NEM Relay Memory Design

Abhinav Gupta
Elad Alon

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-83
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-83.html

May 21, 2009

NEM Relay Memory Design

Abhinav Gupta
University of California, Berkeley
Department of Electrical Engineering and Computer Sciences

May 2009

NEM Relay Memory Design

by Abhinav Gupta

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Elad Alon
Research Advisor

Date

* * * * * * *

Professor Tsu-Jae King Liu
Second Reader

Date

Acknowledgements

In addition to my unforgettable academic journey, the last five years at Berkeley have been one of my most memorable times due largely to the incredible people that I have met. Without their constant support and encouragement, my undergraduate and graduate achievements would not have been possible.

First, I would like to thank my advisor, Professor Elad Alon, who in addition to being the most amazing mentor, has been a figure that one day I hope to become. Despite his vast knowledge and his effortless yet most effective methods of teaching, it is his willingness to learn from all those around that make him truly unique. It has been a privilege being one of his students and truly an honor to learn from him.

I would also like to thank Professor Tsu-Jae King Liu for not only offering her insightful thoughts and comments on this thesis, but more importantly for providing me with an opportunity to research with her at the device level. This allowed me to gain an even deeper perspective into the work presented in this thesis.

The opportunity to work at the Berkeley Wireless Research Center has been another experience of great value to me. Not only have I met a group of enthusiastic faculty members, but more importantly, I've had the pleasure to work with the most

brilliant and talented group of graduate and undergraduate students. First, I would like to thank Chintan Thakkar for *truly* introducing me to circuit design, for being an excellent mentor and an even better friend—his invaluable advice has always improved my work. I will always be thankful to Christian Marcu who has put up with the numerous questions that I have asked him and more importantly for his thorough answers to each one. I would also like to thank Fred Chen for his help on the relay chip by providing his expertise and being my source for clarifications.

My friends Jason Tsai, Adam Abed, and Kenneth Duong whom I met during my graduate work have provided significant support and without whom I could not have completed any of my graduate studies.

In my undergraduate studies, I was extremely fortunate to have met some of my closest friends James Hung, Eric Liaw, Jonathan Wong, Ian Yu, and Glen Wong. Their easygoing yet persistent attitudes helped me accomplish many of my goals.

The always growing encouragement from my best friends Kevin Huang, Christopher Wong, Anand Oza, and Ivan Ilagan has largely shaped me into who I am today, and for this I will be forever thankful to them.

My most influential educators in life have been my father Aditya, my mother Geeta, my sister Megha, and my grandparents (Pitaji, Ammaji, Nanaji, Nani). Their boundless love has been my sole motivation and their immense belief in me my determination. For this, I dedicate my thesis to them.

Abstract

The technology scaling of feature sizes and supply voltages in earlier CMOS designs enabled significant improvements in terms of density, performance, and energy efficiency. With the overall power consumption being largely dominated by the dynamic component, supply scaling drastically reduced the total power albeit exponentially increasing the leakage component. However, today's integrated circuit designs are equally limited by the dynamic and leakage power components halting the trend of further supply scaling and causing designers to examine the use of parallel architectures (e.g. multi-core processors). Although parallelism can continue to improve the energy efficiency achievable by CMOS devices, the finite sub-threshold slope of CMOS transistors will eventually limit any further improvements. Thus, a device with a steeper sub-threshold slope is needed in order to continue the scaling trends beyond those of CMOS technology.

One such device is a nano-electromechanical (NEM) relay consisting of an electrostatically actuated beam that can be positioned to either allow conduction between the source and drain or leave them open-circuited. Because a physical connection determines conduction, relay devices can achieve zero off current and

effectively an infinite sub-threshold slope. However, unlike CMOS technology where delay is largely set by the charging and discharging of capacitances, the mechanical motion of the actuated beam largely dominates the delay of relay-based circuits. Thus, as opposed to traditional CMOS designs where gates are cascaded to construct more complex functions, optimized relay circuit designs arrange for all mechanical motion to occur simultaneously by using large, complex gates.

Considering that a complete system requires both computational blocks and memory structures, this work, in addition to examining logic design, also explores memory designs using relay devices. These designs are then benchmarked to their equivalent CMOS implementations in terms of performance, density, and energy-efficiency. The relay-based logic circuits will be shown to achieve ~10x improvement in energy-efficiency while the memory designs achieve nearly a 3x improvement as compared to CMOS designs for throughputs in the 100 MOPS range with only 20% area overhead.

Table of Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation

During the last 50 years, CMOS technology scaling has provided substantial improvements in terms of density, performance, and energy efficiency. Scaling the minimum feature size of a design by a factor of $1/\sqrt{2}$ has improved the overall density by a factor of 2 ($1/\sqrt{2}$ in the x-dimension and $1/\sqrt{2}$ in the y-dimension). Since the total energy consumption in early designs was largely dominated by the dynamic power dissipation and not by leakage, scaling the supply voltage (V_{DD}) drastically improved energy efficiency. In order to maintain or improve performance at these lower supply voltages, the threshold voltage (V_{th}) was also

scaled, allowing for an increase in on-state current (I_{on}) and reducing the delay. However, this scaling has now reached a point at which an integrated circuit design is limited by its total power consumption—i.e., any further scaling comes at the expense of increased total power.

This has occurred as V_{DD} scaling, although reducing dynamic power consumption, comes at the expense of V_{th} scaling (to maintain performance) that increases the leakage power consumption. The V_{th} of today's transistors is already set to optimally balance a design's dynamic and leakage power consumption, thereby imposing a limit on further supply scaling through conventional means. For further energy efficiency improvements, designers have shifted towards the use of multi-core processors. Each core operates at a lower supply voltage with lower power and at a lower throughput, but as long as parallelism is available, the overall performance is unchanged. However, even at arbitrarily low per-core performance, the energy efficiency achievable by CMOS devices is ultimately limited by their off-state leakage (I_{off}) which will cause even this technique to eventually become ineffective. If a device with a steeper sub-threshold slope (one that achieves a more ideal switching behavior—i.e., lower I_{off} for the same I_{on}) can be identified, this challenge can potentially be overcome, allowing for further scaling improvements in energy efficiency beyond the limits of CMOS technology.

1.2 Thesis Overview

The thesis concentrates on integrated circuit design using nano-electromechanical (NEM) relays, a device that offers zero off-state leakage. More specifically, since a complete digital system requires computational blocks and memory structures, the primary focus of this work is to use these devices in designing such structures and analyzing them in terms of density, performance, and energy per operation.[1] Before describing this device and the respective structures, chapter 2 will provide background information needed to understand the limitations imposed by CMOS scaling, and will further build upon the concepts discussed in the previous section. With the detailed understanding of these limitations, chapter 3 will delve into the design and operation of the NEM relay device. Furthermore, this chapter will explain their electrical and mechanical characteristics, which will then be used to detail digital logic circuit design strategies tailored for these relays. The chapter will conclude by probing further into designing digital logic using these devices by reviewing a NEM relay-based full-adder. Although the relay-based adders have been shown to have a 10x energy efficiency improvement over those of CMOS, such an improvement will pose a challenge for memory structures. This is because logic

[1] The work done in [3] has extensively examined the use of NEM relays in designing computational blocks. Thus, this thesis will briefly review those blocks while the main focus of the work will be to examine the device for the design of memory structures.

optimization techniques can be used in designing relay-based computational blocks that aim to reduce the total number of devices thus mitigating a single relay device's area-overhead. However, the area-overhead encountered in designing relay-based memories is much more pronounced due to CMOS memories already being highly-dense that limit optimization techniques. Thus, chapter 4 will delve further into these challenges as well as discuss two different relay memory designs, their read and write operations, and analyses of their area, throughput, and energy/operation. The chapter will conclude by examining a relay-based sense-amplifier design while investigating the trade-off that it presents between reduced read energy and reduced throughput. Once the models for these metrics have been established, chapter 5 will take these two memory structures and benchmark them against a CMOS SRAM design in terms of throughput, energy (read and write), and area. Chapter 6 will conclude the thesis by providing a summary of the current work as well as discussing future research directions in integrated circuit design with relays.

2

Limitations of CMOS Scaling

This chapter serves to provide a detailed analysis on the limitations of CMOS scaling by discussing the components of energy dissipation as a function of supply voltage.

2.1 Scaling & Dynamic Energy

As mentioned in Chapter 1, V_{th} scaling improves the on-state current, thus improving performance. For a short-channel device and using the velocity saturated model [1], equation 2.1 below shows the relationship between on-current (I_{DSat}) and V_{th} assuming that $V_{GS}=V_{DD}$:

$$I_{DSat} = v_{sat} W C_{ox} \frac{(V_{DD} - V_{th})^2}{(V_{DD} - V_{th}) + E_c L} \quad (2.1)$$

Here, v_{sat} represents the saturated velocity (due to carrier scattering), E_c is the critical value of the electric field at which velocity saturation occurs, W is the width of the transistor, and C_{ox} is the oxide capacitance. As can be seen, a decrease in V_{th} increases I_{DSat}.

The overall performance of a system can be analyzed by examining its delay with equation 2.2 below showing the relationship between delay and on-current [1]:

$$t_p = \frac{K C_L V_{DD}}{I_{DSat}} \quad (2.2)$$

Here, K is a scaling factor[2] and C_L is the load capacitance being driven. By combining equations 2.1 and 2.2, a decrease in V_{th} increases I_{DSat} and decreases t_p. Furthermore, as long as V_{DD} and V_{th} scale by the same factor, V_{DD} can also scale without hindering performance, while drastically reducing the dynamic energy of a design. This occurs because dynamic energy is quadratically related to V_{DD} [1]:

$$E_{dyn} = \alpha C_L V_{DD}^2 \quad (2.3)$$

Here, α refers to the activity or switching factor (the probability that C_L will be charged or discharged in a given cycle).

[2] The scaling factor, K, depends on the output value at which t_p is evaluated, e.g. if t_p is defined to be when the output voltage crosses $V_{DD}/2$, then K would equal ½.

2.2 Scaling & Leakage Energy

For energy-constrained applications with less stringent performance requirements, the sub-threshold regime of operation ($V_{DD}<V_{th}$) offers a technique to further reduce energy consumption. However, in this regime, V_{DD} and V_{th} must be carefully set as V_{th} scaling, despite its performance and dynamic energy benefits, comes at the expense of exponentially increased leakage current (the source to drain carrier transport that occurs even for $V_{GS}<V_{th}$). Figure 2.1 displays the effects of V_{th} scaling:

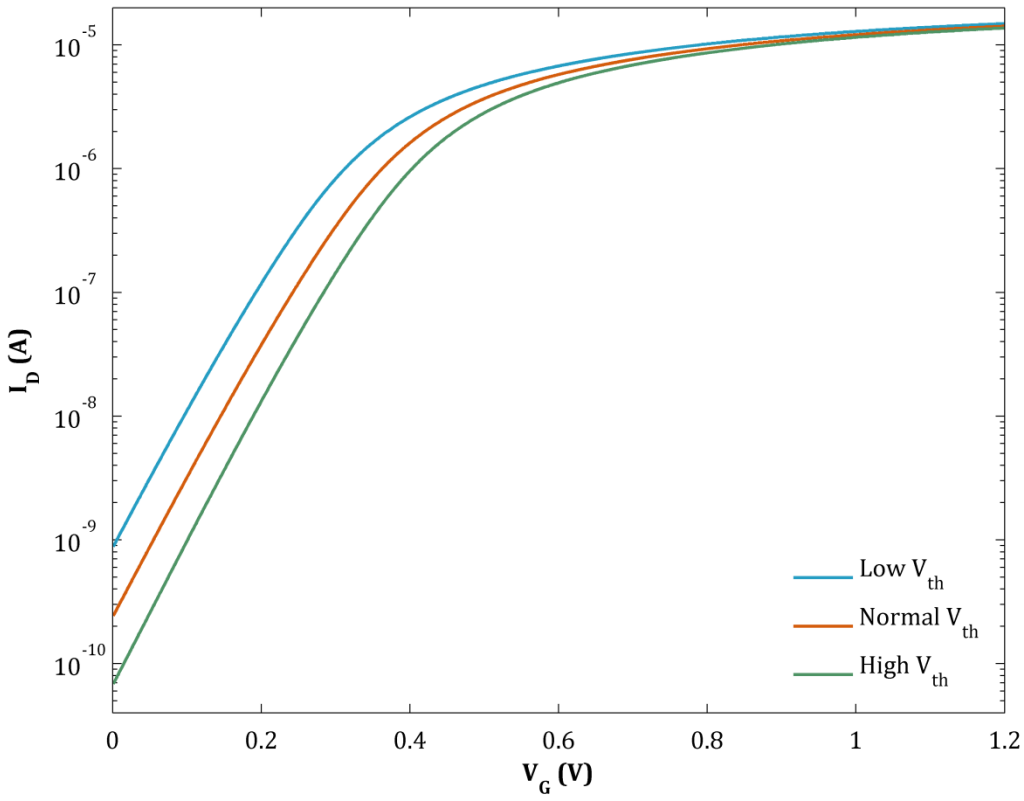

Figure 2.1: I_D vs. V_{GS} for varying V_{th} (V_{DS}=50mV), comparing on and off currents

The plot illustrates that the device with the lower V_{th} experiences a higher on-current, but also a much higher off-current (leakage current, I_D for V_{GS}=0). The opposite is true for the device with the higher V_{th}. The large increase in off-current occurs because in the sub-threshold regime, the drain current ($I_{D,Sub}$) is exponentially related to V_{th} [2]:

$$I_{D,Sub} = I_o e^{\frac{(V_{GS}-V_{th})}{nV_T}}$$
$$where\ I_0 = \mu_0 C_{ox} \left(\frac{W}{L}\right)(n-1)V_T^2 \quad (2.4)$$

Here, V_T is the thermal voltage equal to (kT/q), n is the sub-threshold slope factor equal to ($1+C_{dep}/C_{ox}$), C_{dep} is the depletion capacitance, and μ_0 is the mobility. As can be seen, reducing V_{th} has an exponentially increasing effect on $I_{D,Sub}$. Furthermore, since delay is inversely proportional to the drain current, t_p in the sub-threshold regime is now also exponentially dependent on V_{th} [2] and is given by:

$$t_p = \frac{KC_L V_{DD}}{I_o e^{\frac{(V_{GS}-V_{th})}{nV_T}}} \quad (2.5)$$

By examining equations 2.4 and 2.5 above, in the sub-threshold regime, a decrease in V_{th} increases the leakage current by exactly the same amount as it decreases the delay (and vice versa for an increase in V_{th}). So, V_{th} is set by the desired performance and the total leakage energy can be derived from the product of the leakage power ($V_{DD} \times I_{D,Sub}$) and delay, yielding equation 2.6 below.

$$E_{Leak} \propto KC_L V_{DD}^2 e^{-\frac{V_{DD}}{nV_T}} \quad (2.6)$$

Equation 2.6 shows that E_{Leak} is a function of V_{DD} through a quadratic term and an exponential term. For large V_{DD}, the exponential term approaches 0 and is much smaller than the increase in the quadratic term keeping E_{Leak} at a minimum. However, as V_{DD} decreases further into the sub-threshold regime, the exponential term increases more rapidly than the decrease in the quadratic term, causing E_{Leak} to increase significantly.

2.3 Scaling & Total Energy

From the discussions in the previous two sections and from equations 2.3 and 2.6, supply scaling yields a tradeoff between dynamic and leakage energies—i.e., reducing V_{DD} decreases E_{dyn} at the expense of increasing E_{Leak}. Because of this tradeoff, a particular node and design has a certain V_{DD} that minimizes its energy/operation for any delay [2]. The total energy is given by the summation of E_{dyn} and E_{Leak}:

$$E_{Tot} \propto V_{DD}^2 \left(C_L + KC_L e^{-\frac{V_{DD}}{nV_T}} \right) \quad (2.7)$$

Figure 2.2 below plots the expressions for E_{dyn}, E_{Leak}, and E_{Tot} as a function of V_{DD}.

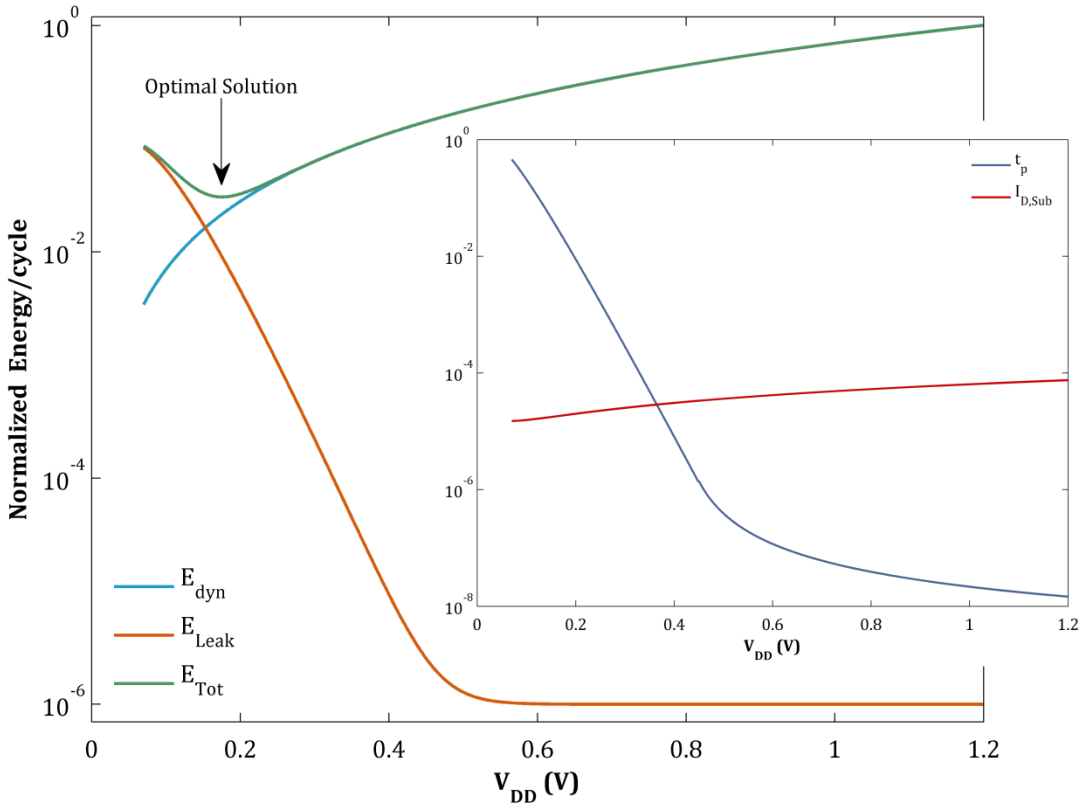

Figure 2.2: Energy vs. V_{DD}, comparing dynamic and leakage energies

The plot shows the dynamic energy decreasing with V_{DD} while leakage energy exponentially increases with V_{DD}, thus yielding a well-defined minimum total energy (the subplot of Figure 2.2 displays how the delay and leakage current change with V_{DD}). Since the minimum total energy point is independent of delay for a design operating in the sub-threshold regime, the energy per operation will eventually level off regardless of how slowly a CMOS design is allowed to run—i.e., the energy

will no longer decrease for any increase in delay. Furthermore, since the total energy is independent of V_{th}, the energy efficiency achievable by CMOS circuits eventually becomes limited.

A conceptual solution to overcome this limitation can be found by re-examining Figure 2.1: if the sub-threshold slope can be made steeper—i.e. the ratio of on-current to off-current can be improved—the same circuit design would experience a lower leakage current at the same supply voltage, making further energy efficiency improvements possible. The following chapter analyzes one such device and examines its structure and operation.

3

Structure and Operation of NEM Relay Devices

As described in the previous section, if a device with a steeper sub-threshold slope can be identified, the point at which the total energy levels off will occur at higher delays, overcoming the challenge of the limitation on energy efficiency imposed by CMOS technology. This chapter begins by examining the structure and the operation of a nano-electromechanical (NEM) relay as a candidate device for potentially an improved energy-efficiency. The turn-on and turn-off characteristics of the device are then explained as well as a delay model to be later used for determining the device's performance in circuit blocks. As mentioned earlier, since computational blocks are a requirement for any digital system, this chapter concludes by reviewing

the design of relay logic circuits. More specifically, a 32-bit relay-based adder is explained, and benchmarked against a 32-bit CMOS adder.

3.1 Structure of NEM Relay

A NEM relay is an electrostatically actuated mechanical switch whose state of operation is set by the voltage difference between a movable gate terminal and a fixed body terminal [3]. Figure 3.1 shows the top and cross-sectional views of this device as well as its circuit symbol. The cantilever gate electrode attaches to the metallic channel via an insulating gate dielectric (cross-sectional view). In the off state, where the gate to body voltage ($|V_{gb}|$) is less than a characteristic "threshold" voltage (V_{th}), an air gap separates the channel from the metallic source and drain. Since there is no path for current to flow, $I_D=0$. In the on state, where $|V_{gb}|>V_{th}$, the electrostatic force is sufficient to bend the cantilever gate enough that the metallic channel comes into contact with the source and drain, allowing for current to flow. Since the device exhibits no leakage current and experiences an abrupt turn-on, it has an extremely steep effective sub-threshold slope. In order to benchmark the performance of relay-based circuits, the complete behavior of the turn-on and turn-off operation must be examined.

