Center for Energy Efficient Electronics Science

Center for Energy Efficient Electronics Science

Theme II - Nanomechanics

CHAPTER 0

Chapter 0: Energy-Efficiency Limit of Digital Computing

Tsu-Jae King Liu

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

Steady advancement in semiconductor process technology over the past 50+ years has enabled ever more functional and affordable electronic devices, resulting in dramatic impacts on virtually every aspect of life in modern society. This chapter will first explain why the capability of electronic computing devices has steadily improved over time, leading to today's high-performance cloud computing services as well as highly functional personal and mobile devices, such as laptop computers and smartphones. Then it will provide an introduction to the metal-oxide-semiconductor (MOS) field-effect transistor (FET) and how it is used as an electronic switch in integrated circuits (ICs) used for digital computing. Non-ideal transistor characteristics, which fundamentally limit the energy efficiency of a digital IC, will then be discussed, motivating the need for alternative switch designs to improve the energy efficiency of electronics in the future.

The reader is expected to have high-school-level knowledge of mathematics (algebra 2) and principles of electricity (current, voltage, electric power, resistance, capacitance), and after finishing the chapter, the reader will have a basic understanding of complementary MOS (CMOS) digital ICs and their energy-efficiency limit.

0.1 INTEGRATED CIRCUIT TECHNCOLOGY ADVANCEMENT

The key to continual improvements in computing performance and cost has been transistor miniaturization: the smaller a transistor is, the more compactly a microprocessor IC chip comprising interconnected transistors can be implemented (Fig. 0.1) [0.1] and hence the more chips that can be yielded from a single silicon wafer substrate of fixed size (*e.g.*, 300 mm diameter in state-of-the-art IC fabrication facilities), resulting in lower manufacturing cost per chip. Additionally, smaller transistors and shorter interconnecting wires (called interconnects) between them have smaller associated capacitance so that voltage signals propagate more quickly through a smaller chip, *i.e.*, the IC can be operated with a higher clock frequency. Advancements in IC manufacturing technology have enabled ever finer control of the thickness of thin (below one micrometer in thickness) films that are used to form the transistors and interconnects, as well as ever finer resolution of the lithographic process that is used to pattern the thin films. Accordingly, the minimum size of transistors has shrunk over time so that the number of transistors incorporated on a single IC chip has roughly doubled every two years according to Moore's Law (Fig. 0.2) [0.2] resulting in improved chip functionality as well as lower cost per function. Presently the most advanced transistors have features with minimum dimension below 10 nanometers, so that the most advanced microprocessors pack more than 100 million transistors within an area of one square millimeter [0.1].

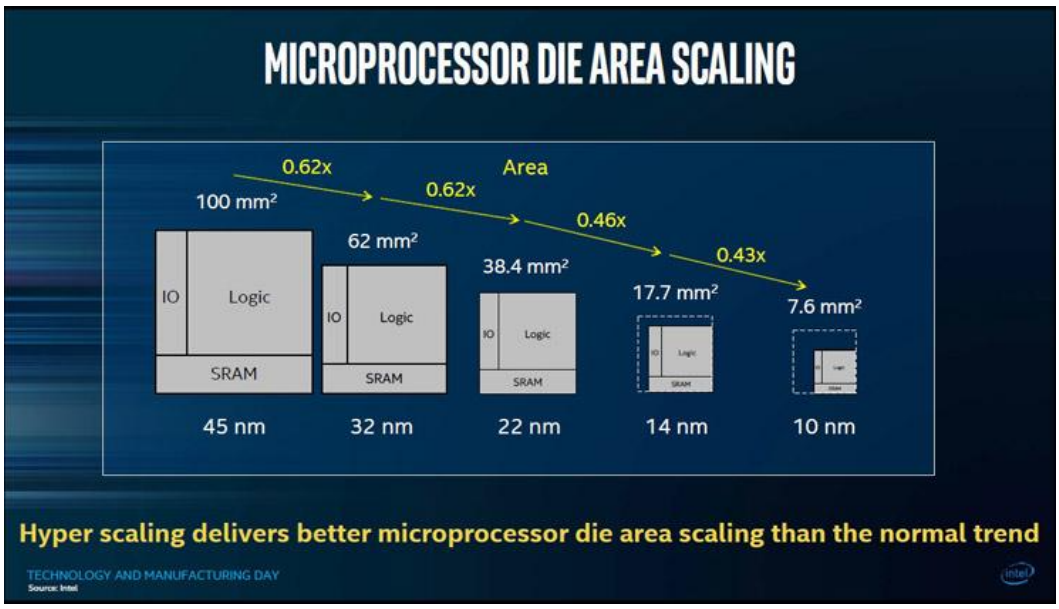


Figure 0.1 - Intel microprocessor chip size reduction with process technology advancement (to smaller minimum feature size) to enable higher transistor density. A microprocessor chip comprises blocks of circuitry that receive and send information (input/output, or I/O), perform computations on data (logic), and temporarily store data (static memory, SRAM). Reprinted with permission from [0.1].

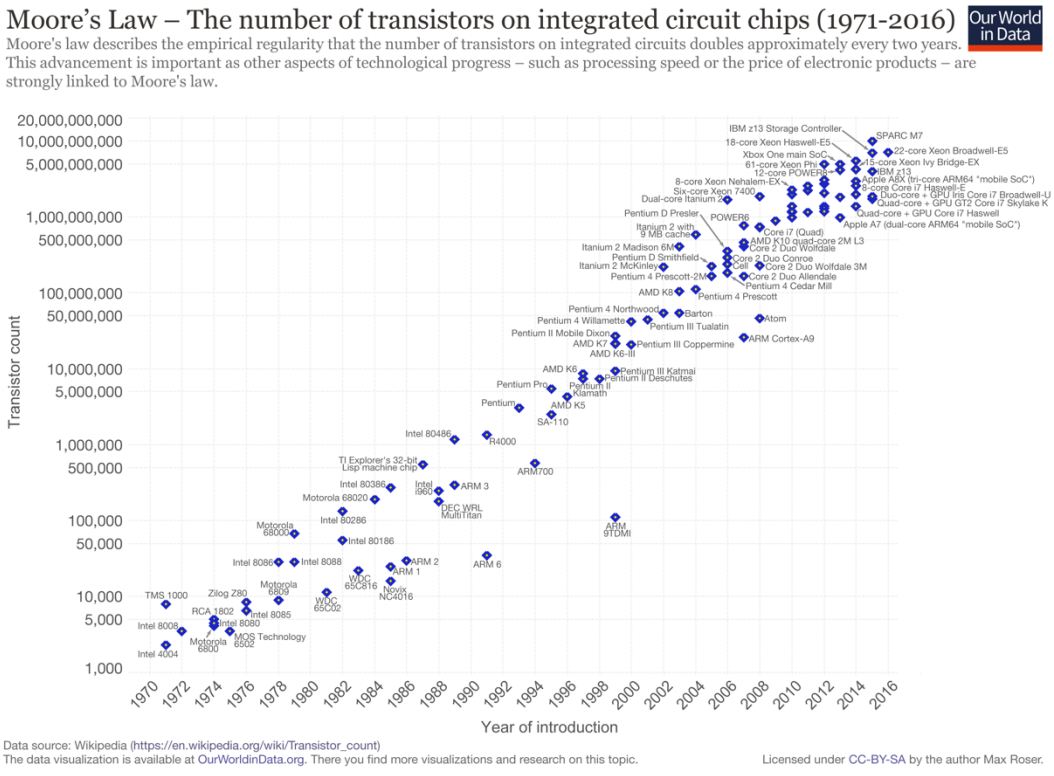


Figure 0.2 - Exponential growth over time in the number of transistors integrated onto the most advanced microprocessor chip, following Moore's Law. Reprinted with permission from [0.2].

