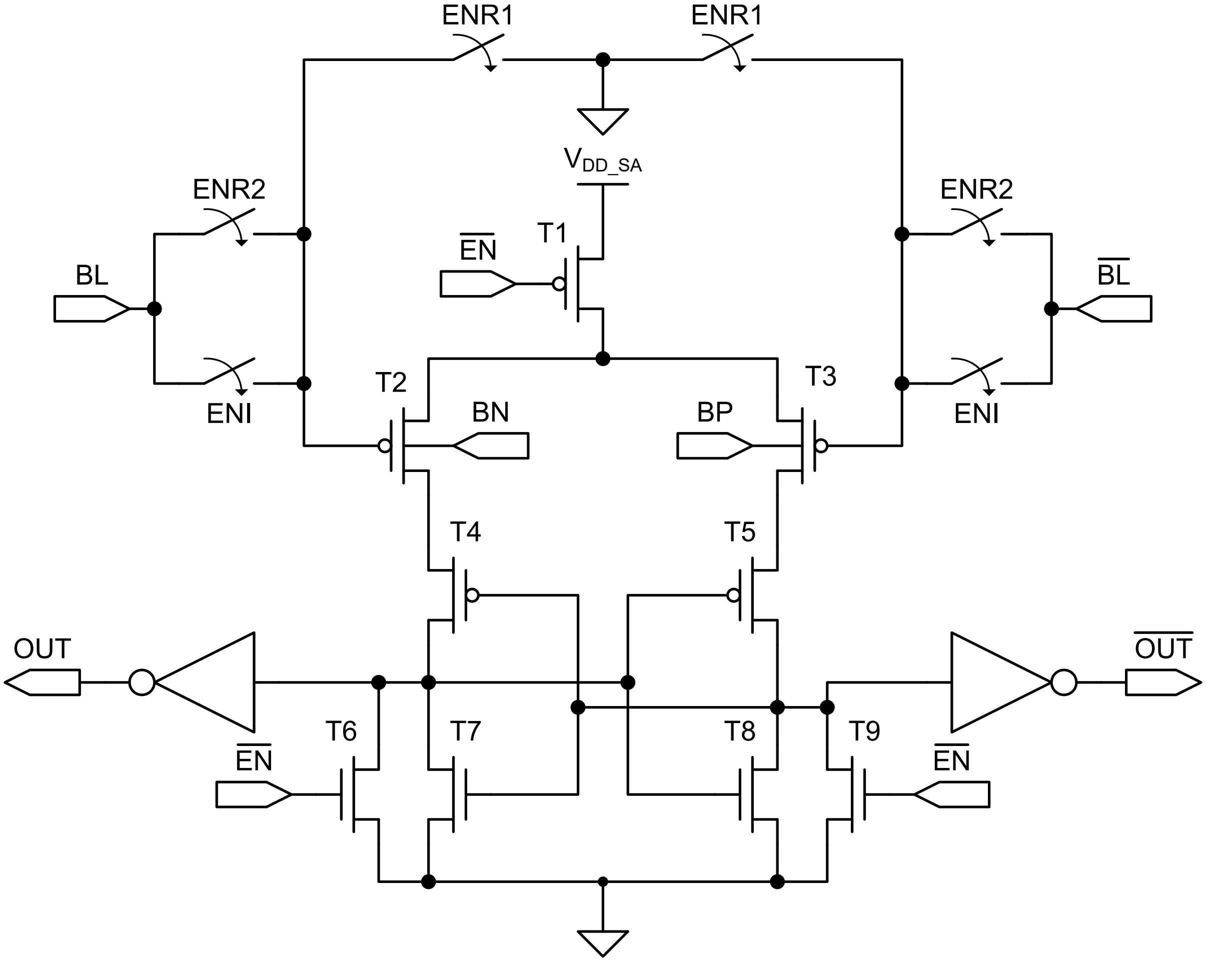

Figure 3.8: Schematic of sense amplifier core

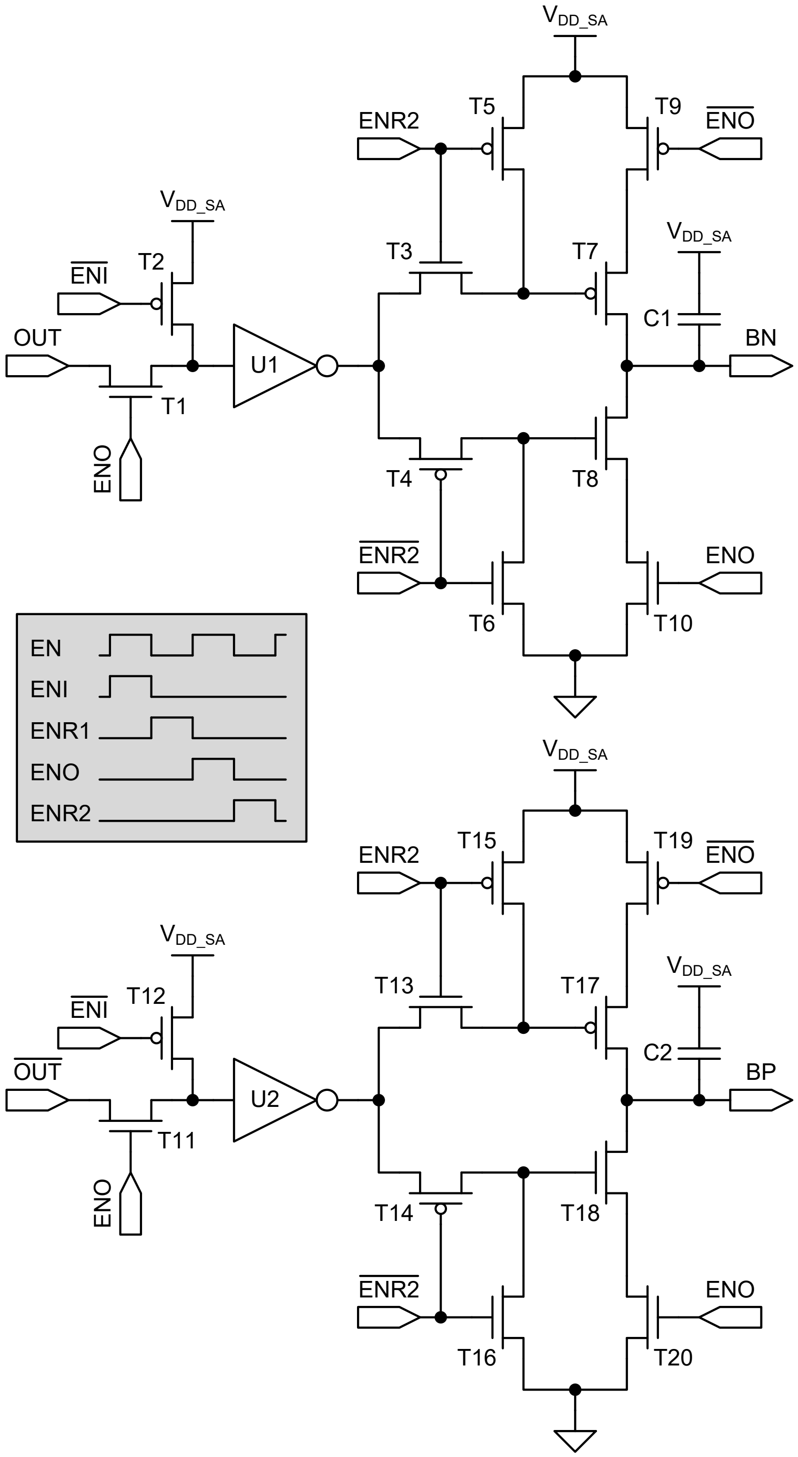

Figure 3.9: Schematic of charge pumps

a 10 mV offset to under 1 mV with a settling time of 13 µs at a 4 MHz clock. A plot of the simulation is shown in Figure 3.10. Power consumption in this case is only 50 nW per sense amplifier. An additional simulation was run at 1 MHz, and its plot is displayed in Figure 3.11. In this case the settling time was about 60 µs which is roughly four times that of the previous simulation. Note that the voltage step size on the capacitors remains constant. This indicates that the charge pump adds or removes a fixed amount of charge per period, and the number of pulses required to compensate for the offset is independent of clock frequency.

