MOSFET Replacement Devices for Energy-Efficient Digital Integrated Circuits

Hei Kam

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-182
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-182.html

December 17, 2009

MOSFET Replacement Devices for Energy-Efficient Digital Integrated Circuits

by

Hei Kam

B.S. (University of California, Berkeley) 2004

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Tsu-Jae King Liu, Chair
Professor Elad Alon
Professor Tarek I. Zohdi

Fall 2009

The dissertation of Hei Kam is approved:

Chair _____ Date_____

_____ Date_____

_____ Date_____

University of California, Berkeley

MOSFET Replacement Devices for Energy-Efficient Digital Integrated Circuits

by

Hei Kam

Abstract

MOSFET Replacement Devices for Energy-Efficient Digital Integrated Circuits

by

Hei Kam

Doctor in Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

Increasing power density is a daunting challenge for continued MOSFET scaling due to non-scalability of the thermal voltage $k_B T/q$. To circumvent this CMOS power crisis and to allow for aggressive supply voltage reduction, alternative switching device designs have been proposed and demonstrated to achieve steeper than 60mV/dec subthreshold swing (S). This dissertation begins with a general overview of the physics and operation of these MOSFET-replacement devices. It then applies circuit-level metrics to establish evaluation guidelines for assessing the promise of these alternative transistor designs.

This dissertation then investigates the abrupt "pull-in" effect of an electrostatically actuated beam to achieve abrupt switching behavior in the nano-electro-mechanical field effect transistor (NEMFET). To facilitate low-voltage NEMFET design, the Euler-Bernoulli beam equation is solved simultaneously with the Poisson equation in order to accurately model the switching behavior of a

NEMFET. The impact of various transistor design parameters on the gate pull-in voltage and release voltage are examined. A unified pull-in/release voltage model is developed.

Finally, this dissertation proposes the use of micro-relays for zero-standby-power digital logic applications. To mitigate the contact reliability issue, it is demonstrated that since relatively high on-state resistance can be tolerated while extremely high endurance is a necessity, hard contacting electrode materials and operation with low contact force are preferred for reliable circuit operation. Using this contact design approach, a reliable relay technology that employs titanium dioxide (TiO_2) coated tungsten (W) electrodes is developed for digital logic applications. Relay miniaturization will lead to improvements in density (for lower cost per function), switching delay (for higher performance), and power consumption. A scaled relay technology is projected to provide >10× energy savings for digital circuits operating at up to ~100MHz.

_____

Professor Tsu-Jae King Liu

Dissertation Committee Chair

Acknowledgements

"Physics isn't the most important thing, Love is." — Richard P. Feynman

Finally, my eight-year-and-a-half journey in Berkeley is approaching to its end. I still remember every little thing in this journey, as if it happened only yesterday. I could never come so far and see so much without the help, love and support of many people. I would like to formally thank them all here.

I would first like to express my deepest and most sincere gratitude to my advisor, Professor Tsu-Jae King Liu. I would like to thank her for giving me the great opportunity to work for her as an undergraduate researcher six years ago. She gave me the freedom to pursue my research interest, while at the same time provided me valuable insights, guidance and support at all stages of my research. Her breadth and depth of knowledge in solid state devices and microfabrication technologies were exceptionally helpful to this thesis. Without her support, this thesis can never be completed.

I would also like to thank Professor Elad Alon in unofficially advising my research even though I am not his student. His intelligence, speed of thought and enthusiasm for research has always been a source of motivation and inspiration.

I am also grateful for Professor Vivek Subramanian for teaching the wonderful EE130 and EE231 in fall 2002 and spring 2003, respectively. I would also thank him for his valuable feedback as a member of my qualifying

examination and insightful advice on my career path. I would like to thank Professor Roger T. Howe and Professor Tarek I. Zohdi as well.

I have also had a lot of pleasure working with all the device group members who "live" in 373 Cory Hall and work endlessly in the microlab. I would like to thank them for their technical discussion and personal friendship. They include Anupama Bowonder, Andrew Carlson, Jaeseok Jeon, Sung Hwan Kim (I particular would like to thank him for sharing his measured data on Ge-source TFET), Joanna Lai, Blake Lin, Donovan Lee, Darsen Lu, Cheuk Chi Lo, Rhesa Nathanael, Pratik Patel, Kinyip Phoa, Vincent Pott, Changhwan Shin, Xin Sun and Reinaldo Vega.

I truly owe a debt to my birthplace, Hong Kong. Her restless energy is always a source of motivation and inspiration. I am also fortunate enough to still be close with friends from my childhood. In particular I would like to thank Yip Kwong Lo, Ken Hui, David Lee, James Wong, Andy Tam, Chris Ng, Anthony Wong, Dennis Chan and May Chui for their friendship.

Most of all, I would like to thank my family. I thank my grandma, my parents, my brother Henry and my sister Kylie for all their unbounded love. This dissertation is dedicated to them.

This work was supported in part by the GAANN fellowship, the FCRP and DARPA.

Hei Kam

December 4, 2009

Cory Hall, Berkeley

"Who sees the future? Let us have free scope for all directions of research."

– Adapted from Ludwig Boltzmann

Contents

Chapter 1

Introduction:

The CMOS Power Crisis

1.1 CMOS Scaling Trend

The steady reduction in the dimensions of complementary metal-oxide-semiconductor (CMOS) transistors from one technology to another has provided for dramatic improvements in the switching speed, density, cost and functionality of CMOS chips. As shown in Fig. 1.1, the physical gate length of a CMOS transistor has been reducing at an exponential rate and is expected to be scaled down to the sub-20nm regime in year 2010 [1.1]. However, due to the fact that the thermal voltage $k_B T/q$ does not scale, the threshold voltage (V_T) of CMOS transistors can no longer be reduced along with their lithographic dimensions. This non-scaling of the threshold voltage forces the supply voltage (V_{dd}) to remain constant across technologies for a given switching speed, as shown in Fig. 1.2, V_{dd} has saturated at around 1V from the 130nm technology forward [1.1]. Therefore, the power density of integrated circuits has increased drastically. As an example, the power dissipation in Intel's state-of-the-art microprocessors has already reached

the level of 100W or more [1.2]. Power is now a major constraint for modern day CMOS chip design.

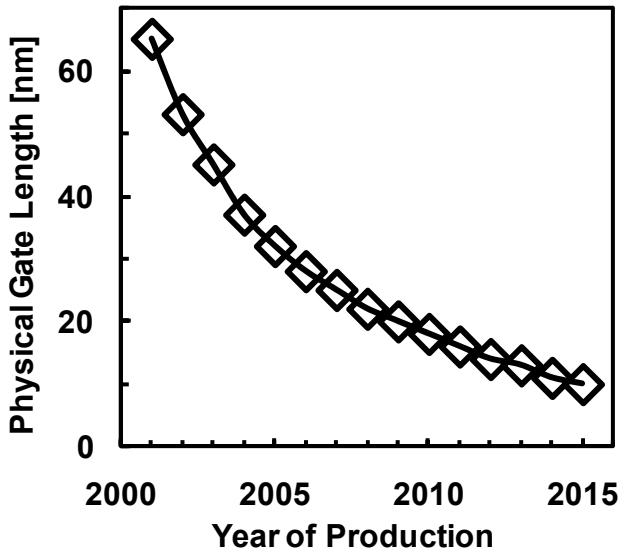


Fig. 1.1 Improvements in the switching speed, density, cost and functionality of CMOS chips have been enabled by the steady miniaturization of the transistor over the past four decades. By the year 2010, the physical gate length of a CMOS transistor is expected to be scaled down to the sub-20nm regime [1.1].

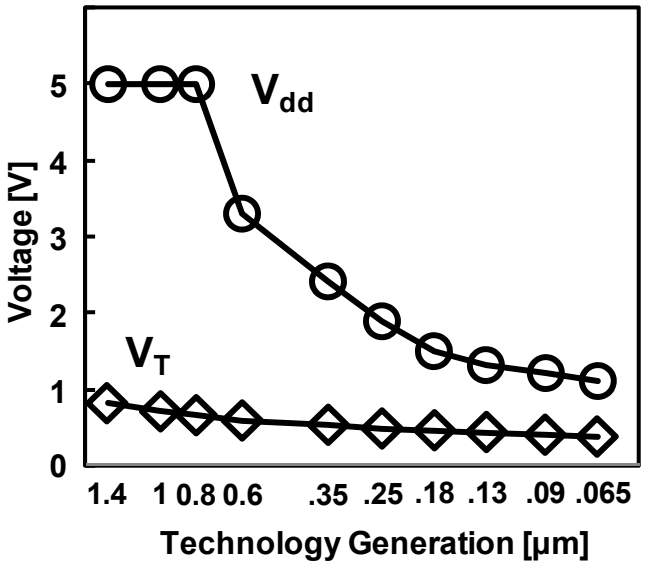


Fig. 1.2 To maintain constant power density, both the supply and the threshold voltages of a CMOS integrated circuit should be reduced along with the lithographic dimension of CMOS transistors. But due to MOSFET subthreshold leakage, both V_{dd} and V_T scaling have slowed down in recent technology generations [1.1].

1.2 MOSFET Physics in the Subthreshold Regime

The CMOS power crisis is fundamentally due to the non-scaling of the thermal voltage $k_B T/q$, which sets a lower limit for the subthreshold swing (S) of a MOSFET. Fig. 1.3 shows the I_{ds}-V_{gs} characteristics of a MOSFET. Below the threshold voltage (V_T), the MOSFET does not turn off completely; instead, I_{ds} decreases exponentially with V_{gs} with an inverse slope ("subthreshold swing") $S \geq$ 60mV/dec at room temperature. Thus, in the off state (V_{gs}=0V), CMOS transistors still dissipate leakage energy.

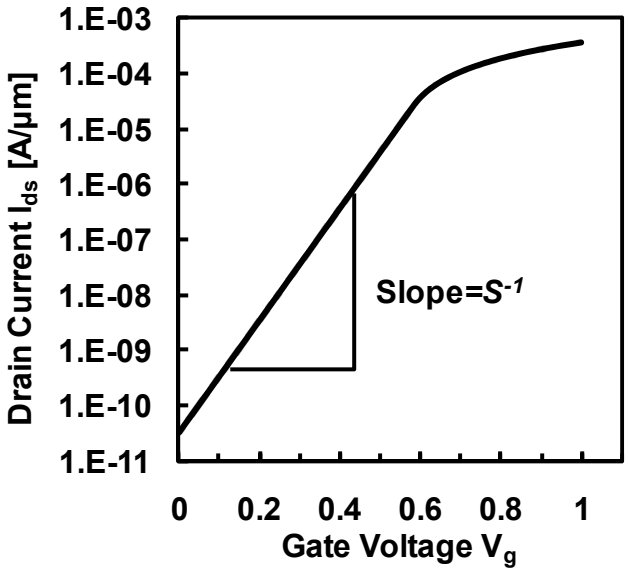


Fig. 1.3 The subthreshold swing *S* of a MOSFET is limited by the thermal voltage k_BT/q; it is greater than or equal to 60mV/dec at room temperature.

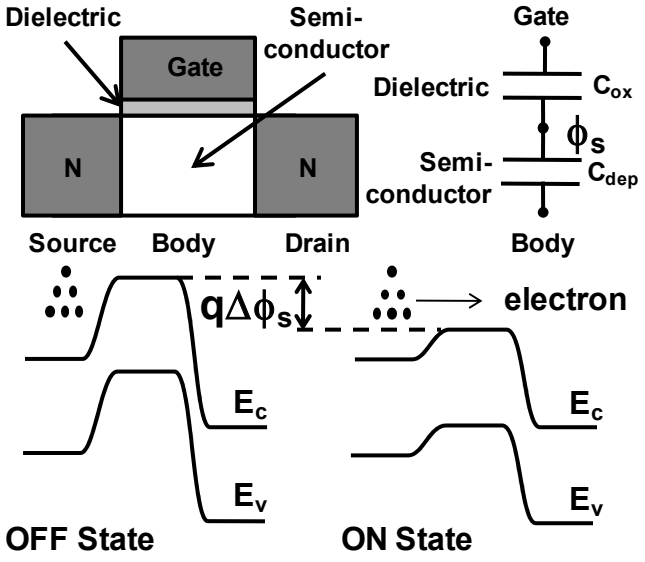


Fig. 1.4 The potential barrier for thermionic emission can be modulated ($\Delta\phi_s$) by the gate voltage; and according to Boltzmann statistics [1.3], electron concentration is exponentially proportional to $q\phi_s/k_BT$. This sets a lower limit for the MOSFET subthreshold swing, which is at least 60mV/dec at room temperature.

The origin of $S \geq 60$mV/dec can be explained by the electron energy band profile of a MOSFET, which is shown in Fig. 1.4. As V_{gs} increases, the channel potential is modulated by the capacitive divider:

$$\frac{d\phi_s}{dV_{gs}} = \frac{C_{ox}}{C_{ox}+C_{dep}} \quad \text{for } V_{gs} < V_T \tag{1.1}$$

where C_{ox} and C_{dep} are the oxide and depletion capacitances, respectively.

Increasing the channel potential (ϕ_s) reduces the potential barrier for electron injection, and hence the electron energy (E) increases. According to the Boltzmann statistics [1.3], electron concentration $n(E)$ increases exponentially with electron energy; therefore, the drain-to-source current flow (I_{ds}) depends exponentially on channel potential:

$$I_{ds} \propto n(E) \propto exp\left(\frac{E}{k_BT}\right) \propto exp\left(\frac{q\phi_s}{k_BT}\right) \propto exp\left(\frac{q}{k_BT}\frac{C_{ox}}{C_{ox}+C_{dep}}V_{gs}\right) \tag{1.2}$$

And the subthreshold swing of the MOSFETs can therefore be expressed as:

$$S \equiv \left(\frac{d\,log_{10}\,I_{ds}}{dV_{gs}}\right)^{-1} = \frac{d\,log_{10}\,I_{ds}}{d\phi_s} \times \frac{d\phi_s}{dV_{gs}} = ln(10)\frac{k_BT}{q}\left(1+\frac{C_{dep}}{C_{ox}}\right) \tag{1.3}$$

which is at least 60mV/dec at room temperature. Due to the non-zero depletion capacitance in the transistor, S is typically ~100mV/dec for state-of-the-art MOSFETs.

The subthreshold swing determines the lower energy limit for CMOS electronics. Fig. 1.5 illustrates the dependence of MOSFET energy consumption on the power supply voltage for a given switching speed. As the supply voltage is scaled down, the dynamic energy ($\sim CV_{dd}^2$) reduces quadratically; but to maintain the same switching speed ($\propto (V_{dd}-V_T)$), V_T must be decreased together with V_{dd} to

maintain the same on-state current. This, as a result, exponentially increases off-state leakage (Eqn. 1.2) and static energy. To reach the minimum operation energy, the dynamic and the leakage energies must be properly balanced; and for most digital designs, this optimal ratio is roughly 30-50% [1.4]. As previously alluded to, both the V_{dd} and V_T have remained roughly unchanged from the 130nm technology node and onwards; therefore the CMOS energy efficiency has not improved proportionately as the transistor dimensions have been scaled down.

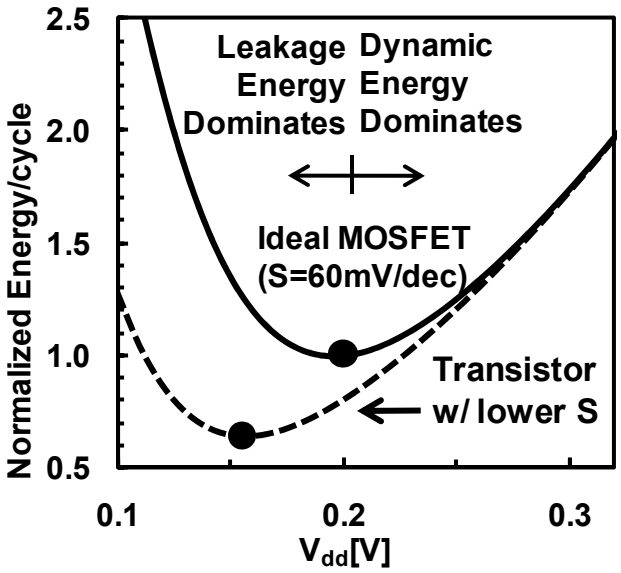


Fig. 1.5 Dynamic energy reduces quadratically as the supply voltage is scaled down; but to maintain a certain switching speed, the threshold voltage of the MOSFET must be scaled down as well, which increases the leakage energy. Therefore there exists an optimal V_{dd} that minimizes the energy dissipation. Transistor designs with lower S value reduce the leakage energy; they therefore improve the energy efficiency.

1.3 Various MOSFET Replacement Devices

To overcome the CMOS energy efficiency limit, alternative transistor designs which can achieve a steeper sub-threshold swing (*i.e.* more abrupt transition between on- and off-states) have been proposed. As shown in Fig. 1.5, transistor designs with lower S value reduce the leakage energy; this allows for more aggressive V_{dd} scaling and improvement in the energy efficiency.

To reach this goal, alternative transistor designs such as the tunneling based field effect transistors [1.5, 1.6], impact ionization MOS [1.7, 1.8], ferroelectric FETs [1.9, 1.10] and electromechanical devices [1.11-1.17] have been proposed and demonstrated to achieve subthreshold swing (S) < 60mV/dec. Among these, the tunneling field effect transistor (TFET) and electromechanical devices show the most promise for low power electronics applications.

1.3.1 Tunneling Field Effect Transistor (TFET)

Among all the alternative transistor designs, the tunnel field effect transistor (TFET) shows the most promise due to its relative simplicity and resemblance to the conventional MOSFET. The TFET utilizes band-to-band tunneling (BTBT) current to achieve a more abrupt on-to-off transition than what is achievable through thermionic emission. Fig. 1.6 shows the energy band diagram of the TFET in on and off states. In the off state, the wide energy barrier prohibits quantum tunneling between the source and channel regions. When a large V_{gs} is applied, the energy barrier narrows, and allowed energy states in the channel conduction band

align with allowed energy states in the source valence band, so that electrons can tunnel from the source to the channel. Since the TFET utilizes a different source injection mechanism from the MOSFET, it can potentially achieve lower S values, which has already been experimentally demonstrated [1.5]. Note, however, that a TFET achieves S<60mV/dec only at low current levels and that S increases as I_{ds} increases. Consequently, at high V_{dd} (~1V) values, a silicon TFET has a significantly lower on-state current I_{on} (~1μA/μm at 1V) than a silicon MOSFET (1mA/μm at 1V). This remains a principal challenge for TFET designers.

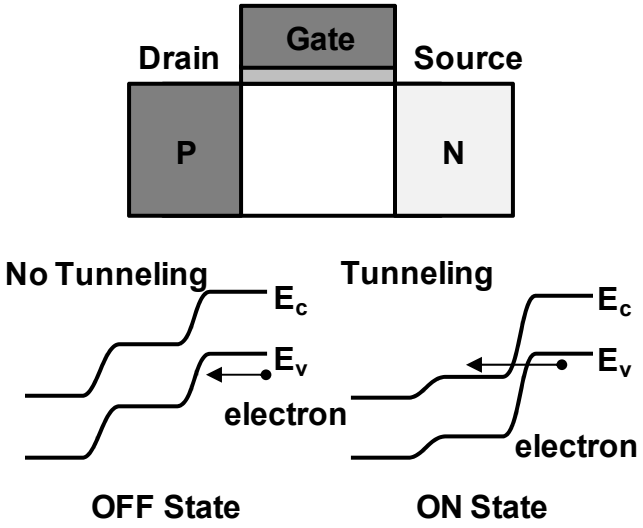


Fig. 1.6 Schematic diagram of a tunneling field effect transistor and its energy band diagram in the off and on states (n-channel operation).

1.3.2 Electromechanical Devices

Besides band-to-band tunneling, the abrupt "pull-in" effect in micro-electromechanical systems (MEMS) has also been harnessed to realize new

switching device designs with higher I_{on}/I_{off} ratio for a given gate voltage swing. These devices utilize a movable beam for switching, and they can roughly be divided into two categories: the nano-electro-mechanical field effect transistor (NEMFET) and the micro-electro-mechanical relay (micro-relay).

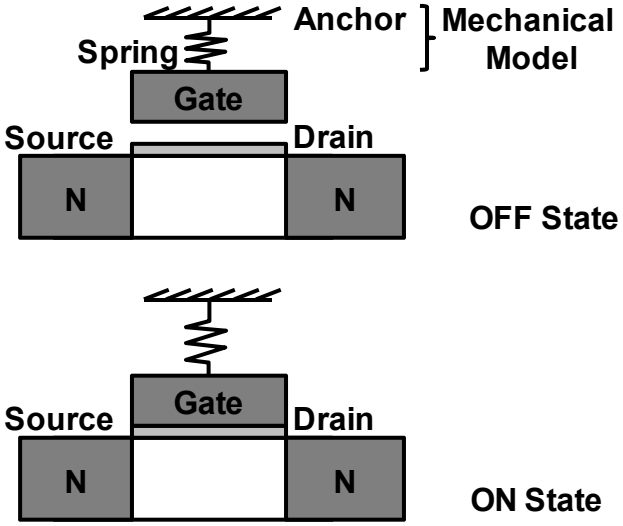


Fig. 1.7 Schematic diagram of a nano-electro-mechanical n-channel enhancement mode field effect transistor in off and on states.

a. Nano-Electro-Mechanical Field Effect Transistor (NEMFET)

A nano-electro-mechanical field effect transistor [1.11-1.13] is essentially a MOSFET with a movable gate electrode which can be physically separated from the gate dielectric layer by an air gap (or vacuum gap). As shown in Fig. 1.7, the gate, which is a mechanical beam anchored on both sides of the semiconducting channel, can be modelled as a simple linear spring (with a characteristic spring constant k) suspended over the semiconductor channel. The gate and the channel form a parallel-plate capacitor with an equivalent air-gap.

In the off-state (V_{gs}=0V), the gate is separated from the gate dielectric; the gate coupling to the channel is weak and the transistor is therefore turned off. When a positive V_{gs} is applied, the electrostatic force attracts the mechanical gate towards the gate dielectric. While the electrostatic force increases quadratically with increasing displacement, the spring restoring force, which counteracts the electrostatic force, increases only linearly with displacement. Hence, there is a critical pull-in voltage (V_{pi}) beyond which the electrostatic force is always larger than spring restoring force, causing the gap to close abruptly. When the gate is in contact with the gate dielectric, the gate coupling to the channel is maximized and the transistor is turned on. Taking advantage of this pull-in phenomenon, a NEMFET with perfectly abrupt switching transition (S=0mV/dec) at V_{gs}=V_{pi} have been utilized for logic, memory and resonator applications [1.11-1.13].

b. Micro-electro-Mechanical Relay (Micro-Relay)

The abrupt pull-in effect has also been harnessed for micro-electro-mechanical relays ("micro-relays") [1.14-1.17]. The attractiveness of micro-relays stems from the fact that a mechanical switch offers nearly ideal switching characteristics: zero off-state drain-to-source and gate leakage currents, and perfectly abrupt off-to-on transition. Since there is no trade-off between off-state leakage current and on-state drive current, the relay threshold voltage and therefore V_{dd} can in principal be reduced much more aggressively than for MOSFETs, potentially leading to improved energy efficiency.

In terms of device structure and operation (shown in Fig. 1.8), a micro-relay for digital logic applications ("logic relays") is very similar to one targeted for radio-frequency signal DC switching applications ("RF relays"). In the off state, an air gap separates the source from the metallic drain electrode so that no current can flow. In the on state where the gate-to-source voltage is greater than the pull-in voltage (V_{pi}), the source, which is a movable beam, comes down and touches the drain electrode, providing a conductive path for current to flow. Since the relay switches on abruptly as V_{gs} is increased above V_{pi}, the I_d-V_g characteristic of the relay exhibits an extremely steep (nearly infinite) subthreshold slope.

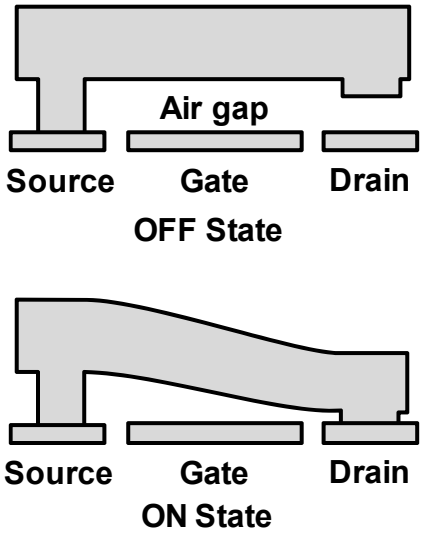


Fig. 1.8 Schematic diagram of a micro-relay in off and on states

1.4 Objectives

This research focuses on the analysis, design and applications of MOSFET-replacement devices, with emphasis on the TFET and electromechanical devices.

As alluded to previously, TFETs often have small S value at low current levels but fail to achieve the required on-to-off current ratio across a range of V_{dd}. To investigate whether TFETs can effectively replace MOSFETs, one needs to compare the energy-performance tradeoff of a TFET with that of a MOSFET. To achieve this goal, chapter 2 first reviews the energy-performance tradeoffs of CMOS, and shows that the optimal I_{on}/I_{off} value for most CMOS replacement devices at the optimum energy depends only on the average activity factor and the logic depth, and that its value is roughly insensitive to all other device parameters. Thus, it is the device's effective subthreshold swing (S_{eff}) over a range of voltage rather than the steepest local subthreshold swing (S) value that determines the energy efficiency. With this in mind, simple guidelines for assessing the energy efficiency of CMOS replacement are established. As a concrete example, we use this methodology to compare TFETs against CMOS, showing that TFETs may offer substantial (~5x) energy savings for performance up to the 100MHz range.

To alleviate the issue of S degradation at high current level, the use of the abrupt gate pull-in effect in NEMFET to achieve the required I_{on}/I_{off} ratio with a smaller V_{dd} appears to be an attractive solution. To facilitate this goal, NEMFET device physics and operation are studied in Chapter 3. Due to the beam bending of the mechanical gate, the channel potential is non-uniform in the semiconductor. The Euler-Bernoulli beam equation is solved simultaneously with the Poisson equation in order to accurately model the switching behavior of NEMFETs. Using this approach, the shape of the movable gate electrode and semiconductor potential

across the width of the channel are derived for the various regimes of transistor operation (before gate pull-in, after gate pull-in, and at the point of gate release). The impact of various transistor design parameters such as the body doping concentration, gate work function, gate stiffness, and as-fabricated actuation gap thickness, as well as source-to-body bias voltage and surface forces, on the gate pull-in voltage and gate release voltage are examined. A unified pull-in/release voltage model is developed, to facilitate NEMFET design for digital and analog circuit applications.

Although the pull-in effect can be harnessed to achieve a perfectly abrupt off-to-on switching transition for NEMFET; the presence of the air-gap in the off-state also severely decreases the gate-to-channel capacitive coupling in the off-state, limiting NEMFET scalability. In light of this limitation, chapter 4 discusses the use of micro-relays for zero-standby power logic applications. Contact design techniques to achieve reliable (high-endurance) micro-relay operation are described. Utilizing TiO_2-coated tungsten contacting electrodes, prototype relays fabricated using a CMOS-compatible process are demonstrated to operate with low surface adhesion force, adequately low on-state resistance (< 100kΩ) over a wide temperature range (20^oC-200^oC), and $>10^9$ on/off switching cycles in N_2 ambient without stiction- or welding induced failure. These results pave the pathway to realizing reliable micro-relays for digital logic applications.

Using calibrated relay delay and energy models, a sensitivity-based relay energy-delay optimization methodology is developed in Chapter 5, in which simple

relay design guidelines are established. The proposed scaling methodology for micro-relays is then presented, which leads to systematic improvements in device density, performance, and energy consumption. Simulation results indicate that scaled relay technology can potentially offer >10× improvement in energy efficiency for applications requiring performance up to ~100MHz.

Chapter 6 summarizes the key results and contributions of this dissertation; future research directions are also suggested.

1.5 References

[1.1] *The International Technology Roadmap for Semiconductors (ITRS)*, 2007. [Online]. Available: http://public.itrs.net

[1.2] B. Nikolic, "Design in the power-limited scaling regime," *IEEE Trans. Electron Devices*, vol. 55, pp. 71, Jan. 2008.

[1.3] R. S. Muller and T. I. Kamins with M. Chan, *Device Electronics for Integrated Circuits*, 3rd ed. New York: Wiley, 2003, pp. 16.

[1.4] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for low-power and high-speed applications," in *Proc. Asia South Pacific Design Automation Conf.*, Jan. 2000, pp. 469–474.

[1.5] S. H. Kim, H. Kam, C. Hu and T.-J. King-Liu, "Germanium-Source Tunnel Field Effect transistors with Record High I_{ON}/I_{OFF}", *Symposium on VLSI Technology Digest of Technical Papers*, pp. 178-179, 2009.

[1.6] K. K. Bhuwalka, J. Schulze, and I. Eisele, "Performance enhancement of vertical tunnel field-effect transistor with SiGe in the δp^+ layer," *Jpn. J. Appl. Phys.*, vol. 43, no. 7A, pp. 4073-4078, Jul. 2004.

[1.7] K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, "I-MOS: A novel semiconductor device with a subthreshold slope lower than *kT/q*," in *IEDM Tech. Dig.*, 2002, pp. 289–292.

[1.8] W. Y. Choi , J. Y. Song , J. D. Lee , Y. J. Park and B.-G. Park "A novel biasing scheme for I-MOS (impact-ionization MOS) devices," *IEEE Trans. Nanotechnol.*, vol. 4, pp. 322, May 2005.

[1.9] S. Salahuddin and S. Datta, "Use of negative capacitance to provide a sub-threshold slope lower than 60 mV/decade," Nanoletters, vol. 8, No. 2, 2008.

[1.10] S. Salahuddin and S. Datta, "Can the subthreshold swing in a classical FET be lowered below 60 mV/decade?," in *IEDM Tech. Dig.*, 2008, pp. 693–696.

[1.11] N. Abele, N. Fritschi, K. Boucart, F. Casset, P. Ancey, and A. M. Ionescu, "Suspended-gate MOSFET: Bringing new MEMS functionality into solid-state MOS transistor," in *IEDM Tech. Dig.*, 2005, pp. 1075–1077.

[1.12] H. Kam, D. T. Lee, R. T. Howe, and T.-J. King, "A new nano-electromechanical field effect transistor (NEMFET) design for low-power electronics," in *IEDM Tech. Dig.*, 2005, pp. 463–466.

[1.13] K. Akarvardar, C. Eggimann, D. Tsamados, Y. Singh Chauhan, G. C. Wan, A.M. Ionescu, R.T. Howe, and H.-S.P. Wong, "Analytical Modeling of the

Suspended-Gate FET and Design Insights for Low-Power Logic,**"** *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 48-59, Jan. 2008.

[1.14] F. Chen, H. Kam, D. Markovic, T.J. King, V. Stojanovic, and E. Alon, "Integrated Circuit Design with NEM Relays," in *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, 2008, pp. 750-757.

[1.15] K. Akarvardar, D. Elata, R. Parsa, G. C. Wan, K. Yoo, J. Provine, P. Peumans, R. T. Howe, H.-S. P. Wong, "Design Considerations for Complementary Nanoelectromechanical Logic Gates," in *IEDM Tech. Dig.*, 2007¸ pp. 299-302.

[1.16] H. Kam, V. Pott, R. Nathanael, J. Jeon, E. Alon and T.-J. King-Liu, "Design and Reliability of a Micro-Relay Technology for Zero-Standby-Power Digital Logic Applications," in *IEDM Tech. Dig.*, 2009, pp. 809–812.

[1.17] R. Nathanael, V. Pott, H. Kam, J. Jeon and T.-J. King-Liu,, "4-Terminal Relay Technology for Complementary Logic," in *IEDM Tech. Dig.*, 2009, pp. 223–226.

Chapter 2

Circuit-Driven Requirements for MOSFET-Replacement Devices

2.1 Introduction

As alluded to in Chapter 1, many alternative switching devices [2.1-2.14] have been proposed and demonstrated to achieve subthreshold swing (S) <60mV/dec to allow for power supply (V_{dd}) and threshold voltage (V_T) scaling to alleviate the CMOS power crisis. However, many of these devices (e.g. the TFET) achieve S<60mV/dec only at low on-current levels, and fail to maintain improved I_{on}/I_{off} across a range of V_{dd}. In addition, some of these devices (e.g., the NEMFET and NEM relay) do not begin to conduct current until sometime after the control voltage arrives, leading to an additional delay. To investigate whether these devices can effectively replace MOSFETs, one needs to compare the energy-performance tradeoff of these new logic devices with that for MOSFETs. To achieve this goal, this chapter starts by describing the energy-performance tradeoffs of CMOS gates in section 2.2. It is then shown in section 2.3 that for a given performance target, the optimal I_{on}/I_{off} value of CMOS at the optimum energy depends only on the average activity factor and the logic depth. In section 2.4, this optimal I_{on}/I_{off} value

is shown to remain roughly the same for most CMOS replacement devices. With this optimal I_{on}/I_{off} fixed, it is shown in section 2.5 that the device's effective subthreshold swing (S_{eff}) over a range of voltage (rather than the steepest local subthreshold swing (S) value) determines the device's energy efficiency. With this in mind, simple guidelines for assessing the energy efficiency of MOSFET replacement devices are then established. Finally, as a concrete example, this methodology is used to compare TFETs against CMOS, showing that TFETs may offer substantial (~5x) energy savings for performance up to the 100MHz range.