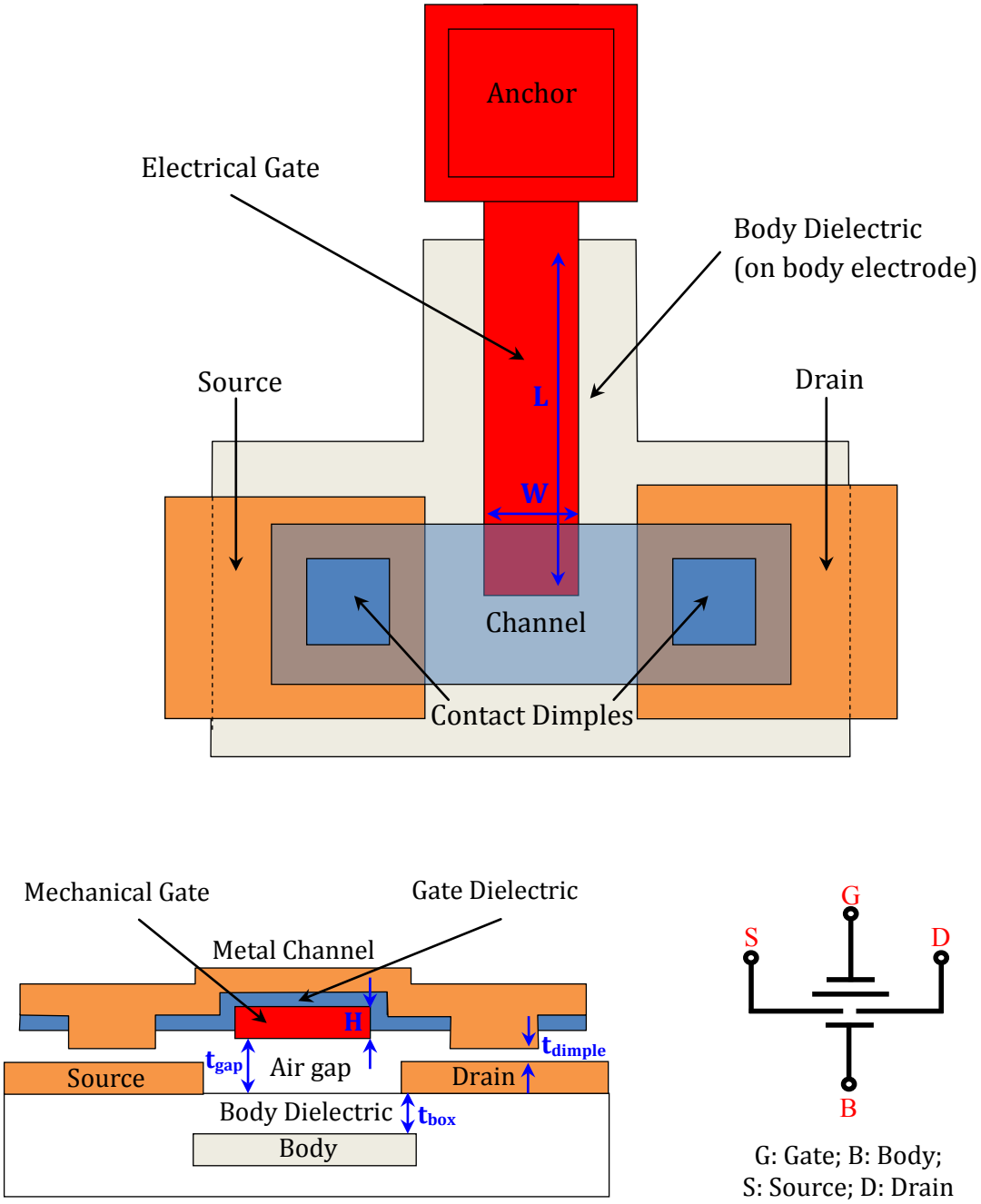

Figure 3.1: Top & cross-sectional view of the NEM relay, including its circuit symbol

3.2 Operation of NEM Relay

3.2.1 Pull-In Behavior and V_{pi}

The mechanically actuated cantilever beam can be modeled as a linear spring-mass system [3]. To aid in understanding the dynamics of this beam during the turn-on operation, Figure 3.2 below illustrates the forces acting on the system.

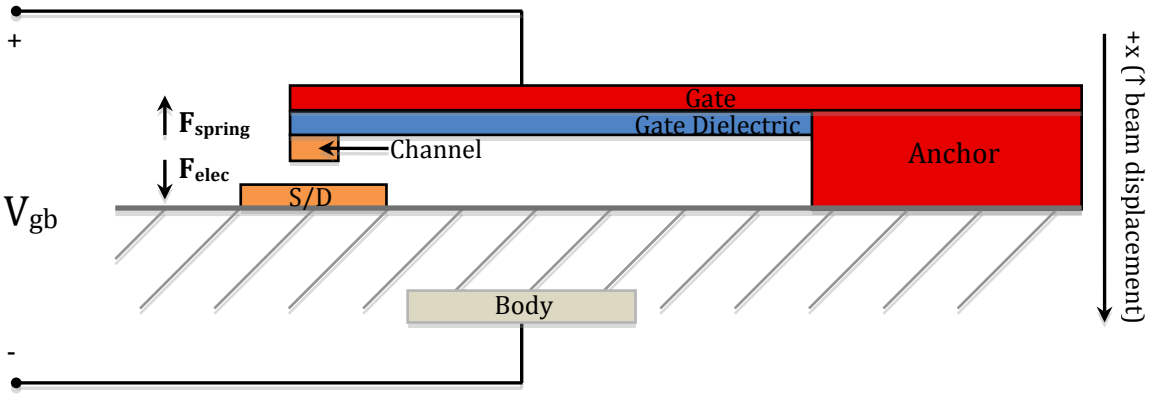

Figure 3.2: Beam dynamics for pull-in operation (Note the direction of +x).

When a gate-to-body voltage (V_{gb}) is applied between the cantilever gate and body electrodes, it results in an increasing electrostatic force (F_{elec}), attracting the cantilever beam towards the body node. This force is given by:

$$F_{elec} = \frac{\varepsilon_0 (WL) V_{gb}^2}{2(d_{eff} - x)^2} \tag{3.1}$$

Here, $d_{eff} = t_{gap} + (t_{box}/\epsilon_{box})$ where t_{gap} and t_{box} refer to the physical thickness of the air gap and the body dielectric thickness, respectively (as illustrated in Figure 3.1), and ϵ_{box} is the relative permittivity of the body dielectric. W and L are defined as illustrated in Figure 3.1 and refer to the width and length of the cantilever beam, respectively. Note that in equation 3.1, x refers to beam displacement and is defined as positive (i.e. increasing) when the beam moves towards the body electrode as illustrated in Figure 3.2.

With the beam displacement increasing, the spring restoring force (F_{spring}), which counteracts F_{elec}, also increases. This spring force is given by:

$$F_{spring} = kx \tag{3.2}$$

Here, $k = \gamma EW(H/L)^3$ where γ is a proportionality constant depending on the beam structure and equals ~0.25 for a cantilever, E is the Young's modulus of the cantilever beam material, and H is the thickness of the beam (as illustrated in Figure 3.1) [3]. Since F_{elec} increases quadratically with increasing x while F_{spring} only increases linearly, there exists a critical beam displacement (x_{crit}) for which F_{elec} is always larger than F_{spring} [4], causing the beam to "pull-in" and close the air gap. Thus, the turn-on operation is also referred to as the pull-in operation.

Finding x_{crit} can be done by examining the stability of the system—i.e. by analyzing the beam's response to a small perturbation from its resting position.[3] The first step in examining the stability is to express the net force (F_{net}) acting on the system. Letting $d = (d_{eff} - x)$ represent the residual thickness of the gap once the beam has displaced by an amount x, and by defining F_{net} to be positive in the upward direction (same as F_{spring}), F_{net} can be expressed as:

$$F_{net} = \frac{-\varepsilon_0 (WL) V_{gb}^2}{2d^2} + k(d_{eff} - d) \tag{3.3}$$

The system is said to be in equilibrium condition when $F_{net}=0$ (i.e. for a given V_{gb}, F_{spring} is equal but opposite to F_{elec}, making the beam stationary). Stability analysis on the system can now be done by perturbing the beam by δd from its equilibrium position and examining δF_{net}. The relationship between the two is given by:

$$\begin{aligned} \delta F_{net} &= \frac{\partial F_{net}}{\partial d} \delta d \\ where\ \frac{\partial F_{net}}{\partial d} &= \frac{\varepsilon_0 (WL) V_{gb}^2}{d^3} - k \end{aligned} \tag{3.4}$$

If δd is negative ($d_{final} < d_{initial}$, implying the gap closes slightly) and results in δF_{net} being positive ($F_{net,final} > F_{net,initial}$, implying the force pushes the beam back up), the system is stable. This means that, if perturbed, the negative relationship between δd and δF_{net} makes the beam return to its equilibrium position. If however,

[3] The detailed analysis for x_{crit} has been done in [5] and many other works, and only the key components of the analysis have been repeated here in order to build intuition.

δd is negative and results in δF_{net} also being negative ($F_{net,final} < F_{net,initial}$ implying the force pushes the beam further down), the system is unstable as slightly closing the gap makes the net force push the beam down even further. In other words, if perturbed, the positive relationship between δd and δF_{net} makes the gap close abruptly.

Since the relationship between δd and δF_{net} is defined by equation 3.4, the system is stable for $\partial F_{net}/\partial d < 0$ and is unstable for $\partial F_{net}/\partial d > 0$. The condition of interest is the boundary condition—the point at which the system goes from being stable to being unstable. This occurs when $\partial F_{net}/\partial d = 0$; the gap thickness and V_{gb} associated with this point are referred to as the pull-in gap (d_{pi}) and the pull-in voltage (V_{pi}), respectively. Setting $\partial F_{net}/\partial d$ in equation 3.4 to 0 yields:

$$k = \frac{\varepsilon_0 (WL) V_{pi}^2}{d_{pi}^3} \tag{3.5}$$

Remembering that in equilibrium, F_{net} equals 0, and by using equations 3.3 and 3.5, the expressions for d_{pi} (including x_{crit}) and V_{pi} can be evaluated.

$$d_{pi} = \frac{2}{3} d_{eff} \implies x_{crit} = \frac{d_{eff}}{3} \tag{3.6}$$

$$V_{pi} = \sqrt{\frac{8}{27} \frac{\gamma E H^3 d_{eff}^3}{\varepsilon_0 L^4}} \tag{3.7}$$

To understand the same concept graphically, Figure 3.3 below plots normalized displacement $(\zeta = x/d_{eff})$ vs. normalized F_{spring} and F_{elec}:

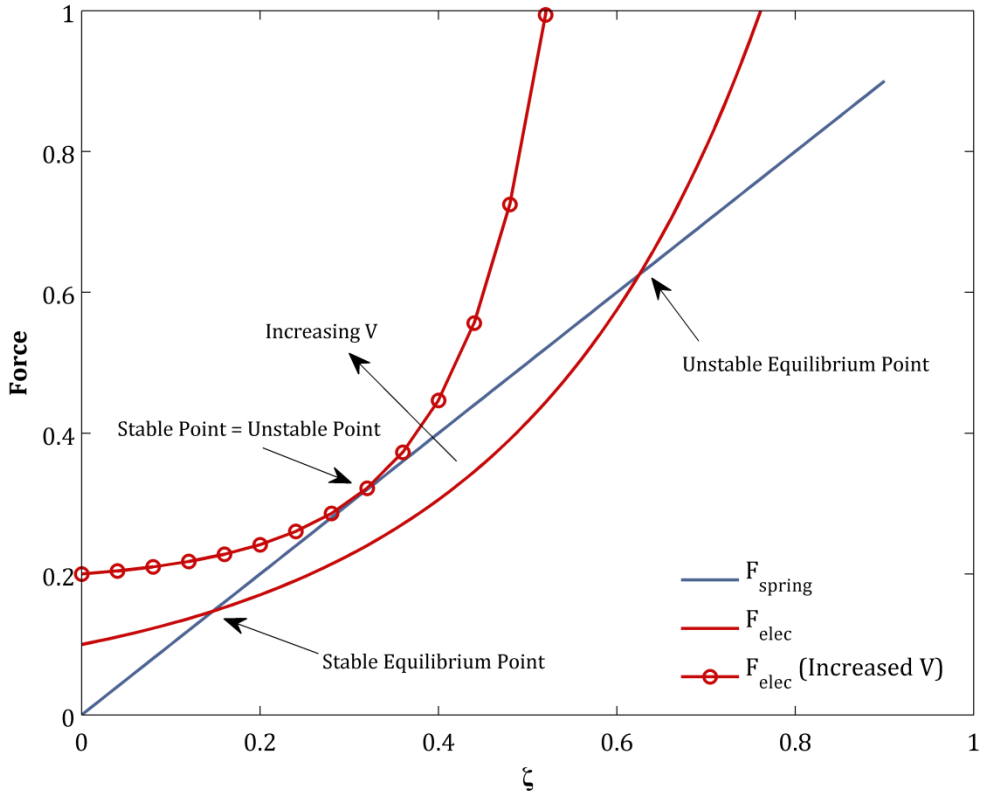

Figure 3.3: Stability analysis—displacement vs. force

The stable equilibrium point for a given V_{gb} occurs when $F_{net}=0$ and any perturbations of the beam cause it to return back to its original position. The unstable equilibrium point also occurs when $F_{net}=0$ but any perturbations of the beam cause it to close the gap. Here, the condition of interest is when the stable

equilibrium point meets the unstable equilibrium point, which occurs when V_{gb} increases, shifting the F_{elec} curve towards the F_{elec} (Increased V) curve. The voltage at which the two points meet is V_{pi}, and by extrapolating from Figure 3.3, this occurs at $\zeta \sim 1/3 \Longrightarrow x_{crit} = d_{eff}/3$.

3.2.2 Pull-Out Behavior and V_{po}

To aid in understanding the dynamics of the NEM relay during the turn-off operation, Figure 3.4 below illustrates the forces affecting the system.

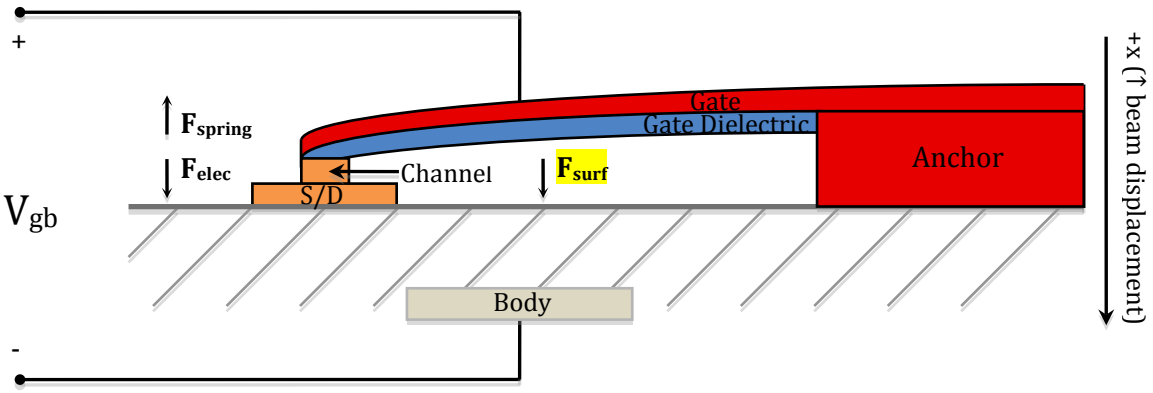

Figure 3.4: Beam dynamics for pull-out operation

The condition for the turn-off case may at first seem similar to the turn-on condition, since reducing V_{gb} beyond a certain point would make F_{spring} always larger than F_{elec} and cause the beam to "pull-out." However, there are two

modifications: (1) the beam needs to displace back up just enough to stop tunneling current between the channel and the source-drain terminals, thus not needing to go all the way back to its initial starting point as in Figure 3.2 and (2) the added surface force (F_{surf}) due to van der Waals attraction between interacting surfaces [5] sets a lower limit on the spring restoring force. The former can be taken into account by using a value for $x=x_{crit,po}$ that ensures very small tunneling current, and the latter can be incorporated by finding an expression for F_{surf} and adding it to the net force equation. The condition for the turn-off or pull-out case then becomes the gate-to-body voltage that makes $F_{net}=0$. The value of this V_{gb} is referred to as the pull-out voltage (V_{po}). The expression for F_{surf} appears in equation 3.8 below followed by the modified $F_{net}=0$ equation used to find V_{po}:

$$F_{surf} = \frac{E_{surf}\left(12\pi d_{0,vdW}^2\right)}{6\pi\left(d_{eff} - x_{crit,po}\right)^3} \tag{3.8}$$

$$\frac{\varepsilon_0 (WL) V_{po}^2}{2\left(d_{eff} - x_{crit,po}\right)^2} + \left[F_{surf} \times (Contact\ Area)\right] = k x_{crit,po} \tag{3.9}$$

Here, E_{surf} is the adhesion surface energy estimated to be $\sim 150\mu J/m^2$ [3] and $d_{0,vdW}$ is the contact distance between atoms estimated to be ~5Å [5]. By examining equations 3.9 and 3.3 (for $F_{net}=0$), it should be noted that V_{pi} is always greater than V_{po} generating a hysteretic effect between the turn-on and turn-off voltages. This is illustrated in Figure 3.5 below that plots the relay beam length (L) vs. V_{pi} and V_{po} for

beam parameters in [3] and also depicts this hysteric effect by plotting the beam displacement (x) vs. V_{gb}.

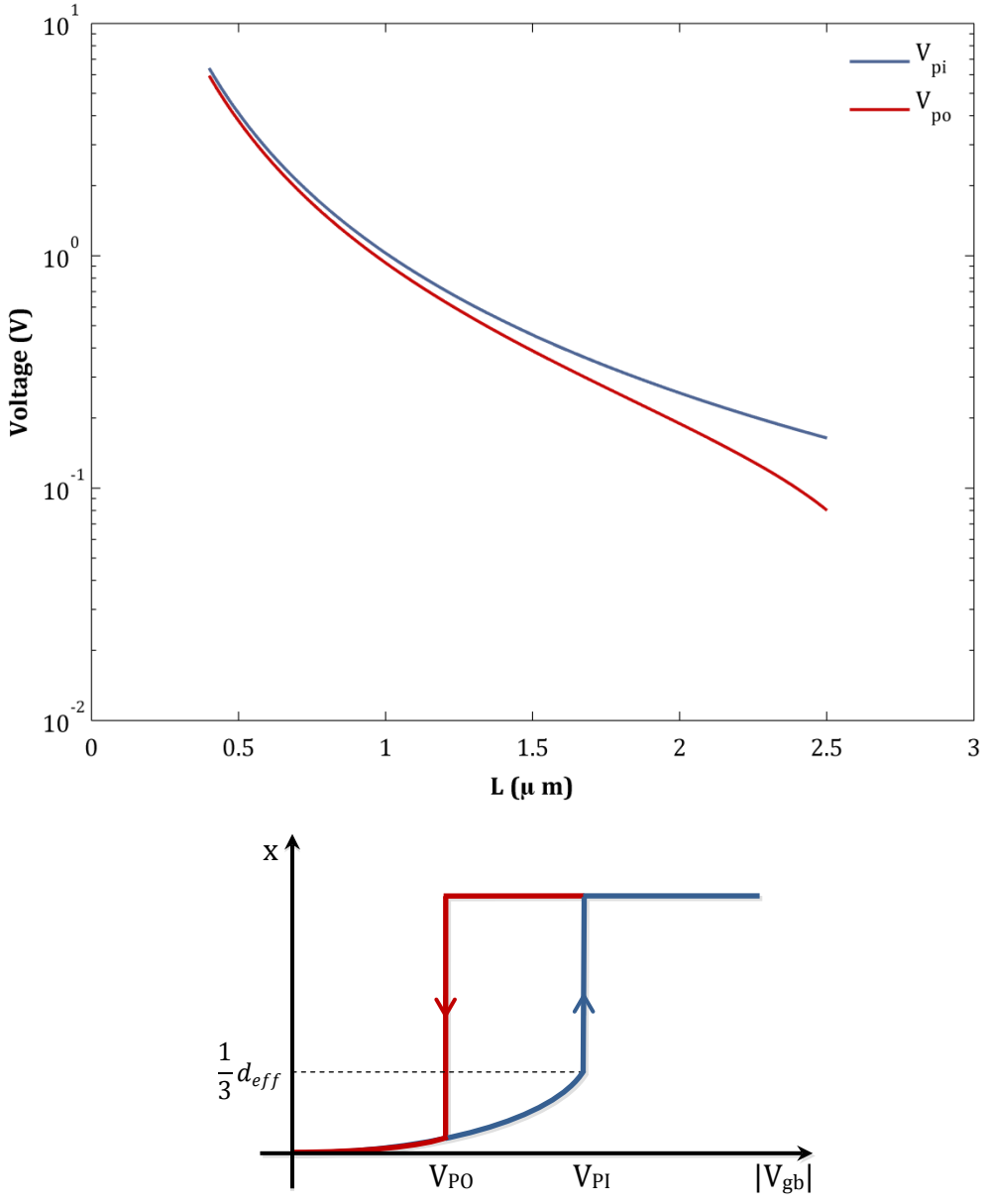

Figure 3.5: L vs. V_{pi} and V_{po} and displacement vs. V_{gb} illustrating hysteretic effect

3.2.3 Delay Model: $t_{p,on}$ and $t_{p,off}$

For numerically benchmarking the performance of relay-based circuits, the exact turn-on and turn-off delays of the device must be examined. The turn-on delay, $t_{p,on}$ is the time the beam takes to move from the off-state to the on-state, and the turn-off delay, $t_{p,off}$, is the opposite. To determine expressions for these, the relay's circuit model capturing its electro-mechanical behavior and its parasitics [3] is used:

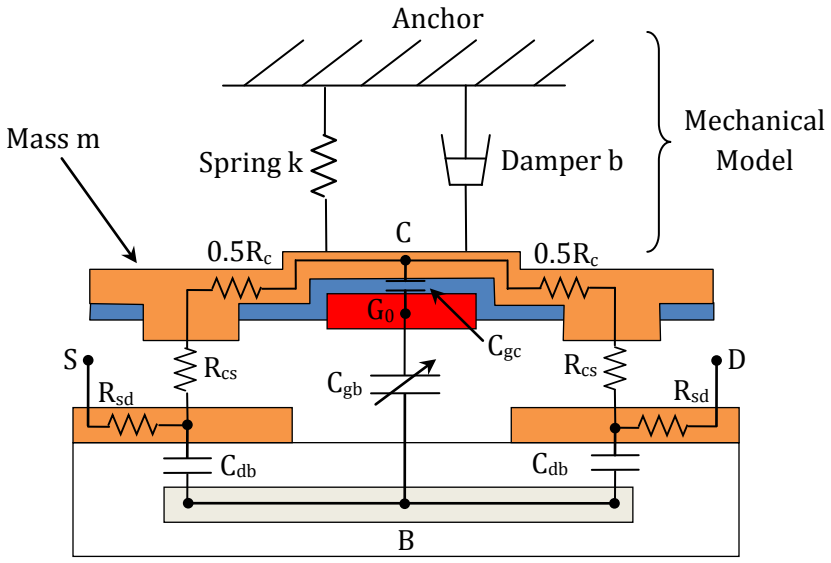

Figure 3.6: Spring-damper-mass system representation of NEM relay

To find $t_{p,off}$, the spring-damper-mass system can be modeled as a 2nd order linear differential equation [6]. Evaluating the time it takes to move $x_{crit,po}$ yields $t_{p,off}$.[4]

[4] The analysis assumes that the surface force disappears once the channel and the source/drain electrodes are no longer in contact. This keeps the dynamics of the beam linear [6].

$$t_{p,off} = \frac{1}{\omega_0 Q} ln\left(\frac{1}{1 - \frac{x_{crit,po}}{t_{dimple}}}\right) \tag{3.10}$$

Here, $\omega_0 = \sqrt{k/m}$ and $Q = \sqrt{km}/b$ are the resonant frequency and the quality factor of the cantilever beam, respectively with equation 3.10 being valid for $Q \approx 0.5$. m is the mass of the beam and equals ρWHL with ρ being the density of the beam material, b is the linear damping factor, and t_{dimple} is defined as in Figure 3.1.