The reference for the sense amplifiers is selectable between internal and external. When set to external, the voltage applied to V_{REF_EXT} is sent to the negative inputs of the sense amps. This voltage will need to be adjusted for different clock frequencies since the bitline voltages reached at the end of a read operation depend on the WL_EN pulse width.

If the reference is set to internal, then a separate individual flash cell is used as the reference. This cell has its own sense amp for programming and erasing separately from the array. There are two advantages to using the internal reference. First, there is one less external voltage source needed. The second and more significant advantage is that no adjustments are necessary when changing clock frequencies. If the reference cell's V_{fg} is written to a higher voltage than that of a programmed array cell, then its bitline voltage will always rise at a slower rate than the array cell. One only needs to program or erase the array cells significant lower or higher than the reference to achieve desired logic readout.

3.7 Output Register and Timer

The outputs of the sense amplifiers are only valid during ENI, so latches are used to hold the value until the data has been read by the mirocontroller.

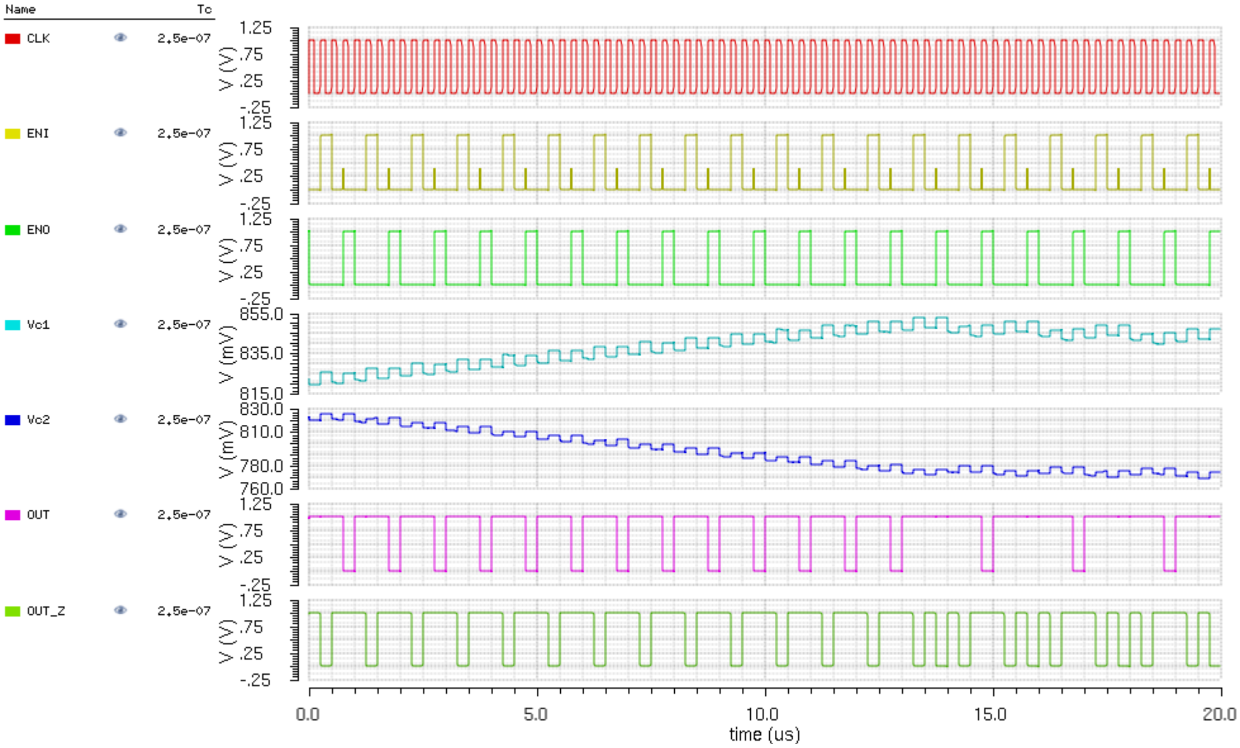

Figure 3.10: Offset cancellation at 4 MHz

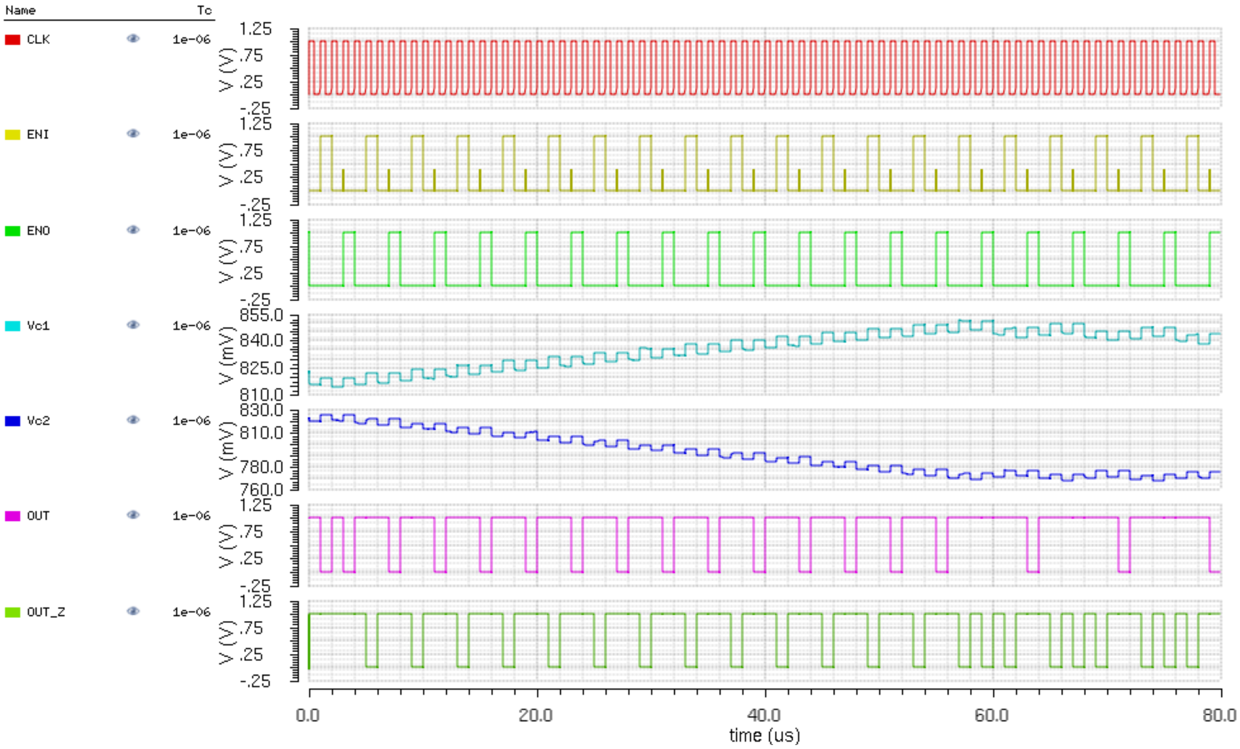

Figure 3.11: Offset cancellation at 1 MHz

3.7.1 Register Timer

It would be a trivial matter to simply latch the data every read cycle regardless of whether or not the previous data was read by the microcontroller. One cannot assume that the microcontroller will have read the previous data in time. This could be due to execution of multi-cycle instructions or simply operating at different clock speeds. For this reason a partially asynchronous scheme is used to initiate and complete the read cycle. The latch timer shown in Figure 3.12 is used to control the timing of the latch operation.