2.2 Simplified Energy-Performance Analysis

Although digital chips clearly consist of a broad variety of circuit types, the tradeoffs between energy and delay for the majority of CMOS gates on a chip are similar to those of an inverter. Therefore, at least for devices whose qualitative behavior is similar to a MOSFET, we can approximately compare the energy and delay tradeoffs by using an inverter chain, as shown in Fig. 2.1. The total energy consumption per operation of an inverter chain with logic depth L_d , average activity factor a , electrical fanout (FO) f, and capacitance/stage C can be computed by adding the dynamic and the leakage energy components:

$$E = E_{dyn} + E_{leak} = aL_dV_{dd}^2C \cdot f + L_dfI_{off}V_{dd}t_{delay} \quad (2.1)$$

where the t_{delay} is simply:

$$t_{delay} = \frac{L_dC \cdot fV_{dd}}{2I_{on}} \quad (2.2)$$

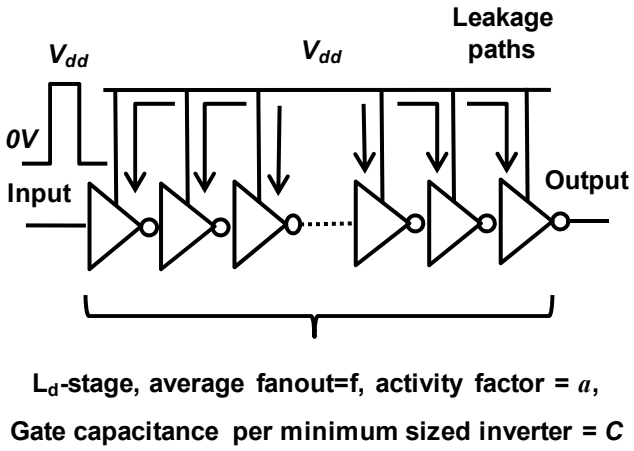


Fig. 2.1 L_d-stage inverter chain considered for energy efficiency. For most digital circuit, the energy delay tradeoffs of CMOS gates are similar to the tradeoff for an inverter chain.

Optimized circuit design entails the choice of parameters such as supply (V_{dd}) and threshold voltage (V_T) that minimize the energy dissipation (Eqn. 2.1) subject to a given delay target (Eqn. 2.2). To reach this goal, the dynamic and static energies must be properly balanced. For CMOS circuits, dynamic energy can be reduced quadratically by decreasing $V_{dd.}$ However, in order to avoid increased circuit delay, V_T must be decreased along with V_{dd} to maintain a high on-state drive current (I_{on}) [2.15]. This results in increased off-state leakage current (I_{off}) and hence increased static energy. For alternative devices that are MOSFET-like [2.1-2.9], [2.12-2.14], the tradeoff between the dynamic and the leakage energies is similar. Note that Nose and Sakurai [2.15] have previously proven that for an optimized CMOS circuit design, the leakage-to-dynamic-energy-ratio of is roughly

0.3-0.5 across wide range of parameters. In this dissertation, we follow a similar derivation approach, but now from a device designer's perspective, to show that such an optimal energy ratio can equivalently be expressed as an optimal I_{on}/I_{off} ratio for CMOS. To reach this goal, in Appendix I, the method of Lagrange multipliers is used to show that a generalized logic device is energy-delay optimized if the device I_{on} and I_{off} values satisfy the following condition:

$$\frac{4a}{L_d f}\frac{I_{on}}{I_{off}} + 1 + \frac{I_{on}}{I_{off}}\frac{dI_{off}}{dI_{on}} = 0 \qquad (2.3)$$

As shown by (Eqn. (2.3)), the optimum I_{on}/I_{off} depends only on the circuit topology and dI_{off}/dI_{on}, a parameter which is related to the subthreshold swing, as will be explained later.

Once the optimal I_{on}/I_{off} value is found, the energy minimum can readily be obtained by Eqn. (2.1), which can equivalently be expressed by:

$$E = V_{dd}^2 C L_d^2 f^2 \left(\frac{a}{L_d f} + \frac{I_{off}}{2I_{on}}\right) \qquad (2.4)$$

where V_{dd} is the supply voltage required to reach the target performance. Since the optimal I_{on}/I_{off} is set by the circuit topology, we can see from Eqn. (2.4) that the minimum energy is proportional to the dynamic energy CV_{dd}^2. This implies that, as will be explained more in detail later, any logic device that can achieve the required I_{on}/I_{off} value at a lower V_{dd} than CMOS is going to be more energy-efficient. To assess the promise of alternative switching devices for replacing the MOSFET, one needs to know how I_{on}/I_{off} of these devices depend on $L_d f/a$. To reach this goal, the optimal I_{on}/I_{off} for CMOS is first derived and then shown to be relatively constant across regions of operation (strong inversion versus subthreshold).

2.3 Optimal I_{on}/I_{off} for CMOS

To find the optimal the I-V characteristics of MOSFETs are first approximated as:

$$I_{ds} = I_s exp\left(\frac{V_{gs}-V_T}{nv_{th}}\right) \quad V_{gs} \leq V_T + \alpha n v_{th} \tag{2.5a}$$

$$I_{ds} = I_s e^{\alpha}\left(\frac{V_{gs}-V_T}{\alpha n v_{th}}\right)^{\alpha} \quad V_{gs} > V_T + \alpha n v_{th} \tag{2.5b}$$

where n is the subthreshold slope factor ($n \cong 1.67$ for S=100mV/dec), $\alpha \cong 1.2$ and v_{th} is the thermal voltage.

Using this I-V model, the optimal I_{on}/I_{off} can be found by first differentiating I_{off} with respect to I_{on} (Eqn. (2.3)):

$$\frac{dI_{off}}{dI_{on}} = \frac{-I_{off}}{nv_{th}}\frac{dV_T}{dI_{on}} \tag{2.6}$$

High performance CMOS digital circuits often operate in the strong inversion region ($V_{gs}>V_T$), therefore dV_T/dI_{on} can be found by differentiating Eqn. (2.5b) with respect to I_{on}, which gives the following expression:

$$1 = I_s\left(\frac{e}{\alpha n v_{th}}\right)^{\alpha}\alpha(V_{dd}-V_T)^{\alpha-1}\left(\frac{dV_{dd}}{dI_{on}} - \frac{dV_T}{dI_{on}}\right) \tag{2.7}$$

Substituting Eqn. (2.6) into Eqn. (2.7), we obtain the expression for dI_{off}/dI_{on}:

$$\frac{dI_{off}}{dI_{on}} = \frac{1}{\alpha n v_{th}}\frac{I_{off}}{I_{on}}(V_{dd}(1-\alpha) - V_T) \tag{2.8}$$

Finally, substituting Eqn. (2.8) into Eqn. (2.3), the optimal I_{on}/I_{off} can be expressed by the following equation:

$$\frac{I_{on}}{I_{off}} = \frac{L_d f}{4a}\left(\frac{V_{dd}(\alpha-1)+V_T-\alpha n v_{th}}{\alpha n v_{th}}\right) \tag{2.9}$$

where V_{dd} and V_T are set by the performance target (Eqn. (2.2)):

$$t_{delay} = \left(\frac{L_d C \cdot f(\alpha n v_{th})^{\alpha}}{2 I_s e^{\alpha}}\right) \frac{V_{dd}}{(V_{dd} - V_T)^{\alpha}} \tag{2.10}$$

Typically, the threshold and supply voltages lie within the following bounds: $1 \geq V_{dd} \geq 0.4$, $0.5 \geq V_T \geq 0.3$; substituting these V_{dd} and V_T values together with $\alpha \cong 1.2$ into Eqn. (2.9), the optimal I_{on}/I_{off} can be expressed by the following equation:

$$\frac{I_{on}}{I_{off}} \approx K_1 \frac{L_d f}{a} \tag{2.11}$$

where K_1 lies within the range 1.6 to 3.2. This verifies that for CMOS circuits operated in strong inversion region, I_{on}/I_{off} at the energy optimum is mainly set by the circuit topology.

Thus far in this discussion, the MOSFET has been assumed to operate in the strong inversion region ($V_{dd} > V_T$); however, many alternative devices such as TFETs achieve low on-current levels (~10μA/μm) at supply voltages similar to those used in current high performance CMOS circuits (~1V). These devices will therefore only be competitive with subthreshold MOSFETs (operated with $V_{dd} < V_T$), which dissipate the minimum energy required for CMOS to perform a given operation [2.16, 2.17]. At the energy optimum, the optimal I_{on}/I_{off} ratio for subthreshold CMOS similarly can be derived by first expressing off-state current in terms on the on-state current (Eqn. (2.5a)) and the delay (Eqn. (2.2)):

$$I_{off} = I_s exp\left(\frac{-V_T}{n v_{th}}\right) = I_{on} exp\left(-\frac{V_{dd}}{n v_{th}}\right) = I_{on} exp\left(-\frac{2 t_{delay}}{n v_{th} L_d C f} I_{on}\right) \tag{2.12}$$

Therefore, dI_{off}/dI_{on} is:

$$\frac{dI_{off}}{dI_{on}} = \left(1 + ln\left(\frac{I_{off}}{I_{on}}\right)\right) \frac{I_{off}}{I_{on}} \tag{2.13}$$

Substituting Eqn. (2.7) into Eqn. (2.3), the optimal I_{on}/I_{off} ratio can be expressed by the following equation:

$$\frac{I_{on}}{I_{off}} = -\frac{L_d f}{4a} lambertW\left(-\frac{4a}{L_d f} e^2\right) \approx \frac{L_d f}{4a} \ln\left(\frac{L_d f}{4a}\right) \approx K_2 \frac{L_d f}{a} \qquad (2.14)$$

where *lambertW(y)* is the x that solves the equation $y=xe^x$ and K_2 lies within the range 1.9 to 3.6. This result is consistent with previous published work [2.16, 2.17]. From Eqn. (2.14), it can be seen that the optimal I_{on}/I_{off} depends only on $f \times L_d / a$, and also that even in subthreshold operation, the optimal I_{on}/I_{off} ratio stays roughly the same as that for super-threshold MOSFETs (Eqn. (2.11)). This is because, as shown in Fig. 2.2, most of the change in I_{on}/I_{off} (as a function of V_{dd}) occurs in the subthreshold region, and therefore even in strong inversion, the ratio is relatively insensitive to small changes in V_{dd}. As a result, the optimal I_{on}/I_{off} depends mainly on $f \times L_d / a$ and is relatively insensitive to all other device parameters.

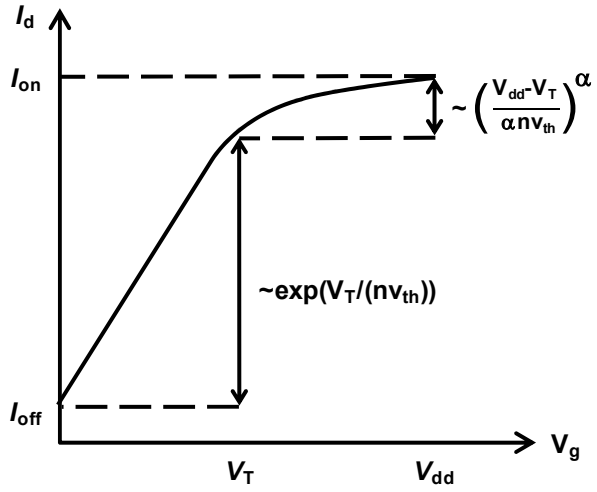


Fig. 2.2. Most of the change in I_{on}/I_{off} (as a function of V_{dd}) for MOSFET occurs in the subthreshold region (approximately five orders of magnitude) rather than in the strong inversion region (approximately 25), and therefore the ratio is relatively insensitive to small changes in V_{dd} . Thus the optimal I_{on}/I_{off} is roughly the same for both sub-threshold and super-threshold MOSFETs.

2.4 Optimal I_{on}/I_{off} for CMOS Replacement Devices

Thus far the discussion has only focused on optimizing MOSFETs. To assess the promise of alternative switching devices for replacing the MOSFET, similar analysis can be applied to CMOS replacement devices. To reach this goal, as derived in Appendix II, Eqn. (2.3) is equivalently expressed by the following equation:

$$\frac{4a}{L_{df}}\frac{I_{on}}{I_{off}} + 2 - ln\left(\frac{I_{on}}{I_{off}}\right)\left(1 - log_{10}\left(\frac{I_{on}}{I_{off}}\right)\frac{dS_{eff}}{dV_{dd}}\right) = 0 \qquad (2.15)$$

where S_{eff} is the effective subthreshold swing:

$$S_{eff} \equiv \left(\frac{1}{V_{dd}} log_{10}\left(\frac{I_{on}}{I_{off}}\right)\right)^{-1} \qquad (2.16)$$

Thus, the optimal I_{on}/I_{off} of the generalized logic device depends only on $f \times L_d / a$ and dS_{eff}/dV_{dd}, where dS_{eff}/dV_{dd} is related to the log-concavity of the transfer characteristics. For a broad variety of different logic devices that are MOSFET-like, the S value degrades as the current level increases, i.e. their transfer characteristics are logarithmically concave with $dS_{eff}/dV_{dd} \sim 0$. Therefore, even without knowing the exact characteristics of a new device, one can still approximate the optimal I_{on}/I_{off} from Eqn. (15), which gives the following expression:

$$\frac{I_{on}}{I_{off}} = K_3 \times \frac{L_d f}{a} \tag{2.17}$$

where K_3 is set by the exact of dS_{eff}/dV_{dd} value, and K_3 lies within the range ~2-8 for most logic devices, as shown in Fig. 2.3.

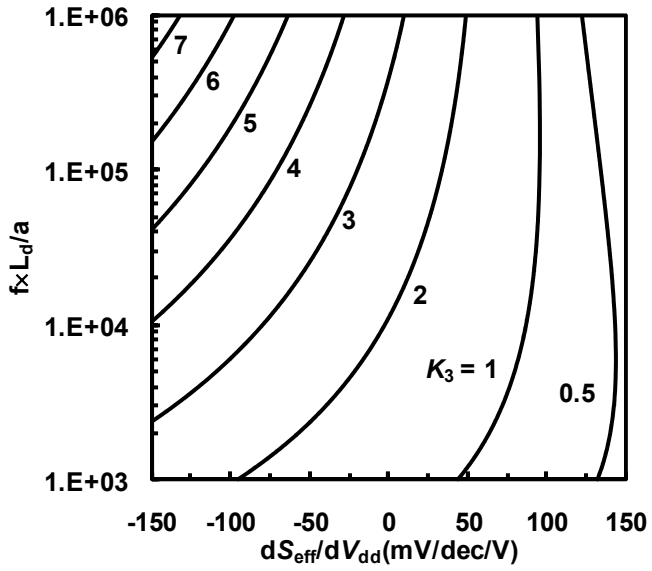


Fig. 2.3. The optimal I_{on}/I_{off} of a generalized logic device depends on the logic style ($f \times L_d/a$) and the log-concavity of the device transfer characteristics (dS_{eff}/dV_{dd}). dS_{eff}/dV_{dd} are approximately zero for most MOSFET-like devices, therefore $I_{on}/I_{off}=K_3 \times f \times L_d/a$, where K_3~ 2-8.

By comparing K_3 against K_1 and K_2 (Eqns. (2.11) and Eqn. (2.14)), the optimal I_{on}/I_{off} ratio is roughly the same for MOSFET-like devices. This is because, as was discussed earlier, most of the change in I_{on}/I_{off} (as a function of V_{dd}) occurs in the region with the largest the steepest effective subthreshold slope, and therefore the ratio is relatively insensitive to small changes in V_{dd}. Using the fact that the optimal I_{on}/I_{off} ratio is fixed across a wide range of switching devices, simple guidelines can be derived to assess the promise of MOSFET replacement devices, which is the focus of the following section.

2.5 Benchmarking CMOS Replacement Devices

2.5.1 General Considerations

As previously alluded to, for a given performance target and logic style, there exists an optimal I_{on}/I_{off} ratio to minimize the total energy, and this value is roughly the same for most MOSFET-like devices. Therefore if a logic device with a small subthreshold swing can reach the required I_{on}/I_{off} at a lower supply voltage, it will achieve the same performance with lower energy dissipation. With this said, merely focusing on the *steepest local subthreshold slope* (S) is misleading, since devices with very small S only at low current levels may not achieve the required performance. To compare the true energy efficiency, Fig. 2.4 summarizes a simple method to assess the promise of alternative devices for replacing MOSFETs. For a given the circuit topology, one first determines the optimal I_{on}/I_{off}~$2fL_d/a$. With a fixed reasonable off-state current (for instance I_{off}~1pA/μm) for both devices, one can then determine the required on state current. The supply voltage required for each device to reach such on state current is then graphically determined. With all these parameters determined, the new device is both faster and energy more efficient than the MOSFETs if it can achieve the required I_{on}/I_{off} at a lower V_{dd} (i.e. with a lower S_{eff}). Notice that a device with a small S value at low current levels but which require a large V_{dd} to reach the required I_{on}/I_{off} for the performance target, like the one shown in Fig. 2.4, does not improve the overall energy efficiency. Furthermore, the point at which the I_{ds}-V_{gs} curves intersect roughly corresponds to the point where the energy-delay curves cross over.

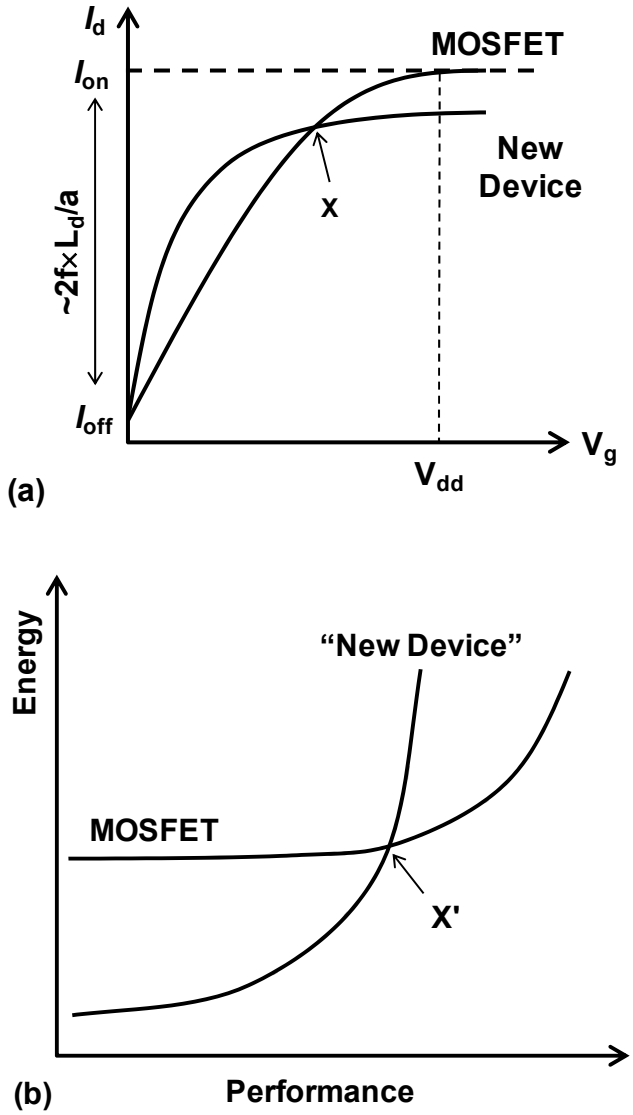


Fig. 2.4 (a) For a given circuit topology, the optimal I_{on}/I_{off} is set (~$2f \times L_d/a$). To compare any new device against the MOSFET, a reasonable off-state current is first fixed. The new device will be energy more efficient if it can achieve the required current ratio at a lower V_{dd}, i.e. at a lower effective subthreshold slope (S_{eff}). **(b)** The V_{dd} value at the point where the transfer characteristics intersect in (a) (denoted as X) is roughly the same for the intersection point in the energy-performance space (point X').

2.5.2 Additional Considerations

Some alternative devices (e.g., the NEMFET and the IMOS) do not begin to conduct current until sometime after the control voltage arrives, leading to an additional setup delay (t_{su}). The delay time is therefore:

$$t_{delay} = t_{RC} + t_{su} = \frac{L_d C f V_{dd}}{2I_{on}}(1 + \gamma_{su}), \text{ where } \gamma_{su} \equiv t_{su}/t_{RC} \quad (2.18)$$

Following a similar procedure as shown in Appendix I, the optimal I_{on} and I_{off} is determined by the following condition:

$$\frac{4a}{L_d f(1+\gamma_{su})}\frac{I_{on}}{I_{off}} + 1 + \frac{I_{on}}{I_{off}}\frac{dI_{off}}{dI_{on}} = 0 \quad (2.19)$$

Comparing Eqn. (2.18) and Eqn. (2.19) with Eqn. (2.2) and Eqn. (2.3), we can see that these devices not only need to have $1+\gamma_{su}$ times higher on current, but also $1+\gamma_{su}$ times higher I_{on}/I_{off} (i.e. $1+\gamma_{su}$ times smaller in S_{eff} than the MOSFET) to compensate for the increased leakage energy.

Furthermore, for devices with a large setup time, circuit topologies with short logic depth but large fan-out per stage are preferred to minimize the overall delay. For the case of a simple buffer chain, the optimal fan out per stage f_{opt} and optimal logic depth L_{opt} can be estimated [2.18] by the following equations:

$$f_{opt} = exp\left(1 + \frac{\gamma_{su}}{f_{opt}}\right), L_{opt} = ln\,(C_L/C_{in})/ln\ f_{opt} \quad (2.20)$$

where C_{in} in the input capacitance of the inverter chain.

In addition to the non-zero setup time, many MOSFET replacement devices may have gate or other parasitic capacitance (denoted as *C'*); the switching delay for these devices are:

$t_{delay} = \frac{L_d C f V_{dd}}{2 I_{on}} \gamma_c$, where $\gamma_c \equiv C'/C$ (2.21)

For these devices, the optimal I_{on}/I_{off} remains unchanged, but the on current can be γ_c times lower for the same performance target. However, it should be noted that for a given switching energy, even if a device has a low gate capacitance, it may not necessarily allow for higher supply voltage. This is because the device layout area impacts capacitance (*e.g.* of interconnect wires) and thereby impacts circuit switching energy and constrains the supply voltage that can be used.

In setting device and circuit design parameters to optimally balance leakage and dynamic energies, it is critical to consider the impact of variations. For example, since I_{off} varies exponentially with the V_T of a MOSFET, the average I_{off} is much higher than $I_{off}(V_{Taverage})$; thus, maintaining the appropriate energy ratio requires a lower nominal I_{off}. In contrast to energy, the performance of a synchronous digital circuit is set by the critical paths. While there is some summing of delay variations along the path, the paths are not very long, so variations remain. Thus I_{on} must be increased to ensure all paths meet the performance target for the worst-case variations.

Applications with low performance demands or large amounts of parallelism can tolerate reduced device performance [2.19], so that V_{dd} can be scaled more aggressively (with margin for variation) to reduce energy. It can be shown that for these applications, I_{on}/V_{dd} is not as critical as the minimum supply voltage $V_{dd,min}$, which depends only on maintaining the optimal I_{on}/I_{off} ratio and is proportional to S_{eff}.

Before moving on, it should be noted that even if a device has low S_{eff} but requires a non-zero output voltage (V_{ds}) [2.6, 2.14] to conduct, it may not improve the overall energy efficiency. This is because digital gates built which such a device would either dissipate significant static power, or would be significantly constrained in terms of the number of devices that can be connected in series.

2.6 TFET Comparisons with CMOS: An Example

To illustrate how the aforementioned methodology can be used to assess the promise of a MOSFET replacement device, we herein use the TFET as an example and compare it against the MOSFET. The TFET (Fig. 2.5) utilizes band-to-band tunneling (BTBT) current to achieve a more abrupt on-to-off transition than what is achievable through thermionic emission. For the purposes of this study, it is adequate to simply approximate I_{ds} using the band-to-band tunneling model [2.20, 2.21], which predicts:

$$I_{ds} = AE_s exp\left(-\frac{\pi\sqrt{m^*}E_g^{3/2}}{2\sqrt{2}q\hbar E_s}\right) = AE_s exp\left(-\frac{B}{E_s}\right) \tag{2.22}$$

where E_{s} is the electric field in the region where the tunneling occurs, which can be estimated for a source-tunneling FET by the following simple expression:

$$E_s = \frac{(V_{gs}+V_{tunnel})}{\kappa t_{ox}} \tag{2.23}$$

where qV_{tunnel} is the minimum energy-band bending needed for band-to-band tunneling to occur, κ is the ratio of the semiconductor permittivity to the gate oxide permittivity and t_{ox} is the gate-oxide thickness over the source [2.22].

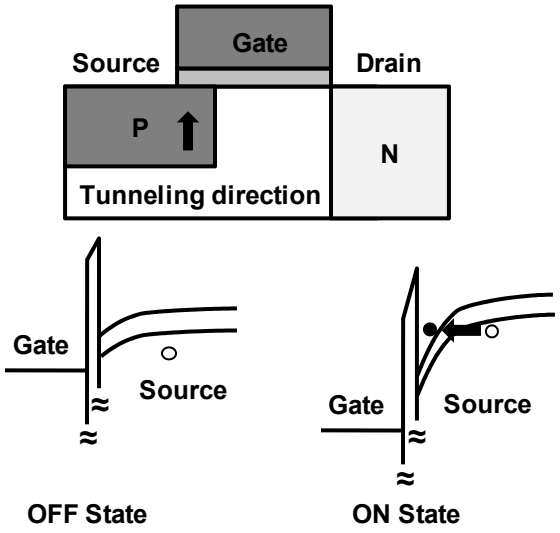


Fig. 2.5 A schematic diagram and the energy band diagram in the OFF/ON state of a source-tunneling field effect transistor.

This band-to-band-tunneling current model is used as it matches well with the measured data of the germanium-source TFET [2.23] and the BTBT off-state leakage current in silicon MOSFETs [2.22]. Note that a TFET has a very small S value at low current levels, but that S increases as I_{ds} increases. Furthermore, at high V_{dd} (~1V) values, a silicon TFET has a significantly lower on-state current I_{on} (~1μA/μm at 1V) than a MOSFET (1mA/μm at 1V). Current research efforts focus on improving I_{on}, for instance, by using a smaller-bandgap material such as germanium [2.23, 2.24]. Therefore, two representative TFET technologies are compared herein: a normal low I_{on} TFET technology ("low I_{on} TFET technology") and an advanced TFET technology that provides a high I_{on} ("high I_{on} TFET technology") to show the implications of the energy-performance analysis. The

device design parameters and the I-V characteristics of both TFET devices are shown in Table I and in Fig. 2.6.

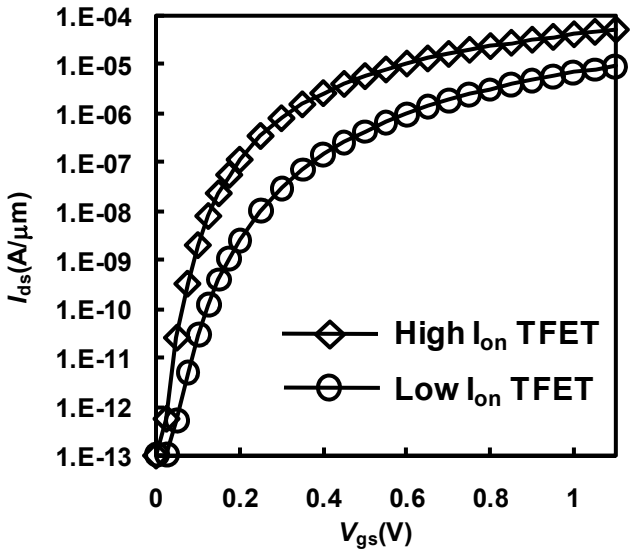


Fig. 2.6 Simulated I-V characteristic of two representative TFET technologies: the "low I_{on}" and "high I_{on}" TFET technologies. The V_{tunnel} values for low I_{on} and high I_{on} TFETs are respectively 0.13V and 0.07V.

Assuming V_{tunnel} is a parameter that can be adjusted by utilizing advanced processing technologies (e.g. gate work function engineering), the optimal I_{on}/I_{off} can readily be found from Eqn. (2.22) and Eqn. (2.23):

$$\frac{I_{on}}{I_{off}}\frac{dI_{off}}{dI_{on}} \approx \frac{(0.1\sim0.3-V_{dd})}{V_{tunnel}}\left(1+\frac{B'}{V_{tunnel}}\right) \qquad (2.24)$$

where $B'=\kappa t_{ox}B$. For typical values of B'=3V, V_{tunnel} =0.2V and V_{dd} =0.5V, $I_{on}/I_{off} \times dI_{off}/dI_{on}$ can be computed from Eqn. (2.24) and its value is roughly -30; the optimal I_{on}/I_{off} is therefore roughly $8\times L_d f/a$, which matches the prediction of Eqn. 2.17.

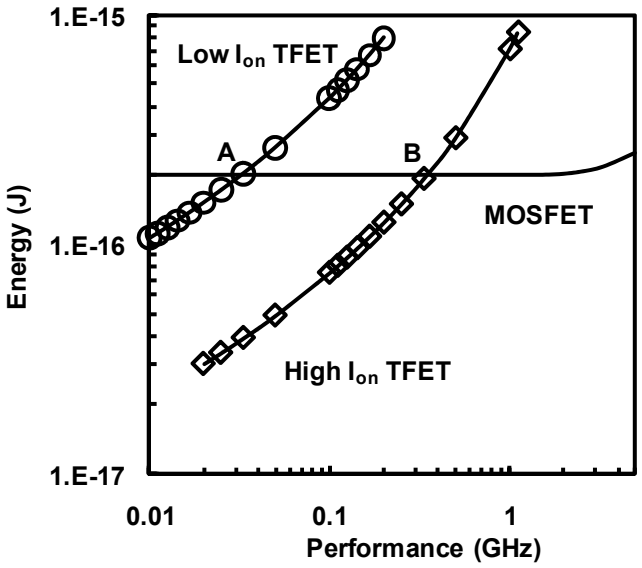


Fig. 2.7 Energy-Performance comparison of 30-stage FO4 65nm CMOS inverter chain with a 65nm-equivalent two different TFET technology.

With this in mind, Fig. 2.7 shows the simulated energy-performance comparison of a 65nm MOSFET vs. a 65nm-equivalent TFET, for a 30-stage fan-out-4 inverter chain (transition probability=0.01, with optimized V_{dd}, V_T, and V_{tunnel} values). The device parameters for the MOSFET (Table 2.2) are chosen according to the ITRS specifications for the 65nm LSTP technology [2.25]. With these circuit design parameters (L_d,=30 a=0.01 and f=4) given, the optimal I_{on}/I_{off} ratio for both CMOS and TFET is approximately 2.4×10^4. The S_{eff} values for the two TFETs to reach this I_{on}/I_{off} ratio at different I_{on}/V_{dd} values are plotted in Fig. 2.8. By overlapping the I-V characteristics of the TFET with that of MOSFET (with I_{off} fixed at 0.1pA/μm for both devices), one can see that for relatively slow (sub-50 MHz) applications where I_{on}/V_{dd} is not critical, both high-I_{on} and low-I_{on} TFETs have smaller S_{eff} values and hence can be more energy-efficient than a MOSFET.

For moderate (50-500MHz) performance applications, only the high-I_{on} TFET can achieve S_{eff} < 100mV/dec ; and for high-performance applications (beyond 1GHz) requiring I_{on} >100μA/μm, both TFET technologies have S_{eff} >100mV/dec and they therefore would consume more energy than a MOSFET. At the I_{on}/V_{dd} where the TFET achieves the same S_{eff} value as a MOSFET (denoted as A' and B' in Fig. 2.8a and A" and B" in Fig. 2.8b), both devices consume roughly the same amount of energy (points A and B in Fig. 2.7). Based on this analysis, the high-I_{on} TFET technology appears to be compelling for low power applications up to ~100MHz.

This simplified energy-performance analysis thus far assumes logic devices just drive other devices; in reality, however, extrinsic wire capacitance (C_w) must be considered in the analysis too, especially if the device has an area overhead. Fig. 2.9 shows the sensitivity of the energy consumption as a function of the wiring capacitance. For slow (50MHz) applications, a TFET operates at a lower V_{dd} than a MOSFET and hence its energy consumptions ($\propto C_w V_{dd}^2$) is less sensitive to C_w; for high performance applications (1GHz), the TFET operates at a higher V_{dd} and hence the energy consumption is more sensitive to C_w.

	Low I_{on} TFET	High I_{on} TFET
Physical Gate Length (nm)	45nm	
A (A/V/µm)	3.82E-13	7. 26E-13
B (MV/cm)	7.78	4.09
Equivalent oxide thickness (nm)	1	
κ	4	
C (fF)	0.741fF	

Table 2.1. Summary of device parameters for the 65nm equivalent TFET technology used in this work.

	Value
Physical Gate Length (nm)	45nm
α	1.18
n	1.667
I_s (µA/µm)	8.259
C (fF)	0.741fF

Table 2.2. Summary of device parameters for the 65nm MOSFET technology used in this work.