0.2 MOSFET BASICS

0.2.1 Transistor Structure and Operation

A transistor is a solid-state device that is used to control the flow of electrons: electric current flows through a resistive path ("resistor") between two conductive terminals, under the control of a voltage signal applied to a third terminal along a direction transverse to that of the current flow. Various transistor designs have been developed over the years, each well-suited to a particular application. In very-large-scale integrated (VLSI) circuits used for digital computing, transistors function very simply as electronic switches. Due to its scalability to nanometer-scale dimensions, the MOSFET is the transistor design of choice for VLSI applications.

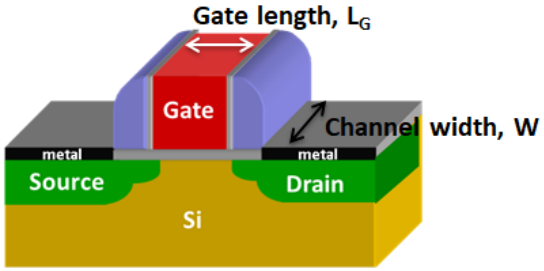


Figure 0.3 - Schematic illustration of a metal-oxide-semiconductor field-effect transistor (MOSFET). The most common semiconductor material used for integrated circuits is silicon (Si).

Fig. 0.3 illustrates the basic MOSFET structure: When a voltage V_{GS} greater than a threshold voltage (V_T) is applied between the Gate and Source terminals, the transistor is in the ON state and current (I_{DS}) can readily flow through the semiconductor channel region between the Source and Drain terminals if their electric potentials are unequal, *i.e.*, if there is a non-zero voltage difference V_{DS} between these terminals to cause electrons to drift from one of these terminals to the other. Note that the metallic gate electrode (shown in red in Fig. 0.3) is electrically insulated from the semiconductor channel region by a thin oxide layer (shown in light gray in Fig. 0.3) so that negligible direct current flows between the gate and semiconductor. The ON-state current (I_{ON}) flowing between the Source and the Drain regions increases super-linearly with increasing gate voltage overdrive $|V_{GS} - V_T|$, which has a maximum magnitude of $|V_{DD} - V_T|$ where V_{DD} is the power supply voltage for the circuit.

When the voltage difference between the Gate and Source terminals is zero (*i.e.*, $V_{GS} = 0$ Volts), the transistor is in the OFF state, meaning that a much lower level of current (I_{OFF}) flows through the semiconductor channel region between the Source and Drain. The Source and Drain semiconductor regions have high concentrations of impurities that make these regions electrically conductive, either with mobile positive charge (p-type conductivity) or with mobile negative charge (n-type conductivity). Typically the conductivity type of the channel region is opposite to that of the Source and Drain regions, so that a large electric potential barrier prevents the diffusion of mobile charges from the Source into the channel region in the OFF state, resulting in low I_{OFF}.

0.2.2 Sub-Threshold Current

The magnitude of the aforementioned potential barrier that impedes the flow of mobile charges from the Source region into the channel region of a MOSFET depends on the Gate voltage relative to the Source voltage (V_{GS}). Specifically, it decreases with increasing $|V_{GS}|$, by a proportionality factor that depends on the capacitance between the channel region and the Gate (C_{ox}) relative to the capacitance between the channel region and the Drain (C_D) and to the capacitance between the channel region and the semiconductor body deeper beneath the channel region (C_{dep}). This

proportionality factor is C_{ox}/C_{total}, where $C_{total} = C_{ox} + C_{dep} + C_D$. In other words, a change in V_{GS} (ΔV_{GS}) results in a change in potential barrier height equal to $\Delta V_{GS} \times (C_{ox}/C_{total})$.

Since the semiconductor is maintained at an absolute temperature (T) greater than 0 K, the mobile charges in the Source region have kinetic energy, with an exponential probability distribution; their population decreases exponentially with (linearly) increasing kinetic energy relative to kT, where k is the Boltzmann constant. As a result, the number of electrons in the Source region with sufficient kinetic energy to surmount the potential barrier (to diffuse into the channel region and drift to the Drain region) increases exponentially as the barrier height decreases. Hence as $|V_{GS}|$ increases, lowering the potential barrier, $|I_{DS}|$ increases exponentially, as illustrated in Fig. 0.4.

When the MOSFET is in the ON state (*i.e.*, when $|V_{GS}| > V_T$), the potential barrier between the Source region and channel region is very small so that I_{DS} is limited not by the rate of mobile charge diffusion into the channel region but by the rate at which mobile charges drift under the influence of a lateral electric field induced by the voltage difference (V_{DS}) between these terminals.

The inverse slope of the $\log(I_{DS})$ *vs.* V_{GS} plot in the subthreshold region (*i.e.*, for $|V_{GS}| < V_T$) is referred to as the subthreshold swing (S), and represents the amount of change in V_{GS} needed to effect 10× (one decade) change in I_{DS}:

$$S = (kT/q)(\ln 10)(C_{total}/C_{ox}) \qquad \text{Eq. 0.1}$$

where q is the electronic charge and kT/q is the thermal voltage. As can be deduced from Equation 0.1, the minimum value of S is $(kT/q)(\ln 10)$, which is approximately 60 mV/decade at room temperature (300 K). For I_{DS} to change by a factor of 10^5, then, the Gate voltage must change (swing) by at least (60 mV/decade)×(5 decades) = 0.30 V.

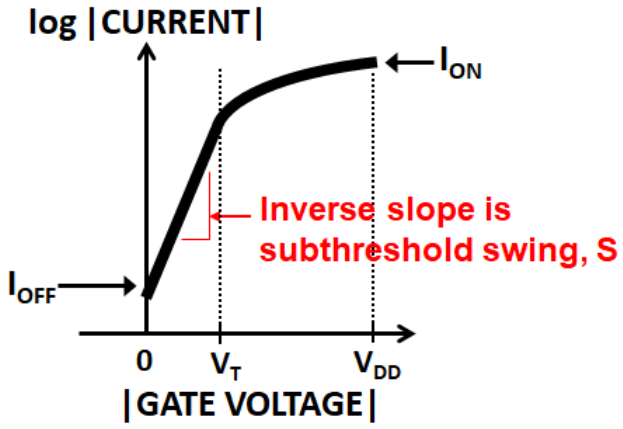


Figure 0.4 - Semi-log plot of MOSFET output (drain) current vs. input (gate) voltage characteristic. Below threshold (i.e., in the sub-threshold region), the current is an exponential function of the gate voltage. Above threshold, the current increases super-linearly with increasing gate voltage.

0.3 COMPLEMENTARY MOS (CMOS) TECHNOLOGY

0.3.1 n-channel and p-channel MOSFETs

As mentioned above, current flow in a MOSFET is controlled through the voltage applied to the Gate electrode relative to the voltage at the Source electrode (V_{GS}). In the ON state, an electrically conductive channel comprising mobile electronic charges of the same type as in the Source and Drain regions is formed at the surface of the semiconductor beneath the Gate electrode, electrically connecting the Source and Drain regions. If the channel comprises negatively charged electrons, induced by applying a positive Gate voltage relative to the source voltage ($V_{GS} > V_T$), then the MOSFET is an n-channel transistor and has n-type Source/Drain regions. On the other hand, if the channel comprises positively charged holes, induced by applying a negative Gate voltage relative to the Source voltage ($V_{GS} < -V_T$), then the MOSFET is a p-channel transistor and has p-type Source/Drain regions.