The microcontroller initiates the read operation by setting the Read Request line (RR) high. If this occurs before entering ENR2, then the output will latch the data at the end of ENI. The Data Ready (DR) signal will be set high at a half clock cycle later to avoid setup time issues. The microcontroller receives the data and sets RR low. It can then set RR high again to initiate a new read cycle.

3.7.2 Output Register

The schematic of the output register is shown in Figure 3.13. The output register core is a flip-flop. Surrounding logic ensures that the latching only occurs once while RR is held high.

3.8 Operation Modes

There are three modes of operation in the flash memory design. Read and write modes allow access to individual bytes. Erase is performed globally.

3.8.1 Read

In read mode both V_{DD_SA} and V_{DD_SL} are at 1 V. Since the sense amplifiers run on a four-phase cycle, the clock frequency must be at least four times the desired read rate.

The timing diagram for the read mode is shown in Figure 3.14. First the address (ADDR) should be set at the beginning of the ENO phase. This ensures that the row and column decoders are set before reading from the array. The ENO phase is also when the bitlines

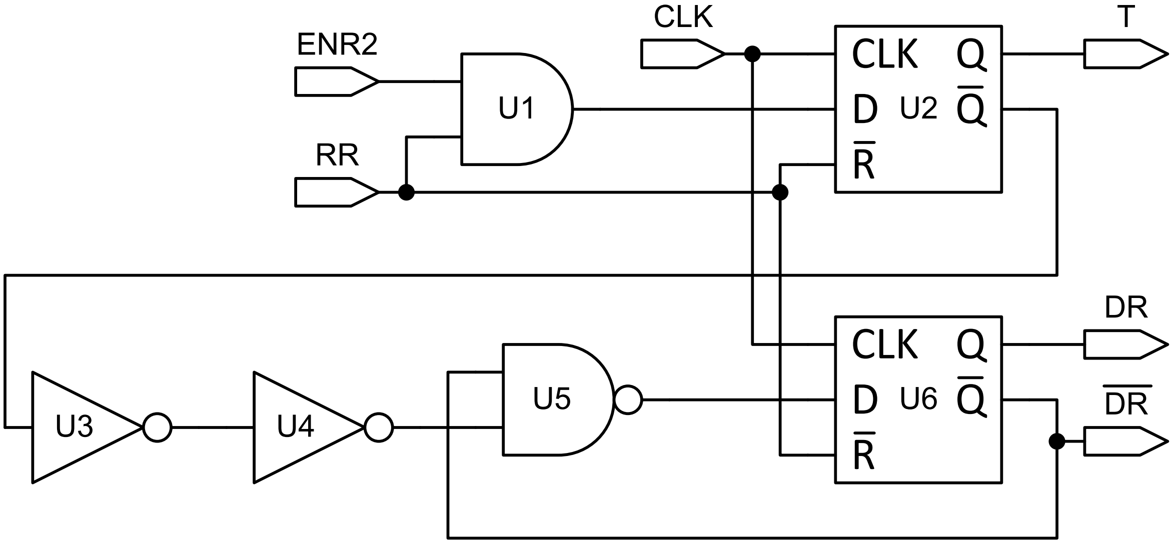

Figure 3.12: Schematic of register timer

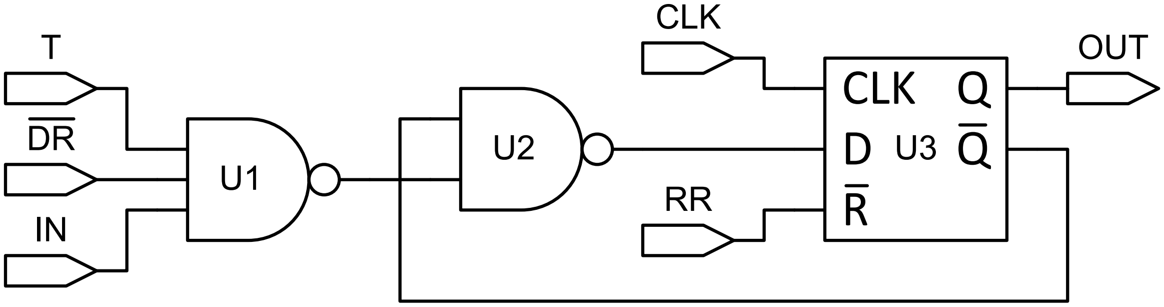

Figure 3.13: Schematic of output register

are discharged by the BLDC signal. During ENR2 the wordline enable (WL_EN) signal is applied to turn on the selected row of the array. If the read request (RR) line is set high before the ENI phase, then the output register latching sequence will be initiated. When ENI is reached, the sense amplifiers will latch the result of the comparison between bitline and reference voltages. The output registers will then store the outputs of the sense amplifiers, and the data will be made available at the beginning of ENR1. The data ready signal (DR) will be set high on the falling edge of the clock. It is delayed to allow the outputs to settle and to avoid setup time issues. DR will remain high until the falling edge of RR. The data outputs reset to logic 0 when RR is set low.

3.8.2 Write

In write mode V_{DD_SL} is increased to 3 V and V_{DD_SA} remains at 1 V. Setting the read-write signal (RW) high causes the clock to be stopped during ENO phase. This is to keep the higher voltages from being presented to the inputs of the sense amplifiers. With RW high the CL line of the selected row is turned on, and the feedback circuits of the byte selected by row and column are activated. Data to be written is applied to the bitline control inputs (BLC) which provide strong paths to ground on the bitlines for logic 1 inputs. The WL_EN line is then pulsed causing drain current to flow in the injectors for those cells whose bitlines are grounded by the BLC inputs.

3.8.3 Erase

During erase operations, all array lines except the global GL line are grounded. The clock is stopped and the two supply voltages are as they were in write mode. To erase the array 7 V is applied to the GL line for the desired erase duration.