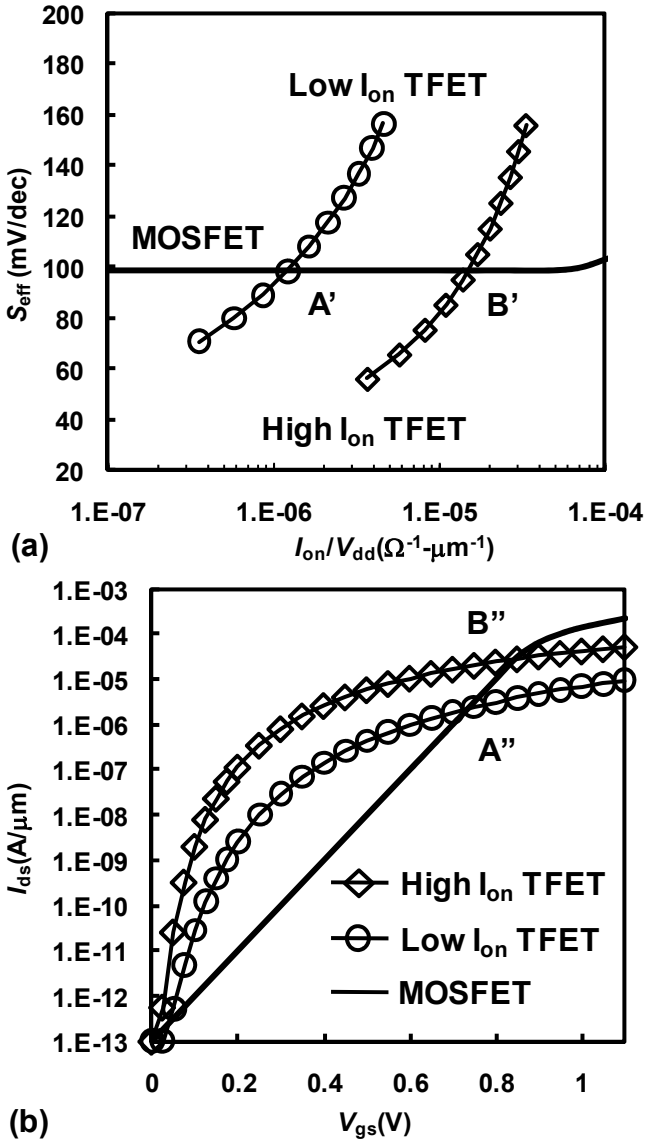


Fig. 2.8 (a) Effective subthreshold values of TFET and MOSFET as different I_{on}/V_{dd} values for a given circuit topology. For low-performance applications, a TFET can achieve the optimal I_{on}/I_{off} ratio with a small S_{eff} and therefore it is more energy efficient than a MOSFET Note that the cross-over points (denoted as A' and B' in the figure) roughly correspond to the cross-over points (denoted as A and B in the Fig. 2.7) in the energy-performance space and in transfer characteristics (denoted as A" and B" (b)).

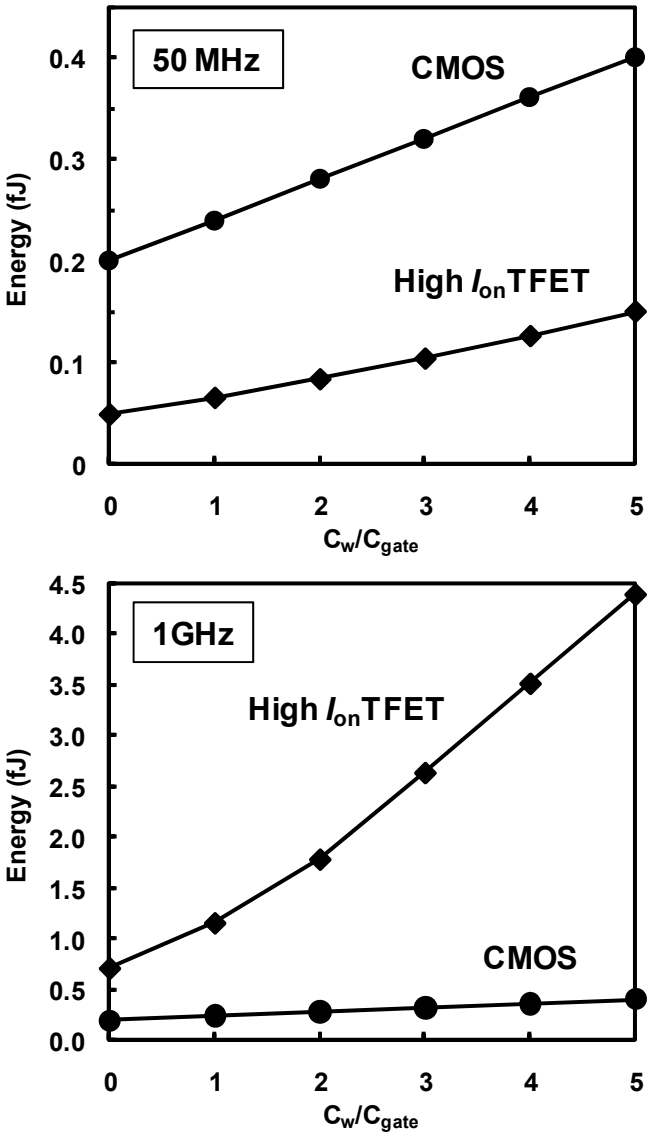


Fig. 2.9 Impact of wiring capacitance on energy dissipation. For low performance application, a TFET is operated at a lower V_{dd} and wiring capacitance is less of an impact to the energy consumption. For high performance application, however, a TFET needs a higher V_{dd} to provide for the high on-state current. Therefore the energy consumption is more sensitive to wiring capacitance.

2.7 Conclusion

In this chapter, a simple evaluation guideline is established to assess the promise of new device technologies. Based upon the energy-performance tradeoff of logic gates, it is shown that the optimal I_{on}/I_{off} ratio for logic devices depends largely only on the circuit topology, and that this optimal I_{on}/I_{off} stays roughly constant across a wide range of switching devices. With this optimal I_{on}/I_{off} ratio in mind, it is then shown that the effective subthreshold swing, rather than the steepest local subthreshold swing determines the energy efficiency of logic devices. As a concrete example, this methodology is used to compare TFETs against MOSFETs, showing that TFETs may offer substantial (~5×) energy savings for performance up to the 100MHz range.

Appendix I. Energy-Delay Optimization of the Generalized Logic Device

To minimize energy consumption subject to a delay constraint, the Lagrange multiplier method is used to set up the following expression:

$$L(V_{dd}, I_{on}) = E(V_{dd}, I_{on}) - \lambda\left(t_{delay} I_{on} - \frac{L_d f C}{2} V_{dd}\right) \quad \text{(A.1)}$$

Where $E(V_{dd},I_{on})$ is the energy consumption

$$E(V_{dd}, I_{on}) = a L_d V_{dd}^2 C f + L_d f I_{off} V_{dd} t_{delay} \quad \text{(A.2)}$$

And $D(V_{dd},I_{on})$ is the delay constraint

$$D(V_{dd}, I_{on}) = t_{delay} I_{on} - \frac{L_d f C}{2} V_{dd} \quad \text{(A.3)}$$

For a broad variety of different alternative devices, the off state leakage current depends only on I_{on} and V_{dd}. Hence differentiating $L(V_{dd},I_{on})$ with respect to V_{dd}:

$$\frac{\partial L}{\partial V_{dd}}: 2 a L_d V_{dd} C f + L_d f I_{off} t_{delay} + L_d f V_{dd} t_{delay} \frac{\partial I_{off}}{\partial V_{dd}} + \lambda \frac{L_d f C}{2} = 0 \quad \text{(A.4)}$$

which can be simplified to:

$$4 a V_{dd} + \frac{2 I_{off} t_{delay}}{C} + \frac{2 V_{dd} t_{delay}}{C} \frac{\partial I_{off}}{\partial V_{dd}} = -\lambda \quad \text{(A.5)}$$

As the total differential of I_{off} is:

$$dI_{off} = \frac{\partial I_{off}}{\partial V_{dd}} dV_{dd} + \frac{\partial I_{off}}{\partial I_{on}} dI_{on} \quad \text{(A.6)}$$

$\frac{\partial I_{off}}{\partial V_{dd}}$ can therefore be expressed by:

$$\frac{\partial I_{off}}{\partial V_{dd}} = \frac{dI_{off}}{dV_{dd}} - \frac{\partial I_{off}}{\partial I_{on}} \frac{dI_{on}}{dV_{dd}} = \frac{dI_{off}}{dI_{on}}\left(\frac{dI_{on}}{dV_{dd}}\right) - \frac{\partial I_{off}}{\partial I_{on}} \frac{dI_{on}}{dV_{dd}}$$

Since $t_{delay} = \frac{L_d f C}{2I_{on}} V_{dd}$, $\frac{dI_{on}}{dV_{dd}} = \frac{I_{on}}{V_{dd}}$. Therefore,

$$\frac{\partial I_{off}}{\partial V_{dd}} = \left(\frac{dI_{off}}{dI_{on}} - \frac{\partial I_{off}}{\partial I_{on}}\right)\frac{I_{on}}{V_{dd}} \tag{A.7}$$

Hence (A.5) is

$$4aV_{dd} + \frac{2I_{off}t_{delay}}{C} + L_d f V_{dd}\left(\frac{dI_{off}}{dI_{on}} - \frac{\partial I_{off}}{\partial I_{on}}\right) = -\lambda \tag{A.8}$$

On the other hand, if $L(V_{dd},I_{on})$ is differentiated with respect to I_{on}

$$\frac{\partial L}{\partial I_{on}} : L_d f V_{dd} t_{delay} \frac{\partial I_{off}}{\partial I_{on}} - \lambda t_{delay} = 0 \tag{A.9}$$

$$L_d f V_{dd} \frac{\partial I_{off}}{\partial I_{on}} = \lambda \tag{A.10}$$

Equating (A.8) and (A.10), we get

$$4aV_{dd} + \frac{2I_{off}t_{delay}}{C} + L_d f V_{dd}\left(\frac{dI_{off}}{dI_{on}} - \frac{\partial I_{off}}{\partial I_{on}}\right) = -L_d f V_{dd}\frac{\partial I_{off}}{\partial I_{on}} \tag{A.11}$$

which is equivalent to:

$$\frac{4aC}{2I_{off}t_{delay}} V_{dd} + 1 + \frac{L_d f V_{dd} C}{2I_{off}t_{delay}} \frac{dI_{off}}{dI_{on}} = 0 \tag{A.12}$$

Substituting the delay expression to (A.12), we finally obtain (2.3):

$$\frac{4a}{L_d f}\frac{I_{on}}{I_{off}} + 1 + \frac{I_{on}}{I_{off}}\frac{dI_{off}}{dI_{on}} = 0 \tag{A.13}$$

Appendix II. Derivation of Equation (2.15)

Equation (2.3) can equivalently be expressed by (2.15) by first knowing that:

$$\frac{I_{on}}{I_{off}}\frac{dI_{off}}{dI_{on}} = \frac{dlog_{10}I_{off}}{dlog_{10}I_{on}} = \frac{dlog_{10}I_{on} - S_{eff}^{-1}V_{dd}}{dlog_{10}I_{on}} = 1 - \frac{dS_{eff}^{-1}V_{dd}}{dlog_{10}I_{on}} \quad \text{(A.14)}$$

And for a fixed performance, (2.2) can be expressed by:

$$log_{10}\, I_{on} = log_{10}\left(\frac{L_d C \cdot f}{2t_{delay}}\right) + log_{10}\, V_{dd} \quad \text{(A.15)}$$

Therefore, using the chain rule, (A.14) can be simplified:

$$1 - \frac{dS_{eff}^{-1}V_{dd}}{dlog_{10}V_{dd}} = 1 - \left(V_{dd}S_{eff}\frac{dS_{eff}^{-1}}{dV_{dd}} + 1\right)S_{eff}^{-1}V_{dd}\, ln\, 10 \quad \text{(A.16)}$$

After simplification, (A.16) becomes:

$$= 1 - ln\left(\frac{I_{on}}{I_{off}}\right)\left(1 - log_{10}\left(\frac{I_{on}}{I_{off}}\right)\frac{dS_{eff}}{dV_{dd}}\right) \quad \text{(A.17)}$$

Substituting (A.17) into (2.3), we obtain (2.15):

$$\frac{4a}{L_d f}\frac{I_{on}}{I_{off}} + 2 - ln\left(\frac{I_{on}}{I_{off}}\right)\left(1 - log_{10}\left(\frac{I_{on}}{I_{off}}\right)\frac{dS_{eff}}{dV_{dd}}\right) = 0 \quad \text{(A.18)}$$

2.8 References

[2.1] B. Nikolic, "Design in the power-limited scaling regime," *IEEE Trans. Electron Devices*, vol. 55, pp. 71, Jan. 2008.

[2.2] T. Baba, "Proposal for surface tunnel transistor", *Jpn J Appl Phys* 31 (1992) (4B), pp. L455–L457

[2.3] W. Y. Choi, B.-G. Park, J. D. Lee, and T.-J. K. Liu, "Tunneling field-effect transistors (TFETs) with subthreshold swing (SS) less than 60 mV/dec," *IEEE Electron Device Lett.*, vol. 28, no. 8, pp. 743–745, Aug. 2007.

[2.4] K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, "I-MOS: A novel semiconductor device with a subthreshold slope lower than *kT/q*," in *IEDM Tech. Dig.*, 2002, pp. 289–292.

[2.5] C. Shen, J.-Q.Lin, E.-H.Toh, K.-F. Chang, P. Bai, C.-H. Heng, G.S.Samudra, and Y.-C.Yeo, 'On the performance limit of impact ionization transistors,' in *IEDM Tech. Dig.*, 2007, pp. 117-120.

[2.6] W. Y. Choi , J. Y. Song , J. D. Lee , Y. J. Park and B.-G. Park "A novel biasing scheme for I-MOS (impact-ionization MOS) devices," *IEEE Trans. Nanotechnol.*, vol. 4, pp. 322, May 2005.

[2.7] N. Abele, N. Fritschi, K. Boucart, F. Casset, P. Ancey, and A. M. Ionescu, "Suspended-gate MOSFET: Bringing new MEMS functionality into solid-state MOS transistor," in *IEDM Tech. Dig.*, 2005, pp. 1075–1077.

[2.8] H. Kam, D. T. Lee, R. T. Howe, and T.-J. King, "A new nano-electromechanical field effect transistor (NEMFET) design for low-power electronics," in *IEDM Tech. Dig.*, 2005, pp. 463–466.

[2.9] K. Akarvardar, C. Eggimann, D. Tsamados, Y. Singh Chauhan, G. C. Wan, A.M. Ionescu, R.T. Howe, and H.-S.P. Wong, "Analytical Modeling of the Suspended-Gate FET and Design Insights for Low-Power Logic," IEEE Trans. Electron Devices, vol. 55, no. 1, pp. 48-59, Jan. 2008.

[2.10] F. Chen, H. Kam, D. Markovic, T.J. King, V. Stojanovic, and E. Alon, "Integrated Circuit Design with NEM Relays," in *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, 2008, pp. 750-757

[2.11] K. Akarvardar, D. Elata, R. Parsa, G. C. Wan, K. Yoo, J. Provine, P. Peumans, R. T. Howe, H.-S. P. Wong, "Design Considerations for Complementary Nanoelectromechanical Logic Gates," in *IEDM Tech. Dig.*, 2007, pp. 299-302.

[2.12] S. Salahuddin and S. Datta, "Use of negative capacitance to provide a sub-threshold slope lower than 60 mV/decade," Nanoletters, vol. 8, No. 2, 2008.

[2.13] S. Salahuddin and S. Datta, "Can the subthreshold swing in a classical FET be lowered below 60 mV/decade?," in *IEDM Tech. Dig.*, 2008, pp. 693–696

[2.14] A. Padilla, C.W. Yeung, C. Shin, M.H. Cho, C. Hu and T.-J. King Liu, "Feedback FET: A Novel Transistor Exhibiting Steep Switching Behavior at Low Bias Voltages," in *IEDM Tech. Dig.*, 2008, pp. 171–174

[2.15] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for low-power and high-speed applications," in *Proc. Asia South Pacific Design Automation Conf.*, Jan. 2000, pp. 469–474.

[2.16] 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. 50 n.9, p.1778-1786 Sept. 2005.

[2.17] S. Hanson , B. Zhai , K. Bernstein , D. Blaauw , A. Bryant , L. Chang , K. K. Das , W. Haensch , E. J. Nowak , D. M. Sylvester, Ultralow-voltage, minimum-energy CMOS, IBM Journal of Research and Development, vol.50 n.4/5, p.469-490, July 2006

[2.18] J. Rabaey, A. Chandrakasan and B. Nikolic, Digital Integrated Circuits: A Design Perspective, second edition, Prentice Hall, NJ, 2003

[2.19] A. P. Chandrakasan , S. Sheng and R. W. Brodersen "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473, Apr. 1992

[2.20] J. L. Moll, *Physics of Semiconductors*, New York: McGraw-Hill, 1964.

[2.21] Q. Zhang , W. Shao and A. Seabaugh "Low-subthreshold-swing tunnel transistors," *IEEE Electron Device Lett.*, vol. 27, pp. 297, Apr. 2006.

[2.22] J. Chen, T. Y. Chan, P. K. Ko, and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Electron Device Lett.*, vol. EDL-8, no. 11, pp. 515–517, Nov. 1987.

[2.23] S. H. Kim, H. Kam, C. Hu and T.-J. King-Liu, "Germanium-Source Tunnel Field Effect Transistors with Record High I_{ON}/I_{OFF}." in Symposium on VLSI Technology *Tech. Dig*, 2009., pp.178 – 179

[2.24] K. K. Bhuwalka, J. Schulze, and I. Eisele, "Performance enhancement of vertical tunnel field-effect transistor with SiGe in the δp^+ layer," *Jpn. J. Appl. Phys.*, vol. 43, no. 7A, pp. 4073-4078, Jul. 2004.

[2.25] *The International Technology Roadmap for Semiconductors (ITRS)*, 2007. [Online]. Available: http://public.itrs.net

Chapter 3

Nano-Electro-Mechanical Field Effect Transistor Design

3.1 Introduction

In Chapter 2, it was shown that the effective subthreshold swing (rather than the steepest, local subthreshold swing) determines a device's energy efficiency. With this consideration in mind, alternative transistor designs which offer perfectly abrupt off-to-on transition to provide for high on/off current ratio with a smaller supply voltage (i.e. small S_{eff} value) are attractive for energy-efficient electronics. One such device is the nano-electro-mechanical field effect transistor (NEMFET) [3.1-3.10], which utilizes the pull-in and release behavior of a mechanical beam to achieve a perfectly abrupt switching transition, and an effective subthreshold swing S that is less than 60mV/dec.

In addition to low-power digital logic applications, NEMFETs also have been proposed for analog circuit applications such as resonators and sensors [3.8-3.10]. The motion of the mechanical gate (or body) changes the equivalent gate-oxide thickness and hence the transistor current, so that a mechanical signal can be

effectively converted into an electrical signal with high transduction efficiency [3.9].

For digital logic applications, the pull-in voltage V_{pi} and the release voltage V_{rl} of a NEMFET are important performance parameters since they determine the turn-on and turn-off voltages of the transistor, respectively [3.1-3.7]. Ideally, pull-in should occur in the sub-threshold regime of operation, *i.e.* V_{pi} should be less than V_{T} (defined as the gate-to-source voltage V_{gs} at which the transistor current becomes linearly dependent on V_{gs}) to achieve the highest on/off current ratio for a given gate-voltage swing. $|V_{\mathrm{rl}}|$ should be greater than zero to ensure that the transistor turns off properly, *i.e.* that it is in the off state for $V_{\mathrm{gs}} = 0\mathrm{V}$. On the other hand, for analog circuit applications [3.8-3.10], V_{pi} sets an upper limit for the bias voltage and should be much higher than V_T to allow for a large DC bias current. Thus, an accurate model for the pull-in/release voltages, as well as the threshold voltage, is needed to guide NEMFET design for various applications.

Previous modeling efforts used a simple lumped parameter model [3.1-3.3] to study the behavior of NEMFETs. While this approach provides intuition for NEMFET design, it does not account for two-dimensional effects, *e.g.* a non-uniform actuation gap thickness due to bending of the gate electrode. Furthermore, previous efforts lacked a discussion of the conditions necessary for pull-in/release to occur in the sub-threshold *vs.* inversion regime of FET operation. To address these shortcomings, in this chapter, the Euler-Bernoulli equation (applicable to mechanical beams and widely used for modeling micro-electromechanical systems

[3.11-3.14]) is solved simultaneously with the Poisson equation to accurately model V_{pi} and V_{rl} of a mechanically gated FET. Using this model, the effects of various device design parameters (*e.g.* body doping concentration, gate stiffness, as-fabricated actuation gap thickness, and source-to-body bias voltage) and surface adhesion force are assessed in Section 3.3. A unified pull-in/release voltage model which accounts for these effects is then provided in Section 3.4.

Although the NEMFET effective subthreshold swing is reduced by utilizing the pull-in effect, the presence of an air-gap in the transistor drastically worsens the short channel effects; the impact of such effects on the NEMFET's scalability for logic applications is discussed in Section 3.5.

3.2 Physics of NEMFET Operation

Fig. 3.1 illustrates the NEMFET structure, which is essentially a metal-oxide-semiconductor field effect transistor (MOSFET) with a movable gate electrode that can be physically separated from the gate dielectric layer by an air gap (or vacuum gap). As shown in the cross-sectional schematic in Fig. 3.2a, the suspended gate is a doubly-clamped beam anchored on each side of the semiconductor channel.

Fig. 3.1 also presents the NEMFET operation. In the off state, an air gap separates the gate from the gate dielectric; as V_{gs} increases, the electrostatic force attracts the gate to the gate dielectric. In the on state (Fig. 3.1d) where the gate-to-source voltage is greater than the pull-in voltage (V_{pi}), the gate is pulled down and is in contact with gate dielectric, which increases the gate-to-channel coupling,

Once the gate is pulled in, the thin dielectric thickness ensures that the electrostatic force is larger than the spring restoring force, and therefore the NEMFET exhibits hysteretic switching (Fig. 3.1 (d)) - i.e. the release voltage V_{gs}=V_{rl} value is lower than V_{pi}.

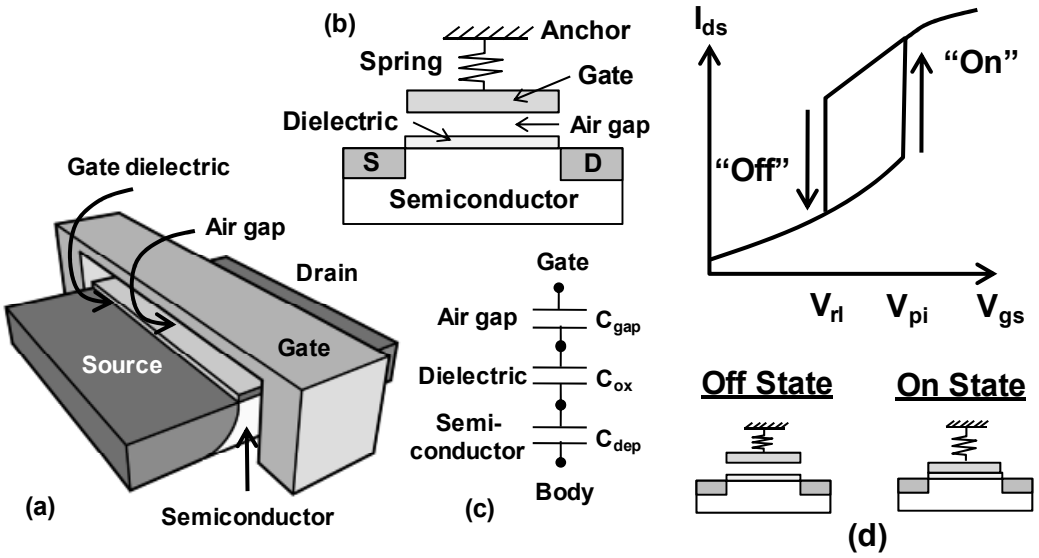


Fig. 3.1. (a) Schematic diagrams of a nano-electro-mechanical field effect transistor (NEMFET). (b,c) Simple lumped-parameter model for a NEMFET. (d) the abrupt pull-in/ release of the gate electrode provide for perfectly abrupt on/ off transitions.

The change in the gap thickness at V_{gs}=V_{pi} (or V_{gs}=V_{rl}) can be equivalently described as a dynamic reduction (or increase) in the threshold voltage V_T. If V_T for a MOSFET is defined as the value of V_{gs} for which the channel is just barely inverted, then the change in V_T due to the movement of the gate at V_{gs}= V_{pi} and V_{gs}= V_{rl} is approximately $|\Delta V_T| \approx t_{gap} \frac{\sqrt{4\varepsilon_{si} q N_a \phi_b}}{\varepsilon_o}$ [3.1], where t_{gap} is the as-fabricated air-gap thickness, N_a is the body doping concentration, and $2\phi_b$ is the channel potential at the onset of strong inversion in the channel. Since the threshold voltage

changes abruptly as $|V_{gb}|$ is increased above V_{pi}, (or decreased below V_{rl}) the I_d-V_g characteristic of the NEMFET exhibits an extremely steep (nearly infinite) subthreshold slope [3.1-3.4] at V_{pi} and V_{rl}. To achieve low S_{eff} value, the supply voltage must be scaled down. To achieve this goal, V_{pi} and V_{rl} need to be minimized. In the pursuit of this goal and to facilitate NEMFET design, this section aims at developing an accurate model for V_{pi} and V_{rl}.

3.2.1 Lumped Parameter Model

A simple lumped parameter model previously has been used to study the behavior of NEMFETs [3.1-3.3]. In this model, as depicted in Fig. 3.1b, the gate is treated as a simple linear spring (with a characteristic spring constant k) suspended over the semiconductor channel. The gate and the channel form a parallel-plate capacitor with an equivalent air-gap that is uniform in thickness across the transistor channel. The details of this lumped parameter model are well covered in the literature [3.1-3.3]; the key results are summarized herein.

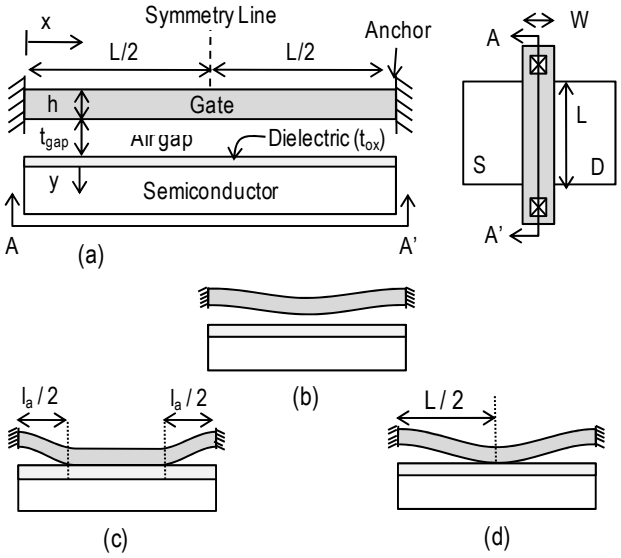


Fig. 3.2: (a) Schematic diagrams illustrating the physical parameters of the NEMFET. (b-d) Shapes of the deflected gate beam, corresponding to the different regions of NEMFET operation: (b) Before pull-in, (c) $V_{gs} > V_{pi}$, (d) $V_{gs} = V_{rl}$.

Current flow in the semiconductor channel is controlled by the voltage applied between the gate and the source, V_{gs}. Because the gate beam is assumed not to bend, the voltage drops across the air gap V_{gap}, the dielectric V_{ox} and the semiconductor surface potential ϕ_s are assumed to be independent of x:

$$V_{gs} + V_{sb} = V_{fb} + V_{gap} + V_{ox} + \phi_s = V_{fb} + V_{eff} + \phi_s \qquad (3.1)$$

where the flat-band voltage V_{fb} is a function of the channel dopant concentration N_a: $V_{fb} = \Phi_m - \Phi_s = C - \frac{kT}{q} ln\left(\frac{N_a}{n_i}\right)$ where C is a constant and Φ_m and Φ_s are the work functions of the gate and the semiconductor channel, respectively. $V_{eff} = V_{gap} + V_{ox}$ is the voltage drop across an equivalent air gap. If the gate beam is not pulled in, the spring restoring force is equal to the electrostatic force:

$$k\left(t_{gap}+\frac{t_{ox}}{\kappa_{ox}}-g\right)=\frac{\varepsilon_o WLV_{eff}^2}{2g^2}, \qquad k\approx 39.5\frac{EWh^3}{L^3} \tag{3.2}$$

where $g=d+t_{gap}+t_{ox}/\kappa_{ox}$ is the equivalent air-gap thickness, d<0 is the displacement of the mechanical gate, t_{ox} and κ_{ox} are the thickness and dielectric constant of the gate dielectric, k is the spring constant for a clamped-clamped beam, h and W and L are the thickness width and length of the gate beam, respectively. E is the Young's modulus of the gate beam material. Non-ideal effects such as dielectric charges and fringing capacitances are assumed to be negligible.

To ensure proper NEMFET operation, the spring restoring force must be significantly greater than the surface adhesion force, F_a. In the absence of capillary forces, the adhesive interactions are dominated by the attractive van der Waals force between non-contacting surfaces [3.15]:

$$F_a\cong\frac{2\Gamma}{d_o}WL,\ for\ 0<|d|\le t_{gap} \tag{3.3}$$

where Γ is the adhesion energy per unit area and d_o is the average distance between the two surfaces.

V_{eff} can be computed from the amount of areal charge in the channel, Q_s:

$$V_{eff}=-\frac{g}{\varepsilon_o}Q_s \tag{3.4}$$

where Q_s depends on the channel potential ϕ_s. To simplify the analysis, we can approximate Q_s as follows [3.16]:

$$Q_s\cong\begin{cases}-\sqrt{2\varepsilon_{si}qN_a\phi_s} & \phi_s<2\phi_b+V_{sb};\ in\ depletion\\ -\sqrt{2\varepsilon_{si}N_ak_bT}e^{\frac{q}{2k_bT}(\phi_s-2\phi_b-V_{sb})} & \phi_s\ge 2\phi_b+V_{sb};\ in\ inversion\end{cases} \tag{3.5}$$

where ε_{si} is the dielectric permittivity of the silicon channel material, N_a is the channel dopant concentration, k_b is the Boltzmann constant, and $\phi_b \equiv (k_b T/q)\ln(N_a/n_i)$. Solving Eqns. (3.1-3.5) either numerically [3.1-3.2] or analytically [3.3] gives the position of the gate and semiconductor surface potential for different gate voltage biases, including V_{pi} and V_{rl}. The details of the solutions are discussed elsewhere [3.1-3.3]. We will use this model as a reference for comparison against the more accurate Euler-Poisson model.

3.2.2 Euler – Bernoulli Equation for the Mechanical Gate Beam Shape

To accurately model the switching behavior of the NEMFET, the non-uniform actuation gap thickness due to bending of the gate electrode must be taken into account. The various voltage drops are functions of the position along the beam length direction (x):

$$V_{gs} + V_{sb} = V_{fb} + V_{eff}(x) + \phi_s(x) \tag{3.6}$$

An applied gate voltage results in an electrostatic force on the mechanical gate. The shape of the gate beam can be found by solving the Euler-Bernoulli equation [3.17] and depends on the state of the beam: not pulled in ($V_{gs}<V_{pi}$), as illustrated in Fig. 3.2b; pulled in, as illustrated in Fig. 3.2c; or just at the point of release ($V_{gs}=V_{rl}$), as illustrated in Fig. 3.2d.

i. Beam not pulled in

Before pull-in occurs, *i.e.* when the gate and the gate-dielectric are not in contact, the shape of the mechanical gate is governed by the following equation:

$$EI\frac{d^4g(x)}{dx^4} = -\frac{\varepsilon_o W V_{eff}(x)^2}{2g(x)^2} - \frac{2\Gamma}{d_o}W \tag{3.7}$$

where I is the moment of inertia of the gate beam. Note that in Eqn. (3.7), non-ideal effects such as residual stress, vertical strain gradient, beam stiffening due to bending, dielectric charges and fringing capacitances are assumed to be negligible.

For a clamped-clamped beam, we can take advantage of symmetry (as depicted in Fig. 3.2a) to establish the boundary conditions at x=0 and x=L/2 as tabulated in Table 3.1.

ii. Beam pulled in

While the electrostatic force (F_{elec}) increases quadratically with increasing displacement, the spring restoring force (F_{spring}, which counteracts the electrostatic force) increases only linearly with displacement. Hence, there is a critical displacement beyond which F_{elec} is always larger than F_{spring}, causing the gap to close abruptly. This critical displacement has a corresponding value of $|V_{\text{gs}}|$ known as the "pull-in" voltage V_{pi}. Upon pull-in, a portion of the gate will zip into contact with the gate dielectric, as illustrated in Fig. 3.2c. Denoting l_{a} as the total length of the gate regions which are not in contact with the gate dielectric, the shape of the gate electrode can be determined in a piecewise manner:

$$\begin{cases} EI\frac{d^4g(x)}{dx^4} = -\frac{\varepsilon_o W V_{eff}(x)^2}{2g(x)^2} - \frac{2\Gamma}{d_o}W \; for \left|x - \frac{L}{2}\right| \geq \frac{L-l_a}{2} \\ g(x) = \frac{t_{ox}}{\kappa_{ox}} \; for \left|x - \frac{L}{2}\right| < \frac{L-l_a}{2} \end{cases} \tag{3.8}$$

iii. Beam at the point of release

With the beam pulled in, l_a increases as V_{gs} decreases. If V_{gs} is reduced to the release voltage (*i.e.* if $V_{gs}=V_{rl}$), then $l_a=L$ and the gate touches the gate dielectric only at $x=L/2$. Thus, Eqn. (3.8) is reduced to Eqn. (3.7).