0.3.2 CMOS inverter circuit

Both n-channel and p-channel MOSFETs are used in the majority of IC chips produced today. In an n-channel MOSFET, a gate voltage that is higher than the source voltage is needed to form a conductive n-type channel to electrically connect the n-type source/drain regions, whereas in a p-channel MOSFET a gate voltage that is lower than the source voltage is needed to form a conductive p-type channel to electrically connect the p-type source/drain regions. If the gate electrodes of a pair of n-channel and p-channel MOSFETs are connected together and their source electrodes are biased as shown in Fig. 0.5, complementary switching behavior is achieved, *i.e.*, only one transistor is turned on at a time when the gate voltage is high (biased at V_{DD}) or low (biased at 0 V). This is advantageous for minimizing static power dissipation, to be only $I_{OFF} \times V_{DD}$ for the complementary MOS (CMOS) inverter circuit shown in Fig. 0.5. Note that the MOSFET drain electrodes are each tied to the output node, and that the n-channel device is used to connect and discharge ("pull down") the output node to 0 V when the input node (tied to the gates) is high; the p-channel device is used to connect and charge ("pull up") the output node to V_{DD} when the input node is low. The larger the transistor ON-state current, the faster the output node voltage is charged or discharged and hence the faster the circuit operates.

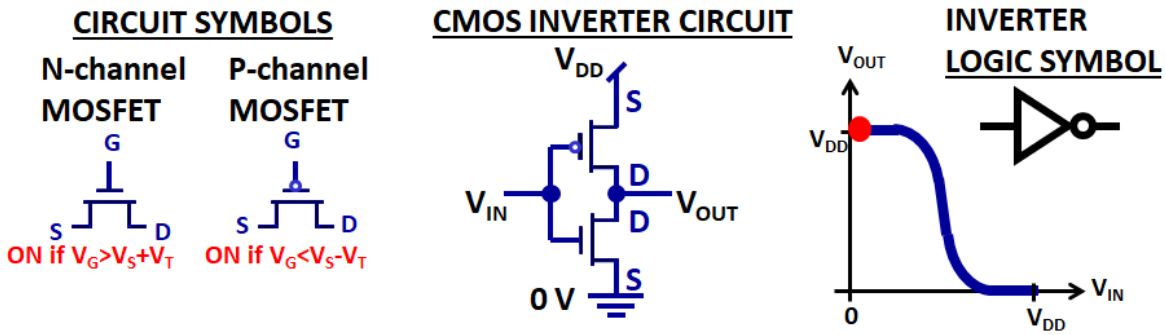


Figure 0.5 - (left) Circuit symbols for n-channel and p-channel MOSFETs, (center) schematic circuit diagram for a CMOS inverter circuit, and (right) voltage transfer curve and logic symbol for the inverter.

0.3.3 V_T Design Trade-off

The magnitude of the transistor OFF-state current (I_{OFF}) at V_{GS} = 0 V is proportional to $10^{-|V_T/S|}$. Therefore, a large value of $|V_T|$ is desirable for low I_{OFF}, for lower static power dissipation; however, a smalle value of $|V_T|$ is desirable for high I_{ON}, for faster circuit operation. Thus, there is a fundamental design tradeoff for V_T, as illustrated in Fig. 0.6.

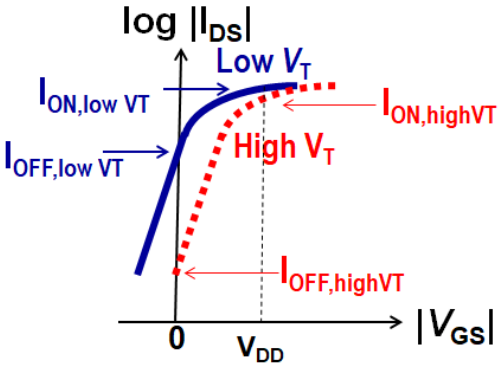


Figure 0.6 - Comparison of current-vs.-voltage (I-V) characteristics for low-V_T vs. high-V_T MOSFET designs.

0.3.4 CMOS Digital Logic and Memory Circuits

Digital computers use a binary number system (*e.g.*, a binary digit (bit) value of '0' is represented by a low voltage, while a bit value of '1' is represented by a high voltage). Computation is implemented with circuits called logic gates, which perform Boolean logic functions. Any logic gate can be constructed with a combination of n-channel pull-down devices and a complementary set of p-channel pull-up devices. (Each input voltage signal drives the gates of a pair of complementary n-channel and p-channel MOSFETs.) For example, a circuit that performs the NOT-AND (NAND) logic function is shown in Fig. 0.7. A microprocessor typically comprises many combinational (time-independent) logic circuits, clocked sequential logic circuits (whose output values depend not only on the present values but also on the past values of the input signals), and memory cells for temporary storage of data. The circuit configuration of a typical static random access memory (SRAM) cell, which has two stable states so that it can store one bit of information, is shown in Fig. 0.8.

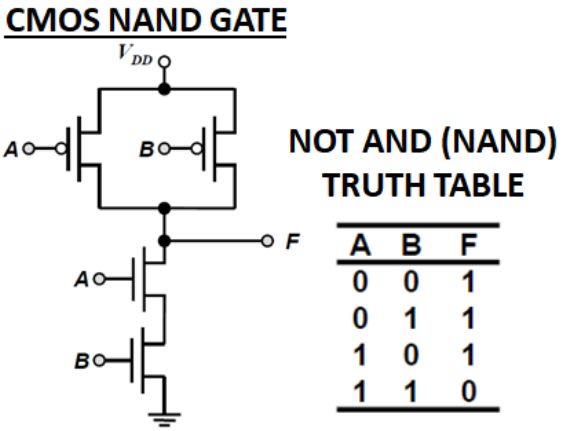


Figure 0.7 - Schematic circuit diagram for a NAND logic gate. When both inputs (signals A and B) are high (logic state '1') the two n-channel MOSFETs are ON, connecting the output to 0 V (logic state '0'); otherwise at least one of the inputs is low so that at least one p-channel MOSFET connects the output to V_{DD} (logic state '1').

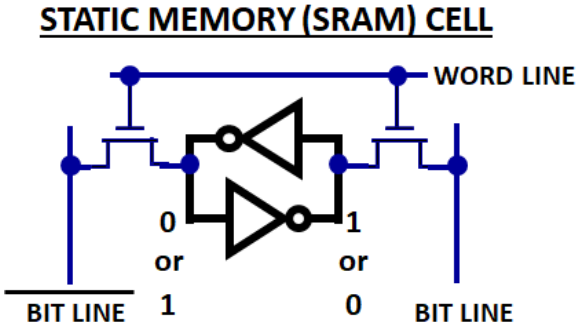


Figure 0.8 - Circuit schematic for a 6-transistor static memory (SRAM) cell comprising two cross-coupled inverters forming a bistable latch, and two n-channel transistors used to access the cell (connecting the storage nodes to the bit lines under the control of the word line voltage signal).