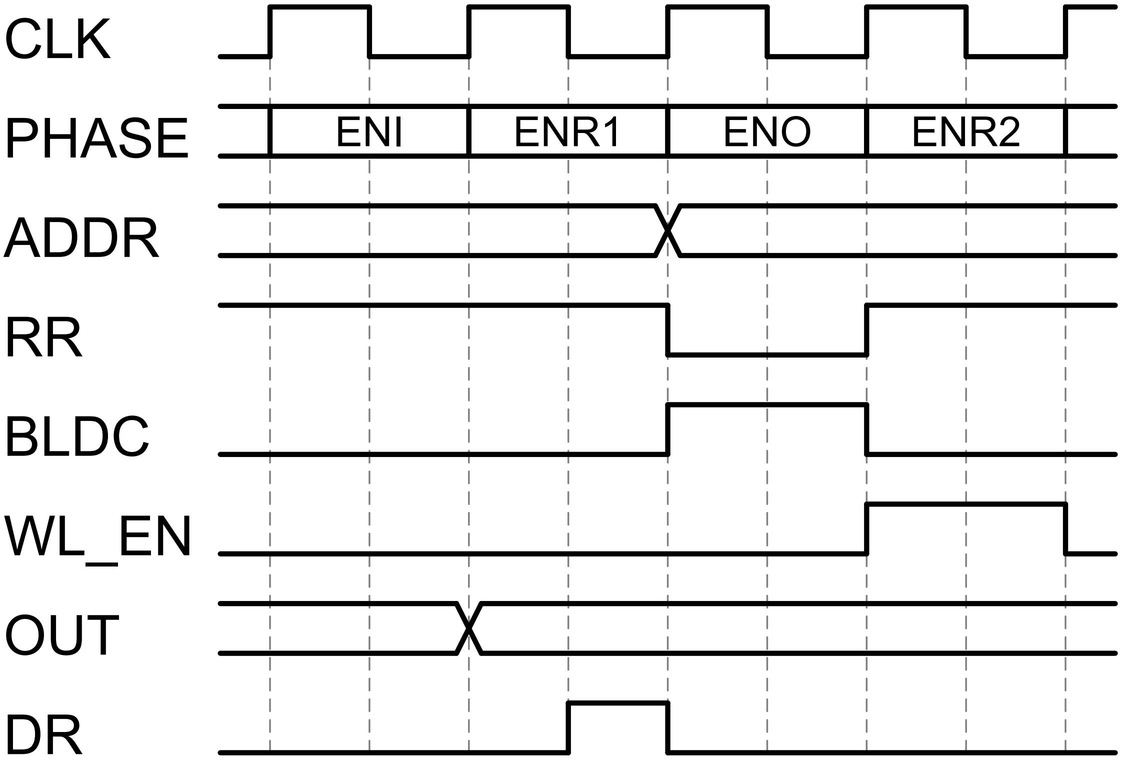

Figure 3.14: Timing diagram of read operation

Chapter 4

Experimental Results

The flash memory design was fabricated in a standard 130 nm CMOS process. A Cypress PSoC 5LP development board was used to test the read, write, and erase functions of the memory. Read performance has been measured at clock speeds from 125 kHz to 4 MHz, and the supply voltage was varied from 0.65 to 1 V. At 4 MHz clock the data throughput was 8 Mb/s. Limitations in the existing test measurement setup prevented reliable testing at higher clock speeds. Write and erase performance were then evaluated, and distributions of the time taken to increase and decrease the target V_{BL} were recorded. Power consumption during all modes is shown in Table 4.1.

4.1 Read Performance

Read access times are dependent on clock frequency because the latching of the output register is synchronous with the ENI phase. The address is changed when entering ENO, and the DR signal is asserted at the falling edge of the clock in ENR1. At 1 MHz this translates to a read access time of 3.5 µs. Given that the reads may be performed at up to one quarter of the clock frequency, this leads to 86.8 pJ per operation. An oscilloscope capture of the memory during read operation is shown in Figure 4.1.

The read performance was then measured as a function of clock frequency. The cells were programmed to a target V_{BL} of 600 mV at 1 MHz. The distributions of V_{BL} are shown

Table 4.1: Power consumption in μW

Supply	**Read**[1]	**Write**[2]	**Erase**
V_{DD_SL}	1.42	12.1	0.31
V_{DD_SA}	23.4	0.03	0.03
Total	24.8	12.1	0.34

[1] at $f_{CLK} = 1$ MHz

[2] at $V_{BIAS} = 475$ mV

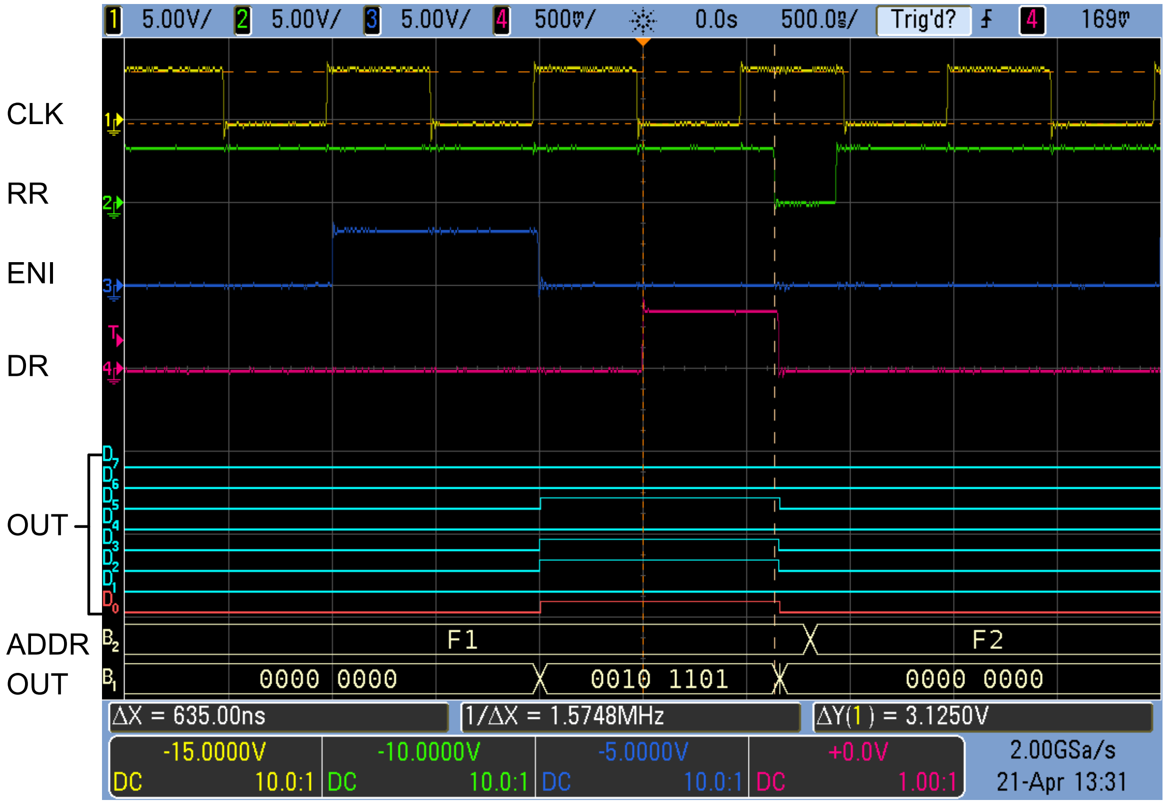

Figure 4.1: Oscilloscope capture of read operation

in Figure 4.2. Note that the distributions of the four lowest clock frequencies appear very similar. It would be expected that V_{BL} would continue to rise as the clock is reduced, but this does not appear to be the case. The limited V_{BL} is likely caused by contamination within the array. Cells that share the same column and row as the selected byte can begin conducting to ground if V_{BL} gets high enough. This effectively clamps V_{BL}. To mitigate this issue a lower target V_{BL} should be used for lower clock frequencies.

Power consumption measurements for the different clock frequencies are shown in Table 4.2. For this set of tests V_{REF} was set to 200 mV so that all of the cells would be read as logic 1. This means that the active power consumption numbers represent the worst case since the all data outputs were switching on and off every read cycle. It can be seen that the consumption is relatively linear with frequency. The active power consumption at 4 MHz is less than expected for two reasons. The first is that the test setup was only able to complete read operations once every two read cycles. The second is that considerably fewer cells were being read as logic 1, so the average number of output switching transitions was reduced.