The governing equations and boundary conditions for the different regimes of gate-beam operation are summarized in Table 3.1. To determine the shape of the gate, $V_{eff}(x)$ must be known. It is related to the semiconductor surface potential which can be found by solving the Poisson equation.

Region of Op.	Governing Equation	x=0	x=L/2 or $l_a/2$
Before Pull-in	$EI\frac{d^4g(x)}{dx^4}=-\frac{\varepsilon_o W V_{eff}(x)^2}{2g(x)^2}-\frac{2\Gamma}{d_o}W$	$g(x)=g_o$; $g'(x)=0$	$g'(L/2)=0$ $g^3(L/2)=0$
At release voltage			$g(L/2)=t_{ox}/\kappa_{ox}$ $g'(L/2)=0$ $g^3(L/2)=0$
Zip-in	$\begin{cases} EI\frac{d^4g(x)}{dx^4}=-\frac{\varepsilon_o W V_{eff}(x)^2}{2g(x)^2}-\frac{2\Gamma}{d_o}W \; for \left\lvert x-\frac{L}{2}\right\rvert \geq \frac{L-l_a}{2} \\ g(x)=\frac{t_{ox}}{\kappa_{ox}} \; for \left\lvert x-\frac{L}{2}\right\rvert < \frac{L-l_a}{2} \end{cases}$		$g(l_a/2)=t_{ox}/\kappa_{ox}$ $g'(l_a/2)=0$ $g^2(l_a/2)=0$

Table 3.1: The governing equation and the boundary conditions for the mechanical gate in different regions of operation (gate not pulled in, gate pulled in, and gate at point of release).

3.2.3 Poisson Equation for the Semiconductor Surface Potential

$V_{eff}(x)$ and $\phi_s(x)$ can be found by solving the Poisson equation [3.16], and the result is as follows:

$$V_{eff}(x) =$$

$$\frac{g(x)}{\varepsilon_o}\sqrt{2q\varepsilon_{si}N_a}\left[\left(\frac{k_bT}{q}e^{-\frac{q}{k_bT}\phi_s(x)}+\phi_s(x)-\frac{k_bT}{q}\right)+\frac{n_i^2}{N_a^2}\left(\frac{k_bT}{q}e^{\frac{q}{k_bT}(\phi_s(x)-V_{sb})}-\right.\right.$$

$$\left.\left.\phi_s(x)-\frac{k_bT}{q}e^{-\frac{q}{k_BT}V_{sb}}\right)\right]^{\frac{1}{2}} \qquad (3.9)$$

Note that Eqn. (3.9) assumes a long-channel MOSFET with small drain-to-source voltage (so that the drain voltage does not significantly impact the channel potential).

For a given $\phi_s(x)$, the channel potential $\phi(x,y)$ can be solved iteratively using the equation:

$$\frac{\partial\phi(x,y)}{\partial y}=\frac{-\sqrt{2q\varepsilon_{si}N_a}}{\varepsilon_{si}}\left[\left(\frac{k_bT}{q}e^{-\frac{q}{k_bT}\phi(x,y)}+\phi(x,y)-\frac{k_bT}{q}\right)+\frac{n_i^2}{N_a^2}\left(\frac{k_bT}{q}e^{\frac{q}{k_bT}(\phi(x,y)-V_{sb})}-\right.\right.$$

$$\left.\left.\phi(x,y)-\frac{k_bT}{q}e^{-\frac{q}{k_BT}V_{sb}}\right)\right]^{\frac{1}{2}} \qquad (3.10)$$

with the following boundary conditions at the channel surface and deep within the semiconductor:

$$\phi(x,y)|_{y=0}=\phi_s(x)\ ,\ \phi(x,y)|_{y\to\infty}=0. \qquad (3.11)$$

The static behavior of a NEMFET in each regime of gate-beam operation is obtained by solving Eqns. (3.6)-(3.11). Numerical simulations utilizing the finite difference and Newton-Raphson methods are implemented using Matlab 7; the details of the simulation methodology used to study the behavior of a NEMFET are summarized in Fig. 3.3

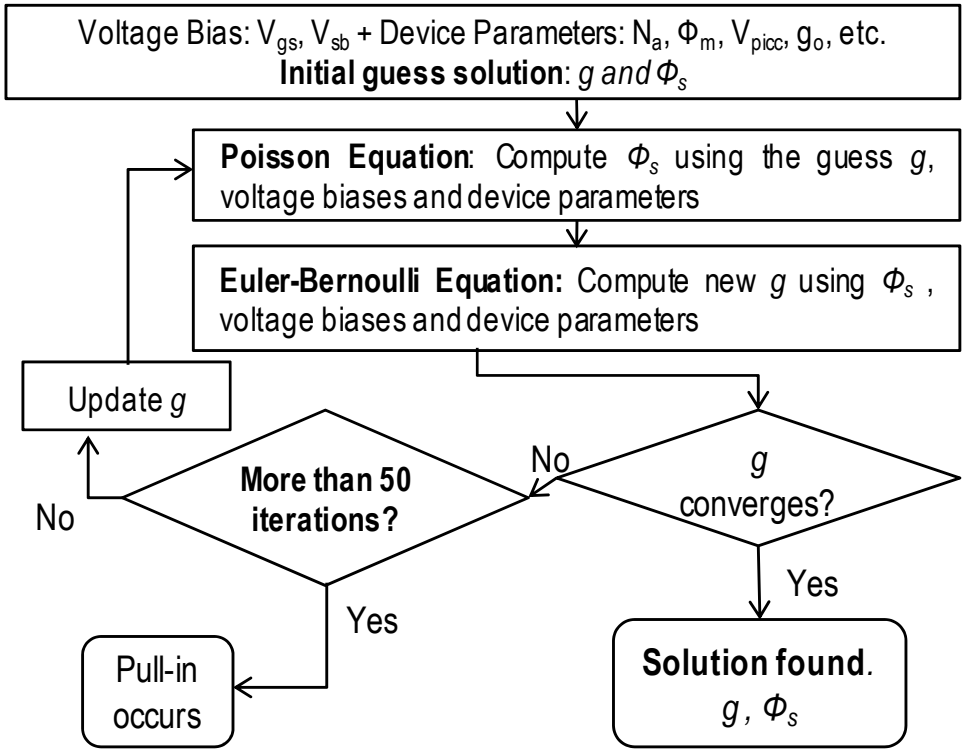


Fig. 3.3: Simulation methodology used to study the behavior of a NEMFET.

Parameter	Value	Parameter	Value
Young Modulus, E	160GPa	Dielectric constant κ_{ox}	3.9
Beam Width Gate Length, W	100nm	Dielectric thickness, t_{ox}	1nm
Beam Length Gate Width, L	800nm	Body biasing V_{sb}	0V
Gate thickness, H	20nm	Body Doping, N_a	1e17cm^{-3}
Gap Thickness, t_{gap}	9.7nm	Surface Roughness, d_o	0.1nm

Table 3.2: NEMFET device parameters used for the Euler-Poisson simulation study in this work.

3.3 Results and Discussion

3.3.1 Mechanical Gate Shape and Channel Potential

The static characteristics of a 100nm (channel length) NEMFET, with device parameters as shown in Table 3.2, were simulated using the Euler-Poisson model. Figs. 4a-4e show the equivalent gap thickness $g(x)$ and the channel potential contours $\phi(x,y)$ for V_{gs}=0V and 1.6V (before pull-in), V_{pi}^{-}=1.814V (just below the pull-in voltage), V_{pi}^{+}=1.814V(just above the pull-in voltage), and V_{rl}=0.308V, for V_{sb}=0V and Γ=0 μJ/m^2. At V_{gs}=0V, the mechanical gate is slightly deflected due to the electrostatic force induced by the built-in voltage ($-V_{fb}$); since the gap thickness $g(x)$ is non-uniform, the channel potential also varies with x.

As V_{gs} increases, the channel eventually becomes inverted (to be n-type). The channel potential contours for V_{gs}=1.6V and V_{gs}=1.814V are shown in Figs. 4b-4c. The contour line for $\phi(x,y)=2\phi_b$=0.8349V is shown in each plot to delineate the inversion region. Since the gate capacitance varies with x, the region of the channel near $x=L/2$ reaches strong inversion first, as V_{gs} is increased. The lateral extent of the inversion region then spreads as V_{gs} is further increased, as can be seen from Fig. 3.4c.

For $V_{gs}>V_{pi}$ (Fig. 3.4d), the gate is pulled in, with its central portion in contact with the gate dielectric. If V_{gs} were to be reduced back toward 0V within this regime of gate-beam operation, the length of the contacting region would decrease; at $V_{gs}=V_{rl}$, the gate would only be in contact with the gate dielectric at

$x=L/2$ (Fig. 3.4e). Any further reduction in V_{gs} would cause the gate to be released from the gate dielectric.

As will be explained below, the magnitudes of V_{pi} and V_{rl} and the channel condition at these gate bias voltages depend on various device parameters including the body doping concentration, gate work function, mechanical gate properties, gap thickness, source-to-body voltage, and surface adhesion force.

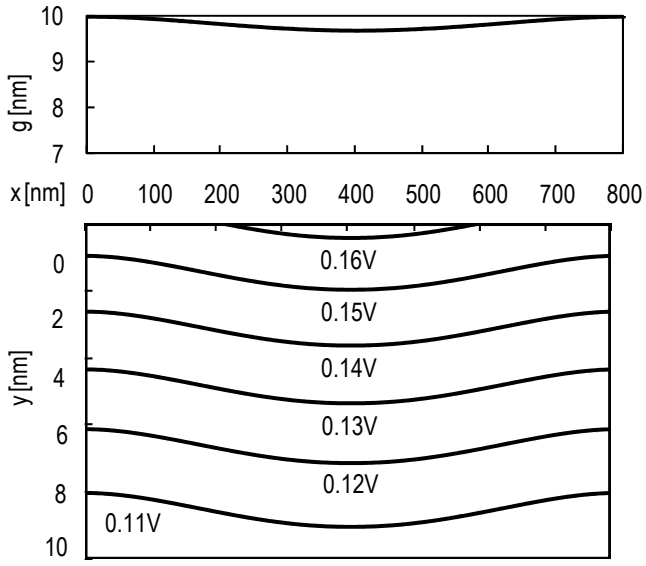


Fig. 3.4 (a)

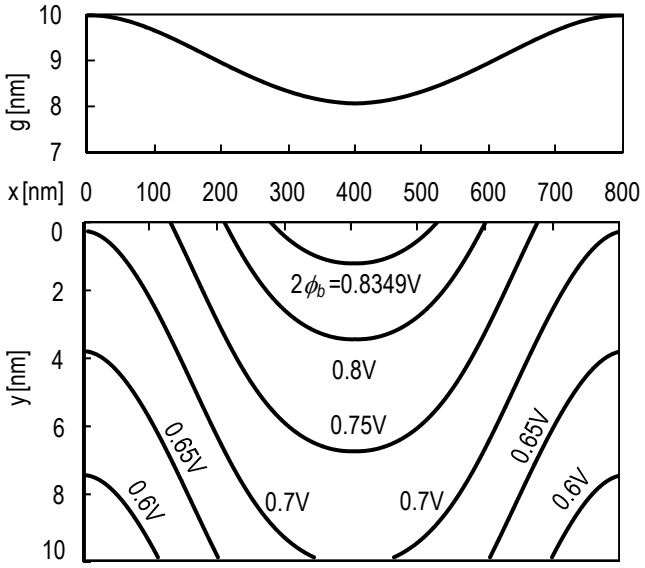


Fig. 3.4 (b)

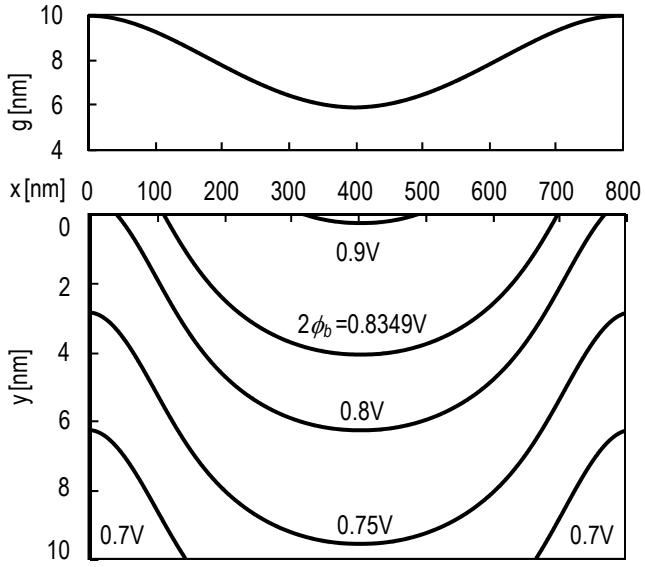


Fig. 3.4 (c)

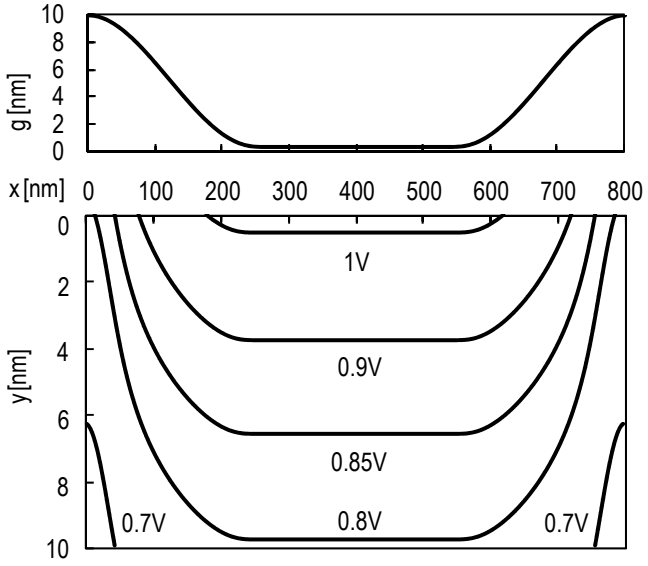


Fig. 3.4 (d)

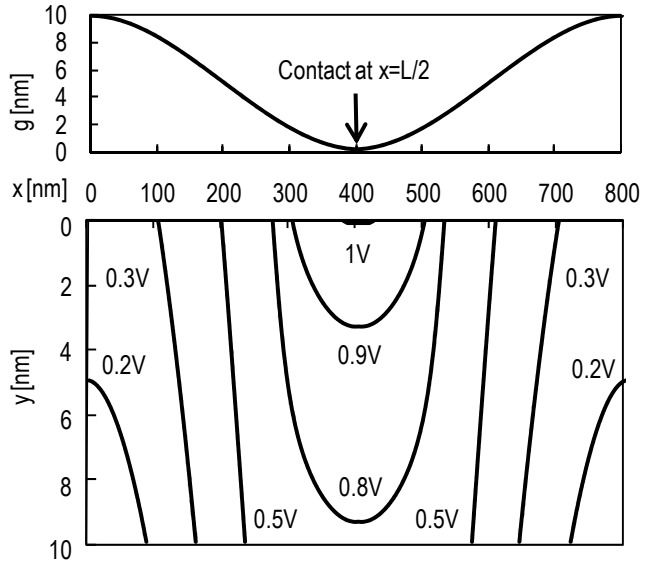


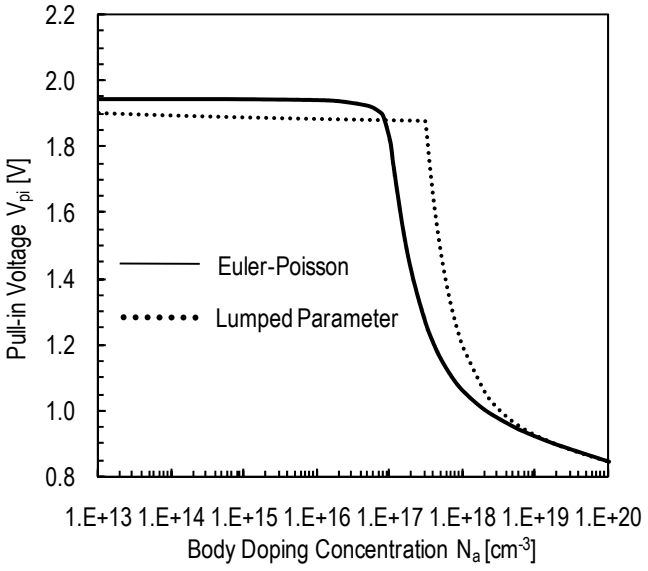
Fig. 3.4 (e)

Fig. 3.4: Simulated equivalent gap thickness, *g(x)*, and semiconductor potential contours $\phi(x,y)$ for different gate voltage biases: a) V_{gs}=0V, b) V_{gs}=1.6V, c) $V_{gs}=V_{pi}^{-}$ =1.8141V, d) $V_{gs}=V_{pi}^{+}$=1.8141V, e) $V_{gs}=V_{rl}$=0.308V.

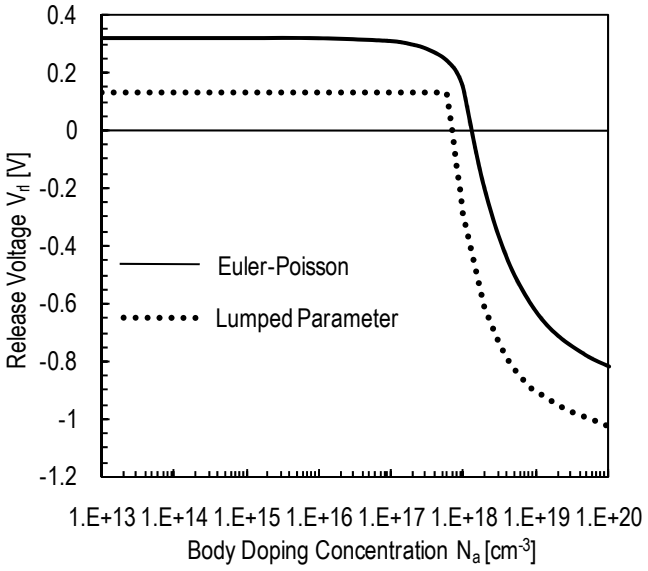
3.3.2 Impact of body doping concentration, N_a

The threshold voltage (V_T) of a conventional MOSFET can be tuned by adjusting the body doping concentration N_a: in order to increase V_T, N_a is increased. In contrast, for a NEMFET, V_{pi} and V_{rl} are independent of N_a below some critical concentration level; above this level, V_{pi} and V_{rl} *decrease* with increasing N_a (Fig. 3.5). To explain this dependence of V_{pi} and V_{rl} on N_a, we plot ϕ_s at $x = L/2$ against N_a, for $V_{gs} = V_{pi}$ and $V_{gs} = V_{rl}$ in Fig. 3.6a and Fig. 3.6b, respectively. The $\phi_s=2\phi_b$ line, which corresponds to the onset of strong inversion in the channel at $x=L/2$, is also shown for reference. The ϕ_s curves for $V_{gs} = V_{pi}$ and $V_{gs} = V_{rl}$ intersect the $2\phi_b$ reference line at the body doping concentrations $N_{FD,pi}$ and $N_{FD,rl}$, respectively. For $N_a>N_{FD,pi}$, the surface channel is "fully depleted" (since $\phi_s<2\phi_b$ for all x) when $V_{gs} = V_{pi}$. If $N_a< N_{FD,pi}$, ϕ_s at $x = L/2$ increases by 60 mV for every 10× increase in N_a; on the other hand, if $N_a>N_{FD,pi}$, ϕ_s at $x = L/2$ decreases very rapidly (>60mV/decade) with increasing N_a. A similar dependence on N_a is seen for ϕ_s at $x = L/2$ with $V_{gs} = V_{rl}$. In the subsequent analysis, we refer to the case where $N_a< N_{FD,pi}$ (or $N_a< N_{FD,rl}$) as "inversion pull-in (or inversion release)" and the case where $N_a>N_{FD,pi}$ (or $N_a> N_{FD,rl}$) as "sub-threshold pull-in (or sub-threshold release)"[1].

[1] Strictly speaking, there also exists the case where pull-in/release occurs when the semiconductor surface is only moderately inverted [3.16]. As was well discussed in [3.16], analytical models for this region of operation are rather complicated; for simplicity, our lumped parameter model does not include the "moderate inversion pull-in/release" case.

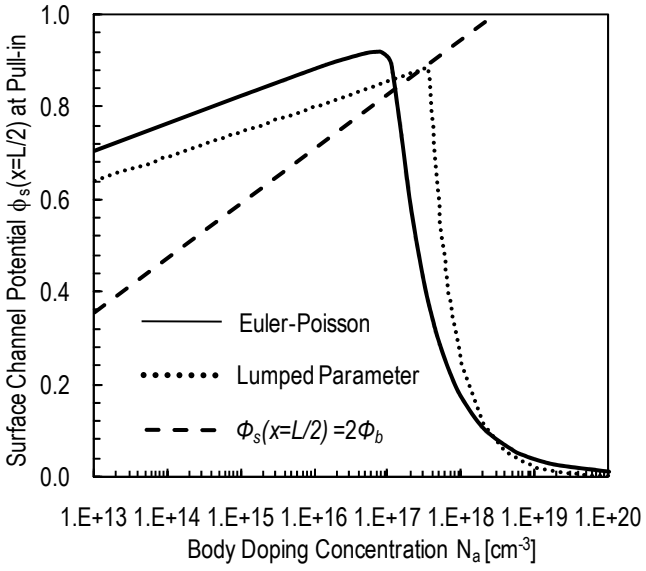


(a)

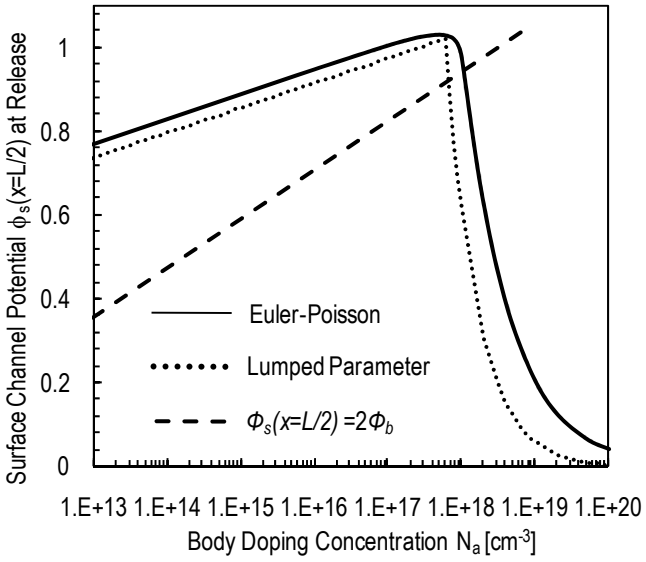


(b)

Fig. 3.5: Dependence of (a) V_{pi} and (b) V_{rl} on the body doping concentration N_a. The results of the lumped-parameter model (dotted line) are also shown for comparison.



(a)



(b)

Fig. 3.6: Dependence of the semiconductor surface potential at $x=L/2$ on N_a, for a gate-to-source bias of (a) V_{pi} and (b) V_{rl}. For reference, $\phi_s(L/2)=2\phi_b$ is also plotted.

a. Inversion pull-in/release

For zero body bias, $V_{gs}=V_{fb}+V_{eff,}+\phi_s$. At $V_{gs}=V_{pi}$, V_{eff} can be estimated using Eqn. (3.4). In strong inversion, the depletion capacitance is screened by the free carriers in the channel; thus, the negative feedback stabilization [3.18] vanishes and therefore

$$V_{eff} = \sqrt{\frac{8k\left(t_{gap}+\frac{t_{ox}}{\kappa_{ox}}\right)^3}{27\varepsilon_o WL}} = V_{picc} \tag{3.12}$$

$$g = \frac{2}{3}\left(t_{gap} + \frac{t_{ox}}{\kappa_{ox}}\right) \tag{3.13}$$

where V_{picc} is the pull-in voltage of a conventional clamped-clamped beam. The total areal charge in the semiconductor at $V_{gs} = V_{pi}$ is given by the equation

$$Q_s = -\frac{\varepsilon_o V_{eff}}{g} = -\frac{3\varepsilon_o V_{picc}}{2\left(t_{gap}+\frac{t_{ox}}{\kappa_{ox}}\right)} \tag{3.14}$$

Then ϕ_s can be found from Eqn. (3.5):

$$\phi_s = \frac{k_b T}{q}\ln\left(\frac{9\varepsilon_o N_a V_{picc}^2}{8\kappa_{si}n_i^2\left(t_{gap}+\frac{t_{ox}}{\kappa_{ox}}\right)^2 k_B T}\right) = \frac{k_b T}{q}\ln\left(\frac{9\varepsilon_o V_{picc}^2}{8\kappa_{si}n_i g_o^2 k_B T}\right) + \frac{k_b T}{q}\ln\left(\frac{N_a}{n_i}\right) \tag{3.15}$$

which implies $\frac{d\phi_s}{dlog(N_a)} \approx 60mV/dec$, consistent with Fig. 3.6a.

By adding V_{fb}, $V_{eff,}$ and ϕ_s together, we obtain the following expression for V_{pi}:

$$V_{pi} = C - \frac{k_b T}{q}\ln\left(\frac{N_a}{n_i}\right) + V_{picc} + \left[\frac{k_b T}{q}\ln\left(\frac{9\varepsilon_o V_{picc}^2}{8\kappa_{si}n_i g_o^2 k_B T}\right) + \frac{k_b T}{q}\ln\left(\frac{N_a}{n_i}\right)\right] \tag{3.16}$$

The expression for V_{rl} can be derived similarly:

$$At\ V_{gs} = V_{rl}: \begin{cases} V_{eff} = \sqrt{\frac{2kt_{gap}(t_{ox}/\kappa_{ox})^2}{\varepsilon_o WL}} \\ Q_s = -\frac{\varepsilon_o V_{eff}}{(t_{ox}/\kappa_{ox})} = -\sqrt{\frac{2kt_{gap}\varepsilon_o}{WL}} \\ \phi_s = \frac{k_b T}{q}\ln\left(\frac{kt_{gap}N_a}{WL\kappa_{si}k_B T n_i^2}\right) \end{cases} \tag{3.17}$$

$$V_{rl} = C - \frac{k_b T}{q}\ln\left(\frac{N_a}{n_i}\right) + \sqrt{\frac{8kt_{gap}(t_{ox}/\kappa_{ox})^2}{27\varepsilon_o WL}} + \left[\frac{k_b T}{q}\ln\left(\frac{kt_{gap}}{WL\kappa_{si}k_B T n_i}\right) + \frac{k_b T}{q}\ln\left(\frac{N_a}{n_i}\right)\right] \tag{3.18}$$

Eqns. (3.16) and (3.18) explain why both V_{pi} and V_{rl} are independent of N_a for inversion pull-in/release: any increase in ϕ_s due to an increase in N_a is compensated by a reduction in the flat-band voltage.

b. Sub-threshold pull-in/release

For sufficiently large N_a, pull-in/release occurs before the semiconductor surface becomes strongly inverted. As shown in Fig. 3.5, both V_{pi} and V_{rl} decrease rapidly with increasing N_a in this case. This is because the depletion capacitance C_{dep} increases with N_a. To understand this qualitatively, we can use the capacitive divider model as shown in Fig. 3.1c:

$$\frac{dV_{eff}}{dV_{gs}} = \frac{C_{dep}}{C_{gap}+C_{dep}} \tag{3.19}$$

Thus, for a given V_{gs}, V_{eff} increases with increasing N_a so that a smaller value of V_{gs} is needed to achieve a certain electrostatic force required for pull-in (or release), and hence V_{pi} (or V_{rl}) decreases with increasing N_a. This is consistent with previously published work [3.3]. Notice also that for intermediate body-doping concentrations, pull-in occurs when the semiconductor surface is moderately inverted; therefore, as depicted in Fig. 3.7, negative feedback stabilization [3.18] reduces the pull-in gap thickness.

For very high body doping concentration, $\phi_s(x)$ approaches zero. Eqn. (3.7) then becomes:

$$EI\frac{d^4g(x)}{dx^4} = -\frac{\varepsilon_o W\left(V-V_{fb}\right)^2}{2g(x)^2} \tag{3.20}$$

and V_{pi}-V_{fb}=V_{ox} converges to $\sqrt{11.7\frac{Eh^3g_o^3}{\varepsilon_o L^4}} = 2V$; the pull-in gap thickness at x=L/2 also converges to 0.603g, as indicated in Figs. 5a and 7. This is exactly the same as the pull-in voltage and the pull-in gap thickness of a clamped-clamped beam with a metallic actuation electrode, which is not surprising (since degenerately doped silicon is a conductive material) and is consistent with the results obtained using the lumped parameter model [3.3].

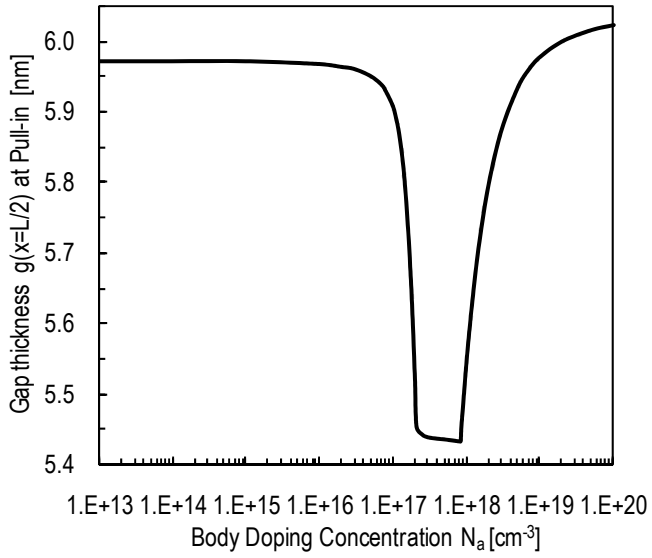


Fig. 3.7: Dependence of the gap thickness at x=L/2 on N_a for a gate-to-source bias equal to V_{pi}.

i. Minimum N_a for sub-threshold pull-in

The values of $N_{FD,pi}$ and $N_{FD,rl}$ are of interest to design the NEMFET for different (digital *vs*. analog) applications. If pull-in occurs at the onset of strong

inversion, ϕ_s at $x = L/2$ for $V_{gs} = V_{pi}$ is $2\phi_b$ and the total areal charge in the semiconductor is

$$Q_s = \sqrt{2\varepsilon_{si} q N_{FD,pi}(2\phi_b)} = \sqrt{4\varepsilon_{si} q N_{FD,pi} \frac{k_b T}{q} \ln\left(\frac{N_{FD,pi}}{n_i}\right)} \tag{3.21}$$

By equating Eqn. (3.12), (3.14) and Eqn. (3.21), we obtain the following equation:

$$\frac{N_{FD,pi}}{n_i} ln\left(\frac{N_{FD,pi}}{n_i}\right) = \frac{k\left(t_{gap}+\frac{t_{ox}}{\kappa_{ox}}\right)}{6 n_i k_b T \kappa_{si} WL} \tag{3.22}$$

Letting $= \frac{k\left(t_{gap}+\frac{t_{ox}}{\kappa_{ox}}\right)}{6 n_i k_b T \kappa_{si} WL}$, the solution to Eqn. (3.22) is

$$N_{FD,pi} = n_i \, \alpha / lambertW(\alpha) \tag{3.23}$$

where *m=lambertW(n)* is the Lambert W function, the solution to the equation $n = me^m$.

Following the same procedure, $N_{FD,rl}$ can similarly be found:

$$N_{FD,rl} = n_i \, \gamma / lambertW(\gamma) \tag{3.24}$$

where $\gamma = \frac{k t_{gap}}{2 n_i k_b T \kappa_{si} WL}$

Although the exact values of $N_{FD,pi}$ and $N_{FD,rl}$ can only be obtained numerically, we can still note that since α and γ are always positive, $N_{FD,pi}$ and $N_{FD,rl}$ are monotonically increasing functions of α and γ, respectively. Thus, the minimum body doping concentration required for sub-threshold pull-in/release operation is larger for a stiffer gate, a larger air-gap, or a smaller actuation area.

3.3.3 Impact of Gate Work Function

In a conventional MOSFET, the gate work function Φ_m can be used to tune the threshold voltage [3.19, 3.20]. Similarly, the gate work function can be used to tune the pull-in and release voltages of a NEMFET, since the built-in voltage between the gate and the semiconductor channel induces an electric field. Since $V_{gs} = V_{fb} + V_{eff} + \phi_s$ and $V_{fb}=\Phi_m-\Phi_s$, both the pull-in voltage and the release voltage shift linearly with Φ_m.