0.3.5 CMOS Digital IC Power Consumption

The total power (P_{total}) dissipated by a CMOS IC comprises two major components: dynamic dissipation associated with charging/discharging of nodal capacitances, which is proportional to the average capacitance charged every clock cycle (C_{eff}), the clock frequency (f), and V_{DD} squared; and static dissipation primarily due to the transistor OFF-state leakage of all of the logic gates:

$$P_{total} = (f)(C_{eff})(V_{DD})^2 + (I_{OFF,all})(V_{DD}) \quad \text{Eq. 0.2}$$

0.4 THE CMOS POWER CRISIS AND THE NEED FOR A NEW LOGIC SWITCH

Since I_{OFF} (hence static power consumption) increases exponentially with decreasing V_T, V_T has not been aggressively scaled down with each new generation of IC technology in recent years (Fig. 0.9) [0.3]. Since I_{ON} (hence circuit operating speed) decreases with decreasing gate overdrive $|V_{DD} - V_T|$, V_{DD} also has not been aggressively scaled down as transistor density has increased.

Because the transistor operating voltage (*i.e.*, V_{DD}) is no longer being reduced commensurately with transistor dimensions, chip power density has grown to be the dominant challenge for continued IC technology advancement. Currently, limitations of chip cooling technology (approximately 300 W/cm^2) constrain CMOS IC design. In order to increase transistor density (for lower cost per function and increased chip functionality) without increasing power density, V_{DD} must be reduced (to reduce active power dissipation) at the expense of performance (*i.e.*, no increase in clock frequency). This has forced the move to multi-core processors, *i.e.*, parallelism, to recover system throughput (Fig. 0.10). (A multi-core processor is a single computing chip with two or more independent processing units called cores.) The number of cores on a processor chip such as a central processing unit (CPU), graphics processing unit (GPU), or tensor processing unit (TPU) has increased in recent years with continued IC technology advancement to increase the number of transistors on a chip (Fig. 0.11) [0.4].

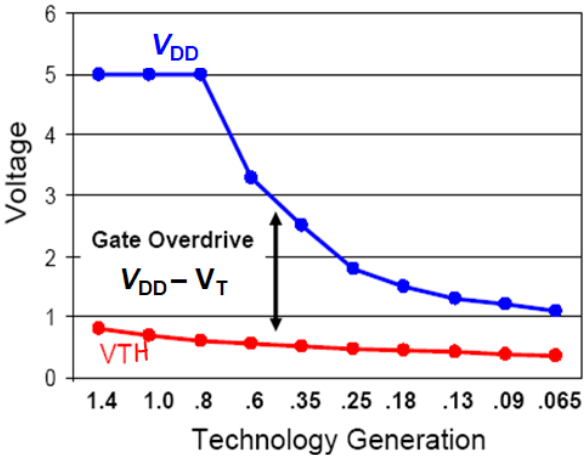


Figure 0.9 - Historical scaling of CMOS supply voltage (V_{DD}) and threshold voltage (V_T) reduction with technology advancement.

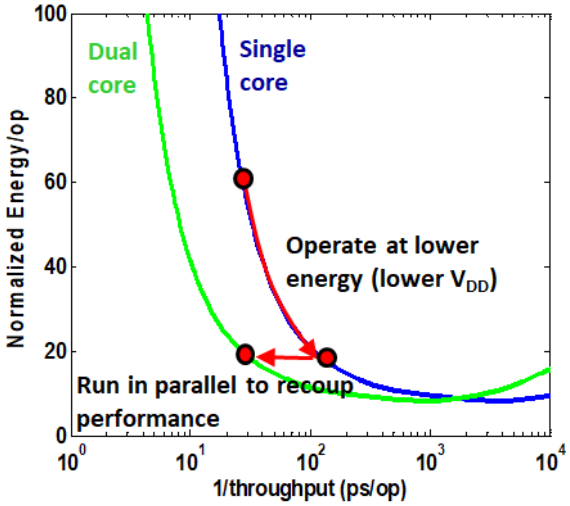


Figure 0.10 - CMOS energy-delay trade-off: To reduce the energy consumed in performing a digital operation (Energy/op), the circuit operating voltage (V_{DD}) should be reduced; since transistor ON-state current decreases super-linearly with V_{DD}, however, this results in slower circuit operating speed. With multiple cores operating in parallel, the information processing throughput of a chip can be improved to compensate for this.

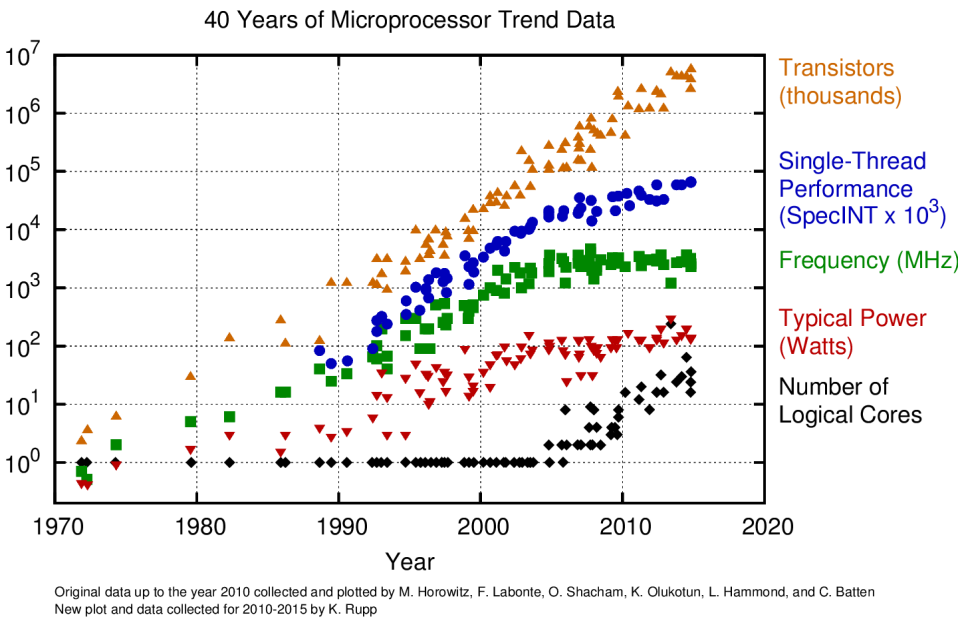


Figure 0.11 - Microprocessor trends over the past 40 years. Reprinted with permission from [0.4].

The degree to which parallelism can be used beneficially will eventually reach a limit because of a fundamental lower limit for the energy required by a CMOS digital IC to perform a digital operation (E/op). This lower limit exists due to transistor off-state leakage current (*i.e.*, non-zero I_{OFF}): As V_{DD} is reduced to lower the amount of dynamic energy required to perform a computation, however, the amount of time required to perform the computation (t_{delay}) increases, resulting in the amount of energy wasted due to static power dissipation (proportional to $t_{delay} \times I_{OFF} \times V_{DD}$) increases. Alternatively, V_T can be reduced together with V_{DD} to maintain t_{delay}, but then I_{OFF} would increase exponentially (cf. Fig. 0.6). When V_{DD} is reduced to V_T, a minimum in total energy (comprising dynamic and static energy components) is reached, *i.e.*, any further reduction in V_{DD} will not result in reduced total energy required to perform the computation (Fig. 0.12). This is because I_{ON} decreases exponentially – and hence t_{delay} and static power dissipation increase exponentially – with decreasing V_{DD} below V_T.

