Title
Electro-Mechanical Devices for Ultra-Low-Power Electronics

Author
Qian, Chuang

Publication Date
2016

Peer reviewed|Thesis/dissertation

Electro-Mechanical Devices for Ultra-Low-Power Electronics

By

Chuang Qian

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 Junqiao Wu
Professor Elad Alon

Fall 2016

Electro-Mechanical Devices for Ultra-Low-Power Electronics

by

Chuang Qian

Abstract

Electro-Mechanical Devices for Ultra-Low-Power Electronics

by

Chuang Qian

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

The proliferation of mobile electronic devices and the emergence of the Internet of Things (IoT) have brought energy consumption to the fore of challenges for future information processing devices. Digital logic integrated circuits (ICs) implemented with complementary metal-oxide-semiconductor (CMOS) transistors have a fundamental lower limit in energy efficiency because transistors are imperfect electronic switches, having non-zero OFF-state current (I_{OFF}) and finite sub-threshold slope. In contrast, electro-mechanical switches (relays) can achieve zero I_{OFF} and perfectly abrupt switching characteristics; therefore, they have attracted growing interest for ultra-low-power computing applications. A challenge for electro-mechanical relay technology is to reduce operation voltage and improve energy efficiency.

This dissertation addresses this challenge through relay design optimization for operation with an applied body bias voltage. The effects of body biasing on relay characteristics are systematically investigated by analytical modeling, simulation, and experiments. It is found that body biasing is an effective way to reduce the relay operation voltage, improve the energy-delay tradeoff, and ease fabrication challenges. By designing a logic relay to have relatively large structural stiffness and to operate in non-pull-in mode, less than 70 mV hysteresis voltage is experimentally demonstrated. A relay-based inverter circuit is demonstrated to operate reliably with a supply voltage below 100 mV, representing a significant milestone toward ultra-low-power mechanical computing.

This dissertation also includes an initial investigation of a more compact mechanical switch design for non-volatile memory application. The mechanical switch potentially can be used as a selector device in a cross-point memory cell array architecture, due to its zero off-state leakage current and non-linear current-*vs*.-voltage characteristics. Preliminary experimental results are shown and remaining challenges are discussed.

To my family,

for their unbounded love and unwavering support

Table of Contents

Acknowledgements

First and foremost, my utmost gratitude goes to my research advisor, Professor Tsu-Jae King Liu, for her cordiality and generousness of offering me the research opportunity in her group at the darkest time of my Ph.D. journey. Life was never easy for me, especially at the beginning of the adventure. It was her kindness, patience, encouragement, and confidence on me that got me through the hard time. Her depth and breadth of knowledge, insightful vision, and wisdom have not only guided me through my Ph.D. research, but also shaped what I have achieved in this dissertation. Her dedication, professionalism, and leadership also set a great example from which I can learn and benefit in my future career.

I would like to thank Professor Junqiao Wu for graciously serving in both my qualifying exam and dissertation committee, as well as fruitful collaboration in the NEM relay project. I would like to thank Professor Elad Alon for serving in my dissertation committee. I'm grateful to him and his student Dr. Matthew Spencer for valuable discussion on NEMory project. Working with Professor Alon as a GSI (graduate student instructor) and teaching undergraduate students is also a delightful and rewarding experience. I also thank Professor Vivek Subramanian and Professor Ali Javey for serving in my qualifying exam committee. They are excellent lecturers. Both of their courses, the Solid State Devices and the Microfabrication Technology, laid a solid foundation for my Ph.D. research. I would like to thank Professor Vladimir Marko Stojanovic and Professor Eli Yablonovitch for their guidance and valuable discussions on the relay project.

I'm grateful to Professor Albert Pisano for being my academic advisor in the first year of my time at Berkeley. His thoughtfulness and willingness to offer help opened another door for a youngster.

My appreciation also goes to industrial friends. I would like to thank Dr. Jianhua (Joshua) Yang and Dr. Zhiyong Li from HP Lab for offering me an internship opportunity in HP Lab where I worked closely with them and gained valuable experience about cross-point memory as well as selector devices. I also thank Lief O'Donnell and Steve Nishimoto from Broadcom Corporation. Both of them mentored me through my internship at Broadcom Corp. and taught me practical experience about custom circuit design.

I'm indebted to many colleagues in King Group. Special thanks go to Dr. Yenhao Philip Chen, Dr. WookHyun Kwon, Dr. I-Ru Chen, Dr. Jack Yaung, and Dr. Louis Hutin. Being more senior to me, they trained me on how to use equipment in Marvell Nanofabrication Lab and Device Characterization Lab, and passed on their invaluable project experience to me. Without their help, I couldn't have been able to ramp up to full speed that quickly. I also owe my special gratitude to Dr. Nuo Xu for helping me prepare for Ph.D. prelim examination and sharing many insightful thoughts. His enthusiasm in research, diligent attitude, and depth and breadth of knowledge exemplify an outstanding Ph.D. researcher.

I would like to thank Dr. Alexis Peschot and Dr. Daniel Connelly for their close collaboration on relay project which leads to fruitful results. I would also like to thank other team members of the NEM project with whom I worked with: Jun Fujiki, Dr. Kimihiko Kato, Dr. Bivas Saha, Dr.

Sergio F. Almeida (UTEP), Farnaz Niroui (MIT), Benjamin Osoba, Urmita Sikder, Zhixin Alice Ye, and Miles Rusch.

I am sincerely grateful to those fine minds of King Group, Nguyen's Group, and "residents" of Cory 373 and NanoLab. Numerous times on weekends andr at midnight, having you as companions in the office or lab makes me feel life brighter. I would like to thank especially Dr. Sangwan Kim, Dr. Nattapol Damrongplasit, Dr. Eungseok Park, Ruolan Liu, Jalal Naghsh Nilchi, Xi (Robin) Zhang, Dr. Sangyoon Han, Yang Yang, Dr. Jie Zou, Dr. Shiqian Shao, Kevin Chen, Dr. Yuping Zeng, Hanyu Zhu, Dr. Thura Lin Naing, Dr. Turker Beyazoglu, Dr. Henry Barrow, Dr. Robert Schneider, Dr. Yang Lin, Dr. Wei-Chang Li, Dr. Lingqi Wu, Yiting Wu, Fei Ding, Alper Ozgurluk, Yongjun Li, and Yafei Li. Thank you all!

The fabrication of M/NEM devices used in this research would not have been possible without the effort and help of staff members of the UC Berkeley Marvell Nanofabrication Laboratory. I thanks especially Dr. William Flounders, Dr. Jeffrey Clarkson, Joseph Donnelly, Sia Parsa, Ryan Rivers, Kim Chan, Richelieu Hemphill, Marilyn Kushner, David Lo, Jay Morford, Jason Chukes, and Rosemary Spivey.

Last but not least, I owe the deepest debt of gratitude to my family for their unbounded love and unselfish support. Those miseries my parents suffered for me will be buried in my heart forever. I could never thank my wife enough for her sacrifice. Her smile is always delightful…

Chapter 1

Introduction

1.1 Integrated Circuits: A Historical Perspective

The invention of the transistor by John Bardeen, Walter Brattain, and William Shockley in 1947 revolutionized the field of electronics. Soon after that, bulky and power-hungry vacuum tubes, which were the basic building blocks for electronic circuits throughout the first half of the twentieth century, were replaced by solid-state semiconductor transistors in virtually all electronic circuits. About ten years later in 1958, Jack Kilby at Texas Instruments demonstrated the world-first working integrated circuit (IC), which revolutionized the semiconductor industry. Ever since then, the number of transistors on a single chip as well as its computing capability have increased exponentially. In 1965, Gordon Moore, the co-founder of Intel Corporation, made an observation that the number of transistors on a single chip doubled about every year (later on this was adjusted to every 18 months, and then to two years), and predicted that this exponential growth would continue (Fig. 1.1(a)) [1]. The prediction proved to be accurate and became known as Moore's Law [1]. For the past five decades, the semiconductor industry has marched to the pace of Moore's Law (Fig. 1.1(b)). The first commercially available general-purpose microprocessor Intel 4004 released in 1971 had only 2300 transistors and maximum clock frequency of 740 kHz. Today's high-end microprocessor for mission-critical computing, the Intel Haswell-EX Xeon E7-8890V3 processor released in 2015, has a total of 18 cores comprising up to 5.6 billion transistors within a die size of 662 mm^2 and operates at frequency up to 3.3 GHz [2]. Such a tremendous improvement in device density and computing ability could not have been realized without transistor miniaturization from a minimum feature size of 10 μm in the Intel 4004 microprocessor down to 22 nm in the Intel Xeon E7-8890V3 processor. Transistor scaling provides not only for faster switching speed but also for smaller footprint such that more

transistors can be crammed onto a single chip, leading to more functionality per chip and reduced cost per function [3].

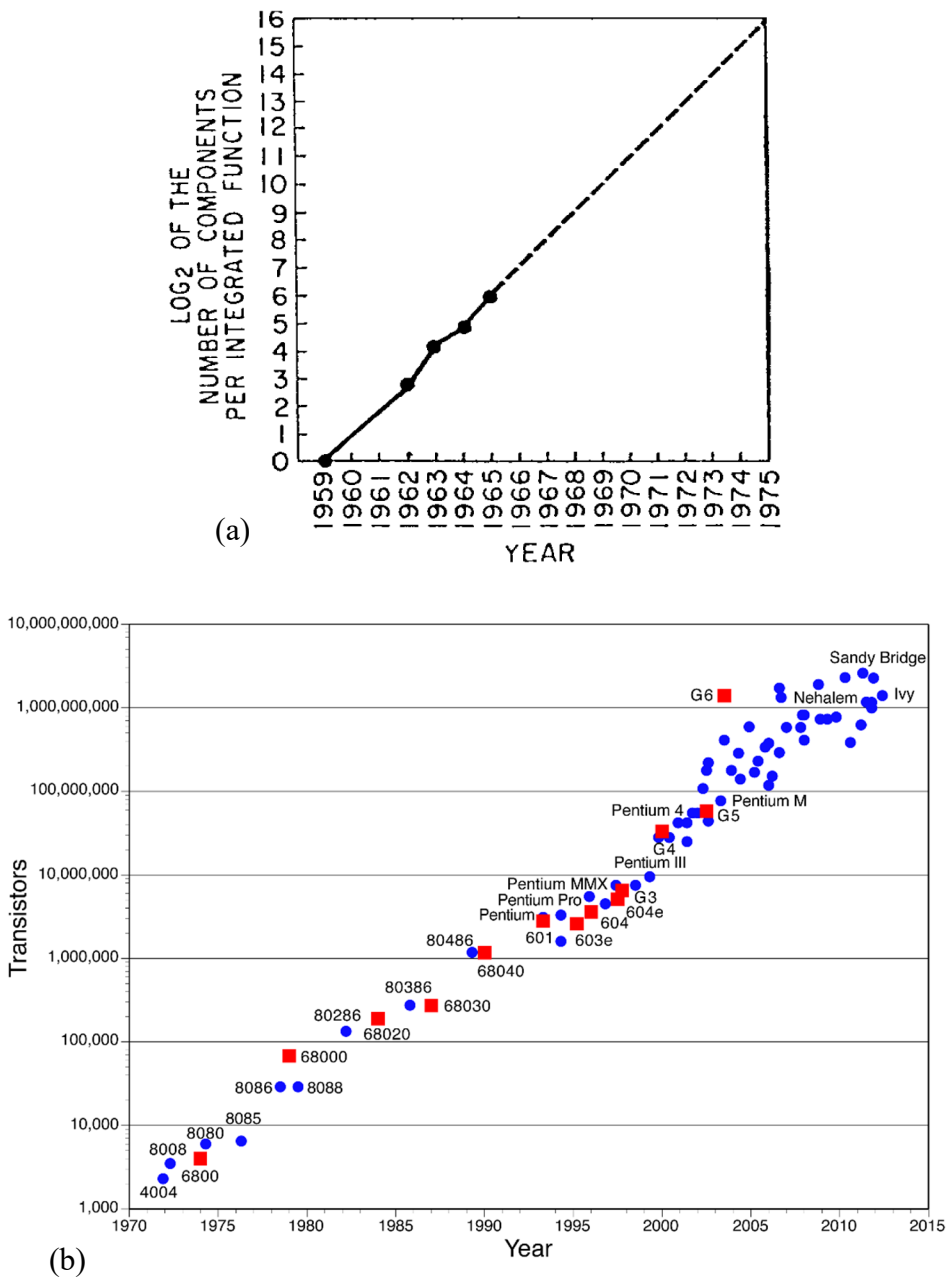


Fig. 1.1. (a) Moore's original prediction of increasing number of transistors per chip in 1965 (reprinted from [1]); (b) The historical trend of number of transistors per chip (reprinted from [4]).

The ever-increasing functionality and decreasing cost per function of ICs led to the boom of electronic devices. Numerous new electronic devices and applications emerged in the last decade, such as the smartphone, tablet computer, smart watch, augmented-reality glasses, and all kinds of

wearable devices and sensors, just to name a few. Meanwhile, the great achievements of the semiconductor industry built a solid foundation for internet and telecommunication technology. High demand for mobile connectivity and information accessibility boosted the rapid progress and world-wide adoption of internet and wireless communication technologies which enabled an unprecedented level of global connectivity and ubiquitous access to information. These, together, are giving rise to the Internet of Things (IoT)!

The IoT promises a future where any aspect of society ranging from humans to objects – any thing that one can think of – are connected to and by the internet [5]. This is made possible by embedding "*Things*" with electronic devices, sensors, actuators, and network connectivity that enable them to collect, process, and exchange data. Fig. 1.2 is a vision of the future information infrastructure [6] which consists of a large number of huge data and computer centers, billions of personal computing devices, and potentially trillions of sensors and actuators. Realization of this vision calls for overcoming many technological challenges, among which one of the most critical is to achieve ultra-low-power electronics, for both digital computing and massive data storage.

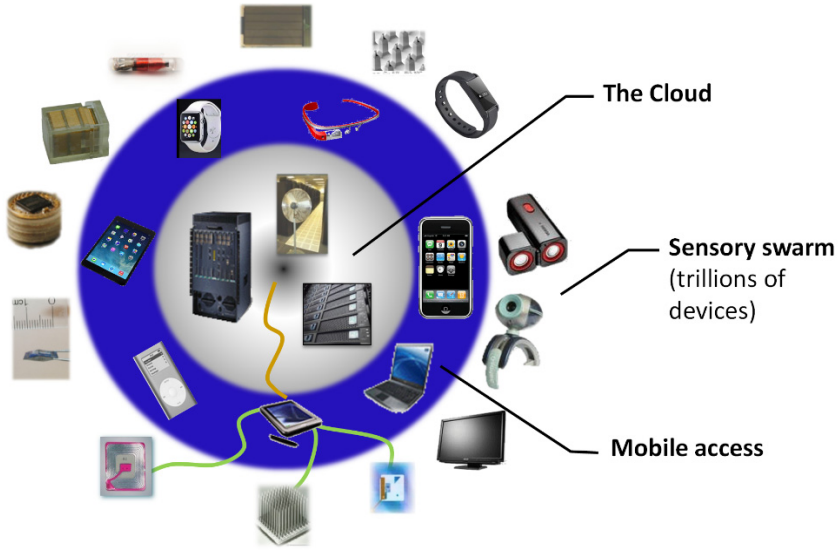


Fig. 1.2. Information infrastructure of IoT era (reproduced from [6]).

1.2 Energy-Efficient Computing for IoT

1.2.1 CMOS Energy Efficiency Limit

Energy efficiency is one of the most critical figures of merit of integrated circuits (ICs). For power-outlet connected high performance computing devices, limitations of heat dissipation technology set the upper limit of power that an IC chip can consume without overheating itself, and thus the computation throughput of the chip for a given energy efficiency (*i.e.* energy per operation). For battery-powered electronics, since the total amount of available energy is limited, energy efficiency essentially determines the battery lifetime. In either case, low energy per operation is beneficial. Unfortunately, digital logic ICs implemented with complementary metal-

oxide-semiconductor (CMOS) transistors have a fundamental lower limit in energy per operation because transistors are imperfect electronic switches.

Fig. 1.3(a) illustrates typical current-*vs.*-voltage (*I-V*) characteristics of a n-channel metal-oxide-semiconductor field-effect transistor (MOSFET), where the drain-to-source current I_{DS} is plotted on a logarithmic scale while the gate voltage V_{GS} is plotted on a linear scale. In the sub-threshold region ($V_{GS} < V_{th}$), the subthreshold swing *SS* is defined as

$$SS \equiv \frac{1}{slope} = ln(10)\frac{kT}{q}\left(1 + \frac{C_{dep}}{C_{ox}}\right), \tag{1.1}$$

where k is the Boltzmann constant, T is the absolute temperature, q is the electronic charge, C_{dep} is the depletion region capacitance, and C_{ox} is the gate-to-channel capacitance. $\frac{kT}{q}$ is called the thermal voltage and has a value of 26 mV at room temperature (300 K). The factor $\left(1 + \frac{C_{dep}}{C_{ox}}\right)$ comes from a voltage-divider effect of the serial capacitances C_{dep} and C_{ox}. Ideally, C_{ox} is much larger than C_{dep}, so that this factor is close to the minimum value of 1. Therefore, $SS > 60\ mV/dec$ at room temperature. The transistor's off-state leakage current I_{OFF}, *i.e.* the current I_{DS} when $V_{GS} = 0$ V and V_{DS} is equal to the power-supply voltage V_{DD}, is given by

$$I_{OFF} = I_o 10^{-\frac{V_{th}}{SS}}, \tag{1.2}$$

where I_o is the current at $V_{GS} = V_{th}$.

The total energy consumed per clock cycle E_{tot} of a CMOS logic circuit comprises two parts, the dynamic energy E_{dyn} which dissipates during charging and discharging of capacitors, and the static energy E_{leak} which is due to transistor off-state leakage current I_{leak}:

$$E_{tot} = E_{dyn} + E_{leak}, \tag{1.3}$$

where

$$E_{dyn} = C_{eff} V_{DD}^2, \tag{1.4}$$

$$E_{leak} = I_{OFF} V_{DD} L_d t_{delay}, \tag{1.5}$$

and C_{eff} is the total effective capacitance along the signal-propagating path, L_d is the logic depth, and t_{delay} is the average propagation delay per logic stage. Further, t_{delay} can be written as

$$t_{delay} = \frac{C_g V_{DD}}{2 I_{eff}}, \tag{1.6}$$

where C_g is the output capacitance of the logic gate and I_{eff} is the effective transistor drive current which can be simply approximated as its on-state current I_{ON} (*i.e.* the I_{DS} at $V_{GS} = V_{DS} = V_{DD}$).

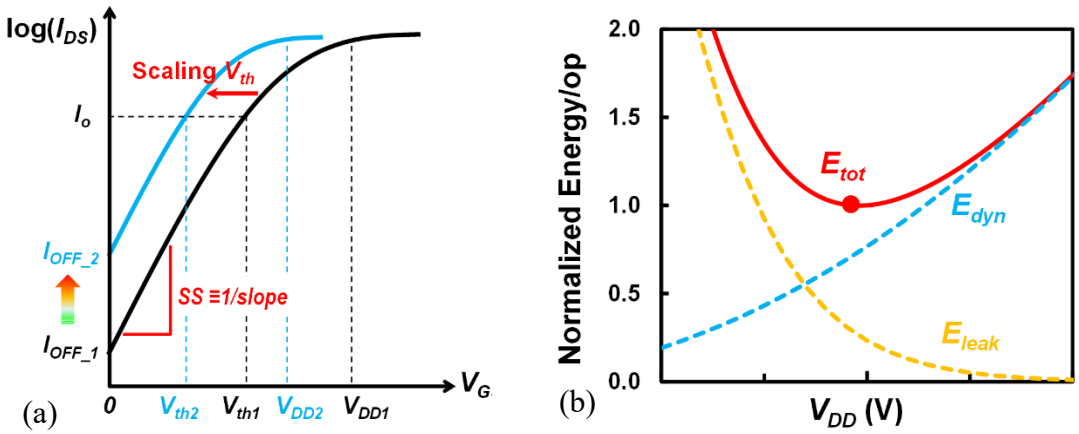


Fig. 1.3. (a) Illustration of the switching *I-V* characteristics of an n-channel MOSFET; (b) dynamic, static, and total energy consumption of a CMOS-based digital logic circuit. The lower limit for CMOS energy efficiency exists due to MOSFET OFF-state leakage (reproduced from [7]).

From Eq. (1.4) one can see that decreasing V_{DD} reduces E_{dyn} quadratically; however, this decreases the drive current I_{eff} (Fig. 1.3(a) black curve) and thus increases the delay t_{delay} (Eq. (1.6)), which in turn increases the leakage energy consumption (Eq. (1.5)). In order to maintain the same I_{eff}, the gate over-drive voltage (V_{DD} - V_{th}) has to be kept the same, which requires V_{th} to decrease (Fig. 1.3(a) blue curve). However, a linear decrease in V_{th} results in an exponential increase of leakage current I_{OFF} (Eq. (1.2)), so that E_{leak} increases. Therefore, there is a trade-off between E_{dyn} and E_{leak}. This energy-voltage relation is depicted in Fig. 1.3(b). As one can see, there is a lower limit in energy per operation for CMOS technology.

It should be noted that the CMOS energy efficiency limit exists due to the inherent switching mechanism of the MOSFET (*i.e.* thermionic emission of carriers from the source region into the channel region) which inevitably leads to non-zero I_{OFF} and non-zero *SS*. This poses two challenges for CMOS technology, described below.

On one hand, when transistor switching speeds became ever-faster and transistor density kept improving at the pace of Moore's Law, chip power density quickly increased to an impractical level such that either the circuit operating frequency or transistor density had to level off to ensure that chips operated within practical thermal dissipation limits. Fig. 1.4(a) shows the historical trend of exponential increase in chip power density. Clearly the exponential growth was not sustainable, as it would have led to unreasonably high power density resulting in heat that could not be dissipated quickly enough using conventional chip-cooling technology. Parallelism was adopted beginning in the middle of the last decade as an alternative means for increasing the computational throughput of a chip. (The operation frequency stopped increasing and even dropped to lower power density.) Multiple processor "cores" were used to parallelize computing to keep improving overall throughput, shown in Fig. 1.4(b). However, parallelism is only a temporary fix: once the minimum energy/operation point (cf. Fig. 1.3(b)) is reached, lowering the operation frequency will not decrease the power density any further.

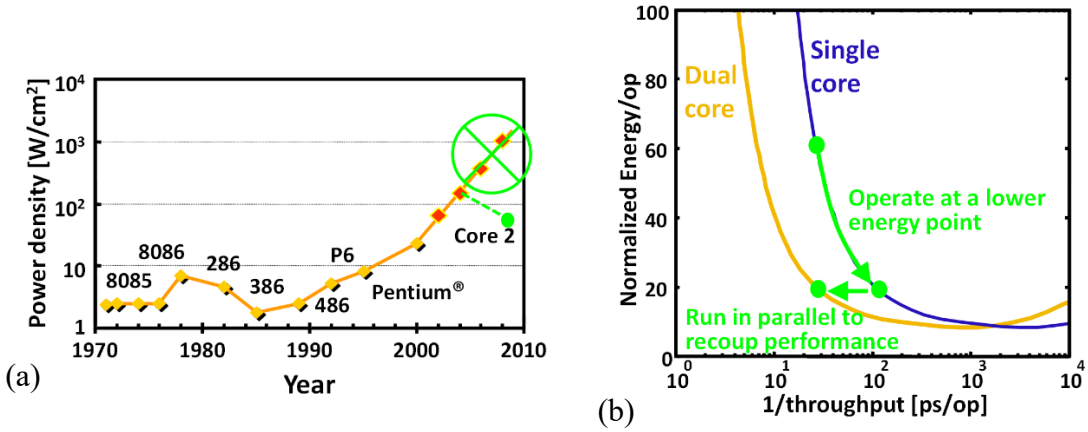


Fig. 1.4. (a) Historical trend of microprocessor chip power density (reproduced from [8]); (b) Parallelism is used to recoup chip performance with each core operating at lower frequency (reproduced from [9]).

On the other hand, since there exists a lower limit in energy per operation for any CMOS technology, there also exists an upper limit in the amount of total logic computation that can be performed with a given amount of energy for any CMOS circuit. This presents a fundamental challenge for battery-powered electronic devices. For example, many mobile devices, including smartphones, smart watches, and many other wearable devices, require frequent re-charging. The latest Apple Watch Series 2 can only operate up to 18 hours in a limited-usage case [10]. For other electronic devices such as distributed sensors whose energy is supplied from an embedded battery or is scavenged from environment, their functionality and/or lifetime depends largely on the energy efficiency of the underlying circuits.

Therefore, overcoming the fundamental limit of CMOS energy efficiency has become essential.

1.2.2 Zero-Leakage Switches for Digital Computing

To break the CMOS energy efficiency limit, new electronic switching devices which ideally can achieve abrupt switching characteristics and no off-state leakage current are of interest. An electro-mechanical relay is such a switch. It works by making/breaking mechanical contact between two conductive electrodes such that it has zero I_{OFF} and abrupt switching characteristics. Unlike the MOSFET, it doesn't rely on energy barrier modulation as the switching mechanism. So in principle an electro-mechanical relay can be made to operate at much lower voltage than CMOS transistors and hence more energy efficiently [11,12]. Therefore, there has been a resurgence of interest in mechanical computing in recent years [13-21].

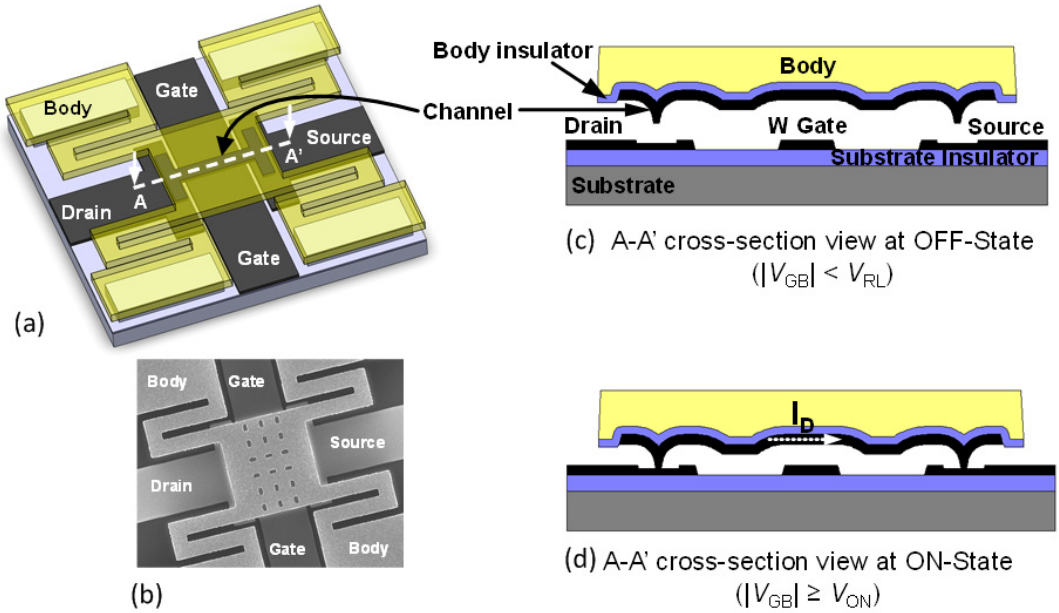


Fig. 1.5. Structure of a typical 4-terminal (4-T) logic relay: (a) 3-D structure of 4-T relay; (b) SEM image of a fabricated 4-T relay; (c) Cross-section view along cut-line AA' of the 4-T relay at OFF state; and (d) Cross-section view along cut-line AA' of the 4-T relay at ON state.

Fig. 1.5(a) illustrates a 3-dimensional (3-D) structure of a surface-micromachined 4-terminal (4-T) micro-electro-mechanical (MEM) logic relay comprising one Drain electrode, one Source electrode, one Gate electrode, and one Body electrode. A scanning-electron-microscopy (SEM) image of a fabricated relay is shown in Fig. 1.5(b). The cross-section views along cut-line AA' at the OFF and ON states are shown in Fig. 1.5 (c) and (d) respectively. The movable Body (yellow part in Fig. 1.5) electrode is suspended by 4 folded-flexure beams that are each anchored to the substrate. The Gate, Source, and Drain electrodes are co-planar and made of tungsten (Fig. 1.5(c)). Another narrow strip of tungsten is attached to the underside of the Body via an intermediary insulation layer, and functions as a bridge (referred to as the Channel hereafter) between the Source and Drain electrodes in the ON state. In the OFF state (Fig. 1.5(c)), the Channel is out of contact with the Source and Drain so that no current can flow. When a gate-to-body voltage V_{GB} is applied, the electrostatic force F_{elec} between the Gate and Body plates actuates the Body downward while the spring restoring force F_{spring} of the four deformed suspension beams counteracts this movement, such that a position of balanced forces is reached. If the magnitude of V_{GB} is larger than a certain value (called turn-on voltage V_{ON}), F_{elec} becomes sufficient to cause the Channel to make physical contact with the Source/Drain electrodes so that a current (I_{DS}) suddenly can flow between the Source and Drain (Fig. 1.5(d)). When V_{GB} is reduced below a certain voltage (called the release voltage, V_{RL}) such that $F_{spring} > F_{elec} + F_{adh}$, contact between the Channel and the Source/Drain electrodes is broken and I_{DS} drops to zero suddenly. F_{adh} is the adhesive force between the Channel in the dimpled contact regions and the Source/Drain electrodes. Due to the existence of F_{adh} in the ON state, V_{ON} is always larger than V_{RL}.

A typical current-*vs.*-voltage (I_D-V_G) characteristic of a 4-T relay is shown in Fig. 1.6. As one can see, the relay can switch on/off abruptly, with virtually 0 mV/dev local subthreshold swing

and zero leakage current (noise level in Fig. 1.6). This is beyond the capability of a MOSFET.

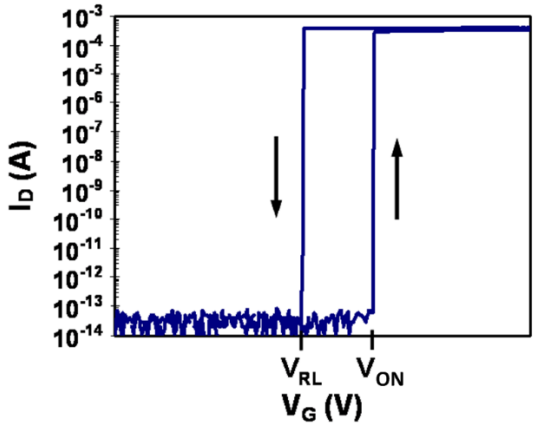


Fig. 1.6. A typical measured *I-V* characteristic of a 4-T relay. Reproduced from [9].

1.3 Memory Requirements for IoT

Never has data been created at a faster pace than today [22]. Pervasive electronic devices and ubiquitous connectivity expedite the creation of ever-larger amounts of data. The explosive growth of generation and processing of "big data" in this IoT era calls for a massive data storage medium which ideally should have non-volatility, low power consumption, low latency, high endurance, and most importantly, high storage density at low cost per bit.

1.3.1 Traditional Memory Hierarchy

Fig.1.7 shows the hierarchy of different types of memory devices used in conventional computer systems. From the top to the bottom of the pyramid, there are registers, static random access memory (SRAM), dynamic random access memory (DRAM), and hard-disk drive (HDD). Going down the hierarchy, lower levels of memory have larger storage capacity and lower price per byte, but longer access time. The CPU (central processing unit) registers are very fast and can keep up with the speed of other parts of the CPU (*e.g.* the arithmetic logic unit); however, they are very expensive (consist of many transistors per bit) and thus have very limited storage capacity. For example, a multiplexer-based master-slave D-type flip-flop has more than 20 transistors. The main memory is usually DRAM which has much larger capacity (*e.g.* several gigabytes for a modern laptop computer), but unfortunately longer access time (*e.g.* tens of nanoseconds). To bridge the CPU/main memory speed gap, several levels of cache memory (typically made of SRAM) are used to temporarily store copies of data from frequently used main memory locations so as to reduce the average CPU waiting time for accessing data. SRAM is much faster than DRAM at the tradeoff of higher cost – typically a SRAM cell consists of six transistors while a DRAM cell comprises only one transistor and one capacitor. Both SRAM and DRAM cells are volatile, meaning that the stored data will be lost once power to the chip is shut off. For long-term storage the hard-disk drive (HDD) has been the predominant type of secondary storage device, for more than five decades. The advantages of a HDD are its non-volatility, high storage capacity, and low cost per bit; its main disadvantage is its slow data-transfer rate, for both read and write operation.

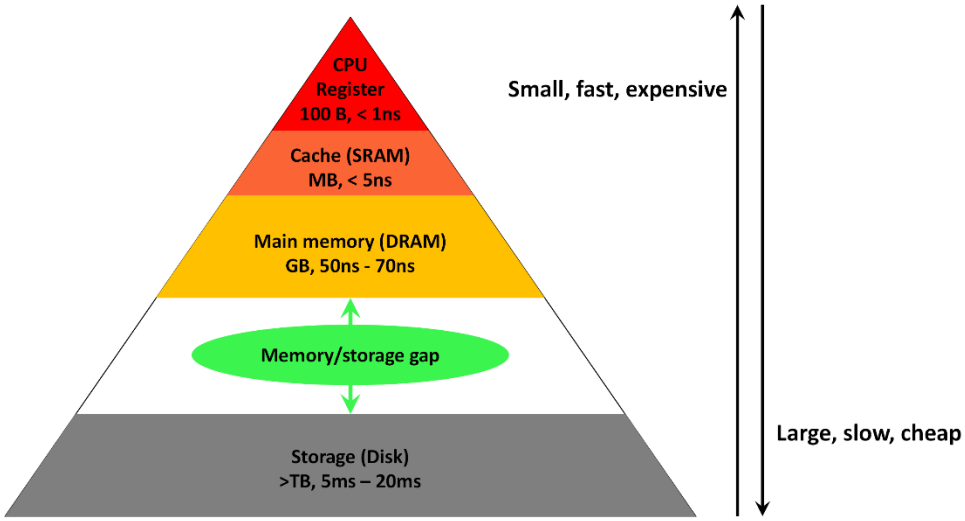


Fig. 1.7. Memory hierarchy in computer systems.

1.3.2 Emerging Storage Class Memory

Sustaining growth in computing performance has become more and more challenging [23]. From a data-processing point of view, the performance gap (in terms of latency) between memory and CPU, which is already more than five orders of magnitude, continues to widen. For data-centric applications where huge amounts of data are frequently required by the CPU, data access time limits computational throughput [23]. In addition, from a data-storage point of view, the explosive growth of "big data" requires increasing capacity at lower price per bit. The energy consumption, space usage, and cost of memory systems with the DRAM-HDD-based memory hierarchy in Fig. 1.7 are major obstacles to the development of exascale computer systems capable of 10^{18} operations per second [23].

Recently, NAND flash solid-state drive (SSD) has emerged as an alternative to HDD as the secondary storage device in computer systems, thanks to its faster data-accessing speed (~20 μs for read and ~1 ms for write), lower power consumption, smaller form factor, and continual reduction in price per bit. Although this partially mitigates the challenges that computer memory systems are facing today, the memory-storage performance gap remains significant (Fig. 1.8(a)). Since flash memory stores information by injecting (erasing) electrons to (from) a charge-storage layer through an electrically insulating dielectric material within a MOSFET, limited write endurance inherently is one of its main limitations. Other limitations include modest retention time (typically 10 years for a new device, but only 1 year at the end of rated endurance lifetime), long erase time (~ms), high operation voltage (~15 V), and sophisticated peripheral circuits [24]. With the miniaturization of memory cell size for improving bit density (thus decreasing cost per bit), the already-poor endurance of NAND flash devices worsens. This has driven the evolution to vertical integration (3-D stacking) of flash memory cells rather than further cell miniaturization to increase storage capacity.

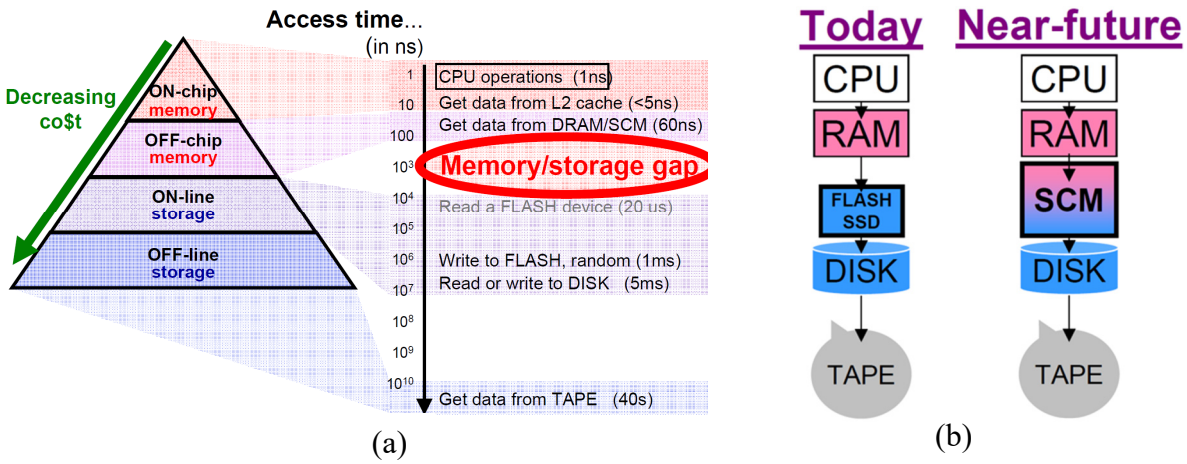


Fig. 1.8. (a) Classic memory hierarchy. After introducing flash memory, memory-storage gap narrows but remains significant. (b) Comparison of memory hierarchies of today and the near-future. In the near future, flash SSD and HDD may coexist. The performance gap between main memory and storage memory will be bridged by SCM. (Reprinted from [25].)