Expressing $t_{p,on}$ is a bit more challenging as the model for the spring-damper-mass system becomes a highly non-linear 2nd order differential equation. This is because the beam is accelerated in the turn-on case with F_{elec} being a non-linear function of x. Although this becomes a challenge, several curve fitting models [5,7] approximate the solution very accurately. Using one such model yields $t_{p,on}$.

$$t_{p,on} = \alpha \sqrt{\frac{m}{k}\frac{t_{dimple}}{t_{gap}}} \left(\frac{V_{pi}}{V_{DD}}\right)^{\beta} \tag{3.11}$$

Here, α is a proportionality constant and β a power constant. For $V_{DD} \geq 1.5V_{pi}$[5]:

$$\begin{aligned} ln(\alpha) &= 0.179\big(ln(Q)\big)^2 - 0.455ln(Q) + 1.651 \\ \beta &= 0.128\big(ln(Q)\big)^2 - 0.333ln(Q) + 1.465 \end{aligned} \tag{3.12}$$

Note that $t_{p,on}$ is a function of the ratio between V_{pi} and V_{DD} and can be changed via V_{DD} (with energy tradeoffs), while $t_{p,off}$ is solely determined by the beam's geometry

[5] For a detailed explanation of why V_{DD} is chosen to be larger than $1.5V_{pi}$, refer to Appendix A: Noise Margins for Relay Circuits.

and properties of the beam material. Figure 3.7 below plots $t_{p,on}$ and $t_{p,off}$ vs. L for $V_{DD}=1.5V_{pi}$. As can be seen from the plot, $t_{p,off}$ is approximately 4x smaller than $t_{p,on}$:

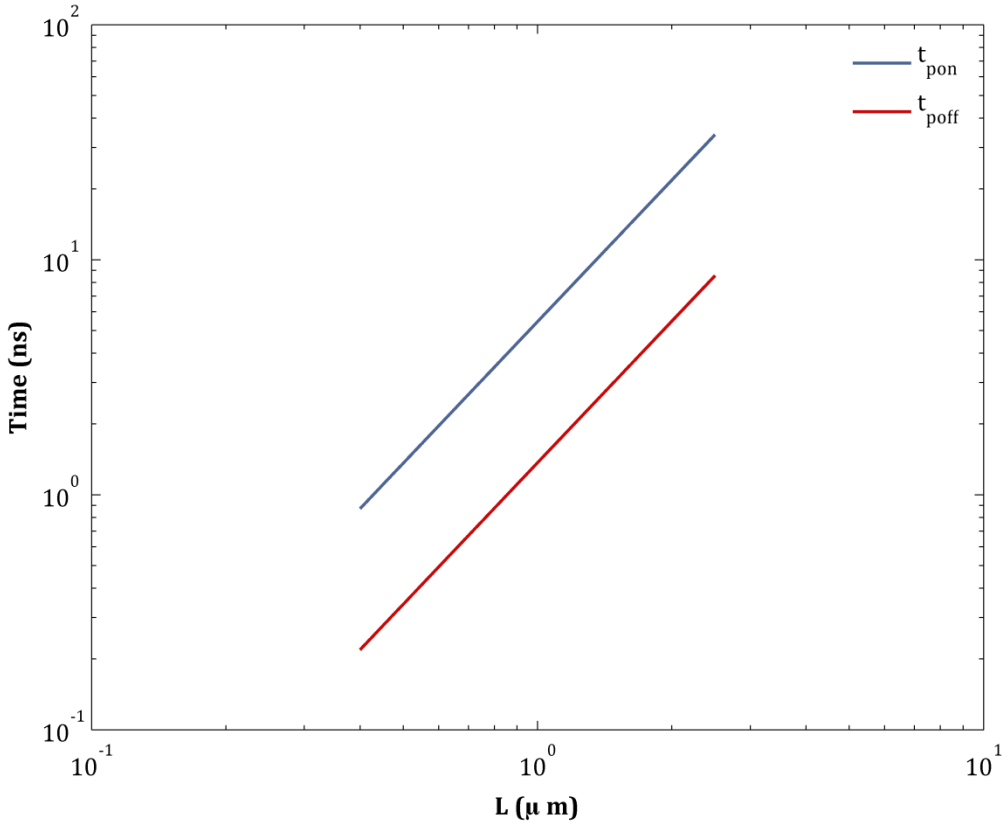

Figure 3.7: L vs. $t_{p,on}$ and $t_{p,off}$ for $V_{DD}=1.5V_{pi}$

The delay expressions above have accounted for just the mechanical delay of the relay while ignoring the electrical delay associated with the relay resistance and any load capacitance. This is generally a fair assumption because the electrical time constant is in the ps range while the mechanical delay is in the ns range [3].

3.3 Digital CMOS vs. Relay Logic

With the understanding of the NEM relay device, the rest of this chapter reviews using these devices in order to design logic circuits [3]. In particular, a brief overview of the techniques required to construct optimized relay-based logic circuits will first be provided. Then, as mentioned at the beginning of this chapter, a 32-bit relay-based adder will be described as a critical building block.

Before moving on, it is important to note that a relay can be turned on by applying a positive or negative V_{gb} beyond V_{pi}, i.e. the relay turns on for $|V_{gb}| \geq V_{pi}$. Thus, the same relay can be operated as an "NMOS" or a "PMOS" transistor by biasing the body node at 0 or V_{DD}, respectively. This is depicted in Figure 3.8.

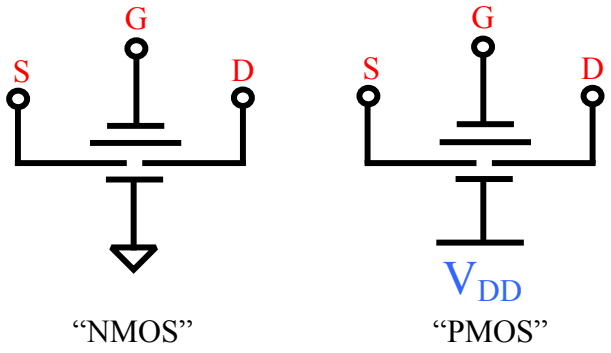

Figure 3.8: Relay switch as an NMOS and PMOS

Optimized relay logic can be now designed by understanding that the relay's mechanical motion dominates its delay. Due to this dominance, unlike traditional

CMOS logic design in which gates are cascaded to construct more complex functions, an optimized relay design arranges for all mechanical motion to occur simultaneously. In other words, each cascaded relay gate on a given path would incur an additional mechanical delay[6] and thus, relay-based designs should instead use a single, large complex gate to implement logic. Figure 3.9 shows an example:

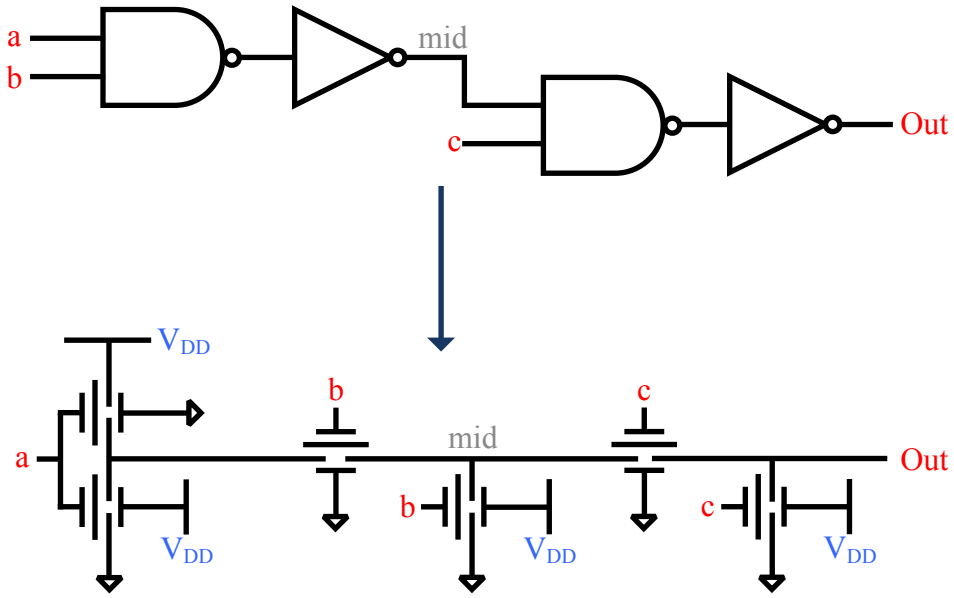

Figure 3.9: Logic gate implementation using relays

Although this design style ensures minimal mechanical delay, each extra series relay increases the electrical delay quadratically because both the path resistance and the

[6] A relay gate is the same as a CMOS gate—i.e., cascaded relay gates refers to a connection between the drain/source of one relay to the gate of a second relay.

capacitance increase. Although many series relays are required for the electrical delay to approach the mechanical delay, it needs to be taken into consideration should the design become very complex.

3.4 Relay Adder

3.4.1 Design and Operation

With knowledge of the issues involved in designing relay-based logic, the design of a relay-based full adder is now analyzed and will then be benchmarked relative to a CMOS adder. Table 3.1 below briefly reviews the truth table of the 1-bit full-adder.

	A	B	C_{in}	S	C_{out}
Kill	0	0	0	0	0
	0	0	1	1	0
Propagate	0	1	0	1	0
	0	1	1	0	1
	1	0	0	1	0
	1	0	1	0	1
Generate	1	1	0	0	1
	1	1	1	1	1

Table 3.1: 1-bit full-adder truth table

With the kill (K), propagate (P), and generate (G) signals defined as highlighted in the table, the carry-out signal, C_{out} can be expressed as:

$$C_{out} = G + PC_{in} \tag{3.13}$$

Here $G = AND(A, B)$ and $P = XOR(A, B)$. Since the XOR function is true when A≠B, the propagate signal can be implemented using a single relay, e.g. by placing A on the gate terminal and B on the body terminal or vice versa. This ensures that the relay will turn on only when the polarity of A is opposite to that of B. Similarly, the sum signal, S can be expressed as:

$$S = \begin{cases} C_{in} & if\ A = B \\ \overline{C_{in}} & if\ A \neq B \end{cases} \tag{3.14}$$

Both of these cases can be implemented as XOR functions: $XOR(A, \overline{B})$ for the former, and $XOR(A, B)$ for the latter. Figure 3.10 below shows the complete full-adder design.

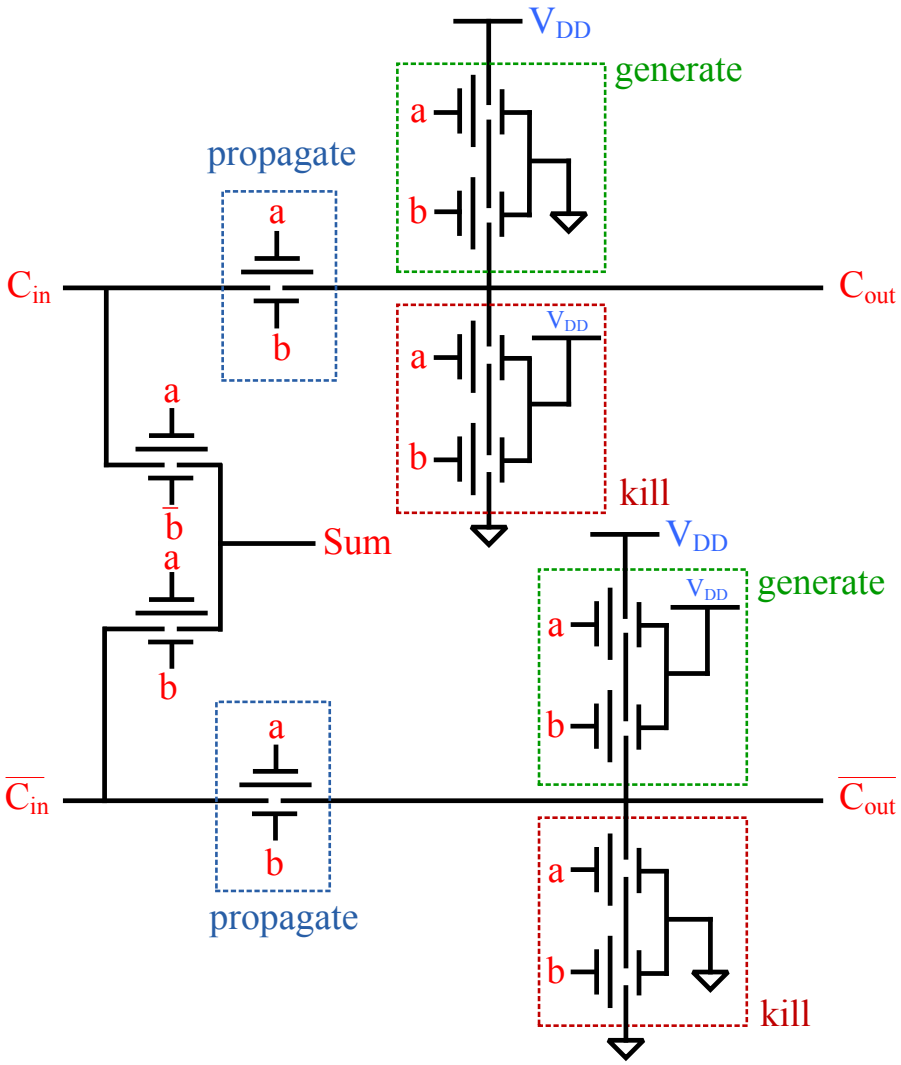

Figure 3.10: 1-bit full-adder cell using relays

This full-adder cell can be used in a ripple-carry configuration to implement a 32-bit adder as a single compound gate with a total delay of $\sim 1t_{p,on}$ [3].[7]

[7] Assuming that all inputs (a_i and b_i) arrive together, all mechanical motion occurs simultaneously, which makes the overall delay $1t_{p,on}$.

3.4.2 Comparison with CMOS Adder

To benchmark the relay adder, a 32-bit CMOS Sklansky adder will be used [7]. As shown in [7], the Sklansky adder is the most energy efficient across a broad range of performance due to its small number of wires and minimum logic depth. Figure 3.11 plots the energy-throughput tradeoffs for the CMOS and relay adders; the results for the relay adder are from [3] and the results for the CMOS adder are from [7].

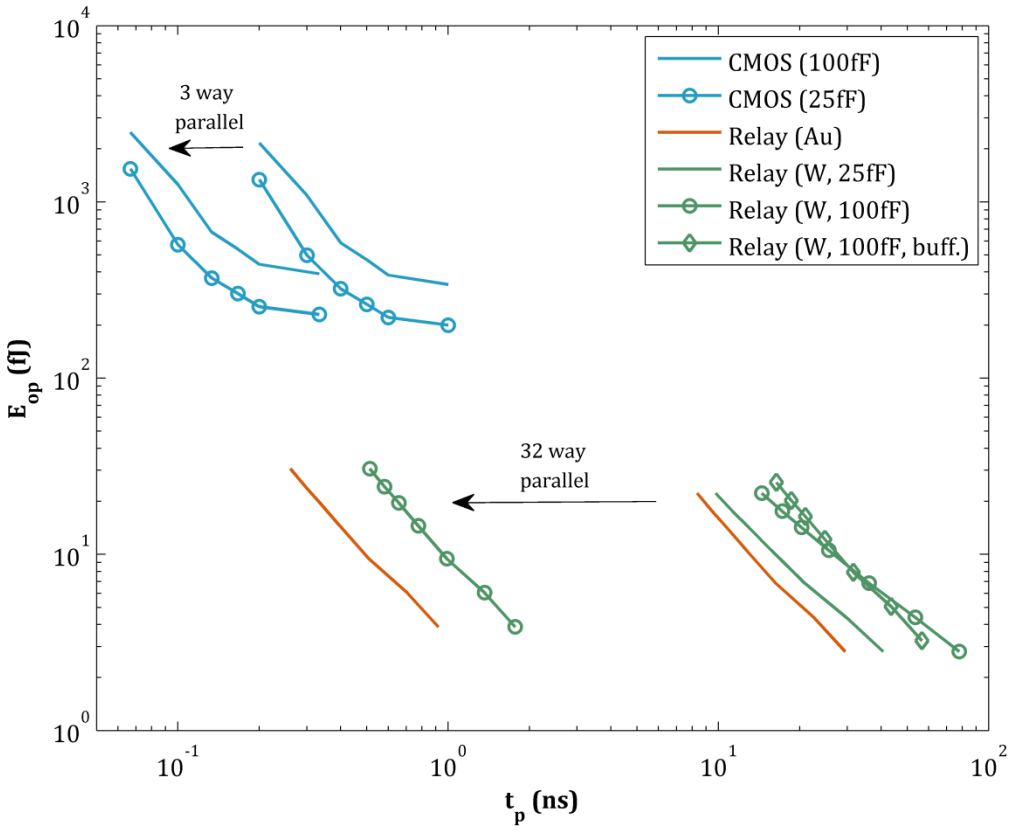

Figure 3.11: Energy-throughput tradeoffs between CMOS and relay adders

The adders have been designed to drive a load capacitance (C_L) of 25fF or 100fF. Two types of relays have been used to implement the relay adders: gold (Au)-based and tungsten (W)-based relays. Due to the low contact resistance of Au (R_{cs}=1Ω), there is hardly any effect on delay due to C_L for the Au-based relays while W-based relays (R_{cs}=1kΩ) do encounter added delay. Thus, when driving a large C_L, adding a buffer stage to the W-based adder improves the energy efficiency. The CMOS adder reaches its minimum energy at ~1ns while the relay-based adder, even at 20ns, is able to achieve ~10x energy efficiency improvements within the same area as the CMOS adder [3]. Furthermore, by using parallelism and increasing the area of the relay adder, this improvement can be extended to higher throughputs.

This area-throughput tradeoff appears in Figure 3.12, which plots the area overhead of relay-based adders over the CMOS adder as a function of throughput. All the curves for this plot are from the results in [3] and maintain a constant energy-efficiency improvement of ~10x as the relay-based adders are targeted for an E_{op} of 20fJ. Comparing the Au relay adder to the 100fF CMOS adder, an area overhead of 5x results in relay adders operating up to throughputs of ~770MOPS. The throughput decreases for the same area overhead in the case of W relay adder due to its higher contact resistance thus increasing the delay. Lastly, beyond 1GOPS, the CMOS adder also needs parallelization causing the area overhead to become constant (i.e. now both CMOS and relay adders are parallelized).