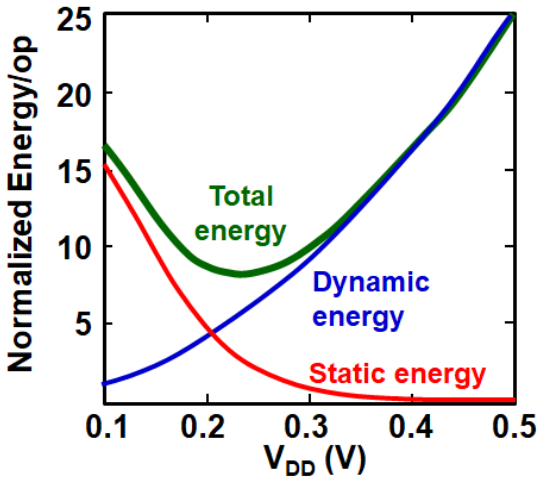


Figure 0.12 - Impact of supply voltage (V_{DD}) scaling on the energy consumed by CMOS IC in performing a digital operation

The solution to this crisis is a new logic switch that can operate with much smaller gate voltage swing than a MOSFET, *i.e.*, smaller S, to allow for lower voltage operation with high ON/OFF current ratio. The lower V_T provides for a lower limit in E/op, *i.e.*, improved energy efficiency. Accordingly, this e-book covers in detail alternative solid-state switch designs that can overcome the fundamental switching steepness limit of 60 mV/decade for the MOSFET. Circuit design co-optimization to maximize the computing performance and energy efficiency of ICs comprising emerging new switch technology is also discussed.

SUMMARY

Non-zero transistor off-state leakage current is the root cause of the computing power crisis that the semiconductor electronics industry faces today. This has motivated research and development into alternative logic switch designs that can operate with much higher ON/OFF current ratio, to enable the operation of digital ICs with much lower voltage below 0.25 V (down to the millivolt range) for more energy-efficient computing in the future.

REFERENCES

[0.1] Mark Bohr, "Leading at the Edge," presentation at Intel Technology and Manufacturing Day (San Francisco, California, USA), March 28, 2017.

[0.2] https://graylinegroup.com/microprocessors-line-spacing-end-era/

[0.3] Paul Packan, Short Course lecture notes, 2007 IEEE International Electron Devices Meeting.

[0.4] https://www.karlrupp.net/2015/06/40-years-of-microprocessor-trend-data/

CHAPTER 1

Chapter 1: Introduction to MEMS/NEMS

Jeffrey H. Lang

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

Nano/Micro ElectroMechanical Systems (NEMS/MEMS) are generally regarded as very small, often microscopic, actuators and sensors fabricated with process technologies borrowed from, or at least related to, those used to fabricate transistors and transistor circuits. MEMS/NEMS have many advantages over conventional actuators and sensors. Their small size and light weight make MEMS/NEMS sensors ideal for many applications. They can, for example, be the most appropriate sensors and actuators with which to interface to tiny worlds, such as microbiological or chemical systems. These sensors and actuators are also generally rugged and low-power devices. Moreover, if fabricated in adequately large quantities, MEMS/NEMS can also be very inexpensive.

MEMS typically have critical dimensions that range from a fraction of a micrometer to hundreds of micrometers, while NEMS typically have critical dimensions that range from a fraction of a nanometer to hundreds of nanometers. Besides their difference in size, a significant difference between MEMS and NEMS is the manner in which their materials are treated. MEMS are large enough that the materials from which they are fabricated are generally treated as bulk materials. On the other hand, the fabrication and operation of NEMS is often carried out on a molecule-by-molecule basis.

MEMS were substantially introduced in the 1980s [1], and subsequently developed for sensing and actuation applications in many sectors. The earliest MEMS were truly electromechanical. That is, they relied on electric fields and forces to sense and actuate rigid-body structures. Early microscale pressure sensors, accelerometers and motors are examples of such devices. Today, however, MEMS are varied in nature. They often involve magnetic, piezoelectric, fluid and/or thermal physics, all in an effort to provide ever better sensors and actuators [2].

Early MEMS were fabricated predominantly from silicon. Silicon remains the most common MEMS material today because, in its crystalline form, it is nearly electrically and mechanically perfect. It is also the second most abundant material on earth, making it relatively inexpensive. Importantly, the electrical conductivity of silicon can be precisely adjusted from nearly insulating to semi-conducting to highly conducting through the use of doping impurities [3]; it is the semiconducting behavior of silicon that makes transistors and modern electronics possible [3]. However, silicon is also an impressive mechanical material.

Silicon is as strong as steel but as light as aluminum. Further, it exhibits almost no mechanical stress-strain hysteresis. Thus silicon structures can be bent an essentially unlimited number of times without fatiguing or breaking, making silicon a highly reliable mechanical material. A lack of hysteresis also means that such deformations occur essentially without loss. Silicon possesses a high thermal conductivity and a low thermal coefficient of expansion in compared to metals, making silicon structures much more immune to thermal shock, and more thermally precise.

Finally, while silicon can be etched, it is immune to many common solvents, etchants and corrosive environments.

From a manufacturing viewpoint, silicon can be "machined" with exceptional precision and repeatability at the micro-scale and below through photolithographically controlled processes [1,2,4] which will be discussed in Chapter 2. It can be both isotropically and anisotropically etched, and silicon wafers can be bonded together nearly seamlessly. This bulk micromachining enables the manufacturing of MEMS exhibiting a wide range of simple to complex geometries. Silicon can also be combined through surface micromachining with deposited and etched layers of poly-crystalline silicon, silicon dioxide, silicon nitride, various metals and other materials, enabling a vast array of tiny sensors and actuators.

MEMS are now ubiquitous in everyday life, and a substantial industry has formed around them as indicated in Figure 1. For example, a common smart phone can contain many MEMS sensors and actuators including accelerometers, gyroscopes, compasses, microphones, pressure sensors and filters. MEMS sensors and actuators are also common in automobiles, printers, and optical and medical devices, for example, among many other products. NEMS are a much newer technology, and only now are undergoing rapid development. NEMS are already beginning to appear in the form of biological and chemical sensors, for example, making use of new materials such as carbon nanotubes, graphene and biochemically-active thin films. They are also likely to appear as interfaces between electronics and biological nervous systems.

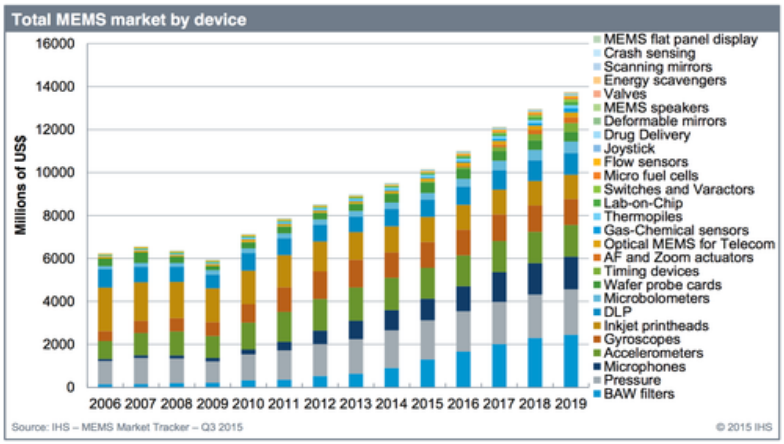


Figure 1: total MEMS market by device/application as of 2015.

Figure 2 shows several illustrative MEMS devices used in everyday products. The digital mirror device in the top row is a component of the TI digital light processor used in projectors. The print head in the top row is a component of an HP ink-jet printer. The Bosch pressure sensor is used in automotive applications, among others. The Knowles microphone, ADI accelerometer and ST gyroscope shown across the bottom row are representative of MEMS devices that are used in a common smart phone.