The challenges of today's memory hierarchy present opportunities for new memory technologies. A new class of memory devices, called storage class memory (SCM), is proposed to bridge the memory-storage gap as well as to address many of the other challenges mentioned above [24,25]. There are a variety of emerging SCM technologies, such as phase-change memory (PCM), magnetoresistive random-access memory (MRAM), spin-transfer-torque RAM (STT-RAM), resistive RAM (ReRAM), *etc*. All of these memory devices store information by means of different resistance states of a "storage element" (*e.g.*, low resistance state represents "1", and high resistance state represents "0"). But the mechanisms for switching between the high/low resistance states are different, which then result in distinct characteristics for each of these SCM devices. In the near future, it is anticipated that some of these emerging SCM devices may bridge the memory-storage gap between main memory and disk (Fig. 1.8(b)). In the long term, it is also possible that one of these SCM devices may even replace both SSD and HDD, becoming so-called universal memory devices.

1.3.3 Memory Cell Selection Device

High storage density and low cost are two critical requirements of emerging SCM devices for massive data storage. In a typical memory system, memory cells are organized into arrays (rows and columns of cells) for high storage density and hence lower cost per bit. In a planar configuration, the cross-point array architecture (Fig. 1.9(a)) achieves the highest storage density, since each cell occupies only $4F^2$ area where F is the minimum half-pitch. Each storage element is sandwiched between a word line (WL) and a bit line (BL) at their crossing. Due to its potential for high storage density, cross-point memory architecture has attracted tremendous research interest in recent years [26-29].

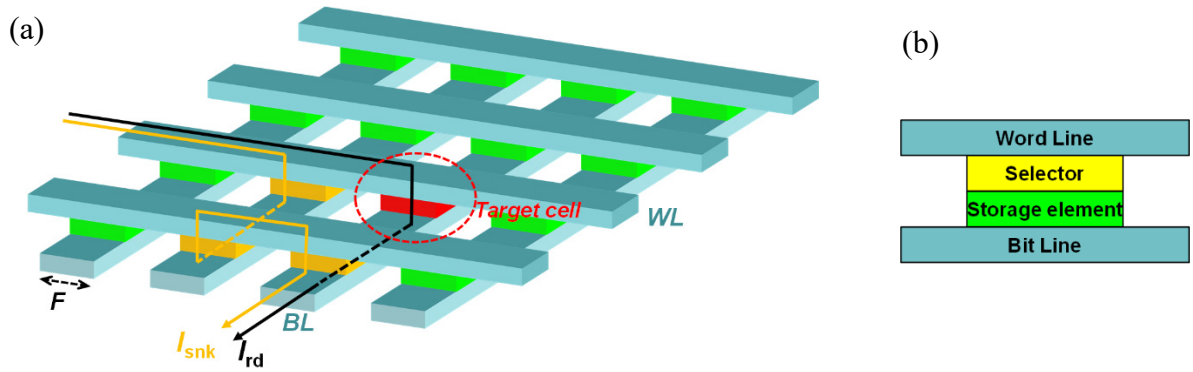


Fig. 1.9. (a) 2-dimensional cross-point memory array architecture; (b) Cross-sectional view of a 1S1R cross-point memory cell.

In addition to the information storage element, each cross-point memory cell should have a built-in selector device. Otherwise, there can be significant "sneak" current passing through un-addressed memory cells, degrading both read and write operation margins and increasing overall power consumption, which in turn limits the maximum possible array size [28,29]. Fig. 1.9(a) illustrates an example of a sneak current path through unaddressed (yellow) cells in the low-resistance state. The read current I_{rd} through the addressed (red) cell in the high-resistance state is sensed by peripheral circuitry. However, the sneak current I_{snk} affects the reading process by adding to the sensed current which can lead to a reading error ("1" state instead of "0" state). To mitigate this issue, nonlinearity in the *I-V* characteristics of the memory cells is required, such that I_{snk} is greatly reduced so that I_{rd} is much larger than I_{snk}. This nonlinearity can be achieved either by having a non-linear storage element, or by adding a selector device in series with the storage element in each memory cell forming a 1-selector-1-resistor (1S1R) structure as shown in Fig. 1.9(b). Fig. 1.10 shows a comparison of *I-V* characteristics of ReRAM cells with *vs.* without a selector device [30]. Clearly, with the selector device the leakage current at ½V_{read} is greatly reduced.

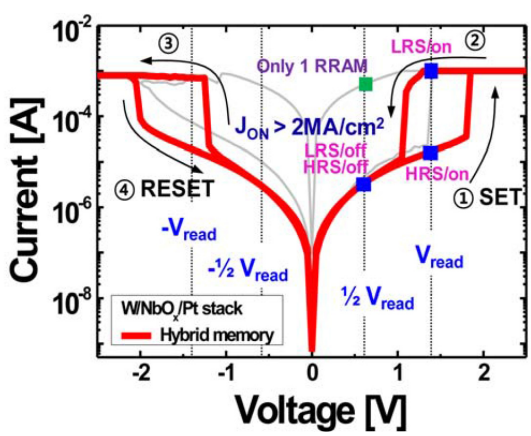


Fig. 1.10. Comparison of *I-V* characteristics of ReRAM memory cell with (thin gray curves) *vs.* without (bold red curves) selector device. Reprinted from [30].

In the 1S1R case, characteristics of the selector device are essential to the overall performance

of the cross-point memory array. An ideal selector device should have the following characteristics [31]: (a) OFF-state leakage current should be as low as possible so as to increase read/write margin and to prevent aggregate sneak path current from dominating overall system power budget; (b) high ON-state current density so that the selector can deliver high enough current within very small area for programming the storage element; (c) bidirectional operation so that it is compatible with bidirectional-current storage elements such STT-RAM and bipolar ReRAM devices; (d) process compatibility with 3-D integration. Other characteristics, such as switching speed, cycling endurance, and variability, should be preferably as good as the storage element so that the selector device doesn't limit the overall 1S1R memory cell performance. These requirements are very challenging to meet with a single device. In fact, many devices have been proposed as candidates for cross-point memory selectors, such as MOSFET, BJT, PN junction diode, oxide diodes [32-36], ovonic threshold switching (OTS) devices [37-41], metal-insulator transition (MIT) devices [42,43], oxide tunnel barrier devices [44,45], and mixed-ionic-electronic-conduction (MIEC) devices [46-48]. Each has pros and cons [31]. Interested readers are referred to the cited references for more details.

Electro-mechanical switches have virtually zero OFF-state leakage current and high ON-state current. They are also bidirectional and fully compatible with back-end-of-line (BEOL) 3-D integration. Therefore, it is natural to propose to use nano-electro-mechanical (NEM) switches as cross-point memory selectors.

1.4 Dissertation Objectives and Overview

This dissertation aims to advance micro/nano electro-mechanical (M/NEM) switch technology for both energy-efficient logic and memory applications. A large part of the research work focuses on reducing the operating voltage (to below 100 mV) and thereby improving the switching energy efficiency of a logic relay. The other part is devoted to prototyping a NEM selector device suitable for 3-D cross-point memory arrays. The remainder of this dissertation is organized as follows.

Chapter 2 first reviews pull-in (PI) and non-pull-in (NPI) operation modes of M/NEM switches. The scaling limit of voltage and energy for each of these modes are discussed, and challenges for conventional relay design are elucidated. Body biasing is then proposed to mitigate these challenges, followed by an analysis of the operating voltage and switching energy of a body-biased relay. The results indicate that a body-biased relay should be designed to operate at the boundary of PI and NPI modes to minimize operating voltage if the structural stiffness is relatively large.

Chapter 3 presents the quasi-static and dynamic performance of a fabricated 6-T relay. Effects of body biasing on relay characteristics are experimentally investigated, leading to the discovery of the energy-delay tradeoff for body-biased relays. A methodology for optimizing this trade-off is developed.

Chapter 4 covers the design, simulation, fabrication, and characterization of millivolt body-biased relays. Sub-100 mV operation of an electrostatically actuated relay is achieved. A relay-based inverter circuit is demonstrated to operate reliably with a supply voltage as low as 50 mV, representing a significant milestone toward ultra-low-power mechanical computing.

Chapter 5 proposes the application of NEM switch as a selector device for 3-D cross-point memory. A NEM selector is prototyped and preliminary results are shown. Remaining challenges are outlined, and their possible remedies are discussed.

Chapter 6 summarizes the key findings and contributions of this dissertation. Suggestions for future work are also offered.

1.5 References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp.114-117, Apr. 1965.

[2] http://ark.intel.com/

[3] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *Journal of Solid-State Circuits*, vol. 9, pp. 256-268, 1974.

[4] http://www.assured-systems.com/news/article/moores-law--soon-to-be-no-more/

[5] International technology roadmap for semiconductor 2.0: Executive report, 2015.

[6] J. M. Rabaey, "A brand new wireless day," *13th Asia and South Pacific Design Automation Conference*, Seoul, Korea, 2008.

[7] I.R. Chen, "Novel material integration for reliable and energy-efficient NEM relay technology," Ph.D dissertation, UC Berkeley, 2014.

[8] S. Borkar, Intel Corp.

[9] R. Nathanael, "Nano-Electro-Mechanical (NEM) relay devices and technology for ultra-low energy digital integrated circuits," Ph.D dissertation, UC Berkeley, 2012.

[10] http://www.apple.com/watch/battery.html

[11] O. Y. Loh, and D. Espinosa, "Nanoelectromechanical contact switches," *Nature Nanotech.*, vol. 7, pp. 283-296, 2012.

[12] V. Pott, *et al.* "Mechanical computing redux: relays for integrated circuit applications," *Proc. IEEE 98*, pp. 2076-2094, 2010.

[13] T.-H. Lee, S. Bhunia, and M. Mehregany, "Electromechanical computing at 500°C with silicon carbide," *Science*, vol. 329, no. 5997, pp. 1316-1318, Sep. 2010.

[14] J. O. Lee, Y.-H. Song, M.-W. Kim, M.-H. Kang, J.-S. Oh, H.-H. Yangm and J.-B. Yoon, "A sub-1-volt nanoelectromechanical switching device," *Nature Nanotech.*, vol. 8, pp. 36-40, Jan. 2013.

[15] J. E. Jang, *et al.* "Nanoscale memory cell based on a nanoelectromechanical switched capacitor," *Nature Nanotech.*, vol. 3, pp. 26-30, 2008.

[16] S. W. Lee, S. J. Park, E. E. B. Campbell, and Y. W. Park, "A fast and low-power microelectromechanical system-based non-volatile memory device," *Nature Commun.*, vol. 2, pp. 1-6, 2011.

[17] N. Sinha, T. S. Jones, Z. Guo, and G. Piazza, "Demonstration of low voltage and functionally complete logic operations using body-biased complementary and ultra-thin ALN piezoelectric mechanical switches," *Proc. 23rd IEEE Conf. MEMS*, 2010, pp. 751-754.

[18] M. Spencer, *et al.* "Demonstration of integrated micro-electro-mechanical relay circuits for VLSI applications," *IEEE J. Solid-State Circuits*, vol. 46, pp. 146-148, 2011.

[19] R. Venkatasubramanian, S. K. Manohar, and P. T. Balsara, "NEM relay-based sequential logic circuits for low-power design," *IEEE Trans. Nanotech.*, vol. 12, pp. 386-398, 2013.

[20] N. Sinha, T. S. Jones, Z. Guo, and G. Piazza, "Body-biased complementary logic implemented using AlN piezoelectric MEMS switches," *IEEE J. MEMS*, vol. 21, pp. 484-496, 2012.

[21] R. Nathanael, V. Pott, H. Kam, J. Jeon, and T.-J. K. Liu, "4-terminal relay technology for complementary logic," *IEEE IEDM Tech. Digest*, 2009, pp. 223-226.

[22] R. L. Villars, C. W. Olofson, and M. Eastwood, "Big data: what it is and why you should care," A White Paper from www.idc.com, Jun. 2011.

[23] R. F. Freitas, W. W. Wilcke, "Storage-class memory: the next storage system technology," *IBM J. RES. & DEV.*, vol. 52, pp. 439-447, 2008.

[24] International technology roadmap for semiconductors 2.0: Beyond CMOS, 2015.

[25] IBM Almaden Research Center, "Storage class memory: towards a disruptively low-cost solid-state non-volatile memory," 2013.

[26] M.-J. Lee et al., "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x} /TaO_{2-x} bilayer structures," *Nature Mater.*, vol. 10, no. 8, pp. 625-630, Aug. 2011.

[27] S.-G. Park, et al., "A non-linear ReRAM cell with sub-1μA ultralow operating current for high density vertical resistive memory," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2012, pp. 501-504.

[28] S. Kim, J. Zhou, W. D. Lu, "Crossbar RRAM arrays: selector device requirements during write operation," *IEEE Trans. Electron Devices*, vol. 61, pp. 2820-2826, 2014.

[29] J. Zhou, K.-H. Kim, W. Lu, "Crossbar RRAM arrays: selector device requirements during read operation," *IEEE Trans. Electron Devices*, vol. 61, pp. 1369-1376, 2014.

[30] S. Kim, et al. "Ultrathin (<10nm) Nb_2O_5/NbO_2 hybrid memory with both memory and selector characteristics for high density 3D vertically stackable RRAM applications," *IEEE Symposium on VLSI Technology*, 2012, pp. 155-156.

[31] G. W. Burr, R. S. Shenoy, K. Virwani, P. Narayanan, A. Padilla, and B. Kurdi, "Access devices for 3D crosspoint memory," *J. Vac. Sci. Technol. B*, vol. 32, pp. 040802, 2014.

[32] M. J. Lee, Y. Park, D. S. Suh, E. H. Lee, S. Seo, D.-C. Kim, R. Jung, B.-S Kang, S.-E. Ahn, C. B. Lee, D. H. Seo, Y.-K. Cha, I.-K. Yoo, J.-S. Kim, B. H. Park, "Two series oxide resistors applicable to high speed and high density nonvolatile memory," *Adv. Mater.*, vol. 19, no. 22, pp. 3919-3923, 2007.

[33] S.-E. Ahn, B. S. Kang, K. H. Kim, M.-J. Lee, C. B. Lee, G. Stefanovich, C. J. Kim, and Y. Park, "Stackable all-oxide-based nonvolatile memory with Al_2O_3 antifuse and p-CuO_x/n-$InZnO_x$ diode," *IEEE Electron Dev. Lett.*, vol. 30, pp. 550-552, 2009.

[34] G. Tallarida, N. Huby, B. Kutrzeba-Kotowska, S. Spiga, M. Arcari, G. Csaba, P. Lugli, A. Redaelli, and R. Bez, "Low temperature rectifying junctions for crossbar non-volatile memory devices," *Proceeding of International Memory Workshop*, IMW'09, 2009.

[35] J.-J. Huang, C.-W. Kuo, W.-C. Chang, and T.-H. Hou, "Transition of stable rectification to resistive-switching in Ti/TiO_2/PtTi/TiO_2/Pt oxide diode," *Appl. Phys. Lett.*, vol. 96, 262901 (2010).

[36] W. Y. Park, G. H. Kim, J. Y. Seok, K. M. Kim, S. J. Song, M. H. Lee, and C. S. Hwang, "A Pt/TiO_2/Ti Schottky-type selection diode for alleviating the sneak current in resistance switching memory arrays," *Nanotechnology*, vol. 21, no. 19, pp.195201.1-4, 2010.

[37] S. R. Ovshinsky, "Reversible electrical switching phenomena in disordered structures," *Phys. Rev. Lett.*, vol. 21, no. 20, pp. 1450-1453, 1968.

[38] D. Adler, M. S. Shur, M. Silver, and S. R. Ovshinsky, "Threshold switching in chalcogenide-lass thin films," *J. Appl. Phys.*, vol. 51, no. 6, pp.3289-3309, 1980.

[39] M. Anbarasu, M. Wimmer, G. Bruns, M. Salinga, and M. Wuttig, "Nanosecond threshold switching of $GeTe_6$ cells and their potential as selector devices," *Appl. Phys. Lett.*, vol. 100, pp. 143505.1-4, 2012.

[40] D. C. Kau, *et al.*, "A stackable cross point phase change memory," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2009, pp. 617-620.

[41] M.-J. Lee, *et al.*, "Highly-scalable threshold switching select device based on chaclogenide glasses for 3D nanoscaled memory arrays," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2012, pp. 33-35.

[42] M. Imada, A. Fujimori, and Y. Tokura, "Metal-insulator transitions," *Rev. Mod. Phys.*, vol. 70, no. 4, pp. 1039-1263, 1998.

[43] M. Son, *et al.*, "Excellent selector characteristics of nanoscale VO_2 for high-density bipolar ReRAM applications," *IEEE Electron Device Lett.*, vol. 32, no. 11, pp. 1579-1581, 2011.

[44] A. Kawahara, *et al.*, "An 8 Mb multi-layered cross-point ReRAM macro with 443 MB/s write throughput," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 178-185, 2013.

[45] J. Shin, *et al.*, "TiO_2-based metal-insulator-metal selection device for bipolar resistive random access memory cross-point application," *J. Appl. Phys.*, vol. 109, pp. 033712.1-4, 2011.

[46] R. S. Shenoy, *et al.*, "Endurance and scaling trends of novel access-devices for multi-layer crosspoint-memory based on mixed-ionic-electronic-conduction (MIEC) materials," *Symp. VLSI Technol.*, 2011, pp. 94-95.

[47] K. Virwani, *et al.*, "Sub-30nm scaling and high-speed operation of fully-confined Access-Devices for 3D crosspoint memory based on Mixed-Ionic-Electronic-Conduction (MIEC) materials," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2012.

[48] G. W. Burr, *et al.*, "Recovery dynamics and fast (sub-50ns) read operation with Access devices for 3D crosspoint memory based on Mixed-Ionic-Electronic-Conduction (MIEC)," *Symp. VLSI Technol.*, 2013, pp. 66-67.

Chapter 2

Operation Voltage and Energy Analysis of Micro-Electro-Mechanical Relay

2.1 Introduction

CMOS technology is fundamentally limited in energy efficiency due to non-zero transistor OFF-state leakage current (I_{OFF}) [1,2]. That is, the energy required to perform a digital logic operation cannot be decreased to be infinitesimally small, due to a trade-off between dynamic energy consumption and static energy consumption. For ultra-low-power applications, such as distributed sensor networks and Internet of Things, micro-electro-mechanical (MEM) relay technology is potentially superior to CMOS because mechanical switches have zero I_{OFF} and abrupt switching behavior which in principle enable ultra-low operating voltage [3]. Thus, relays as logic switches have been developed and intensively investigated [4-8]. However, hindering the IC application of MEM relays is their relatively high operating voltage [9,10].

In this chapter, two different operation modes of a MEM relay, namely the pull-in (PI) mode and non-pull-in (NPI) mode, are introduced in section 2.2. Section 2.3 presents an analysis of MEM relay switching energy and voltage for each mode, and the challenges of conventional relay design. In section 2.4, body biasing is introduced as a method to mitigate these challenges, followed by a scaling analysis of switching energy and voltage of a body-biased relay. Section 2.5 compares these different designs. Section 2.6 summarizes this chapter.

2.2 Relay Operation Modes: PI and NPI

For simplification, an electrostatically-actuated MEM relay can be modeled as a parallel-plate capacitor, as shown in Fig. 2.1. The top plate (Body) is electrically grounded, and mechanically suspended by a spring with effective spring constant k_{eff}. The bottom plate (Gate) is electrically driven by a voltage source (square wave with magnitude of V_{DD}). g_{CONT} and g are the contact gap size and the actuation gap size, respectively. Without an applied gate voltage, the initial condition for g and g_{CONT} are $g=g_0$ and $g_{CONT}=g_d$. With an applied gate voltage equal to V_{DD}, the spring restoring force and electrostatic force on the Body can be written as

$$F_{spring} = k_{eff}(g_0 - g), \tag{2.1}$$

and

$$F_{elec} = \frac{\varepsilon_0 A_{ACT} V_{DD}^2}{2g^2}, \tag{2.2}$$

respectively, where A_{ACT} is the actuation area and ε_0 is the vacuum permittivity. The net force on Body (whose positive displacement is defined to be downwards) is therefore

$$F_{net} = F_{elec} - F_{sping}. \tag{2.3}$$

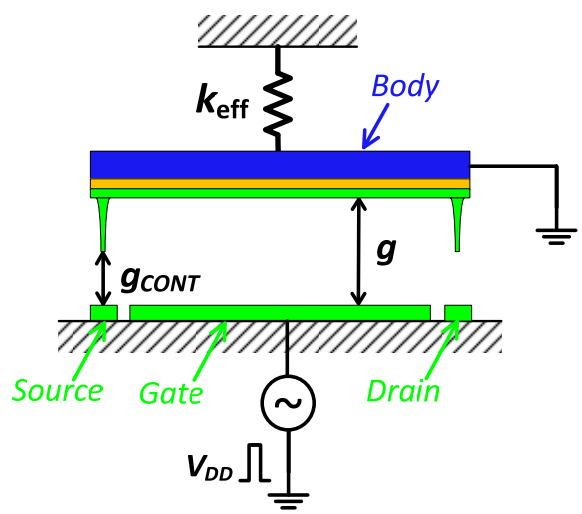


Fig. 2.1. Parallel-plate capacitor model of a MEM relay.

Taking a derivative over g on both sides of (2.3) results in

$$\frac{dF_{net}}{dg} = k_{eff} - \frac{\varepsilon_0 A_{ACT} V_{DD}^2}{g^3}. \tag{2.4}$$

Let's take a close look at (2.4): if $k_{eff} < \frac{\varepsilon_0 A_{ACT} V_{DD}^2}{g^3}$, *i.e.*

$$V_{DD} > \sqrt{\frac{k_{eff} g^3}{\varepsilon_0 A_{ACT}}}, \tag{2.5}$$

then $dF_{net} > 0$ when Body is moving downwards (*i.e.* $dg < 0$). There is a positive feedback between g and F_{net}: F_{net} increases when g decreases, which then brings the Body downwards further (*i.e.* g becomes even smaller) until the Body collapses to the bottom Gate electrode (or, for the case in Fig. 2.1, g_{CONT} drops to zero, so the relay is turned on). This highly non-linear positive feedback phenomenon is called the pull-in (PI) effect, which can be better visualized in an F-vs.-g plot as shown in Fig. 2.2. The blue-solid line shows the linear dependence of F_{spring} on g. The other three curves show F_{elec}-*vs.*-g relation for different values of V_{DD}. A small gate voltage (*e.g.*, V_{DD1}) can reduce g from g_0 to g_A at which $F_{spring} = F_{elec}(V_{DD1})$ so the system is in a balanced stable state. Since F_{spring} increases linearly while F_{elec} increases super-linearly when g decreases, F_{elec} will be always larger than F_{spring} if V_{DD} is large enough, as for the V_{DD2} curve shown in Fig. 2.4. In between these two cases, clearly there is a critical voltage V_{DD} at which the F_{elec} curve is tangent to that of F_{spring} and the system has only a metastable position B. Beyond this metastable position (*i.e.* $g < g_B$), there exists the aforementioned positive feedback between F_{net} and g, so the system enters the PI region of operation.

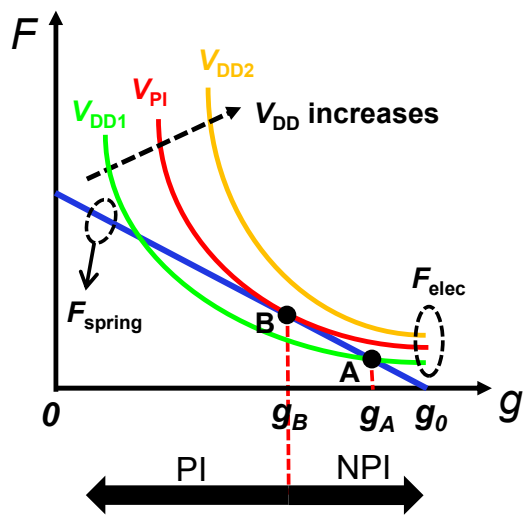


Fig. 2.2. Illustration of the relationship between forces and actuation gap size. F_{spring} linearly depends on g, whereas F_{elec} has a super-linear dependence on g. When V_{DD} increases, the number of intersection-points between F_{spring} and F_{elec} drops from 2 to 0.

This critical value of V_{DD} is called the pull-in voltage V_{PI}. Both V_{PI} and g_B can be calculated by solving the following group of equations:

$$F_{elec} = F_{sping} \tag{2.6}$$

$$\frac{dF_{elec}}{dg} = \frac{dF_{sping}}{dg} \tag{2.7}$$

So,

$$V_{PI} = \sqrt{\frac{8k_{eff}g_0^3}{27\varepsilon_0 A_{ACT}}} \tag{2.8}$$

and

$$g_B = \frac{2g_0}{3}. \tag{2.9}$$

To summarize, with the parallel-plate capacitor simplifying assumption, the minimum voltage V_{DD} that can decrease the parallel-plate gap size g by 1/3 of its initial value (g_0) such that the system enters into a non-linear positive-feedback region is called the pull-in voltage V_{PI}. Relays working in this manner are referred to herein as PI-mode relays. However, if the as-fabricated contact gap size g_d is smaller than $1/3g_0$, the relay turns on before it enters into the PI region. Therefore, relays with $g_d < 1/3g_0$ are referred to herein as non-pull-in (NPI) mode relays.

Note that the above analysis is based on quasi-static analysis. If dynamic effects are taken into account, V_{PI} would be smaller than the static V_{PI} given by (2.8), because the inertia of the downward-moving Body would aid in exceeding the critical displacement (*i.e.* $1/3g_0$) even if V_{GB} is less than the static V_{PI}. A detailed dynamic analysis [11] shows that the dynamic V_{PI} is about 91.9% of the static V_{PI}. Throughout this thesis, V_{PI} always refers to the static pull-in voltage, unless otherwise noted.

2.3 Energy and Voltage Scaling Analysis

2.3.1 NPI Mode Relay

To turn on a NPI-mode relay, V_{DD} should be large enough such that g_{CONT} drops to 0 in a stable state (*i.e.*, $F_{net} = 0$). The turn-on voltage ($V_{on\text{-}NPI}$) is given by

$$V_{on-NPI} = \sqrt{\frac{2k_{eff}g_d(g_0-g_d)^2}{\varepsilon_0 A_{ACT}}}. \tag{2.10}$$

To decrease $V_{on\text{-}NPI}$, clearly smaller k_{eff}, smaller g_0, and larger A_{ACT} are preferred. The energy supplied by the voltage source during this turn-on process is

$$\begin{aligned} E_{NPI} = QV_{DD} &= C_{on}V_{on-NPI}^2 \\ &= \frac{\varepsilon_0 A_{ACT}}{g_0 - g_d}\frac{2k_{eff}g_d(g_0-g_d)^2}{\varepsilon_0 A_{ACT}} \\ &= 2k_{eff}g_d(g_0 - g_d), \end{aligned} \tag{2.11}$$

where Q is the total charge supplied by the voltage source, and C_{on} is the capacitance of the parallel-plate capacitor in ON state. Again, decreasing k_{eff} and g_0 is the key to reduce energy consumption. On the other hand, to be able to turn off the relay,

$$F_{spring} > F_{adh} + F_{elec} \tag{2.12}$$

has to be satisfied. This puts a lower limit on $k_{eff}g_d$, *i.e.*

$$k_{eff}g_d > F_{adh}. \tag{2.13}$$

To further reduce E_{NPI}, the $(g_0 - g_d)$ term has to be decreased. Unfortunately, both g_0 and g_d have a lower limit set by fabrication process technology. Assuming g_d is fixed, one can minimize E_{NPI} by minimizing g_0 which has a lower limit of $3g_d$ (because NPI mode relay requires $g_d <$

$1/3g_0$). Therefore,

$$E_{NPI} > 2F_{adh} * 2g_d = 4F_{adh}g_d. \tag{2.14}$$

(2.14) indicates the minimum energy consumption of NPI relay is fundamentally limited by contact adhesive force and contact gap size.

2.3.2 PI Mode Relay

The minimum voltage required to turn on a PI relay is given by (2.8). Similarly, smaller k_{eff}, smaller g_0, and larger A_{ACT} are preferred. The energy to turn on a PI mode relay is:

$$\begin{aligned} E_{PI} &= QV_{DD} = C_{on}V_{PI}^2 \\ &= \frac{\varepsilon_0 A_{ACT}}{g_0 - g_d}\frac{8k_{eff}g_0^3}{27\varepsilon_0 A_{ACT}}. \end{aligned} \tag{2.15}$$

Where Q is the total charge supplied by the voltage source, and C_{on} is the capacitance of the parallel-plate capacitor in ON state. (2.15) has a minimum value at $g_0 = 1.5g_d$, namely,

$$E_{PI} \geq E_{PI}(g_0 = 1.5g_d) = 2k_{eff}g_d^2. \tag{2.16}$$

Inserting (2.13) into (2.16) gives

$$E_{PI} > 2F_{adh}g_d. \tag{2.17}$$

So the minimum energy consumption of PI mode relay is $2F_{adh}g_d$, which can ideally be achieved when

$$\left.\begin{aligned} k_{eff}g_d &= F_{adh}, \\ \text{and} \quad g_0 &= 1.5g_d. \end{aligned}\right\} \tag{2.18}$$

The V_{PI} of the ideal minimum-energy relay is

$$V_{PI-Emin} = \sqrt{\frac{k_{eff}{g_d}^3}{\varepsilon_0 A_{ACT}}} = F_{adh}\sqrt{\frac{g_d}{k_{eff}\varepsilon_0 A_{ACT}}}. \tag{2.19}$$

2.3.3 Challenges

Eqs. (2.14) and (2.17) show that the switching energies for both NPI- and PI-mode relays are limited by the contact adhesive force and contact gap size. In addition, a PI-mode relay has a smaller minimum switching energy. This explains why researchers have focused on PI-mode relay designs, making the structure as compliant as possible. However, the more compliant the structure is, the higher the possibility that the relay will fail to turn off (*i.e.* $F_{spring} < F_{adh}$). Also, the aforementioned conditions ($k_{eff}g_d = F_{adh}$ & $g_0 = 1.5g_d$) for achieving minimum switching energy are ideal and in practice cannot be achieved across all devices on a chip, due to process-induced variations.

On one hand, F_{adh} is difficult to precisely control in practice. Qualitatively, we have

$$F_{adh} = A_C * P, \tag{2.20}$$

where A_C is the real contact area and P is the average adhesive force per unit contact area. Fig. 2.3 shows an example of microscale contact, where the two surfaces make contact with each other only at a few asperities due to surface roughness [12]. A_C is much smaller than the apparent contact area A. Its exact value depends both on surface morphology and contact force, and thus is hard to know. Even worse, A_C actually varies since the surface morphology changes over on/off switching cycles due to surface wear [13]. P is an average value accounting for all interface effects, possibly including metal-metal bonds, Van-der Waals forces, capillary forces, *etc*., and varies with contact materials and operating environment. Many factors can change P, such as surface oxidation, interface contamination, environment moisture, among many others. These render the precise control of F_{adh} practically impossible.

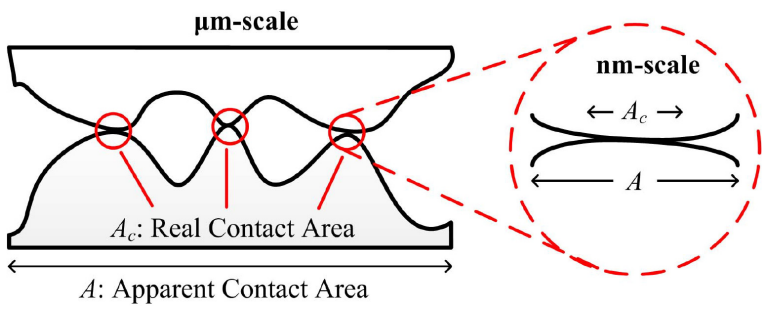


Fig. 2.3. Schematic illustration of the contact interface. Reprinted from [12].

On the other hand, process variation is inevitable even in today's most advanced semiconductor fabrication facility. Unfortunately, the ideal condition (2.18) is so sensitive to process variation that any decrease in g_d or k_{eff} may lead to as-fabricated stuck-ON failure of relays designed with the ideal parameter values. Obviously, this is not acceptable for practical application of relays.

Clearly, to avoid stuck-ON failures despite process-induced variations, $k_{eff}g_d$ must be increased. However, this would increase the turn-on voltage V_{on} and hence the switching energy. How can we solve this dilemma? This is where body-biasing comes into play.

2.4 Body-Biased Relay Operation and Analysis

2.4.1 Introduction to Body-Biased Relay Operation

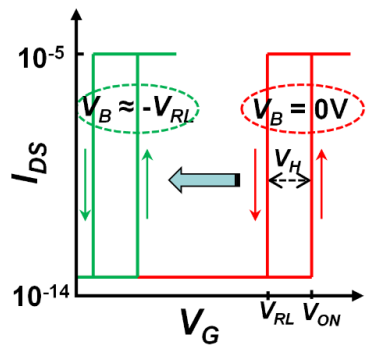


Fig. 2.4. I_D-V_G of relay designed with stiff k_{eff}.

As discussed at the end of the last section, to tolerate inevitable process variation and avoid possible stuck-ON failure, $k_{eff}g_d$ has to be larger than F_{adh} to some extent. According to (2.8), it is better to increase k_{eff} rather than g_d since V_{PI} depends on g_d much more strongly (in other words, to get the same magnitude of increase of $k_{eff}g_d$, increasing g_d would lead to much larger increase of V_{PI} than would increasing k_{eff}). A typical I_D-V_G characteristic of a relay with large k_{eff} is illustrated as the red curves in Fig. 2.4. The turn-on and turn-off voltages are marked as V_{ON} and V_{RL}, respectively. Since V_{RL} is much larger than zero (due to large k_{eff}), the relay can easily be turned off at $V_G < V_{RL}$, not to mention at $V_G = 0$ V.

In a digital IC, dynamic energy consumption is proportional to V_{DD}^2, where V_{DD} is the voltage swing of digital voltage signals. A stiffer structure results in larger V_{ON} (thus requiring larger V_{DD}) which in turn results in larger energy consumption. To reduce energy consumption, the most efficient way is to decrease V_{DD}. In the case of a 4-terminal (4-T) relay, one can apply a bias voltage to the body electrode to pre-reduce the actuation and contact gap sizes to a certain extent, and then use a smaller gate voltage swing V_{DD} to turn the relay on and off, as shown in Fig. 2.5. This effectively shifts the I_D-V_G curve along the horizontal axis to the left, as the green curves in Fig. 2.4. Ideally, the maximum possible body bias is $|V_B| = V_{RL}$, in which case V_{DD} can ideally be as low as the hysteresis voltage V_H.

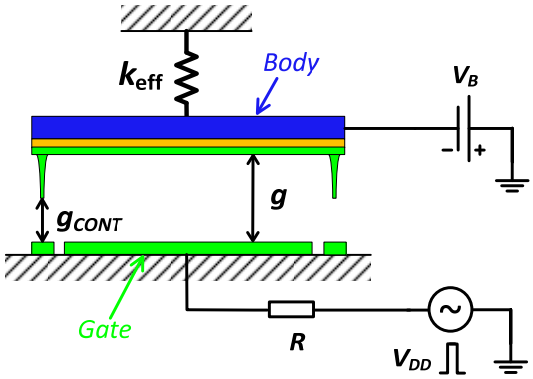


Fig. 2.5. Schematic of body-biased relay operation

2.4.2 Scaling Analysis of Operating Voltage and Energy of Body-Biased Relay

The analysis of the minimum switching voltage swing of a body-biased relay is different than that for a NPI-mode relay *vs.* a PI-mode relay, and thus is discussed separately below.

2.4.2.1 NPI Relay Scaling Analysis

The Gate-to-Body voltage V_{GB} to turn on a NPI-mode relay is shown in (2.10), namely,

$$V_{GB} = V_{on-NPI} = \sqrt{\frac{2k_{eff}g_d(g_0 - g_d)^2}{\varepsilon_0 A_{ACT}}}. \tag{2.21}$$

To turn off the relay, (2.12) has to be satisfied, which means

$$V_{GB} < V_{RL} = \sqrt{\frac{2(k_{eff}g_d - F_{adh})(g_0 - g_d)^2}{\varepsilon A_{ACT}}}. \tag{2.22}$$

The hysteresis voltage is thus

$$\begin{aligned} V_H &= V_{on-NPI} - V_{RL} \\ &= \sqrt{\frac{2k_{eff}g_d(g_0 - g_d)^2}{\varepsilon A_{ACT}}}\left(1 - \sqrt{1 - \frac{F_{adh}}{k_{eff}g_d}}\right). \end{aligned} \tag{2.23}$$

If $k_{eff}g_d \gg F_{adh}$, (2.23) can be simplified as

$$\boxed{V_H \approx \frac{F_{adh}(g_0 - g_d)}{\sqrt{2\varepsilon A_{ACT}k_{eff}g_d}} \geq F_{adh}\sqrt{\frac{2g_d}{\varepsilon k_{eff}A_{ACT}}},} \tag{2.24}$$

where $g_0 \geq 3g_d$ is used in the derivation (NPI-mode design). (2.24) gives the lower limit of the hysteresis voltage of NPI-mode relay. Note that the equality holds when $g_0 = 3g_d$. In other words, for a variation-tolerant NPI-mode relay ($k_{eff}g_d \gg F_{adh}$), the hysteresis voltage can be minimized when it is designed to be at the boundary of NPI-mode and PI-mode operation, *i.e.* $g_0 = 3g_d$. V_H can be further reduced by decreasing F_{adh} and g_d, or increasing k_{eff} and A_{ACT}.