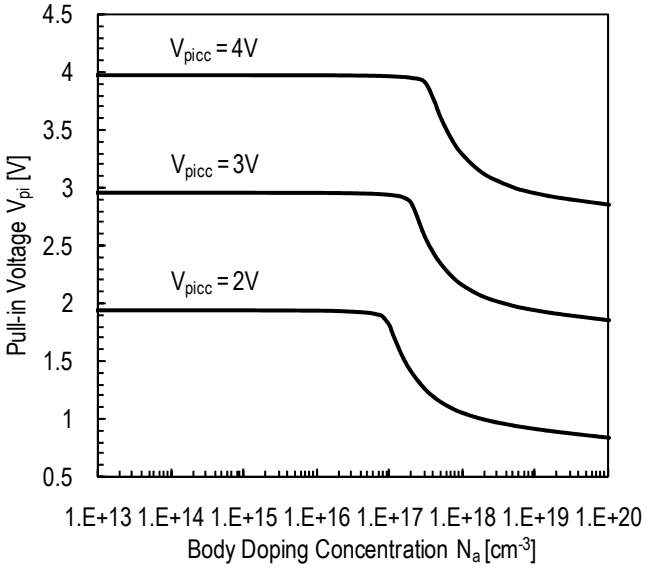
3.3.4 Impact of Gate Stiffness

The stiffness of the mechanical gate depends on the gate dimensions *(W,h,L)* and the gate material Young's modulus (*E*). As was discussed in [3.21], these parameters can be lumped into a single parameter $V_{picc} = \sqrt{11.7\frac{Eh^3g_o^3}{\varepsilon_o L^4}}$, the pull-in voltage of a conventional gap-closing actuator with metallic electrodes [3.22]. Fig. 3.8 plots V_{pi} and V_{rl} for different values[2] of V_{picc}. Increasing V_{picc} increases the gate stiffness, hence both V_{pi} and V_{rl} increase with V_{picc} over the entire range of body doping concentration. Note that $N_{FD,pi}$ and $N_{FD,rl}$ also increase with the gate stiffness, as expected from Eqns. (3.23)-(3.24). This is because a higher body doping concentration is needed to increase the value of V_{gs} at which the surface becomes strongly inverted to be equal to V_{pi} (or V_{rl}).

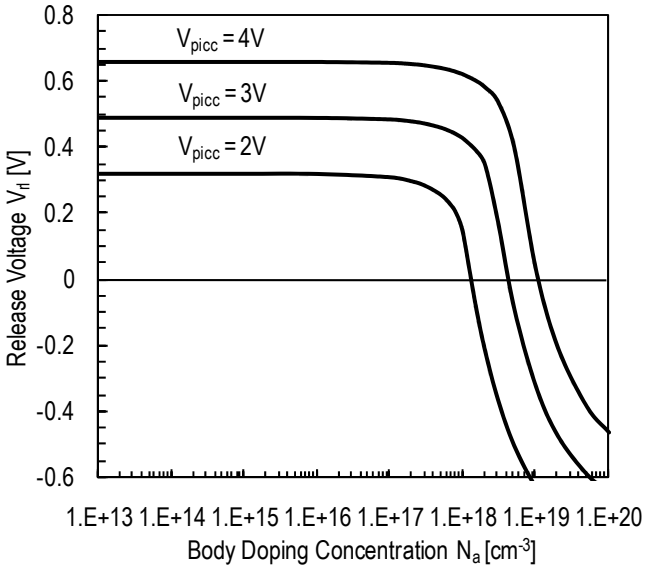
[2]Note that V_{picc} depends on g_o, the as-fabricated equivalent gap thickness. But g_o not only changes V_{picc} , but also the equivalent gate-oxide thickness of the built-in transistor, as indicated by the Poisson equation (Eqn. (3.9)). In this subsection, we assume a constant g_o as we change V_{picc}. The impact of g_o on V_{pi} will be discussed in detail in the following subsection.

3.3.5 Impact of As-Fabricated Air-Gap Thickness, t_{gap}

It is well known that a reduction in the gap thickness can improve the transduction efficiency of an electrostatically actuated device [3.23]. Similarly, a reduction in the gap thickness reduces the pull-in voltage of a NEMFET. Fig. 3.9 plots V_{pi} and V_{rl} for various values of effective as-fabricated actuation gap thickness, g_o. A decrease in g_o reduces V_{pi} in two ways: first, it reduces V_{picc} of the doubly-clamped beam; second, for a given applied gate voltage, it increases the channel surface potential ϕ_s and hence the electric field across the equivalent air gap (Eqn. 3.9). V_{rl} also decreases because the spring restoring force decreases linearly with decreasing t_{gap}.

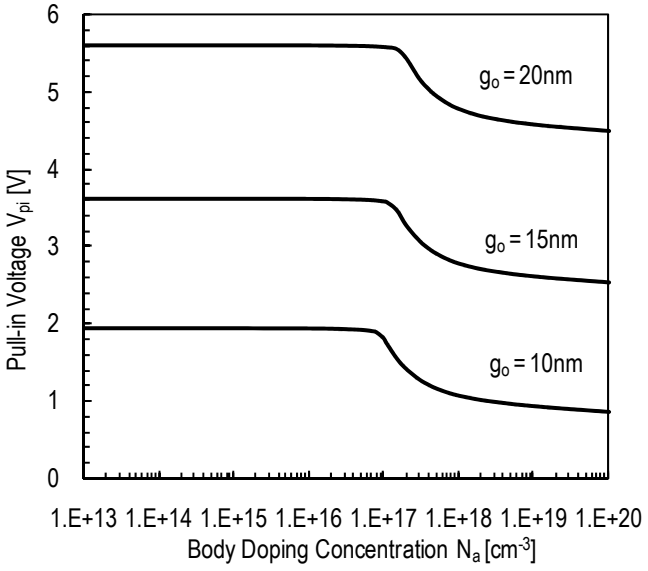


(a)

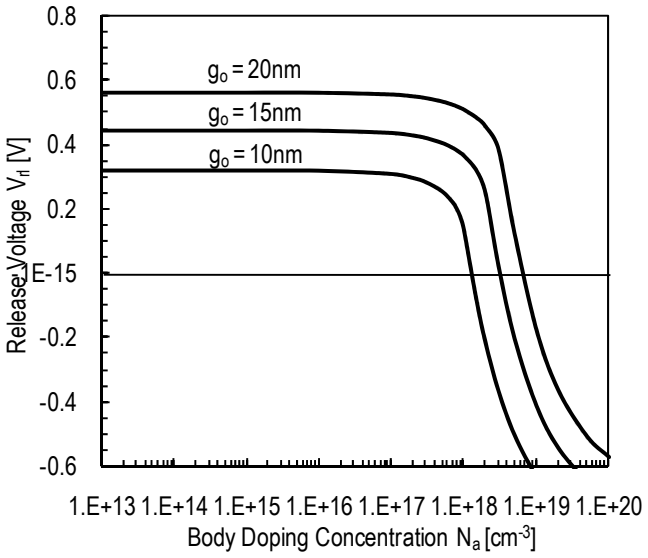


(b)

Fig. 3.8: Impact of beam stiffness (correlated to the pull-in voltage of a conventional doubly clamped beam, V_{picc}) on V_{pi} and V_{rl}. An increase in beam stiffness increases the voltages required for gate pull-in and release.



(a)



(b)

Fig. 3.9: Impact of as-fabricated equivalent gap thickness ($t_{gap} + t_{ox}/\kappa_{ox}$) on V_{pi} and V_{rl}. A reduction in the gap thickness increases the electrostatic force on the gate and hence reduces V_{pi}. It also reduces the spring restoring force and V_{pi}.

3.3.6 Impact of Source-to-Body Bias Voltage, V_{sb}

Thus far in our discussion, we have assumed that the source and body electrodes are biased at the same potential; however, this is not necessarily the case, since transistors may be connected in series. Since the body terminal of an n-channel FET is usually grounded, the source of an n-channel NEMFET can be at a potential that is higher than the body potential. The source-to-body voltage V_{sb} due to a "floating" source can significantly affect the threshold voltage of a conventional transistor. The impact of V_{sb} on V_{pi} and V_{rl} of a NEMFET depends on whether pull-in/release occurs in the sub-threshold or inversion regime of operation.

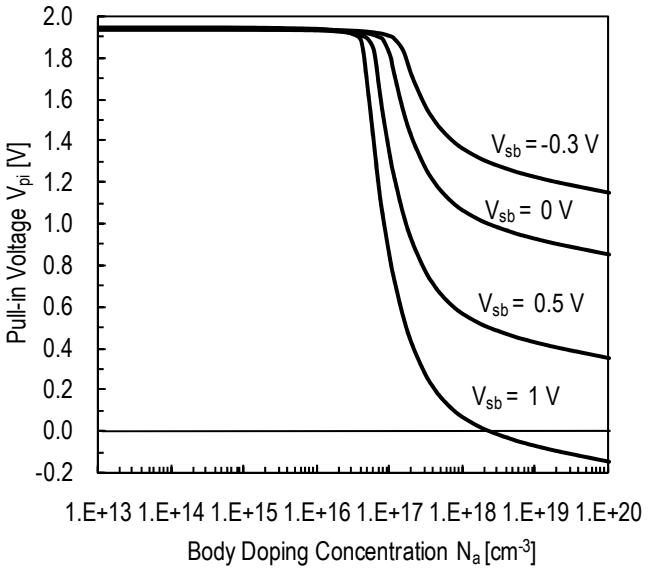
For the inversion pull-in/release case, both V_{pi} and V_{rl} are independent of V_{sb}, as shown in Fig. 3.10. This is because in strong inversion, although V_{sb} causes a redistribution of charge between the inversion layer and depletion region, the total charge in the semiconductor (Q_{inv} + Q_{dep}) remains unchanged [3.16]: $Q_{inv} + Q_{dep} = C_{ox}(V_{gs} - V_{fb} - 2\phi_b)$. Thus, V_{eff} (which determines the actuation force) and therefore V_{pi} and V_{rl} are all independent of V_{sb}.

For the sub-threshold pull-in/release case, however, an increase in V_{sb} increases the electric field across the air gap; thus, the value of V_{gs} required to achieve a given surface potential and V_{eff} decreases:

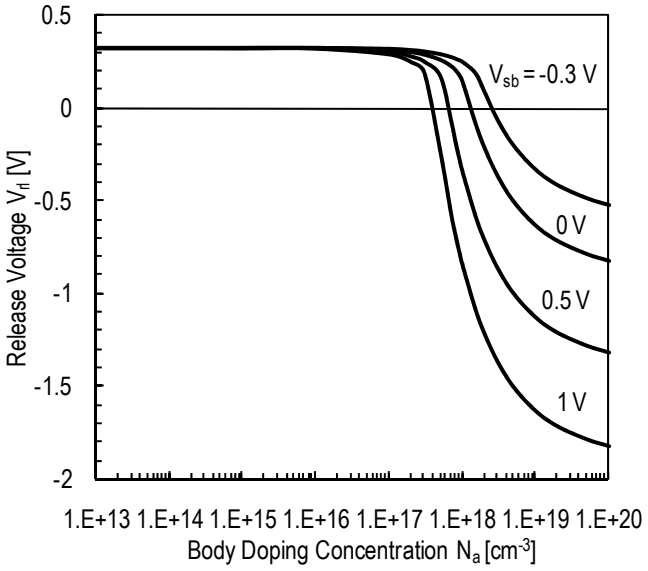
$$V_{gs}=V_{fb}+\phi_s+V_{eff}-V_{sb} \tag{3.25}$$

Therefore, V_{pi} and V_{rl} decrease linearly with V_{sb} :

$$V_{pi}=V_{pi}(V_{sb}=0) - V_{sb} \quad ; V_{rl}=V_{rl}(V_{sb}=0) - V_{sb} \tag{3.26}$$



(a)



(b)

Fig. 3.10: Impact of source-to-body voltage, V_{sb}, on V_{pi} and V_{rl}. Both V_{pi} and V_{rl} show no dependence on V_{sb} if pull-in/release occurs in the inversion region of operation. V_{pi} and V_{rl} decrease with increasing V_{sb} if pull-in/release occurs in the inversion region of operation.

As can be seen from Fig. 3.10, $N_{FD,pi}$ and $N_{FD,rl}$ also depend on $V_{sb.,}$ for the sub-threshold pull-in/release case. At the onset of inversion, $\phi_s=2\phi_b+V_{sb}$. Following the derivation in Eqns. (3.28)-(3.30), $N_{FD,pi}$ can be determined from the following equation:

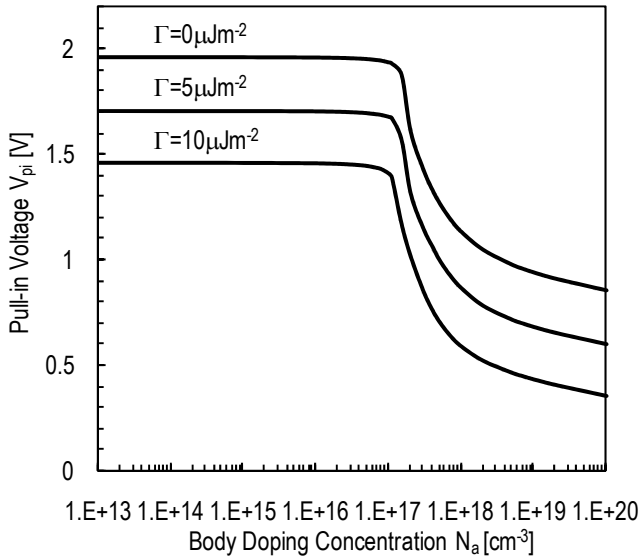
$$N_{FD,pi} = n_i \alpha / lambertW\left(\alpha e^{\frac{qV_{sb}}{2k_bT}}\right) \tag{3.27}$$

Similarly, $N_{FD,rl}$ can be derived: $N_{FD,rl} = n_i\, \gamma / lambertW(\gamma e^{\frac{qV_{sb}}{2k_bT}})$ (3.28)

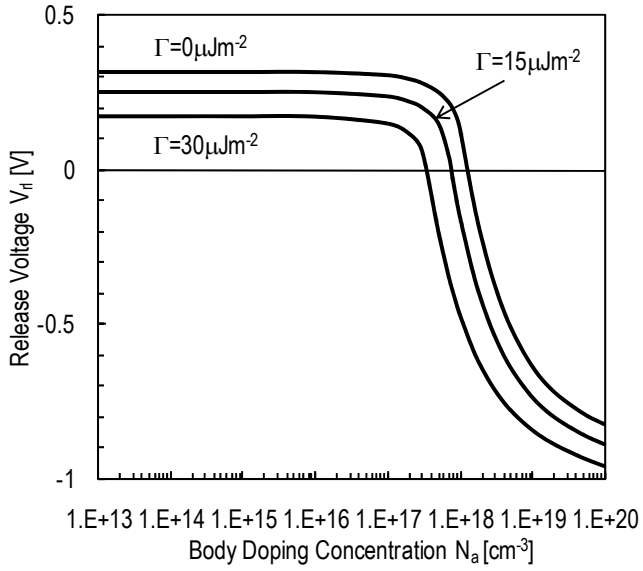
If $V_{sb}> 0$, the gate voltage required for the semiconductor channel surface to reach inversion is increased. Thus, the range of body doping concentration for sub-threshold pull-in/release is increased, *i.e.* $N_{FD,pi}$ and $N_{FD,rl}$ each decrease with increasing V_{sb}.

3.3.7 Impact of Surface Adhesion Force

Thus far in our discussion, surface forces have been ignored; but for the small air-gap thicknesses assumed herein, the Van der Waals force should be considered. Fig. 3.11 plots V_{pi} and V_{rl} for various values of interfacial adhesion energy per unit area. As expected, an increase in the adhesion energy decreases the V_{pi} and V_{rl} values. For the NEMFET device parameters used in this work, the spring restoring force is roughly 96nN, which is significantly greater than the surface adhesion force (~3.2nN for $\Gamma=2\mu J/m^2$ [3.15] with $d_o=0.1nm$). Thus, the impact of surface force on V_{pi} and V_{rl} is negligible.



(a)



(b)

Fig. 3.11: Impact of surface adhesion energy, Γ, on V_{pi} and V_{rl}. An increase in the surface adhesion force decreases both V_{pi} and V_{rl} .

3.4 Unified Model for V_{pi} and V_{rl}

A primitive sub-threshold pull-in model was derived in [3.3]. Here we add the impact of body biasing and expressions for $N_{FD,pi}$ and $N_{FD,rl}$, in order to make the model complete:

3.4.1 Pull-in Voltage V_{pi}

$$V_{pi}=V_{fb}+\phi_s+V_{eff}-V_{sb} \tag{3.29}$$

where

$$\phi_s=\begin{cases}\frac{k_bT}{q}\ln\left(\frac{9\varepsilon_o N_a V_{picc}^2}{8\kappa_{si}n_i^2 g_o^2 k_B T}\right)+V_{sb} & N_a\le N_{FD,pi}\\ \frac{k^2\varepsilon_o}{18q^3\kappa_{si}^3N_a^3(WL)^2}\left(1+\sqrt{1+\frac{6WL(\kappa_{si}qN_a)^2g_o}{k\varepsilon_o}}\right)^2 & N_a>N_{FD,pi}\end{cases} \tag{3.30}$$

$$V_{eff}=\begin{cases}V_{picc} & N_a\le N_{FD,pi}\\ \frac{2}{3}\left(\frac{g_o\sqrt{2q\varepsilon_{si}N_a\phi_s}}{\varepsilon_o}-\phi_s\right) & N_a>N_{FD,pi}\end{cases} \tag{3.31}$$

$$N_{FD,pi}=n_i\alpha/lambertW\left(\alpha e^{\frac{qV_{sb}}{2k_bT}}\right),\ \alpha=\frac{k\left(t_{gap}+\frac{t_{ox}}{\kappa_{ox}}\right)}{6n_ik_bT\kappa_{si}WL},\ V_{pi,cc}=\sqrt{\frac{8kg_o^3}{27\varepsilon_oWL}}$$

3.4.2 Release Voltage V_{rl}

$$V_{rl}=V_{fb}+\phi_s+V_{eff}-V_{sb} \tag{3.32}$$

where

$$\phi_s=\begin{cases}\frac{k_bT}{q}\ln\left(\frac{kt_{gap}N_a}{WL\kappa_{si}k_BTn_i^2}\right)+V_{sb} & N_a\le N_{FD,rl}\\ \frac{kt_{gap}}{WL\kappa_{si}qN_a} & N_a>N_{FD,rl}\end{cases} \tag{3.33}$$

$$V_{eff}=\sqrt{\frac{2kt_{gap}(t_{ox}/\kappa_{ox})^2}{\varepsilon_oWL}} \tag{3.34}$$

$$N_{FD,rl} = n_i\gamma/lambertW\left(\gamma e^{\frac{qV_{sb}}{2k_bT}}\right), \gamma = \frac{kt_{gap}}{2n_i k_b T \kappa_{si} WL}$$

To ensure that the effect of surfaces forces is insignificant, the spring restoring force must be much greater than the surface adhesion force. Eqns. (3.29)-(3.34) can be used to design the NEMFET in order to achieve the desired values of V_{pi} and V_{rl} for a particular application.

3.5 NEMFET Scalability

The introduction of a thin air-gap in a NEMFET drastically improves the effective subthreshold slope. However, such an air-gap also drastically decreases the gate-to-channel capacitive coupling in the off-state, making the device more susceptible to short channel effects [3.16] than the conventional MOSFET. Results from quasi 2-D analysis [3.16] indicate that the minimum channel length of a MOSFET for a given short channel control is correlated to the characteristic length (λ), which is given by the following equation:

$$\lambda = \sqrt{\frac{\varepsilon_s}{\varepsilon_{ox}} t_{eot} X_j} \qquad (3.35)$$

where t_{eot} is the equivalent gate oxide thickness, X_j is the source-drain junction depth. Assuming the NEMFET and the MOSFET have the same source-drain junction depth, the NEMFET channel length needs to be at least $\lambda_{NEMFET}/\lambda_{MOSFET}$ $=\sqrt{10nm \times 3.9/1nm} \sim 6$ times longer than a MOSFET with a 1nm thick SiO_2 gate oxide for the same short channel control. This would drastically increase the circuit layout area and hinder compact logic implementation.

The NEMFET limited scalability remains an issue for logic applications; but the built-in transistor gain makes the device very attractive for sensor and resonator applications where layout area is not as critical. To fully harness the benefits of the NEMFET, the designs of the mechanical gate and the intrinsic transistor must be co-optimized. As an example, consider a NEMFET resonator where the resonant motion of the gate is sensed by the modulated drain-to-source current. The resonant frequency can be adjusted by changing the gate dimensions and the motional current can be maximized by optimizing the transistor transconductance. To achieve this goal, though, techniques are required that are beyond the scope of this thesis.

3.6 Summary

The Euler-Bernoulli equation is solved consistently with the Poisson equation to model the behavior of a nano-electro-mechanical field effect transistor. Using this Euler-Poisson model, the shape of the deflected gate electrode and channel potential profile across the width of the channel for various gate voltages can be studied. The dependence of the pull-in voltage (V_{pi}) and gate release voltage (V_{rl}) on the body doping, gate work function engineering, gate stiffness, as-fabricated gap thickness, the body bias and surface force are examined. The Euler-Poisson model well matches the results of the unified model for V_{pi} and V_{rl}, which is provided to aid NEMFET design for digital and analog applications.

To circumvent the limited scalability issue of the NEMFET, a better transistor design is needed to fully harness the benefits of the abrupt pull-in effects

without sacrificing short channel control; with this goal in mind, micro-relays for logic applications are discussed in Chapter 4.

3.7 References

[3.1] A. M. Ionescu, V. Pott, R. Fritschi, K. Banerjee, M. J. Declerq, P. Renaud, C. Hibert, P. Fluckiger, and G. A. Racine, "Modeling and design of a low-voltage SOI suspended-gate MOSFET (SG-MOSFET) with a metal over-gate architecture," in *Proc. ISQED*, 2002, pp. 496–501.

[3.2] H. Kam, D. T. Lee, R. T. Howe, and T.-J. King, "A new nano-electromechanical field effect transistor (NEMFET) design for low-power electronics," in *IEDM Tech. Dig.*, 2005, pp. 463–466.

[3.3] K. Akarvardar, C. Eggimann, D. Tsamados, Y. S. Chauhan, G. C. Wan, A. M. Ionescu, R. T. Howe, and H.-S. P. Wong, "Analytical modeling of the suspended-gate FET and design insights for low-power logic," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 48–59, Jan. 2008.

[3.4] N. Abele, R. Fritschi, K. Boucart, F. Casset, P. Ancey, and A. M. Ionescu, "Suspended-gate MOSFET: Bringing new MEMS functionality into solid-state MOS transistor," in *IEDM Tech. Dig.*, 2005, pp. 479–481.

[3.5] N. Abele, A. Villaret, A. Gangadharaiah, C. Gabioud, P. Ancey, and A. M. Ionescu, "1T MEMS memory based on suspended gate MOSFET,"in *IEDM Tech. Dig.*, 2006, pp. 509–512.

[3.6] Y.S. Chauhan, D. Tsamados, N. Abelé , C. Eggimann , M. Declercq, A.M. Ionescu, "Compact modeling of suspended gate FET," in *IEEE int. Conf. on VLSI design*, 2008, pp. 119–24.

[3.7] D. Tsamados, Y. S. Chauhan, C. Eggimann, K. Akarvardar, H.-S. Philip Wong and A.M. Ionescu, "Finite element analysis and analytical simulations of Suspended Gate-FET for ultra-low power inverters," *Solid State Electronics*, vol. 52, issue 9, pp1374-1381, Sep. 2008.

[3.8] H. C. Nathanson, W. E. Newell, R. A. Wickstrom, and J. R. Davis, "The resonant gate transistor," *IEEE Trans. Electron Devices*, vol. ED-14, no. 3, pp. 117–133, Mar. 1967.

[3.9] N. Abele, V. Pott, K. Boucart, F. Casset, K. Segueni, P. Ancey, and A. M. Ionescu, "Comparison of RSG-MOSFET and capacitive MEMS resonator detection," *Electron. Lett.*, vol. 41, no. 5, pp. 242–244, Mar. 2005.

[3.10] D. Grogg, D. Tsamados, N-D. Badila, AM. Ionescu, "Integration of MOSFET transistors in MEMS resonators for improved output detection," in *18th Int. Solid-State Sens. Actuators Conf. Tech. Dig.*, Lyon, France, June 10-14, 2007, pp. 1709–12.

[3.11] E. S. Hung and S. D. Senturia, "Generating efficient dynamical models for microelectromechanical systems from a few finite-element simulations runs," *IEEE/ASME J. Microelectromech. Syst.*, vol. 8, pp. 280-289, 1999.

[3.12] J. A. Pelesko and D. H. Bernstein, *Modeling MEMS and NEMS* Boca Raton, FL: Chapman & Hall/CRC, 2003

[3.13] P. M. Zavracky, S. Majumder, and N. E. McGruer, "Micromechanical switches fabricated using nickel surface micromachining," *J. Microelectromech. Syst.*, vol. 6, pp. 3-9, 1997.

[3.14] K. Wang, A.-C. Wong, and C. T.-C. Nguyen, "VHF free-free beam high-*Q* micromechanical resonators," *IEEE/ASME J. Microelectromech. Syst.*, vol. 9, no. 3, pp. 347-360, Sept. 2000.

[3.15] J. A. Knapp and M. P. de Boer, "Mechanics of microcantilever beams subject to combined electrostatic and adhesive forces," *J. Microelectromech. Syst.*, vol. 11, no. 6, pp. 754–764, Dec. 2002

[3.16] Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed. New York: McGraw-Hill, 1999.

[3.17] S. P. Timoshenko and J. M. Gere, Mechanics of Materials. Pacific Grove: Brooks/Cole, 2001.

[3.18] J. I. Seeger and S. B. Crary, "Stabilization of electrostatically actuated mechanical devices," in *Proc. Int. Conf. Solid-State Sens. Actuators*, 1997, pp. 1133–1136.

[3.19] Q. Lu, Y.-C. Yeo, P. Ranade, H. Takeuchi, T.-J. King, C. Hu, S. C. Song, H. F. Luan, D.-L.Kwong, "Dual-metal gate technology for deep-submicron CMOS transistors," *Symp. VLSI Tech. Digest of Technical Papers*, 2000 pp. 72-73.

[3.20] T. J. King, J. R. Pfiester, J. D. Shott, J. P. McVittie, and K. C. Saraswat, "Polycrystalline-$Si_{1-x}Ge_x$-gate CMOS technology," in *IEDM Tech. Dig.,* 1990, pp. 253–256.

[3.21] S. Gorthi, A. Mohanty and A. Chatterjee, "Cantilever beam electrostatic MEMS actuators beyond pull-in," *Micromechanics and Microengineering*, 2006, 16, pp. 1800-1810.

[3.22] P. M. Osterberg and S. D. Senturia, "M-TEST: A test chip for MEMS material property measurement using electrostatically actuated test structures," *IEEE/ASME J. Microelectromech. Syst.,* vol. 6, no. 2, pp.107–118, 1997.

[3.23] Y. Xie, S.-S. Li, Y.-W. Lin, Z. Ren, and C. T.-C. Nguyen, "UHF micromechanical extensional wine-glass-mode ring resonators," in *Proc. IEEE Int. Electron Devices Meeting* Washington, DC, 2003, pp. 953-956.

Chapter 4

Design and Reliability of Micro-Relays for Logic Applications

4.1 Introduction

Chapter 3 presented the operation and design of the NEMFET, which harnesses the pull-in effect to achieve a perfectly abrupt off-to-on switching transition. The presence of the air-gap in the off-state severely limits NEMFET scalability, however. In face of this limitation, micro-electro-mechanical relays ("micro-relays") [4.1-4.5] appear to be an attractive alternative for zero-standby power logic applications. The attractiveness of micro-relays stems from the fact that a mechanical switch offers nearly ideal switching characteristics: zero off state drain-to-source and gate leakage currents, and perfectly abrupt off-to-on switching transition. Since there is no trade-off between off-state leakage current and on-state drive current, the relay threshold voltage and therefore V_{dd} can in principle be reduced much more aggressively than for a MOSFET in order to improve the energy efficiency.

In terms of device structure and operation, a micro-relay for digital logic (dubbed "logic relays") applications is very similar to that for used for radio-

frequency signal DC switching (dubbed "RF relays") applications [4.6-4.9]. However, the relay contact resistance requirements for these two applications are drastically different. For RF relays in which achieving ultra-low on-state resistance (R_{ON} < 1Ω) is the primary target, the relay design (*e.g.* device dimensions, contacting electrode material, contact force, *etc.*) is optimized to achieve the required metal-to-metal contact conductance. In contrast, for logic relays, R_{ON} can be as high as 10-100kΩ (for load capacitance of 10-100fF) because the switching delay of a relay-based circuit is dominated by the mechanical pull-in time (t_{PI}, typically 10- 100ns) rather than the electrical RC delay (t_{RC}) [4.2]. Since extremely high endurance, fast switching speed, high energy efficiency and high layout density are necessities in this application while relatively high R_{ON} can be tolerated, the design space for RF and logic relays are drastically different.

To date, no systematic design, optimization, and scaling methodology for logic relays has been proposed. To remedy this issue, this chapter begins with a general description of relay structure and operation in Section 4.2, followed by contact design considerations and a tungsten-based relay technology discussed in Section 4.3. Section 4.4 then presents the calibrated relay delay and energy models which are used to develop a sensitivity-based energy-delay optimization and scaling methodology in Chapter 5.

4.2 Relay Structure and Operation

In terms of device structure and operation, a logic relay is very similar to larger micrometer-scale mechanical switches that have been developed for radio

frequency electronics. Fig. 4.1 shows the schematic 3D view of the electrostatically-actuated relay structure and indicates several important relay features and design parameters. As shown, the relay is comprised of a movable actuation plate (the "source") supported at the four corners by four support beams, each of which is anchored to the substrate. Folded-flexure beams are purposely used to ensure the relay design is robust against residual thermal stress and vertical strain gradient effects. The position of this actuation plate depends on the electric field across the actuation gap (thickness *g*) between the gate and the source.

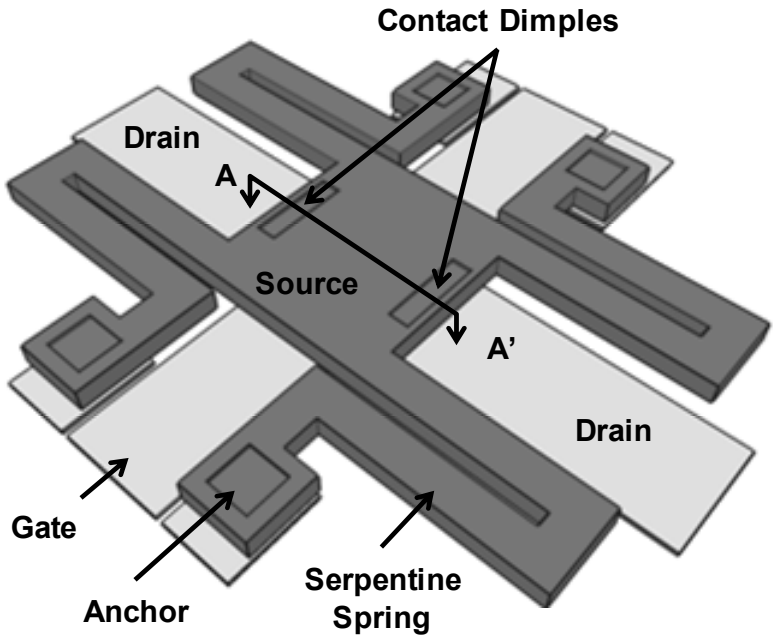


Fig. 4.1 A schematic 3D view of the electrostatically-actuated relay structure.

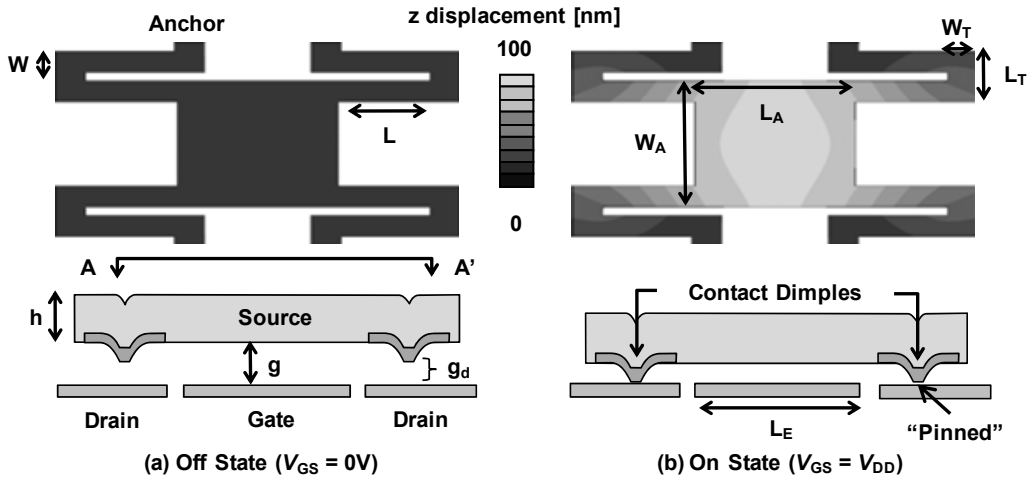


Fig. 4.2 ANSYS simulated displacement contours and the schematic cross-sectional view of the source electrode in the (a) Off state and (b) On state.