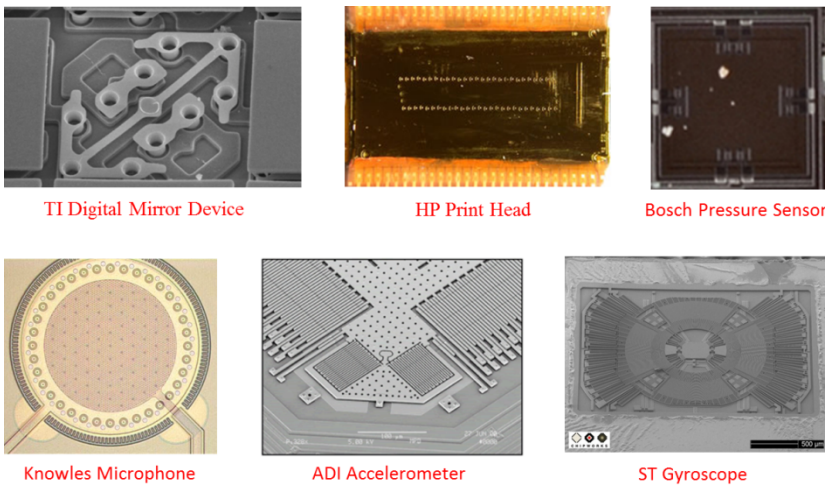


Figure 2: illustrative MEMS devices used in everyday products.

As a further example of a MEMS/NEMS device, one which is a focus of this eBook, consider the greatly simplified four-terminal MEMS relay, or switch, illustrated in Figure 3. The relay comprises four metallic conductors that are labeled A, B, C and D. Conductors A and D are the switched terminals of the relay while conductors B and C are its actuation terminals. Conductor A takes the shape of a cantilever, and overhangs the other three conductors. As fabricated, the relay is as shown to the left, and all four conductors are electrically isolated from one another. In particular, the electrical path between conductors A and D is broken by an airgap. However, when a voltage is applied between conductors B and C as shown to the right in the figure, charges are induced in the conductors. The attraction between these charges across the airgaps that separate them causes the cantilevered conductor A to bend down until it contacts conductor D; see Chapter 3. At this point the relay closes, allowing current to flow along the dotted path in Figure 3. When the electrostatic actuation is removed, the spring-like behavior of the cantilever opens the relay, returning it to its as-fabricated state. The fabrication of such relays, and others like it, through surface micromaching is discussed in Chapters 4, 7 and 8, and also in [2,4].

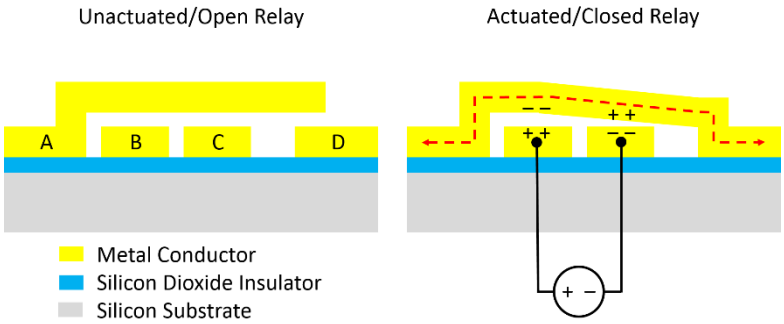


Figure 3: an unactuated/open relay is shown to the left while an actuated/closed relay is shown to the right. The relays comprise four metallic conductors labeled A, B, C and D. Actuation is carried out electrostatically. Once closed, the actuated relay forms a closed electrical path between conductors A and D along the dotted red path.

Given the manner in which the relay operates, it is important to ask whether MEMS/NEMS relays can switch fast enough to support modern computing. Indeed, transistors today can switch on (closed) and off (open) billions of times per second, and it is surely challenging for an electromechanical relay to switch that quickly. However, as MEMS/NEMS relays become smaller, they also become faster for reasons rooted in their scaling. To illustrate this, let x be a characteristic scaling by which all relay dimensions are varied. In this case, the mass of the cantilevered relay conductor scales as x^3, while the stiffness of the cantilever spring scales as x [2]. Thus, the

cantilever resonance frequency ($\sqrt{stiffness/mass}$), which determines the time required to open the relay, scales as x^{-1}; the time itself scales as x. Similarly, the surface area over which the electrostatic actuation acts scales as x^2. Thus, the actuated acceleration (force/mass) of the cantilever scales as x^{-1}, assuming that the electrostatic pressure remains constant. Further, the distance which the cantilever must travel to close scales as x. Taken together, the time required to close the relay ($\sim\sqrt{distance/acceleration}$) also scales as x. The x-scaling of the opening and closing times shows that smaller relays switch faster in proportion to the downwards scaling of their size. Indeed, it is argued in [5,6] that very small NEMS relays might become capable of switching 100 million times per second. While this is not quite as fast as transistor switching, modifications to the style of computing-logic design employing them can make MEMS/NEMS switching yet more competitive [6].

Figure 3 also indicates two significant advantages of MEMS/NEMS relays. The first advantage is that, when open, Conductors A and D are separated by an airgap. This results in an off state that exhibits very-low leakage. The second advantage, as discussed above, comes from relay scaling. The smaller they are the faster they are. As discussed in Chapter 0, both advantages are critical to implementing low-energy computation.

While the basic operation of MEMS/NEMS relay is quite simple, as illustrated in Figure 3, there are significant challenges to making MEMS/NEMS relays successful. The most significant of these challenges is stiction, that is, the tendency of van der Waals forces to hold the relay closed once its contacts touch [7,8]; see Chapters 5 and 6. It is the spring stiffness of the cantilever in Figure 3 that must open the relay against stiction, and this spring stiffness scales as x. Thus, the stiffness reduces with the size scale. Of course, the spring can be made stiffer to overcome stiction by making the cantilever thicker, that is, by not scaling all dimensions equally. However, a stiffer spring then requires a greater electrostatic force to close the relay, and hence a larger voltage. This is directly contrary to the objective of designing of a relay exhibiting a low switching energy. Therefore, stiction must be carefully managed in micro/nano-scale relays. As discussed in Chapters 4, 7 and 8, it is done so here through a combination of careful dimensional relay design, molecular anti-stiction coatings, and/or the use of tunneling or strain-modulated conductivity to avoid contact.

The remainder of this e-book examines the development of digital electronic switches, or relays, based on MEMS and NEMS technologies. Special focus is placed on low-voltage, and hence low-energy, switching. It culminates in the implementation of digital logic circuits using such relays in Chapter 9.

REFERENCES

[1] K. E. Petersen; “Silicon as a mechanical material”; *Proceedings of the IEEE*, *70*, (5), May 1982.
[2] S. D. Senturia; *Microsystem Design*; Kluwer Academic Publishers, 2001.
[3] S. M. Sze; *Physics of Semiconductor Devices*; John Wiley & Sons, 1981.
[4] R. Ghodssi and P. Lin, editors; *MEMS Materials and Processes Handbook*; Springer 2011.

[5] F. Niroui, A. I. 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*, August 5, 2015; doi: 10.1021/acsnano.5b02476.
[6] F. Chen, H. Kam, D. Markovic, T.-J. King Liu; V. Stojanovic and E. Alon; "Integrated circuit design with NEM relays"; IEEE/ACM International Conference on Computer-Aided Design, 750-757, 2008.
[7] R. T. Howe and R. Maboudian; "Adhesion in surface micromechanical structures; *Journal of Vacuum Science and Technology*", *B*, *15*, 1, 1997; doi: 10.1116/1.589247.
[8] C. Pawashe, K. Lin, and K. J. Kuhn; "Scaling Limits of Electrostatic Nanorelays"; *IEEE Transactions on Electron Devices*, *60*, 9, 2936-2942, September 2013.