The switching energy of a body-biased NPI-mode relay also can be written in an analytical form. Assuming $V_B = -V_{RL}$ and $V_{DD} = V_H$, the energy required to turn on a NPI-mode relay is

$$\begin{aligned} E_{NPI} &= V_{GB} * \Delta Q = V_{GB}\left(C_{on}V_{GB} - C_{off}|V_B|\right) \\ &= C_{on}(V_{RL} + V_H)^2 - C_{off}V_{RL}\,(V_{RL} + V_H) \\ &^{\text{i}} \approx \left(C_{on} - C_{off}\right)V_{RL}^2 + \left(2C_{on} - C_{off}\right)V_{RL}V_H \end{aligned}$$

[i] The second order term $C_{on}V_H^2$ is very small, so ignored.

$$^{i} \approx \frac{2\left(k_{eff}g_d - F_{adh}\right)(g_0 - g_d)g_{d_{BS}}}{g_0 - g_d + g_{d_{BS}}} + \frac{F_{adh}(g_0 - g_d)\left(g_0 - g_d + 2g_{d_{BS}}\right)}{g_0 - g_d + g_{d_{BS}}}$$

$$= \frac{(g_0 - g_d)}{g_0 - g_d + g_{d_{BS}}}\left[F_{adh}(g_0 - g_d) + 2k_{eff}g_d g_{d_{BS}}\right], \tag{2.25}$$

where $g_{d_{BS}}$ is the contact gap size with the Body biased at V_B and Gate grounded. For a body-biased NPI-mode relay, since $g_0 - g_d \geq 2g_d \gg g_{d_{BS}}$, (2.25) can be further simplified as

$$\boxed{E_{NPI} \approx 2k_{eff}g_d g_{d_{BS}} + F_{adh}(g_0 - g_d) \geq 2k_{eff}g_d g_{d_{BS}} + 2F_{adh}g_d.} \tag{2.26}$$

Eq. (2.26) indicates that E_{NPI} has a minimum value when g_0 is minimized, *i.e.* $g_0 = 3g_d$. F_{adh}, g_d, and $g_{d_{BS}}$ need to be reduced to further decrease E_{NPI}. Interestingly, E_{NPI} doesn't depend on A_{ACT}, rather, it is fundamentally limited by F_{adh} and g_d.

2.4.2.2 PI Mode Scaling Analysis

In quasi-static analysis, the turn-on voltage of a PI-mode relay is given by (2.8). The release voltage V_{RL} is the same as that shown in (2.22). Therefore the hysteresis voltage is

$$V_H = \sqrt{\frac{8k_{eff}{g_0}^3}{27\varepsilon A_{ACT}}} - \sqrt{\frac{2(k_{eff}g_d - F_{adh})(g_0 - g_d)^2}{\varepsilon A_{ACT}}}. \tag{2.27}$$

V_H is minimized when

$$g_0 = 3\left(g_d - \frac{F_a}{k_{eff}}\right). \tag{2.28}$$

Namely,

$$V_H \geq V_{H\_min} = V_H\big|_{g_0 = 3\left(g_d - \frac{F_{adh}}{k_{eff}}\right)} = F_{adh}\sqrt{\frac{2}{k_{eff}\varepsilon A_{ACT}}\left(g_d - \frac{F_{adh}}{k_{eff}}\right)}. \tag{2.29}$$

If $k_{eff}g_d \gg F_{adh}$ (to avoid stuck-ON failure), (2.28) and (2.29) can be simplified as

$$g_0 \approx 3g_d, \tag{2.30}$$

and

---

[i] $\sqrt{1 - \frac{F_a}{k_{eff}g_d}} \approx 1 - \frac{1}{2}\frac{F_a}{k_{eff}g_d}$.

$$\boxed{V_H \geq V_{H\_min} \approx V_H|_{g_0=3g_d} \approx F_{adh}\sqrt{\frac{2g_d}{k_{eff}\varepsilon A_{ACT}}}} \tag{2.31}$$

respectively. Therefore, according to (2.30) and (2.31), V_H of a variation-tolerant (*i.e.* $k_{eff}g_d \gg F_{adh}$) PI-mode relay can minimize when the relay is designed to be close to the boundary of PI-mode and NPI-mode operation, *i.e.* $g_0 \approx 3g_d$. The minimum V_H can be further decreased by decreasing F_{adh} and g_d, or increasing k_{eff} and A_{ACT}. In addition, it is very interesting to note that (2.31) has the same form as (2.24). This indicates that V_H of both a PI-mode relay and a NPI-mode relay converges to the same minimum value if the relay is designed to operate at the boundary of PI-mode and NPI-mode operation (*i.e.* $g_0 \approx 3g_d$) and has a large effective spring constant (*i.e.* $k_{eff}g_d \gg F_{adh}$). This provides an important design guideline for achieving an ultra-low-voltage logic relay.

The analysis of switching energy of a PI-mode relay is similar to that for a NPI-mode relay. Again assuming $V_B = -V_{RL}$ and $V_{DD} = V_H$, the energy to turn on a PI-mode relay is

$$E_{PI} = V_{GB} * \Delta Q = V_{GB}\left(C_{on}V_{GB} - C_{off}|V_B|\right)$$

$$= C_{on}(V_{RL} + V_H)^2 - C_{off}V_{RL}\,(V_{RL} + V_H)$$

$$\approx \left(C_{on} - C_{off}\right)V_{RL}^2 + \left(2C_{on} - C_{off}\right)V_{RL}V_H$$

$$^{\text{i}} \approx \frac{2\left(k_{eff}g_d - F_{adh}\right)(g_0 - g_d)g_{d_{BS}}}{g_0 - g_d + g_{d_{BS}}} +$$

$$\frac{g_0 - g_d + 2g_{d_{BS}}}{(g_0 - g_d)(g_0 - g_d + g_{d_{BS}})}\sqrt{\frac{2(k_{eff}g_d - F_{adh})(g_0 - g_d)^2}{\varepsilon A_{ACT}}}\left(\sqrt{\frac{8k_{eff}{g_0}^3}{27\varepsilon A_{ACT}}} - \sqrt{\frac{2(k_{eff}g_d - F_{adh})(g_0 - g_d)^2}{\varepsilon A_{ACT}}}\right). \tag{2.32}$$

To gain insight into the dependence of E_{PI} on relay design parameters, numerical simulation is needed and will be presented in Chapter 4.

2.5 Comparison of Minimum Voltage and Energy of Different Relay Designs

The above analytical analyses are summarized in Table 2.1. The idealistic design is the conventional one which has zero V_{RL} (and hence doesn't need Body biasing), but requires the spring constant k_{eff} to be as small as possible. It can achieve smaller switching energy E, but is susceptible to stuck-ON failure. The body-biasing designs, in contrast, have stiffer spring constant, and thus can be turned off properly. Note that both NPI-mode and PI-mode body-biased relays have similar V_{DD_min} (see (2.24) and (2.31)) which can be achieved if $k_{eff} \gg F_{adh}/g_d$ and $g_0 \approx 3g_d$. The performances of NPI-mode relay and PI-mode relays become similar if g_0 approaches $3g_d$, which is not surprising because both V_{DD_min} and E_{min} are continuous functions of g_0.

[i] The second-order term $C_{on}V_H^2$ is very small, hence ignored.

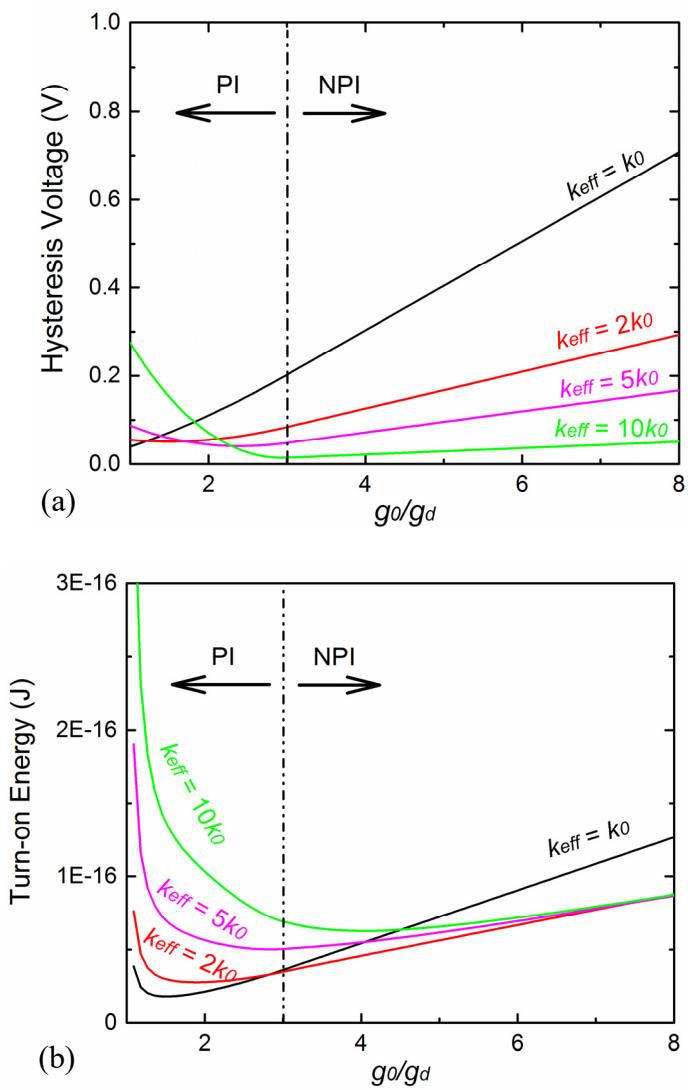


Fig. 2.6. Effects of g_0/g_d and k_{eff} on V_H (a) and E_{min} (b). When k_{eff} is small (close to k_0), PI mode has smaller V_H and E_{min}. Increasing k_{eff} shifts the minimum-V_H and minimum-E_{min} points towards the NPI region of relay design. When $k_{eff} \gg F_{adh}/g_d$, V_H is minimized when g_0 is near $3g_d$.

The value of g_0 to minimize energy of a body-biased relay is not very obvious. On one hand, the term $F_{adh}(g_0 - g_d)$ on the right-hand-side of Eq. (2.26) decreases monotonically with decreasing g_0 and reaches a minimum value at $g_0 = 3g_d$. This indicates that an even smaller g_0 (*e.g.* $< 3g_d$, *i.e.* in PI mode) may be beneficial for achieving smaller energy. On the other hand, since k_{eff} is relatively large ($k_{eff} \gg \frac{F_{adh}}{g_d}$), the other term $2k_{eff}g_d g_{d_{BS}}$ in (2.26) can be significant unless the relay is designed to be operated in NPI-mode such that $g_{d_{BS}}$ is very small with body biasing. These two competing facts indicate that g_0 should be somewhere close to $3g_d$, but the exact value needs numerical simulation (discussed in Chapter 4). The above analysis can be more clearly illustrated with Fig. 2.6 which shows the effects of both k_{eff} and g_0/g_d on V_H and E_{min}. The following values are used for plotting Fig. 2.6: g_d = 5nm, F_{adh} = 1.81 nN, and A_{ACT} = 1

μm^2. Clearly, when k_{eff} is close to F_{adh}/g_d (*i.e.* similar to the idealistic case), the PI-mode relay has smaller V_H and E_{min}. Increasing k_{eff} reduces V_H but increases E_{min}, and shifts the minimum-V_H and minimum-E points towards the NPI-mode design region. Specifically, when $k_{eff} \gg F_{adh}/g_d$, V_H is minimized when $g_0 \approx 3g_d$.

Please note that, since k_{eff} of a body-biased relay is larger than that of the idealistic design, the body-biased relay can operate with a smaller V_{DD}. In practical ICs, parasitic capacitance-induced energy consumption (*i.e.* CV_{DD}^2) can be more significant than switching energy of the logic devices themselves. Therefore, smaller V_{DD} is more beneficial for achieving ultra-low-power circuits.

TABLE 2.1
COMPARISON OF VOLTAGES AND ENERGY CONSUMPTION OF DIFFERENT DESIGNS

Design		V_{PI}	V_{RL}	V_{DD_min}	E_{min}	Conditions	
Idealistic (w/o V_B)		$\sqrt{\frac{8k_{eff}g_0^3}{27\varepsilon_0 A_{ACT}}}$	0	$F_{adh}\sqrt{\frac{g_d}{k_{eff}\varepsilon A_{ACT}}}$	$2F_{adh}g_d$	$g_0 = 1.5g_d$	$k_{eff} = F_{adh}/g_d$
With body bias	NPI	$\sqrt{\frac{2k_{eff}g_d(g_0-g_d)^2}{\varepsilon_0 A_{ACT}}}$	$\sqrt{\frac{2(k_{eff}g_d-F_{adh})(g_0-g_d)^2}{\varepsilon A_{ACT}}}$	$\sim F_{adh}\sqrt{\frac{2g_d}{k_{eff}\varepsilon A_{ACT}}}$	$\sim 2k_{eff}g_d g_{d_{BS}} + 2F_{adh}g_d$	$g_0 \approx 3g_d$	$k_{eff} \gg F_{adh}/g_d$
	PI	$\sqrt{\frac{8k_{eff}g_0^3}{27\varepsilon_0 A_{ACT}}}$	$\sqrt{\frac{2(k_{eff}g_d-F_{adh})(g_0-g_d)^2}{\varepsilon A_{ACT}}}$	$\sim F_{adh}\sqrt{\frac{2g_d}{k_{eff}\varepsilon A_{ACT}}}$	$\leq E_{NPI\_min}$. Need simulation.	$g_0 \approx 3(g_d - F_{adh}/g_d)$	

2.6 Summary

A MEM relay can be designed for operation in one of two modes: PI and NPI. Using the parallel-plate approximation, $g_0/g_d \geq 3$ for a NPI-mode relay whereas $g_0/g_d < 3$ PI for a PI-mode relay. The conventionally idealistic design, which requires low k_{eff}, has both small V_{PI} and small switching energy, but unfortunately is susceptible to stuck-ON failure and thus is not practical. A stiffer structure is needed to solve this problem, which results in increased turn-on voltage. Body biasing can be used to achieve a low gate voltage swing V_{DD} (as low as V_H), to avoid significantly increasing the switching energy. The minimum V_{DD} is limited by V_H, so smaller V_H is preferred to minimize V_{DD}.

A body-biased relay can be designed to operate in either PI mode or NPI mode. When k_{eff} is small (*e.g.* close to F_{adh}/g_d), the PI mode design has smaller V_H. Increasing k_{eff} reduces V_H (at the trade-off of increasing $|V_B|$) and shifts the minimum-V_H design point (*i.e.* g_0/g_d) towards the NPI design. When $k_{eff} \gg F_{adh}/g_d$, the minimum-V_H design point approaches $g_0/g_d \approx 3$, *i.e.* the boundary of PI-mode and NPI-mode designs.

2.7 References

[1] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. K. Das, W. Haensch, E. J. Nowak, and D. M. Sylvester, "Ultralowvoltage, minimum-energy CMOS," *IBM J. Res. & Dev.*, vol. 50, no. 4/5, pp. 469-490, Sep. 2006.

[2] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in sub-threshold circuits," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778-1786, Sep. 2005

[3] F. Chen, H. Kam, D. Marković, T.-J. King Liu, V. Stojanović, and E. Alon, "Integrated circuit design with NEM relays," *Proc. IEEE/ACM Int'l Conf. Comput.-Aided Des.*, 2008, pp. 750-757.

[4] R. Nathanael, V. Pott, H. Kam, J. Jeon, and T.-J. King Liu, "4-Terminal relay technology for complementary logic", *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2009, pp. 223-226

[5] J. Fujiki, N. Xu, L. Hutin, I-R. Chen, C. Qian, and T.-J. K. Liu, "Microelectromechanical relay and logic circuit design for zero crowbar current," *IEEE Trans. Electron Devices*, vol. 61, no.9, pp. 3296-3302, Sep. 2014

[6] N. Xu, J. Sun, I-R. Chen, L. Hutin, Y. Chen, J. Fujiki, C. Qian, and T.-J. K. Liu, "Hybrid CMOS/BEOL-NEMS technology for ultra-low-power IC applications," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2014, pp. 677-680.

[7] T.-J. K. Liu, N. Xu, I.-R. Chen, C. Qian, and J. Fujiki, "NEM Relay Design for Compact, Ultra-Low-Power Digital Logic Circuits," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2014, pp. 327-330.

[8] H. Kam, T.-J. K. Liu, V. Stojanovic, D. Markovic, and E. Alon, "Design, optimization and scaling of MEM relays for ultra-low-power digital logic," *IEEE Trans. Electron Devices*, vol. 58, no. 1, pp. 236-250, 2011.

[9] O. Y. Loh, and D. Espinosa, "Nanoelectromechanical contact switches," *Nature Nanotech.*, vol. **7**, pp. 283-296, May 2012.

[10] J. O. Lee, Y.-H. Song, M.-W. Kim, M.-H. Kang, J.-S. Oh, H.-H. Yangm and J.-B. Yoon, "A sub-1-volt nanoelectromechanical switching device," *Nature Nanotech.*, vol. 8, pp. 36-40, Jan. 2013.

[11] G. N. Nielson, and G. Barbastathis, "Dynamic pull-In of parallel-plate and torsional electrostatic MEMS actuators," *IEEE J. MEMS*, vol. 15, pp. 811-821, 2006.

[12] C. Pawashe, K. Lin, and K. J. Kuhn, "Scaling limits of electrostatic nanorelays," *IEEE Trans. Electron Devices*, vol. 60, no. 9, pp. 2936-2942, Sep. 2013

[13] Y. Chen, R. Nathanael, J. Jeon, J. Yaung, L. Hutin, and T.-J. K. Liu, "Characterization of contact resistance stability in MEM relays with tungsten Electrodes," *IEEE J. MEMS*, vol. 21, no. 3, pp. 511-513, 2012.

Chapter 3

Body-Biased Micro-Electro-Mechanical Relay

3.1 Introduction

As discussed in chapter 2, body biasing is an effective way to reduce the operating voltage and power consumption for relays designed with large stiffness in order to avoid stuck-ON failure. This chapter focuses on the study of body-biased relays.

Section 3.2 introduces an improved relay structure which has 6-terminals and can be operated in a similar fashion as the 4-T relay introduced in the first chapter. Both quasi-static and dynamic performance of fabricated 6-T relays are characterized in section 3.2. Effects of body biasing on relay performance are experimentally investigated in section 3.4, leading to the discovery of the energy-delay tradeoff for a body-biased relay. The optimization of this trade-off has two scenarios and both are discussed in section 3.5. Section 3.6 summarizes this chapter.

3.2 Structure and Operation of 6-Terminal Relay

Fig. 3.1(a) shows the structure of a 6-terminal (6-T) logic relay developed for digital logic applications. It has one Gate, one Body, and two pairs of Source/Drain (S/D) electrodes. The structure is similar to that of the 4-T relay shown in Fig. 1.5 except that the 6-T relay has one more set of Channel and Source/Drain electrodes. The advantages of the 6-T relay design are: 1)

there are two switches incorporated into the same footprint, which effectively doubles device density; and 2) the two Channels (corresponding to the two pairs of Source/Drain electrodes) are located only at the edges of the Gate-to-Body actuation area (rather than spanning across the actuation area in the 4-T relay case), which significantly decreases the asymmetry of Gate and Body electrodes [1]. Fig. 3.1(b) is a scanning electron microscope (SEM) image of a fabricated 6-T relay, with as-fabricated gap thicknesses g_d = 135 nm and g_0 = 260 nm. Details of the fabrication process are provided in [1].

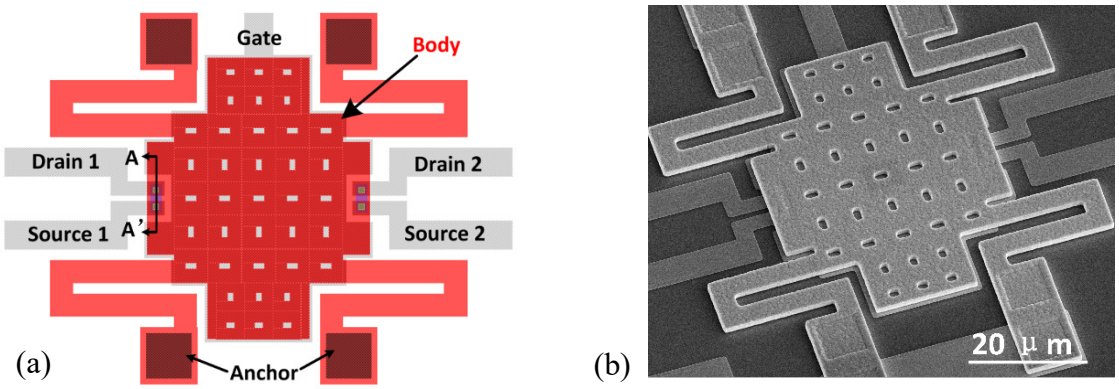


Fig. 3.1. (a) Plan-view schematic of a 6-T relay; (b) Birds-eye SEM image of a fabricated 6-T relay.

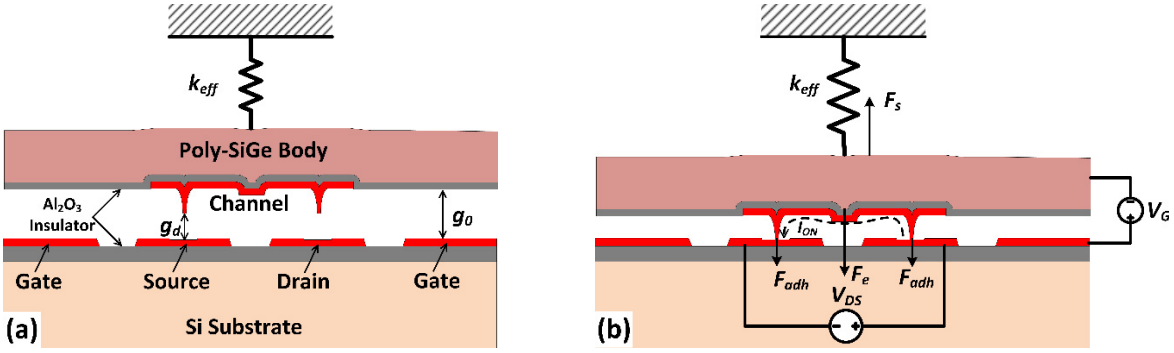


Fig. 3.2. Operation of a logic relay: (a) OFF-state, with as-fabricated actuation gap size g_0 and contact gap size g_d; (b) ON-state, with Gate-to-Body voltage V_{GB}, Source-to-Drain voltage V_{DS}, and current I_{DS}.

Fig. 3.2 shows cross-sectional views along cutline AA' in Fig. 3.1(a). In the OFF state (Fig. 3.2(a)), the Body electrode (formed in a layer of poly-$Si_{0.4}Ge_{0.6}$) is suspended by four folded-flexure beams, so that the Channels (narrow strips of tungsten) attached to its underside via an Al_2O_3 insulating layer are separated from their respective S/D electrodes, which are formed together with the Gate electrode in an underlying layer of tungsten over the insulating substrate. When the magnitude of the Gate-to-Body voltage (V_{GB}) is increased to be equal to or larger than that of the pull-in voltage (V_{PI}), the electrostatic force (F_{elec}) is sufficient to actuate the Body downward such that each Channel contacts its respective pair of S/D electrodes (Fig. 3.2(b)) so that output current (I_{DS}) suddenly can flow. When $|V_{GB}|$ is reduced below the magnitude of the release voltage (V_{RL}), the spring restoring force (F_{spring}) of the suspension beams is sufficient to

actuate the Body upward to separate the Channels from the S/D electrodes so that I_{DS} suddenly drops to zero. By applying a negative body voltage (V_B), the Gate voltage required to turn ON the relay (V_{DD}) can be decreased to $V_{PI} - |V_B|$. For proper relay operation (*i.e.* to ensure that the relay turns OFF at $V_G = 0$ V), the maximum value of $|V_B|$ is V_{RL}; thus the minimum V_{DD} is $V_{PI} - V_{RL}$.

3.3 Characterization of 6-T Relay Performance

3.3.1 Quasi-Static *I-V* Characteristics

A fabricated 6-T relay was characterized in vacuum (0.5 μTorr) at room temperature (~20 ºC) using an Agilent B1500 Semiconductor Parameter Analyzer. The Gate voltage was swept with a step size of 33 mV. Measured current-*vs*.-voltage (*I-V*) characteristics are shown in Fig. 3.3. Note that 1) the switching characteristics are abrupt; 2) the OFF state-current I_{off} is at the noise floor of the test setup; and 3) the ON-state current I_{DS} is limited by a current compliance limit in order to avoid too much Joule heating which may damage the contacts [2]. The black curves are the measured characteristics with zero body bias. The values of V_{PI}, V_{RL} and V_H are 5.43 V, 4.3 V and 1.13 V, respectively, so that ideally an applied Gate voltage $V_G = 1.13$ V would be sufficient to turn on the relay when a body bias $V_B = -4.3$ V is applied. The red curves are the measured characteristics with $V_B = -4$V. As expected, the body-biased relay can be switched on/off with smaller gate voltage swing.

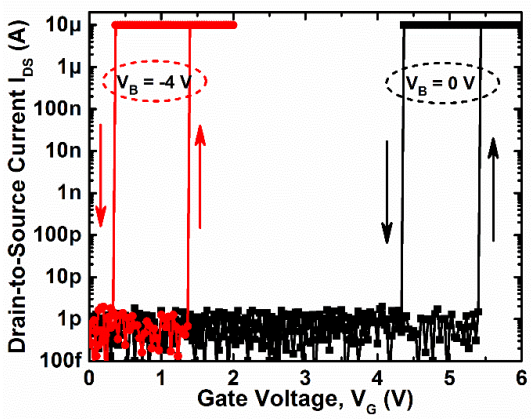


Fig. 3.3. Measured relay *I-V* characteristics.

3.3.2 Dynamic Characteristics

In a relay-based IC the voltage waveforms are more square than triangular (with slowly changing voltage). As briefly mentioned at the end of section 2.2, dynamic switching characteristics are different than quasi-static characteristics and should be separately measured. Fig. 3.4(a) shows a setup for characterizing relay dynamic performance, wherein the 4-T relay can be replaced by a 6-T relay (using one set of Channel/Source/Drain electrodes). The equivalent circuit is shown in Fig. 3.4(b).

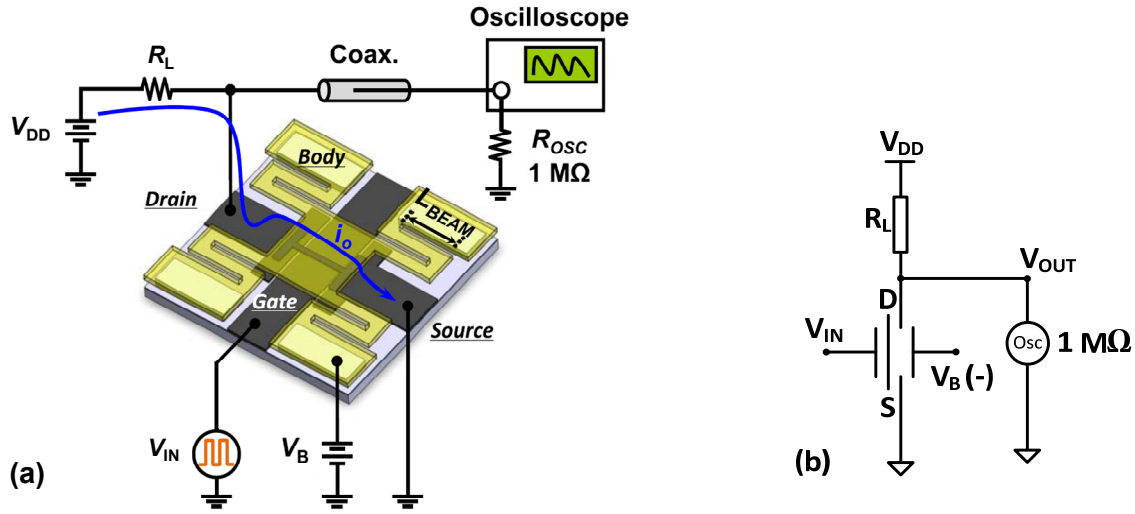


Fig. 3.4. (a) Schematic of dynamic performance characterization setup [3]; (b) Equivalent circuit.

3.3.2.1 Dynamic Pull-in Voltage V_{PI}

A square-waveform voltage signal with gradually increasing magnitude V_{IN} can be used to drive the Gate electrode of a relay, in order to determine dynamic V_{PI}, as shown in Fig. 3.5. If the magnitude of V_{IN} is not high enough, the relay never switches on and the output voltage V_{OUT} is constant at V_{DD}. However, when V_{IN} is equal to or larger than the dynamic V_{PI}, the relay switches on/off so that V_{OUT} is pulled down toward 0 V whenever the relay is in the ON state. Finer increments of V_{IN} result in more accurate determination of dynamic V_{PI}. The measured dynamic V_{PI} of the fabricated 6-T relay is 5.08 V, which is about 93.6% of the quasi-static V_{PI}. This is close to the theoretical prediction (*i.e.* 91.9%) [4]. As was mentioned at the end of section 2.2, the dynamic V_{PI} is smaller than the quasi-static V_{PI} because of physical momentum. If the speed of Body plate is decreased (*e.g.* due to air damping), the dynamic V_{PI} will be closer to the quasi-static V_{PI}.

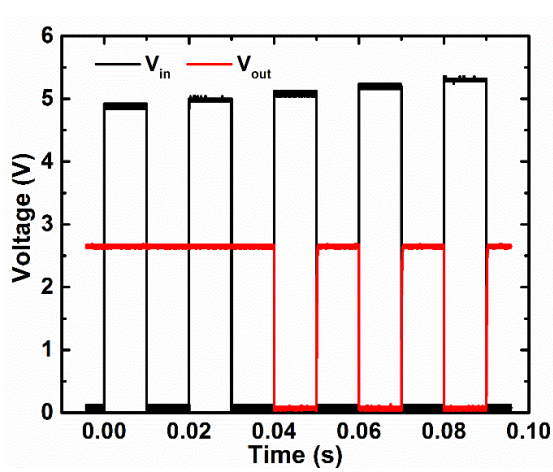


Fig. 3.5. Dynamic V_{PI} measurement.

3.3.2.2 Release Voltage V_{RL}

The release voltage V_{RL} is equal to the Gate input voltage V_{IN} at which the Channel-to-Source/Drain contacts break such that the relay turns off. Immediately after the relay is turned off, V_{DD} starts to charge the output node such that V_{OUT} increases. As shown in Fig. 3.6, using a sloped-falling edge V_{IN}, one can measure V_{RL} by monitoring the rising edge of V_{OUT}. The measured V_{OUT} for this 6-T relay is about 4.3 V.

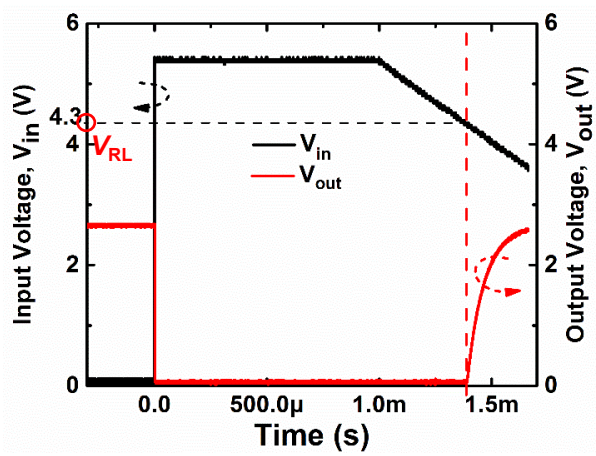


Fig. 3.6. V_{RL} measurement.

3.3.2.3 Turn-on Delay τ_{ON}

By zooming in on the rising edge of V_{IN} in Fig. 3.5 and Fig. 3.6, one can see that there is a delay between the rising edge of V_{IN} and the falling edge of V_{OUT}, as shown in Fig. 3.7. This turn-on delay (τ_{ON}) essentially is the time it takes for the Body to move from its suspended position in the OFF state to its (Channel) contacting position in the ON state. Note that the RC delay (*i.e.* electrical charging time) is comparable to τ_{ON} in Fig. 3.7. Such a large RC delay is mainly due to large parasitic load capacitance at the output node which consists of a large probe pad, long and wide routing wires, probe tip, cables, and input capacitance of the oscilloscope. Especially, the W pads/wires form parasitic capacitors with the Si substrate through a thin (50 nm) Al_2O_3 dielectric layer, so the parasitic capacitance is significant. In a properly designed relay-based circuit, the load capacitance should be much smaller, so the RC delay would be orders of magnitude smaller than the mechanical delay τ_{ON}. Therefore, τ_{ON} limits relay-based integrated-circuit operating speed.

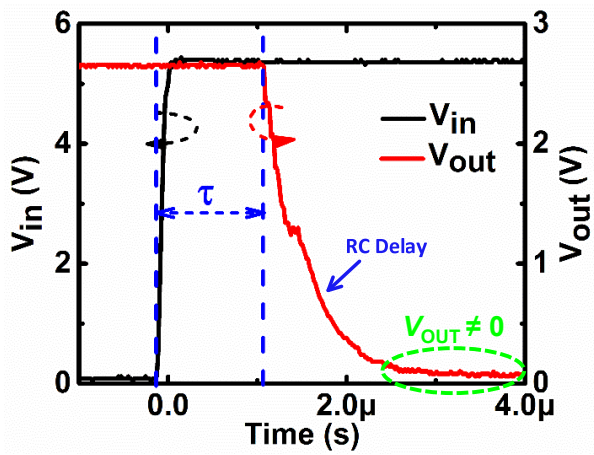


Fig. 3.7. Turn-on delay measurement.

3.3.2.4 Contact Resistance R_C

Close examination of the V_{OUT} waveform in Fig. 3.7 reveals that it doesn't reach 0 V. By re-visiting Fig. 3.4(b) one can readily find the reason, that the ON-resistance R_{ON} of the relay is not zero. R_{ON} can be extracted from the following voltage-divider equation:

$$V_{OUT} = \frac{R_{ON}R_{osc}}{R_{ON}R_{osc} + R_L(R_{ON} + R_{osc})}V_{DD}, \tag{3.1}$$

where V_{OUT} is the minimum output voltage (Fig. 3.7), R_{OSC} is the input resistance of oscilloscope, and R_L is the load resistor (Fig. 3.4). Please note that the conductive Channel of relay is a W strip which has negligible resistance. The dominant component of R_{ON} comes from the contact resistance R_C. Hereinafter, R_C and R_{ON} are used interchangeably unless otherwise noted.

3.4 Effects of Body Biasing on Relay Characteristics

3.4.1 Contact Impact Velocity

Before a discussion of body biasing effects, the concept of contact impact velocity is introduced in this sub-section. It is the speed of the Body (as well as the Channel attached to it) at the moment when the Channel comes into contact with its Source/Drain electrodes. It can depend strongly on the body bias voltage and has a distinct effect on relay performance characteristics.

As mentioned previously, an electrostatically actuated relay can be modeled as a parallel-plate capacitor, as shown in Fig. 3.8 (replica of Fig. 2.5 for the readers' convenience). The top plate (Body electrode) is electrically biased at a voltage V_B, and mechanically supported by suspension beams whose composite effective stiffness is k_{eff}. The bottom plate (Gate electrode) is driven by a square-wave voltage signal with amplitude V_{DD}, through a resistor R. In this simplified model, the movable Body behaves as a time-dependent one-degree-of-freedom oscillator whose governing equation is [5]:

$$m_{eff}\ddot{g} + b\dot{g} - k_{eff}(g_0 - g) = \frac{-\varepsilon_0 A_{ACT}(V_G - V_B)^2}{2\left(g + \frac{d}{\varepsilon_r}\right)^2} \tag{3.2}$$

where g is time-dependent actuation gap, g_0 is as-fabricated actuation gap, m_{eff} is effective dynamic mass, b is damping coefficient, A_{ACT} is effective actuation area, ε_0 is vacuum permittivity, d is film thickness of the Body insulating Al_2O_3 layer, and ε_r is the dielectric constant of Al_2O_3. Since the capacitor is charged/discharged through a resistor, the Gate voltage V_G is also time-dependent and governed by the equation:

$$\frac{d\left(\frac{\varepsilon_0 A_{ACT}(V_G - V_B)}{g + d/\varepsilon_r}\right)}{dt} = \frac{V_{DD} - V_G}{R}. \tag{3.3}$$

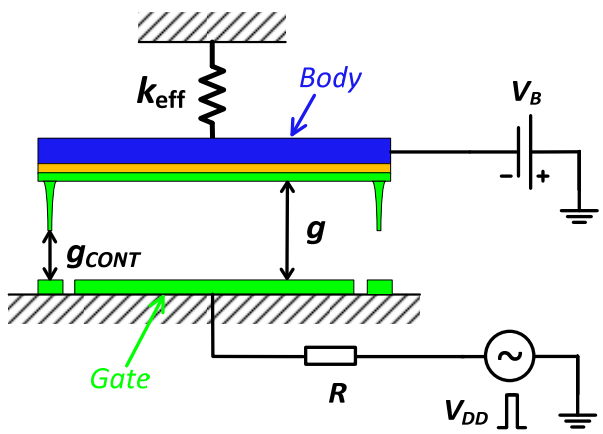


Fig. 3.8. Simplified parallel-plate model of an electrostatically actuated relay.