4.1.1 Low-Voltage Read Operation

Although this flash memory was designed to operate at 1 V for read operations, it is possible to reduce this supply voltage. Experiments were conducted to determine how low the supply voltage could be reduced while still remaining functional. As it turns out, the limiting factor was the level shifters on the test board. They are specified to operate down to 0.9 V, but in fact they worked down to 0.65 V. Below this point synchronization between the PSoC and the flash memory ceased to function.

To establish a baseline for these measurements, all 2048 cells were programmed at the nominal V_{DD} of 1 V to a target V_{REF} of 600 mV at 1 MHz clock frequency. Figure 4.3 shows the distribution of all 2048 cells at different supply voltages divided into 5 mV bins. It can be seen that the voltages that all of the bitlines reach decreases with supply voltage. By the time the supply was reduced to 0.80 V, nearly all of the cells' V_{BL} had reduced to less than

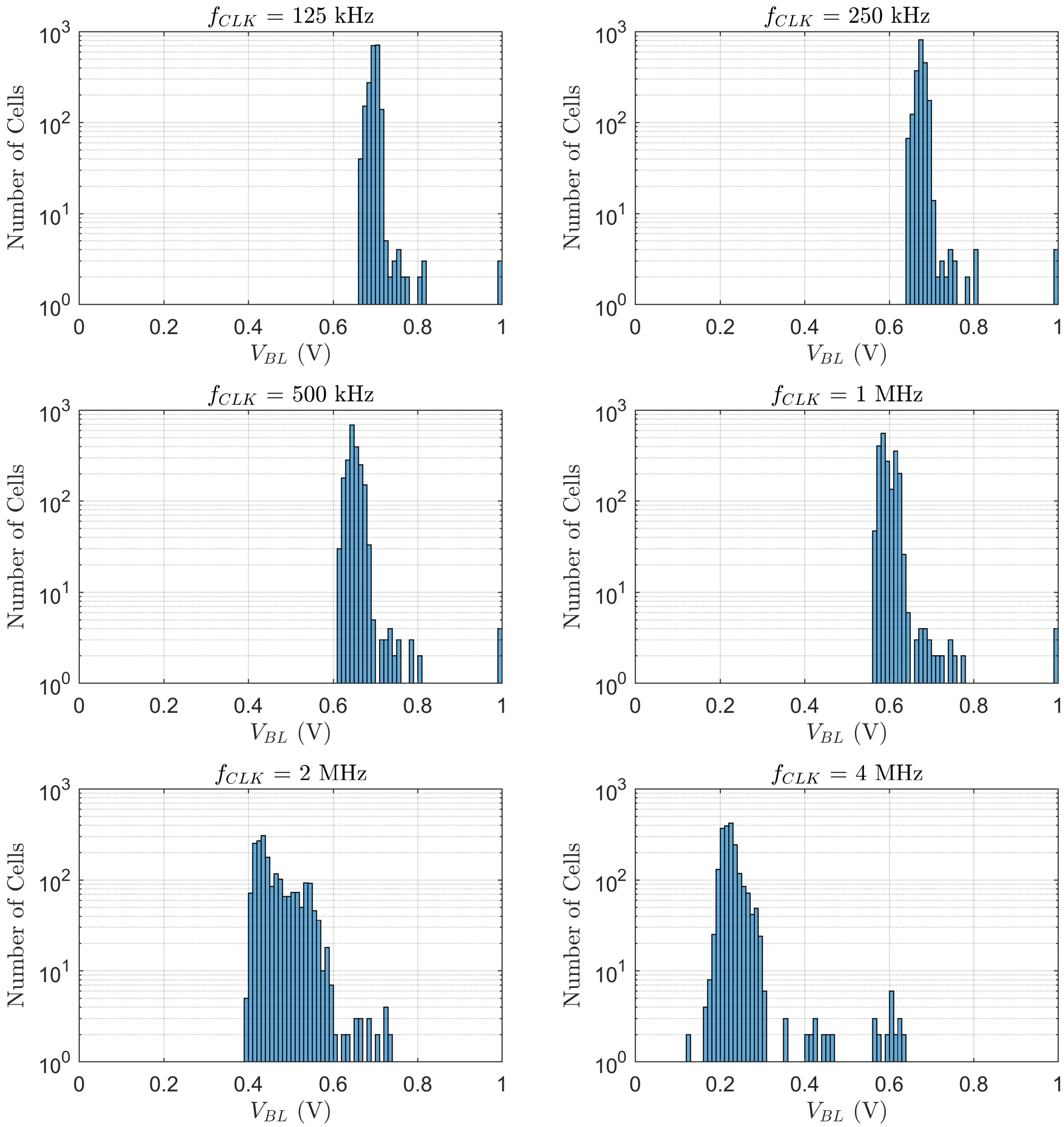

Figure 4.2: Distributions of V_{BL} vs f_{CLK}

Table 4.2: Read power consumption in μW vs f_{CLK}

$\mathbf{f_{CLK}}$ **(MHz)**	$\mathbf{P_{Idle}}$	$\mathbf{P_{Active}}$[1]
0.125	0.74	3.20
0.25	1.45	6.37
0.50	2.86	12.7
1.0	5.36	24.8
2.0	10.0	49.1
4.0	19.3	55.0

1. at $V_{REF} = 200$ mV

50 mV. However there are a few outliers around 450 mV, and this strongly suggests that programming the cells to a higher value will compensate for lower V_{DD}.

Power consumption was also measured at the same V_{DD} values. The results are shown in Table 4.3. The idle values are when no read requests are being issued, and the active values are when read requests are being continuously issued every read cycle. During these measurements V_{REF} was set to 200 mV, and it can be seen that there is a large jump in active power consumption above 0.90 V. This is because most of the bits were read as logic high, so output register's pins were switching repeatedly. The output register was driving the inputs of a level shifter on the test board as well as bond wires and PCB traces. In the typical embedded application that this flash memory is designed for, the output register would only be driving the inputs of a microcontroller, so the actual increase in power consumption would not be as drastic.

4.2 Write and Erase Performance

Write and erase times depend on the voltages applied and on the desired degree of separation between logic 1 and 0. For initial testing purposes with a 1 MHz read clock, a 200 mV reference voltage was chosen. Logic 0 was defined as $\leq$ 100 mV, and logic 1 was defined as $\geq$ 300 mV. With a V_{BIAS} of 400 mV, a typical cell required 325 ms to increase V_{BL} from 100 to 300 mV. Likewise the erase time was 2 s to decrease the cell back to 100 mV. Further testing was performed with V_{BIAS} at 475 mV, V_{REF} at 350 mV, and the lower and upper logic levels were changed to $\leq$ 300 and $\geq$ 400 mV respectively. This reduced the write time to 129 ms and erase time to 710 ms. The total energy consumed to write one byte was 1.56 μJ, and erasing the whole array consumed 241 mJ.