In the off state (Fig. 4.2a), an air gap separates the source from the metallic drain electrode so that no current can flow. In the on state (Fig. 4.2b) where the gate-to-source voltage is greater than the pull-in voltage (V_{pi}), the actuation plate comes down and rests upon the metal-coated contact dimples; the source is in contact with drain electrode, providing a conductive path for current to flow. Since the relay switches on abruptly as $|V_{gb}|$ is increased above V_{pi}, the I_d-V_g characteristic of the relay exhibits an extremely steep (nearly infinite) subthreshold slope. Note that since electrostatic force is ambipolar, the equivalent of an n-channel MOSFET (which turns on when a positive bias voltage is applied to the gate) or a p-channel MOSFET (which turns on when a negative bias voltage is applied to the gate) can be achieved with the same switch design by appropriately biasing the source electrode. Thus, electro-mechanical switches can mimic CMOSFETs and can be used to implement digital logic circuitry accordingly.

Although three terminal relays are discussed in this section, many types of micro-relays have been described in the literature, with variations in the mechanical spring design, the location of the air-gap actuator, number of electrode terminals [4.3] and the orientation of beam deflection. While this paper, unless specified, mainly focuses on the three-terminal relay design, the analysis and design techniques presented here can easily be extended and applied to any specific relay design.

4.3 Reliable Micro-Relay Technology

Despite the relay ideal switching behavior alluded to previously, the principal challenge facing micro-relay designers is achieving good reliability while maintaining sufficiently low contact resistance. The contact resistance is determined by material properties and is limited by asperities on the contacting metallic surfaces, and can be computed by the following equation [4.10]:

$$R_{ON} = \frac{4\rho\lambda_p}{3\pi a^2} + \left(\frac{1+0.83\lambda_p/a}{1+1.33\lambda_p/a}\right)\frac{\rho}{2a} \tag{4.1}$$

where a is the radius of the contact asperities, ρ and λ_p are the resistivity and electron mean free path of the contact material, respectively.

The area of the contact asperities A_r is a function of the material hardness H, the deformation coefficient ξ at the contact and also the loading force F_c, which is approximately the electrostatic force:

$$A_r = \pi a^2 = \frac{F_c}{\xi H} \tag{4.2}$$

where $1>\xi>0$ and ξ is inversely proportional to the loading force.

Because the on-state conductance of a micro-relay is limited by asperities on the contacting metallic surfaces, for RF relays applications, one can infer from Eqns. (4.1) and (4.2) that soft contacting electrode materials (esp. gold) and a large applied load to plastically deform [4.11-4.14] the contacting surfaces (*i.e.* such that the metallic material liquefies) are preferred in order to minimize asperities and achieve the required contact resistance (<1Ω). However, this makes high endurance very difficult to achieve with such designs [4.13]. In contrast, for logic relays, R_{ON} can be as high as 10-100kΩ (for load capacitance of 10-100fF) because the switching delay of a relay-based circuit is dominated by the mechanical pull-in time (t_{PI}, typically 10-100ns) rather than the electrical RC delay (t_{RC}) [4.2]. Since extremely high endurance is a necessity in this application while relatively high R_{ON} can be tolerated, hard contacting electrode materials [4.15-4.17] (*e.g.* tungsten) and operation with low contact force [4.15-4.17] are preferred. In addition, a surface treatment can be applied to reduce adhesion [4.16, 4.18] as long as R_{ON} is not increased beyond ~10kΩ. With these considerations in mind, we have developed a reliable relay technology suitable for digital logic applications using titanium dioxide (TiO_2) coated tungsten (W) electrodes. The TiO_2 coating limits the current flow at the contact asperities, and the relay contact can therefore better avoid issues with contact welding. TiO_2 is also a high electron affinity dielectric ($q\chi_{TiO2}$=4.2eV), and thus presents only a moderate potential barrier to electron conduction from W, *i.e.* it degrades on-state conductance least among common dielectric materials.

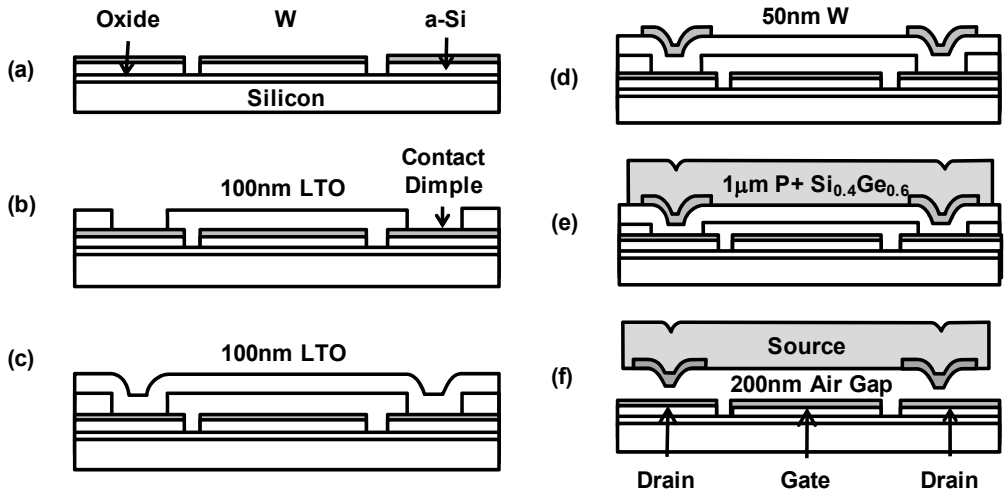


Fig. 4.3 4-mask process flow used to fabricate prototype micro-relays (A-A' cross-section).

Parameter	Value	Parameter	Value
Young Modulus, E	145GPa	Electrode Length, L_E	15μm
Shear Modulus, G	57GPa	Truss Width, W_T	5μm
Density	4126kg·m^{-3}	Truss Length, L_T	12μm
Beam Width, W	5μm	Beam Thickness, h	1μm
Beam Length, L	{10,…,50}×μm	Fabricated Gap Thickness, g	200nm
Actuation Plate Width, W_A	30μm	Dimple Gap thickness, g_d	100nm
Actuation Plate Length, L_A	27μm	Dimple Area, A_D	2×{4,10,15,25}×μm^2

Table 4.1. Relay device parameters

Relays with parameters shown in Table I were fabricated on oxidized Si wafer substrates as follows. Amorphous silicon (which promotes adhesion of W to SiO_2) and W layers, each 50 nm thick, were sequentially deposited by sputtering and then patterned to form the drain and gate electrodes (Fig. 4.3a). 100nm-thick LTO was then deposited at 400^{o}C as the first sacrificial layer. Contact dimple regions

were then formed by optical lithography and dry etching (Fig. 4.3b). After the deposition of a 2nd 100nm-thick LTO sacrificial layer (Fig. 4.3c), a 50nm-thick W layer was sputtered and patterned to form metallic contacting electrodes (Fig. 4.3d). Next, a 1μm-thick structural layer of *in-situ* boron-doped polycrystalline $Si_{0.4}Ge_{0.6}$ was deposited at 410°C [4.19]. The $Si_{0.4}Ge_{0.6}$ layer was then patterned (Fig. 4.3e) and the structures were released (Fig. 4.3f) with a timed isotropic oxide etch using vapor 49% hydrofluoric acid at 27°C. Immediately afterwards, the entire relay structure was coated with TiO_2 at 275°C using 12 cycles of atomic layer deposition (ALD). One ALD cycle consists of one pulse of titanium tetrachloride ($TiCl_4$) followed by Ti oxidation, and deposits ~0.25Å of TiO_2. Note that since the maximum process temperature is 410°C, this relay technology is suitable for fabrication over CMOS circuitry [4.20] or on glass substrates.

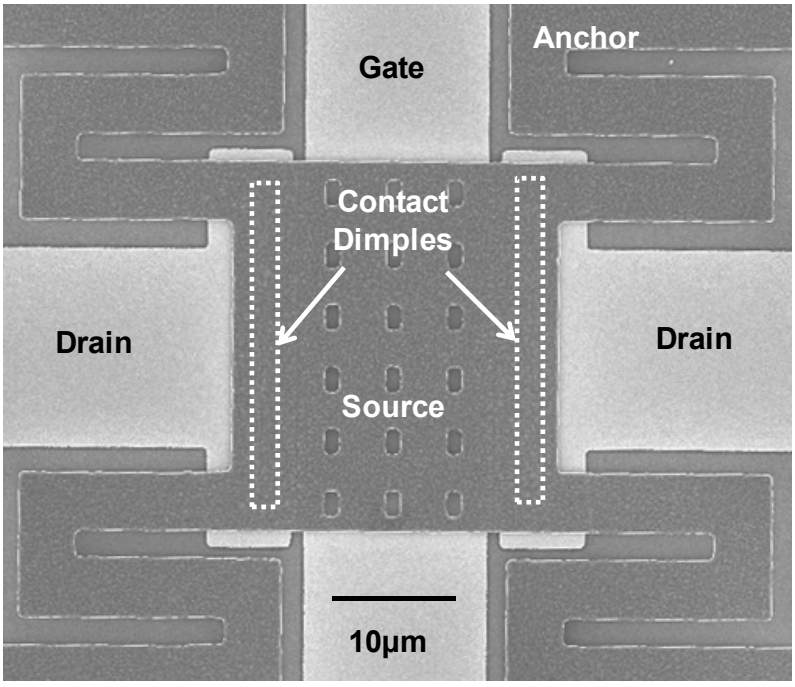


Fig. 4.4 Plan view scanning electron micrograph of a relay before the release step.

4.4 Results and Discussion

Several logic relays with beam length L in the range from 10 to 50μm were fabricated using the process flow detailed in Section 4.3; Fig. 4.4 presents a plan-view scanning electron micrograph (SEM) of a prototype L =20μm relay. All measured relays (>200) were found to be functional. Before actual measurement, each fresh device is first electrically "burned in" by switching it on/off with a relatively high drain-to-source voltage (V_{DS}=1~2.5V) several times [4.21]. This will break the native tungsten oxide and other contaminants on the contacting surfaces to obtain a clean and stable contact resistance. Measured I_D-V_{GS} and I_D-V_{DS} characteristics of a L=40μm relay are plotted in Fig. 4.5. As expected, the relays have zero off state drain-to-source and gate leakage currents. Abrupt switching is seen at V_{GS}=5.35V and R_{ON}=8.1kΩ; the measured R_{ON} matches well with the contact resistance model, which predicts R_{ON} ~10kΩ for F_c=0.89μN at V_{GS}=5.35V, using TiO_2 resistivity of 0.26Ω-cm, tungsten hardness of 1.1GPa [4.22], λ_p=33nm and ξ=0.3 [4.22]. As depicted in Fig. 4.6, the relay can endure 1.25 billion on/off "hot" switching cycles with V_{DS}=1V in N_2 ambient without stiction or welding-induced failure, making tungsten an attractive candidate for reliable relay contact materials.

With a pathway to enable reliable micro-relay for digital logic applications experimentally demonstrated, the following section aims to develop and calibrate relevant models and analytical formulations for relay performance (e.g pull-in/release voltages, switching speed and energy) in order to facilitate design optimization for relay-based circuits.

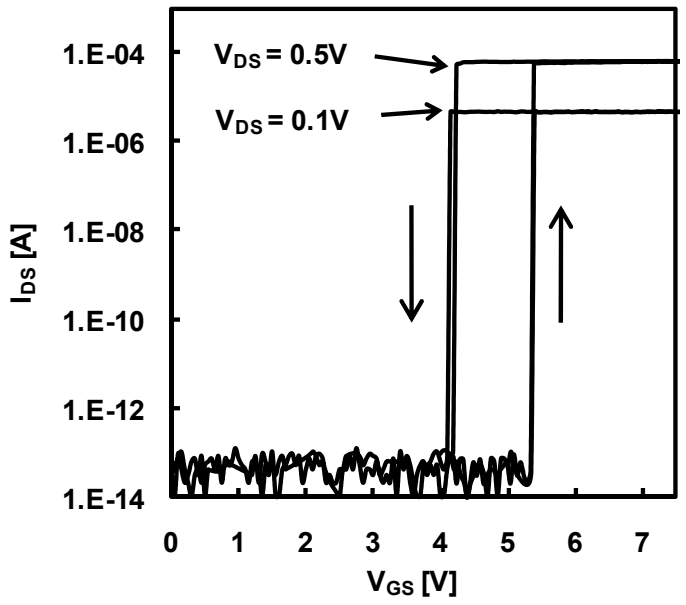
4.4.1 Static Performance Analysis

a. Pull-in and Release Voltages

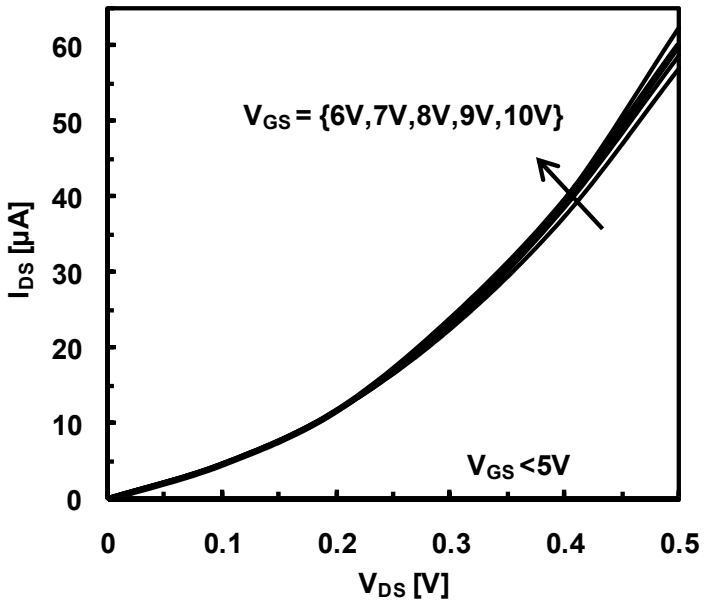
Since the relay is operated in the pull-in mode ($g_d > g/3$), hysteretic switching behavior is observed as indicated by Fig. 4.5. Analytically, the pull-in voltage V_{pi} and release voltage V_{rl} can be calculated as:

$$V_{pi} = \sqrt{\frac{8k_{eff}g^3}{27\varepsilon_o A}},\ V_{rl} = \sqrt{\frac{2(k_{eff}g_d - F_A)(g-g_d)^2}{\varepsilon_o A}} \tag{4.3}$$

where k_{eff} is the effective spring constant of the folded-beam, A is the actuation area : $A = L_E \times W_A$ and F_A is the surface adhesion force. Note that in Eqn. 4.3, non-ideal effects such as fringing capacitance and actuation area reduction due to the release holes are assumed to be negligible.



(a)



(b)

Fig. 4.5(a) Measured I_D-V_{GS} and (b) I_D-V_{DS} characteristics for a relay with L =40 mm. The two drains are externally connected .

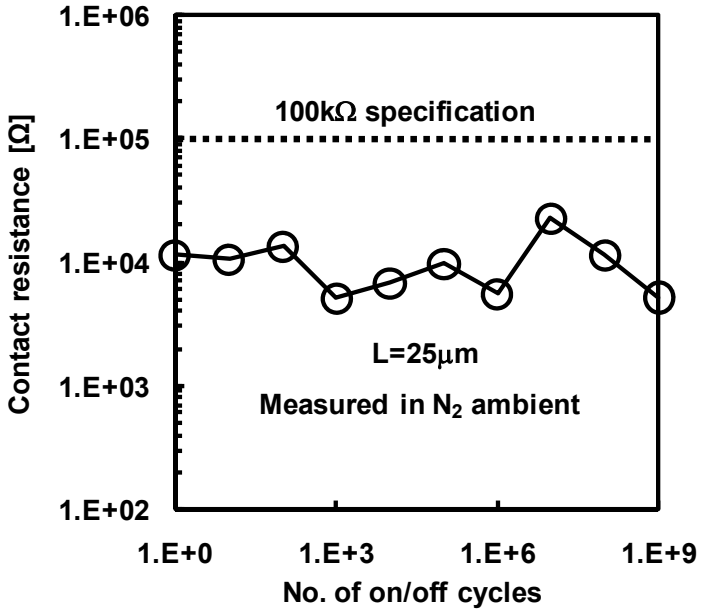


Fig. 4.6 Measured contact resistance *vs.* number of on/off hot switching cycles (at V_{DS}=1V to mimic scaled-relay operation). Due to the formation of tungsten native oxide, contact resistance is unstable.

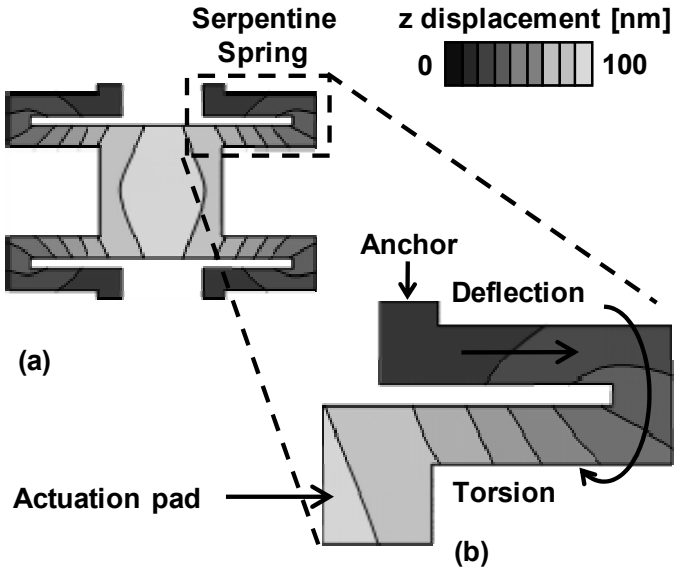


Fig. 4.7 (a) ANSYS-simulated displacement contours of the source electrode in the on-state. (b) the spring exhibits both bending and torsional motions when the actuation pad moves.

As indicated by the ANSYS-simulated displacement contours of the actuation plate shown in Fig. 4.7, the spring exhibits both bending and rotational motions when the plate moves. The exact k_{eff} model that accounts for shear displacements and rotary inertias is rather complicated [4.23]; by sacrificing some degree of accuracy, k_{eff} can be rendered into a more intuitive form, which consists of flexural ($\propto 1/L^3$) and torsional ($\propto 1/L$) [4.24] terms:

$$\frac{1}{k_{eff}} \cong \left(\gamma_f \frac{EWh^3}{L^3}\right)^{-1} + \left(\gamma_t \frac{GWh^3}{L}\right)^{-1} \tag{4.4}$$

where γ_f and γ_t are the flexural and torsional constants. By using ANSYS, γ_f and γ_t are found to be 3.66 and $1.341\times10^{10}\text{m}^{-2}$ respectively. As indicated in Fig. 4.8, the analytical model is within 10% of ANSYS simulation; Eqn. 4.3 predicts V_{pi} values within 10% of the measured data, as shown in Fig. 4.9

To ensure that the relay can be turned off, i.e. $V_{rl}>0$, the spring restoring force must be sufficient to overcome surface adhesion force, F_A. F_A is dominated by the metal-to-metal contacts and it can be estimated by expressing V_{rl} as a function of V_{pi}:

$$V_{rl}^2 = \frac{27}{4}\frac{g_d}{g}\left(1-\frac{g_d}{g}\right)^2 V_{pi}^2 - \frac{2(g-g_d)^2}{\varepsilon_o A}F_A \tag{4.5}$$

The average F_A is extracted to be 0.45μN for a dimple area (A_D) of $2\times10\mu\text{m}^2$, with individual relays' extracted F_A varying by ±1μN around this value, as indicated in Fig. 4.10. The extracted F_A value matches well with results obtained by atomic force microscopy [4.25]. Using the extract F_A value, Eqn. 4.3 predicts V_{rl} values within 10% of the measured data, as shown in Fig. 4.11.

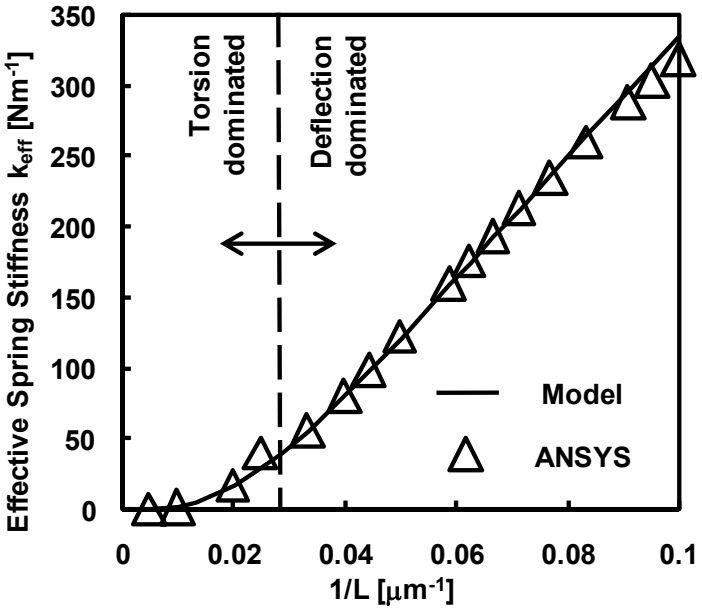


Fig. 4.8 Modeled k_{eff} values match ANSYS simulation. As L decreases, torsional motion dominates.

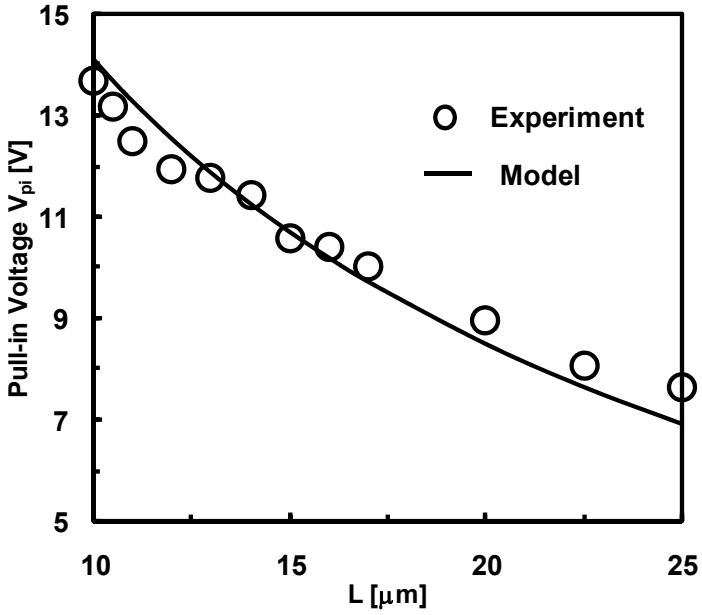


Fig. 4.9 Measured relay V_{pi} *vs. L*. V_{pi} decreases as L increases, as expected. The measured data is within 10% of the analytical model

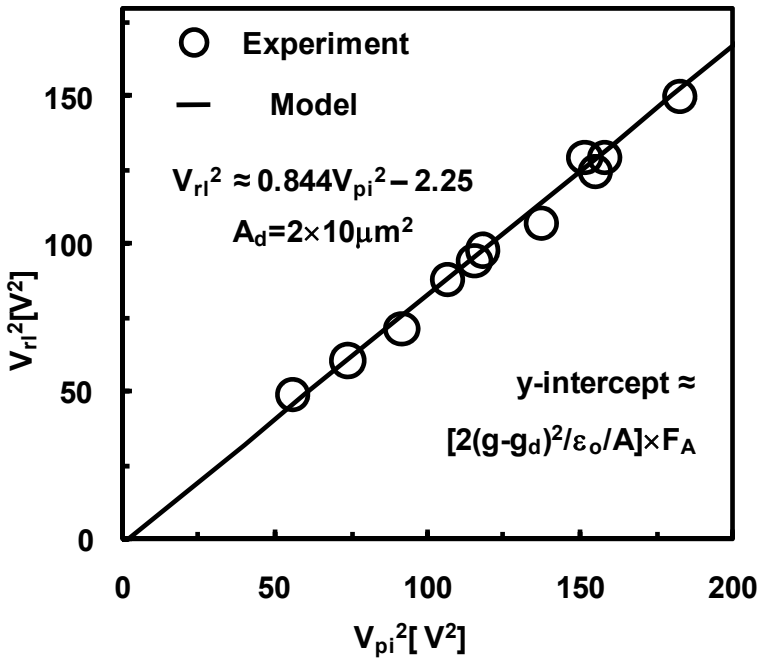


Fig. 4.10 V_{rl}^2 *vs.* V_{pi}^2 for the relays of Fig. 4.9. F_A is extracted to be 0.45μN.

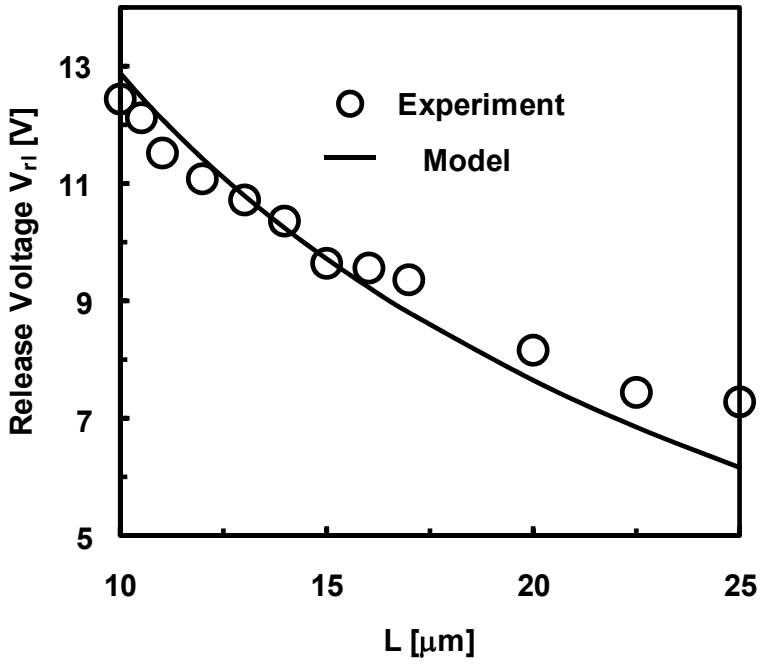


Fig. 4.11 Measured relay V_{rl} *vs. L*. The measured data is within 10% of the analytical model.

Finally, as previously alluded to, the serpentine spring design ensures that the relay is robust against thermal stress effects through expansion; therefore, the measured values of V_{pi}, V_{rl} and R_{ON} do not significantly change with temperature, as shown in Fig. 4.12 and 4.13. The negative Young's modulus temperature coefficient account for the slight reduction in V_{pi} and V_{rl} values.

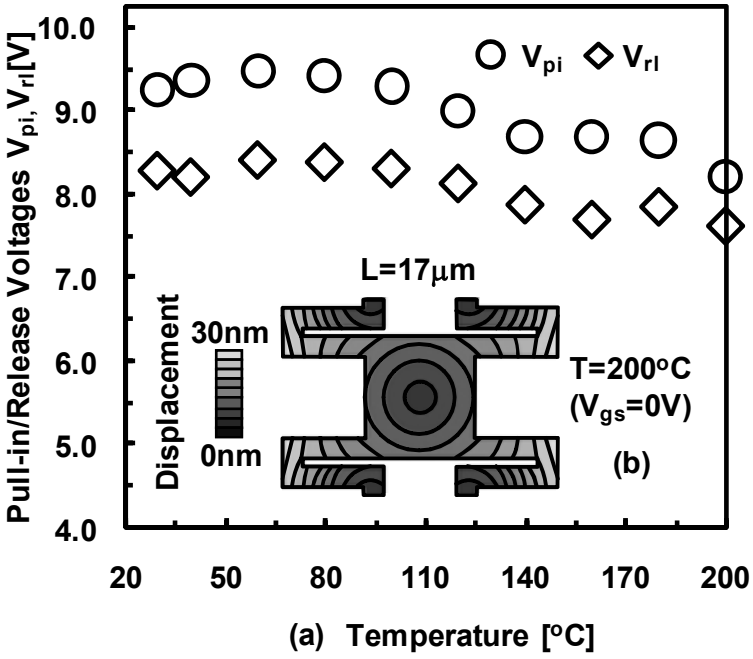


Fig. 4.12 a) Measured V_{pi} and V_{rl} *vs*. T for L=17μm. (b) ANSYS shows that the folded spring design releases the thermal/residual stress by expansion.

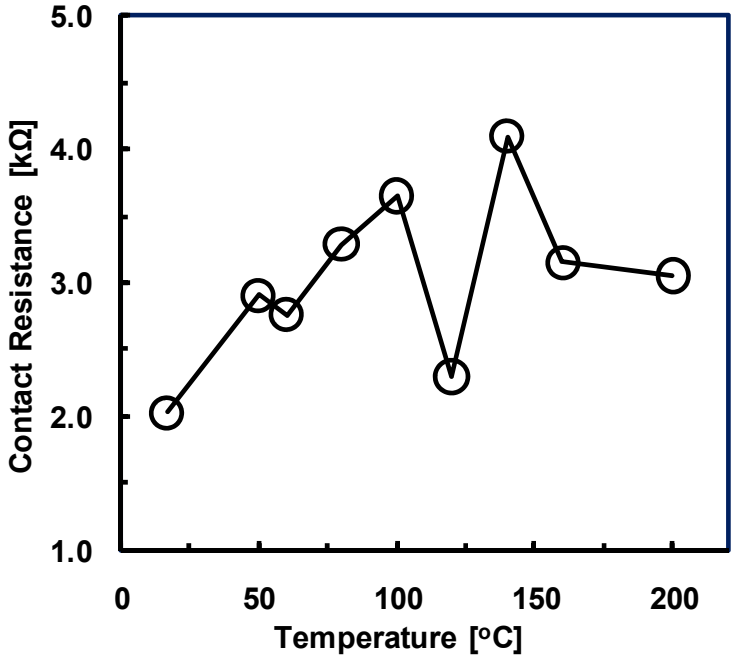


Fig. 4.13 Measured contact resistance *vs*. temperature for a L=40μm relay.

b. Dimple Support Design

When the relay is turned on, the actuation plate comes down and rests upon the contact dimples (Fig. 4.2b); proper relay design involves not only the choice of relay device dimensions that give the desired pull-in and release voltages, but also includes device dimensions that prevent the actuation pads from catastrophically pulling in and shorting the electrodes. To determine the catastrophic pull-in voltage (V_{cpi}) value, exact information for the contact dimples (e.g. surface roughness, contact friction, *etc*.) is needed, which is complicated. Fortunately, sacrificing some degree of accuracy, a conservative V_{cpi} estimation can once again be obtained by modeling the actuation pad as a simply-supported beam pinned at its two dimples,

as indicated in Fig. 4.2b. The bending of the actuation plate can be estimated by the Euler-Bernoulli equation [4.24]:

$$\left(\frac{E}{1-v^2}\right) I_A \frac{d^4 z^*}{dx^4} = \begin{cases} -\frac{\varepsilon_o W_A V^2}{2(g-g_d-z^*)^2} & in\ actuation\ region \\ 0 & otherwise \end{cases} \quad (4.6)$$

where v is the Poisson ratio of the structural material. $E/(1-v^2)$ instead of the Young Modulus E is used to account for the plate effects; I_A is the moment of inertia $=W_A h^3/12$, and $z^* = z - g_d$.

Eqn. (4.6) is solved with the appropriate boundary conditions at both dimple ends:

$$z^*(x=0) = z^*(x=L_A) = 0\ ;\ \left.\frac{d^2 z^*}{dx^2}\right|_{x=0} = \left.\frac{d^2 z^*}{dx^2}\right|_{x=L_A} = 0 \quad (4.7)$$

which gives the analytical formulation for V_{cpi} :

$$V_{cpi} \cong 1.516\sqrt{\frac{Eh^3(g-g_d)^3}{\varepsilon_o L_A^4}} \quad (4.8)$$

Eqn. 4.8 predicts a V_{cpi} value of 14.6V; ANSYS simulation gives V_{cpi}=18.9V. As previously alluded to, Eqn. 4.8 is a conservative estimate of V_{cpi}; therefore, the calculated V_{pci} is lower than the measured value, which is 23V.

4.4.2 Dynamic Performance Analysis

In addition to the static voltages, circuit designers are particularly interested in the mechanical delay time as it determines the circuit performance. When a bias voltage (V) is applied between gate and source electrodes, the motion of the actuation plate is governed by Newton's Second Law of Motion, which yields the following second order differential equation [4.26]:

$$m_{eff}\ddot{z} + \frac{\sqrt{k_{eff} m_{eff}}}{Q}\dot{z} + k_{eff} z = \frac{\varepsilon_o A V^2}{2(g-z)^2} \quad (4.9)$$

The right side of the equation is the electrostatic force, Q is the quality factor; z is the actuation plate displacement, and m_{eff} is the effective mass of transport.