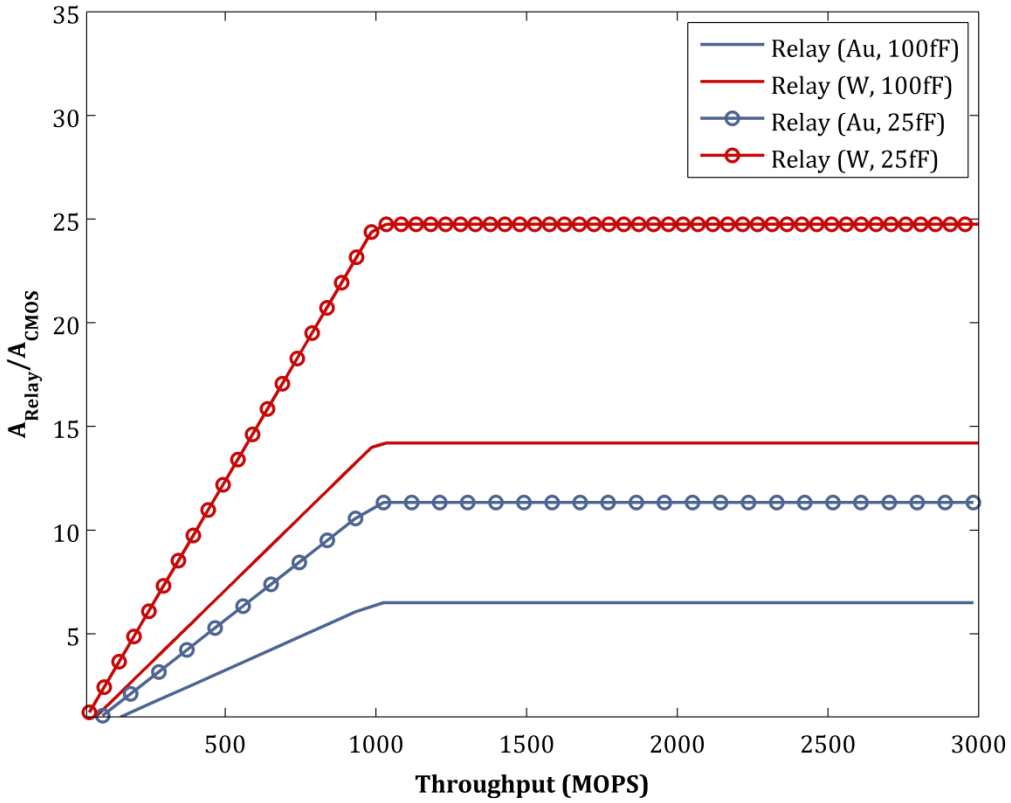

Figure 3.12: Throughput vs. area overhead at a constant energy-efficiency (~10x)

With such energy-efficiency improvements at low throughputs and without a large area overhead at high throughputs, relay-based logic circuits appear capable of significant potential benefits over CMOS.

However, unlike a 1-bit CMOS full-adder that uses 24 transistors in its design allowing for logic optimization and implementing the same functionality through only 12 relays, CMOS memories are already highly dense with very few transistors.

Therefore, the increased area of a single device is significantly mitigated in the overall design of logic blocks, but becomes a larger concern in the design of memory structures. This is because the few number of transistors in a CMOS memory cell limits optimization techniques that aim to reduce the number of relay devices while maintaining the cell's functionality. Thus, to ensure that the advantages of relay-based logic blocks remain for another key component of a system, memory, the next chapter examines relay-based memory structures.

4

Relay Memories

4.1 Motivation for Memories

A complete digital system requires logic building blocks for computation, dense memories to store the results, and I/O for communication. The last chapter analyzed the tradeoffs for relay-based logic blocks; the discussion in [3] analyzes the tradeoffs for I/O by implementing mixed-signal building blocks. As mentioned in the previous section and in addition to being a key component, memory structures are also of a significant interest due to the area-overhead challenges encountered while designing them using relay technology. This is because the increased area of a single relay device was offset by a reduced total number of devices within a logic block, but

the already few number of transistors within a CMOS memory cell make it harder to balance this area overhead. This issue is exacerbated as density is a much more vital metric for memory structures than it is for logic blocks. Therefore, the remainder of this thesis examines techniques to implement relay-based memories and their area-energy-throughput tradeoffs while benchmarking them to 6T CMOS SRAMs.

Before proceeding to relay memory design, the next section briefly reviews the 6T CMOS SRAM structure and presents the delay, energy, and area models to be used later for comparisons with the relay-based structures.

4.2 6T CMOS SRAM

4.2.1 SRAM Cell Delay and Energy Model

Figure 4.1 below displays the standard 6T CMOS SRAM cell and its Elmore delay model [1]. The widths of the transistors (access, pull-up, and pull-down) are based on 90nm technology with cell area being $1.80\mu m \times 0.69\mu m$ as in [8].

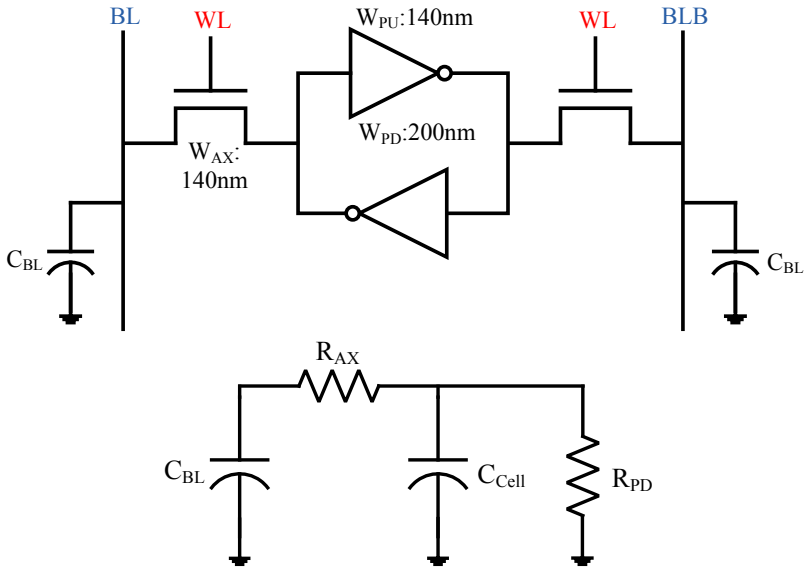

Figure 4.1: 6T CMOS SRAM cell (above) and its Elmore delay model (below)

In the Elmore model, $R_{AX} = R_{sqn}(L/W_{AX})$ and $R_{PD} = R_{sqn}(L/W_{PD})$ are the equivalent "on" resistances of the access and pull-down transistors, respectively (R_{sqn} is the normalized "on" resistance of an NMOS, i.e. an NMOS with W/L=1). $C_{Cell} = C_D(W_{AX}) + C_G(W_{PU} + W_{PD}) + C_D(W_{PU} + W_{PD})$ is the capacitance of a single cell where C_D is the diffusion capacitance per unit width and C_G is the gate capacitance per unit width. Assuming an NxN SRAM array, $C_{BL} = NC_D W_{AX} + C_{BL,wire}$ where $C_{BL,wire} = C_{pp} W_{wire} NL_{cell} + 2C_{fr} NL_{cell}$ is the wire capacitance on the bitline (BL) and C_{pp} is the parallel plate capacitance per unit area, C_{fr} the fringe capacitance per unit length, W_{wire} the wire width, and L_{cell} the "y-dimension" of a single SRAM cell. With the Elmore model, the delay can be expressed as in equation 4.1 below:

$$\tau = (R_{AX} + R_{PD})C_{BL} + R_{PD}C_{Cell} \quad (4.1)$$

Assuming that the SRAM array employs sense-amplifiers to reduce the swing on the BLs to V_{swing}, the SRAM cell delay becomes:

$$t_{p,cell} = ln(2)\left(\frac{V_{swing}}{V_{DD}/2}\right)\tau \quad (4.2)$$

The dynamic read energy of the SRAM array includes a component from the low swinging BLs and from the full-swinging wordlines (WL):

$$E_{read,dyn} = N\left[(C_{BL} + C_{Cell})V_{Swing}\,V_{DD} + C_{WL}V_{DD}^2\right] \quad (4.3)$$

Here, $C_{WL} = 2C_G W_{AX} + \left(C_{pp}W_{wire}\,W_{cell} + 2C_{fr}W_{cell}\right)$ is the wordline capacitance of a single cell where W_{cell} is its "x-dimension". The dynamic write energy of the SRAM array is similar to the read case except that the BLs also swing full rail:

$$E_{write,dyn} = N\left[\left(C_{BL} + \frac{C_{cell}}{2}\right)V_{DD}^2 + C_{WL}V_{DD}^2\right] \quad (4.4)$$

The total read and write energies of an SRAM array can often be dominated by the leakage energy, so it must be taken into account and added to equations 4.3 and 4.4:

$$E_{leak} = N^2\left(V_{DD}I_{leak,cell}\right)\cdot t_{op} \quad (4.5)$$

Here, $I_{leak,cell}$ is the leakage current of a single SRAM cell and t_{op} is the operation frequency of the overall SRAM structure determined by its total delay[8].

[8] The total delay is composed of the SRAM cell delay, the decoder delay, and the delays of any peripheral circuitry (e.g. sense-amplifiers).

4.2.2 Decoder and Sense-Amp Delays

Figure 4.2 shows the basic SRAM decoder structure for N=128. The initial stage pre-decodes 3 or 4 inputs and their results are sent to the final decoding stage. For pre-decoding done near the bottom of the SRAM array and final decoding done in front of the appropriate WLs, C_{par} in the figure is the capacitance on the wire connecting the two stages. C_L in the figure equals C_{WL} as mentioned in the previous section:

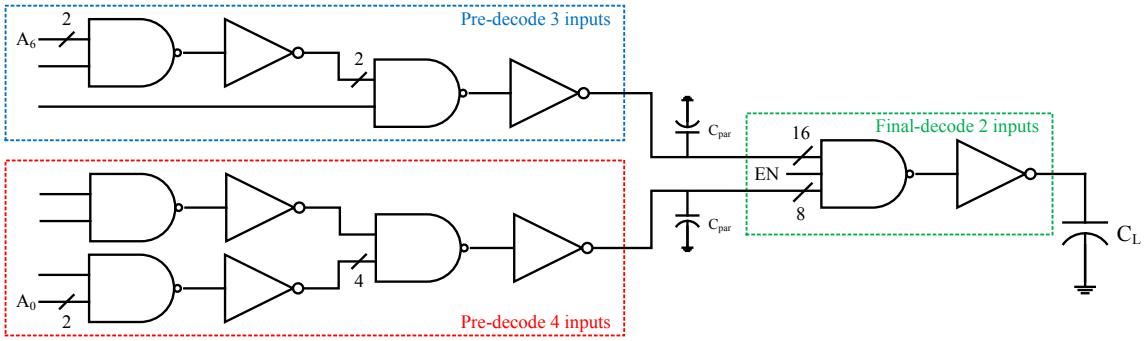

Figure 4.2: The decoder structure for N=128

Assuming that inputs A_0-A_6 can each drive C_{in} up to 5fF and by using the logical effort method to size each gate [1], the delay across the decoder equals:

$$t_{p,dec} = 33.789 t_{inv} \tag{4.6}$$

Here $t_{inv} = 3ln(2)LR_{sqn}C_G$ is the intrinsic delay of an unloaded inverter for $\frac{W_p}{W_n} = 2$.

The sense-amp delay is the time required to take the low swinging BLs and resolving it to full-rail making it a function of V_{swing}. It has been approximated to be:

$$t_{p,SA} = 3t_{p,FO4} \tag{4.7}$$

Here $t_{p,FO4} = t_{inv}(\gamma + 4)$ is the delay across an inverter driving four copies of itself (i.e. the fanout of 4 delay) where $\gamma = C_D/C_G$.

The delay for the complete SRAM structure is the summation of equations 4.2, 4.6, and 4.7. For simplicity, the read and write energies and the area of the decoder and sense-amp will not be taken into account and only equations 4.3 and 4.4 (in summation with equation 4.5) will be used for comparisons with the relay-based memories. The next section will now introduce a three-relay memory cell (3R MC) design followed by a two-relay memory cell (2R MC) design. It should also be noted from Figure 3.7 that $t_{p,off}$ was ~4x smaller than $t_{p,on}$, so relay-based memories should attempt to incorporate $t_{p,off}$ in place of $t_{p,on}$ in order to increase throughput.

4.3 3R DRAM Design & Operation

4.3.1 Overall Structure

With density being a critical metric for memory cells, the number of relay devices within a cell must be minimized. Since the write operation requires the charging or discharging of a cell's internal node, a "pass-gate relay" can be used. As mentioned earlier, in order to increase throughput, the read operation can be implemented

using a NAND flash topology—i.e., by "stacking" the memory cells and making sure that transistors switch off when accessed [1]. Figure 4.3 displays a 3R MC design:

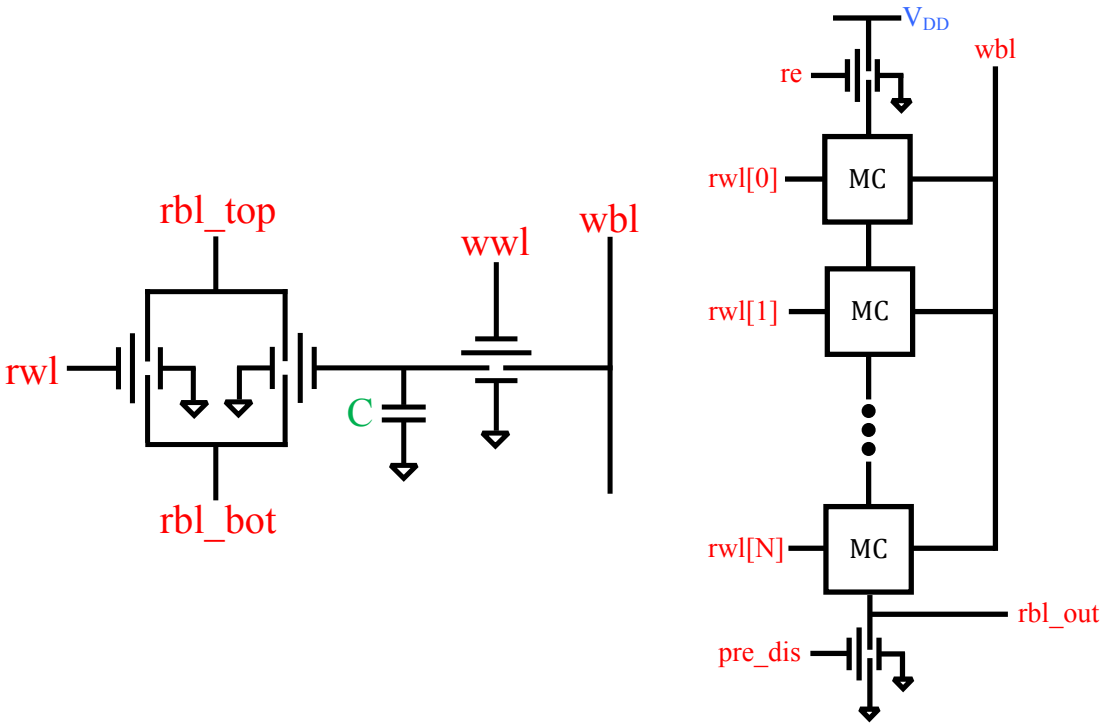

Figure 4.3: 3R memory cell (left) and column configuration (right)

Note that the memory cell has a DRAM-like functionality as opposed to an SRAM one, which is acceptable due to the zero leakage current of relay devices[9]. The sections that follow examine the 3R MC's read and write operations followed by its area, throughput, and read and write energy calculations.

[9] CMOS-based DRAMs need periodic refreshing even if the cells are not accessed due to the higher leakage currents of CMOS transistors that discharge a cell's internal node. Since relay devices exhibit zero leakage, once a cell has been written, it seems as if it is "statically" storing its value.

4.3.2 Write Operation

To understand the write operation of the 3R MC, only the right half of the cell, as highlighted in Figure 4.4 below, needs to be examined.

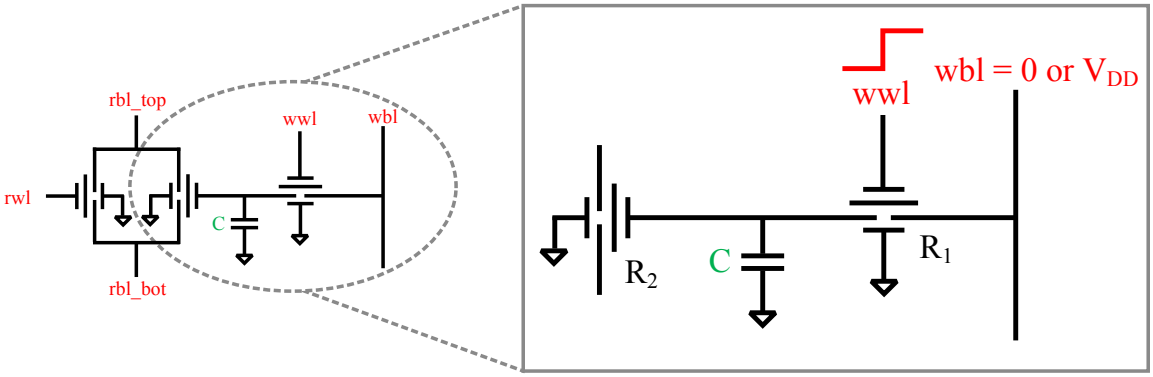

Figure 4.4: Write operation with a 3R MC

The write operation is similar to writing a conventional CMOS DRAM cell, and the following steps are taken:

1. Write bitline (wbl) is driven to 0V (writing a "0") or to V_{DD} (writing a "1").
2. Write wordline (wwl) is pulsed high (i.e. from 0V → V_{DD}).

When wwl is pulsed, the "write" relay (R_1) turns on and the capacitor, $C = C_{gb,R2} + C_{db,R1} \approx C_{gb,R2}$, either charges, discharges, or remains unchanged depending on the previous value on C. Depending on if C charges, discharges, or remains unchanged, R_2 will then turn on, off, or maintain its previous state, respectively.

4.3.3 Read Operation

To understand the read operation, only the left half of the cell and a single column, as highlighted in Figure 4.5, need analysis (for simplicity, only two cells are shown).

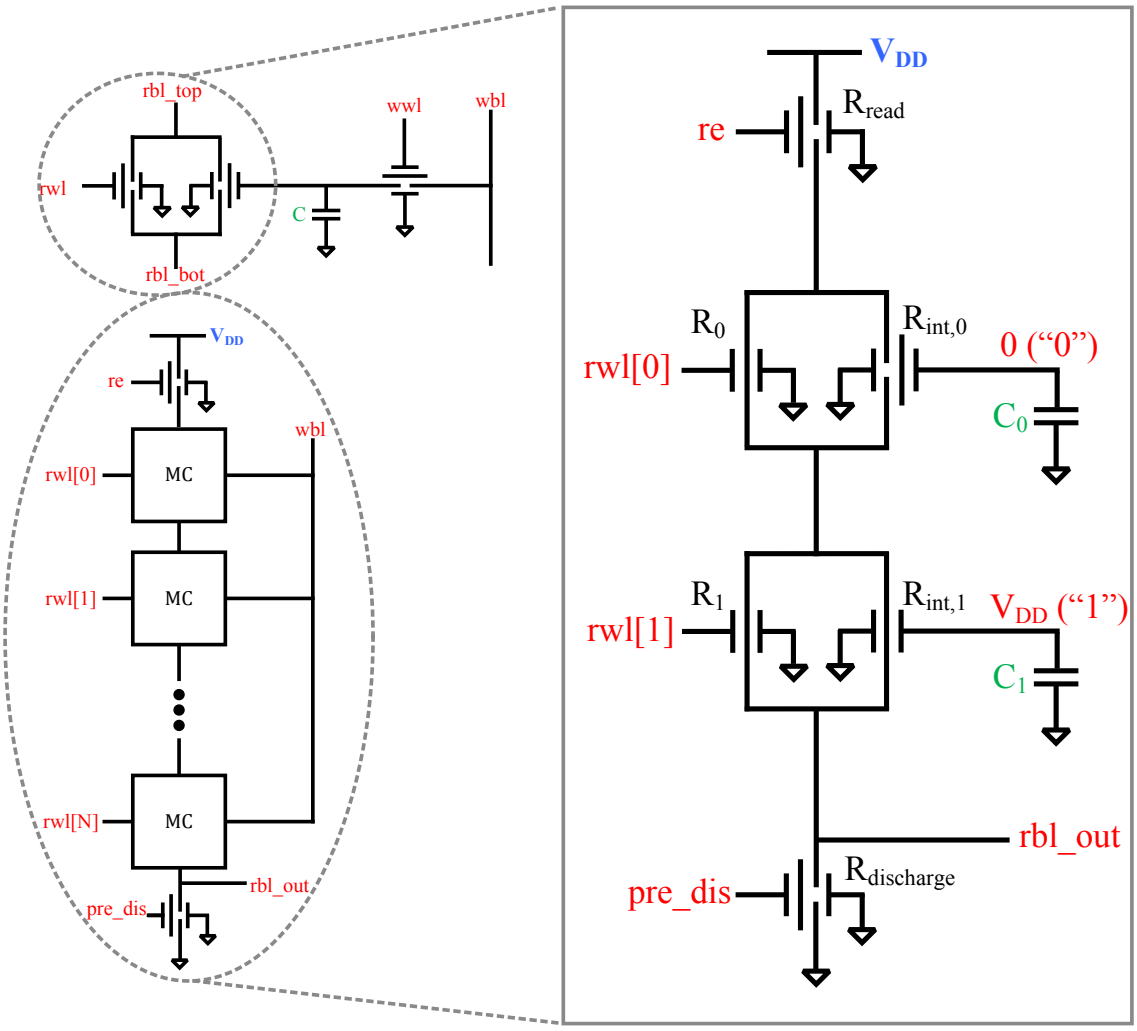

Figure 4.5: Read operation with a 3R MC

Note that for all cells storing a 0, their "internal" relays will be off ($R_{int,0}$) while all cells storing a 1 will have their "internal" relays on ($R_{int,1}$). Remembering that every attempt should be made to incorporate $t_{p,off}$ in place of $t_{p,on}$, the steps that appear below outline the read operation and can be followed using Figure 4.6.[10]

1. All read wordlines (rwl) are set high (i.e. at V_{DD}).
2. The pre-discharge signal (pre_dis) is pulsed high turning on $R_{discharge}$ and discharging rbl_out to 0V.[11]
3. Once discharged, pre_dis is driven back to 0V turning off $R_{discharge}$ making rbl_out float.
4. The selected rwl goes low (from $V_{DD} \rightarrow 0V$) turning off the appropriate "read" relay (R_0 or R_1 in this case).
5. The read signal (re) is pulsed high turning on R_{read}.