CHAPTER 2

CHAPTER 2: FABRICATION OF MICRO/NANO SYSTEMS

Mayuran Saravanapavanantham[1], Jatin Patil[1], Farnaz Niroui[1],
Jeffrey H. Lang[1], Vladimir Bulović[1]

[1.]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

For a technology to be mass produced and industrially relevant, it is imperative that there exist a set of reliable tools and techniques which allow for large scale, high throughput fabrication. Microfabrication has its uses in a variety of applications, having started with the fabrication of integrated circuits. In recent years, it has also enabled the advent of certain MEMS devices (accelerometers, gyroscopes) as integral components in consumer electronic devices. In this chapter, we will discuss conventional microfabrication techniques, as well as novel approaches to defining structures not achievable through standard fabrication processes.

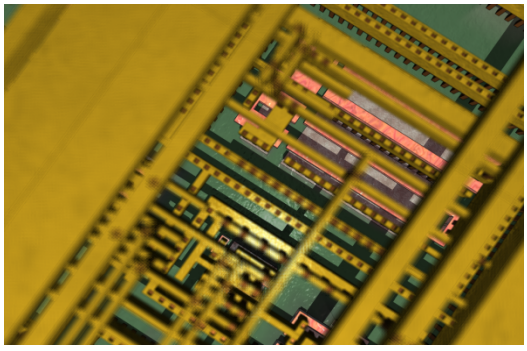


Figure 2.1 - A computer sketch of a real microfabricated chip design with interconnects. This graphic shows color contrast between copper interconnects, silicon, silicon oxide, and other metals. (Source: David Carron, Wikimedia Commons)

Essentially, the term *microfabrication* refers to a set of techniques employed to mass-produce mechanically and electrically active systems ranging in size from a few microns to tens of nanometers. First we shall discuss three major branches of microfabrication (namely lithography, thin-film deposition and etching) which together allow for the realization of **m**icro**e**lectro**m**echcanical **s**ystems (MEMS).

The goal of any fabrication technique is to manifest a design through pattern transfer – whether it be 3D printing, where a CAD pattern is extruded into a 3D structure, or Computer Numerical Control (CNC) machining where through selective material removal a 3D pattern is formed. A distinguishing feature of microfabrication is that it enables the replication and **simultaneous** fabrication of the same structure.

In microfabrication, the medium of choice to transfer patterns is light – lending the name **photolithography** to this pattern transfer technique. Through photolithography, the pattern is transferred to a light sensitive polymer which then acts as a mask on the substrate of choice upon which a combination of thin-film deposition and subtractive etching techniques help manifest a 3D mechanically and electrically active structure. Prior to delving into the details of the various processes, let us consider a simple example of how microfabrication is used to transfer a pattern onto a substrate (eg. silicon).

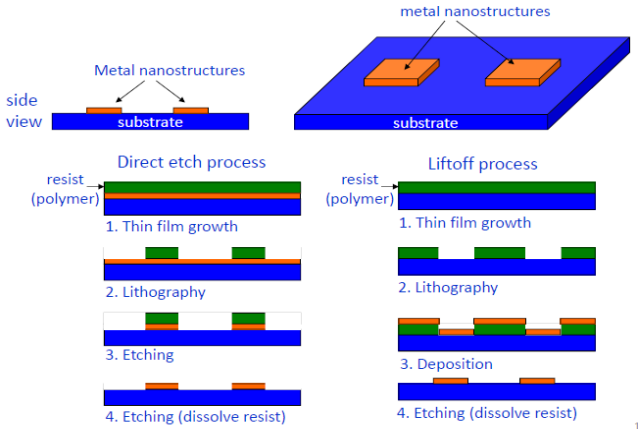


Figure 2.2 - A schematic description of how the same pattern can be realized through two general processes (direct etch vs. lift-off). (Source: Dr. Bo Cui, University of Waterloo)

The idea behind the **direct etch process** is to deposit a layer of the desired material directly onto the substrate and realize patterns in it through removal of certain regions. After depositing the material, a light-sensitive polymer is deposited and is exposed to light in a way that it remains on top of the regions of interest where the material should remain intact. The rest of the polymer is washed away and yields a masked structure which can then be etched (material removal). Upon etching the regions not protected by the polymer mask is stripped away. Once the protective polymer is dissolved away we are left with the desired structures on the substrate.

On the contrary, the **lift-off process** involves patterning the photo-sensitive polymer layer in such a way so that regions of interest on the substrate where we want to have the structures is left exposed and the material of choice is deposited on top of this. This way, the material only comes in contact with the substrate where there is no polymer and once the polymer is removed, it takes away all the material that happened to be deposited above it, leaving behind a structure identical to what was achieved through the direct-etch process. Now that we have seen an example of a simple microfabrication process, let us delve into the details of each aspect of it.

2.1 Lithography

From studying the simple process explained in Figure 1 we are aware that photolithography is used to define the patterns of interest. As the name implies, information of the pattern and its features is transferred by means of light (hence the prefix *photo-)*. A photolithography system simply shines light (most commonly in the ultraviolet regime) onto a photo mask (as shown in Figure 2.3) which is placed on top of the substrate coated with a photoactive compound. The **photomask** has your pattern of interest and thus controls which parts of the substrate below receive the incoming light. After light exposure, the substrate is immersed in a **developer** and depending on the chemistry of your polymer, the region that was exposed to the light will either be dissolved off the substrate while the unexposed region remains, or vice versa.

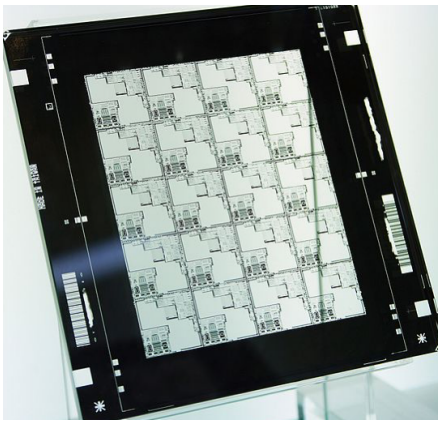


***Figure 2.3** - Photograph of a photo mask. The photomask limits light transmission onto the underlying photoresist as per the patterning interests of the user. Photomasks are generally made of glass or quartz and patterned in chrome. (Source: Peelldeen, Wikimedia Commons)*

The underlying photoactive polymer is known as a **photoresist**. Photoresists can be of two kind, **positive** and **negative**. Positive resists, when exposed to light, dissolve at a higher rate in the developer than the unexposed regions. Negative resists work in the opposite way, where once exposed to light the regions become hardened and remain on the substrate while the unexposed regions get washed away.

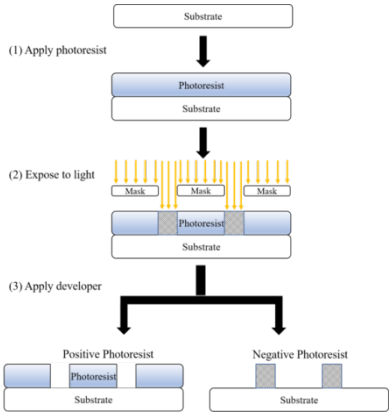


Figure 2.4 - *Positive vs. Negative photoresists. (Source: May Lam, Wikimedia Commons)*

As briefly mentioned earlier, light in the ultraviolet regime is mainly employed in photolithography. This light can be generated from a mercury arc lamp where the characteristic electronic transitions of excited mercury vapor yields distinct wavelengths of light which carry sufficient energy to induce chemical changes in the absorbing photoresist. Diffraction limitations imply that, to define smaller features, light of smaller wavelength (higher energy) must be used. **Excimer** (excited dimer – of a halogen and a noble gas) lasers serve as bright sources for such smaller wavelengths.