To develop a rigorous delay model, it is worthwhile to spend some effort in deriving an accurate analytical formulation to the effective mass. To reach this goal, we first note that the effective mass consists of the mass of the actuation plate and the loaded mass from the springs, and can be determined from the total kinetic energy in the relay (KE_{tot}) and the velocity of the actuation plate (v_{p}) [4.27]:

$$m_{eff} = \frac{KE_{tot}}{\frac{1}{2}v_p^2} = \frac{\frac{1}{2}m_p v_p^2 + \frac{1}{2}m_t v_t^2 + \frac{1}{2}\int v_p^2 dm_b}{\frac{1}{2}v_p^2} = m_p + m_t\left(\frac{v_t}{v_p}\right)^2 + \frac{4\rho Wh}{[z_b(y)]^2}\int [z_b(y)]^2 dL \tag{4.10}$$

where m's and v's are the masses and velocities, and the subscripts p, t and b denote the actuation plate, trusses and the folded beams, respectively. The trusses are displaced by a distance $z/2$, and therefore,

$$v_t = \frac{1}{2}v_p \tag{4.11}$$

$z_{\text{b}}(\text{y})$ is the beam deflection. For beam AB, as depicted in Fig. 4.14,

$$z_b(y) = z\left[\frac{3}{2}\left(\frac{x}{L}\right)^2 - \left(\frac{x}{L}\right)^3\right] \tag{4.12}$$

And for beam CD

$$z_b(y) = z\left[1 - \frac{3}{2}\left(\frac{x}{L}\right)^2 + \left(\frac{x}{L}\right)^3\right] \tag{4.13}$$

Substituting Eqns. (4.11)-(4.13) into (4.10), m_{eff} is given by:

$$m_{eff} = m_p + \frac{1}{4}m_t + \frac{12}{35}m_b \tag{4.14}$$

Substitute the plate, truss and beam dimensions into Eqn. (4.14), m_{eff} can be simplified and expressed by:

$$m_{eff} = \alpha_o \rho A h + \alpha_1 \rho W L h \tag{4.15}$$

where α_o =1.93 and α_1 =2.74; ANSYS simulation predicts α_o=1.43and α_1=3.83.

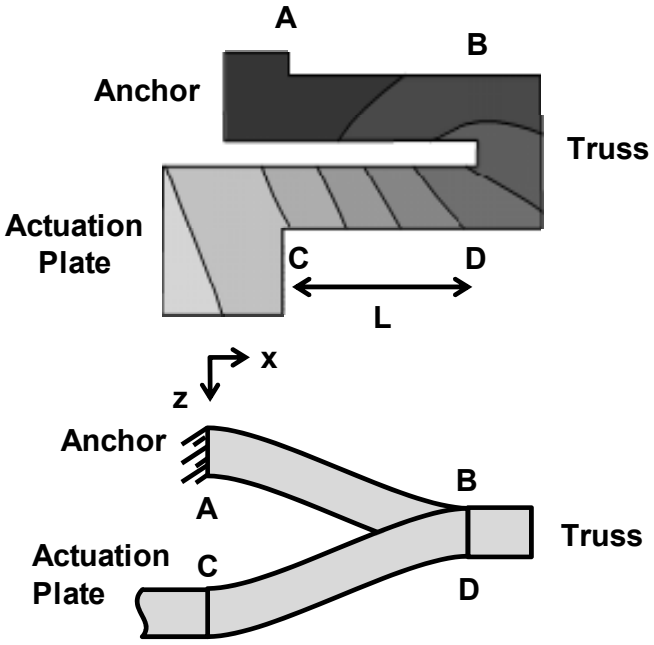


Fig. 4.14. The effective mass of the folded spring can be estimated by the beam kinetic energy, which is determined by the mode shape of segment AB and CD.

Substituting Eqn. (4.4) and Eqn. (4.15) into Eqn. (4.9), the plate position is obtained by solving the resultant equation. But in order to provide insight for relay design, we analytically approximate t_{delay} by the following expression:

$$t_{delay} \cong \alpha \sqrt{\frac{m_{eff}}{k_{eff}}} \left(\frac{g_d}{g}\right)^{\gamma} \left(\frac{V_{dd}}{V_{pi}} - \chi\right)^{-\beta} \quad \text{for} \quad 5V_{pi} \geq V_{dd} > 1.1V_{pi}\,, g_d \geq g/3 \tag{4.16}$$

where $\chi \cong 0.8$; α , β and γ , to the first order, depend only on Q and their values are plotted in Fig. 4.15. The details of the derivation are discussed in Appendix I. Over the range of interest, Eqn. (4.16) predicts t_{delay} values within 20% of Eqn. (4.9), and the accuracy improves as V_{dd}/V_{pi} increases. Note that Eqn. 4.16 shows that t_{delay} not only depends on the relay resonant frequency $\sqrt{k_{eff}/m_{eff}}$, but also on the V_{dd}/V_{pi} ratio, which from now on we denote it as the "gate overdrive".

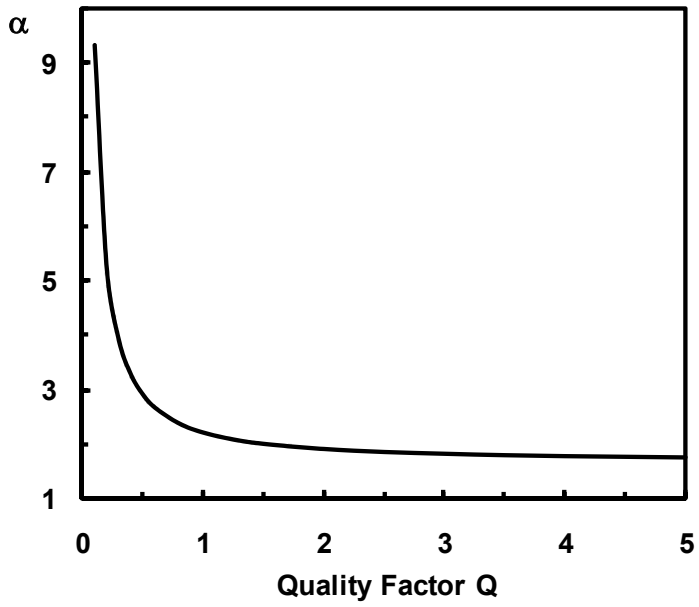


Fig. 4.15a. Dependence of α on quality factor, for $5V_{pi} \geq V_{dd} > 1.1V_{pi}$, $g_d \geq g/3$

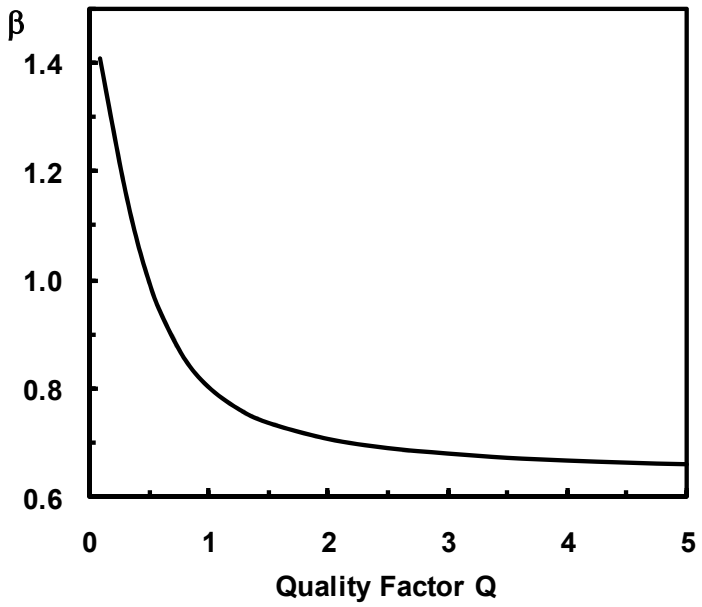


Fig. 4.15b. Dependence of β on quality factor, for $5V_{pi} \geq V_{dd} > 1.1V_{pi}$, $g_d \geq g/3$

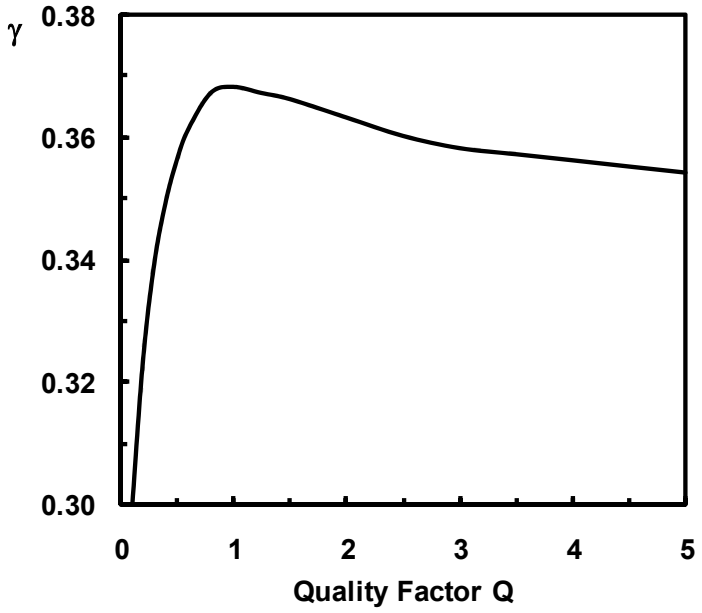


Fig. 4.15c. Dependence of γ on quality factor, for $5V_{pi} \geq V_{dd} > 1.1V_{pi}$, $g_d \geq g/3$

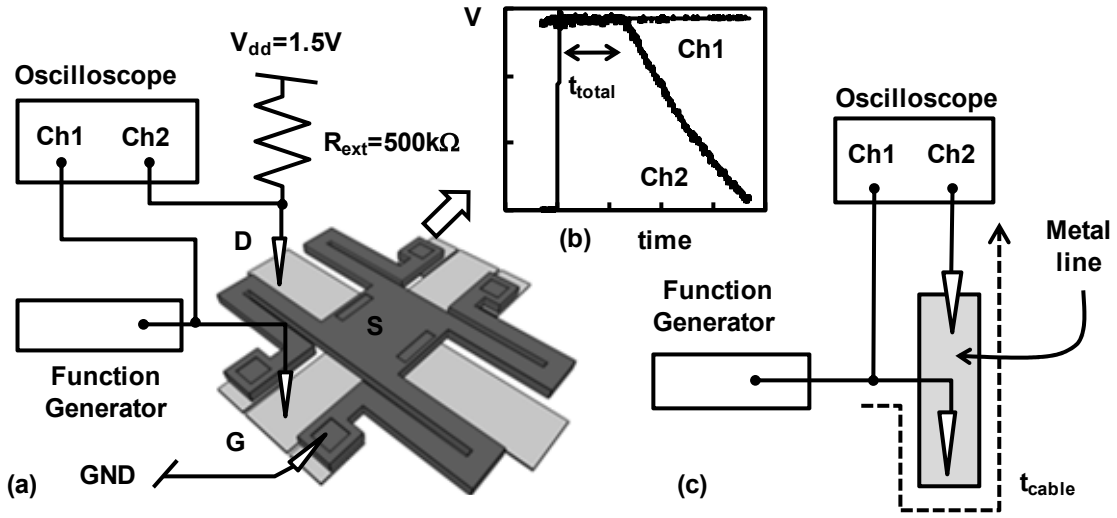


Fig. 4.16 (a) Relay delay measurement setup. Measurements were made in N_2 ambient (b) The total delay t_{total} can be extracted from the difference between the input and output signals. (c) t_{delay}= t_{total} –$t_{cable,}$ where t_{cable} is the electrical delay of the cable.

To calibrate the delay model, relay delay measurements were taken in nitrogen ambient using the test setup shown in Fig. 4.16. As shown in the figure, a load resistor is utilized in this measurement setup to allow for application of a step voltage to the gate electrode and measure to trigger a voltage drop (delayed by t_{total}) at the drain electrode (Fig. 4.16b). To separate the cable delay (t_{cable}) from the pull-in time, t_{cable} is first measured using an on-chip, dummy metal line (Fig. 4.16c); t_{delay} can then be extracted from the difference between t_{total} and t_{cable}. Utilizing this measurement setup, Fig. 4.17 shows the measured delay of three relays with L=14μm, 40μm and 50μm at different V_{dd} values. t_{delay} decreases with increasing V_{dd} and $1/k_{eff}$, as expected. Note that no contact bounce is observed; measured t_{delay}

values match predictions using Eqn. 4.9 within 20%, with an extracted quality factor value $Q = 0.3$.

Finally, since the relay is purely electrostatically actuated, all of the energy consumed by switching it on and off is supplied by the voltage source driving the gate electrode and any wiring and load capacitance at the drain electrode. The total energy consumed in switching the relay on and off is simply set by the charge supplied by the supply voltage:

$$E_s \cong \left(\frac{\varepsilon A}{g-g_d} + C_L\right) V_{dd}^2 \tag{4.17}$$

where C_L is extrinsic load capacitance. With a reliable relay technology, together with accurate static and dynamic models properly developed, we are now ready to optimize relay based digital circuits and project the performance of scaled relays. These topics will be addressed in Chapter 5.

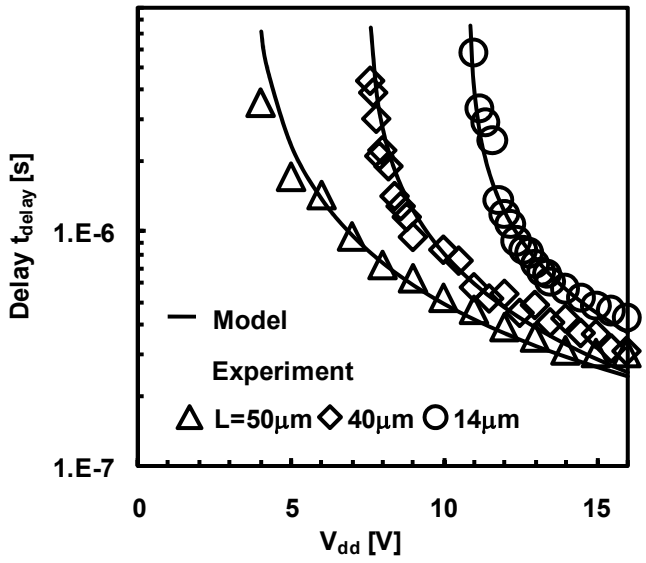


Fig. 4.17 Measured t_{PI} *vs.* V_{DD} for three different relays.

4.5 Conclusion

A pathway to enable reliable micro-relays for digital logic applications through proper contact design is proposed and demonstrated with TiO_2-coated W contacting electrodes. Prototype relays fabricated utilizing a CMOS-compatible, poly-$Si_{0.4}Ge_{0.6}$ surface micromachining process have been demonstrated to operate at near 5V supply voltage with reasonable (~10kΩ) contact resistance over a wide temperature range (20 ^{o}C -200^{o}C). The devices have low surface adhesion energy (~4μJ/m^2) and can endure more than 10^9 on/off hot-switching cycles in N_2 ambient without stiction- or welding-induced failure. In order to facilitate design optimization and scaling for relay-based circuits, relevant models and analytical formulations for relay performance are also developed in this chapter. It should be note that the relatively high supply voltages demonstrated in this chapter do not necessarily represent the ultimate voltage limit for relay technology. Through proper relay energy-delay optimization and device scaling, scaled relays will offer substantial reduction in supply voltage and improvements in energy efficiency over CMOS. These topics will be addressed in chapter 5.

4.6 References

[4.1] K. Akarvardar, D. Elata, R. Parsa, G. C. Wan, K. Yoo, J. Provine, P. Peumans, R. T. Howe, H.-S. P. Wong, "Design Considerations for Complementary Nanoelectromechanical Logic Gates," in *IEDM Tech. Dig.*, 2007, pp. 299-302.

[4.2] F. Chen, H. Kam, D. Markovic, T.J. King, V. Stojanovic, and E. Alon, "Integrated Circuit Design with NEM Relays," in *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, 2002, pp. 750-757.

[4.3] R. Nathanael, V. Pott, H. Kam, J. Jeon and T.-J. King-Liu,, "4-Terminal Relay Technology for Complementary Logic," to appear IEEE International Electron Device Meeting, Dec. 2009.

[4.4] H. Kam, V. Pott, R. Nathanael, J. Jeon, E. Alon and T.-J. King-Liu, "Design and Reliability of a Micro-Relay Technology for Zero-Standby-Power Digital Logic Applications," to appear IEEE International Electron Device Meeting, Dec. 2009.

[4.5] H. Kam, T.-J. King-Liu, E Alon, M. Horowitz, "Circuit Level Requirements for MOSFET Replacement Devices," in *IEDM Tech. Dig.*, 2008, pp. 427.

[4.6] G.-L. Tan and G.M. Rebeiz, "A DC-contact MEMS shunt switch," *IEEE Microwave Wireless Compon. Lett.,* vol. 12, pp. 212-214, Jun. 2002.

[4.7] P. M. Zavracky, S. Majumder, and N. E. McGruer, "Micromechanical switches fabricated using nickel surface micromachining," *IEEE J.Microelectromech. Syst.,* vol. 6, no. 1, pp. 3–9, 1997.

[4.8] C. Goldsmith, J. Randall, S. Eshelman, T.H. Lin, D. Denniston, S. Chen and B. Norvell, "Characteristics of micromachined switches at microwave frequencies," in *1996 IEEE MTT-S Int. Microwave Symp. Dig.*, San Francisco, CA, Jun. 1996, pp. 1141-1144

[4.9] A.Q. Liu, M. Tang, A. Agarwal and A. Alphones, " Low-loss lateral micromachined switches for high frequency applications," in *J. Micromech. Microeng.* vol. 15 pp.157-167, 2005

[4.10] B. Nikolic and P. B. Allen, "Electron transport through a circular constriction," *Phys. Rev. B*, vol. 60, no. 6, pp. 3963-3969, 1999.

[4.11] J.B. Muldavin and G.M. Rebeiz, "Inline capacitive and DC-contact MEMS shunt switches," *IEEE Microwave Wireless Compon. Lett.*, vol. 11, pp. 334-336, Aug 2001

[4.12] R. E. Mihailovich, M. Kim, J. B. Hacker, E. A. Sovero, J. Studer, J. A. Higgins, and J. F. DeNatale, "MEM relay for reconfigurable rf circuits," *IEEE Microw. Wireless Compon. Lett.*, vol. 11, pp. 53–55, Feb. 2001.

[4.13] G.M. Rebeiz, *RF MEMS Theory, Design and Technology*, New York: Wiley, 2003.

[4.14] E. J. J. Kruglick and K. S. J. Pister, "Lateral MEMS microcontact considerations,"*J. Microelectromech. Syst.*, vol. 8, pp. 264–271, Sept. 1999.

[4.15] B. D. Jensen, K. Huang, L. L. W. Chow, and K. Kurabayashi, "Adhesion effects on contact opening dynamics in micromachined switches," *J. Appl Phys.*, vol. 97, no. 10, p. 103 535, May 2005.

[4.16] S. P. Sharma, “Adhesion coefficients of plated contact materials,” *J. Applied Phys.*, vol. 47, no. 8, pp. 3573-3576, Aug. 1976.

[4.17] M.E. Sikorski, “The adhesion of metals and factos that influence it,” *Wear,* **7**, 144, 1964

[4.18] M. R. Houston, R. Maboudian, and R. T. Howe, “Ammonium fluoride surface treatments for reducing in-use stiction in polysilicon microstructures,” *Proc. 8th Int. Conf. Solid-State Sensors and Actuators (Transducers'95)* Stockholm, Sweden, pp. 210-213, June 25–29, 1995.

[4.19] C. W. Low, "Novel Processes for Modular Integration of Silicon-Germanium MEMS with CMOS electronics," *Ph.D. Thesis, Dept. of EECS, UC Berkeley, 2007*

[4.20] A. E. Franke, Y. Jiao, M. T. Wu, T.-J. King, and R. T. Howe, “Post-CMOS modular integration of poly-SiGe microstructures using poly-Ge sacrificial layers,” in *Solid-State Sensor and Actuator Workshop* Hilton Head, S.C, June 2000, pp. 18-21

[4.21] B. D. Jensen, L. L. W. Chow, K. Huang, K. Saitou, J. L. Volakis, and K. Kurabayashi, “Effect of nanoscale heating on electrical transport in RF MEMS switch contacts,” *J. Microelectromech. Syst.*, vol. 14, no. 5, pp. 935–946, Oct. 2005.

[4.22] R. Holm and E. Holm, *Electric Contacts; Theory and Application*, 4th ed. Berlin, Germany: Springer-Verlag, 1967.

[4.23] W.Weaver, Jr., S. P. Timoshenko, and D. H. Young, *Vibration Problems in Engineering*, 5th ed. New York: Wiley, 1990.

[4.24] S. P. Timoshenko and J. M. Gere, Mechanics of Materials. Pacific Grove: Brooks/Cole, 2001.

[4.25] D. Lee, V. Pott, H. Kam, R. Nathanael and T.-J. King Liu, "AFM Characterization of Adhesion Force in Micro-Relays," *submitted to MEMS 2010*

[4.26] R. K. Gupta and S. D. Senturia, "Pull-in time dynamics as a measure of absolute pressure," in *Proc. MEMS 1997*, pp. 290–294.

[4.27] R. A. Johnson, *Mechanical Filters in Electronics*. New York: Wiley, 1983.

Chapter 5

Optimization and Scaling of Micro-Relays for Logic Applications

5.1 Introduction

With a pathway to reliable micro-relay for digital logic applications experimentally demonstrated in Chapter 4, the focus of this Chapter is on achieving fast, energy-efficient, and compact relay-based digital circuits. To date, no systematic optimization and scaling methodology for logic relays has been proposed. To remedy this issue, this Chapter begins with a sensitivity-based energy-delay optimization methodology in Section 5.2, allowing the establishment of simple relay design guidelines. Based upon these guidelines, together with the measured adhesion force scaling with contact dimple area, we then propose a scaling methodology for micro-relays in Section 5.3, which leads to systematic improvements in device density, performance, and energy consumption. Simulation results indicate that scaled relay technology may offer >10×

improvement in energy efficiency for applications requiring performance up to ~100MHz. Finally, section 5.4 concludes this chapter .

5.2 Relay Energy-Delay Optimization

Having developed the relevant relay energy-delay models in Chapter 4, the focus of this chapter is now on optimizing relay-based circuits. As was described in [5.1]-[5.3], the optimal topology for relay circuits is drastically different from that for CMOS circuits. Specifically, because the delay of relay-based circuits is dominated by the mechanical delay rather than the electrical RC delay, an optimized relay-based circuit should (as shown in Fig. 5.1) consist of single stage complex gates [5.2, 5.4] such that the mechanical motions of all the relays in the circuit occur simultaneously[3].

In optimizing the relay design for the appropriate circuit topology, it is important to note that, as for CMOS circuit design [5.5- 5.10], relay energy and performance trade off against each other. The goal of the optimization is to minimize relay circuit delay subject to a given energy budget, which in essence boils down to solving the following constrained optimization problem for an N-relay stack:

Minimize: $t_{delay} \cong \alpha \sqrt{\frac{m_{eff}}{k_{eff}}} \left(\frac{g_d}{g}\right)^{\gamma} \left(\frac{V_{dd}}{V_{pi}} - \chi\right)^{-\beta}$

Subject to: $E_{tot} = \sum_{i=1}^{N} \left(\frac{\varepsilon A_i}{g - g_d} + C_{L,i}\right) V_{dd}^2$ (5.1)

[3] Note that this only work for four-terminal relays.

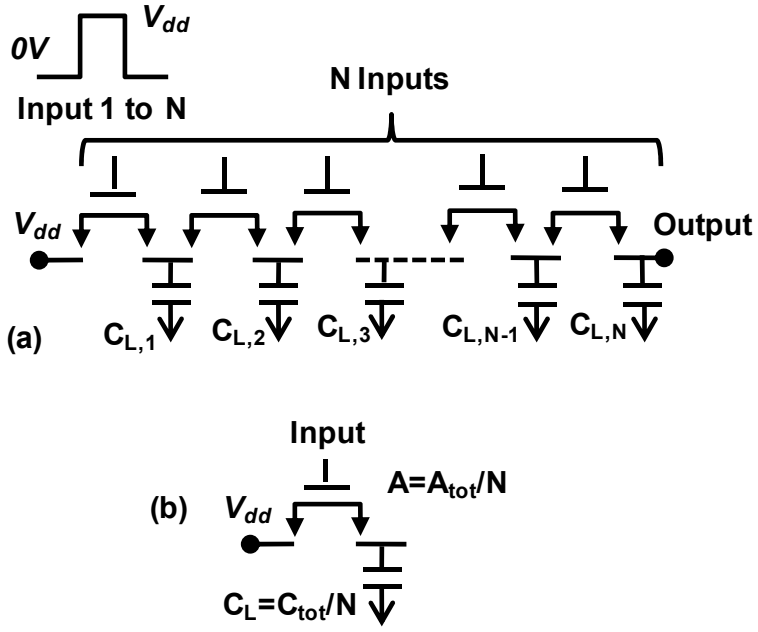


Fig. 5.1 (a) Optimal relay circuit topology. (b) The energy-delay optimization problem can be simplified to optimizing one single relay driving the average capacitance C_{tot}/N

where E_{tot} is the total energy; A_i and C_{Li} are the area and the load capacitance, respectively, of the i-th relay in the stack. Assuming all relays have the same fabricated gap thickness g and dimple gap thickness g_d, the total energy can be rewritten as:

$$E_{tot} = \left(\frac{\varepsilon A_{tot}}{g-g_d} + C_{tot}\right) V_{dd}^2 \tag{5.2}$$

where A_{tot} and $C_{,tot}$ are the total actuation area and load capacitance, respectively. Given that the electrical delay is negligible when compared against the mechanical delay, then by symmetry, an optimized relay circuit design would size all relays identically so that the devices have the same switching delay and energy consumption:

$$E_s = \frac{E_{tot}}{N} = \left(\frac{\varepsilon A}{g - g_d} + C_L\right) V_{dd}^2 \quad \text{where } A_{tot} = N \times A,\ C_L = C_{tot}/N \tag{5.3}$$

Therefore, the energy-delay optimization problem is now reduced to optimizing the energy-delay of a single relay driving one capacitive load C_L, where C_L is the average capacitance driven by each relay. To solve this constraint optimization problem, we first apply the sensitivity-based analysis to explore the relay energy-delay tradeoff.

5.2.1 Sensitivity analysis

Sensitivity-based analysis for CMOS circuit optimization has been extensively explored in [5.5-5.10] to explore digital integrated circuit energy-delay optimization. The optimization can equally applied to optimize relay circuits, therefore the key concepts are here briefly reviewed. The energy-delay sensitivity of a given tuning variable *var* is defined as:

$$S_{var} \equiv \frac{\partial t_{delay}/\partial var}{\partial E_s/\partial var} \tag{5.4}$$

which is interpreted as the delay reduction per energy cost by adjusting variable *var*. For most relay designs, the beam width W is the minimum feature size set by photolithography; thickness h and the minimum dimple gap thickness g_d are set by process constraints. Once the thickness is fixed, the quality factor and therefore the α, β and γ values are known. Therefore, the supply voltage V_{dd}, actuation area A, fabricated gap thickness g and the beam length L are the available design variables for optimization. By adjusting these variables, the optimal relay design is reached when the sensitivities to all tuning variables are balanced [5.5, 5.6, 5.9]. With these

goals in mind, we herein derive the analytical formulations for sensitivities to these variables.

i. Sensitivity to Supply Voltage

We begin our analysis by exploring the sensitivity of delay to energy due to the change in V_{dd}. As V_{dd} increases, the switching delay decreases because the electrostatic force increases. Of course, the switching energy also increases with V_{dd}. With the power-law dependences of both delay and switching energy on V_{dd}, the resultant negative sensitivity to supply voltage is given by Eqn. 5.5:

$$\frac{\partial t_{delay}/\partial V_{dd}}{\partial E_s/\partial V_{dd}} = -\frac{\beta t_{delay}}{2E_s} \times \left(\frac{V_{dd}}{V_{pi}}\right) \Big/ \left(\frac{V_{dd}}{V_{pi}} - \chi\right) \tag{5.5}$$

Where t_{delay} and E_s are respectively the nominal delay and energy at a given V_{dd}. Note that, as will be discussed later, typical a V_{dd}/V_{pi} value lies within the range 1.5-3. Therefore the normalized sensitivity [5.11], which is defined as $\frac{E_s}{t_{delay}} \times \frac{\partial t_{delay}/\partial V_{dd}}{\partial E_s/\partial V_{dd}}$, is roughly –(0.7-1.1)×β. Hence, as a simple rule of thumb, every 2X energy increase can be sacrificed for ~1.6-1.8X reduction in relay delay by V_{dd} adjustment.

ii. Sensitivity to Actuation Area

In addition to V_{dd} adjustment, relay sizing is also an effective means to adjust the tradeoff between energy and delay. The sensitivity of delay to energy due to the change in the actuation area is given by Eqn. 5.6:

$$\frac{\partial t_{delay}/\partial A}{\partial E_s/\partial A} = \frac{t_{delay}}{2E_s}\left[\frac{\alpha_o \rho h A}{m_{eff}} - \beta\left(\frac{V_{dd}}{V_{pi}}\right) \Big/ \left(\frac{V_{dd}}{V_{pi}} - \chi\right)\right]\left(1 + C_L \Big/ \left(\frac{\varepsilon_o A}{g - g_d}\right)\right) \tag{5.6}$$

As the actuation area A increases, the on-state capacitance and therefore the switching energy increases. On the other hand, increasing A has a two-fold impact on the relay switching speed: the pull-in time decreases due to the reduction in V_{pi}; but the increase in the actuation mass counteracts this effect. For high-Q relays (i.e. small β), as will be discussed later, the resonant frequency term dominates switching delay. Therefore, reducing the actuation area is attractive because it decreases both the switching delay and energy, leading to a positive sensitivity. But it is important to note that the sensitivity diminishes with reduction in actuation area because, the gate overdrive V_{dd}/V_{pi} decreases and the spring-loaded mass becomes increasingly significant in determining the resonant frequency. All these eventually lead to a negative sensitivity.

iii. Sensitivity to Fabricated Gap Thickness

In designing micro-relays, it is desirable to use use the thinnest dimple gap (g_d) to minimize the travelling distance of the actuation plate. A thinner fabricated actuation gap (g) also provides for larger actuation force. However, it is important to note that reducing the fabricated gap thickness increases the on-state capacitance and therefore the switching energy. This results in a negative sensitivity to g, as shown by Eqn. 5.7:

$$\frac{\partial t_{delay}/\partial g}{\partial E_s/\partial g} = -\frac{t_{delay}}{E_s}\left(1-\frac{g_d}{g}\right)\left(1+C_L/\left(\frac{\varepsilon_o A}{g-g_d}\right)\right)\left(-\gamma+\frac{3\beta}{2}\left(\frac{V_{dd}}{V_{pi}}\right)/\left(\frac{V_{dd}}{V_{pi}}-\chi\right)\right) \quad (5.7)$$

iv. Sensitivity to Beam Length

Finally, if the beams do not contribute substantial capacitance, then to first order, relay switching energy is independent of the beam length. Therefore, the

beam length can be optimized to maximize the switching speed with no energy penalty. Before doing so, it is important note that the pull-in time depends not only on the resonant frequency, but also on the gate overdrive. Therefore, increasing the beam length has a dual impact on the relay switching speed: the pull-in time decreases due to the reduction in V_{pi}, but the increase in the loaded mass counteracts this effect. Therefore, there exists an optimal beam length that properly balances the resonant frequency and the gate overdrive:

$$\frac{dt_{delay}}{dL} = 0 \Rightarrow \left[\left(\frac{1}{\omega_o}\frac{\partial \omega_o}{\partial L}\right)/\left(\frac{1}{V_{pi}}\frac{\partial V_{pi}}{\partial L}\right)\right] \times \left[1 - \chi\left(\frac{V_{dd}}{V_{pi}}\right)^{-1}\right] = \beta \qquad (5.8)$$

where the left-hand side if the equation is proportional to $m_{eff}\times[1\text{-}\chi(V_{dd}/V_{pi})^{-1}]$. Therefore, for high-$Q$ relays (i.e. low β value), short beams are preferred to decrease the mass and increase the relay resonant frequency. For low-Q relay, on the other hand, one would prefer to use long beams to decrease V_{pi} and therefore increase the gate overdrive.

5.2.2 Relay Design Optimization

With analytical expressions for the sensitivities, the relay design can now be optimized. As previously alluded to, this goal can be reached by balancing the sensitivities of all of the design variables. This means that we can use the normalized sensitivity to supply voltage (–(0.7-1.1)×β), to pick *all* other design variables. It also implies that for an optimized relay, every 2× energy increase can be sacrificed for ~1.6-1.8× reduction in relay delay by changing *any* design

variable. Using this result, simple guidelines for energy-efficient relay design are established in this section.

i. Optimal Gap Thickness Ratio

We begin the optimization process by exploring the optimal g_d/g ratio. By balancing the sensitivities to V_{dd} and to g,

$$-\frac{1}{2}\varphi = -\left(1-\frac{g_d}{g}\right)(1+f)\left(-\gamma+\frac{3}{2}\varphi\right) \tag{5.9}$$

where $\varphi = \beta\left(\frac{V_{dd}}{V_{pi}}\right)\Big/\left(\frac{V_{dd}}{V_{pi}}-\chi\right)$, $f = C_L\Big/\left(\frac{\varepsilon_o A}{g-g_d}\right)$ is the fan-out of the relay. Since $\gamma\sim0.3$ and $\beta\sim1$, $\gamma<<1.5\varphi$, the optimal g_d/g ratio is approximated by Eqn. 5.10:

$$\frac{g_d}{g} \approx \frac{2+3f}{3+3f} \tag{5.10}$$

Therefore the optimal g_d/g ratio is roughly 0.66-0.75 across a large fan-out range, and this value is largely independent of the quality factor value (β). This implies that pull-in operation is preferred for optimal energy efficiency. These results are consistent with the simulation results shown in **Fig. 5.2**.