In reading from a cell storing a 0, when rwl goes low turning off the appropriate "read" relay (rwl[0] and relay R_0 in this case), there is no path from V_{DD} to rbl_out once R_{read} turns on. As can be seen from the left half of Figure 4.6, this makes rbl_out stay at its discharged value of 0V. However, in reading from a cell storing a 1, when rwl goes low turning off the appropriate "read" relay (rwl[1] and relay R_1 in this case), there is a path from V_{DD} to rbl_out once R_{read} turns on. This makes rbl_out

[10] To conduct the simulation of Figure 4.6, the verilogA model as developed in [3] was used.

[11] During the simulation of figure 4.6, rbl_out was initially made to float at 0.5V, so the discharging effect could be seen.

charge up to V_{DD}, which can be seen in the right half of Figure 4.6. Note that the turning off of the appropriate read relays (as opposed to turning them on during access) enables for increases in throughputs due to $t_{p,off}$ being ~4x smaller than $t_{p,on}$.

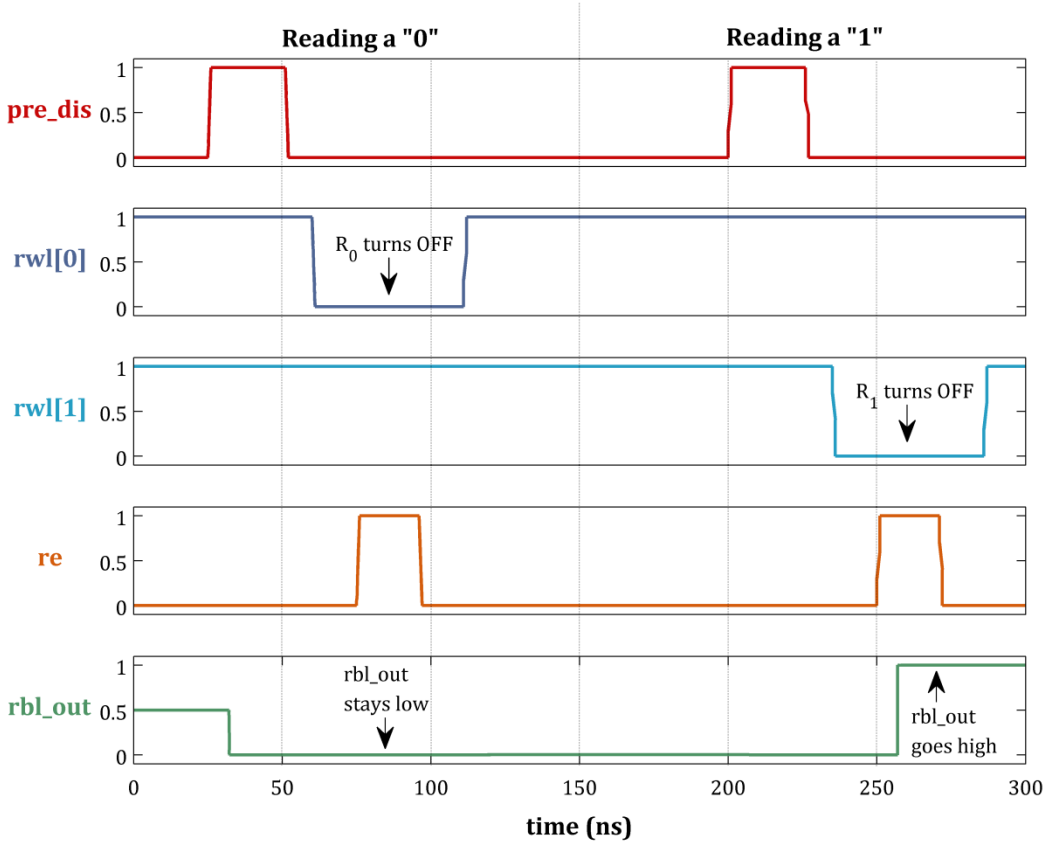

Figure 4.6: 3R MC read simulation for reading a 0 (left) and reading a 1 (right)

By understanding the read and write operations of the 3R MC, the next section moves on to analyze the cell's area, throughput, and read and write energies.

4.3.4 Area, Throughput, and Read & Write Energies

To find the dimensions for estimating area, the stick diagram of Figure 4.7 is used that accounts for appropriate spacing between traces and vias[12].

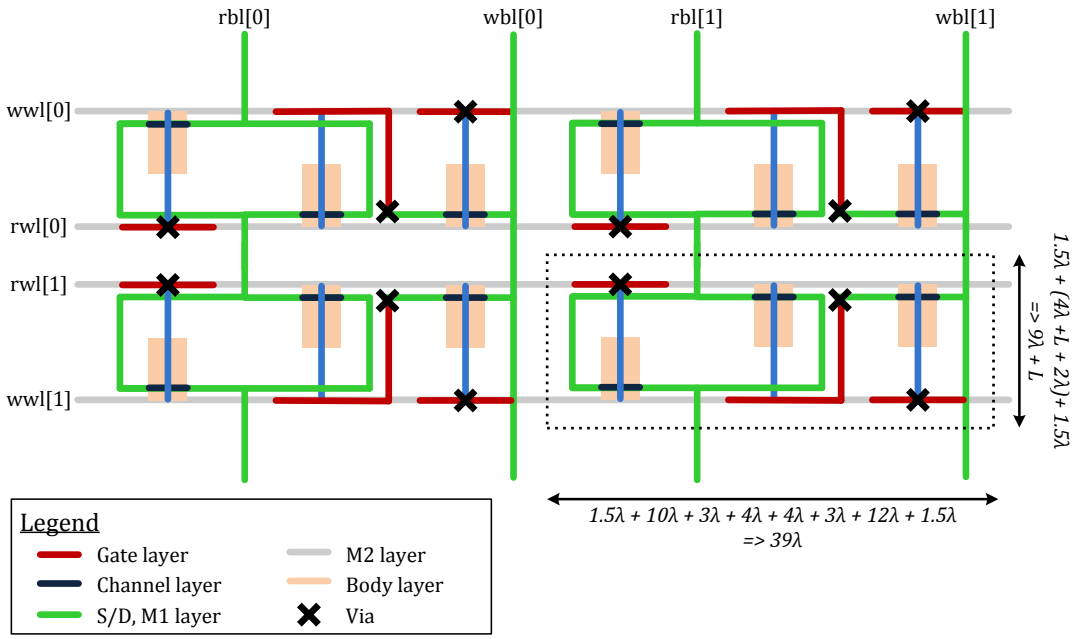

Figure 4.7: Stick diagram for estimating area

The area of a cell depends on the relay's length (L) and width (W). For $\lambda = W/2$,

$$Area_{3R} = 39\lambda \times (9\lambda + L) \tag{4.8}$$

Like the CMOS SRAM, the read energy includes a WL and a BL component. For an NxN array and assuming that the BL swings full rail[13], it can be expressed as:

[12] For an explanation on how the spacing was evaluated, refer to Appendix B: Relay Dimensions.

[13] Although a sense-amp (SA) can lower read energy, it would also lower throughput by introducing an extra delay component. Thus, for simplicity, the discussion of an SA is held off until section 4.6.

$$E_{read\,,3R} = N[(C_{WL} + C_{BL})V_{DD}^2] \tag{4.9}$$

The wordline capacitance per cell, C_{WL}, is made up of a "read" relay and wire capacitance and equals: $C_{WL} = C_{gb} + \left(C_{pp}\,W_{wire}\,W_{cell} + 2C_{fr}\,W_{cell}\right)$ with $W_{cell} = 39\lambda$. Finding an expression for the bitline capacitance is a bit more challenging as the total switched capacitance depends on the value of the cell being accessed as well as the exact wordline that it resides in. In other words, if the accessed cell stores a 1, the entire bitline capacitance switches, but if storing a 0 and with the cell being at the top of the array, a lot less capacitance switches as opposed to switching a cell at the bottom. With this "fragmented" capacitance and assuming that a cell stores a 0 or 1 with a probability of 1/2, the average switched capacitance on the BL becomes:

$$C_{BL} = \frac{1}{2N}\left\{NC_{BL,tot} + \sum_{i=0}^{N-1}\left[3C_{db} + i\left(4C_{db} + C_{wire\,-cell\,,BL}\right)\right]\right\} \tag{4.10}$$

$C_{BL,tot} = 3C_{db} + (N-1)4C_{db} + NC_{wire\,-cell\,,BL} + 3C_{db}$ is the total BL capacitance that switches when the accessed cell stores a 1. The second term is the "fragmented" capacitance that switches when the accessed cell stores a 0. Here $C_{wire\,-cell\,,BL} = C_{pp}\,W_{wire}\,L_{cell} + 2C_{fr}\,L_{cell}$ with $L_{cell} = 9\lambda + L$ and is the BL wire capacitance per cell.

Similarly, the write energy also has a wordline and a write bitline component as well as a component originating from the "internal" relay that sometimes switches (relay R_2 in Figure 4.4). The wordline component is the same as in the read case. To help reduce the write energy, the write bitlines can be held to their

previous values; by assuming that half of the cells retain their previous values and of the other half that do change, half will be written to 0 and the other half to 1, the write energy can be expressed as:

$$E_{write\ ,3R} = N\left[\left(C_{WL} + \frac{1}{4}C_{BL,write} + \frac{1}{4}C_{internal}\right)V_{DD}^2\right] \tag{4.11}$$

Here $C_{BL,write} = NC_{db} + C_{pp}W_{wire}NL_{cell} + 2C_{fr}NL_{cell}$ and $C_{intern\ al} = C_{gb}$.

In a back-to-back read operation, the throughput is set by one $t_{p,on}$ and one $t_{p,off}$; $t_{p,on}$ arises from the decoder needing to access the appropriate read wordline and $t_{p,off}$ arises from the "read" relay (R_0 and R_1 in Figure 4.5). In a back-to-back write operation, the throughput is set by two $t_{p,on}$'s: the first is due to the decoder and the second due to the "write" relay (R_1 in Figure 4.4)[14]. However, due to the large C_{BL}, the electrical delay of the read and write bitlines ($t_{p,RBL-3R}$ and $t_{p,WBL-3R}$) must also be taken into account. Using the Elmore delay model, this is the product of the bitline resistance[15] and capacitance. Thus, the throughput equals:

$$\begin{aligned} Freq_{3R,read} &= \frac{1}{\left(t_{p,on} + t_{p,off} + t_{p,RBL-3R}\right)} \\ Freq_{3R,write} &= \frac{1}{\left(2t_{p,on} + t_{p,WBL-3R}\right)} \end{aligned} \tag{4.12}$$

[14] Note that the pre_dis signal can be overlapped with the decoder evaluation—i.e., pre_dis can go high while the decoder is evaluating the WL that needs to be accessed.

[15] The bitline resistance is composed of the relay "on" resistance [3] and the wire resistance.

Although the 3R MC achieves memory functionality, its density is based upon the use of three relays. More specifically, the use of two relays to implement the NAND-like read operation causes an increased area overhead, which can be mitigated through the reduction of one relay. Since the write operation requires a pass-gate relay and the NAND topology at least one more relay[16], the crucial factor of density in memory design sets the motivation to design a 2R memory cell.

[16] As mentioned earlier, the NAND flash topology is desirable as it allows for the read access delay to be set by $t_{p,off}$ as opposed to $t_{p,on}$.

4.4 2R DRAM Design & Operation

4.4.1 Overall Structure

The column configuration for the 2R MC design remains the same as in Figure 4.3 and its cell structure appears in Figure 4.8 below.

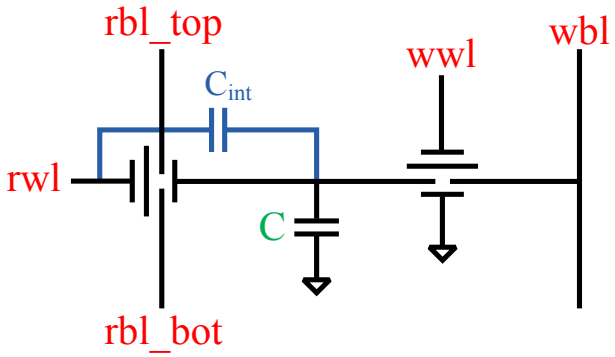

Figure 4.8: 2R memory cell design

Other than using one less relay, the main change in this structure is the use of an extrinsic capacitor, C, which unlike the 3R version was only composed of the intrinsic C_{gb} and C_{db} of the relays (C_{int} in Figure 4.8 refers to C_{gb} of the relay on the left). An extrinsic C is needed because unlike in the 3R design, the read relay and the cell's internal storage relay (relays R_0 and $R_{int,0}$ in Figure 4.5) are the same. This means that when the read relay is accessed, the "coupling" capacitor C_{int} will create a glitch on the cell's internal node. The complete details of this are presented in the next few sections that describe the write operation followed by the read operation.

4.4.2 Write Operation

Figure 4.9 can be used to examine the write operation of the 2R MC. Note that during the write operation, the read wordline of a cell (rwl) is set to 0V, and thus C and C_{int} appear in parallel.

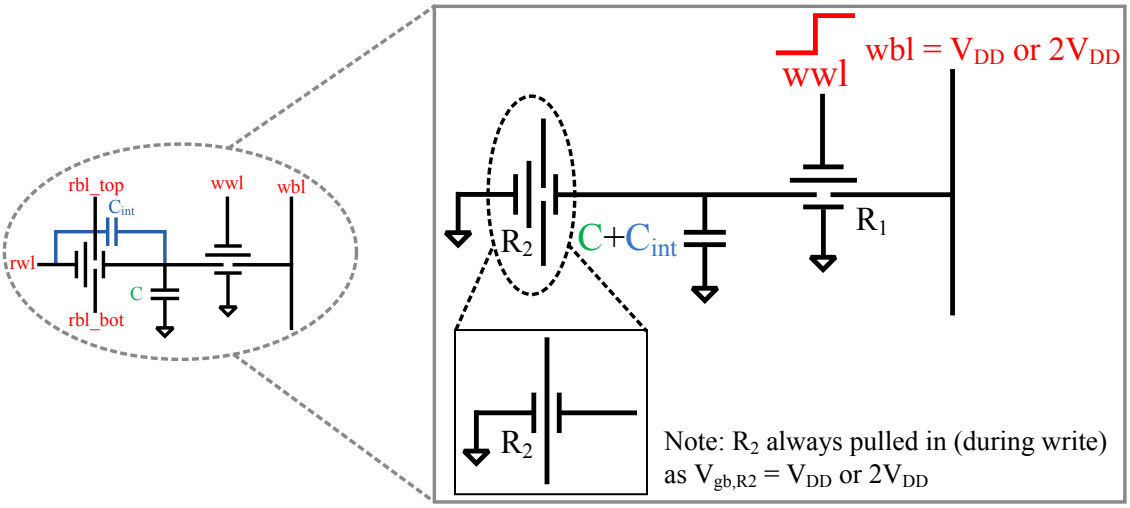

Figure 4.9: Write operation with the 2R MC

The steps involved in writing are similar to the 3R MC, but with two key differences:

1. wbl is set to V_{DD} in order to write a "0" (as opposed to 0V in the 3R version) or to $2V_{DD}$ in order to write a "1 (as opposed to V_{DD} in the 3R version).
2. Unlike in the 3R structure in which R_2 is turned on, off, or stays in the same state depending on the behavior of C, R_2 here always remains on. This is because its V_{gb} is always larger than V_{pi} (V_{DD} or $2V_{DD}$).

4.4.3 Read Operation

As before, to understand the read operation of the 2R MC, only the left half of the cell and a single column, as highlighted in Figure 4.10, need examination.

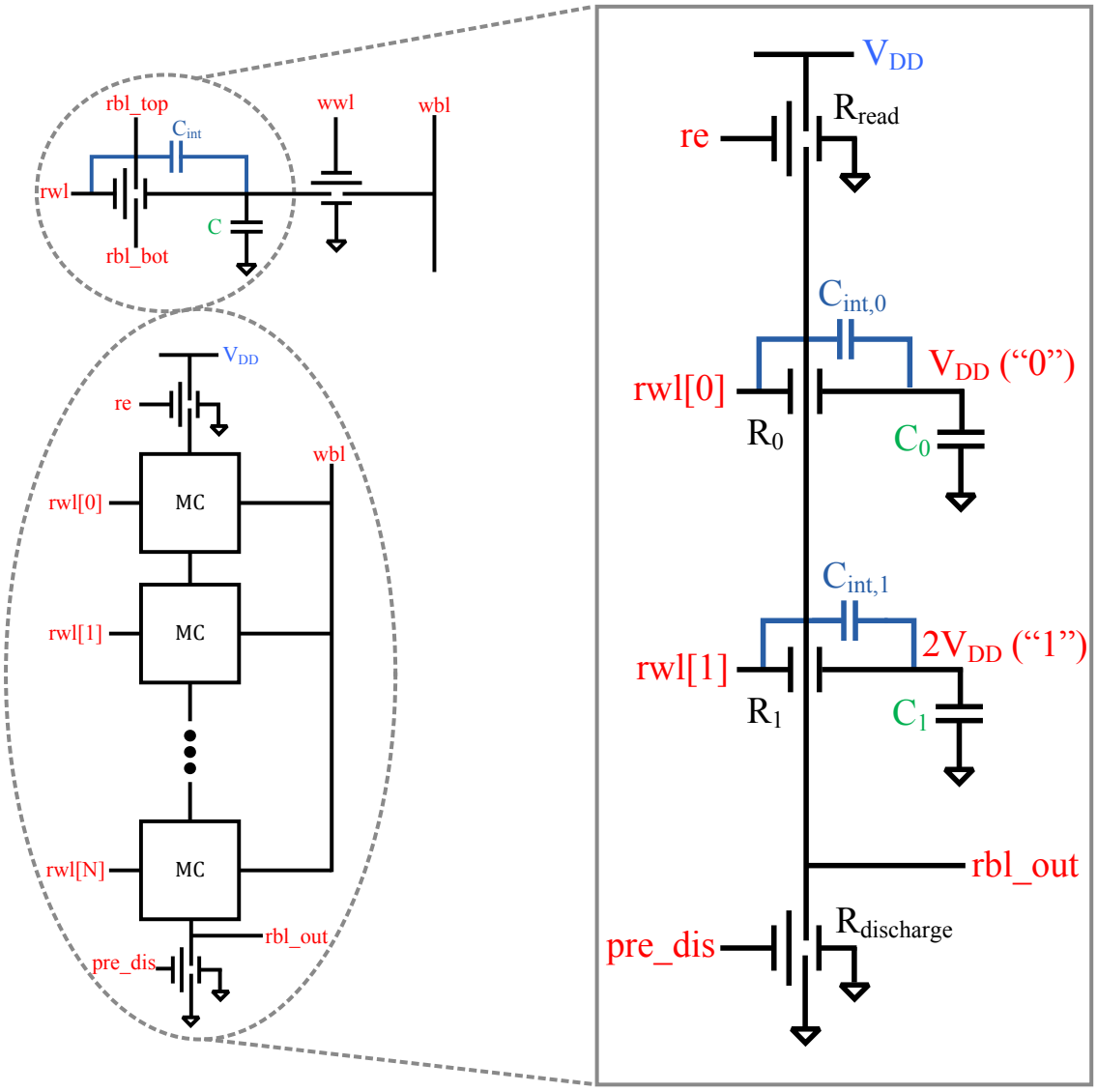

Figure 4.10: Read operation with the 2R MC

The read operation is very similar to the 3R case except for two key differences:

1. All read wordlines are initially low unlike step 1 of the 3R case in which they were all high. This keeps all the "read" relays on (R_0 and R_1 in this case).
2. The selected rwl is driven to V_{DD} (step 4 of the 3R case drove it to 0V).

The steps involved in reading from the 2R MC can be followed using Figure 4.11 below.[17] In the case of reading from a cell storing a 0 and assuming that the extrinsic capacitance is much larger than the intrinsic one ($C_0 >> C_{int,0}$ in this case), when the selected rwl is driven to V_{DD}, the "read" relay turns off (rwl[0] and R_0). There is no path from V_{DD} to rbl_out keeping it discharged when the R_{read} relay turns on as can be seen from the left half of Figure 4.11. When reading from a cell storing a 1 and regardless of the ratio between the extrinsic and intrinsic capacitors, when the selected rwl is driven to V_{DD}, the "read" relay remains on (rwl[1] and R_1). There is a path from V_{DD} to rbl_out and it is charged to V_{DD} once R_{read} relay turns on as shown in the right half of Figure 4.11.