Given the initial actuation gap size $g(t = 0)$ and initial Body speed $\dot{g}(t = 0)$, the contact impact velocity can be calculated by numerically solving equations (3.2) and (3.3). For a rough estimation, the following approximations can be made: 1) m_{eff} is equal to the actual mass of the Body plate because i) the mass of the suspension beams is much smaller than that of the Body, ii) most parts of the suspension beams move much more slowly than the Body, and iii) deformation of the Body plate is negligible; 2) the damping factor b is zero since the relay was tested in a vacuum chamber (0.5 μTorr); and 3) d is ignored since d/ε_r is much smaller than g. In addition, since the capacitive charging delay (*i.e.* the "*RC*" delay of charging the parallel-plate capacitor through the series resistor R) of a relay is negligible compared to its mechanical switching delay [6], the Gate voltage can be approximated as $V_G = V_{DD}$. So the above two equations can be replaced by:

$$m_{eff}\ddot{g} + b\dot{g} - k_{eff}(g_0 - g) = \frac{-\varepsilon_0 A_{ACT}(V_{DD} + |V_B|)^2}{2g^2}. \tag{3.4}$$

The following parameters which are used in the simulation are extracted from experimental results for fabricated relays: k_{eff} = 78.5 N/m, A_{ACT} = 1584 μm^2, m_{eff} = 7.79 ng, g_0 = 260 nm, V_{DD} + $|V_B|$= 5.4 V, and b = 0. The simulated impact velocity dependence on body bias voltage is shown in Fig. 3.9. Clearly, impact velocity decreases with increasing $|V_B|$ if V_{GB} is fixed. This is because with body biasing: 1) there is less distance for the Body to move and accelerate to a high speed before contact is made, and 2) the average net force on the Body is smaller so that the average acceleration is smaller.

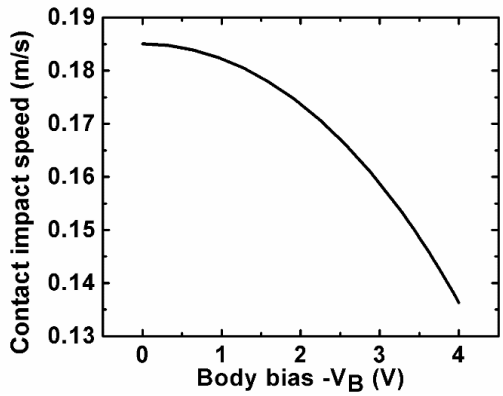


Fig. 3.9. Dependence of contact impact velocity on body bias voltage, for V_{GB} fixed at 5.4 V.

3.4.2 Effects of Body Biasing on V_H and R_C

For the purpose of this study, V_H is defined as the difference between the quasi-static V_{PI} (as opposed to dynamic V_{PI}) and the release voltage V_{RL} (see section 3.3.2.2). The usage of quasi-static rather than dynamic V_{PI} is because the latter parameter is not a fixed value. As explained in section 3.3.2.1, dynamic V_{PI} approaches quasi-static V_{PI} as the Body velocity decreases. Since contact impact velocity decreases with increasing $|V_B|$, the dynamic V_{PI} approaches quasi-static V_{PI} with increasing body bias.

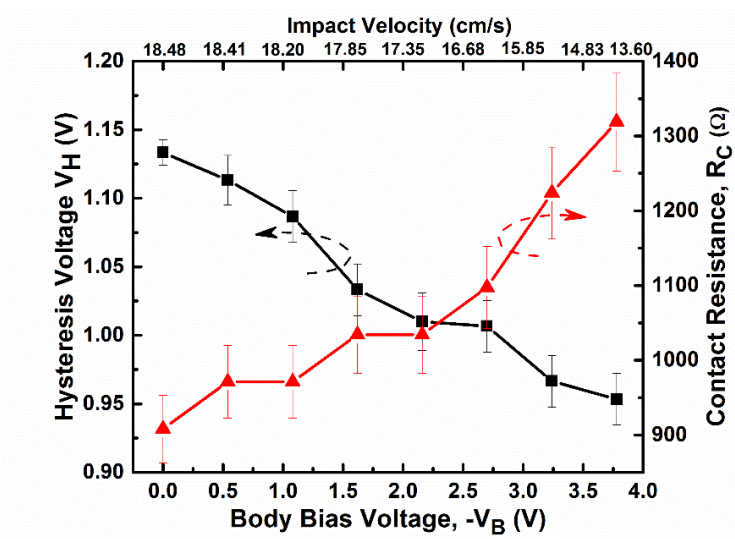


Fig. 3.10. Effects of body bias voltage on hysteresis voltage and contact resistance.

Fig. 3.10 shows the effects of body biasing on V_H and R_C. Interestingly, V_H decreases steadily with the increase of $|V_B|$, and is reduced by approximately 0.2 V with $V_B = -4$ V. R_C, however, increases with increasing $|V_B|$. For easy comparison, the simulated impact velocity for various values of $|V_B|$ are indicated along the top of Fig. 3.10. Higher impact velocity usually results in higher F_{adh} due to both physical and chemical surface changes [7]. Therefore, the observed dependence of V_H on $|V_B|$ can be explained by decreasing F_{adh} with increasing $|V_B|$. The relatively small R_C at low $|V_B|$ (high impact velocity) indicates that there are more/stronger charge conduction paths formed in the ON state due to either increased contact area or more metallic bonds, which coincides with larger F_{adh} and V_H. The fact that V_H decreases with increasing $|V_B|$ is beneficial for minimizing the switching energy, since smaller V_{DD} ($\geq V_H$) can be used.

Fig. 3.11 shows experimental measurements of adhesive force as a function of impact velocity, reported in [7]. Clearly, larger impact velocity results in larger adhesive force.

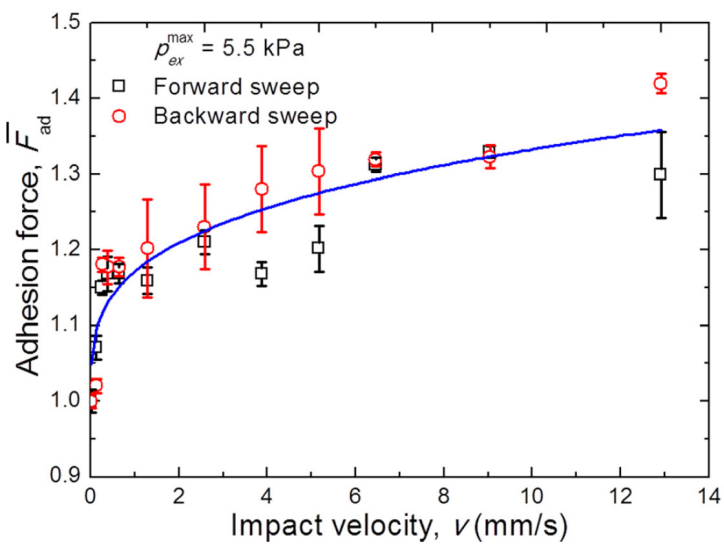


Fig. 3.11. Effect of impact velocity on adhesive force. Reprinted from [7].

3.4.3 Effect of Body Biasing on τ_{ON}

Since the impact velocity decreases with body bias, it is natural to inquire how the turn-ON delay τ_{ON} of a relay changes with body biasing. Fig. 3.12 shows that V_B can significantly affect τ_{ON}. Interestingly, this effect is dependent on V_{GB}. τ_{ON} decreases slightly when the gate-overdrive voltage (*i.e.* $V_{GB} - V_{PI}$) is large, whereas it increases quickly when the gate-overdrive voltage is very small. This is due to two competing factors: the contact air-gap thickness in the OFF state decreases with increasing $|V_B|$; the average velocity of Body movement decreases with increasing $|V_B|$. Fig. 3.13 compares the contact impact velocity for two values of V_{GB}. It is clear that the contact velocity decreases with increasing $|V_B|$ for both cases, but from the normalized velocity curve one can see clearly that the velocity is a more sensitive function of V_B when $V_{GB} = V_{PI}$, that is, for minimum-V_{GB} operation. For small V_{GB} (low gate-overdrive voltage), the decrease in contact velocity dominates the decrease in air-gap thickness, so τ_{ON} increases.

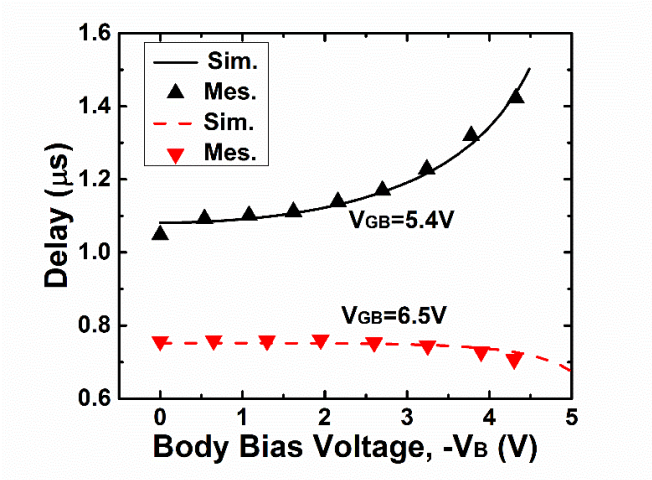


Fig. 3.12. Effect of body bias voltage on the turn-ON delay of a relay with V_{PI} = 5.4V: triangle symbols (measured) and solid/dashed lines (simulated). With increasing $|V_B|$, delay increases for V_{GB} = 5.4V, but decreases for V_{GB} = 6.5V.

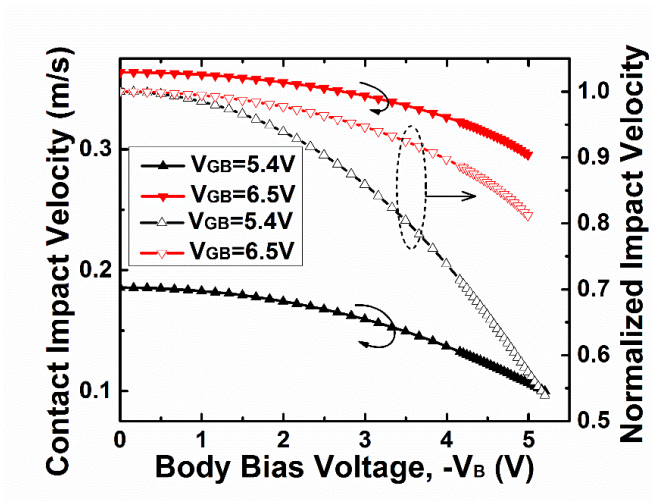


Fig. 3.13. (Simulated) Effect of body bias on contact impact velocity. The open triangle lines are the velocity data normalized to that of $|V_B|$ = 0V. For both values of V_{GB} (6.5V and 5.4V), the contact impact velocity decreases with increasing $|V_B|$; but for smaller V_{GB}, the normalized velocity decreases more rapidly.

3.4.4 Energy-Delay Trade-Off for Body-Biased Relay

The relationship between relay switching energy and turn-on delay for each V_{GB} case is compared in Fig. 3.14. The switching energy is calculated as follows:

$$E = V_{GB}(C_{ON}V_{GB} - C_{OFF}|V_B|) = \frac{\varepsilon_0 A_{ACT}}{g_0 - g_d}V_{GB}^2 - \frac{\varepsilon_0 A_{ACT}}{g_0 - g_x}V_{GB}|V_B| \tag{3.5}$$

where g_x is the top plate displacement of the body-biased relay for V_G = 0 V. The triangular symbols represent experimental data for different values of $|V_B|$ with the same value of V_{GB}. The dashed/solid lines are the calculated energy-delay curves. The experimental data well match simulations which show that, for minimum-V_{GB} operation, body-biasing is effective for reducing energy at a tradeoff of increasing delay, but that for larger-V_{GB} operation both energy and delay decrease with increasing $|V_B|$. Note that a lower energy limit exists for each of the curves in Fig. 3.14, since $|V_B|$ is limited to be not larger than V_{RL}. In fact, the minimum energy is achieved for minimum-V_{GB} operation. But for a higher switching energy, minimum-V_{GB} operation is not necessarily optimal since it has a relatively large τ_{ON}. Parameter optimization via numerical simulation is needed to achieve the minimum delay for a given switching energy.

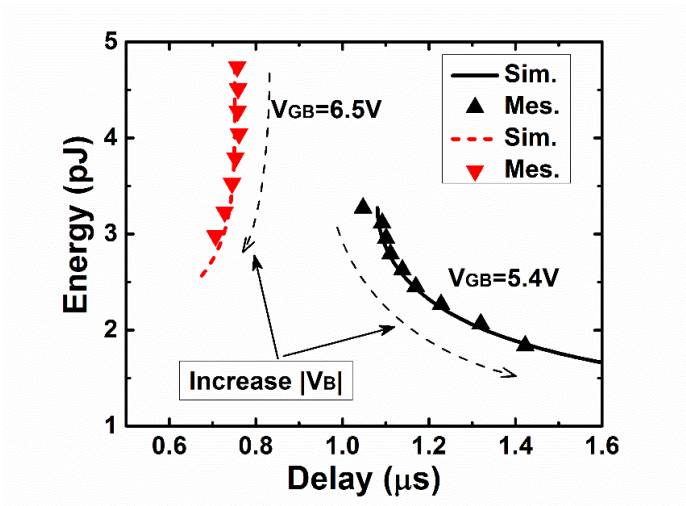


Fig. 3.14. Energy-delay performance of a relay with V_{PI} = 5.4V: triangle symbols (measured) and solid/dashed lines (simulated). With increasing $|V_B|$, the switching energy always decreases, but the delay increases when V_{GB} is low while it decreases when V_{GB} is high.

In addition to V_B and V_{DD} (= V_{GB} - $|V_B|$), another important parameter deserving careful consideration is the as-fabricated actuation gap g_0. From Eq. (3.5) it can be seen that E can be reduced by making the OFF-state gate capacitance (C_{OFF}) approach the ON-state gate capacitance (C_{ON}), *i.e.* by reducing the initial (V_G = 0) contact dimple gap thickness $g_d - g_x$ to be nearly zero. This is not possible, however, unless the relay is designed to operate at the boundary or inside the NPI-mode design space, *i.e.* $g_d \leq g_0/3$ (see section 2.4.3). The solid line in Fig. 3.15 is the calculated energy-delay curve for a relay that is identical to the fabricated device except with $g_0 = 3g_d$ = 405 nm instead of 260 nm. It can be seen that the energy-delay trade-off is significantly improved for this theoretical design, even though it has a larger value of V_{PI} (10.5 V). For comparison, one point on each curve in this figure is highlighted, for the same gate operating voltage ($V_{DD\text{-}B} = V_{DD\text{-}A}$ = 1.1 V). Although a larger body bias voltage is required for design B, this does not directly affect the dynamic power consumption of a relay-based circuit since the Body does not conduct current. To further improve the energy efficiency of a relay, design optimization is necessary.

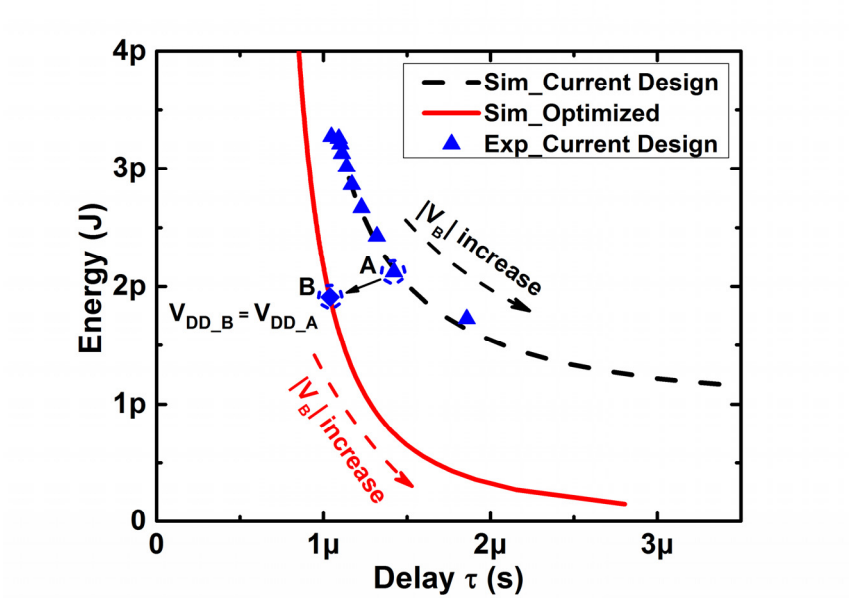


Fig. 3.15. Improvement of energy efficiency by using $g_0 = 3g_d$. The solid optimized curve has constant V_{GB} = 10.5 V, and the dashed curve has constant V_{GB} = 5.4 V.

3.5 Energy-Delay Performance Optimization of NEM Relay

A methodology for optimizing the energy-delay performance of a nano-electro-mechanical relay is developed herein. Contrary to popular belief that structural stiffness should be minimized to achieve minimum switching energy, it is shown that the effective spring constant should be relatively large for better energy-delay performance. The optimal energy-delay of an aggressively scaled NEM relay (with 5 nm contact gap as fabricated) is presented. Simulations of nanoscale relay designs indicate that body biasing can be used to mitigate relay manufacturing challenges while enabling ultra-low-voltage (sub-100 mV) operation with relatively fast switching speed.

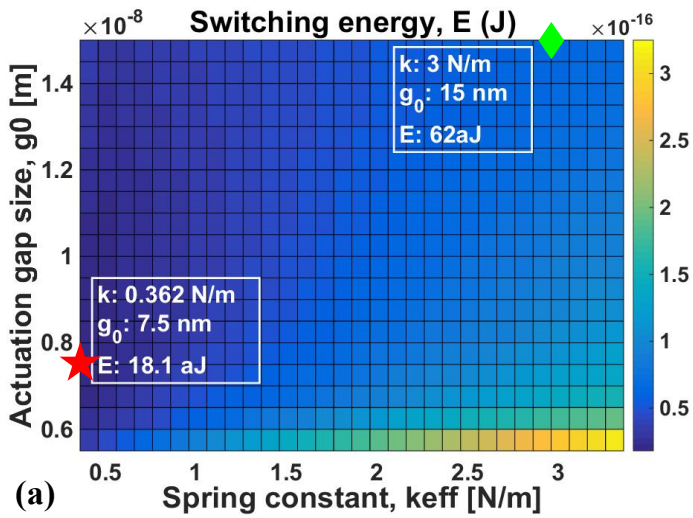
3.5.1 Parameters to Optimize

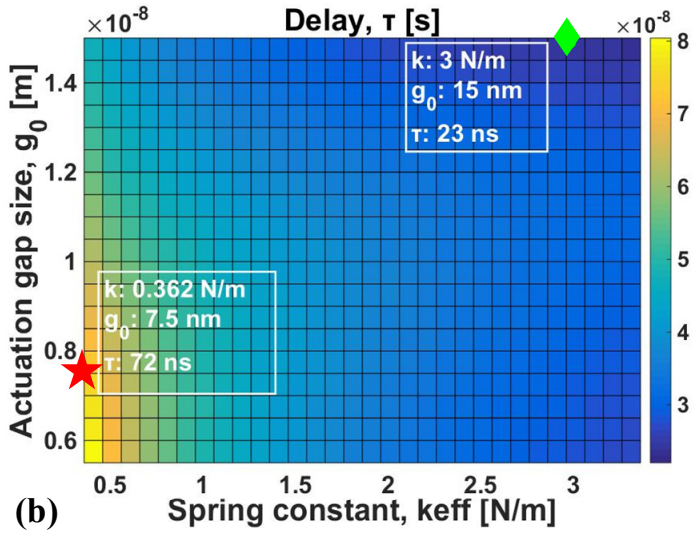
Key parameters for designing a logic relay are: contact adhesive force F_{adh}, as-fabricated contact air gap thickness g_d, as-fabricated actuation air gap thickness g_0, actuation area A_{ACT}, structural poly-$Si_{0.4}Ge_{0.6}$ layer thickness t, effective spring constant of the suspension beams k_{eff}, body bias voltage V_B, and operating voltage V_{DD}. Fundamentally, F_{adh} and g_d are two limiting factors of energy consumption (See Table 2.1). They should be as small as possible. A_{ACT} is usually limited by device density requirement. Based on process technology limitations, the following parameters are used for relay simulation: g_d = 5 nm; A_{ACT} = 1 μm^2; t = 30 nm; and F_{adh} = 1.81 nN (which corresponds to electrodes with real contact area A_C = 1 nm^2 and coated with an ultra-thin (3 Å) layer of TiO_2) [8]. Therefore the following parameters remain to be co-optimized for relay design: g_0, k_{eff}, V_B, and V_{DD}.

3.5.2 Minimization of V_{DD}

If minimization of V_{DD} is a design objective (relevant for the case of interconnect-dominated energy consumption), then ideally V_B should be always set to $-V_{RL}$ so that V_{DD} can be set to the minimum value (*i.e.* $V_{PI} - V_{RL}$). In this case, the remaining design parameters to be co-optimized are k_{eff} and g_0.

The contour plots in Fig. 3.16 show how the switching energy, delay and V_{DD} change with k_{eff} in the range from 0.362 N/m (= F_{adh}/g_d, *i.e.* the minimum value required to ensure $F_S > F_{adh}$) to 3.62 N/m, and with g_0 in the range from 5 nm ($g_0 > g_d$) to 15 nm (the boundary of PI and NPI mode). It can be seen that for low values of k_{eff}, delay and V_{DD} are relatively large although energy is small. Qualitatively, the effects of k_{eff} and g_0 can be more clearly seen from the energy-*vs*.-delay plot in Fig. 3.17: larger k_{eff} is beneficial for lower turn-ON delay, and larger g_0 (below $3g_d$) is beneficial for lower switching energy.





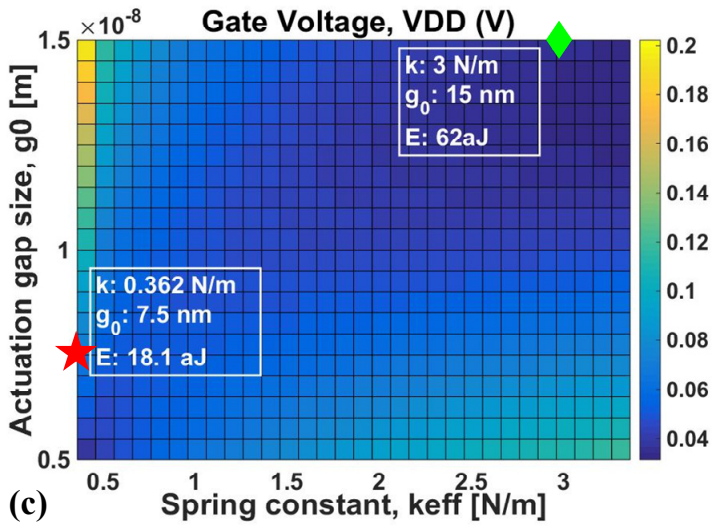


Fig. 3.16. Dependence of (a) switching energy, (b) delay, and (c) minimum operating voltage on k_{eff} and g_0. $V_B = -V_{RL}$. V_{DD} is set to be $V_{PI} - V_{RL}$. Under this condition, the energy and delay are only functions of k_{eff} and g_0. The red star (small k_{eff} and g_0) corresponds to the conventional minimum energy design, while the green diamond (large k_{eff} and g_0) corresponds to a body-biased relay design for comparison.

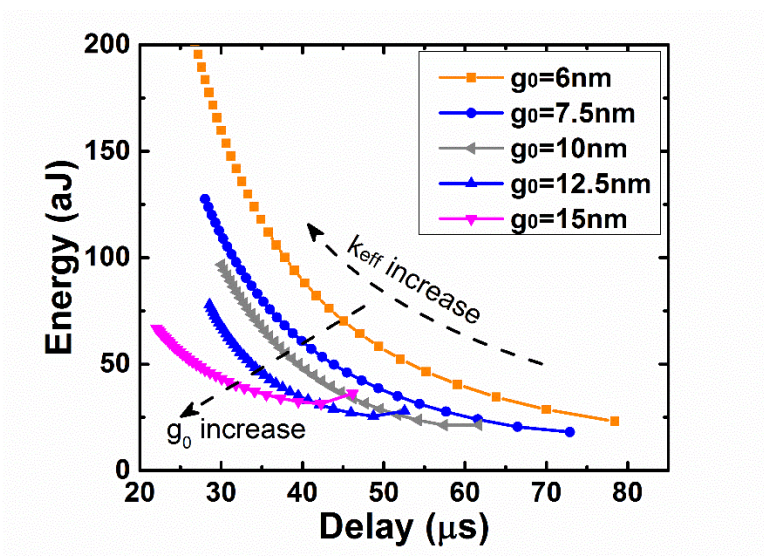


Fig. 3.17. (Simulated) Effect of k_{eff} and g_0 on the energy-delay performance of a body-biased relay, from Fig. 3.16. For a fixed g_0, increasing k_{eff} increases energy and reduces delay. Increasing g_0 tends to improve energy-delay performance.

Conventional wisdom suggests that k_{eff} should be as small as possible to minimize the switching energy of a relay, but this does not consider body biasing as a means to reduce V_{DD} and energy. In Fig. 3.16, the conventional optimal (idealistic) relay design (small k_{eff} and g_0) for minimum switching energy is highlighted with a star symbol, which requires $g_0 = 1.5g_d = 7.5$ nm

and $k_{eff} = F_{adh}/g_d = 0.362$ N/m (see section 2.3.2). An alternative body-biased relay design (large k_{eff} and g_0) is highlighted with a diamond symbol, which has more relaxed values of g_0 and k_{eff}: $g_0 = 15$ nm is chosen so that the relay operates at the boundary of PI and NPI mode; $k_{eff} = 3$ N/m is chosen to avoid the possibility of stuck-ON failure. This alternative design eases manufacturing challenges and can provide for reduced device footprint, since shorter suspension beams can be used. The energy-delay performance of these two designs is compared in Fig. 3.18. For both designs, the delay can be reduced by increasing V_{DD} at a tradeoff of higher energy consumption, as shown by the dashed and dash-dot lines. It is clear that the body-biased relay design is more energy efficient for small delay. Note that the solid curve (body-biased design) cannot extend further down beyond the diamond point since $|V_B|$ is limited to be no larger than V_{RL} which is given by

$$V_{RL} = \sqrt{\frac{2(k_{eff}g_d - F_{adh})(g_0 - g_d)^2}{\varepsilon_0 A_{ACT}}} = 546\ mV \tag{3.6}$$

In one sense the conventional design looks better since it can achieve smaller switching energy, but this comes at the cost of significantly increased delay. For a target switching delay, a body-biased relay design generally is more energy efficient. In addition, the low k_{eff} for the conventional design makes the relay susceptible to stuck-ON failure (*i.e.* low endurance) since the spring restoring force is small.

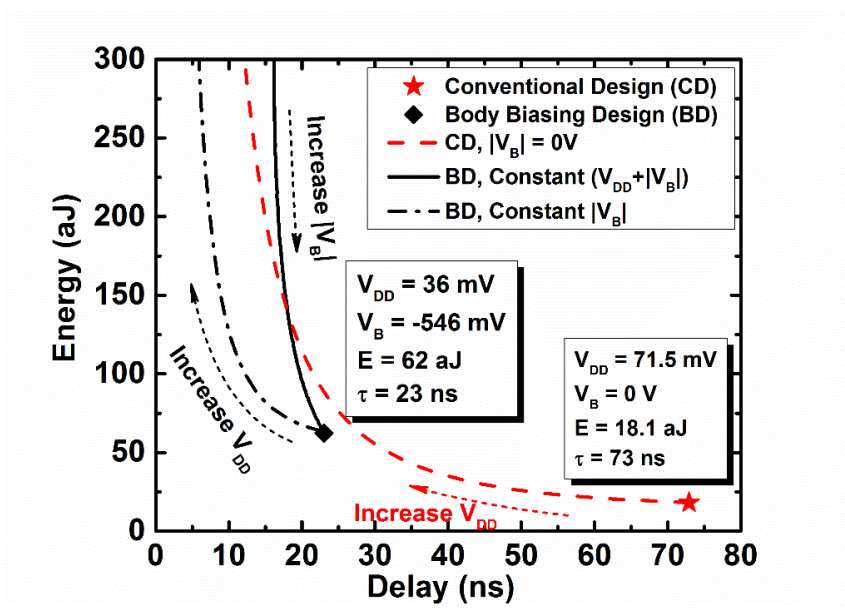


Fig. 3.18. (Simulated) Energy-delay performance comparison for the two relay designs highlighted in Fig. 3.16. Clearly the body-biased relay design has lower V_{DD} and smaller delay than the conventional design. For each design, the delay can be reduced further by increasing V_{DD}.

The key design parameters and simulated performance for the two nano-electro-mechanical (NEM) relay designs are compared in Table 3.1. The performance of the relay with relaxed g_0 and k_{eff} but without body biasing (*i.e.* for $V_B = 0$ V) is also included for comparison. It can be seen that the body-biased relay design provides for the lowest gate operating voltage and better

energy efficiency at 23 ns delay, as compared to the idealistic relay design for operation with the theoretical minimum switching energy. Considering that interconnect and parasitic capacitances may be comparable to or even larger than the intrinsic capacitance of a NEM relay, lower-voltage operation ultimately may provide for more energy efficient relay-based integrated circuits. With body biasing, it is possible to maintain a low gate operating voltage while further scaling down A_{ACT} to reduce the device footprint.

TABLE 3.1
NEM RELAY DESIGN PARAMETERS AND ENERGY-DELAY PERFORMANCE

Designs		Design Parameters					Simulated Performance			
		Constraints			Knobs					
		g_d (nm)	F_{adh} (nN)	A_{ACT} (μm^2)	g_0 (nm)	k_{eff} (N/m)	$\|V_B\|$ (mV)	V_{DD} (mV)	E (aJ)	τ(ns)
Idealistic		5	1.81	1	7.5	0.362	0	72	18.1	73
Relaxed	w/o V_B	5	1.81	1	15	3	0	582	300	16
	w/ V_B	5	1.81	1	15	3	546	36	62	23

3.5.3 Minimization of Energy and Delay

If low-voltage operation is not the primary design objective, V_{DD} would not need to be fixed at $V_{PI} - V_{RL}$. In other words, V_B and V_{DD} should be co-optimized. The co-optimization of g_0, k_{eff}, V_B and V_{DD} for minimum energy-delay is a differential-equation-based multi-dimensional optimization problem with both linear and non-linear constraints, which can be solved numerically. The mathematical procedure is briefly described as follows.

The governing equations of the electro-mechanical-coupled one-degree-of-freedom oscillation system have been detailed in Eq. (3.2) and Eq. (3.3). For simplification, Eq. (3.4) can be used instead. For the reader's convenience, it is repeated here:

$$m_{eff}\ddot{g} + b\dot{g} - k_{eff}(g_0 - g) = \frac{-\varepsilon_0 A_{ACT}(V_{DD} + |V_B|)^2}{2g^2}, \tag{3.4}$$

where m_{eff}, b, ε_0, and A_{ACT} are known; k_{eff}, g_0, V_{DD}, and V_B are to be optimized; and g is a function of time. The boundary conditions are:

$$\left.\begin{aligned} g(\mathrm{t} = 0) &= x_0 \\ \dot{g}(t = 0) &= 0 \end{aligned}\right\} \tag{3.7}$$

where x_0 satisfies

$$k_{eff}x_0 = \frac{\varepsilon_0 A_{ACT} V_B^2}{2(g_0 - x_0)^2}. \tag{3.8}$$

Conceptually, solving differential Eq. (3.4) for $g(t)$ would give an implicit expression

$$g(t) = g\left(t, k_{eff}, g_0, V_{DD}, V_B\right). \tag{3.9}$$

The mechanical delay τ is the time t at which contact gap drops to zero, *i.e.*

$$g(t = \tau) = g_0 - g_d, \tag{3.10}$$

where g_d is known. By solving (3.9) and (3.10), one would be able to get

$$\tau = g^{-1}\left(k_{eff}, g_0, g_d, V_{DD}, V_B\right) \tag{3.11}$$

Remember that the goal here is to optimize k_{eff}, g_0, V_{DD}, and V_B so as to minimize the delay τ. For this optimization, the following constraints must be met:

$$\left.\begin{aligned}
E &= V_{GB}(C_{ON}V_{GB} - C_{OFF}|V_B|) \\
C_{ON} &= \frac{\varepsilon_0 A_{ACT}}{g_0 - g_d} \\
C_{OFF} &= \frac{\varepsilon_0 A_{ACT}}{g_0 - x_0} \\
V_{GB} &= V_{DD} + |V_B| \\
V_B &< -V_{RL} \\
V_{DD} &> V_{PI} - |V_B| \\
V_{PI} &= \sqrt{\frac{8k_{eff}g_0^3}{27\varepsilon_0 A_{ACT}}} \\
V_{RL} &= \sqrt{\frac{2(k_{eff}g_d - F_{adh})(g_0 - g_d)^2}{\varepsilon_0 A_{ACT}}} \\
k_{eff} &> \frac{F_{adh}}{g_d}
\end{aligned}\right\} \tag{3.12}$$

To rephrase, the optimization problem is mathematically equivalent to the following problem: for given constraints (3.12) and boundary condition (3.7), optimize k_{eff}, g_0, V_{DD}, and V_B, such that for each given energy, the target function (3.11) is minimized. Since (3.11) is an implicit expression, there is no closed-form analytical solution. A numerical method must be used to solve this ordinary differential equation based optimization problem. Computational software such as MATLAB [9] can handle this well.

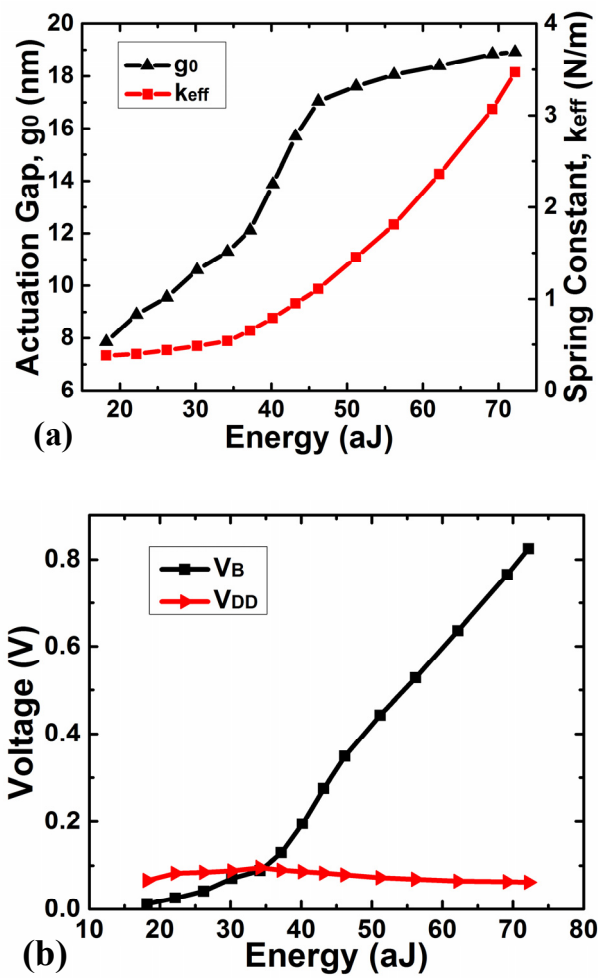


Fig. 3.19. Numerical optimization results of (a) k_{eff}, g_0, and (b) V_{DD}. For a given switching energy, there is an optimal combination of k_{eff}, g_0, and V_{DD} to minimize turn-on delay. In general, one should use high k_{eff}, high g_0, and high $|V_B|$ to minimize delay.

The optimal values of g_0, k_{eff}, V_B and V_{DD} to minimize delay across a range of targeted switching energies are shown in Fig. 3.19. It can be seen that the optimal value of g_0 increases quickly across the boundary between PI-mode and NPI-mode design regions (*i.e.* $g_0 = 3g_d = 15$ nm) and then more gradually with increasing energy. This is due to the trade-off between energy and delay: qualitatively, pull-in mode operation yields smaller switching energy while non-pull-in mode operation yields smaller delay. The optimal value of k_{eff} increases with energy, since delay improves with $1/\sqrt{k_{eff}}$ [10]. Accordingly, the optimal value of $|V_B|$ must increase as k_{eff} increases to take full advantage of body biasing to reduce V_{DD}. (Note that V_{DD} is less than 0.1 V, even at relatively high switching energies.) The final energy *vs.* minimum-delay curve is plotted in Fig. 3.20.

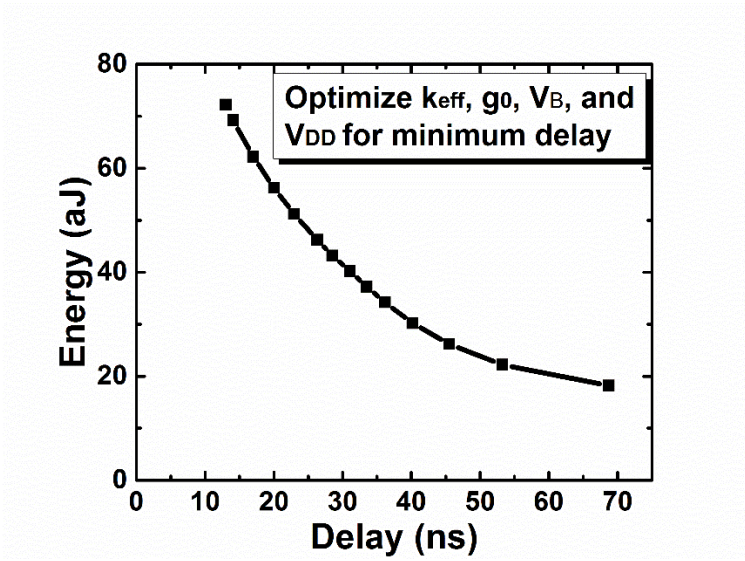


Fig. 3.20. (Simulated) The ultimate energy-delay performance of a NEM logic relay for a given technology (A_{ACT} = 1 um^2, g_d = 5 nm, and F_{adh} = 1.81 nN.)