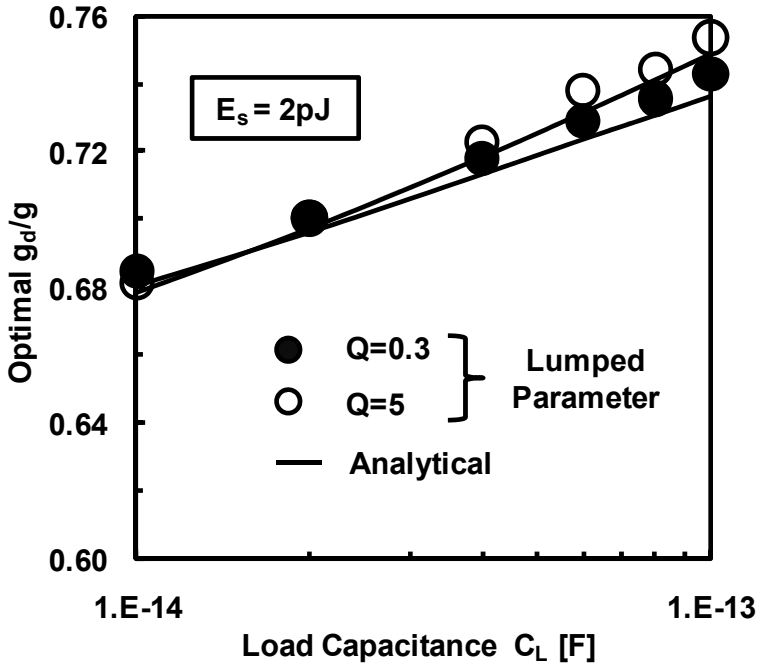


Fig. 5.2. The optimal relay g_d/g ratio stays roughly at 2/3, i.e. pull-in operation is preferred, for a wide range of design parameters.

ii. Optimal Actuation Area and Supply Voltage

Once the g_d/g ratio is known, the relay can be sized optimally if the optimal relay fan-out is known. Although the exact analytical formulation for optimal fan-out is complex, since mechanical delay dominates, the effect of the relay sizing on the RC delay of the gate input signal is negligible. As a result, all that is left is the quadratic dependence of energy on V_{dd} vs. the linear dependence on gate capacitance. If the relay has to drive a load capacitance under a given energy constraint, it would be preferable to upsize the relay and lower V_{dd} to reduce the energy spent on the load capacitance. In doing so, the pull-in voltage ($V_{pi} \propto A^{-1/2}$) drops at a faster rate than the supply voltage ($V_{dd} \propto \sqrt{\frac{E_S}{C_L+\frac{\varepsilon_0 A}{g-g_d}}} \propto (1+\mathrm{A})^{-0.5}$), and

results in an increase in the gate overdrive. Together with the fact that delay is relatively insensitive to sizing ($t_{delay} \propto A^{(1-\beta)/2} \sim A^{0.1}$), it can be concluded that it is worthwhile to upsize the relay gate in order to allow for an overall reduction in delay. Therefore the optimal fan-out is less than one, as shown in Fig. 5.3.

Once the optimal values for the fabricated gap thickness and actuation area are obtained, the optimal V_{dd} is simply set by the energy constraint. The upper bound for V_{dd} value is the catastrophic pull-in voltage.

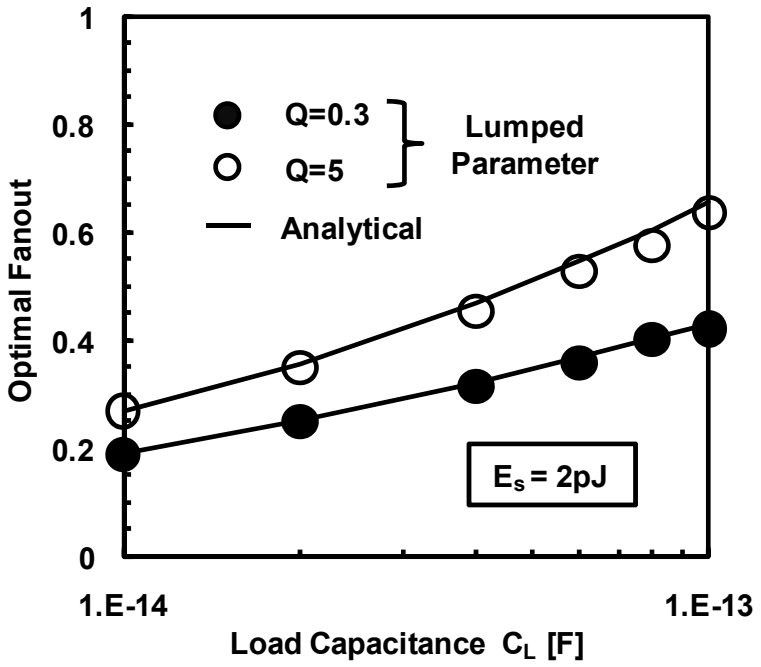


Fig. 5.3. Optimal fan-out for a relay for a range of design parameters.

iii. Optimal V_{dd}/V_{pi} and beam length

As previously alluded to, the beam length dictates the balance between the relay gate overdrive and the resonant frequency. It can be shown at the optimal beam length, the gate overdrive and the resonant frequency are balanced when

$$\varphi = 1 - \kappa\left(1 - \frac{\alpha_o \rho h A}{m_{eff}}\right) \tag{5.11}$$

Where $\varphi = \beta\left(\frac{V_{dd}}{V_{pi}}\right)\Big/\left(\frac{V_{dd}}{V_{pi}} - \chi\right)$ and $\kappa = \left(\frac{k_{eff}}{L}\right)\Big/\left(\frac{dk_{eff}}{dL}\right)$. κ is roughly -1/3 for most relay designs.

To compute the optimal gate over drive, once needs to know $\alpha\rho hA/m_{\text{eff}}$ ratio. Optimal $\alpha\rho hA/m_{\text{eff}}$ ratio can be obtained by balancing the sensitivities to area and to fabricated gap thickness:

$$\frac{1}{2}\left[\frac{\alpha_o \rho h A}{m_{eff}} - \beta\left(\frac{V_{dd}}{V_{pi}}\right)\Big/\left(\frac{V_{dd}}{V_{pi}} - \chi\right)\right] = -\left(1 - \frac{g_d}{g}\right)\left(-\gamma + \frac{3\beta}{2}\left(\frac{V_{dd}}{V_{pi}}\right)\Big/\left(\frac{V_{dd}}{V_{pi}} - \chi\right)\right)$$

Since the optimal g_d/g value is roughly 0.7 and γ~0.3 is negligible, the optimal $\alpha\rho hA/m_{\text{eff}}$ ratio is proportional to φ, as shown in Eqn. 5.12:

$$\frac{\alpha_o \rho h A}{m_{eff}} \cong \varphi\left(-2 + 3\frac{g_d}{g}\right) = 0.1\varphi \tag{5.12}$$

Substituting Eqn. 5.12 into Eqn. 5.11 the optimal gate overdrive can be obtained:

$$\beta\left(\frac{V_{dd}}{V_{pi}}\right)\Big/\left(\frac{V_{dd}}{V_{pi}} - \chi\right) = \varphi = \frac{1-\kappa}{1-0.1\kappa} \approx 1.29 \tag{5.13}$$

For low-Q relays with β > 0.8, $V_{\text{dd}}/V_{\text{pi}}$>>2 and therefore long beams are preferred. However, it is important to keep in mind that the longest beam length will be set by the surface adhesion energy, or by layout area constrains. On the other hand, for high-Q relays with β≈0.6, the optimal $V_{\text{dd}}/V_{\text{pi}}$ ratio lies within the range 1.5-2. Therefore, shorter beams are preferred. These results are consistent with the simulation results shown in Fig. 5.4 and Fig. 5.5.

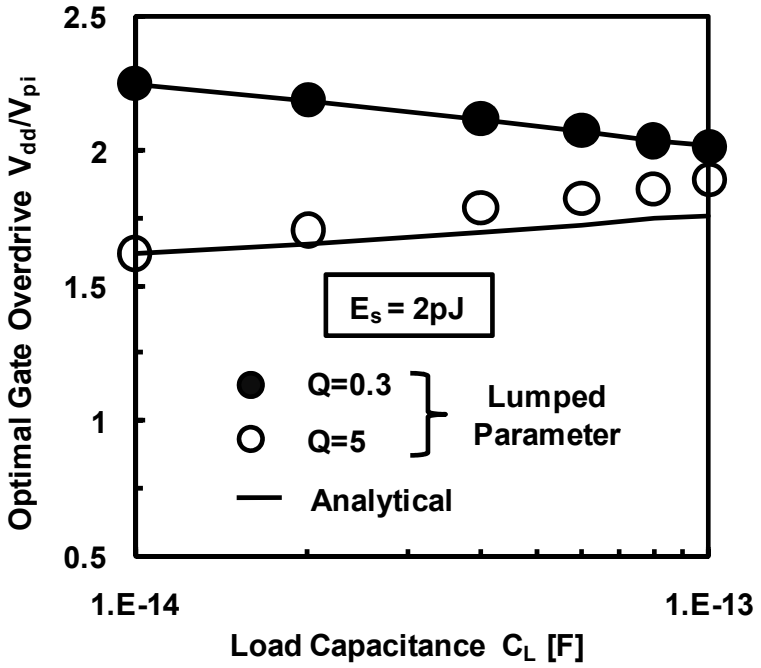


Fig. 5.4 The optimal gate overdrive (V_{dd}/V_{pi}) value for high and low-Q relays.

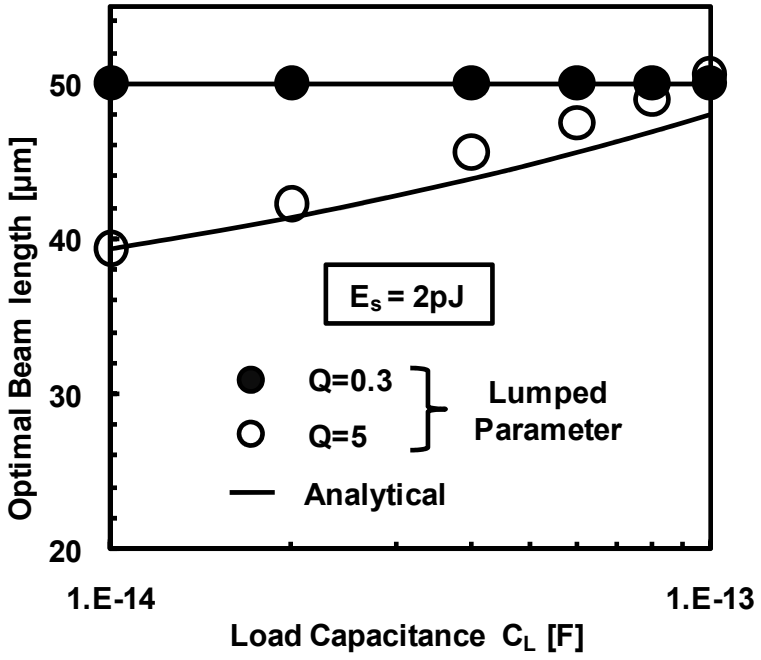


Fig. 5.5 The optimal relay beam length is chosen to achieve the required gate overdrive (**Fig. 5.4**); low-*Q* relays have higher gate overdrive and therefore longer beams are preferred. Note that we assume the longest beam length is 50μm in this study.

5.2.3 Relay Optimization Example

Simple guidelines for designing energy-efficient relays have already been established in this section. As a concrete example, the previously described methodology is used to optimize a 5μm wide relay (with parameters shown in Table I). The relay delay is optimized for E_s=2pJ with total load capacitance ranging from 10fF to 100fF, and the results are shown in Fig. 5.6. As expected, the delay increases with increasing load capacitance. In addition, as also previously alluded to, relays with high-Q values do not provide substantial (~2X) speed improvement over their low-Q counterparts.

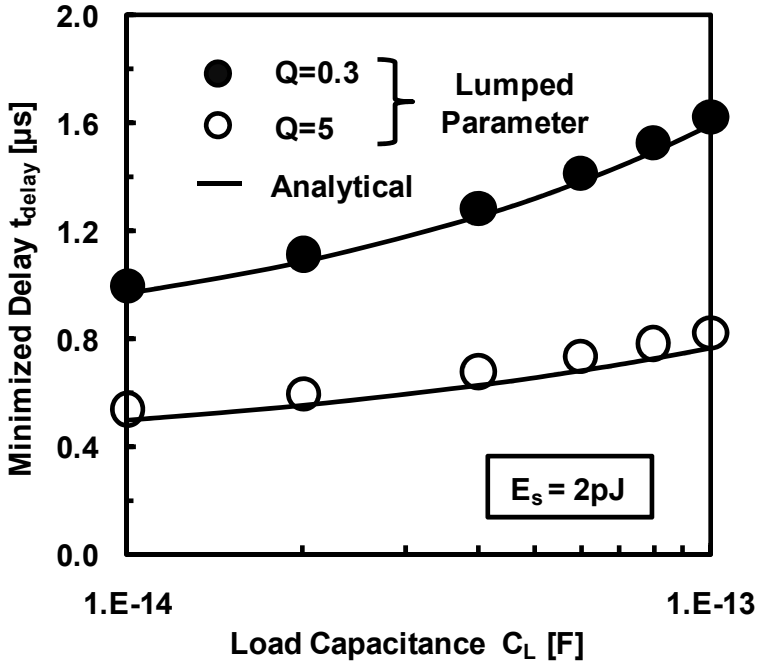


Fig. 5.6 For a given energy constraint and load capacitance, relay delay can be minimized by properly choosing the relay design parameters. The lumped parameter model is in close agreement with the analytical model.

For ultra-low power electronics applications such as wireless sensor networks, switching energy [**5.12**] (rather than speed) is the primary concern, and hence the relay's minimum switching energy is the most important metric. This minimum switching energy is dictated by the need to overcome surface adhesion energy (Γ) in order to break physical contact:

$$0.5\, k_{\text{eff}} g_{\text{d}}^2 \geq \Gamma \tag{5.14}$$

This gives the minimum pull-in and supply voltages:

$$V_{dd,min} = V_{pi,min} = \sqrt{\frac{16\Gamma g^3}{27\varepsilon_o A g_d^2}} \tag{5.15}$$

If the load capacitance is ignored, the relay switching energy is expressed by:

$$E_s = \left(\frac{\varepsilon A}{g-g_d}\right) V_{dd,min}^2 = \left(\frac{16\Gamma}{27}\right) \frac{1}{(1-g_d/g)(g_d/g)^2} \tag{5.16}$$

which has a minimum value of 4Γ at $g_d/g = 2/3$. Of course the switching energy will be higher for any practical relay designs because a relay operating at exactly $V_{dd}=V_{pi}$ has no noise margin and is highly susceptible to process or environmental variations.

5.3 Relay Scaling

The need for large supply voltage and layout area remains an issue for the relays demonstrated in this work, but can be alleviated if the relay dimensions are properly scaled down. In a manner very much analogous to the classic scaling theory developed for MOSFETS [5.13], extensive treatments of constant-field scaling for micro-electromechanical-systems (MEMS) have been reported [5.14]. This scaling methodology maintains the electric field across the actuation gap at a

constant value while all of the dimensions of the device are scaled by a factor *S*. Although this simple scaling methodology provides useful relay scaling insights, the scaled relay may not correspond to the optimal design for a given operating situation. To remedy this, we expand upon prior work by developing an optimized relay scaling methodology and assess its implications for relay switching speed, energy, and layout area.

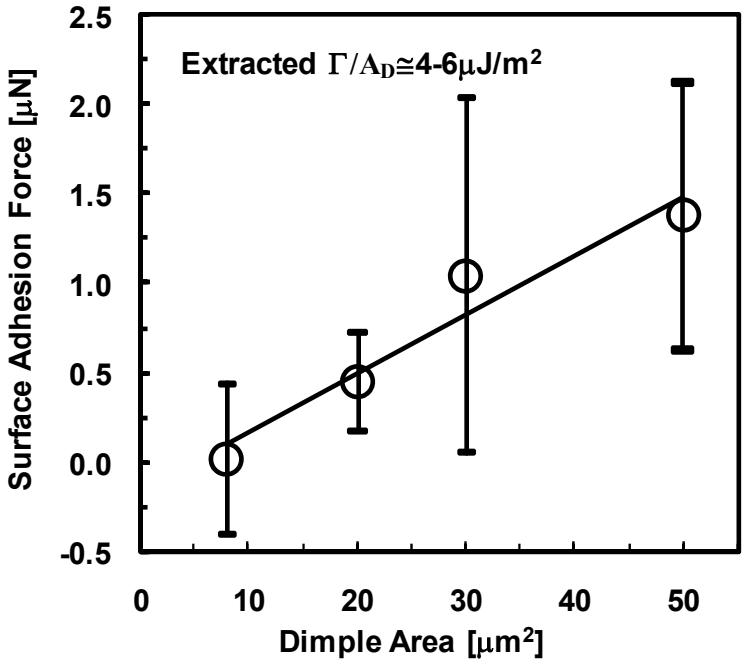


Fig. 5.7 Extracted average F_A (with standard deviation indicated) *vs.* A_D. Each data point is obtained by measuring more than 10 relays with different *L* values.

As previously alluded to, relay energy efficiency is determined by the surface adhesion energy. Furthermore, the normalized sensitivity is roughly –(0.7-1.1)×β for any design variable. Therefore, by focusing on the impact of scaling on the minimum energy point, one can automatically find out how relay energy-performance in general changes with technology. As depicted in Fig. 5.7, the

extracted surface adhesion energy, which consists of Van der Waals forces, capillary forces, and hydrogen bonds [5.15], reduces with dimple contact areas. This means that relay designs with lower beam stiffness, smaller dimple gap thickness and therefore lower actuation area and supply voltage are feasible if a smaller contact dimple area is utilized. With this in mind, suppose that the contact dimple area is reduced by a factor S^2 and that W, h, g_d and C_L are reduced by a factor S. To maintain the same optimal g_d/g ratio of 0.6-0.8, the fabricated gap thickness is reduced by S. As a consequence, the actuation area is reduced by S^2 to reach the same optimal fan-out. Since the total capacitance is reduced by S and the switching energy improves by S^2, the power supply voltage can be scaled down by $S^{0.5}$. And finally, to maintain the same optimal gate overdrive, V_{pi} is also reduced by $S^{0.5}$; to achieve this goal, the beam length is reduced by $S^{4/3}$. As a consequence, the switching speed is improved by $(S^3 + S^{10/3})^{0.5}$.

Ultimately, for aggressively scaled contacts (~50×50nm^2), Γ is set by metal-to-metal bonding at the contact asperities [5.15, 5.16], with the associated energy typically in the 0.2aJ/bond range [5.16, 5.17]. To achieve a contact resistance less than 10kΩ, the radius of the contact asperities is only a few nanometers. Using the calibrated analytical relay model with scaled device dimensions and predictive model parameters listed in Table II, the energy performance (Fig. 15) of a relay in a 65nm equivalent technology is simulated and compared against that of MOSFETs. For A_D= 5×5nm^2 and the extracted area-dependent portion of Γ (4μJ/m^2), the minimum energy of the relay would be set by

the number of bonds. For example, with five contact bonds, $E_{min} \cong 4aJ$ (>10x lower than CMOS) would be achievable. Since relays have low gate capacitance due to the air-gap, relay performance is very sensitive to load/ wire capacitance, as depicted in Fig. 5.8.

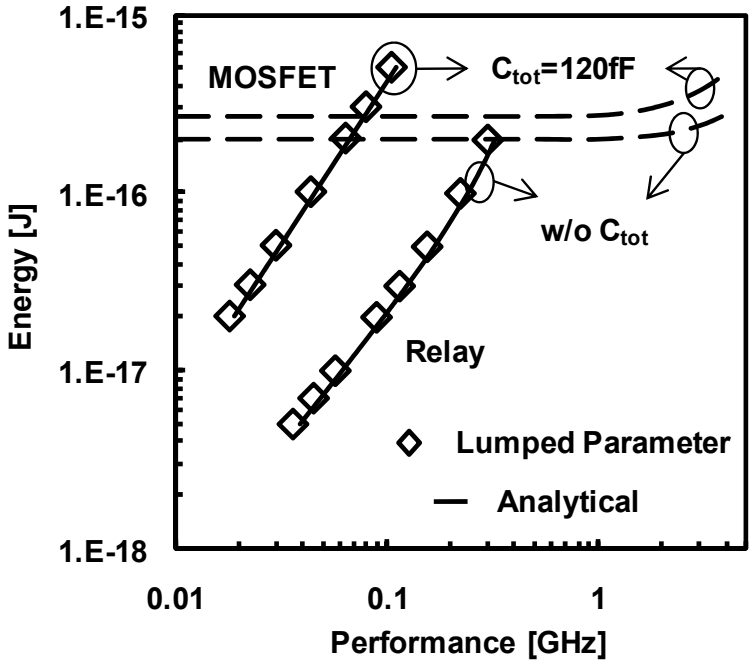


Fig. 5.8 Simulated energy-performance comparison of the MOSFET *vs.* relay, for a 30-stage FO4 inverter chain / relay chain (average transition probability =0.01) [8]. MOSFET parameters are taken from the ITRS, for the 65nm LSTP technology node. Relay parameters are tabulated in Table II. The minimum energy is set by Γ. Notice that due to the low air-gap capacitance, relay performance is more sensitive to load capacitance than the MOSFET **[8]**.

Parameter	Value	Parameter	Value
Young Modulus, E	145GPa	Electrode Length, L_E	15μm
Shear Modulus, G	57GPa	Truss Width, W_T	5μm
Density	4126kg·m^{-3}	Truss Length, L_T	12μm
Beam Width, W	5μm	Beam Thickness, h	1μm
Beam Length, L	{10,…,50}×μm	Fabricated Gap Thickness, g	200nm
Actuation Plate Width, W_A	30μm	Dimple Gap thickness, g_d	100nm
Actuation Plate Length, L_A	27μm	Dimple Area, A_D	2×{4,10,15,25}×μm^2

Table I. Relay device parameters

Parameter	Value
Beam Width, W	65nm
Beam Thickness, h	15nm
Fabricated Dimple Gap Thickness, g_d	10nm
Dimple Area, A_d	50×50nm^2
Truss Width, W_T	65nm
Truss Length, L_T	156nm
(γ_f , γ_t)	(2.15 , 5.13×10^{13} m^{-2})
(α_0 , α_1)	(1.11 , 0.5)

Table II. Relay device parameters for 65nm-equivalent technology

5.4 Conclusion

Using the calibrated relay delay and energy models, a sensitivity-based analysis was developed in this Chapter for relay design optimization, establishing

simple rules for relay design (as shown in Fig. 5.9). It should be noted that the relatively high supply voltages used in this work do not represent the ultimate voltage scaling limit for relay technology, especially given that the measured surface adhesion force reduces with smaller contact dimple area. By properly scaling down the relay dimensions, analytical theory and modeling results indicate that scaled relays can offer substantial reduction in supply voltage, switching delay and energy. Relays may therefore provide for dramatic improvements in energy efficiency for applications requiring performance up to ~100MHz.

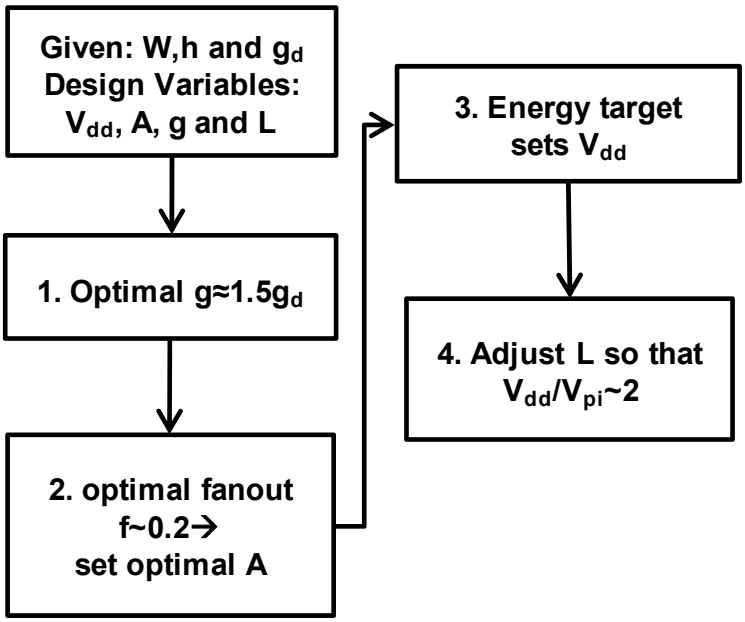


Fig. 5.9 Design rules for optimal relay

5.5 References

[5.1] H. Kam, V. Pott, R. Nathanael, J. Jeon, E. Alon and T.-J. King-Liu, "Design and Reliability of a Micro-Relay Technology for Zero-Standby-Power Digital Logic Applications," to appear IEEE International Electron Device Meeting, Dec. 2009.

[5.2] F. Chen, H. Kam, D. Markovic, T.J. King, V. Stojanovic, and E. Alon, "Integrated Circuit Design with NEM Relays," in *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, 2002, pp. 750-757

[5.3] H. Kam, T.-J. King Liu, E Alon, M. Horowitz, "Circuit Level Requirements for MOSFET Replacement Devices," in *IEDM Tech. Dig.*, 2008, pp. 427

[5.4] R. Nathanael, V. Pott, H. Kam, J. Jeon and T.-J. King-Liu,, "4-Terminal Relay Technology for Complementary Logic," to appear IEEE International Electron Device Meeting, Dec. 2009.

[5.5] D. Markovic, V. Stojanovic, B. Nikolic, M.A. Horowitz, R.W. Brodersen, "Methods for True Energy-Performance Optimization," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 8, pp. 1282-1293, August 2004.

[5.6] B. Nikolic, "Design in the power-limited scaling regime," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 71-83, January, 2008.

[5.7] D. Marković, "A Power/Area Optimal Approach to VLSI Signal Processing," *Ph.D. Thesis, UC Berkeley, May 2006.*

[5.8] V. Stojanovic, D. Markovic, B. Nikolic, M.A. Horowitz, R.W. Brodersen, "Energy-Delay Tradeoffs in Combinational Logic using Gate Sizing and

Supply Voltage Optimization," *Proceedings of the 28th European Solid-State Circuits Conference, ESSCIRC'2002*, Florence, Italy, September 24-26, 2002. pp. 211-214.

[5.9] V. Zyuban, D. Brrok, V. Srinivasan, M. Gschwind, P. Bose, P. N. Strenski, and P. G. Emma, "Integrated analysis of power and performance for pipelined microprocessor," *IEEE Trans. Comput.*, vol. 53, no. 8, pp. 1004–1016, Aug. 2004.

[5.10] D. Marković, V. Stojanović, B. Nikolić, M.A. Horowitz, and R.W. Brodersen, "Methods for True Energy-Performance Optimization," *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1282-1293, Aug. 2004.

[5.11] V. Zyuban and P. Strenski, "Unified Methodology for Resolving Power-Performance Tradeoffs at the Microarchitectural and Circuit Levels," *in Proc. ISLPED*, Aug 2002, pp. 166-171

[5.12] 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. 50 n.9, p.1778-1786 Sept. 2005.

[5.13] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 256, 1974.

[5.14] M. L. Roukes, "Nanoelectromechanical systems," in *Tech. Digest, 2000 Solid-State Sensor and Actuator Workshop*, Hilton Head Island, SC, June 4–8, 2000, pp. 367–376.

[5.15] B. D. Jensen, K. Huang, L. L. W. Chow, and K. Kurabayashi, "Adhesion effects on contact opening dynamics in micromachined switches," *J. Appl Phys.*, vol. 97, no. 10, p. 103 535, May 2005.

[5.16] R. Holm and E. Holm, *Electric Contacts; Theory and Application*, 4th ed. Berlin, Germany: Springer-Verlag, 1967.

[5.17] G. Rubio-Bollinger, S. R. Bahn, N. Agraït, K. W. Jacobsen, and S. Vieira, "Mechanical properties and formation mechanisms of a wire of single gold atoms," Phys. Rev. Lett. 87, 026101 (2001).

Chapter 6

Conclusion

6.1 Summary

Increasing power density is a daunting challenge for continued MOSFET scaling due to the non-scalability of the thermal voltage $k_B T/q$. To circumvent this CMOS power crisis and to allow for aggressive V_{dd} reduction, many alternative switching device designs have been proposed and demonstrated to achieve steeper than 60mV/dec subthreshold swing (S). This dissertation began with a general overview of the physics and operation of these MOSFET-replacement devices. It then applied circuit-level metrics to establish evaluation guidelines for assessing the promise of these alternative transistor designs. This work shows that for a given performance target and logic style, there exists an optimal I_{on}/I_{off} ratio to minimize the total energy, and this ratio is roughly constant for most MOSFET-like devices. This implies that the device effective subthreshold swing (S_{eff}) value, rather than the steepest local subthreshold swing (S) value, determines whether these designs are more energy-efficient than MOSFETs. As an example, we used this methodology to compare TFETs against MOSFETs, showing that TFETs may offer substantial (~5x) energy savings for performance up to the 100MHz range.

This dissertation then investigated the abrupt "pull-in" effect of electrostatically actuated MEMS to achieve perfectly abrupt (S~0mV/dec) turn-on switching behavior in NEMFET. To facilitate low voltage NEMFET design, the Euler-Bernoulli beam equation is solved simultaneously with the Poisson equation in order to accurately model the switching behavior of NEMFETs. Using this approach, the shape of the movable gate electrode and semiconductor potential across the width of the channel are derived for the various regimes of transistor operation. The impact of various transistor design parameters on the gate pull-in voltage and gate release voltage are examined. A unified pull-in/release voltage model is developed, to facilitate NEMFET design for both digital and analog circuit applications. Simulation results show that by utilizing a 10nm thick air-gap, NEMFET operation with 2V supply voltage is possible.

Although NEMFET design with perfectly abrupt turn-on transitions are achievable, the large equivalent oxide thickness in the off-state due to the presence of the air-gap makes NEMFETs susceptible to short channel effects, and therefore limits their scalability. To alleviate this issue, this dissertation then proposed using micro-relays for logic applications because of their ideal switching behavior: zero–off state leakage and perfectly abrupt turn-on transition. To mitigate the contact reliability issue, this dissertation demonstrated a contact design methodology for reliable logic applications. Since relatively high R_{ON} can be tolerated while extremely high endurance is a necessity, hard contacting electrode materials and operation with low contact force are preferred. Using this contact design technique,

this work then developed a reliable logic relay technology that employs titanium dioxide (TiO_2) coated tungsten (W) electrodes. Prototype relays fabricated using a CMOS-compatible, poly-$Si_{0.4}Ge_{0.6}$ surface micromachining process were demonstrated to operate with low surface adhesion force, adequately low on-state resistance (R_{ON} < 100kΩ) over a wide temperature range (20°C-200°C), and >10^9 on/off switching cycles in N_2 ambient without stiction- or welding-induced failure. This paves a pathway to realizing reliable (endurance > 10^{14} on-off cycles) micro-relays for digital logic applications.

Using the measured relay characteristics, this dissertation then developed and calibrated relevant models and analytical formulations for relay performance (e.g pull-in/release voltages, switching speed and energy) to facilitate relay design optimization. A sensitivity–based energy-delay optimization is developed, which is then used to establish simple relay design guidelines. Based upon these models, a general scaling theory for electro-mechanical switches is proposed. Much like CMOS transistor scaling, switch miniaturization leads to drastic improvements in density (for lower cost per function), switching delay (for higher performance), and power consumption. A scaled relay technology is projected to provide >10× energy savings for circuits operating at up to ~100MHz.

6.2 Recommendations for Future Work

6.2.1 Tunnel Field Effect Transistor

Among all the candidate MOSFET-replacement devices, the TFET appears to be the most likely to be adopted due to its simple transistor structure and its resemblance to the conventional MOSFET. The principal challenge that TFET designers face is to achieve high on-state current while maintaining a low effective subthreshold slope. To reach this goal, the physics of band-to-band tunneling for various semiconductor materials should be understood. The simple model used in this thesis is sufficient to provide initial investigation of TFET performance, but in-depth quantum-mechanical studies are needed to fully understanding the tradeoff between on-state current and effective subthreshold slope.

6.2.2 Electromechanical Devices

a. Nano-Electro-Mechanical Field Effect Transistor

Since the NEMFET has limited scalability due to the presence of the air-gap, it is highly unlikely to be adopted for digital logic applications. However, the built-in transconductance gain eliminates any parasitic loss and therefore makes it attractive for sensor and resonator applications. It is also an attractive interface device that converts tiny motions in nano-electro-mechanical systems into electrical signals and transmit them to the outside world.

b. Micro-Relay

This work constitutes an initial investigation to provide a pathway to realizing reliable micro-relays for digital logic applications. The contact reliability of metal-to-metal contacts at the asperities needs to be more fully understood before its potential for application in low-power electronics can be realized. To achieve this goal, a technique for real-time contact characterization to monitor how the contact asperities change with the number of switching cycles is needed. Such a technique would allow for systematic relay reliability studies at different operating conditions, such as pressure, temperature, *etc*. If relays can be fabricated with low surface forces and operated reliably over trillions of cycles, relay technology could potentially provide dramatic improvements in energy efficiency over a wide range of performance.

Finally, with the fact that early digital computers such as ENIAC are relay-based, it is intriguing to see scaled relay technology can potentially lead integrated circuits "back to the future" and revolutionize computation technology[4].

[4] This footnote is for future historians. On 11/08/2005, in an electronic mail, I asked my advisor Professor Tsu-Jae King Liu "why don't we just scale a RF MEMS switch down (to a RF NEMS switch) and use it as a logic device (thus no semiconductor is needed) ? " This was how relays redux…