[17] As done for Figure 4.6, the simulation of Figure 4.11 was done using the verilogA model developed in [3].

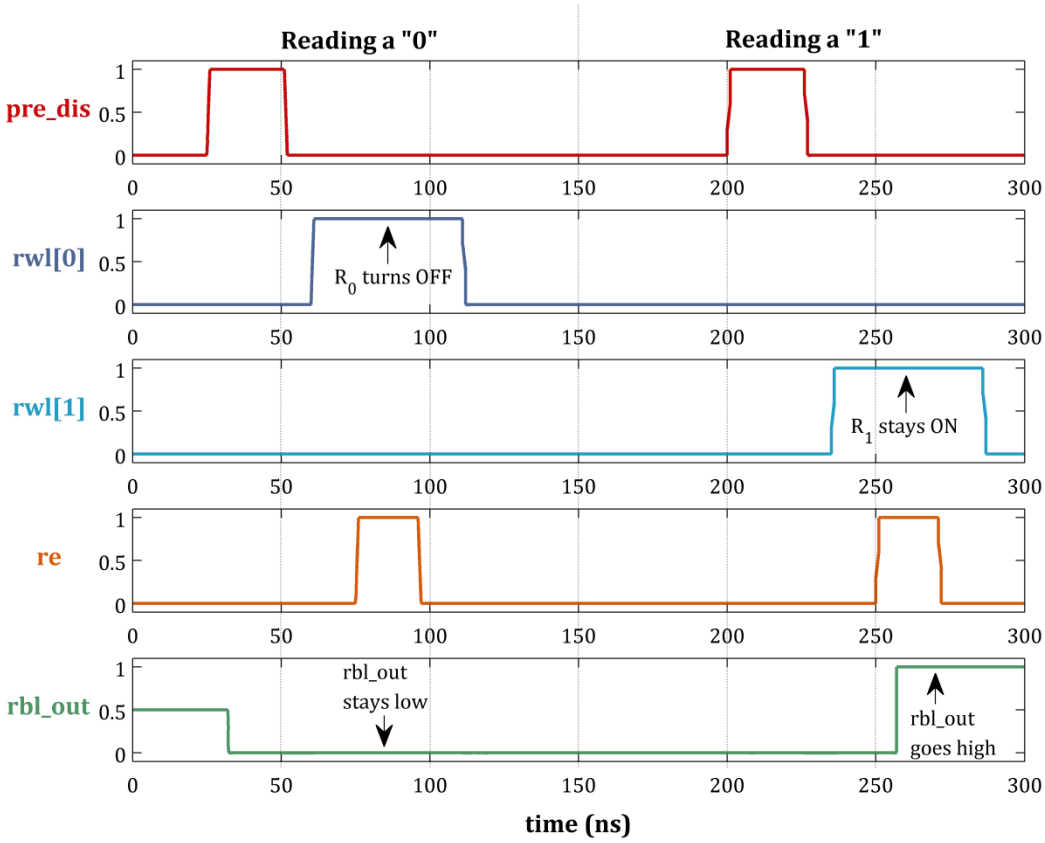

Figure 4.11: 2R MC read simulation for reading a 0 (left) and reading a 1 (right)

As mentioned at the beginning of this section, during the read operation of the 2R MC, the cell's internal node will experience a glitch due to the coupling between the read wordline and the internal node via C_{int}. Figure 4.12 shows this glitch for some extrinsic capacitor value, displaying that the internal node rises by Δ from its initial voltage V_{init}, when a rising step is placed on rwl. Without the extrinsic capacitor C—i.e., for a completely floating internal node—a rising step on rwl will

completely propagate to the cell's internal node causing V_{gb} across the read relay to remain unchanged. Since the read relay is initially on, reading a 1 is successful even without an extrinsic C because the read relay will remain on due to this unchanged V_{gb}. However, in reading a 0, the lack of an extrinsic C keeps the read relay on instead of turning it off, causing an improper operation. Thus, the ability to read a 0 successfully, sets a lower bound on the value of C—i.e., when reading from a cell storing a 0 and to be able to turn the "read" relay off, its body node, in the worst case, can only swing by V_{po}.[18] Utilizing the capacitive divider function at the body node, the value for the extrinsic capacitor evaluates to:

$$C_{min} = C_{gb}\left(\frac{V_{DD}}{V_{po}} - 1\right) \tag{4.13}$$

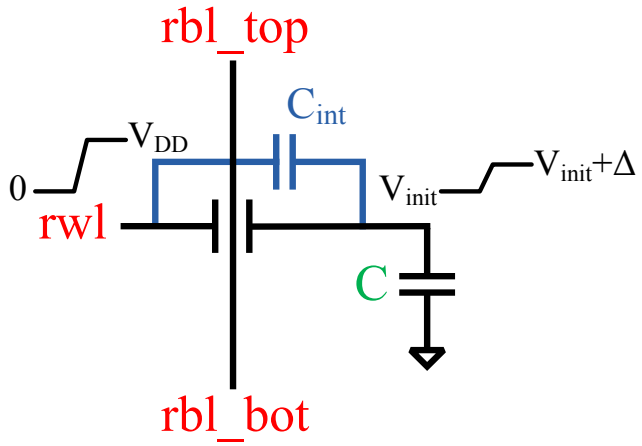

Figure 4.12: 2R MC internal node glitch

[18] Since $V_{gb,final} \leq V_{po}$ when reading from a cell storing a 0 and due to $V_{gb,initial} = V_{DD}$, the maximum swing allowed on the body node is $\Delta V_{B,max} = V_{po}$.

4.4.4 Area, Throughput, and Read & Write Energies

As done for the 3R case, the stick diagram of Figure 4.13 for the 2R design provides the dimensions and allows estimating its area.

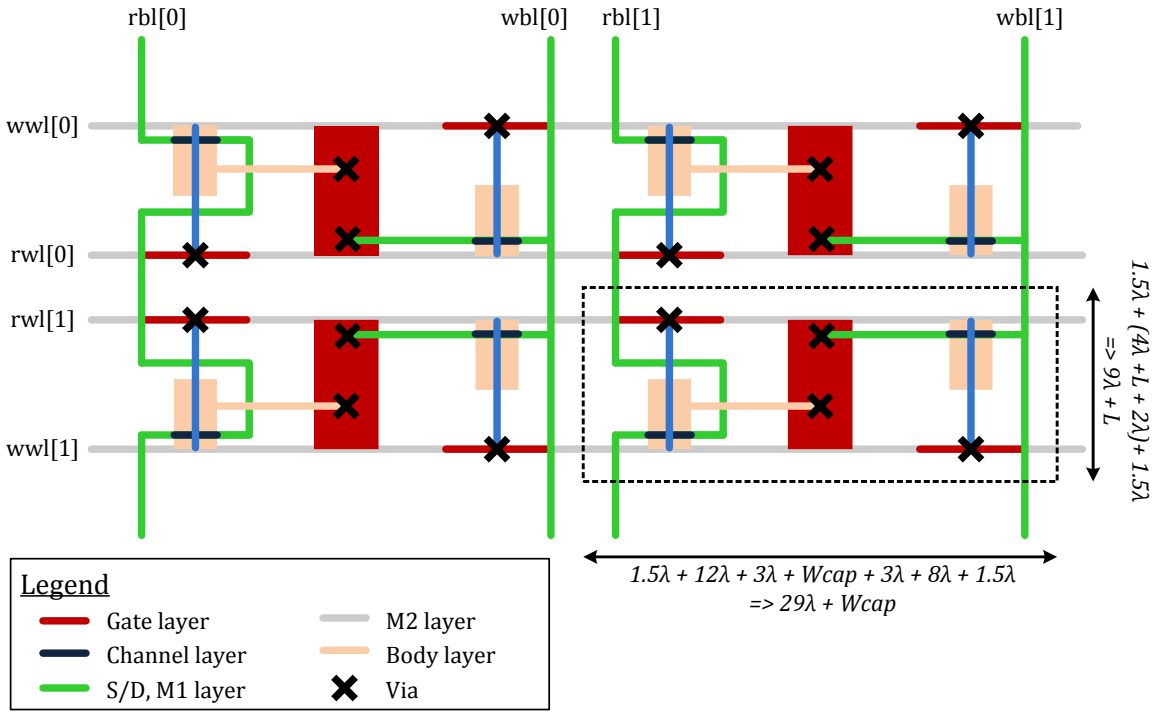

Figure 4.13: Stick diagram for estimating area

The area of the cell can be expressed as:

$$Area_{2R} = (29\lambda + W_{cap}) \times (9\lambda + L) \tag{4.14}$$

Here $W_{cap} = C_{min} / (C_{p-p,cap} \cdot L)$ refers to the width of the extrinsic capacitor, C_{min}, with the material used to build it having a capacitance per unit area of $C_{p\text{-}p,cap}$.

The read and write throughputs are similar to the 3R design except that one less relay reduces the read BL electrical delay ($t_{p,RBL}$) and the parallel combination of C_{min} and C_{int} (Figure 4.9) slightly increases the write BL electrical delay ($t_{p,WBL}$).

$$Freq_{2R,read} = \frac{1}{\left(t_{p,on} + t_{p,off} + t_{p,RBL-2R}\right)}$$
$$Freq_{2R,write} = \frac{1}{\left(2t_{p,on} + t_{p,WBL-2R}\right)} \quad (4.15)$$

The read energy can be expressed as in equation 4.9 (repeated below), but has different WL and BL components (due to C_{min} on rwl and one less relay on rbl).

$$E_{read,2R} = N\left[\left(C_{WL,read} + C_{BL}\right)V_{DD}^2\right] \quad (4.16)$$

Since rwl sees the series combination of C_{int} and C_{min}, $C_{WL,read}$ can be modified to $C_{WL,read} = \left[C_{gb}C_{min}/\left(C_{gb}+C_{min}\right)\right] + \left(C_{pp}W_{wire}W_{cell} + 2C_{fr}W_{cell}\right)$ with W_{cell} being expressed as in Figure 4.13. Much like the 3R cell, C_{BL} can be expressed as:

$$C_{BL} = \frac{1}{2N}\left\{NC_{BL,tot} + \sum_{i=0}^{N-1}\left[2C_{db} + i\left(2C_{db} + C_{wire-cell,BL}\right)\right]\right\} \quad (4.17)$$

Here, $C_{BL,tot} = 2C_{db} + NC_{db} + NC_{wire-cell,BL}$ is the modified total capacitance on the bitline that switches when the accessed cell stores a 1. Since the length of the cell remains unchanged from the 3R MC, $C_{wire\text{-}cell,BL}$ also remains unchanged.

The write energy for the 2R MC has the same expression as equation 4.11. However, expressions for C_{WL} and the "internal" capacitance need modifications due

to a different W_{cell} and added C_{min}, respectively. Since the cell's length has not changed, $C_{BL,write}$ remains unchanged. Assuming that a $2V_{DD}$ supply is available:

$$E_{write\ ,2R} = N\left[\left(C_{WL} + \frac{1}{4}C_{BL,write} + \frac{1}{4}C_{internal}\right)V_{DD}^2\right] \quad (4.18)$$

Since only W_{cell} has changed, C_{WL} maintains the same expression as for the 3R case, but with $W_{cell} = 29\lambda + W_{cap}$. Remembering that during the write operation, C_{min} and C_{int} appear in parallel (Figure 4.9), $C_{internal} = C_{gb} + C_{min}$. Note that C_{WL} in the read and write energy expressions for the 3R case is the same, but this is not the case in the read and write energy expressions for the 2R case. This is again due to the capacitive divider from C_{int} and C_{min} that originates during a read in the 2R MC.

By examining the energies of the 3R and 2R memory cells (read or write), and the expressions for C_{WL} and C_{BL}, it can be seen that the energy component due to the WL capacitance is proportional to N while that due to the BL capacitance is proportional to N^2. This causes the BL component to dominate both the overall read and write energies. Since minimized energy is another key goal of memory design, techniques that reduce this dominating BL component should also be examined.

Since devices with longer beam lengths have a lower V_{pi} and thus require a lower V_{DD} for proper operation, the use of such devices can significantly lower the total energy. However, longer devices imply a highly increased area and a large penalty in terms of density. Thus, another technique that aims to increase the area

overhead just slightly yet provides for a drastic decrease in energy needs to be realized. Understanding that the bitlines need to swing full-rail during the write operation (in order to have a successful write), and do not need to do so during the read operation (in order to have a successful read) sets the motivation for examining a relay-based sense-amplifier. The next section discusses such a design.

4.5 Relay Sense-Amplifier

As mentioned earlier, a sense-amplifier (SA) can lower the read energy by allowing a lower swing on the read BLs and without a significant increase in the overall area of the memory array. However, due to the added delay in resolving the final output voltage, it also lowers the overall throughput. This section analyzes the structure of a relay-based SA and the energy-throughput tradeoffs that it presents.

4.5.1 Overall Structure

Figure 4.14 below displays the relay-based SA design based on the work in [3].

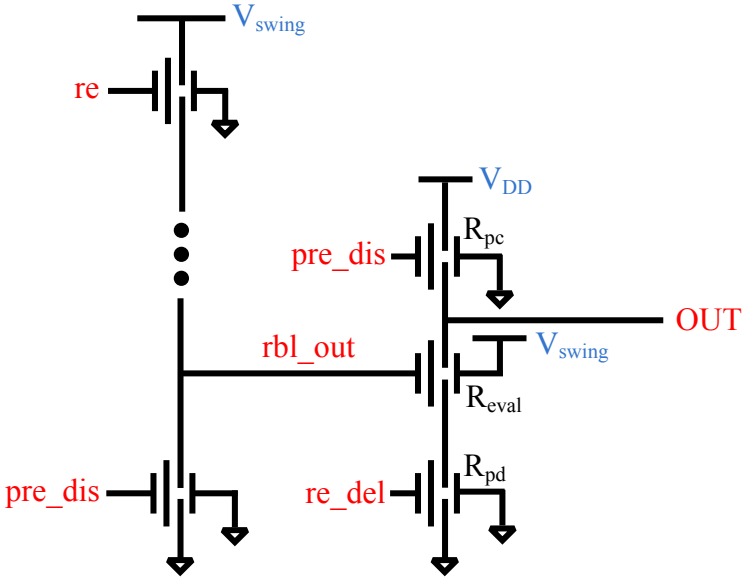

Figure 4.14: Relay-based sense-amplifier design

The left half of Figure 4.14 has the same column configuration as Figure 4.3 except it uses V_{swing} instead of V_{DD}. The three relays in the right half implement the SA.

4.5.2 Read Operation

During the read operation, the left half of Figure 4.14 keeps the same functionality as before, except that rbl_out is driven to a lower voltage, V_{swing}[19] (instead of V_{DD}) when the accessed cell stores a 1. Thus, only the right half of Figure 4.14 needs analysis. The steps below outline the process and are illustrated in Figure 4.15.[20]

1. The pre_dis signal is pulsed high, turning on the "pre-charge" relay R_{pc} in order to charge the out node to V_{DD}. Note that when pre_dis goes high, rbl_out gets discharged to 0V and turns on the "evaluation" relay, R_{eval}[21].
2. The pre_dis signal falls, turning off R_{pc} making out and rbl_out nodes float.
3. The accessed cell is allowed to evaluate, causing rbl_out to either stay at 0V or charge up to V_{swing} once the re signal goes high.
4. The re_del signal, which is a delayed version of the re signal, goes high turning on the "pull-down" relay, R_{pd}.

[19] The R_{eval} relay can operate at this lower V_{swing} by making its length longer, thus keeping its V_{pi} low.

[20] As mentioned earlier, the simulation of Figure 4.15 was done using the verilogA model in [3].

[21] In the simulation of figure 4.15, out and rbl_out were initially made to float so the charging and discharging effects could be seen. Also, V_{swing} was chosen to be 200mV for the simulation.

If the accessed cell stored a 0, rbl_out would remain at 0V keeping R_{eval} turned on. This would cause the out node to fall to 0V since both R_{eval} and R_{pd} would be on as can be seen in the left half of Figure 4.15. But, if the accessed cell stored a 1, rbl_out would charge up to V_{swing}, causing R_{eval} to turn off. This would keep the out node at its pre-charged value of V_{DD} since no path would exist from out to ground as shown in the right half of Figure 4.15. The added delay from this sense-amp structure is thus one extra turn-off delay (originating from R_{eval} having to turn-off).

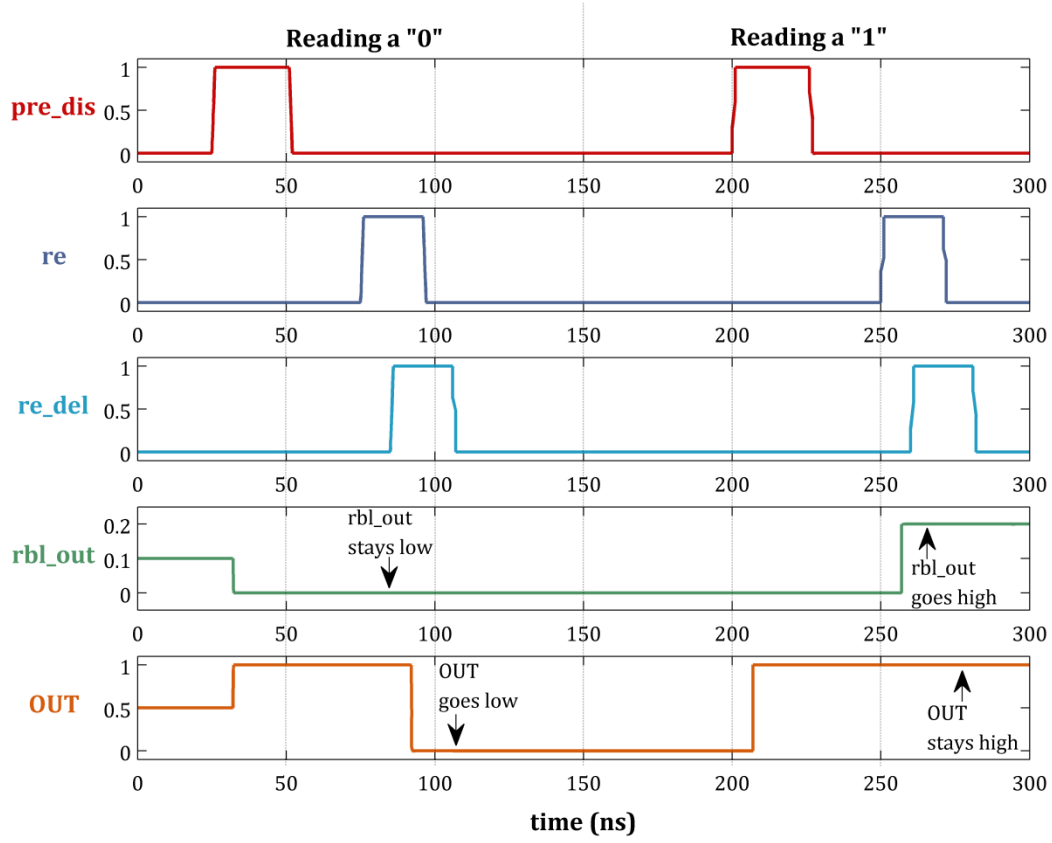

Figure 4.15: SA read simulation for reading a 0 (left) and reading a 1 (right)

4.5.3 Energy-Throughput Tradeoffs

Since only one extra turn-off delay is added to the read throughput, it becomes:

$$Freq_{3R(2R),read} = \frac{1}{\left(t_{p,on} + 2t_{p,off} + t_{p,RBL-3R(2R)}\right)} \tag{4.19}$$

As can be seen, the throughput degrades by $t_{p,off}$. Similarly, the only modification to the read energy comes from the bitlines swinging by the reduced V_{swing} (ignoring the input capacitance of SA). Thus, equations 4.9 and 4.16 can be expressed as:

$$E_{read} = N\left[C_{WL}V_{DD}^2 + C_{BL}V_{swing}^2\right] \tag{4.20}$$

Since there haven't been any design changes within the memory structure itself, the expressions for C_{WL} and C_{BL} also remain unchanged from their 3R and 2R discussions from sections 4.3.4 and 4.4.4, respectively.[22] As was the case with the 6T CMOS SRAM, for simplicity, the energy of the relay-based SA will not be taken into account and only the energy from the memory arrays will be compared.

With this, the analysis of all of the key components within a memory architecture has been completed. Therefore, the next section compares the 6T CMOS SRAM and the 2R and 3R relay DRAMs with and without the sense-amp in terms of the read and write energies (from just the memory array), the area (of just a single cell), and the throughput (taking into account the decoder, array, and SA).

[22] For the 2R design, $C_{WL} = C_{WL,read}$.

5

Results

5.1 Relay Memories: Sense-Amp Tradeoff

Section 4.5 mentioned that the use of an SA can lower the read energy (E_{read}) without much of an increase in the overall area but at the expense of decreasing the read throughout (f_{read}). Thus, before comparing the relay memories to the CMOS SRAM, this tradeoff must be examined. Figure 5.1 displays f_{read} vs. E_{read} for the 3R and 2R MC designs with and without the SA. The 2R design has a lower E_{read} than the 3R design for the same f_{read} due to its lower WL and BL capacitance[23] and as f_{read} decreases, the difference in E_{read} between the SA and non-SA cases also decreases.