3.6 Summary

The effects of body biasing on relay performance characteristics were investigated in this chapter. The switching hysteresis voltage, which sets a lower limit for the relay operating voltage, is experimentally found to decrease with increasing body bias voltage, due to reduced contact adhesive force. The effect of body biasing on mechanical turn-on delay depends on gate-overdrive voltage: with increasing $|V_B|$, delay increases for small gate-overdrive voltage but decreases for large gate-overdrive voltage. It is demonstrated that the switching energy of a relay can be reduced by body biasing, at a trade-off of increased mechanical turn-on delay.

Simulations of nanoscale relay designs indicate that body biasing can be used to substantially mitigate relay manufacturing challenges while enabling ultra-low-voltage (sub-100 mV) operation with relatively fast switching speed. A methodology for optimizing the energy-delay performance of a nano-electro-mechanical relay is developed. Contrary to popular belief that structural stiffness should be minimized to achieve minimum switching energy, this work shows that the effective spring constant should be relatively large for optimal energy-delay performance. The optimal energy-delay curve for an aggressively scaled NEM relay (with 5 nm contact gap as fabricated) is presented.

3.7 References

[1] R. Nathanael, J. Jeon, I-R. Chen, Y. Chen, F. Chen, H. Kam and T.-J. K. Liu, "Multi-input/multi-output relay design for more compact and versatile implementation of digital logic with zero leakage," *Proc. of the IEEE 19th Int'l Symp. on VLSI Technology, Systems, & Applications (VLSI-TSA'12)*, Hsinchu, Taiwan, 2012.

[2] Y. Chen, R. Nathanael, J. Jeon, J. Yaung, L. Hutin, and T.-J. K. Liu, "Characterization of contact resistance stability in MEM relays with tungsten electrodes," *IEEE J. MEMS*, vol. 21, pp. 511-513, 2012.

[3] Y. Chen, I.-R. Chen, C. Qian, A. Peschot, and T.-J. K. Liu, "Experimental studies of contact detachment delay in microrelays for logic applications," *IEEE Trans. Electron Devices*, 62, pp. 2695-2699, 2015.

[4] G. N. Nielson, and G. Barbastathis, "Dynamic pull-In of parallel-plate and torsional electrostatic MEMS actuators," *IEEE J. MEMS*, vol. 15, pp. 811-821, 2006.

[5] Senturia, S. D. *Microsystem Design* Ch. 6 (Kluwer Academic Publishers, 2002).

[6] F. Chen, H. Kam, D. Marković, T.-J. King Liu, V. Stojanović, and E. Alon, "Integrated circuit design with NEM relays," *Proc. IEEE/ACM Int'l Conf. Comput.-Aided Des.*, 2008, pp. 750-757.

[7] H. Xiang, and K. Komvopoulos, "The effect of impact velocity on interfacial adhesion of contact-mode surface micromachines," *App. Phy. Lett.*, vol. 101, 053506 (2012).

[8] C. Pawashe, K. Lin, and K. J. Kuhn, "Scaling limits of electrostatic nanorelays," *IEEE Trans. Electron Devices*, vol. 60, no. 9, pp. 2936-2942, Sep. 2013.

[9] http://www.mathworks.com/

[10] H. Kam, et al., "Design, optimization and scaling of MEM relays for ultra-low-power digital logic," *IEEE Trans. Electron Devices*, vol. 58, no. 1, pp. 236-250, 2011.

Chapter 4

Millivolt Relay Technology

4.1 Introduction

Electrostatically actuated mechanical switches have zero I_{OFF} and abrupt switching characteristics, so that they in principle can be made to operate at much lower voltages than transistors [1,2]. However, the operation voltage of electromechanical switches generally has been higher than that for MOSFET, *i.e.* > 1 V, despite recent progress [1,3,4]. To fulfill the promise of mechanical computing, a main challenge is to lower the operation voltage (*e.g.*, to be less than 0.1 V) so as to outperform CMOS technology in terms of energy efficiency.

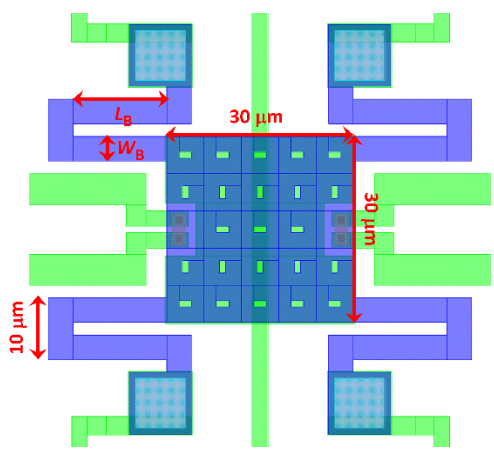
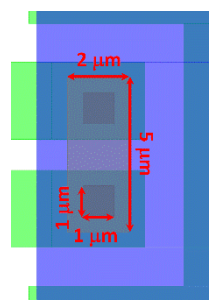
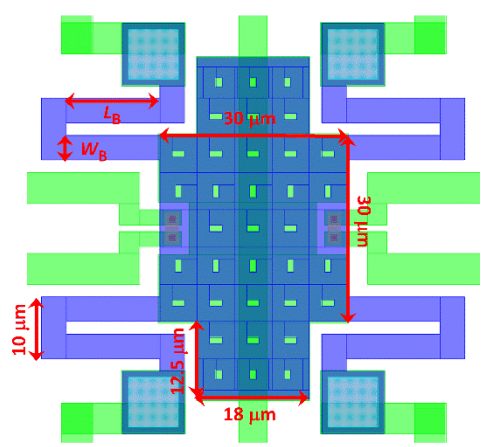
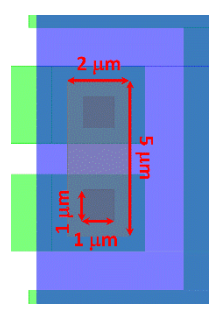
In chapter 3, body biasing was introduced as a means for improving relay switching energy efficiency and discussed in detail. With body biasing, it is possible to achieve ultra-low-voltage operation with relatively fast switching speed. In this chapter, body-biased relay designs targeting millivolt switching are first presented, with key design parameters discussed in detail, in section 4.2. Both process and device simulations are performed to predict relay performance in section 4.3. The fabrication process is then detailed in section 4.4, and followed by performance characterization in section 4.5. Sub-100 mV relays are successfully demonstrated. A relay-based inverter circuit is demonstrated to operate reliably with a supply voltage below 100 mV, representing a significant milestone toward ultra-low-power mechanical computing.

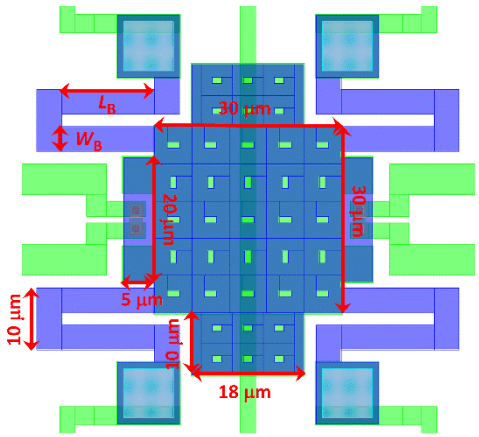
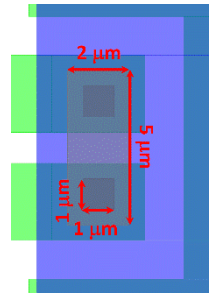
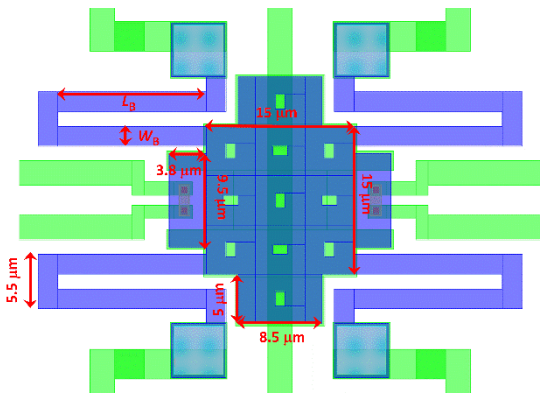
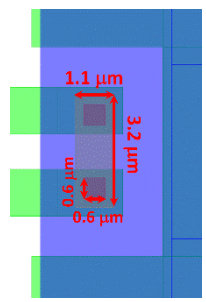
4.2 Millivolt Relay Design

4.2.1 Relay Structure

A 6-terminal (6-T) relay structure, similar to that in Fig. 2.8, was used as the starting point for the body-biased relay design. Table 4.1 shows four designs with differences in the actuation area, beam length/width, and contact area. Key dimensional parameters also are listed in Table 4.1. Designs A, B, and C are very similar, except in shape and size of the actuation area. Design D is a scaled version of C, with each lateral dimension reduced roughly by a factor of 2. For each of the four designs, there are 6 experimental splits in beam length L_B, ranging from 8 μm to 28 μm.

TABLE 4.1
6-T BODY-BIASED RELAY STRUCTURE

Design	Relay Layout	Contact Layout	Main Parameters
A			A_{ACT} = 810 μm^2 A_{CONT} = 1 μm^2 W_B = 4 μm L_B = 8, 12, 15, 20, 24, or 28 μm
B			A_{ACT} = 1236 μm^2 A_{CONT} = 1 μm^2 W_B = 4 μm L_B = 8, 12, 15, 20, 24, or 28 μm

C			A_{ACT} = 1341 µm^2 A_{CONT} = 1 µm^2 W_B = 4 µm L_B = 8, 12, 15, 20, 24, or 28 µm
D			A_{ACT} = 335 µm^2 A_{CONT} = 0.36 µm^2 W_B = 2 µm L_B = 8, 12, 15, 20, 24, or 28 µm

4.2.2 Key Design Parameters

As mentioned in section 3.5.1, the critical parameters of a logic relay are: contact adhesive force F_{adh}, as-fabricated contact gap thickness g_d, as-fabricated actuation gap thickness g_0, actuation area A_{ACT}, structural poly-$Si_{0.4}Ge_{0.6}$ layer thickness t, effective spring constant of the suspension beams k_{eff}, body bias voltage V_B, and operating voltage V_{DD}. Design considerations for these parameters are briefly discussed below:

A. F_{adh} & A_{CONT}

To first order, adhesive force is proportional to real contact area A_C (see Eq. (2.20)). Typically, A_C is only a fraction of the apparent contact area A_{CONT} since physical contacts are made only at certain asperities rather than over the entire area A_{CONT}. Fig. 4.1 shows a piecewise model for the relation between A_C and A_{CONT} [5]. As one can see, in the multi-asperity contact region (A_{CONT} > 100 nm), A_C/A_{CONT} is very small. Since A_C/A_{CONT} is approximately constant in this region, decreasing A_{CONT} is an effective way to decrease F_{adh} hence hysteresis voltage, which in turn allows for lower V_{DD} and improved energy efficiency. However, decreasing contact area would also increase contact resistance, which is not desirable. Another drawback of aggressively

minimizing A_{CONT} comes from fabrication challenge (*e.g.,* to etch and re-fill small vias) [6] and possible degradation of contact reliability (smaller contact is more mechanically fragile and more susceptible to welding-induced failure). In designs A-C, A_{CONT} = 1 μm^2; in the scaled design D, A_{CONT} = 0.36 μm^2.

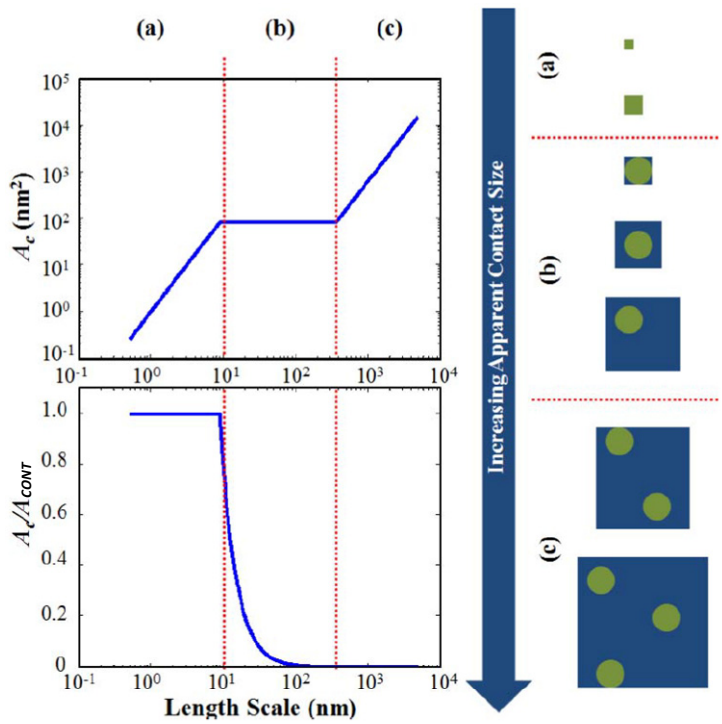


Fig. 4.1. Piecewise model of A_C/A_{CONT} with three regions: (a) full contact, (b) single asperity, and (c) multi-asperity. In the diagrams on the right side, the circles represent asperities that contact the opposing surface. Reprinted from [5].

B. g_0 & g_d

g_0 and g_d are critical design parameters. g_d, together with F_{adh}, determines the minimum relay switching energy (see Table 2.1) and should be as small as possible. As will be discussed in section 4.3, both g_0 and g_d are formed by selectively removing sacrificial material (*e.g.* SiO_2) between the Gate and Body electrodes. Small g_d and g_0 impose challenges on this so-called "release process" and can easily lead to stiction-induced relay failure (*e.g.* the channel gets stuck onto the Source/Drain electrode due to capillary forces). g_d = 60 nm is used in this work to ease this fabrication challenge. The nominal g_0 is set to 220 nm, because g_0 should be larger than $3g_d$ for minimum V_{DD}, as discussed in section 3.5. (Due to out-of-plane deflection caused by a non-zero strain gradient in the poly-$Si_{0.4}Ge_{0.6}$ film, both g_d and g_0 increase – by up to 20 nm – upon release. The nominal g_0 is set to ensure NPI-mode operation despite this.)

C. A_{ACT}

According to section 2.4, both turn-on voltage V_{ON} and hysteresis voltage V_H are inversely proportional to $\sqrt{A_{ACT}}$. Larger actuation area is helpful to reduce operation voltage. The upper limit of A_{ACT} is set by the maximum device footprint, which is dictated by the device density requirement. Designs A, B, and C have the same footprint but with incrementally increasing A_{ACT}. Design C is expected to result in the smallest values of V_H and V_{ON}.

D. k_{eff}

The effective spring constant affects many aspects of relay performance, including V_{ON}, V_{RL}, V_H, switching speed, switching energy, and reliability. There are often competing requirements for k_{eff} with respect to these performance parameters. (See chapter 2 for details.) Rules of thumb for body-biased relay design are: 1) k_{eff} should be large enough to avoid stuck-ON failure (*i.e.* $k_{eff} \gg F_{adh}/g_d$) and achieve small V_H; and 2) k_{eff} should not be so large as to make $|V_B|$ too large to be practical.

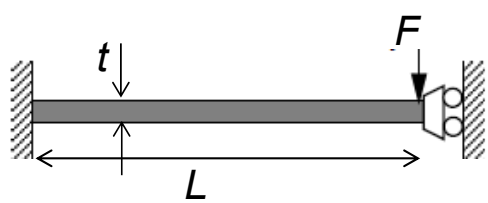


Fig. 4.2. A concentrated-load, guided-end beam [7].

According to Euler-Bernoulli beam theory [8], the spring constant k of a concentrated-load guide-end beam (Fig. 4.2) is

$$k = \frac{Ewt^3}{L^3}, \tag{4.1}$$

where E is the Young's modulus of the beam material, w is beam width, t is beam thickness, and L is beam length. For qualitative estimation, k_{eff} of the relay design in this work can be derived by modeling the mechanical structure as four such beams in parallel, so

$$k_{eff} \propto \frac{Ewt^3}{L^3}. \tag{4.2}$$

In addition to bending, torsional and stretching terms can be added to result in a more elaborate expression for k_{eff} [9]. Interested readers are encouraged to read [10,11]. As one can see from Eq. (4.2), k_{eff} is a strong function of L, t, and w. To achieve a suitable k_{eff}, six experimental splits for L were included for each relay design. The beam width w of design D is one half that of the other designs.

E. t

In this work, two experimental splits (1.5 μm and 1.75 μm) were included for the structural layer thickness t. It affects not only k_{eff} (see Eq. (4.2)) but also the total mass of the movable structure. To the first order, τ_{ON} is determined by the fundamental oscillation frequency of the movable structure [12], *i.e.*

$$\tau_{ON} \propto \sqrt{\frac{m_{eff}}{k_{eff}}}, \tag{4.3}$$

where m_{eff} and k_{eff} are the effective dynamic mass and the effective spring stiffness, respectively. Substituting k_{eff} with Eq. (4.2) and m_{eff} with $m_{eff} \approx m = \rho St$ (where ρ is the structural material density and S is the movable plate area), it can be shown that $\tau_{ON} \propto 1/t$. Therefore, increasing t is beneficial to decrease turn-on delay.

F. V_B & V_{DD}

The optimal values of the operating voltages V_B and V_{DD} are determined by other parameters (see section 3.5). Generally, larger A_{ACT}, smaller F_{adh}, and smaller g_d are preferred to minimize both V_{DD} and $|V_B|$. However, k_{eff} has opposite effects on V_B and V_{DD}: increasing k_{eff} allows for V_{DD} to decrease by increasing $|V_B|$.

4.3 Simulation

4.3.1 Process Emulation

The fabrication process was emulated by CoventorWare (a finite-element-method simulator) [13] using the layouts (as in Table 4.1) which were designed using Cadence Virtuoso [14]. The process recipes and materials used were specified in Process Editor and Material Properties Editor, respectively. The material and thickness used for each layer is listed in Table 4.2. With these, CoventorWare can emulate the fabrication process and build up a 3-D relay structure. Fig. 4.3 shows an emulated 3-D structure of relay design C.

TABLE 4.2
MATERIALS AND FILM THICKNESS OF EMULATED RELAYS

Layer	Material	Thickness
Substrate insulation	Al_2O_3	50 nm
Electrode (Source, Drain, Gate)	W	50 nm
1st sacrificial layer	SiO_2	60 nm
2nd sacrificial layer	SiO_2	160 nm
Channel	W	60 nm
Channel-to-Body insulation oxide	Al_2O_3	50 nm
Structure	Poly-$Si_{0.4}Ge_{0.6}$	1.75 μm

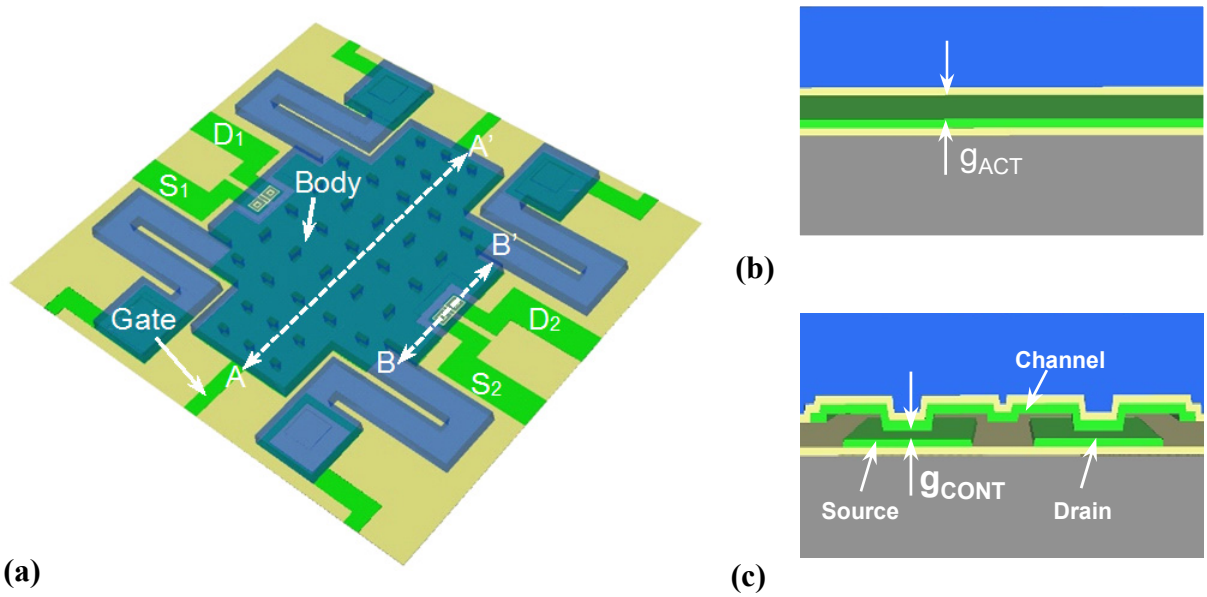
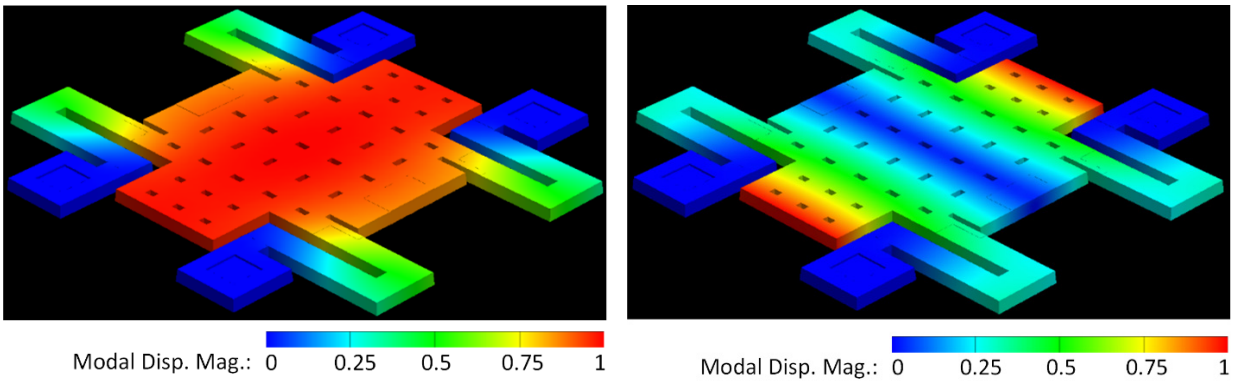


Fig. 4.3. Emulated 3-D structure of relay design C. (a) 3-D structure of a 6-terminal logic relay, comprising two pairs of source/drain electrodes (S_1/D_1 and S_2/D_2), one body electrode (translucent blue region), and one gate electrode (green region under the body electrode). The movable body is suspended over the fixed gate electrode by four flexure-suspension beams. (b) Cross-sectional view along cut-line AA'. A parallel-plate capacitor is formed by the body and gate electrodes with an air gap of thickness g_{ACT} and a thin layer of Al_2O_3 dielectric layer in-between. (c) Cross-sectional view along cut-line BB'. A strip of tungsten is attached to the underside of body electrode via the Al_2O_3 insulation layer, acting as a current-conducting channel between the source and drain when the relay is turned on. In the OFF state, the air gap between source/drain and the channel in the dimpled contact regions is g_{CONT}.

4.3.2 Device Simulation

4.3.2.1 Modal Analysis

Modal analysis of the above 3-D model was done by using the Analyzer module of CoventorWare software. The 1st and 2nd vibration modes in the mechanical domain are shown in Fig. 4.4. Note that the 1st mode has a frequency of about 942 kHz. As a rough estimation, the turn-on delay τ is approximately ¼ of the period of the 1st mode, so τ is approximately 0.27 μs.



(a) 1st mode. f = 941.7 kHz. (b) 2nd mode. f = 1.39 MHz.

Fig. 4.4. Modal analysis in mechanical domain. (a) 1st mode; (b) 2nd mode.

4.3.2.2 Structural Analysis

Upon application of a Gate-to-Body voltage V_{GB}, the movable Body plate is actuated downward to the extent that either F_{spring} balances F_{elec} or the Channel makes contact with Source/Drain electrodes. Fig. 4.5 shows a color contour map of the body plate displacement in the ON state, wherein V_{GB} is just high enough to turn on the relay (*i.e.* g_{CONT} drops to zero), simulated using ConventorWare. As one can see, maximum displacement occurs at the center of the body plate, with a value of 69 nm, which is 15% more than the displacement at the dimpled contact (*i.e.* 60 nm). The deformation of the rectangular body plate is fairly small, and 85% of the bending takes place in the suspension beams, which ensures that the relay can be turned on correctly (*i.e.* the dimpled contacts can make contact with source/drain electrodes before the center of the body plate contacts the underlying gate electrode). This is realized by designing the beams to have much lower effective spring constant (narrow width) than the relatively rigid body plate. The small deformation of the body plate also justifies the parallel plate approximation.

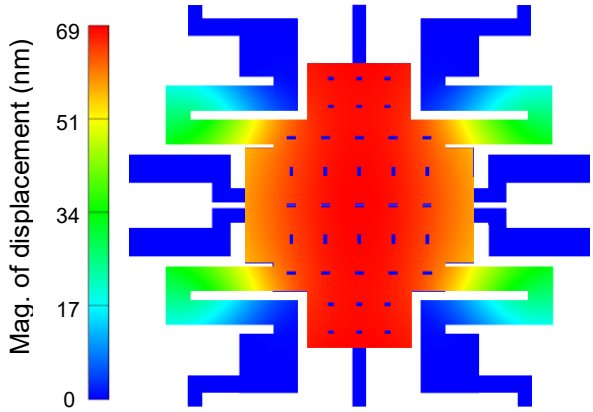


Fig. 4.5. Simulated displacement of the Body plate of for relay design C in the ON state ($V_{GB} = V_{ON}$). Most of the deformation happens in the flexure beams. The body plate is relatively flat.

4.3.2.3 Turn-on Voltage

Simulation methods for determining turn-on voltage V_{ON} are different for NPI-mode *vs.* PI-mode relays. For a NPI-mode relay, V_{ON} can be found by gradually increasing V_{GB} until a value at which the contact air gap g_{CONT} in the stable state reaches zero[i]. (For each value of V_{GB}, the movable plate stabilizes to a certain position and the corresponding g_{CONT} can be simulated. The minimum V_{GB} at which g_{CONT} drops to zero is the turn-on voltage.) However, for a PI-mode relay, once V_{GB} is increased to cause the relay to enter into the highly nonlinear positive feedback region of operation (*e.g.* such that the actuation gap size g_{ACT} is less than $1/3(g_0)$, for the parallel-plate capacitor model), the simulation does not converge. To solve this problem, CoventorWare Analyzer has a specialized analysis option called *Detect Pull-In* which can handle the pull-in-induced divergence. The idea is as follows. If V_{GB} is so high (called V_{GB_high}) that the relay enters into the divergence region, it takes a step back by reducing V_{GB} by a certain step size (user sets) and performs the simulation again with the new V_{GB} (called V_{GB_new}). If V_{GB_new} doesn't result in divergence, it is concluded that $V_{GB\_new} < V_{ON} < V_{GB\_high}$. Another trial value between V_{GB_new} and V_{GB_high} can then be picked for simulation to further narrow down the range. By repeating the above process, the simulation tool can eventually find V_{ON} to a specified precision.

The body-biased relays used in this work operate in NPI-mode, so the first method was used for V_{ON} simulation. Fig. 4.6 shows the simulated V_{ON} values for the four designs A, B, C, and D with the same value of L_B = 15 μm. Note that designs A, B, and C have the same parameters except for actuation area; from A to C, A_{ACT} increases so that V_{ON} decreases, as predicted by the simplified analytical Eq. (2.10). Design D is a scaled version of design C, by a factor of 2 in lateral dimensions. Since k_{eff} scales down linearly with lateral dimension while A_{ACT} scales quadratically, V_{ON} increases from design C to design D. To maintain a constant V_{ON}, the structural layer thickness for design D should be scaled down as well.

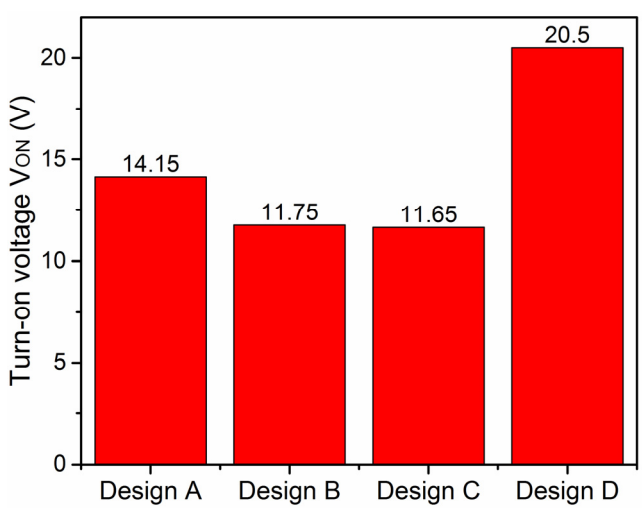


Fig. 4.6. Simulated V_{ON} values for the four relay designs A, B, C, and D with L_B = 15 μm.

[i] The V_{ON} simulated in this way is actually the quasi-static turn-on voltage.

Fig. 4.7(a) shows the effect of beam length on V_{ON}. As expected, longer beam length leads to smaller V_{ON} due to smaller k_{eff}. Fig. 4.7(b) shows a plot of V_{ON}-vs.-$L_B^{-0.5}$. Interestingly, one can see that in certain range of beam length, V_{ON} has a roughly linear dependence on $L_B^{-0.5}$. This provides guidance for relay beam length design.

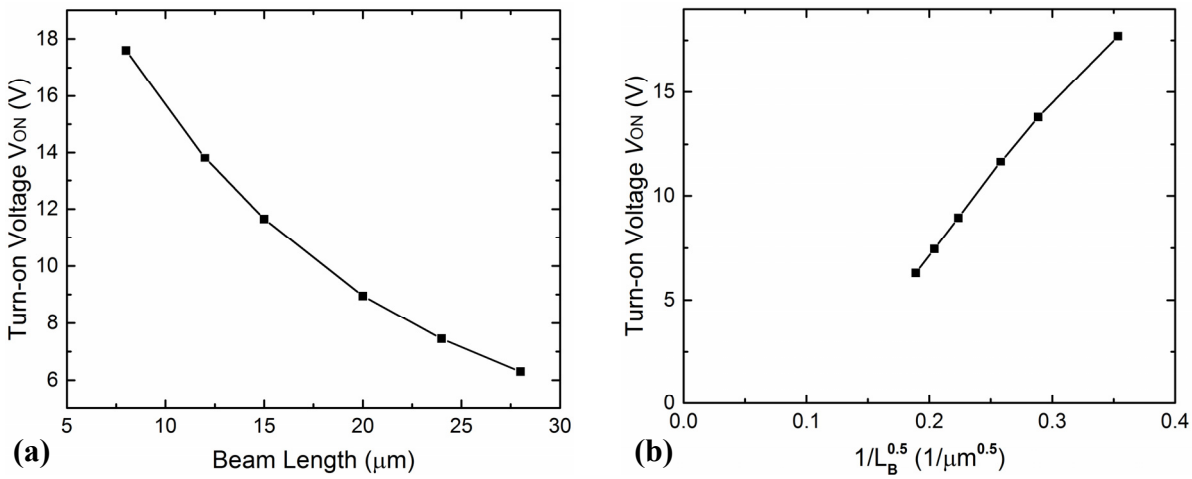


Fig. 4.7. (a) Effect of beam length on V_{ON}; (b) V_{ON} has a linear dependence on $L_B^{-0.5}$.

4.3.2.4 Hysteresis Voltage

As introduced in Chapter 2, the two sources contributing to hysteresis voltage V_H are the adhesive force between contacting surfaces and the pull-in operation mode. It is very challenging, if not impossible, to accurately simulate F_{adh} since it depends on many complicated factors including material properties, surface roughness, contact force, surface oxidation/contamination, capillary force, *etc*. Instead of simulation, F_{adh} is usually measured by atomic-force microscopy (AFM) [15] or extracted from measured values of V_H [6].

The other component of V_H originates from pull-in operation mode and can be eliminated by designing the relay to operate in non-pull-in mode. To see how significantly PI-mode operation contributes to V_H, simulations of V_H for both PI-mode and NPI-mode relays were performed with F_{adh} ignored (*i.e.* assuming $F_{adh} = 0$). A relay of design C and a control relay with exactly the same parameters except $g_d = 100$ nm ($g_d > 1/3g_0$ to ensure pull-in operation) were simulated. Fig. 4.8 shows how the displacement in the dimpled contact region changes with the actuation voltage V_{GB}, for both relays. For the control device (PI-mode relay), the displacement increases smoothly as V_{GB} increases to approximately 11.1 V, after which the relay enters the PI region of operation wherein the structure snaps down to close the contact gap and turn on the relay. When V_{GB} is subsequently decreased, the contacts are not broken (*i.e.* the relay doesn't turn off) until V_{GB} falls below 10.3 V, indicating a hysteresis voltage of 0.8 V. (Note that this hysteresis is solely due to pull-in mode operation, since F_{adh} was set to zero in the simulation.) For the non-pull-in mode relay, the displacement-*vs*.-voltage curve for turn-on and turn-off traces are exactly the same, as there is no hysteretic switching effect. Clearly, PI-mode operation mode itself contributes significantly to V_H. To achieve ultra-low voltage operation, then, relays should be designed to operate in NPI mode.

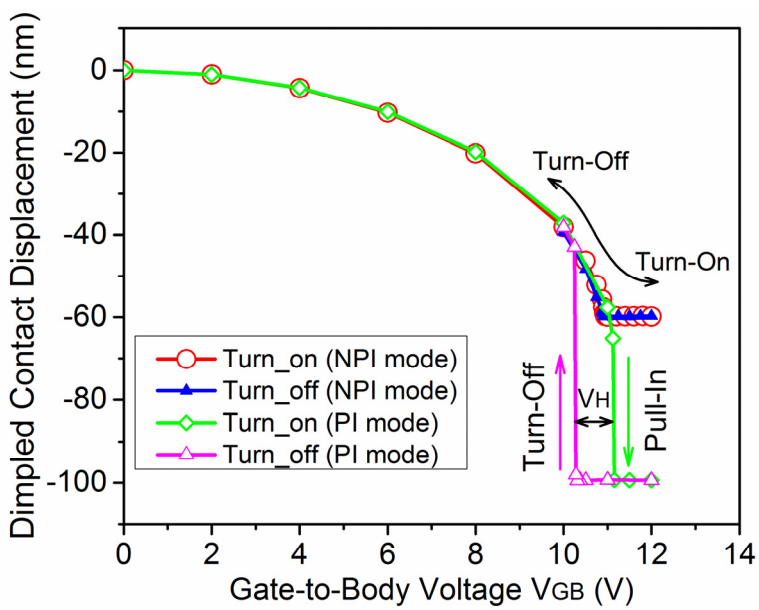
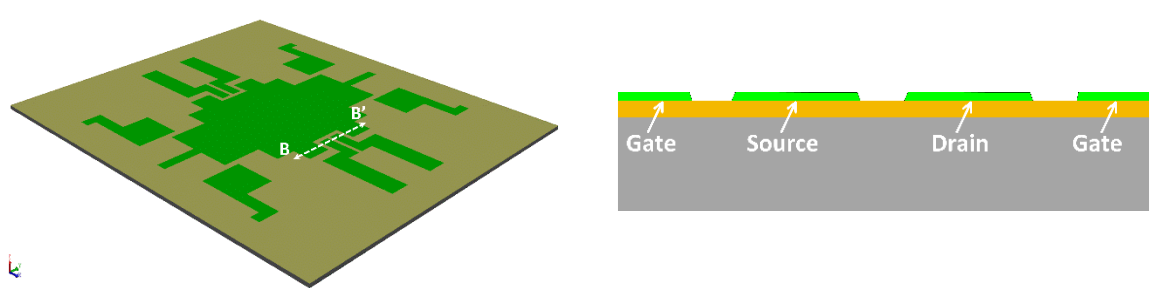


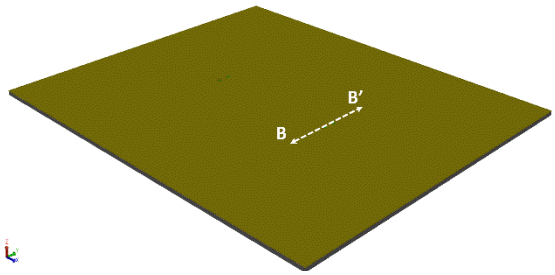
Fig. 4.8. Simulated displacement in the dimpled contact regions as a function of the applied actuation voltage, for pull-in (PI) mode and non-pull-in (NPI) mode relays. F_{adh} is ignored in these simulations. During the turn-on process of a PI-mode relay, the displacement increases smoothly with increasing V_{GB} until the relay enters into the pull-in region of operation, after which the contact gap closes to zero (green open-diamond curve). To turn off the relay, V_{GB} must be reduced below the turn-on voltage to cause the contacts to open (pink open-triangle curve). For the NPI-mode relay, the turn-off curve (blue solid-triangle curve) traces the turn-on curve (red open-circle curve), so there is no hysteresis.

4.4 Relay Fabrication Process

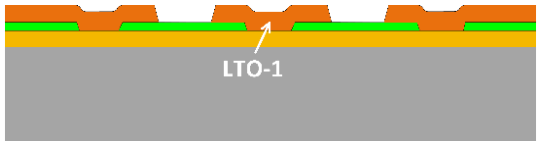
One of the main advantages of electrostatic actuation over other actuation mechanisms used in miniature mechanical switches suitable for integrated circuit applications is its relatively simple actuator structure (an air-gap capacitor). The process for fabricating 6-terminal logic relays, illustrated in Fig. 4.9 (comprising bird-eye views of 3-D relay structure and cross-sectional views along the cut-lines BB' and CC'), involves neither exotic materials nor very high process temperatures and therefore can be used to fabricate relays integrated with CMOS circuitry on the same substrate.