4.2.1 Write Performance

The distribution of write times is shown in Figure 4.4. The write times for all 2048 cells are measurements of the time required to increase the target V_{BL} by 100 mV. Although

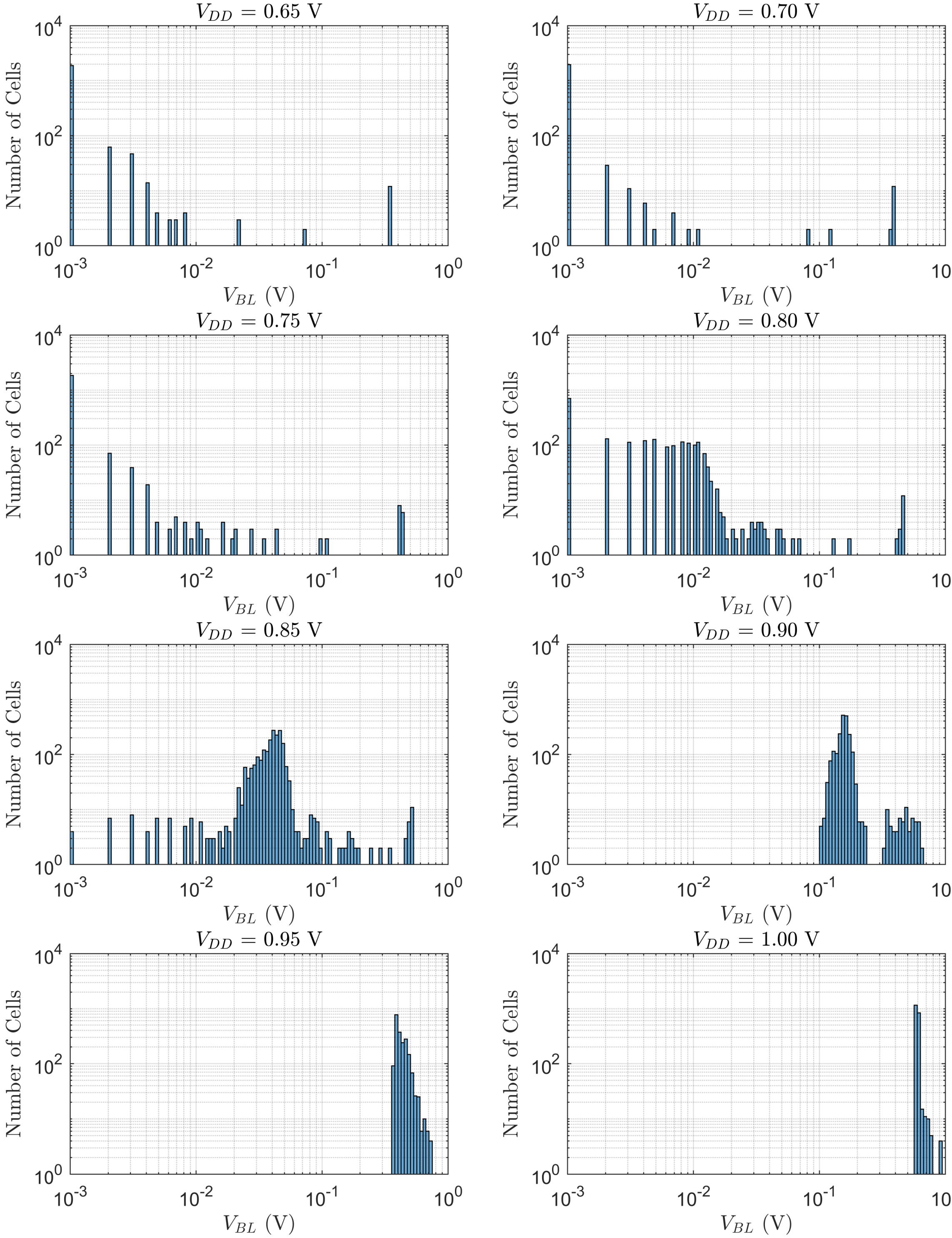

Figure 4.3: Distributions of V_{BL} vs V_{DD} at 1 MHz

Table 4.3: Read power consumption in μW vs V_{DD}

$\mathbf{V_{DD}}$ **(V)**	$\mathbf{P_{Idle}}$	$\mathbf{P_{Active}}$[1]
0.65	1.93	3.36
0.70	2.24	3.88
0.75	2.59	4.47
0.80	2.96	5.14
0.85	3.39	5.92
0.90	3.89	7.15
0.95	4.55	22.1
1.00	5.39	24.8

[1] at V_{REF} = 200 mV

the majority of times are within an order of magnitude, there were times as short as 10 ms and as long as 5 seconds. This large spread of write times makes it difficult to establish a single write time to use when writing to the memory. Fortunately, it is possible to write to individual bits, so a record of each bit's write performance could be used to control write times.

4.2.2 Erase Performance

The distribution of erase times is shown in Figure 4.5. The method of measurement is similar to that of the write time test. However, since erasing is global, the erase times were measurements of how many tunneling pulses were required for each cell's target V_{BL} to decrease by 100 mV. Like the write time test, there was a large spread of erase times. Erase is performed globally, so it may be desirable to compensate by adding write pulses to those cells whose erase times are short. Additionally, caution needs to be exercised to avoid over-tunneling the cells. If too much charge is added to a floating gate, the biasing of the feedback transistor during write operation may be severely affected.

4.3 Performance Comparison

A comparison to existing designs is provided in Table 4.4. Unfortunately some quantities are hard to compare directly since vital information was not provided. In the case of [20], no power consumption values were available. Neither [20] nor [21] stated the clock frequency used during read operation, so direct comparisons of read access times is not possible. However, [21] did provide read and write performance in terms of bitrate, so these were translated to worst case time values assuming that each operation was performed immediately following the previous operation.

There is one important observation to make in this comparison, and that is that there was significantly less write and erase power consumed by this design versus [21]. The inclusion of on-die high-voltage charge pumps and switches carries a significant power penalty. Combined