[23] C_{WL} is lower due to the capacitive divider between C_{gb} and C_{min} as well as due to the decreased W_{cell} resulting in a lowered WL wire capacitance. C_{BL} is lower due to the one less relay.

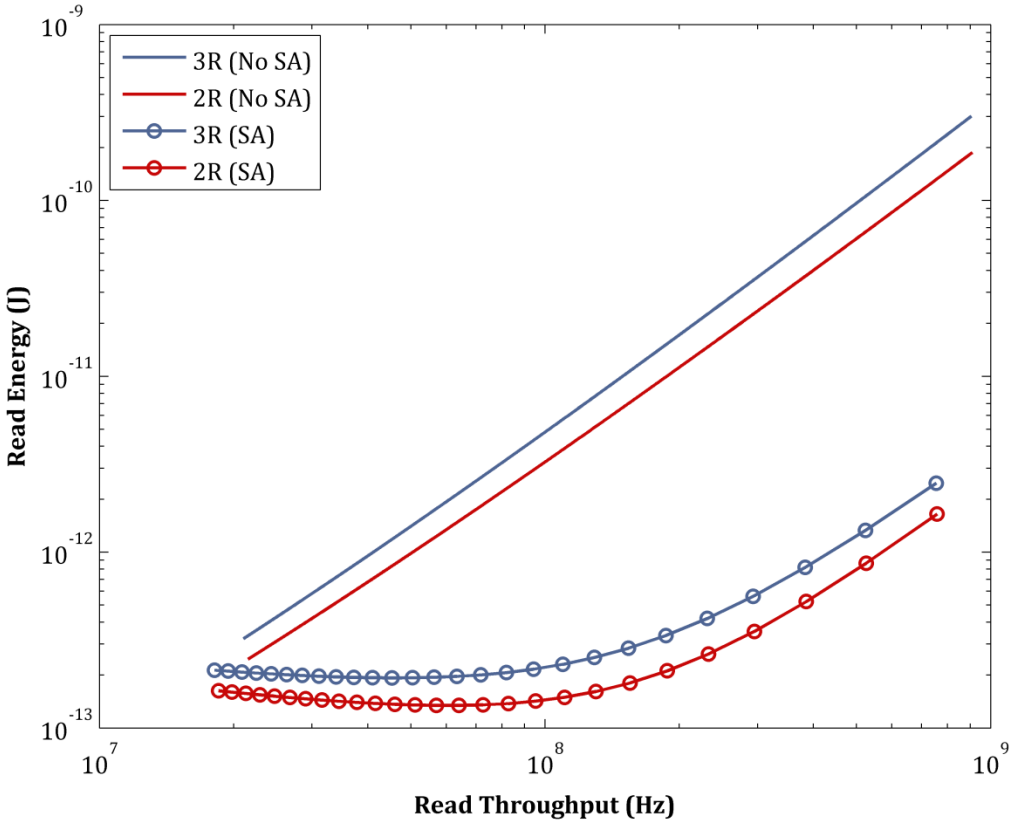

Figure 5.1: Read Throughput vs. Read Energy for SA and non-SA 3R & 2R MCs

To understand the trends of Figure 5.1, E_{read} must be decomposed into its WL and BL components for the SA and non-SA cases. Since the 3R and 2R designs experience the same trends, Figure 5.2 shows this decomposition for the 3R design only.

Since f_{read} and $V_{DD}=1.5V_{pi}$ are both proportional to $1/L^2$ (equations 3.10, 3.11, and 3.7), it means that at low read throughputs, V_{DD} is also low and approaches V_{swing}=200mV. Due to the decreased difference in V_{DD} and V_{swing}, the difference in E_{BL}

for the SA and non-SA cases is also reduced. Although at higher read throughputs, a lower L decreases C_{BL}, the drastic increase in V_{DD} increases E_{BL} for the non-SA case. Since V_{swing} remains constant[24], a reduced C_{BL} lowers E_{BL} for the SA case. Lastly, increasing V_{DD} increases E_{WL} as WLs swing full rail for both cases. This means that the non-SA design is always BL dominated while the SA case is BL dominated for low read throughputs and becomes WL dominated at higher read throughputs.

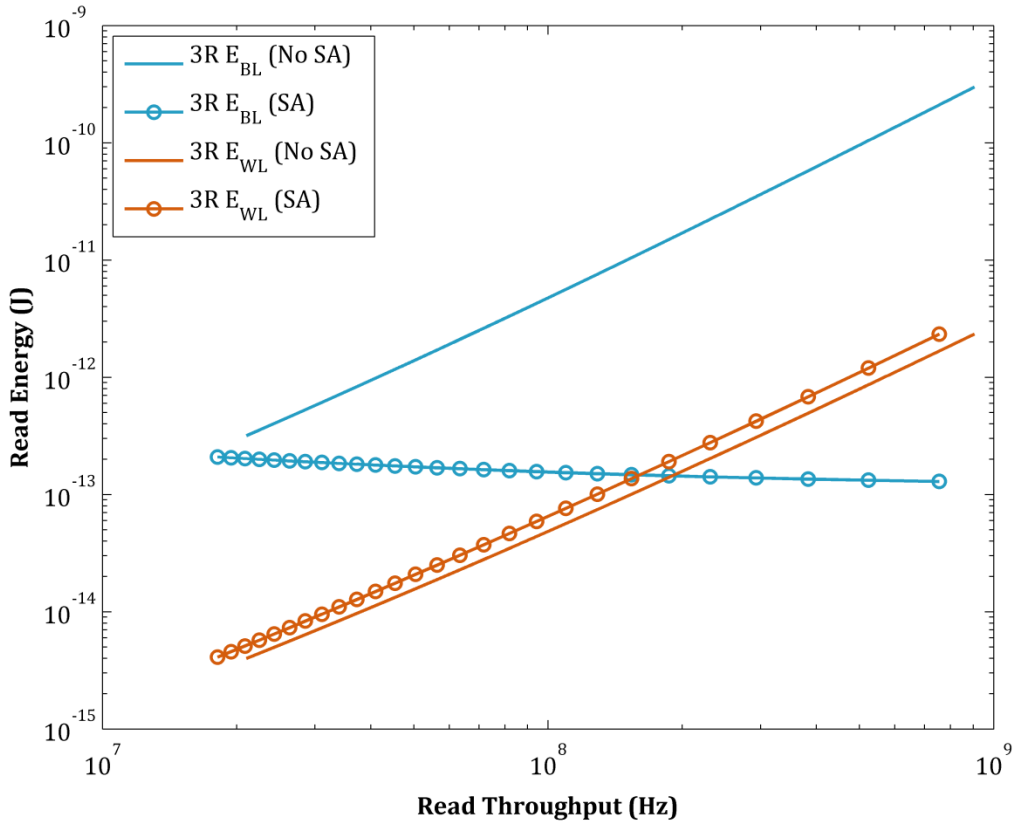

Figure 5.2: Read Throughput vs. BL & WL Read Energy (3R design only)

[24] As mentioned earlier, V_{swing} can remain the same even at higher throughputs by making L of relay R_{eval} in Figure 4.14 longer thus keeping its V_{pi} low.

5.2 CMOS and Relay Memories

Since the read energy is always lower for the same throughput in the SA-based designs, relay memories incorporating the sense amplifier will be the ones used for comparison with the CMOS SRAM.

5.2.1 Area, Read Energy, Read Throughput Tradeoffs

The left half of Figure 5.3 shows the read energy (E_{read}) as a function of area overhead (A_{Relay}/A_{CMOS}) while the right half shows it as a function of f_{read} for the different relay memories and the CMOS SRAM. The read energy of the CMOS SRAM ($E_{read,CMOS}$) reaches its minimum point at throughputs below ~200MHz, while at this same throughput, the 3R design takes ~2/3 of $E_{read,CMOS}$ with an area overhead of ~1.6x and the 2R design takes ~1/3 of $E_{read,CMOS}$ with an area overhead of ~1.3x. For applications with less stringent throughput and density requirements (~100-150MHz), the improvement in energy efficiency offered by relay-based memories can be extended by increasing the area overhead to ~2x.

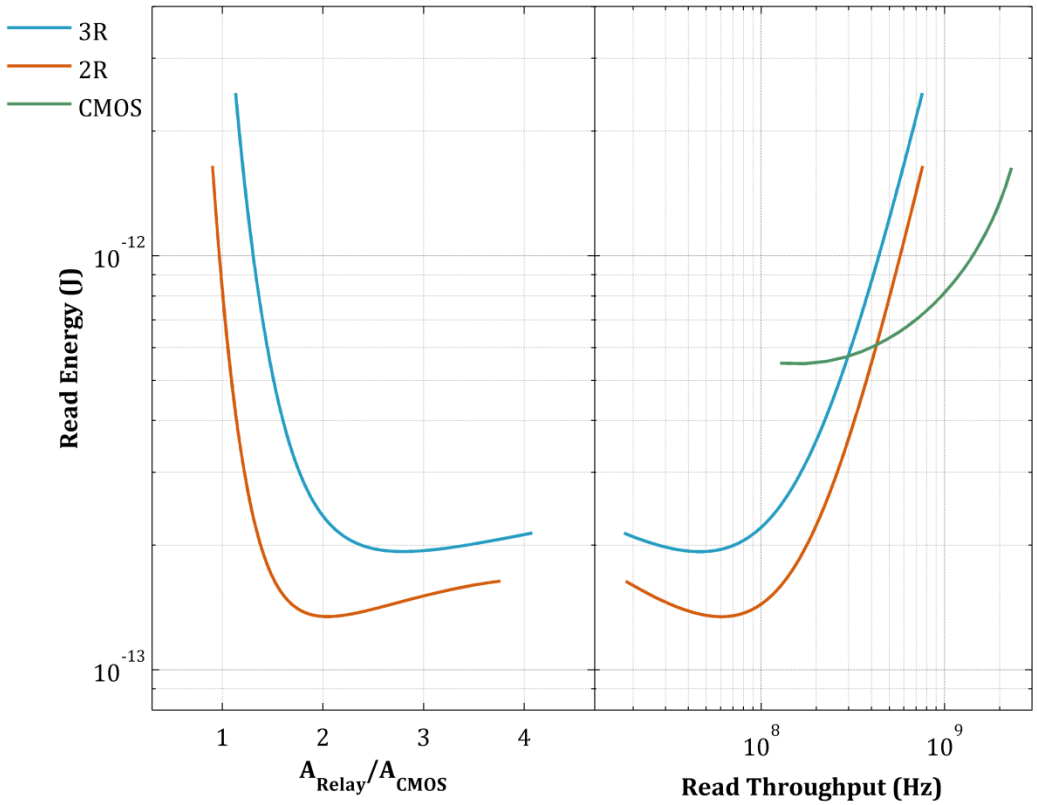

Figure 5.3: Area overhead vs. E_{read} (left) and read throughput vs. E_{read} (right)

To support the throughput analysis of section 4.5.3, Figures 5.4 and 5.5 below both need to be used to demonstrate that the relay memory's maximum frequency of operation as given in equation 4.19 is achievable. Figure 5.4 plots the area overhead as a function of the beam length (left) as well as the read throughput for that length (right). Relay devices with a beam length of 0.5μm yield no area overhead for the 2R MC and an area overhead of 1.25 for the 3R MC. Furthermore,

this length requires $V_{DD}=1.5V_{pi}$ of 6V (according to Figure 3.5) and yields a maximum frequency of operation (read throughput) of 500MHz for both arrays.

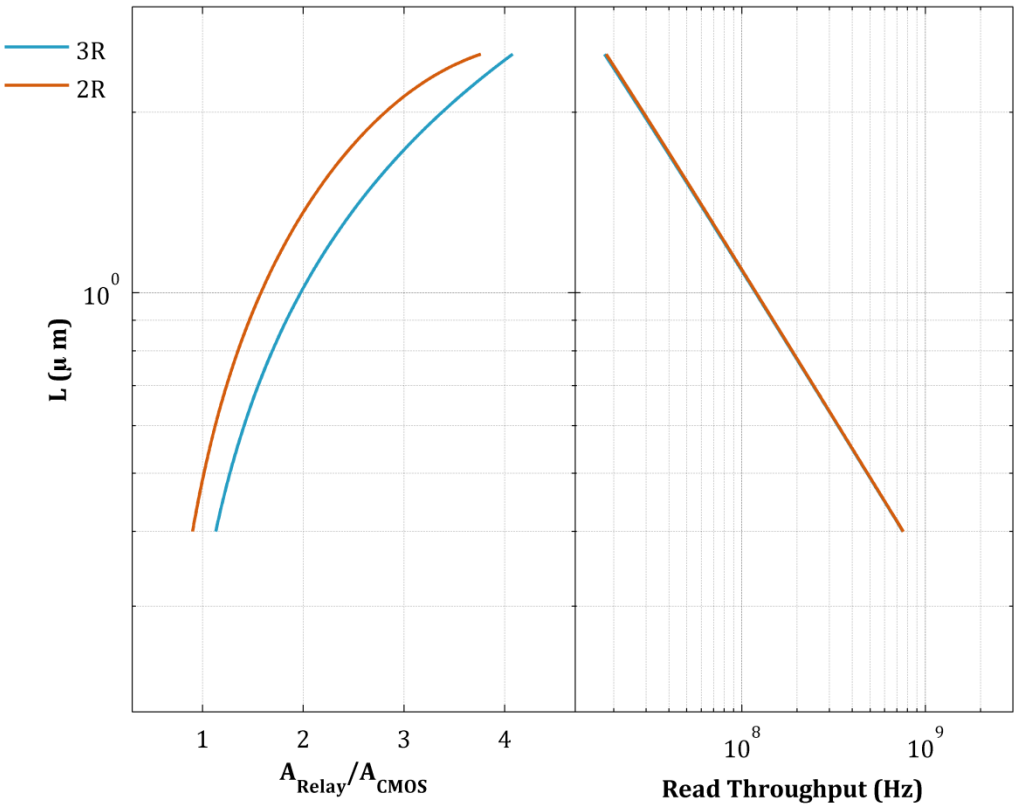

Figure 5.4: L vs. area overhead (left) and L vs. read throughput (right)

Figure 5.5 is a reproduction of Figure 4.15 and uses the verilogA model developed in [3] for its simulation; relay devices with a beam length of 0.5μm are used. First, the pre_dis signal pulses, charging the OUT node to 6V, and discharging the rbl_out node to 0V. This turns the SA's R_{eval} relay on. Next, the appropriate read relay is accessed

after which the re signal pulses. In reading a 0 (left half of Figure 5.5), since there is no path from V_{swing} to rbl_out, rbl_out stays at 0V keeping R_{eval} on; when the re_del signal pulses, OUT discharges to 0V. In reading a 1 (right half of Figure 5.5), there is a path from V_{swing} to rbl_out, causing rbl_out to charge to V_{swing}. This turns off R_{eval} causing OUT to no longer have a path to ground. Thus, it remains at its pre-charged value of 6V when re_del pulses. Since the two read operations occur successfully with a 2ns period, the memory can operate at its maximum frequency of 500MHz.

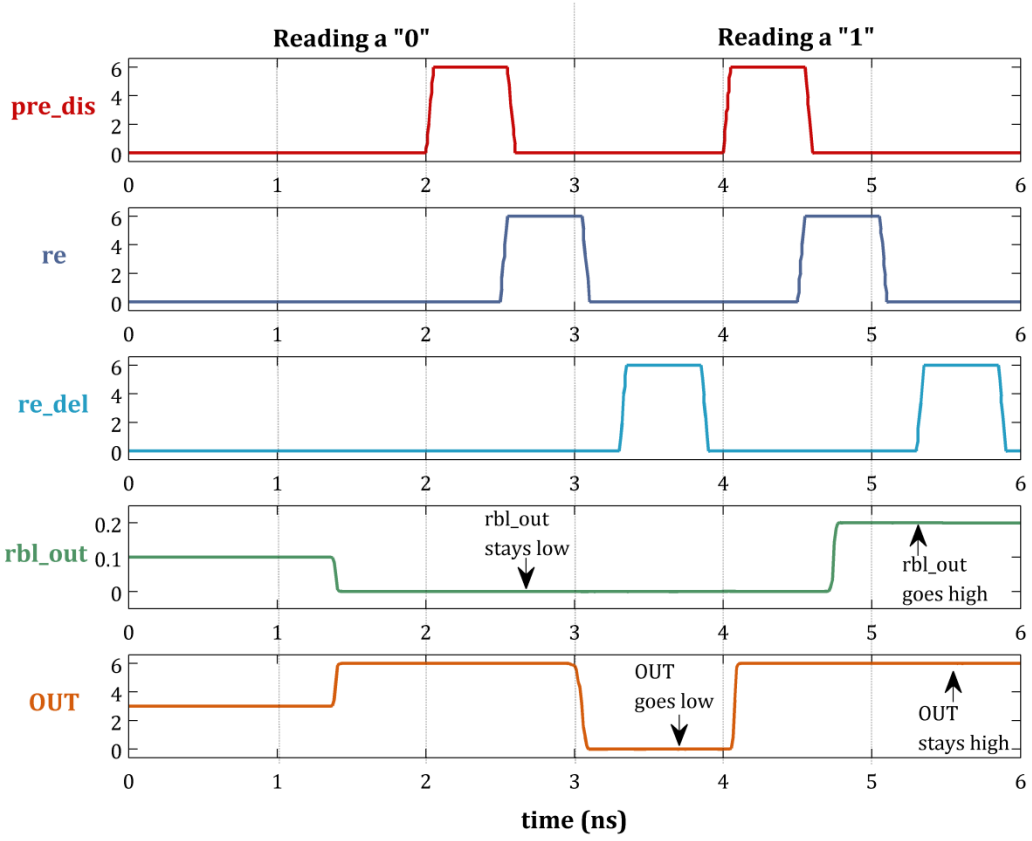

Figure 5.5: Consecutive reads occurring at the maximum theoretical throughput

5.2.2 Area, Write Energy, Throughput Tradeoffs

As done for the read energy, Figure 5.6 shows the write energy (E_{write}) as a function of area overhead and write throughput (f_{write}) for the relay and CMOS memories. Note that for a given f_{write}, $E_{write,3R} \approx E_{write,2R}$. This is because the full swinging *write* bitline capacitance ($C_{BL,write}$) dominates E_{write} with $C_{BL,write,3R}=C_{BL,write,2R}$.[25] As with $E_{read,CMOS}$, $E_{write,CMOS}$ begins leveling off at ~200MHz, but unlike in the read case, the relay designs, at this same throughput, exhibit ~6x larger E_{write} with an area overhead of ~1.5x (3R) and ~1.2x (2R). However, as with read energy, designs with less stringent throughput requirements can obtain lower E_{write} using relay-based memories by increasing their area overhead to ~2x or even ~3x.

[25] This is unlike E_{read} in which the *read* bitline capacitance was different due to the one less relay.

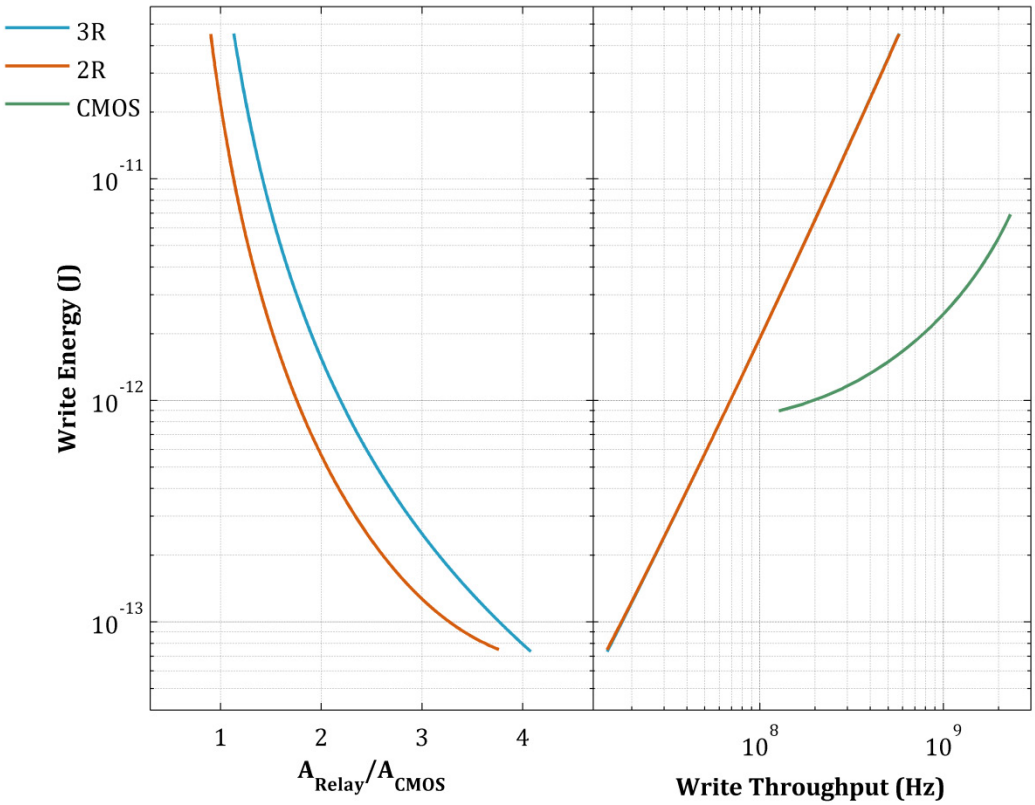

Figure 5.6: Area overhead vs. E_{write} (left) and write throughput vs. E_{write} (right)

6

Conclusion

6.1 Summary of Results

In this work, IC design using NEM relays was evaluated through the analysis of energy efficiency improvements beyond the limitations of CMOS technology. Thus, the sub-threshold regime of operation in CMOS circuits was examined while introducing its dynamic and leakage energy models. The tradeoff in these energies was analyzed through V_{DD} and V_{th} scaling. However, the sub-threshold leakage limited further V_{th} scaling, causing current CMOS designs to be limited by their total power consumption. With the NEM relay's zero off-state leakage, it provided for a potential solution to overcome this limitation. Therefore, its structure as well as its

pull-in and pull-out operations were examined to allow for the design of relay-based circuits.