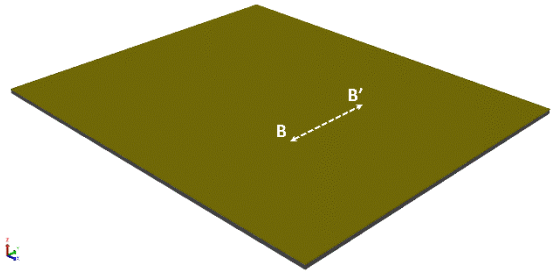
(A) (a)


B'
B

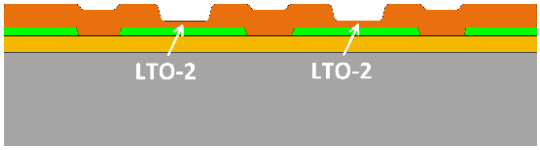
(B)


LTO-1

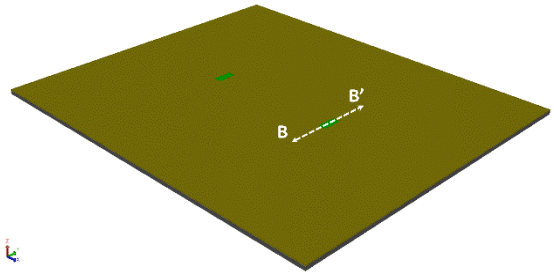
(b)


B'
B

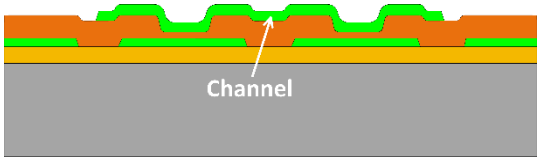
(C)


LTO-2
LTO-2

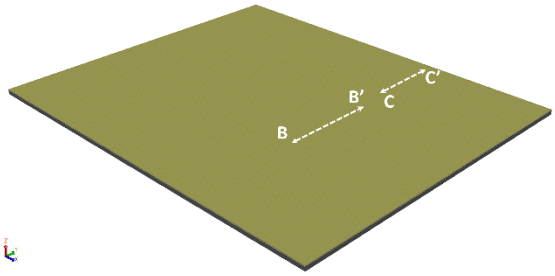
(c)


B'
B

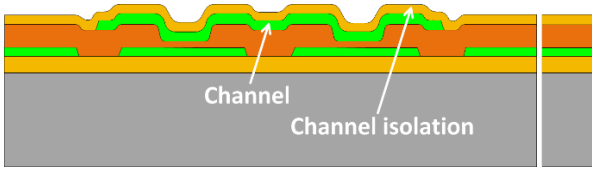
(D)


Channel

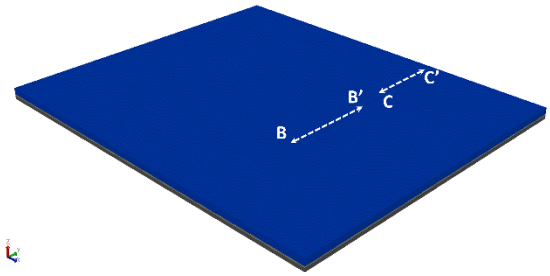
(d)


C'
B'
C
B

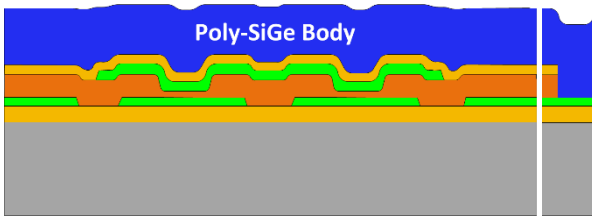
(E)


Channel
Channel isolation

(e)


C'
B'
C
B

(F)


Poly-SiGe Body

(f)

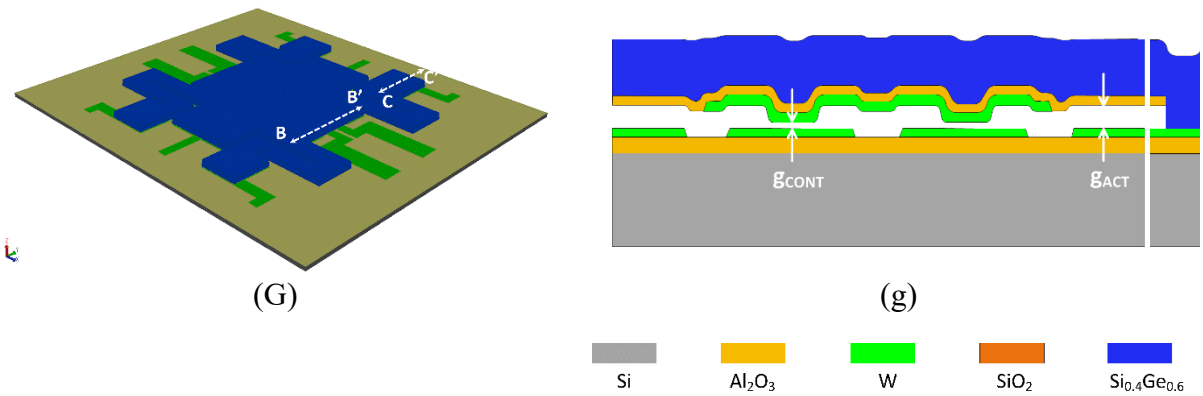


Fig. 4.9. Schematic illustration of the fabrication process for a 6-T logic relay. (a)-(d) and the left side of (e)-(g) are cross-sectional views along the cut-line BB' in (A)-(G) respectively; the right-most sides of (e)-(g) are cross-sectional views along the cut-line CC'.

A. Substrate Insulation

The fabrication process described herein is based on standard silicon (Si) surface micromachining technology. A prime-quality 6-inch p-type (Boron doped to resistivity ~20 Ω•cm) Si wafer of <100> crystalline surface orientation is used as the substrate, which serves as a physical platform on which relays are going to be built. Other substrates could be used so long as they are compatible with standard planar fabrication processes. Since the Si substrate is electrically conductive, an insulating layer is needed to electrically isolate it. As hydrofluoric acid (HF) is used to release the relays at the end of the fabrication process, the insulating material should be resistant to HF. In this work, Al_2O_3 is used.

An 80 nm-thick Al_2O_3 layer was deposited onto the bare Si wafer by atomic layer deposition (ALD) at 300 °C. The precursors used were trimethylaluminum (TMA) and vapor H_2O. The deposition rate was about 1 Å per cycle (*i.e.* TMA pulse → N_2 purge → H_2O pulse → N_2 purge) which was roughly 9 s in the Picosun ALD System[i]. Since the ALD deposition rate is relatively slow, a bilayer comprising a thermally-grown or chemical-vapor-deposited layer of SiO_2 (which is not resistant to HF) and a thinner capping layer of Al_2O_3 could be used for substrate insulation.

B. Bottom Electrodes Formation

The choice of bottom electrode material is very critical, since it is used not only for electrical-signal routing wires, but also for the Source/Drain electrodes whose surface properties determine contact resistance, adhesive force, and reliability. An ideal electrode material (a) is electrically conductive for low contact resistance; (b) provides for low contact adhesive force (between two pieces of the same material, as the channel is usually made from the same material); (c) is resistant to welding, mechanical wear and material transfer, chemical oxidation and surface contamination [16]; and (d) is compatible with CMOS processing. Unfortunately these requirements are difficult to meet simultaneously. In this work, W is used because of its superior hardness, high melting point, and good compatibility with CMOS processing [17].

[i] System dependent. The cycle length can be adjusted as needed.

A 50 nm-thick tungsten layer was deposited by pulsed-DC sputter deposition on top of the Al_2O_3 coated Si wafer directly after the ALD process[i]. The deposition chamber was kept at room temperature, but the temperature on the wafer surface might have been higher due to self-heating during the sputtering process, as there was no feedback temperature control system. The base pressure (chamber pressure before flowing the sputtering gas Ar) was 0.5 μTorr. Such a high vacuum is necessary to ensure a high-quality W film. Base pressure higher than 1 μTorr is generally not acceptable since significant oxidation may happen in the presence of H_2O vapor. The DC source power used was 1 kW and the sputtering pressure (pressure of Ar gas during plasma sputtering process) was 8 mTorr. Each of these process parameters has significant impact on residual stress in the sputtered-W film[ii] [18-20]. The measured average sheet resistance of the deposited 50 nm W film was 3.6 Ω/□ with standard deviation 0.26 Ω/□; the intrinsic film stress was compressive, with an average value of ~540 MPa.

The blanket W film was then patterned by photolithography and reactive ion etching (RIE) with SF_6 plasma to form the Source/Drain/Gate electrodes and routing wires (Fig. 4.9(a)). Note that the area of the gate electrode is much larger than that of the source/drain electrodes. The photoresist masking layer was then removed by ashing in oxygen plasma at 250 ºC and the wafer was further cleaned by soaking in PRS-3000 photoresist stripper[iii] at 80 ºC. (Note that wafer must be dry before going into PRS-3000 bath, otherwise W may be etched because there exists a small amount of caustic base[iv].)

C. Sacrificial Oxide Deposition and Contact Region Formation

The sacrificial material used in this work is low-temperature deposited silicon dioxide (LTO) because it can be easily and selectively removed in HF vapor. There are two layers of sacrificial LTO used in the relay fabrication process.

The first layer (160 nm thick) of sacrificial oxide (LTO-1) was deposited by low pressure chemical vapor deposition (LPCVD) at 400 ºC with SiH_4 (90 sccm flow rate) and O_2 (135 sccm flow rate) precursor gases at 300 mTorr. The deposition rate was about 10 nm/min. The contact regions were then defined by patterning the LTO-1 layer (using lithography and RIE) to expose the underlying the source/drain electrodes (Fig. 4.9 (b)). The RIE was carried out in the Centura MXP tool, a magnetically enhanced RIE etcher from Applied Materials, Inc., with a gas mixture of 10 sccm CF_4, 50 sccm CHF_3, and 120 sccm Ar, at a chamber pressure of 200 mTorr. The etching rate was about 4.6 nm/s.

Afterwards, LPCVD was used again to deposit the second layer (60 nm thick) of sacrificial oxide (LTO-2), whose thickness determines the nominal as-fabricated contact gap size g_{CONT}.

---

[i] Cleanliness of the Al_2O_3 surface is of utmost importance. Avoid any contamination if possible. Delamination was observed for W film sputtered on Al_2O_3-coated Si wafer which went through 15 minutes soak in PRS-3000 photoresist stripper, de-ionized wafer rinse, and N_2 gas dehydration immediately before the sputtering, although it was not systemically verified. Use with caution!

[ii] Large intrinsic film stress can result in film delamination; it can also bow the Si wafer such that the bent wafer may not be able to be processed through the remaining steps. For example, typically wafers with bow > 50 μm cannot be handled by the lithographic exposure tool (stepper).

[iii] J.T.Baker® PRS™-3000 from Avantor Performance Materials company.

[iv] This is usually hard to control in a shared research lab environment, in which case it is recommended to avoid long time soaking and to check the wafer status frequently. A freshly filled PRS-3000 bath with water is more caustic.

The total thickness of the LTO-1 and LTO-2 layers determines the nominal as-fabricated actuation gap size g_{ACT}, as shown in Fig. 4.9 (c).

D. Current-conducting Channel Formation

To avoid the formation of a built-in electric field due to charge transfer upon contact, which would enhance contact adhesion, the material used for the channel should be the same as for the Source/Drain electrodes. The process integration of W channels is relatively straight forward, compared to Ru channels [21].

A 60 nm-thick W layer was deposited and patterned into small strips above the Source/Drain electrodes, as shown in Fig. 4.9 (d). Sputtering rather than evaporation was used for better step coverage. Depending on the aspect ratio and step height of the contact holes, a thicker W film may be needed. A thicker W Channel is actually beneficial for improving the robustness of the dimpled contacts. However, the W should not be too thick such that a significant amount of the underlying sacrificial LTO is removed during the etching process. Since the selectivity between W and LTO for the SF_6-based RIE recipe is not very high, significant over-etching of the LTO can occur during the W Channel patterning step, which would result in smaller as-fabricated actuation gap than the designed value.

E. Channel Insulation Oxide Deposition

To electrically insulate the Channel from the overlying Body electrode, an insulating Al_2O_3 layer (50 nm thick) was deposited by ALD (Fig. 4.9 (e)). Its thickness was set to ensure a sufficiently large breakdown voltage. As reported in the literature, the breakdown electric field of thermally-grown ALD Al_2O_3 usually falls in the range of 5-10 MV/cm [22 - 24]. Conservatively, 50 nm Al_2O_3 should be able to sustain 25 V Body-to-Channel voltage (V_{ON} can be as high as ~20 V as shown in Fig. 6). Note that the Al_2O_3 should not be thicker than necessary because the ALD process is slow, and (more importantly) the residual stress in a thick Al_2O_3 layer can result in undesirable out-of-plane deflection of the structure upon release. (See the next section for more discussion about stress mismatch issues.)

F. Structural Layer Deposition

Polycrystalline silicon-germanium (poly-$Si_{0.4}Ge_{0.6}$) was used as the structural material because of its excellent mechanical properties (comparable to those of poly-Si) and low deposition temperature, compatible with post-CMOS processing [25,26].

Before poly-$Si_{0.4}Ge_{0.6}$ deposition, contact holes to the bottom electrodes (see left side of Fig. 9(f)-(g)) were opened by etching through the top Al_2O_3 layer, LTO-2 layer and LTO-1 layer to expose the underlying tungsten electrodes in the anchor regions. The Al_2O_3 etching was done in Centura-MET, a DPS (decoupled plasma source) etcher specially designed for etching metal film by Applied Materials, Inc. The etching process used a gas mixture of 60 sccm BCl_3 and 30 sccm Cl_2 with chamber pressure 10 mTorr, RF power 1000 W, and bias power 100 W. The etch rate was about 1 nm/s.

The structural layer was then deposited by a two-step LPCVD process. The first step is to deposit an ultra-thin (~5 nm-thick) Si seed layer with Si_2H_6 (100 sccm) as the precursor gas at 410 ^{o}C and 300 mTorr pressure for 5 minutes. This layer affects the microstructure and hence the stress gradient of the following poly-$Si_{0.4}Ge_{0.6}$ layer [25]. The boron-doped poly-$Si_{0.4}Ge_{0.6}$ was

then deposited by using SiH_4 (140 sccm), $GeHe_4$ (60 sccm), and BCl_3 (45 sccm of 1% v/v BCl_3/He) precursor gases at 410 °C and 600 mTorr, in the same LPCVD furnace without breaking vacuum. It took about 5 hours to deposit 1.75 μm-thick poly-$Si_{0.4}Ge_{0.6}$ (Fig. 4.9 (f)).

It should be noted that the deposition of poly-$Si_{0.4}Ge_{0.6}$ is one of the most critical steps in the entire relay fabrication process. The average residual stress and stress gradient within the as-deposited film has enormous impact on relay performance. Fig. 4.10 shows two potential failure modes of relays due to stress gradient. If there is a negative stress gradient along the vertical direction (from the bottom to the top of the film), the structure would bend as in (a), where the contact gap size g_{CONT} is much smaller than the designed value. In the worst case, the g_{CONT} is zero so the relay is stuck ON as-fabricated. In the case of positive stress gradient as shown in Fig. 4.10 (b), the downward curvature increases g_{CONT} such that it may be larger than g_{ACT}, in which case the relay will fail to turn on because the center of the Body electrode makes contact with the bottom Gate electrode, preventing further actuation.

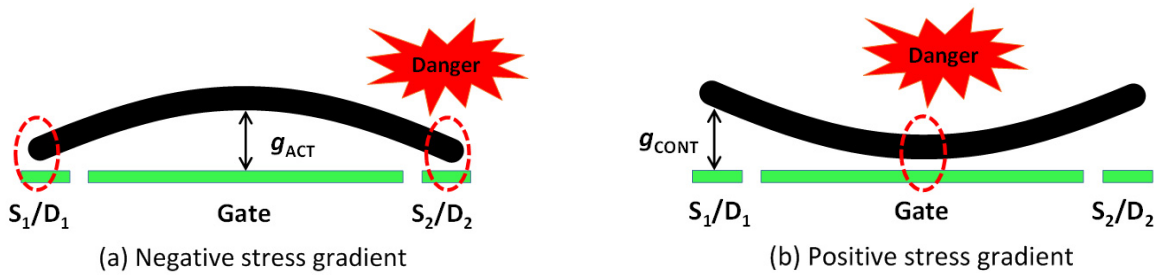


Fig. 4.10. Two potential failure modes due to stress gradient within the structural layer. (a) Negative stress gradient. Relay may end up stuck ON as fabricated if g_{CONT} drops to zero. (b) Positive stress gradient. g_{CONT} may be larger than g_{ACT}, in which case the relay would not be able to turn on before the Body makes contact with the Gate.

The residual stress of the poly-$Si_{0.4}Ge_{0.6}$ film is affected by its microcrystalline structure, which depends on the deposition process conditions such as temperature, chamber pressure, precursor gas type and flow rate, doping species and concentration, *etc.* [25,27,28]. The microcrystalline structure usually varies along the thickness direction of LPCVD poly-SiGe, which results in varying stress with depth and hence a stress gradient along the depth direction. Fig. 4.11 shows the crystalline structure and stress depth profile for films deposited under various conditions [26]. Clearly the crystalline structure can change dramatically with the deposition recipe, and from the lower portion to the upper portion of the film. The lower portion (near the interface with the underlying substrate) is amorphous and has larger compressive stress, whereas the upper portion is polycrystalline. The crystal grains (columnar in Fig. 4.11(b) *vs.* conical in Fig. 4.11(c)) also affect stress gradient significantly. Conical gains are generally not preferred [29]. A thicker film has less overall stress gradient because the effect of the amorphous lower portion is averaged out.

It should be noted that the intrinsic stress of the Al_2O_3 channel isolation layer should be taken into account when optimizing the stress gradient in the structural layer. Fortunately, ALD Al_2O_3 generally has tensile stress, and can therefore compensate the compressive stress of poly-SiGe.

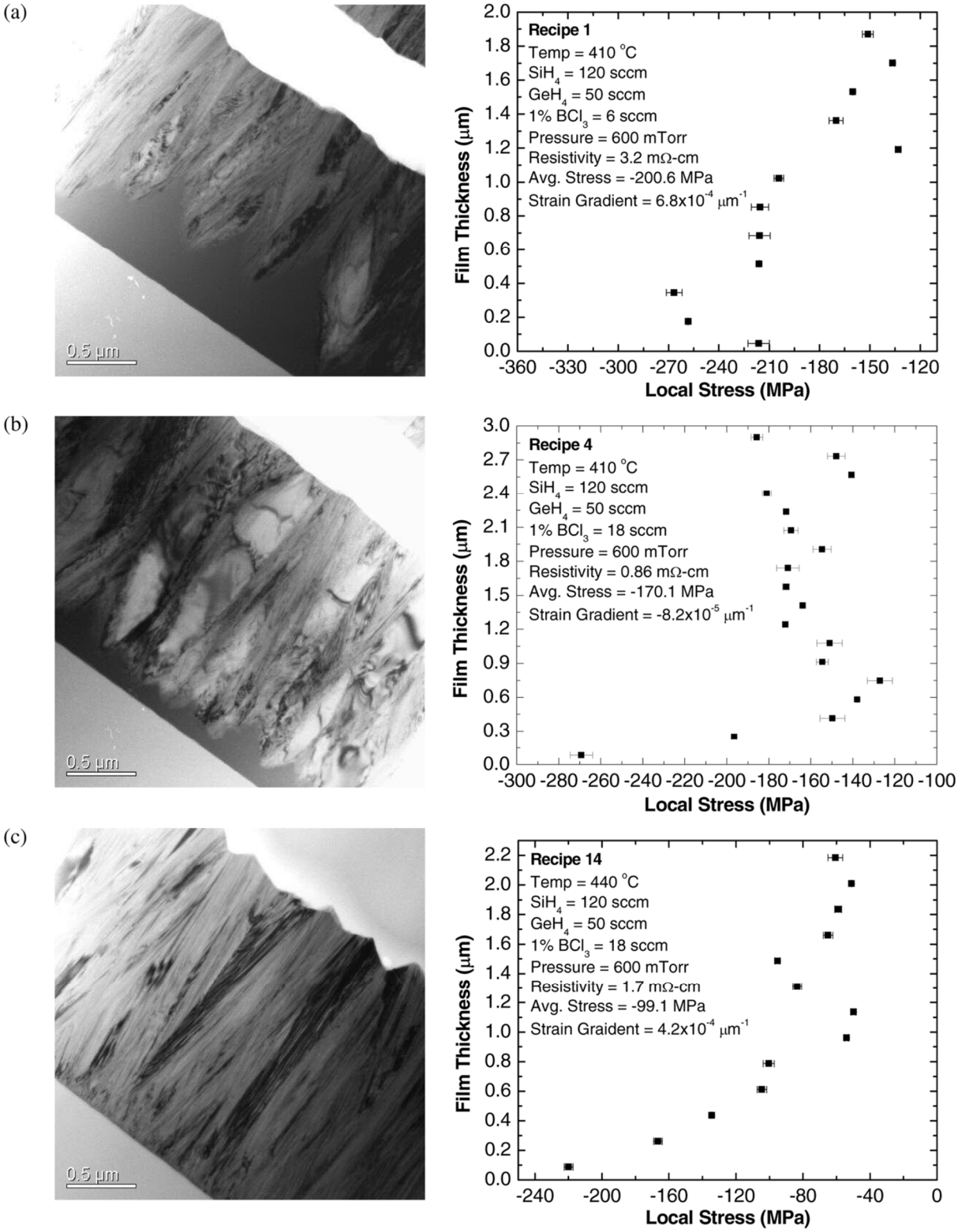


Fig. 4.11. Cross-sectional TEM images and stress depth profiles for films deposited with different recipes. Reprinted from [26].

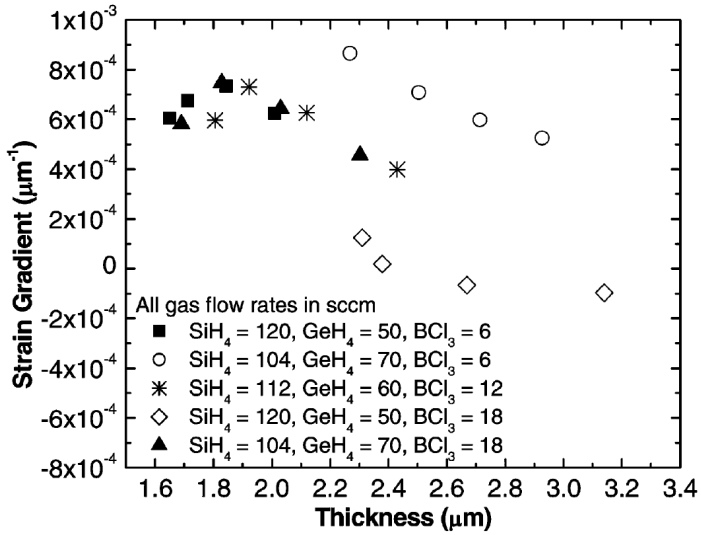


Fig. 4.12. Effect of film thickness on strain gradient of poly-SiGe. Reprinted from [26].

G. Structure Patterning

A 400 nm-thick LTO layer deposited by LPCVD was used as a hard mask for patterning the thick structural layer. The LTO hard mask, poly-$Si_{0.4}Ge_{0.6}$ layer, and the Body-insulating Al_2O_3 layer were patterned by lithography and RIE, leaving the sacrificial LTO-2 layer exposed. The etching of poly-$Si_{0.4}Ge_{0.6}$ was done using a TCP (Transformer Coupled Plasma) RIE etcher from Lam Research with etching gases HBr (150 sccm) and Cl_2 (50 sccm), chamber pressure 12 mTorr, TCP RF power 300 W, and bias RF power 40 W. The etching of hard-mask LTO and Body-insulating Al_2O_3 is the same as in step (C).

H. Structural Release

The final step was to remove the sacrificial LTO-1 and LTO-2 layers to release the movable structure. To avoid capillary-force induced stiction, ethanol-assisted anhydrous HF vapor was used to selectively and isotropically etch away the sacrificial layers. The released relay structure is illustrated in Fig. 4.9 (g). SEM images of two fabricated relays (from design C and D) are shown in Fig. 4.13.

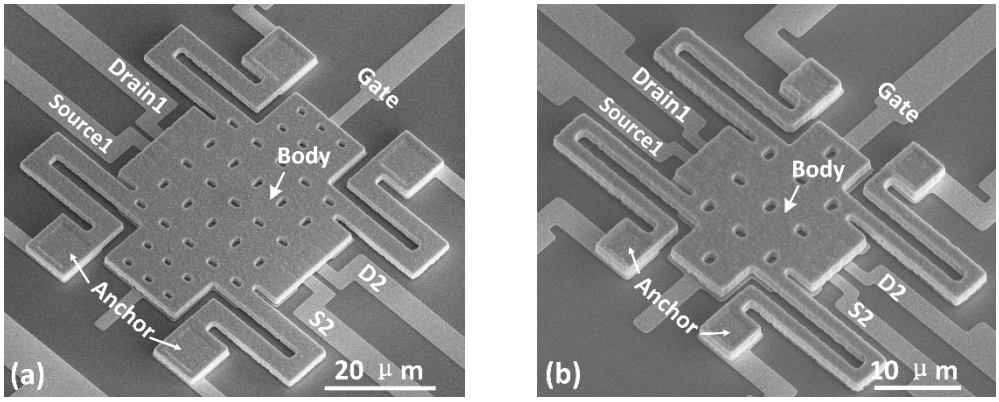


Fig. 4.13. SEM images of two fabricated relays with L_B = 15 μm: (a) Design C; (b) Design D.

4.5 Relay Performance Characterization

Fabricated relays were electrically characterized in vacuum (0.5 μTorr) at room temperature (~ 20 ºC).

4.5.1 Switching Voltages

Fig. 4.14 shows measured switching voltages (*i.e.* V_{ON}, V_{RL}, and V_H) for fabricated relays of different designs but the same beam length L_B = 15 μm. At least five randomly selected devices of each design were characterized, and each device was tested more than five times. The average values are labeled in the histogram. The length of the error bar equals two standard deviations. Since $V_{ON} \propto (k_{eff}/A_{ACT})^{1/2}$ and the nominal k_{eff} of designs A ~ C are the same, V_{ON} decreases from design A to C due to increasing actuation area. Design D is a scaled version of C (by 1/2 in lateral dimensions). Since $k_{eff} \propto W_B$, V_{ON}(D) should be roughly equal to $\sqrt{2}V_{ON}$(C), which is verified by the experimental results.

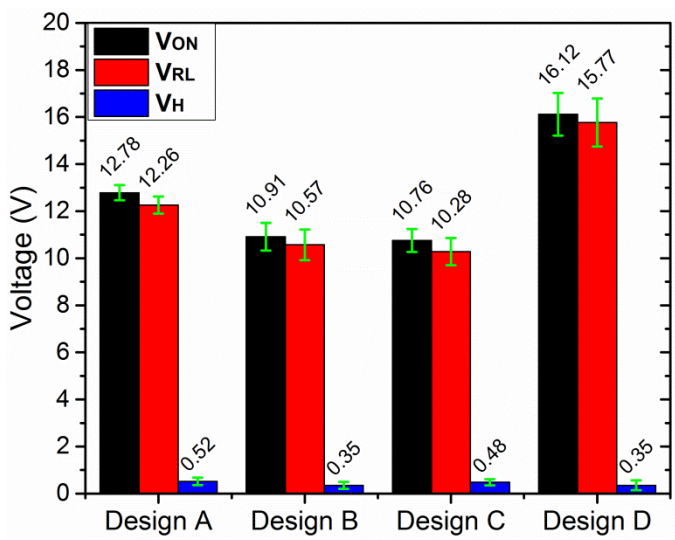


Fig. 4.14. Measured switching voltages for the different relay designs. L_B = 15 μm.

Fig. 4.15 shows a comparison of V_{ON} between simulated and experimental results. Although the exact values are different, the dependencies of V_{ON} on design parameters (*i.e.* A_{ACT} and W_B) show the same trend. The fact that the simulated switching voltages are consistently larger than the measured switching voltages may be due to the following reasons: a) the material properties (*e.g.* Young's modulus of poly-SiGe and dielectric constant of Al_2O_3) used for the simulations were not calibrated; b) the effect of the silicon substrate, which formed a parasitic capacitor with the poly-SiGe suspension beams, was not taken into account in the simulations; c) geometric parameter values used in the simulations were different than actual experimental values due to process variations; and d) g_{ACT} and g_{CONT} are different from their nominal values due to out-of-plane deflection upon structural release.

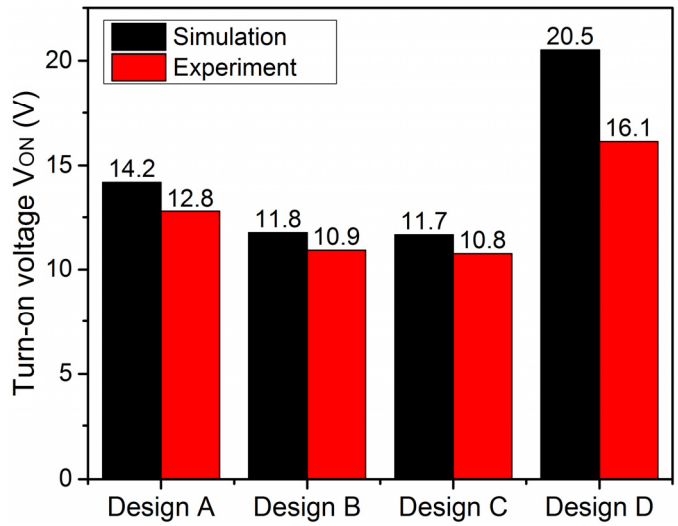


Fig. 4.15. Comparison of V_{ON} between simulated and experimental results.

The effect of beam length is also experimentally investigated. Fig. 4.16 shows measured switching voltages for relay design C as a function of beam length. As expected, longer beam length results in smaller V_{ON} and V_{RL}, but larger V_H. A comparison of simulated *vs.* experimentally measured V_{ON} values is shown in Fig. 4.17. Again, the trends match very well except that there is a systematic discrepancy in exact values, which can be explained by the reasons noted above.

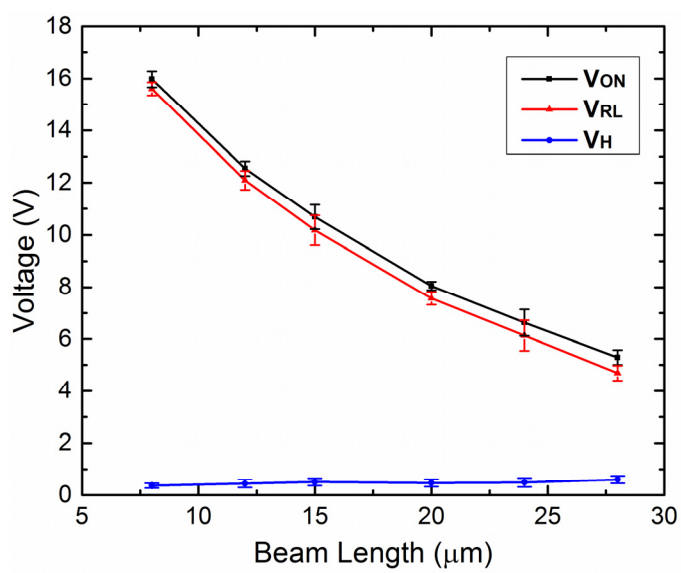


Fig. 4.16. Effect of beam length on relay voltage characteristics (design C).

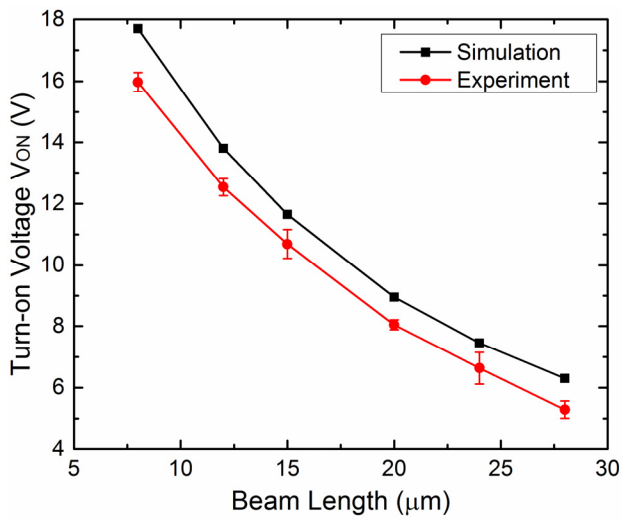


Fig. 4.17. Comparison of simulated *vs.* experimentally measured V_{ON} values.

4.5.2 *I-V* Characteristics

Measured current-*vs*.-voltage (I_{DS}-V_G) characteristics are shown in Fig. 4.18 for a relay operating similarly as an n-channel MOS (NMOS) transistor, *i.e.* turning on with positive Gate voltage sweep. The blue curve shows the *I-V* characteristic with zero body bias. It can be seen that I_{DS} changes abruptly between the OFF state (immeasurably low leakage current) and the ON state (current subjected to a compliance limit of 10 μA) at about 11.1 V. With a negative body bias, the gate voltage required to turn on/off the relay decreases significantly, shifting the *I-V* curve to the left (red curve). To observe the hysteresis effect more clearly, a zoomed-in *I-V* curve is shown in the inset. With Body bias voltage V_B = -11 V, the relay turns on at V_G = 92 mV and turns off at V_G = 22 mV, indicating a hysteresis voltage (V_H) of 70 mV. Considering that the applied voltage step size is 2 mV and ON/OFF ratio is greater than 7 decades, the local sub-threshold swing (*SS*) and overall equivalent *SS* is less than 0.3 mV/decade and 10 mV/decade, respectively.

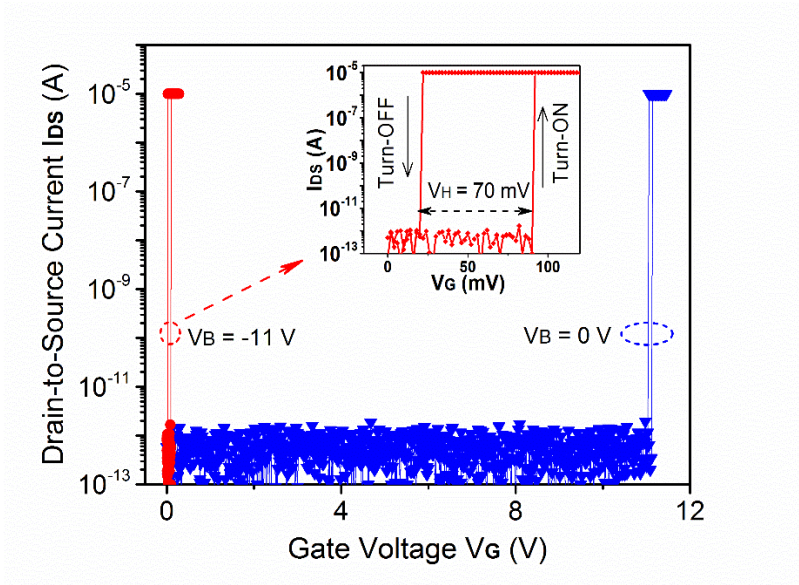


Fig. 4.18. Measured *I-V* characteristics of a fabricated 6-T logic relay (design C, L_B = 15 μm) body-biased so as to operate as a pull-down switch.

To mimic the behavior of a p-channel MOS (PMOS) transistor, which turns on with negative Gate voltage sweep, the applied Body bias voltage should be positive. Accordingly, Fig. 4.19 shows measured I_{DS}-V_G characteristics for V_B = 11.2 V. With a positive body bias, the relay can be turned off at high V_G and turned on at low V_G, similarly to the characteristic of a conventional p-type MOSFET. The hysteresis and on/off ratio of the measured p-relay are 62 mV and 10^7, indicating an equivalent SS < 9 mV/decade. The ground level of gate current I_g is mainly due to the charging/discharging displacement current of the Gate-to-Body parallel capacitor.

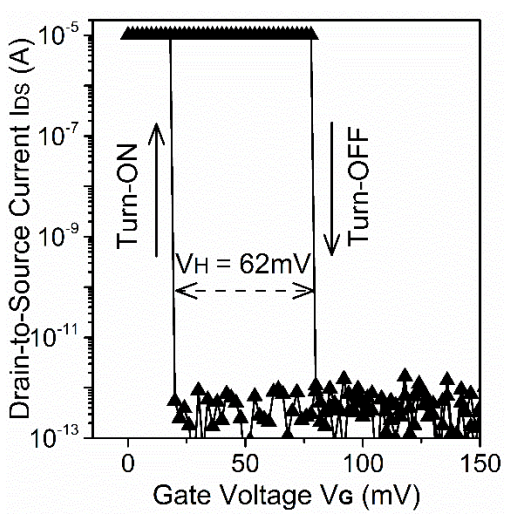


Fig. 4.19. Measured *I-V* characteristic of a fabricated 6-T logic relay (design C, L_B = 15 μm) body-biased so as to operate as a pull-up switch.

4.5.3 Millivolt Inverter Demonstration

In a CMOS digital logic circuit, NMOS transistors are used as "pull-down" devices (*i.e.* to pass a low voltage, 0 V, from the source to the drain), whereas PMOS transistors are used as "pull-up" devices (*i.e.* to pass a high voltage, V_{DD}, from the source to the drain). As intimated above, a logic relay can function either as a pull-down device or as a pull-up device, depending on the value of the Body bias voltage. To prove the viability of low-voltage relay-based digital logic circuits, we demonstrate herein low-voltage operation of a logic relay as a pull-down device and also as a pull-up device.