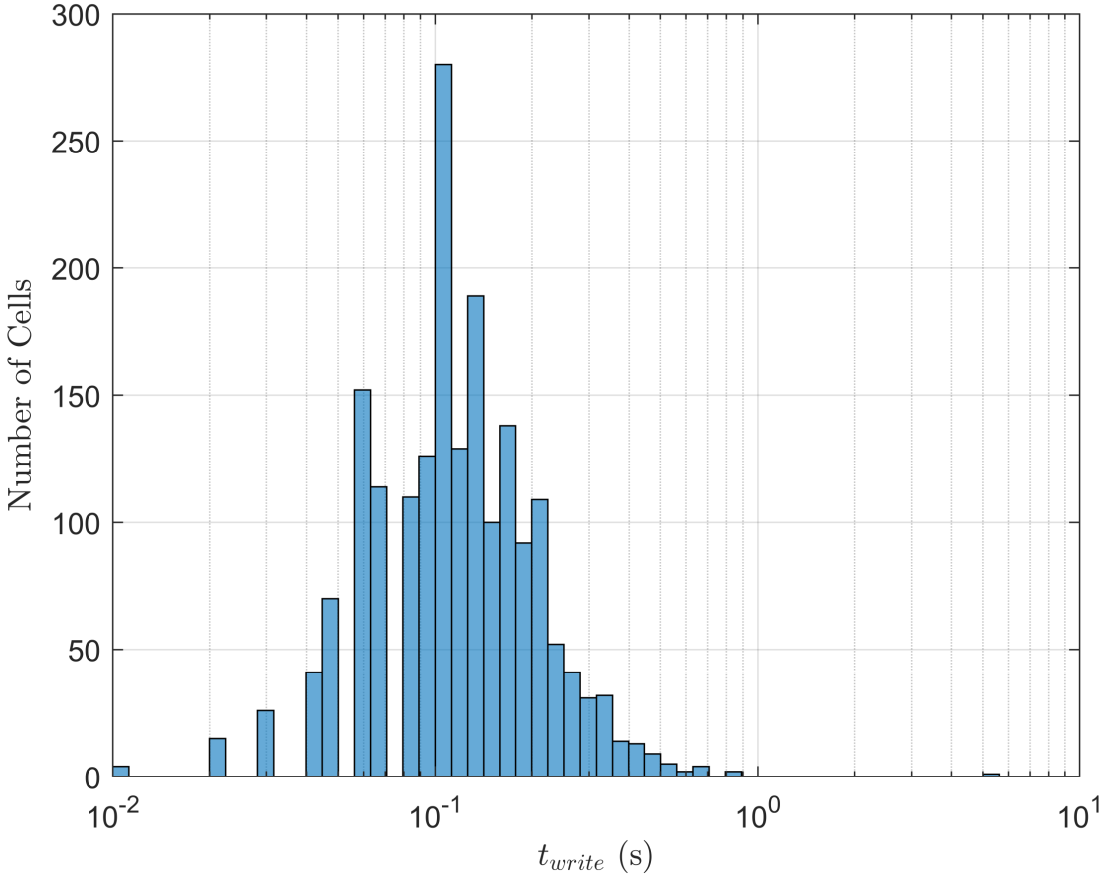

Figure 4.4: Distribution of write times for $\Delta V_{BL,target} = 100$ mV

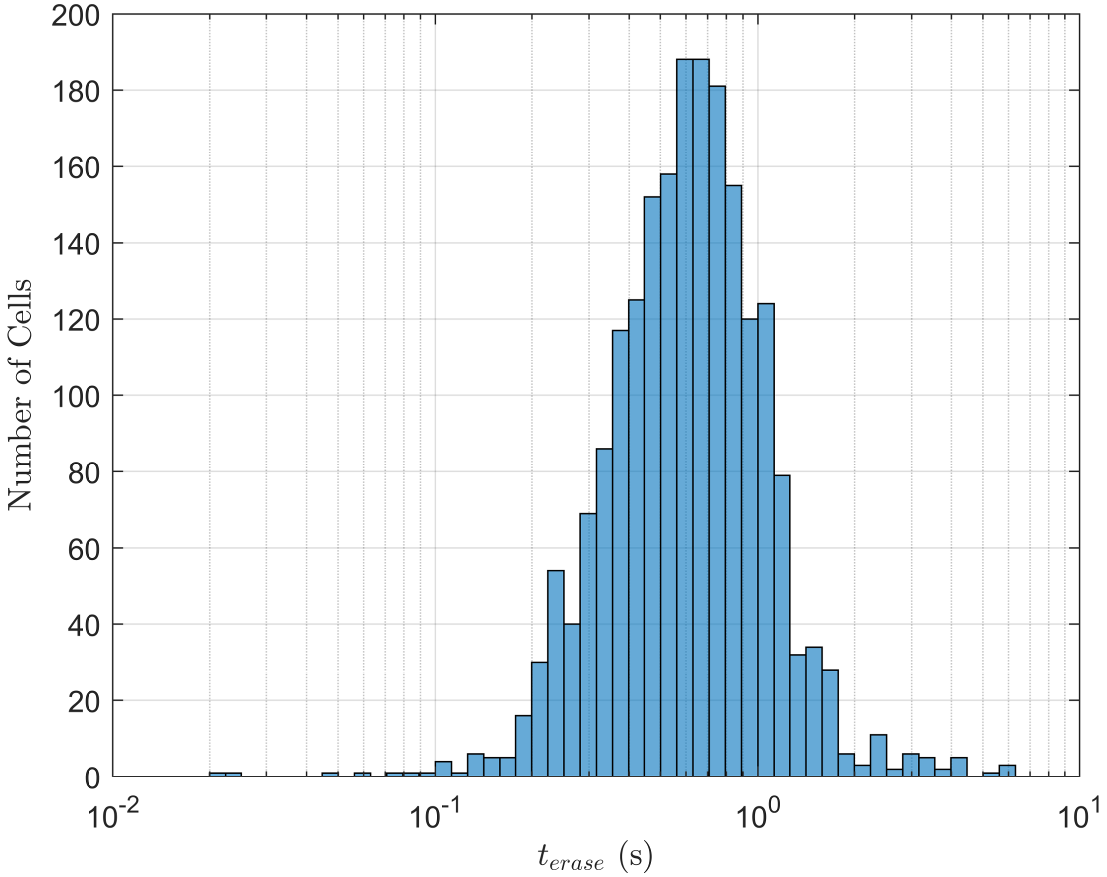

Figure 4.5: Distribution of erase times for $\Delta V_{BL,target}$ = -100 mV

with the area differences shown in Table 3.1, it is clear that the elimination of the extra circuitry should be considered when low write/erase power consumption and high density are of great importance.

Table 4.4: Performance comparison

Parameter	**This Design**	**Raszka** [20]	**Wang** [21]
Read voltage (V)	1	1.2	1.2
Read access time (µs)	3.5	10	0.147
Read power consumption (µW)	24.8[1]	N/A	9.2
Write method	IHEI	FN	FN
Write voltage (V)	3	7	8.8
Write time (ms)	129	10	0.125[3]
Write power consumption (µW)	12.1[2]	N/A	22.9
Erase Method	FN	FN	FN
Erase voltage (V)	7	7	8.8
Erase time (ms)	710	10	0.125[4]
Erase power consumption (µW)	0.34	N/A	22.9

1. at $f_{CLK} = 1$ MHz

2. at $V_{BIAS} = 475$ mV

3. based on 6.8 Mb/s rate

4. based on 8 kb/s rate

Chapter 5

Conclusion and Future Work

Extensive measurements of injection and tunneling currents were performed on the 130 nm CMOS process used in the design on the flash memory. The measurements were used to derive a set of simulation models. Testing of retention characteristics proved to be beneficial in determining that the source of leakage was tunneling current.

The complete flash memory was designed and fabricated. It was tested for performance, and it achieved a 3.5 μs read access time with a 1 MHz clock while consuming less than 25 μW from a 1 V supply. Although the power consumed was higher than the design goal, this includes the power required to drive the level shifters on the test board. Tests at clock frequencies from 125 kHz to 4 MHz revealed that power consumption is linearly dependent on frequency. Read tests were also performed at supply voltages from 0.65 to 1 V. Read operation at 0.65 V was obtained while reducing power to only 3.4 μW. Write time was 129 ms at 12.1 μW, and erase time was 710 ms at 0.34 μW.

A new revision of this flash memory was fabricated. The new design is currently undergoing testing, but results were not yet ready for inclusion in this report. The new revision includes several changes that are intended to improve read performance. During testing of the current version, it was observed that some transient spikes were occurring on the data outputs. This was caused by skew between phases of the phase generator due to loading by the sense amps. In order to eliminate those spikes, buffers were added to the outputs of the phase generator in the new revision. The current design uses flip-flops in the output register, and this required that a half-cycle delay be added to the DR signal to avoid setup time issues. The new revision uses transparent latches instead of flip-flops, so the DR delay was able to be removed. This gives the microcontroller more time to acquire the output data and issue a new read request. Finally, diode-connected NMOS devices were added to clamp the bitlines to a lower voltage. This was done to alleviate two existing issues. Lower clock frequencies allows the bitlines to reach higher voltages, and this causes read accuracy to be compromised when the common-mode input range of the sense amps is exceeded. The higher bitline voltage also causes cross-conduction within the array. Cells that were programmed to a strong logic 1 could cause other cells that share the same row or column to be read high. The diode clamps aim to prevent these two issues by limiting the V_{BL} to less than 300 mV. The diodes can be switched out of the circuit if desired by setting the RMODE signal low.