Since a complete system requires computational, memory, and I/O blocks, a relay-based full-adder was initially reviewed from the work in [3] and benchmarked to a CMOS Sklansky adder in terms of area, throughput, and energy. The relay-based design had an order of magnitude improvement in energy over the CMOS design without any area overhead. This was because the increased area of a single relay device was offset through logic optimization that reduced the overall number of relay devices within the adder. Furthermore, the upscaling of CMOS transistors dependent on the load capacitance being driven also reduced the area overhead when designing relay-based computational blocks. This is because as the load capacitance increases, CMOS gates also need to be sized up in order to keep delay at its minimum; however, relay devices are dominated by their mechanical delay, and increasing the load capacitance without any upscaling has little effect on the overall delay. Thus, when driving large loads, the area overhead is even less pronounced due to the larger CMOS gates and the relatively unchanged size of the relay gates.

CMOS memories however, are already highly dense with very few transistors, limiting logic optimization. The issue is further exacerbated because unlike computational blocks whose delay was largely set by the load capacitance, CMOS memories are parasitic-delay dominated and do not yield the same

improvements in the overall delay when upscaling the transistors. As a result, the transistors in CMOS memories are usually sized to be as small as possible. Understanding these challenges and with density being a critical issue in memory structures, the main goal of this research involved designing relay-based memories while ensuring little or no area penalty.

Two relay-based memory designs were presented (3R and 2R) with an analysis of their read and write operations. This was followed by assessing their area overhead, achievable throughput, and their read and write energy consumption. By slightly reducing the throughput and slightly increasing the overall area, the read energy of these structures can be significantly reduced, thus motivating the implementation of a relay-based sense amplifier. The complete memory architecture was then benchmarked to a standard 6T CMOS SRAM also in terms of area, throughput, and energy. The read energy of the relay-based designs was nearly 3x lower than CMOS with only a 20% area overhead while the write energy remained the same with an area overhead of 2x. Despite the increased area overhead, applications restricted by their total power consumption but with less stringent throughput and density requirements can significantly benefit from the use of relay-based architectures. Furthermore, with the energy efficiency benefits achievable for digital computation and memory structures, NEM relay devices can provide a means to overcome the limitations imposed by CMOS technology.

6.2 Future Work

The relay-adder and the 3R MC were used to tape-out a chip to test the functionality and performance of these structures. The full-adder was used for 1-bit, 2-bit, 4-bit, and 8-bit adders in a ripple-carry configuration while the 3R MC was used to construct a 16-bit array. With fabrication and reliability issues being resolved, the energy-throughput measurements taken from the chip will more accurately predict the tradeoffs between relay-based and CMOS designs. The 2R MC benefits over the 3R MC will also need to be realistically examined, requiring another test chip.

These test chips are only an initial step in determining the benefits of these relay devices with the eventual objective of this research being the design of a complete relay-based microcontroller. Thus, the design techniques of logic optimization for computational blocks (enabling a 10x improvement in energy) and concepts of NEM relay operation for memory structures (enabling a 3x improvement in energy) will need to be extended to other building blocks. These include arithmetic logic units (also incorporating multipliers), other types of memories (combination of DRAMs and SRAMs), interface modules between these structures (serial and parallel buses), as well as various I/O structures. By achieving similar energy efficiency improvements in the design of all these structures, NEM relay technology may become a viable alternative to CMOS technology.

Appendix A: Noise Margins for Relay Circuits

This section serves to provide an understanding of the noise margins (NM) of digital relay circuits. More specifically, it first establishes how noise margins for a relay circuit are defined and then explains the constraints on the supply voltage (V_{DD}) that must be met for proper circuit functionality.

The low and high noise margins (NM_L and NM_H, respectively) for a relay circuit can be defined by examining the DC transfer characteristic of a relay-based "buffer." Figure A.1 shows the circuit schematic of the relay buffer (left) and its properly functioning DC transfer curve (right). The buffer uses a "PMOS" relay to

pull the output node down to 0V and an "NMOS" relay to pull the output node up to V_{DD}.[26]

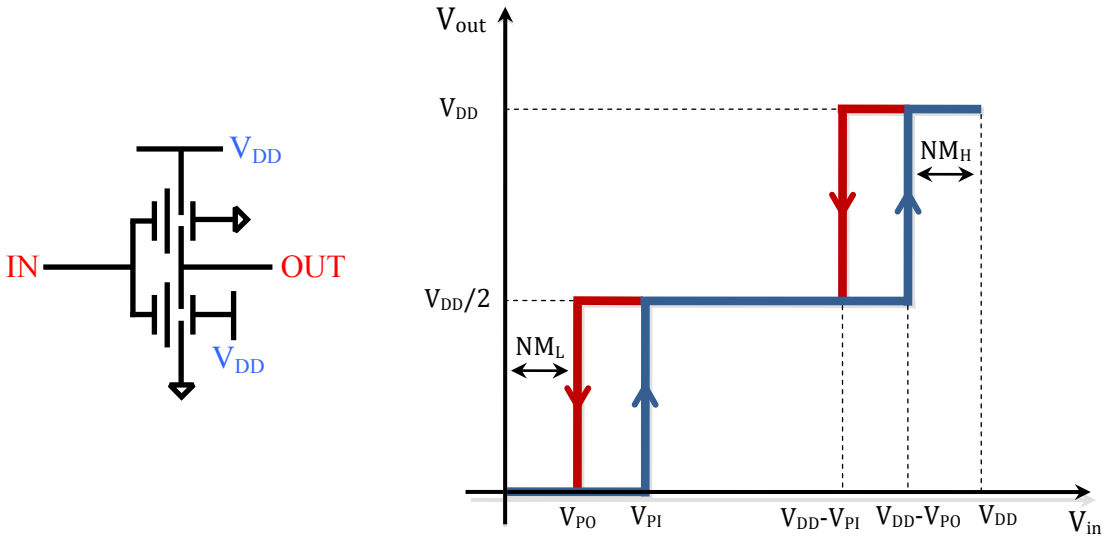

Figure A.1: The relay buffer (left) and its DC transfer characteristic (right)

When ramping the input node (V_{in}) from 0V to V_{DD}, the transfer characteristic can be understood by examining the mode of operation of each device:

$$V_{out} = \begin{cases} 0 \quad for \quad 0 \le V_{in} \le V_{pi} & (NMOS\ off, PMOS\ on) \\ \frac{V_{DD}}{2} \quad for \quad V_{pi} \le V_{in} \le V_{DD} - V_{po} & (NMOS\ on, PMOS\ on) \\ V_{DD} \quad for \quad V_{DD} - V_{po} \le V_{in} \le V_{DD} & (NMOS\ on, PMOS\ off) \end{cases} \tag{A.1}$$

Similarly, ramping V_{in} from V_{DD} down to 0V yields:

[26] Note that in traditional CMOS gates, a PMOS transistor is usually not used as a "pull-down" device due to its inability to pull the output node all the way down to 0V. However, a "PMOS" relay, when used as a pull-down device, can pull the output node all the way down to 0V. This is because the CMOS-based PMOS transistor's turn-on and turn-off characteristics are set by V_{GS} as opposed to V_{GB} for a relay-based PMOS transistor. For the same reason, an "NMOS" relay can be used as a "pull-up" device and is able to fully pull the output node up to V_{DD}, unlike the CMOS-based NMOS transistor.

$$V_{out} = \begin{cases} V_{DD} \;\; for \;\; V_{DD} - V_{pi} \leq V_{in} \leq V_{DD} & (NMOS\ on, PMOS\ off) \\ \frac{V_{DD}}{2} \;\; for \;\; V_{po} \leq V_{in} \leq V_{DD} - V_{pi} & (NMOS\ on, PMOS\ on) \\ 0 \;\; for \;\; 0 \leq V_{in} \leq V_{po} & (NMOS\ off, PMOS\ on) \end{cases} \quad \text{(A.2)}$$

The "staircase" shape of the transfer curve exists only for $V_{DD}>V_{pi}+V_{po}$, causing NM_L and NM_H to equal V_{po} as illustrated in Figure A.1. For values of $V_{DD}<V_{pi}+V_{po}$, the NM begins to decrease to values below V_{po}. This is because at $V_{DD}=V_{pi}+V_{po}$, the V_{pi} point in Figure A.1 becomes the $V_{DD}-V_{po}$ ($=V_{pi}+V_{po}-V_{po}$) point and the V_{po} point becomes the $V_{DD}-V_{pi}$ ($=V_{pi}+V_{po}-V_{pi}$) point. For further reductions in V_{DD}, the $V_{DD}-V_{po}$ and $V_{DD}-V_{pi}$ points "fold over" the V_{pi} and V_{po} points, respectively, reducing the NM to below V_{po}. In the extreme case, for $V_{DD}=V_{pi}$, Figure A.2 shows that the V_{pi} point becomes the V_{DD} point and the $V_{DD}-V_{pi}$ point becomes 0, completely eliminating NM_L and NM_H:

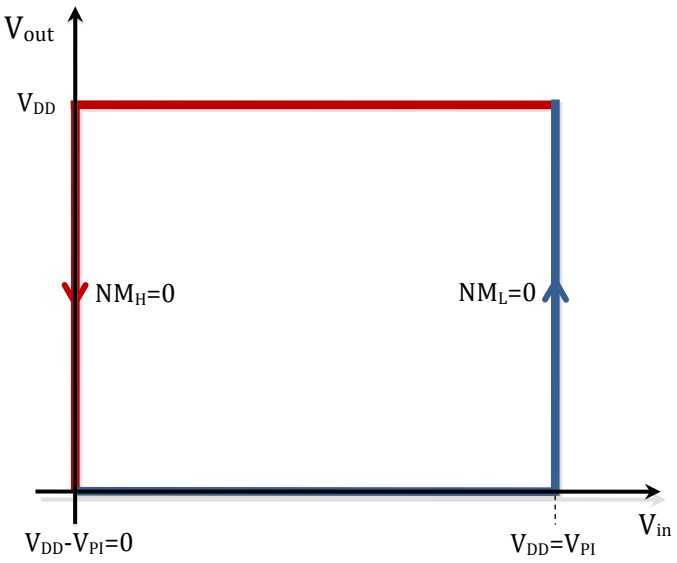

Figure A.2: Relay buffer's DC transfer characteristic for $V_{DD}=V_{pi}$ ($NM_L=NM_H=0$)

As done earlier, the transfer characteristic of Figure A.2 can also be understood by examining the mode of operation of each device in the relay buffer. The "$V_{DD}/2$" portion of Figure A.1 no longer exists in Figure A.2 because there is no voltage at which both the NMOS and PMOS relays are on. More specifically, when ramping V_{in} from 0V to V_{DD} (=V_{pi}), like Figure A.1, the PMOS relay is initially on and the NMOS relay is initially off. The NMOS relay only turns on for $V_{in} \geq V_{pi}$, but the PMOS relay is no longer on for this voltage range.[27] Similarly, when ramping V_{in} down from V_{DD} (=V_{pi}) to 0V, the NMOS relay is initially on and the PMOS relay is initially off (like Figure A.1). The PMOS relay only turns on for $V_{in} \leq V_{DD} - V_{pi}$, but the NMOS relay is no longer on for this voltage range.[28] Therefore, to maintain proper circuit functionality and have any NM, V_{DD} should be chosen to be greater than V_{pi}.

Using the concepts illustrated in Figures A.1 and A.2 and from the discussion above, an expression for NM as a function of V_{DD} for the relay buffer results in:

$$NM = \begin{cases} V_{DD} - V_{pi} & for \quad V_{pi} \leq V_{DD} \leq V_{pi} + V_{po} \\ V_{po} & for \quad V_{DD} \geq V_{pi} + V_{po} \end{cases} \tag{A.3}$$

Figure A.3 below plots the above expression and can be used to find the minimum V_{DD} that ensures a sufficient noise margin. For V_{DD}=V_{pi}, the noise margin is 0 while

[27] If V_{DD}=V_{pi}, when ramping V_{in} up, V_{in} reaches the V_{DD}-V_{po} point in Figure A.1 before the V_{pi} point. The PMOS relay turns off for $V_{in} \geq V_{DD} - V_{po}$. Thus, when V_{in} reaches V_{pi}, the NMOS relay turns on while the PMOS relay remains off.

[28] If V_{DD}=V_{pi}, when ramping V_{in} down, V_{in} reaches the V_{po} point in Figure A.1 before the V_{DD}-V_{pi} point. The NMOS relay turns off for $V_{in} \leq V_{po}$. Thus, when V_{in} reaches V_{DD}-V_{pi}, the PMOS relay turns on while the NMOS relay remains off.

for $V_{DD}>V_{pi}+V_{po}$, the noise margin is V_{po}. For the beam lengths of interest in this work, $V_{pi}+V_{po}$ can be approximated to be $1.5V_{pi}$. Thus, in this work V_{DD} has been set to $1.5V_{pi}$.

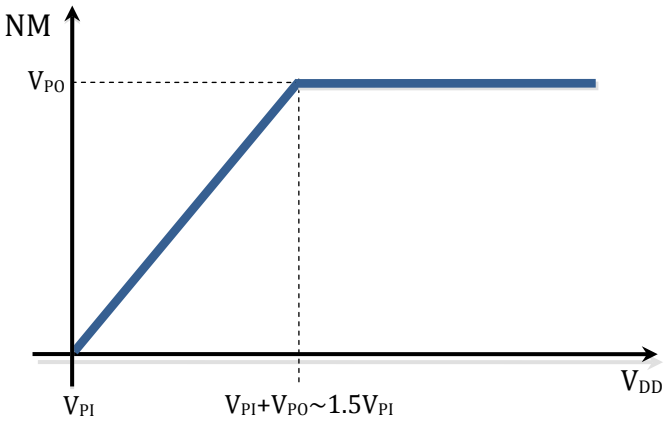

Figure A.3: Noise margin as a function of supply voltage

Appendix B: Relay Dimensions

This section serves to detail how the dimensions for the area of the relay memories were calculated. More specifically, it will illustrate the assumptions used for the spacing between the relay device's contacts.

For $\lambda = W/2$ with W being the width of the relay's channel (as illustrated in Figure 3.1), Figure B.1 below shows the complete dimensions of the relay assumed in this thesis. The channel width is 2λ by definition and the S/D contacts and the anchor are each $4\lambda \times 4\lambda$. The spacing on either side of the channel is 1λ with respect to the edges of the S/D regions. Although not shown in Figure B.1, the minimum spacing between any two adjacent paths or tracks on the same layer has been assumed to be 3λ (e.g. M1-M1 minimum spacing = M2-M2 minimum spacing = 3λ).

With these assumptions, the dimensions for the area calculations in Figures 4.7 and 4.13 can be evaluated.

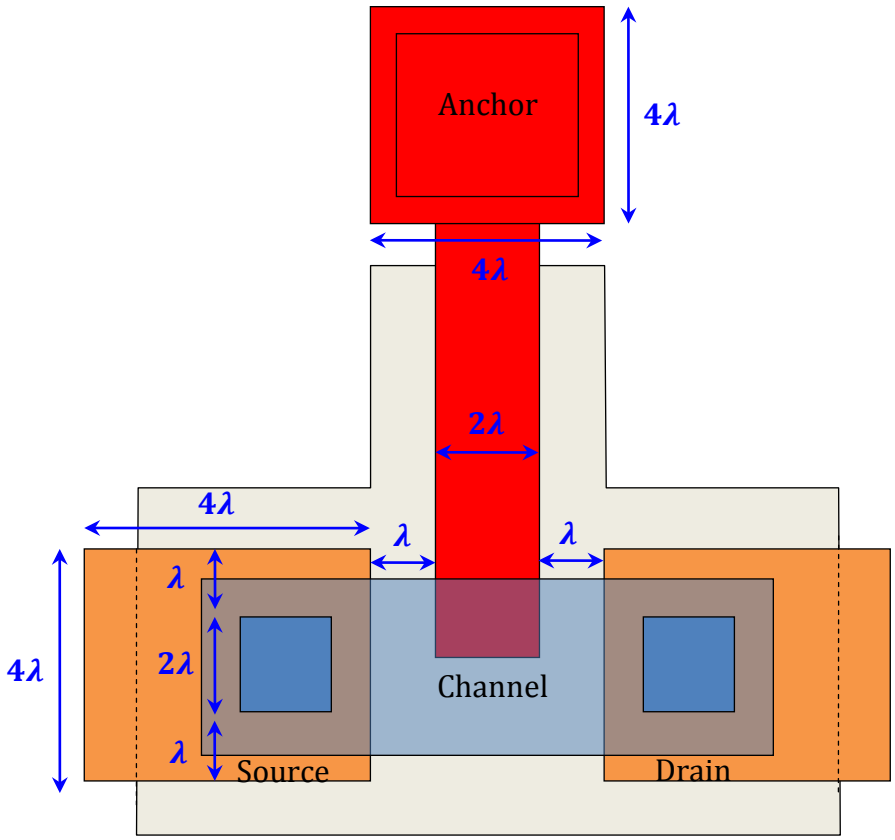

Figure B.1: Relay dimensions and contact spacing

Bibliography

[1] J. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital Integrated Circuits: A Design Perspective*, 2nd ed. Prentice-Hall International, 2003.

[2] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1778-1786, 2005.

[3] F. Chen, H. Kam, D. Marković, T. J. King, V. Stojanović, and E. Alon, "Integrated Circuit Design with NEM Relays," *IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2008, pp. 750-757.

[4] S. D. Senturia, *Microsystem Design*, Kluwer Academic, 2002.

[5] R. Maboudian and R. T. Howe, "Critical Review: Adhesion in Surface Micromechanical Structures," *AVS Journal of Vacuum Science and Technology B*, vol. 15, no. 1, pp.1-20, 1997.

[6] H. Kam, E. Alon, and T. J. King, "Generalized Scaling Theory for Electro-Mechanical Switches," (UCB internal, unpublished work)

[7] D. Patil, O. Azizi, M. Horowitz, and R. Ho, "Robust Energy-Efficient Adder Topologies," in *IEEE Symposium on Computer Arithmetic*, June 2007, pp. 16-28.

[8] K. Nii et al., "A 90nm Low-Power 32-kB Embedded SRAM with Gate Leakage Suppression Circuit for Mobile Applications," *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 684-693, 2004.

[9] H. C. Nathanson et al., "The Resonant Gate Transistor," *IEEE Transactions on Electron Devices*, vol. 14, pp. 117-133, 1967.

[10] L. M. Castaner and S. D. Senturia, "Speed-Energy Optimization of Electrostatic Actuators based on Pull-In," *Journal of Microelectromechanical Systems*, vol. 8, pp. 290-298, 1999.

[11] I. Schiele et al., "Micromechanical Relay with Electrostatic Actuation," *International Conference on Solid State Sensors and Actuators*, vol. 2, pp. 1165-1168, 1997.

[12] R. K. Gupta, and S. D. Senturia, "Pull-In Time Dynamics as a Measure of Absolute Pressure," *IEEE International Workshop on Micro Electro Mechanical Systems*, pp. 290-294, 1997.

[13] L. A. Rocha et al., "Pull-In Dynamics: Analysis and Modeling of the Traditional Regime," *IEEE International Conference on Micro Electro Mechanical Systems*, pp. 249-252, 2004.

[14] N. Abele et al., "1T MEMS Memory Based on Suspended Gate MOSFET," *IEEE International Electron Devices Meeting*, pp. 1-4, 2006.

[15] Woo Young Choi et al., "Compact Nano-Electro-Mechanical Non-Volatile Memory (NEMory) for 3D Integration," *IEEE International Electron Devices Meeting*, pp. 603-606, 2007.

[16] B. H. Calhoun and A. Chandrakasan, "A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 680-688, 2007.

[17] B. S. Amrutur and M. A. Horowitz, "Speed and Power Scaling of SRAM's," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 175-185, 2000.

[18] K. Zhang et al., "SRAM Design on 65-nm CMOS Technology with Dynamic Sleep Transistor for Leakage Reduction," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 895-901, 2005.

[19] S. Borkar, "Design Challenges of Technology Scaling," *IEEE Micro*, vol. 19, pp. 23-29, 1999.

[20] S. Borkar, "Low Power Design Challenges for the Decade," *Proceedings of the ASP Design Automation Conference*, pp. 293-296, 2001.

[21] Z. Zhang and Z. Guo, "Active Leakage Control with Sleep Transistors and Body Bias," (UCB internal, unpublished work)

[22] D. Ho et al., "Ultra-Low Power 90nm 6T SRAM Cell for Wireless Sensor Network Applications," *IEEE International Symposium on Circuits and Systems*, pp. 4-8, 2006.

[23] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mV Robust Schmitt Trigger Based Subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 2303-2313, 2007.