Fig. 4.20(a) shows the circuit schematic for an inverter in which the relay functions as a pull-down device and a load resistor (R_L = 140 kΩ) serves as a passive pull-up device. The output voltage (V_{OUT}) is monitored using an oscilloscope which has 1 MΩ input resistance (R_{osc}). As a result, V_{OUT} cannot reach V_{DD} when V_{IN} is low (*i.e.* when the relay is OFF) because a voltage divider is formed by R_L and the oscilloscope input resistance. The measured voltage waveforms in Fig. 4.20(b) show that, with V_B = -11 V, this relay-based inverter can be operated with V_{DD} = 0.2 V. The sloped rising edge of the V_{OUT} waveform is due to the "RC" delay associated with charging the capacitance of the output node through the load resistor. When V_{IN} goes high, the relay turns ON and the output node is discharged through the relay, causing V_{OUT} to drop. From the minimum value of V_{OUT} the ON-state resistance (R_{ON}) of the relay is calculated to be approximately 4.8 kΩ (See Eq. (3.1)).

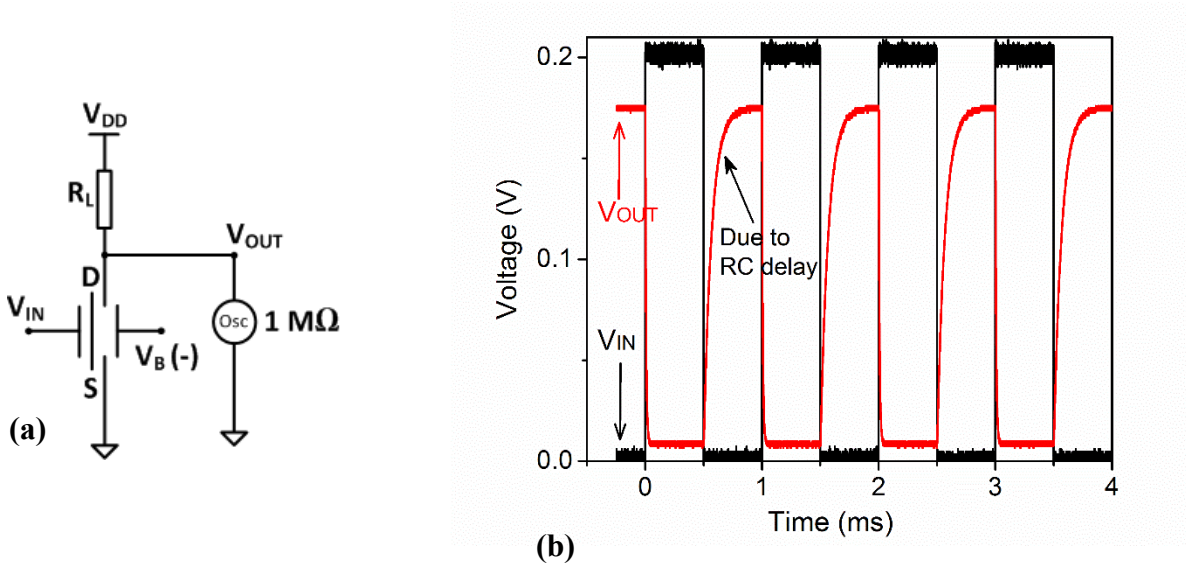


Fig. 4.20. (a) Schematic of an inverter circuit comprising a relay that functions as a pull-down switch. The output voltage is monitored by an oscilloscope which has 1 MΩ input resistance. (b) Measured voltage waveforms of the inverter in (a), with V_B = -11 V and V_{DD} = 0.2 V.

Fig. 4.21(a) shows the circuit schematic for an inverter in which the relay functions as a pull-up device and a load resistor (R_L = 140 kΩ) serves as a passive pull-down device. Again, the output voltage (V_{OUT}) is monitored using an oscilloscope which has 1 MΩ input resistance. The measured voltage waveforms in Fig. 4.21(b) show that, with V_B = 11.2 V, this relay-based inverter can be operated with V_{DD} = 0.2 V. From the maximum value of V_{OUT} the ON-state resistance (R_{ON}) of the relay is calculated to be approximately 4.2 kΩ (See Eq. (3.1)).

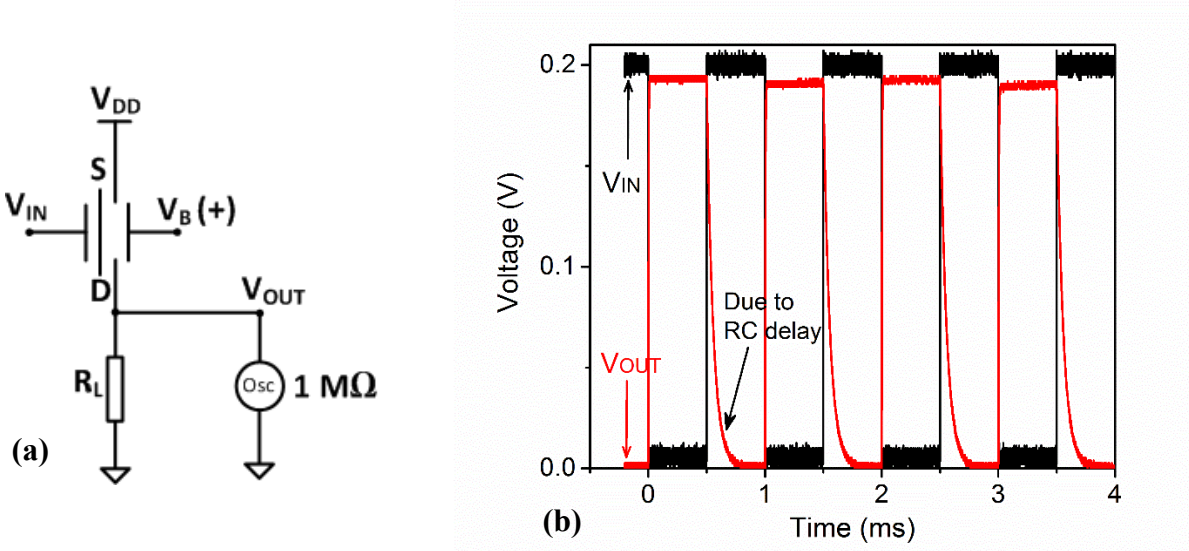


Fig. 4.21. (a) Schematic of an inverter circuit comprising a relay that functions as a pull-up switch. (b) Measured voltage waveforms of the inverter in (a), with V_B = 11.2 V and V_{DD} = 0.2 V.

The ability of the relay to operate with sub-100 mV voltage is demonstrated in Fig. 4.22. Compared to the inverter voltage waveforms in Fig. 4.20(b) where V_{DD} = 0.2 V, R_{ON} increases to 28 kΩ when V_{DD} is reduced to 80 mV (which explains why V_{OUT} doesn't drop to ground level when V_{DD} goes high). Fig. 4.23 shows that the relay-based inverter can operate at 50 mV, at the trade-off of increased R_{ON} (≈ 330 kΩ). The high R_{ON} is mainly attributed to oxidation (*e.g.* tungsten oxide) and contamination between W-W contact surfaces. It should be noted that during the fabrication process, Source/Drain tungsten electrodes were exposed to O_2 at 400 °C during the LTO sacrificial layer deposition steps, thus their surfaces was oxidized to some extent. The contact surfaces were further oxidized when they were exposed to air after the LTO sacrificial layer was removed in anhydrous HF vapor. At high operation voltage, R_{ON} is low because high V_{DD} helps to break down thin dielectric layers. High V_G also helps to generate larger contact force and thus larger intimate contact area [30], which in turn results in smaller R_{ON}. At low voltage, however, both effects diminish so that R_{ON} increases.

Note that the higher R_{ON} would not be problematic for implementing complementary logic circuits because in practice: a) the OFF-state resistance of a relay is infinite and the input impedance of a relay also is very large (much larger than the input impedance of the oscilloscope which is used to monitor the output voltage waveforms in this work) so that the output voltage swing would be V_{DD}, and b) the load capacitance for an integrated relay circuit would be much smaller than the inverter load capacitance in this work, so that the RC delay would be much smaller than the mechanical switching delay. For pass-gate logic circuits, smaller R_{ON} is desirable [31]. To this end, an alternative process integration scheme or alternative metallic electrode material which avoids the formation of non-conductive oxide on the contacting electrode surfaces would be beneficial.

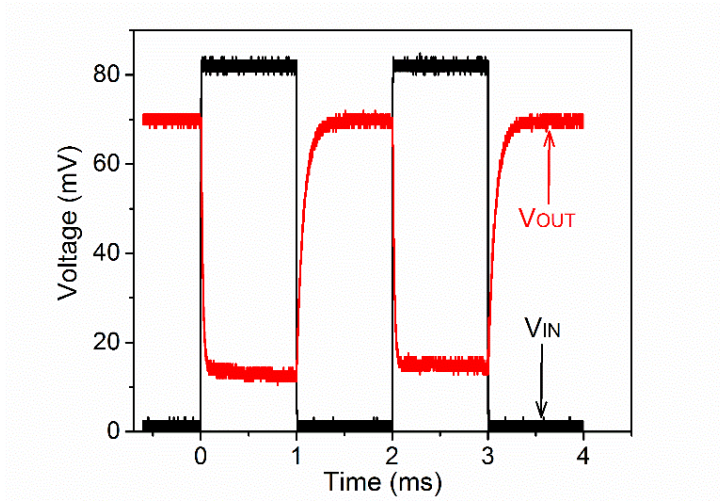


Fig. 4.22. Measured voltage waveforms for an inverter circuit with a pull-down relay (the same configuration as Fig. 4.20(a)), with V_B = -11.02 V and V_{DD} = 80 mV.

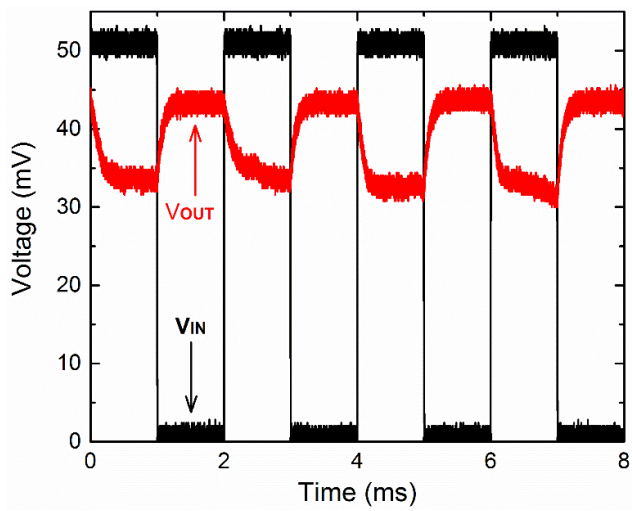


Fig. 4.23. Measured voltage waveforms for an inverter circuit with a pull-up relay (the same configuration as Fig. 4.20(a)), with V_B = -11.05 V and V_{DD} = 50 mV.

4.5.4 Mechanical Switching Delay

Fig. 4.24 shows zoomed-in voltage waveforms at the rising edge of the input signal of Fig. 4.20(b). The relay's mechanical turn-on delay (τ_{ON}) can be distinguished from the RC delay (output capacitance discharging) upon closer examination of the V_{OUT} waveform: notice that V_{OUT} starts to decrease after V_{IN} rises, but only after a delay of τ_{ON} = 440 ns. This measured value reasonably matches the simulation result (which predicts a quarter period of base mode is 0.27 μs) in section 4.3.2.1. Note that as discussed in section 3.4.3, the operating voltages V_B and V_G have significant effect on τ_{ON}, which can explain the discrepancy between simulation (which

doesn't take into account V_B and V_G for modal analysis) and experiment results. It should be noted that the large RC delay seen in Fig. 4.24 is mainly due to the large capacitive load consisting of large parasitic capacitance from the probe pad, probe tip, cables, and input capacitance of oscilloscope. In practical circuit, the parasitic capacitance would be much smaller so the RC delay should be much less than the mechanical delay.

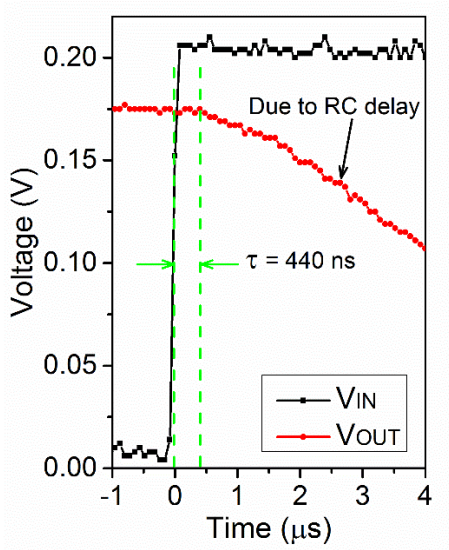


Fig. 4.24. Measurement of the mechanical turn-on delay (τ_{ON}) of a 6-T relay operated with V_B = -11 V and V_{DD} = 0.2 V.

4.5.4 Variability

Finally, variability of relay switching voltages is investigated. Fig. 4.25 shows the cumulative distribution function (CDF) of V_{ON}, V_{RL}, and V_H for fabricated relays of design C with L_B = 15 μm. 20 devices were measured. Each device was tested more than 5 times so as to obtain an average voltage value. With body bias, the median V_H drops to less than 0.2 V, thanks to smaller contact impact velocity and thus smaller adhesive force [32].

Many sources contribute to the overall variability. For example, the polycrystalline nature of the poly-$Si_{0.4}Ge_{0.6}$ structural layer (*e.g.* variation of grain size, shape, and orientation, *etc.*) leads to variation in effective spring constant. Variation of stress gradient in the structural layer leads to variation in contact and actuation gap sizes, which in turn results in variation in V_{ON} and V_H. The variation of actual contact area and thus surface adhesive force also results in variation in V_H. Furthermore, process-induced variation in film thickness and line-edge roughness of the suspension beams contribute to overall variability as well. To reduce variability in relay performance parameters, better control of physical and geometrical properties are needed.

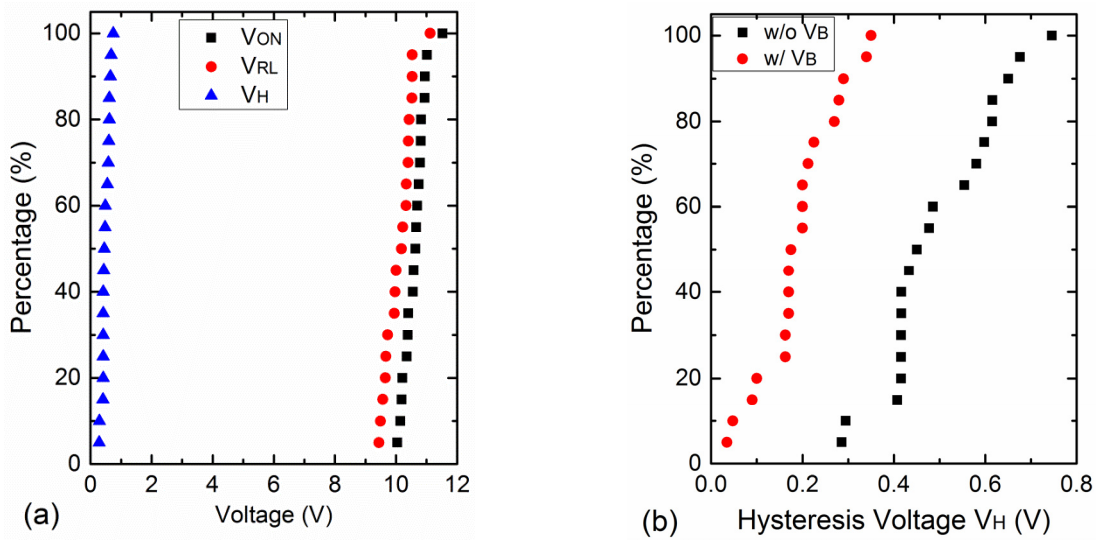


Fig. 4.25. Measured cumulative distributions of switching voltages, for relays of design C and 15 μm beam length: (a) CDF of V_{ON}, V_{RL}, and V_H; (b) CDF of V_H for relays with (red dots) and without (black square) body bias.

4.6 Summary

Body biasing is an effective away to lower the operation voltage of a relay. By designing the relay for non-pull-in mode operation and with relatively large spring stiffness, sub-100 mV operation is successfully demonstrated. Key design parameters and their effects on relay characteristics are discussed in detail. Relays of four structural designs and varying values of spring stiffness are fabricated and systematically characterized. The fabrication process is described with material choices and process parameters discussed in detail. Both process and device simulations are carried out by using CoventorWare FEM software. The simulation results and experimental results match well, and confirm predicted trends from the analytical analysis in chapter 2. Relay-based inverter circuits operating below 100 mV are demonstrated with body biasing. These results indicate that electro-mechanical relay technology is promising for ultra-low-power mechanical computing.

4.7 References

[1] O. Y. Loh, and D. Espinosa, "Nanoelectromechanical contact switches," *Nature Nanotech.*, vol. **7**, pp. 283-296, 2012.

[2] V. Pott, *et al.*, "Mechanical computing redux: relays for integrated circuit applications," *Proc. IEEE 98*, pp. 2076-2094, 2010.

[3] T.-H. Lee, S. Bhunia, and M. Mehregany, "Electromechanical computing at 500°C with silicon carbide," *Science*, vol. 329, pp. 1316-1318, 2010.

[4] J. O. Lee, *et al.*, "A sub-1-volt nanoelectromechanical switching device," *Nature Nanotech.*, vol. 8, pp. 36-40, 2013.

[5] C. Pawashe, K. Lin, and K. J. Kuhn, "Scaling limits of electrostatic nanorelays," *IEEE Trans. Electron Devices*, vol. 60, no. 9, pp. 2936-2942, Sep. 2013.

[6] J. Yaung, L. Hutin, J. Jeon and T.-J. K. Liu, "Adhesive force characterization for MEM logic relays with sub-micron contacting regions," *IEEE/ASME J. Microelectromech. Syst.*, vol. 23, no. 1, pp. 198-203, Feb. 2014.

[7] G. K. Fedder, "Simulation of microelectromechanical systems", Ph.D. thesis, UC Berkeley, 1994.

[8] S. Timoshenko, *History of strength of materials*, McGraw-Hill New York, 1953.

[9] H. Kam, *et al.*, "Design, optimization, and scaling of MEM relays for ultra-low-power digital logic," *IEEE Trans. Electron Devices,* vol. 58, pp. 236-250, 2011.

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

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

[12] G. M. Rebeiz, *RF MEMS: Theory, Design, and Technology* Ch. 3, John Wiley & Sons, 2003.

[13] http://www.coventor.com/mems-solutions/products/coventorware/

[14] https://www.cadence.com

[15] D. Lee, V. Pott, H. Kam, R. Nathanael, T.-J. K. Liu, "AFM characterization of adhesion force in micro-relays," *IEEE MEMS*, 2010, pp. 232-235.

[16] Y. Chen, R. Nathanael, Y. Yaung, L. Hutin, and T.-J. K. Liu, "Reliability of MEM relays for zero leakage logic," *Proc. of SPIE*, vol. 8614, pp. 861404.1-7, 2013.

[17] C. Chen, R. Nathanael, J. Jeon, J. Yaung, L. Hutin, and T.-J. K. Liu, "Characterization of contact resistance stability in MEM relays with tungsten electrodes," *J. MEMS*, vol. 21, pp. 511-513, 2012.

[18] A. M. Haghiri-Gosnet, F. R. Ladan, C. Mayeux, H. Launois, and M. C. Joncour, "Stress and microstructure in tungsten sputtered thin films," *J. Vac. Sci. Technol. A*, vol. 7, pp. 2663-2669, 1989.

[19] T. Karabacak, C. R. Picu, J. J. Senkevich, G.-C. Wang, and T.-M. Lu, "Stress reduction in tungsten films using nanostructured compliant layers," *J. Applied Phys.*, vol. 96, pp. 5740-5746, 2004.

[20] H. Windischmann, "Intrinsic stress in sputtered thin films," *J. Vac. Sci. Technol. A*, vol. 9, pp. 2431-2436, 1991.

[21] I.-R. Chen, Y. Chen, L. Hutin, V. Pott, R. Nathanael, T.-J. K. Liu, "Stable ruthenium-contact relay technology for low-power logic," *IEEE Transducers*, pp. 896-899, 2013.

[22] P. Ericsson, S. Bengtsson, and J. Skarp, "Properties of Al2O3-films deposited on silicon by atomic layer epitaxy," *Microelectron. Eng.*, vol. 36, pp. 91-94, 1997.

[23] H. C. Lin, P. D. Ye, and G. D. Wilk, "Leakage current and breakdown electric-field studies on ultrathin atomic-layer-deposited Al_2O_3 on GaAs," *App. Phys. Lett.*, vol. 87, 182904, 2005.

[24] M. D. Groner, J. W. Elam, F. H. Fabreguette, and S. M. George, "Electrical characterization of thin Al_2O_3 films grown by atomic layer deposition on silicon and various metal substrates," *Thin Solid Films*, vol. 413, pp. 186-197, 2002.

[25] A. E. Franke, J. M. Heck, T.-J. King, and R. T. Howe, "Polycrystalline silicon germanium films for integrated microsystems," *J. Microelectromech. Syst.*, vol. 12, pp. 160-171, Apr. 2003.

[26] C. W. Low, T.-J. K. Liu, and R. T. Howe, "Characterization of polycrystalline silicon-germanium film deposition for modularly integrated MEMS applications," *IEEE J. MEMS*, vol. 16, pp. 68-77, 2007.

[27] S. Sedky, A. Witvrouw, A. Saerens, P. V. Houtte, J. Poortmans, and K. Baert, "Effect of in situ boron doping on properties of silicon germanium films deposited by chemical vapor deposition at 400 °C," *J. Mater. Res.*, vol. 16, pp. 2607-2612, 2001.

[28] C. W. Low, M. L. Wasilik, H. Takeuchi, T.-J. King, and R. T. Howe, "In-situ doped poly-SiGe LPCVD process using BCl for post-CMOS integration of MEMS devices," *Proc. Electrochem. Soc. SiGe Mater.,Process., Devices Symp.*, Honolulu, HI, Oct. 3–8, 2004, pp. 1021–1032.

[29] C. W.-Z. Low, "Novel processes for modular integration of silicon-germanium MEMS with CMOS electronics," Ph.D. thesis, UC Berkeley, 2007.

[30] Y. Chen, I.-R. Chen, C. Qian, A. Peschot, T.-J. K. Liu, "Experimental studies of contact detachment delay in microrelays for logic applications," *IEEE Trans. Electron Devices*, vol. 62, pp. 2695-2699, 2015.

[31] M. Spencer, F. Chen, C. C. Wang, R. Nathanael, H. Fariborzi, A. Gupta, H. Kam, V. Pott, J. Jeon, T.-J. K. Liu, D. Markovic´, E. Alon, and V. Stojanovic´, "Demonstration of integrated micro-electro-mechanical relay circuits for VLSI applications," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 146-148, Jan. 2011.

[32] C. Qian, A. Peschot, I. R. Chen, Y. Chen, N. Xu, and T.-J. K. Liu, "Effect of body basing on the energy-delay performance of logic relays," *IEEE Electron Dev. Lett.*, vol. 36, no. 8, pp. 862-864, Aug. 2015.

Chapter 5

NEM Selector for Cross-point Memory

5.1 Introduction

Cross-point arrays of resistive random access memory (ReRAM) or phase change memory (PCM) cells require highly nonlinear selector devices to reduce sneak path currents for larger read and write margins and lower power consumption. Nano-electro-mechanical (NEM) switches inherently have non-linear *I-V* characteristics and high ON/OFF current ratio, and hence are prospective candidates for selector devices.

Section 5.2 proposes a structure of NEM selector suitable for 3-D cross-point memory integration, describing the device operation principle and *I-V* characteristics. Section 5.3 presents a fabrication process and preliminary characterization results for the NEM selector. Challenges for miniaturization are discussed in section 5.4. To address the challenge of stuck-ON failure of 2-terminal NEM selectors, section 5.5 introduces a 3-terminal NEM selector with one additional terminal used for actively resetting the selector. Section 5.6 summarizes this chapter.

5.2 Proposal of NEM Selector

Fig. 5.1 shows the structure of a proposed 1S-1R cross-point memory cell comprising a NEM selector in series with a storage element (*e.g.*, ReRAM or PCM). The NEM selector consists of two orthogonally running linear metal electrodes with an air gap in-between, in the OFF state (Fig. 5.1(b)). The red thin films on the top surface of the middle electrode (ME) and on the

bottom surface of the top electrode (TE) represent native oxide layers. The overlap region between the TE and ME forms a parallel plate capacitor. When a memory cell is selected, the voltage applied between the TE and ME is high enough such that the electrostatic force F_{elec} is larger than the spring restoring force F_{spring} of the TE beam so that it bends and makes contact with the ME (Fig. 5.1(c)); when the cell is un-selected, F_{elec} decreases such that F_{spring} overcomes the adhesive force F_{adh} between the TE and ME so that the TE detaches from the ME (Fig. 5.1(b)). Note that in the selected state (*i.e.* when the TE contacts the ME), there still exists a voltage V_{TM} between the TE and ME because of the existence of contact resistance R_C which forms a voltage divider with the storage element R_S. This voltage V_{TM} is referred to as the hold voltage V_{hold} which, together with F_{adh}, holds TE in contact with ME when the memory cell is selected for read/write operation. By engineering R_C (*e.g.* by changing the contact area and/or surface oxide thickness), R_C and hence V_{hold} can be adjusted. The proposed selector structure is very simple and easy to integrate into a cross-point memory array.

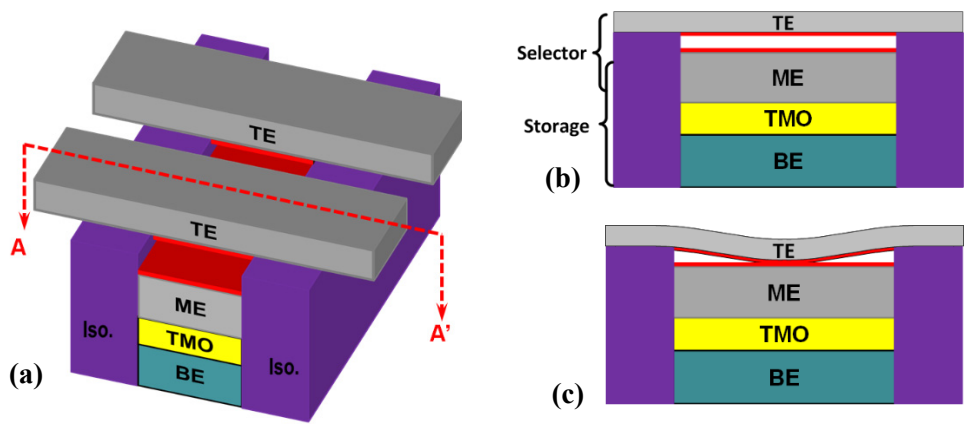


Fig. 5.1. Proposed 1S1R cross-point memory cell with NEM selector. (a) 3-D cross-point memory cell structure; (b) cross-section view of OFF-sate selector; and (c) cross-section view of ON-state selector.

The current-*vs.*-voltage characteristics of the proposed selector are as shown in Fig. 5.2. When the voltage across the selector V_{SL} is larger than a threshold voltage V_{th}, the selector current I_{SL} increases suddenly from zero to V_{SL}/R_C; when V_{SL} decreases, due to the existence of F_{adh}, I_{SL} doesn't drop to zero until V_{SL} is smaller than the release voltage V_{rl}. The *I-V* characteristics are centrosymmetric about the coordinate origin *O*, so the NEM selector is also suitable for bipolar ReRAM.

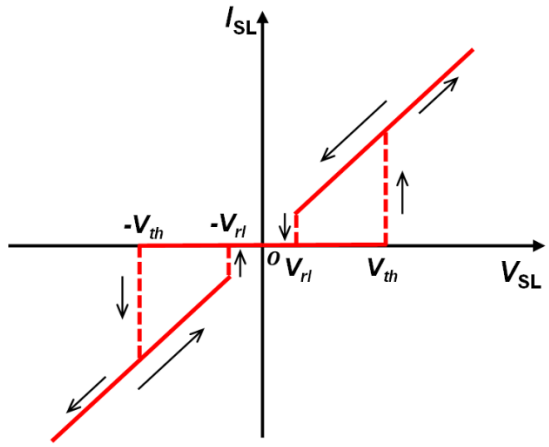


Fig. 5.2. *I-V* characteristics of the proposed NEM selector (linear scale). The selector turns on when $V_{SL} > V_{th}$, and turns off when $V_{SL} < V_{rl}$. The inverse of the *I-V* slope is equal to contact resistance R_C.

Fig. 5.3 shows a simplified *I-V* characteristic of a bipolar ReRAM device which has two resistance states: a high resistance state (HRS) and a low resistance state (LSR). It can be set (from HRS to LRS) when the voltage on the storage element V_{ST} is larger than the set voltage V_{set}, and be reset (from LRS to HRS) when V_{ST} goes negatively below $-V_{rset}$. Note that R_{HRS} should be much larger than R_C so that R_C doesn't decrease the read margin substantially.

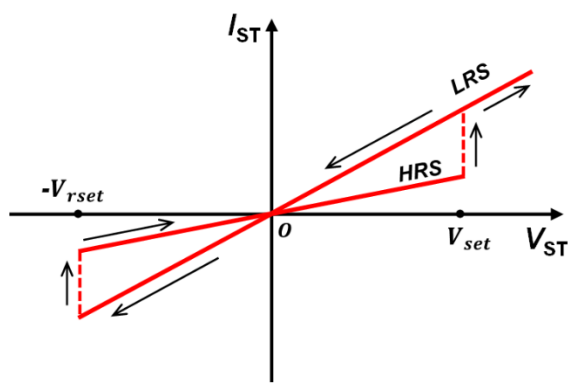


Fig. 5.3. Hypothetical *I-V* characteristics of a bipolar ReRAM storage device which has a high resistance stage (HRS) and low resistance state (LRS). It can be set from HRS to LRS when $V_{ST} > V_{set}$, and reset from LRS to HRS when $V_{ST} < -V_{rset}$.

Fig. 5.4 shows the overall *I-V* characteristics of the combined 1S1R memory cell for both set and reset cycles. When the write voltage V_W is less than V_{th}, the selector is off so no current flows (*i.e.* $I_W = 0$); when $V_W > V_{th}$, the selector is turned on and the current is

$$I_W = \frac{V_W}{R_C + R_{HRS}} \approx \frac{V_W}{R_{HRS}}. \tag{5.1}$$

When V_W increases further such that $V_W > V_{set}$, the storage element is set from HRS to LRS. The current through the memory cell is

$$I_W = \frac{V_W}{R_C + R_{LRS}}, \tag{5.2}$$

which doesn't drop to zero until $V_W < V'_{rl}$, where V'_{rl} is

$$V'_{rl} = V_{rl}\left(1 + \frac{R_{LRS}}{R_C}\right). \tag{5.3}$$

To reset the cell, the *I-V* loop in the third quadrant of Fig. 5.4 is traversed, where

$$V'_{rset} = V_{rset}\left(1 + \frac{R_C}{R_{LRS}}\right), \tag{5.4}$$

and

$$V''_{rl} = V_{rl}\left(1 + \frac{R_{HRS}}{R_C}\right) \approx V_{rl}\frac{R_{HRS}}{R_C}. \tag{5.5}$$

Equations (5.3)~(5.5) take into account the voltage divider comprising selector and storage elements.

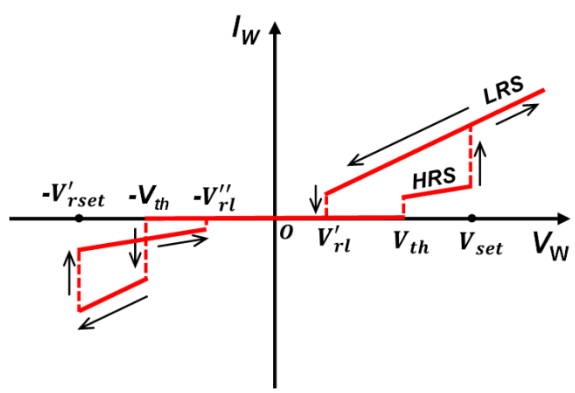


Fig. 5.4. *I-V* characteristics of the proposed 1S1R memory cell for both set and reset cycles.

The read margin of a cross-point memory array depends on the read scheme [1]. Two popular schemes, *i.e.* 1/2 V and 1/3 V schemes, are shown in Fig. 5.5. For the case of 1/2 V in Fig. 5.5(a), the word line (WL) and bit line (BL) of a selected cell are connected to *V* and ground, respectively. All other lines are biased at *V*/2, so that every cell along the selected WL(red)/BL(green) is halfway biased (at *V*/2) except for the cell at the intersection point, which is fully biased. The total sneak current in the array is

$$I_{sneak} = (m + n - 2) * I\left(\frac{V}{2}\right), \tag{5.6}$$

where m and n are the number of WL and BL in the array, respectively. For the case of 1/3 V read scheme in Fig. 5.5(b), the WL and BL of the selected cell are biased at V and ground, respectively, but every other WL is biased at $V/3$ and every other BL is biased at $2V/3$. So the total sneak current is

$$I_{sneak} = (mn - 1) * I\left(\frac{V}{3}\right). \tag{5.7}$$

From (5.6) and (5.7), one can clearly see that $I(V/2)$ or $I(V/3)$ should be very small in order to limit the total sneak current and power consumption.

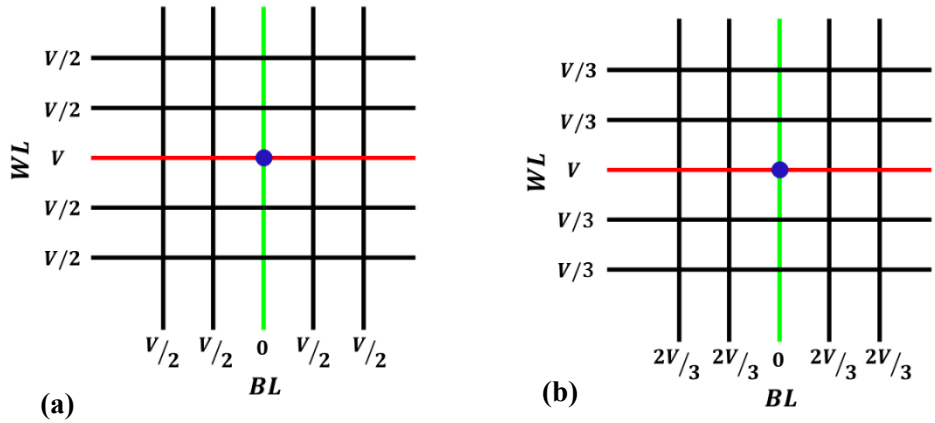


Fig. 5.5. (a) $V/2$ read scheme; (b) $V/3$ read scheme.

Fig. 5.6 indicates the read current for fully-selected and half-selected cells. When V_{rd} is larger than V_{th}, the selector of a fully-selected cell is turned on so the state (HRS or LRS) of the storage element can be differentiated; whereas the selectors of half-selected cells are in the OFF state so sneak current is virtually zero. Thanks to the high nonlinearity of a NEM selector, sneak current in a cross-point memory array can be significantly decreased so that the array size (m and n) can be increased and power consumption can be kept low.

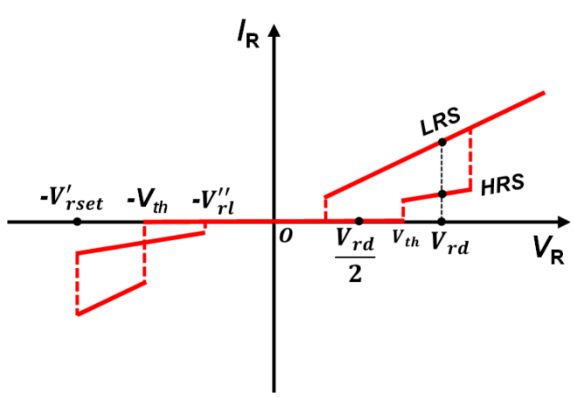


Fig. 5.6. *I-V* characteristics of 1S1R memory cell in read state.

5.3 Experimental Investigation of NEM Selector

To investigate the feasibility of NEM selector, prototype devices were fabricated and characterized. The fabrication process and preliminary results are shown in this section. Some challenges and possible remedies are also discussed.

5.3.1 Device Fabrication

The fabrication process[i] of a cross-point NEM selector is illustrated in Fig. 5.7. Starting from a silicon wafer, a 150nm-thick SiO_2 layer was grown by thermal oxidation, followed by ALD deposition of a layer of 80 nm-thick Al_2O_3. Both SiO_2 and Al_2O_3 were used for substrate insulation; Al_2O_3 layer also protected the underlying SiO_2 from being etched by vapor HF in the final releasing step. Sputtering deposition was then applied to deposit a layer of 100nm-thick tungsten, on top of which a very thin (10~20 nm) low-temperature SiO_2 (LTO) sacrificial layer was deposited by LPCVD at 400 ºC. Both layers were then patterned by photolithography and RIE etching, forming the bottom electrode stack (called bit line hereafter) as shown in Fig. 5.7(a). Bit-line width determined the length of doubly-clamped beam (*i.e.* the suspended part of word line) which determined the beam stiffness and threshold voltage V_{th}. Different widths of bit lines were fabricated, ranging from 250 nm to 6 μm. Note that the LTO thickness is critical since it determines air gap thickness which then determines F_{elec}, F_{spring}, V_{th}, V_{rl}, etc. To achieve a low-V_{th} selector (to be compatible with the read/write voltage of the storage element, and also for lowering power consumption), thinner air gap is preferred. It is, however, quite challenging to fabricate ultra-thin (*e.g.* sub 10 nm) air gaps.