The consistency of write performance for the existing design needs improvement. The write speed varied over a range of nearly three orders of magnitude. Some cells took more than 5 seconds to write. The problem with writing this long is that other cells are subject to leakage. Even though they are not being written to, turning on WL_EN for extended periods will still cause some cells' V_{BL} to increase slowly. There are two possible explanations for this phenomenon. The first is that the feedback transistors in the memory cells are minimum length devices. These devices are always on in write mode for the selected byte even when WL_EN is low. Their V_{SD} is reduced compared to the injection transistors, but is still possible for low-level injection to occur. Increasing the channel length of the feedback devices should eliminate this, and it may be possible to do so without increasing the footprint of the memory cell. The other possible source of leakage is through the tunneling devices. During write operations GL is grounded, and there is about 2.6 V across the oxide. It is possible for reverse tunneling to occur. However, this is much lower than the voltage used for erasing, so the tunneling rate would be minuscule. It would only be an issue if the memory was idling in write mode for several hours or more.

Overall the results of this project were exceptional. Characterization of the 130 nm CMOS process led to a greater understanding of the physics behind floating-gate operation, and the result was an improved flash cell design. The derived simulation models helped determine the optimal voltages and currents to be used in writing and erasing. Designing the memory in a standard CMOS process allowed it to be fabricated without special structures. This flash memory compares favorably with existing single-poly designs in both performance and areal density making it an excellent design for on-die firmware storage in low-power embedded applications.

References

[1] K. Ohsaki, N. Asamoto, and S. Takagaki, "A single poly EEPROM cell structure for use in standard CMOS processes," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 311–316, Mar 1994. 1, 2, 4, 44, 45

[2] L. Larcher, P. Pavan, and A. Maurelli, "Flash memories for SoC: an overview on system constraints and technology issues," in *Fifth International Workshop on System-on-Chip for Real-Time Applications (IWSOC'05)*, pp. 73–77, July 2005. 1

[3] M. H. White, Y. Yang, A. Purwar, and M. L. French, "A low voltage SONOS nonvolatile semiconductor memory technology," *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part A*, vol. 20, pp. 190–195, Jun 1997. 2

[4] M. French, H. Sathianathan, and M. White, "A SONOS nonvolatile memory cell for semiconductor disk application," in *[1993 Proceedings] Fifth Biennial Nonvolatile Memory Technology Review*, pp. 70–73, Jun 1993. 2

[5] P. Sun, B. Chen, F. Chen, G. Chen, W. Chen, M. Cho, S. Kwon, E. Liu, M. D. Pia, L. Qu, W. Si, J. Sun, C. Tan, A. Tsai, J. wang, D. Xu, V. Yang, C. Yen, K. Miu, C. Yi, E. Lusky, Y. Polanski, I. Nistan, Y. Betser, and B. Eitan, "An high density data flash based on trapped charge programming and erasing," in *2006 8th International Conference on Solid-State and Integrated Circuit Technology Proceedings*, pp. 784–787, Oct 2006. 2

[6] S. A. Jackson, J. C. Killens, and B. J. Blalock, "A programmable current mirror for analog trimming using single poly floating-gate devices in standard CMOS technology," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 48, pp. 100–102, Jan 2001. 2

[7] E. Rosenbaum and L. F. Register, "Mechanism of stress-induced leakage current in MOS capacitors," *IEEE Transactions on Electron Devices*, vol. 44, pp. 317–323, Feb 1997. 10, 12

[8] K. Rahimi, C. Diorio, C. Hernandez, and M. D. Brockhausen, "A simulation model for floating-gate MOS synapse transistors," in *2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No.02CH37353)*, vol. 2, pp. II–532–II–535 vol.2, 2002. 10, 13, 15

[9] J. Lu and J. Holleman, "A floating-gate analog memory with bidirectional sigmoid updates in a standard digital process," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, pp. 1600–1603, May 2013. 13, 18, 47

[10] M. Song, K. P. MacWilliams, and J. C. S. Woo, "Comparison of NMOS and PMOS hot carrier effects from 300 to 77 K," *IEEE Transactions on Electron Devices*, vol. 44, pp. 268–276, Feb 1997. 13, 14

[11] C. Li, X. Gu, J. Li, L. Liu, and G. Li, "A model for single poly EEPROM cells," in *2014 12th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, pp. 1–3, Oct 2014. 13

[12] J. C. Ranuárez, M. Deen, and C.-H. Chen, "A review of gate tunneling current in MOS devices," *Microelectronics Reliability*, vol. 46, no. 12, pp. 1939 – 1956, 2006. 15, 16, 17

[13] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *Journal of Applied Physics*, vol. 40, no. 1, pp. 278–283, 1969. 15

[14] Z. A. Weinberg, "On tunneling in metal-oxide-silicon structures," *Journal of Applied Physics*, vol. 53, no. 7, pp. 5052–5056, 1982. 15

[15] K. F. Schuegraf and C. Hu, "Hole injection SiO_2 breakdown model for very low voltage lifetime extrapolation," *IEEE Transactions on Electron Devices*, vol. 41, pp. 761–767, May 1994. 16

[16] W.-C. Lee and C. Hu, "Modeling CMOS tunneling currents through ultrathin gate oxide due to conduction- and valence-band electron and hole tunneling," *IEEE Transactions on Electron Devices*, vol. 48, pp. 1366–1373, Jul 2001. 16

[17] J. P. Gambino and B. Cunningham, "Junction leakage due to $CoSi_2$ formation on As-doped polysilicon," *Journal of The Electrochemical Society*, vol. 140, no. 9, pp. 2654–2658, 1993. 35

[18] T. Drummond, "Work functions of the transition metals and metal silicides," *Journal of Applied Physics*, Feb 1999. 41

[19] R. B. Merrill, "CMOS image sensor employing silicide exclusion mask to reduce leakage and improve performance." Patent, 12 2000. US 6160282 A. 43

[20] J. Raszka, M. Advani, V. Tiwari, L. Varisco, N. D. Hacobian, A. Mittal, M. Han, A. Shirdel, and A. Shubat, "Embedded flash memory for security applications in a 0.13 µm CMOS logic process," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, pp. 46–512 Vol.1, Feb 2004. 44, 45, 46, 80, 84

[21] Y. Wang, J. Xiang, X. Chen, T. Yang, N. Yan, and H. Min, "A fully logic CMOS compatible non-volatile memory for low power IoT applications," in *Internet of Things (IOT), 2015 5th International Conference on the*, pp. 98–103, Oct 2015. 45, 46, 80, 84

[22] P. Beshay, B. H. Calhoun, and J. F. Ryan, "Sub-threshold sense amplifier compensation using auto-zeroing circuitry," in *Subthreshold Microelectronics Conference (SubVT), 2012 IEEE*, pp. 1–3, Oct 2012. 60

[23] J. Lu and J. Holleman, "A low-power dynamic comparator with time-domain bulk-driven offset cancellation," in *2012 IEEE International Symposium on Circuits and Systems*, pp. 2493–2496, May 2012. 60

Vita

David Andrew Basford was born in Oak Ridge, Tennessee. He began experimenting with electronics when he was only eight years old. He continued to learn electronics on his own and has built many projects for his own use. He completed his Associate's Degree in Computer and Electrical Engineering Technology at ITT Technical Institute in 2009 while earning Highest Honors, Academic Excellence, and Valedictorian awards. He then earned his Bachelor of Science in Electrical Engineering at the University of Tennessee in 2014 with Summa Cum Laude honors. He is currently finishing his Master of Science degree at UT and plans to graduate in December 2017.