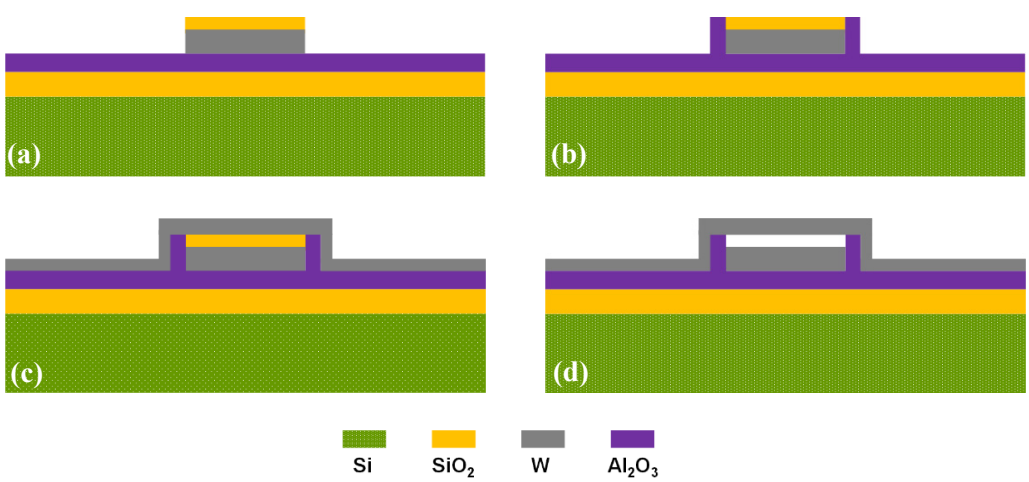


Fig. 5.7. Fabrication process flow for a cross-point NEM selector.

The Al_2O_3 spacers along the sidewalls of the bit line (Fig. 5.7(b)) were formed firstly by ALD conformal deposition followed by RIE anisotropic etch-back. This was also a challenging step

[i] Note that all materials used here are the same as those used in chapter 4 for fabricating logic relays. The deposition/etching recipes are all the same. Please refer to section 4.4 for recipe details.

because there was no available RIE etch recipe with good selectivity between Al_2O_3 and SiO_2 (*i.e.* no recipe which could etch Al_2O_3 much faster than SiO_2). Slight underetching of the Al_2O_3 spacer layer would not fully expose the underlying sacrificial SiO_2, preventing successful release in HF vapor subsequently (Fig. 5.8(a)); however, slight overetching could easily remove the thin layer of sacrificial SiO_2 (Fig. 5.8(b)) and results in no air gap. Both scenarios would result in a non-functional device. Thus, not only the etching rate but also the etching time must be precisely controlled in the spacer etch step. This was quite tricky due to inevitable process variation (*e.g.* etching rate was different between edge and center region of a wafer; it also varied from wafer to wafer, batch to batch, *etc*.). A SEM image of patterned bit lines with ~50 nm-thick Al_2O_3 spacers is shown in Fig. 5.9. SiO_2 also could be used as the spacer material, in which case the spacers together with the top layer of sacrificial LTO would be removed by HF vapor in the final releasing step, leaving air gaps at the spacer regions also.

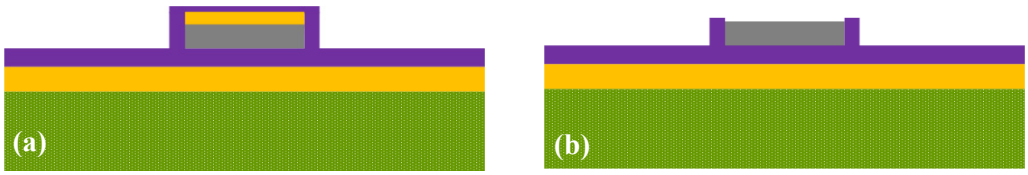


Fig. 5.8. Two failure modes associated with the bit-line sidewall spacer etch step. (a) Al_2O_3 underetch (b) Al_2O_3 overetch.

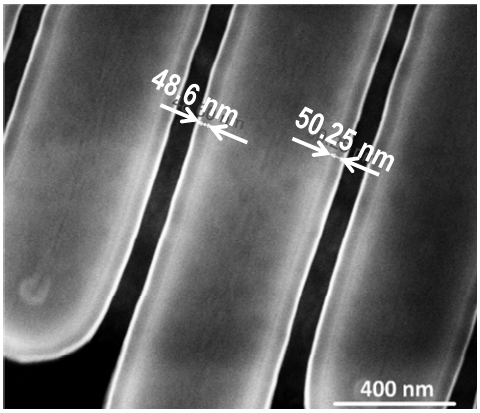


Fig. 5.9. SEM image of fabricated bit lines with Al_2O_3 spacers.

As an additional remark, silicon nitride cannot be used as the spacer material since it reacts with HF vapor. Fig. 5.10 shows a SEM image of a sample which used Si_3N_4 for both the spacers and the substrate insulation layer and was exposed to anhydrous HF vapor (assisted by ethanol) for about 15 minutes. Clearly the Si_3N_4 spacers were attacked and formed debris mainly comprised of ammonium-fluorine salts, such as NH_4HF_2 and $(NH_4)_2SiF_6$ [2,3].

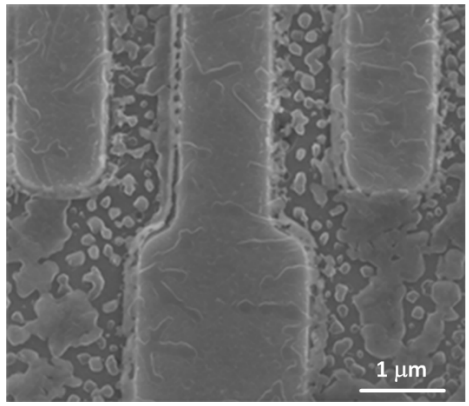


Fig. 5.10. SEM image of a sample using Si_3N_4 as the spacer and substrate insulation material, after being exposed to HF vapor for 15 minutes.

The next step after spacer formation was to form the top electrode (word line). The same material (tungsten) that was used for the bit lines was used for the word lines. A 50 nm-thick tungsten layer was deposited by sputtering and then patterned to form parallel word lines (running orthogonally to the bit lines) by photolithography and RIE (Fig. 5.7(c)). It should be noted that stress control of the tungsten film (thus word line beams) is critical. Compressive stress in a beam can cause it to buckle up or down (depending on the stress gradient) after releasing it by selectively removing the sacrificial LTO. Buckling up would significantly increase the pull-in voltage V_{th} ($\propto g^{3/2}$); buckling down would possibly close the air gap and cause stuck-ON failure. For this reason, tensile stress is preferred over compressive stress. However, a tensile beam tends to contract upon release. For the case of SiO_2 spacers (which are removed during the release process), beam contraction can cause the beam to come into contact with the bit line, as illustrated in Fig. 5.11.

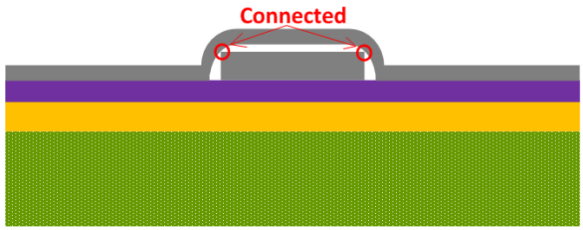


Fig. 5.11. A tensile word-line beam contracting upon release can come into contact with the bottom electrode.

This failure mode has been experimentally observed in fabricated poly-Si beam/SiO_2 spacer NEM selector arrays, as shown in Fig. 5.12, where the poly-Si WL beam is ~50 nm thick, the actuation air gap is ~40 nm, and the spacer air gap width is only ~18 nm. The spacer width was purposely designed to be very small in order to reduce V_{th}. Unfortunately the intrinsic tensile stress in the poly-Si beam caused the beam to shrink upon release and closed the air gap at the BL edges. This issue applies to W beams as well. Therefore, care must be taken to engineer the thin film stress. A reasonable design margin for spacer width is necessary, trading off manufacturing yield for low V_{th}.

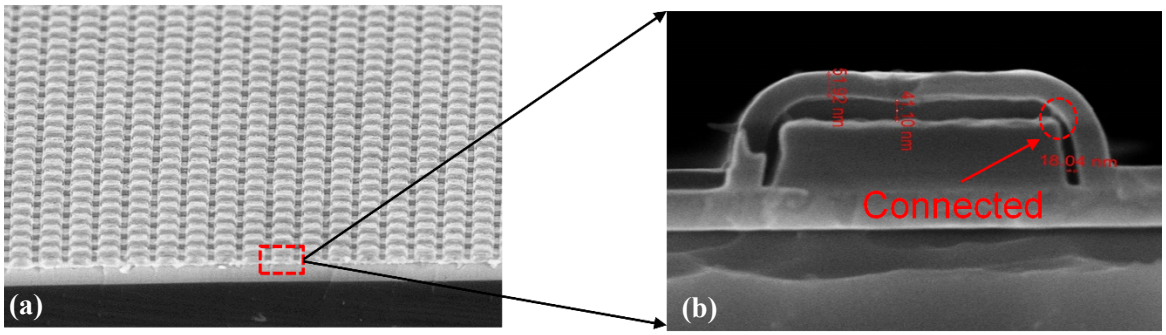


Fig. 5.12. (a) Fabricated NEM selector array with poly-Si beams and SiO_2 spacers; (b) cross-sectional view of a selector showing as-fabricated connection between WL and BL at the bit-line corners.

The final step was to release the beams. After dicing the wafer into smaller chips, an individual chip was exposed to a mixture of anhydrous HF vapor (etchant) and ethanol (catalyst) to remove sacrificial silicon oxide (Fig. 5.7(d)). SEM images of a fabricated cross-point NEM selector array (for which SiO_2 was used as the spacer material) are shown in Fig. 5.13. Fig. 5.14 shows a bird-eye-view SEM image of a selector with Al_2O_3 spacers. An air gap between the top and bottom electrodes can be clearly seen.

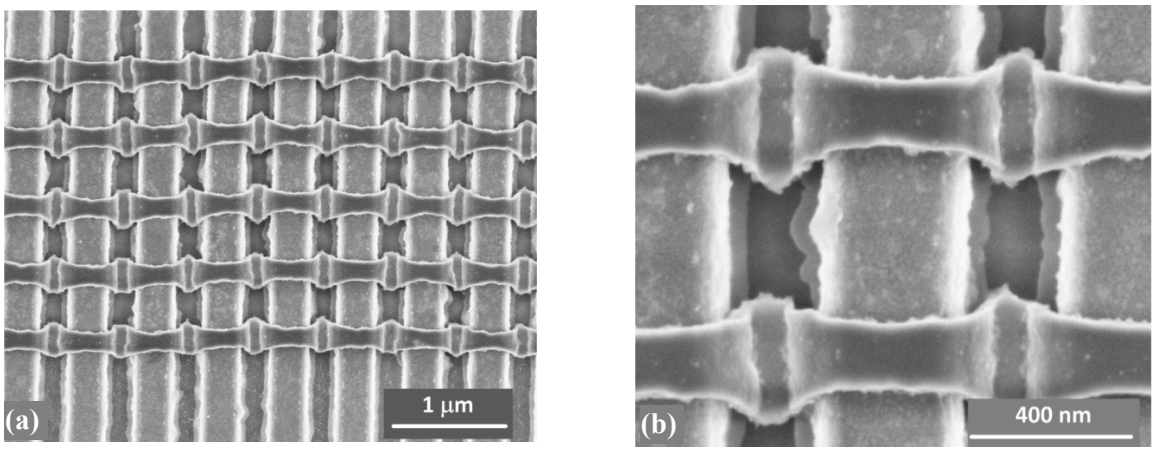


Fig. 5.13. (a) SEM image of fabricated cross-point NEM selector array with ~400 nm beam length. SiO_2 was used as the spacer material. (b) Zoomed-in image of (a). It can be seen that the SiO_2 spacers have been removed after exposure to HF vapor.

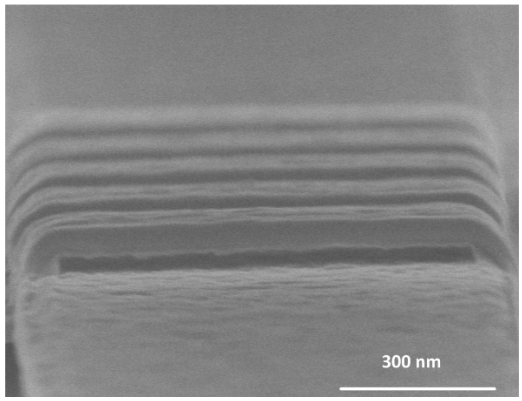


Fig. 5.14. Bird's-eye SEM image of a selector array with Al_2O_3 spacers. Air gap is successfully fabricated.

5.3.2 Electrical Characterization

Electrical characteristics of fabricated cross-point NEM selectors were measured using an Agilent 4156 semiconductor analyzer in atmosphere at room temperature. Fig. 5.15 shows the *I-V* characteristic of a selector with 2 μm beam length and about 20 nm air gap size. It turns on at about 2 V. Note that the ON-current was limited by a compliance limit to reduce Joule-heating-induced oxidation [4] for improved reliability. During the first 20 cycles, no measurable performance degradation was observed.

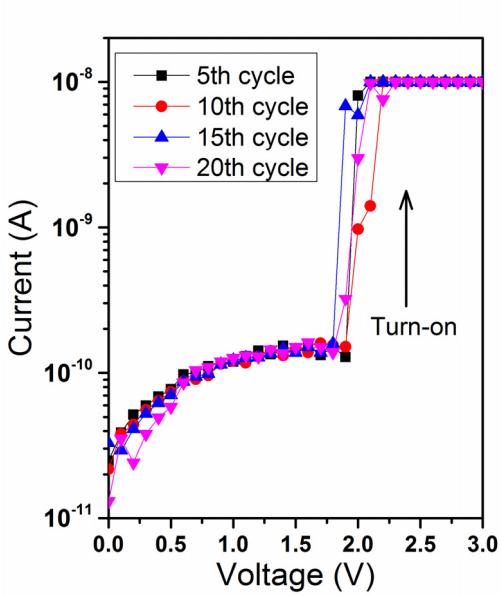


Fig. 5.15. *I-V* characteristics of a cross-point NEM selector comprising a 2μm-long beam.

5.4 Challenges for Selector Miniaturization

Although the concept of NEM selector is straight forward, some challenges still exist for practical implementation. Two major challenges, namely stuck-ON failure and strain limit, are discussed in this section.

5.4.1 Stuck-ON Failure

Similarly as for logic relays discussed in chapters 2-4, stuck-ON failure is one of the most critical challenges for NEM selector devices. Since the NEM selector is a two-terminal device, it cannot take advantage of the body-biasing technique discussed in previous chapters. This poses a great challenge for achieving high manufacturing yield.

With the parallel-plate approximation introduced in section 2.2, the threshold voltage V_{th} of a NEM selector is equal to the pull-in voltage shown in Eq. (2.8), repeated here for the readers' convenience:

$$V_{th} = V_{PI} = \sqrt{\frac{8k_{eff}g_0^3}{27\varepsilon_0 A_{ACT}}}, \tag{2.8}$$

where k_{eff} is the effective spring constant, g_0 is the air gap size, ε_0 is the vacuum permittivity, and A_{ACT} is the overlapping electrode area of the cross-point structure. To turn on a NEM selector, the applied voltage V_{WB} between word line and bit line must be larger than V_{th}, hence,

$$k_{eff} < \frac{27\varepsilon_0 A_{ACT} V_{WB}^2}{8g_0^3}. \tag{5.8}$$

Assuming that the contacting surfaces are comprised of the same material, there is no charge transfer between the two surfaces hence no built-in potential in ON state [5]. In this case, a properly functioning NEM selector should turn off when V_{WB} is zero. This requires $F_{spring} > F_{adh}$, where $F_{spring} = k_{eff}g_0$, and $F_{adh} = A_C * P$ (see Eq. (2.20)). A_C is the real contact area, and P is the average adhesion pressure (*i.e.* force per unit real contact area). So,

$$k_{eff} > \frac{PA_C}{g_0}. \tag{5.9}$$

Combining (5.8) and (5.9) one has

$$\frac{A_C}{A_{ACT}} < \frac{27\varepsilon_0 V_{WB}^2}{8Pg_0^2}. \tag{5.10}$$

(5.10) is actually a harsh constraint. Let's put in some numbers to see this. Table 5.1 lists adhesion pressure values of some materials commonly used in mechanical switches. Assuming that the NEM selector is coated with a thin TiO_2 surface layer (so that P is relatively small), the target operation voltage is V_{WB} = 1 V, and the air gap is g_0 = 20 nm. Inserting these values into Eq. (5.10) gives the upper limit of the area ratio:

$$\frac{A_C}{A_{ACT}} < 4.13 * 10^{-5}. \tag{5.11}$$

TABLE 5.1
ADHESION PRESSURE BETWEEN FLAT SELF-SIMILAR SURFACES OF SOME COMMON MATERIALS [6]

Material	P (nN/nm^2)
SiO_2	1.04
TiO_2	1.81
C	3.5
Au	59.3
Pt	100
W	130

Although it is true that the real contact area A_C is usually much smaller than the apparent contact area due to surface roughness, the ratio given by Eq. (5.11) is still too small to be achievable in practice. This explains why most of the fabricated NEM selectors could only be actuated once and then became stuck ON. To satisfy Eq. (5.10), clearly one should decrease g_0, P, and/or A_C/A_{ACT}. A practical lower limit for g_0 is set by fabrication technology; achieving sub-5 nm air gap is very challenging nowadays. Since P is a material property, it is hard to decrease unless a better material (preferably with low adhesion pressure, low contact resistance, good stability, and high endurance) is used. A_C/A_{ACT} can be reduced by increasing surface roughness or decreasing the apparent contact area. Fig. 5.16 shows a reported cross-point NEM structure [7] where the authors proposed to use the edges of the bottom electrode (versus the middle flat portion) for the contacting regions. This is actually very similar to the case in Fig. 5.11. Instead of ending up with as-fabricated WL-BL connections as in Fig. 5.11, the authors carefully controlled the spacer size and shape, such that the edge gap size was smaller than that in the flat region, so that the word line beam makes contact with the bottom electrode only at the edges.

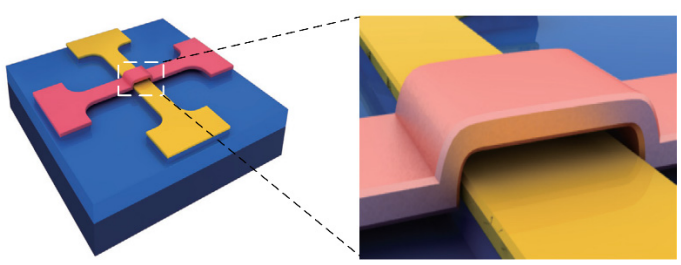


Fig. 5.16. A proposed cross-point structure where two electrodes make contact only along the edges of the lower electrode. Reprinted from [7].

5.4.2 Material Strain Limit

Mechanical strain in a beam can potentially cause plastic deformation or even mechanical fracture if it exceeds the beam material's yield point or fracture limit. Hence, close attention should be paid to the strain limit.

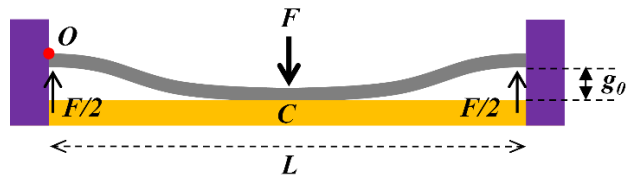


Fig. 5.17. Schematic illustration of a bent clamped-clamped beam.

The beam of a NEM selector device can be modeled as a classical clamped-clamped beam (Fig. 5.17). The maximum strain (and also stress) inside the beam exists at the top/bottom surfaces of the beam ends (*e.g.* point *O*). For analytical estimation, the downward force on the beam can be simplified as a concentrated load *F* applied at the middle of the beam (*i.e.* position *C*). Each of the two anchors provides an upward force *F*/2. According to the classical Euler-Bernoulli beam theory [8], the beam displacement at position *C* in the ON state is

$$g_0 = \frac{FL^3}{16EWt^3}, \tag{5.12}$$

where *E* is the Young's modulus of the beam, *L* is beam length, *W* is beam width, and *t* is beam thickness. The radius of curvature at point *O* is

$$R_O = \frac{EI}{M}, \tag{5.13}$$

where *M* is the internal bending moment, and *I* is the moment of inertia of the beam cross-section. *M* and *I* can be written as

$$\mathrm{M} = \frac{1}{8}FL, \tag{5.14}$$

and

$$I = \frac{1}{12}Wt^3, \tag{5.15}$$

respectively. Inserting (5.12), (5.14), and (5.15) into (5.13) leads to

$$\frac{1}{R_O} = \frac{24g_0}{L^2}. \tag{5.16}$$

So the strain at point *O* is

$$\epsilon_O = \frac{t/2}{R_O} = \frac{12tg_0}{L^2}. \tag{5.17}$$

For given dimensional parameters, the maximum strain inside the beam can be estimated by (5.17). For example, for t = 50 nm, L = 1 μm, and g_0 = 20 nm, one obtains $\epsilon_o = 0.12\%$. Depending on the strain limit of the beam material, this strain may induce plastic deformation or even fracture at the beam ends. Indeed, fracture was observed in fabricated W-beam NEM selectors (Fig. 5.18).

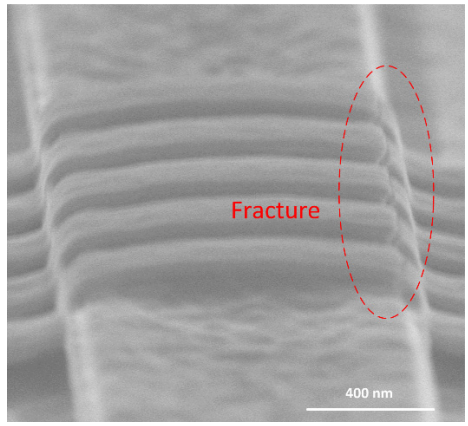


Fig. 5.18. SEM image of W-beam NEM selectors after electrical testing. Fracture happens at the beam ends where maximum strain exists.

To mitigate strain-induced fracture issue, thinner and longer beam, smaller air gap, and beam material with larger strain limit are preferred. For given material and process technology, the minimum beam length of a cross-point NEM selector can be estimated by the following formula:

$$L_{min} = \sqrt{\frac{12tg_0}{\epsilon_{limit}}}, \tag{5.18}$$

where ϵ_{limit} is strain limit of beam material. Therefore, Eq. (5.18) can be used to predict the scaling limit of a NEM selector. Table 5.2 shows a group of materials and their minimum beam length predicted by Eq. (5.18). For a 5 nm-thick polycrystalline-silicon beam (with a strain limit of 0.93%), the minimum beam length is ~114 nm for an actuation gap of 2 nm. Alternative structural materials such as TiNi with superior yield strain will be needed to scale the beams to shorter length. Although seemingly interesting due to ease of integration, Al beams should not be scaled below 258 nm due to their particularly low tolerance to strain.

TABLE 5.2
SCALING LIMIT OF NEM SELECTOR WITH DIFFERENT BEAM MATERIALS

Beam material	Poly-Si	Aluminum	TiNi
Strain limit (ϵ) [9, 10, 11]	0.93 %	0.18 %	~ 8 %
Thickness (t)	5 nm	5 nm	5 nm
Actuation gap (g_0)	2 nm	2 nm	2 nm
Beam length	114 nm	258 nm	39 nm

5.5 Actively-Reset NEM Selector Design

To avoid the stuck-ON failure issue discussed in section 5.4.1, another way is to add a reset-electrode (RSE) which can actively pull the WL up to reset the NEM selector by electrostatic force, as illustrated in Fig. 5.19. Both RSE and BL are very stiff so that only WL is movable during the set/reset process.

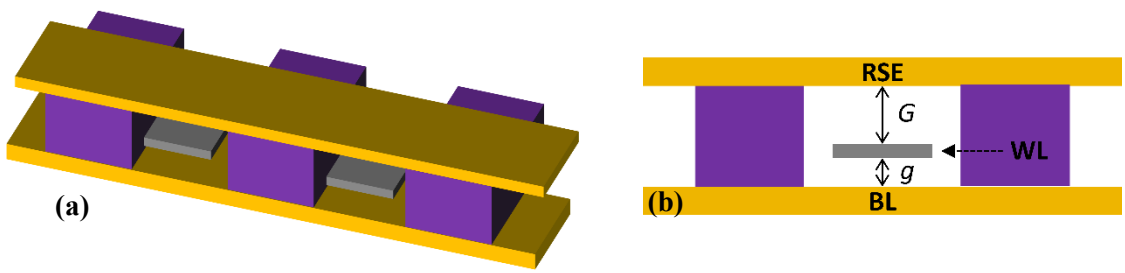


Fig. 5.19. Proposed NEM selector with a reset electrode. (a) 3-D structure; (b) cross-sectional view.

To read/write a cross-point memory cell, RSE is grounded, and BL/WL are biased in the similar way as for the normal 2-terminal NEM selectors discussed in previous sections. To reset a NEM selector (not the storage element sandwiched in-between WL and BL), RSE needs to be charged to a high voltage level such that $F_{elec} + F_{spring} > F_{adh}$. Again, the parallel-plate approximation is utilized to estimate the reset voltage V_{rst}. The above inequality can be written as

$$k_{eff}g + \frac{\varepsilon_0 A_{ACT} V_{rst}^2}{2(g+G)^2} > F_{adh}, \tag{5.19}$$

where G is the thickness of the air gap between RSE and WL (Fig. 5.19(b)). So the required reset voltage is

$$V_{rst} > \sqrt{\frac{2(g+G)^2}{\varepsilon_0 A_{ACT}}(F_{adh} - k_{eff}g)}. \tag{5.20}$$

Clearly G must be minimized to reduce V_{rst}. However, if G is too small (*e.g.* $G < g$) WL may stick to RSE upon reset, in which case a larger read/write voltage on BL would be needed to address a selected cell (*i.e.* to close the air gap between WL and BL of the selected cell). This high voltage on BL not only increases the power consumption but also can damage the storage element, and therefore should be avoided. From the 3-D structure in Fig. 5.19(a), one can infer that all NEM selectors under the same RSE are reset at the same time with a high V_{rst} voltage.

The drawbacks of this actively-reset NEM selector design are: a) more complicate fabrication process (involving the formation of two air gaps per cross-point memory cell, and b) one more photolithography step, which increases process cost.

5.6 Summary

A NEM switch is proposed to be used as selector device for cross-point memory arrays. A prototype NEM selector is fabricated and characterized. Both the fabrication process and preliminary characterization results are presented. Challenges remain for scaling down the operation voltage and selector dimensions. A remedy (by adding additional reset electrode) to prevent stuck-ON failure is proposed.

5.7 References

[1] J. Zhou, K.-H. Kim, and W. Lu, "Crossbar RRAM arrays: selector device requirements during read operation," *IEEE Trans. Electron Devices*, vol. 61, pp. 1369-1376, 2014.

[2] G. Vereecke, M. Schaekers, K. Verstraete, S. Arnauts, M. Heyns, and W. Plante, "Quantitative analysis of trace metals in silicon nitride films by a vapor phase decomposition/solution collection approach," *J. Electrochem. Soc.*, vol. 147, no. 4, pp. 1499-1501, 2000.

[3] A. Witvrouw, B. D. Bois, P. D. Moor, A. Verbist, C. V. Hoof, H. Bender, and K. Baert, "A comparison between wet HF etching and vapor HF etching for sacrificial oxide removal," *Proc. SPIE Micromachining and Microfabrication Process Technology VI*, vol. 4174, p. 130, 2000.

[4] Y. Chen, R. Nathanael, J. Jeon, Y. Yaung, L. Hutin, and T.-J. K. Liu, "Characterization of contact resistance stability in MEM relays with tungsten electrodes," *IEEE J. MEMS*, vol. 21, pp. 511-513, 2012.

[5] L. Hutin, W. Kwon, C. Qian, and T.-J. K. Liu, "Electro-mechanical diode cell scaling for high-density non-volatile memory," *IEEE Trans. Electron Devices*, vol. 61, pp. 1382-1387, 2014.

[6] C. Pawashe, K. Lin, and K. J. Kuhn, "Scaling limits of electrostatic nanorelays," *IEEE Trans. Electron Devices*, vol. 60, no. 9, pp. 2936-2942, Sep. 2013

[7] J. O. Lee, Y.-H. Song, M.-W. Kim, M.-H. Kang, J.-S. Oh, H.-H. Yang, and J.-B. Yoon, "A sub-1-volt nanoelectromechanical switching device," *Nature Nanotech.*, vol. 8, pp. 36-40, 2012.

[8] S. Timoshenko, *History of strength of materials*, McGraw-Hill New York, 1953.

[9] A. Ishida, M. Sato, and S. Miyazaki, "Mechanical properties of Ti–Ni shape memory thin films formed by sputtering," *Mater. Sci. Eng. A*, vol. 273–275, pp. 754-757, 1999.

[10] S.-H. Lee, J. W. Evans, Y. E. Pak, J. U. Jeon, and D. Kwona, "Evaluation of elastic modulus and yield strength of Al film using an electrostatically actuated test device," *Thin Solid Films*, vol. 408, pp. 223-229, 2002.

[11] Y. C. Tai and R. S. Muller, "Fracture strain of LPCVD polysilicon," *Proc. Solid-State Sens. Actuator Workshop*, 1988, pp. 88-91.

Chapter 6

Conclusion

6.1 Summary and Contributions of This Work

This dissertation aims to advance progress in M/NEM switch technology for digital logic and memory applications.

As a logic switch, an electro-mechanical relay is potentially superior to CMOS transistors because it has virtually zero leakage current and abrupt switching behavior, which in principle enable lower-power digital circuits. This dissertation focuses on reducing the operating voltage to improve the energy efficiency of logic relays.

Conventional wisdom suggests that the effective spring constant k_{eff} should be as small as possible to minimize the switching energy of a logic relay, but this doesn't take into consideration the possibility of using body biasing as a means of reducing V_{DD} and thereby switching energy. Relays designed in the conventional way (*i.e.* with a compliant structure) are susceptible to stuck-ON failure. Analytical modeling shows that body biasing is an effective way to reduce V_{DD}. The hysteresis voltage V_{H}, which sets the lower limit for V_{DD}, can be decreased by using a stiffer structure. When k_{eff} increases, the optimal ratio of actuation gap/contact gap (*i.e.* g_0/g_{d}) shifts from the pull-in mode design region (*i.e.* $g_0/g_{\mathrm{d}} < 3$ for parallel-plate capacitor approximation) towards the non-pull-in mode design region (*i.e.* $g_0/g_{\mathrm{d}} \geq 3$); when $k_{eff} \gg$

F_{adh}/g_d, optimal design point approaches $g_0/g_d \approx 3$, *i.e.* at the boundary of the pull-in mode and non-pull-in mode regions.

The effects of body biasing on logic relay performance characteristics are experimentally investigated. It is found that V_H slightly decreases with increasing $|V_B|$, due to smaller contact impact velocity and reduced contact adhesive force. The mechanical turn-on delay, which limits relay-based circuit operating speed, may increase or decrease with increasing $|V_B|$, depending on the gate overdrive voltage. There is a trade-off between switching energy and mechanical turn-on delay; that is, the switching energy can be reduced at the trade-off of increasing delay. A methodology for optimizing the energy-delay performance of logic relay is developed. Contrary to popular belief that structural stiffness should be minimized to achieve minimum switching energy, this dissertation shows that k_{eff} should be relatively large for better energy-delay performance.

By designing a logic relay to operate near the boundary of pull-in and non-pull-in operation regions, and with a relatively large effective spring constant, its operating voltage can be minimized. Relay design guidelines, fabrication process, and performance characteristics are discussed in detail. Finite-element-method simulations are carried out, and the simulated performance characteristics well match experimental measurements. Finally, a relay-based inverter circuit is demonstrated to operate reliably with a supply voltage below 100 mV, representing a significant milestone toward ultra-low-power mechanical computing.

Nano-electro-mechanical switches also can be integrated into cross-point memory cells as selector devices, offering virtually zero leakage current and superior current nonlinearity. Part of this dissertation research effort is devoted to the prototyping of NEM selectors. A cross-point memory cell comprising a ReRAM storage element in series with a NEM switch selector is proposed. The theoretical *I-V* characteristics of this 1S-1R cell are analyzed for both read and write operations. Prototype NEM selectors are fabricated. Initial experimental results are presented. Challenges for practical application of NEM selectors are noted, including: a) formation of nm-size air gap; b) stuck-ON failure due to large adhesive force and insufficient spring restoring force; c) material strain limit; and d) reliability issues.

6.2 Suggestions for Future Work

While this research advances M/NEM switch technology for both logic and memory applications, there is still room for further improvements:

<u>Contact material</u>. Although W has superior hardness, high melting point, and good endurance to wear and tear, it readily forms non-conductive native oxide when exposed to air. For millivolt operation, this issue becomes more problematic: even very thin native oxide built up on the contacting surfaces can cause resistance to increase significantly, because the small drain-to-source voltage and slow contact impact velocity are less effective to break down native oxide for improved current-conduction. An alternative metallic electrode material which avoids the formation of non-conductive oxide on the contacting electrode surfaces would be beneficial. Please note that since the contact impact velocity is relatively low with body biasing, the

contacting material is less subject to mechanical wear and tear. This can relieve the requirements for a new contacting material.

Reliability. As reported in [1,2], mechanical fatigue of the poly-$Si_{0.4}Ge_{0.6}$ structural layer is not the limiting factor; rather, contact resistance instability limits logic relay reliability. For a properly functioning relay in pass-gate logic circuits [3], the contact resistance should remain below a certain limit (*e.g.* 10 kΩ) within a targeted device operating lifetime (*e.g.* 10^{15} cycles, which corresponds to 10 years for a relay operating at 100 MHz frequency with 1% activity factor). Therefore, the reliability issue is related to the choice of contact material. To remedy this challenge, a new contacting material and hermetic packaging can be beneficial.

Variability. For practical application in very large-scale integrated (VLSI) circuits, variability in device performance should be minimized. (Although the body bias voltage can be adjusted to compensate for variability in turn-on voltage V_{ON}, *i.e.* by applying different V_B for different relays, this puts a burden on the power supply circuits.) Many factors contribute to device variability, including variations in poly-$Si_{0.4}Ge_{0.6}$ film thickness and stress gradient, uncertainty of actual contact area and adhesive force, and other process-induced variations. To reduce the overall variability, these factors should be more precisely controlled.

Further scaling down of V_H. Further scaling of V_H is possible if surface adhesive force can be reduced. One possible way is to coat the contacting surface with lower-adhesive-force material. This potentially can increase contact resistance and deteriorate reliability if the coated material is not very conductive or not very stable, however. Some preliminary results have been reported [4]. Another possible way is to operate the relay as a "squitch," in which case the contact gap of logic relay is filled with organic molecules through which the tunneling current can be eletromechanically modulated by gate voltage [5]. The drawback of the "squitch" approach is that the *I-V* characteristic is no longer abrupt; rather, the sub-threshold swing relies on tunneling current and molecular properties, *etc*.

3-D integration. To leverage the ultra-low-power merit of M/NEM logic relay and the high-speed advantage of CMOS transistors, integration of M/NEM relays on top of CMOS circuitry using a standard back-end-of-line (BEOL) process is a promising pathway for achieving ultra-low-power circuits [6]. Considering that the footprint of a planar M/NEM switch is relatively large, a vertically oriented logic relay can have an advantage of better area efficiency [7]. A functional logic relay implemented using a conventional BEOL process, with millivolt switching capability, remains to be demonstrated.

NEM selectors. Although the zero leakage and highly nonlinear *I-V* characteristics of a 2-terminal NEM switch make it an attractive candidate for cross-point memory selector, significant challenges exist for its adoption in large cross-point memory arrays. These are summarized at the end of section 6.1, and must be overcome to advance NEM selector technology.

6.3 Reference

[1] Y. Chen, R. Nathanael, Y. Yaung, L. Hutin, and T.-J. K. Liu, "Reliability of MEM relays for zero leakage logic," *Proc. of SPIE*, vol. 8614, pp. 861404.1-7, 2013.

[2] Y. Chen, "Reliability studies of micro-relays for logic applications," Ph.D. Dissertation, UC Berkeley, 2015.

[3] F. Chen, H. Kam, D. Markovic, T.-J. K. Liu, V. Stojanovic, and E. Alon, "Integrated circuit design with NEM relays," *IEEE/ACM Computer-Aided Design*, Nov. 2008.

[4] B. Osoba, B. Saha, L. Dougherty, J. Edgington, C. Qian, F. Niroui, J. H. Lang, V. Bulović, J. Wu, T.-J. K. Liu, "Sub-50 mV NEM relay operation enabled by self-assembled molecular coating," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2016.

[5] F. Niroui, A. Wang, E. M. Sletten, Y. Song, J. Kong, E. Yablonovitch, T. M. Swager, J. H. Lang, and V. Bulovic, "Tunneling nanoelectromechanical switches based on compressible molecular thin films," *ACS Nano*, vol. 9, no. 8, pp. 7886-7894, 2015.

[6] G. K. Fedder, R. T. Howe, T.-J. King Liu and E. P. Quevy, "Technologies for cofabricating MEMS and electronics," *Proc. IEEE*, vol. 96, no. 2, pp. 306-322, 2008.

[7] N. Xu, J. Sun, I-R. Chen, L. Hutin, Y. Chen, J. Fujiki, C. Qian, and T.-J. K. Liu, "Hybrid CMOS/BEOL-NEMS technology for ultra-low-power IC applications," *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2014, pp. 677-680.