Likewise, for extremely fine features one can transition to even smaller wavelength light sources such as x-rays from a beam line or electrons. **Electron-beam lithography (EBL)** is a vital technique for NEMS processes where features on the order of a few nanometers are necessary. Unlike photolithography, EBL involves a single finely focused beam of electrons drawing out the pattern of interest on the photoresist rather than the entire pattern transfer taking place in one shot.

In EBL, thin-films of long polymeric chains are used as the photosensitive medium for patterning. First, polymethylmethacrylate (PMMA) thin films are deposited via spin-coating onto the surface of interest. Next, the user defines the pattern of interest on a design software and as per the design, an electron beam is focused onto the substrate and drawn along the user-defined path. The high-energy electron beam has sufficient energy to break chemical bonds in the polymeric chain, thereby reducing the chain-length of the polymers. This in turn, increases the solubility of the polymer in the **developer** (methyl isobutyl ketone - MIBK). Therefore, upon exposure and development, we are left with nanoscale features defined along the path of the electron beam.

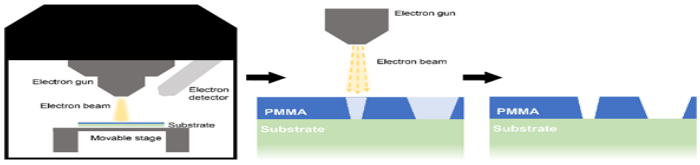


Figure 2.5 - *A schematic of the electron beam lithography chamber is shown to the left. The substrate (usually silicon coated with poly-methyl methacrylate (PMMA) is placed in a low-pressure chamber (~1E-6 Pa). Then, an electron gun is scanned across the sample. The electrons break the covalent bonds of the PMMA, making it into smaller molecules. These electron-exposed regions are soluble in a "developer" solution, while the unexposed regions are not. This leaves a pattern where the substrate is not exposed to electrons.*

An alternative lithography technique to define nanoscale features involves physical pattern transfer - that is, using a mold and pressing it into a polymer film to leave behind a 3D pattern. **Nanoimprint lithography** is carried out with a master mold that is created with previously

described lithography techniques (e-beam). Once prepared, the master is impressed upon a heated (soft) resist which is the cooled down before releasing the mold. The cooled resist retains the shape left behind by the mold. Once the resist has been patterned, it can be permanently transferred into the underlying substrate via post-processing (e.g. reactive ion etching). One advantage of nanoimprinting is that it allows for high-throughput processing of nanoscale features. Once the mold has been manufactured, it can be used over and over, thereby patterning all nanoscale features simultaneously, rather than sequentially as done with e-beam lithography.

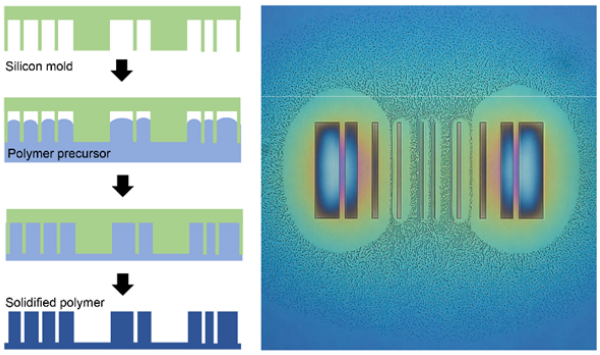


Figure 2.6 - *To the left is a schematic of nanoimprint lithography. A mold (green) is impressed upon a polymer film (blue), and removed to leave behind a pattern mirroring the recesses in the mold. To the right is an image of such patterns defined by imprinting. (Source: Felix Trier and Adam Andersen Læssøe, Wikimedia Commons)*

2.2 Thin-Film Deposition

In the pattern transfer process, it is imperative one has a set of techniques at their disposal which allow for the deposition of various materials on the bulk substrate. In general, when compared to the substrate upon which the pattern is defined, the subsequently deposited layers are much thinner (hence the name, thin-film deposition). Thin-films can be deposited either through **chemical** or **physical** means. Chemical means involve the formation of the material of interest on the substrate through a reaction. Whereas, physical means involve the deposition of a material that is located elsewhere and simply transported onto the substrate (for example in the form of an evaporated gas, which solidifies on the substrate). Let us briefly look at a series of chemical and physical deposition techniques that are commonly employed.

2.2.1 Spin-coating

In general, spin coating is the process of depositing a thin-film of a material of interest (polymers, molecules, nanoparticles, etc.) by dissolving it in a solvent, dispensing a small amount of solution on a solid surface and spreading it over the surface using centrifugal forces. The machine used to carry out this process is known as the spin-coater. The thickness of the resulting film is a function of many parameters: volatility of the solvent, solution concentration, viscosity, and spinning speed. Dilute solutions, and high spinning speeds result in thinner films. Commonly, solvents with low boiling points are used to allow for rapid solidification of the film. Heating the substrate after spin-coating allows for removal of residual solvent. Spin-coating is the deposition technique of choice for depositing photoresist prior to lithography.

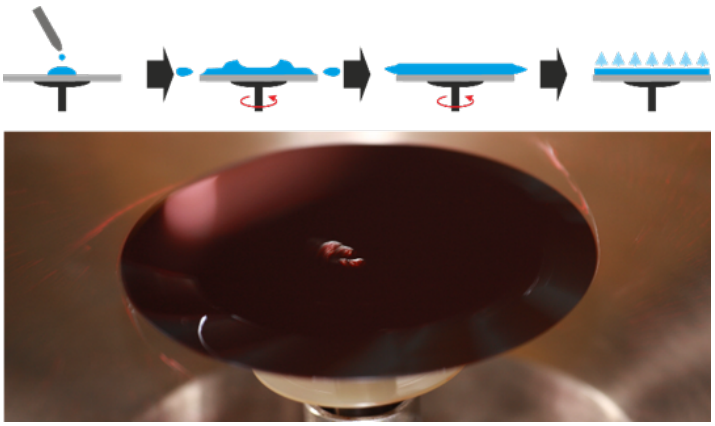


Figure 2.7 - A schematic of the spin-coating process, and a photograph of the process underway. First the substrate is mounted on the spin coater and the material is deposited. Next, as the spin-coater starts spinning the material spreads out due to centrifugal forces and some of it also get spun off the substrate. As the solvent evaporates the film solidifies and reaches an even thickness across the wafer. (Source: Stefan Reich and Sei, Wikimedia Commons)

2.2.2 Evaporation

As the name suggests, evaporation simply involves heating the material of interest from a source to the point of evaporation and then having it condense on your substrate in the form of a thin-film. Evaporation can be achieved in two ways – **thermal** heating of the entire source or **electron-beam** local heating of the source. In both situations, the process takes place in vacuum to ensure that once the material has evaporated it does not collide with particles in the atmosphere and deviate from its path towards the substrate. Of course, the process is not as simple as described here. One must take into account the way the material comes off the substrate (the evaporated plume) as well as how it reaches the substrate and what factors contribute to the formation of an even film free of defects. Figure 2.8, below, gives a schematic of this deposition technique. Evaporation is a physical technique as there is no chemical reaction involved – simply a series of physical phase changes.

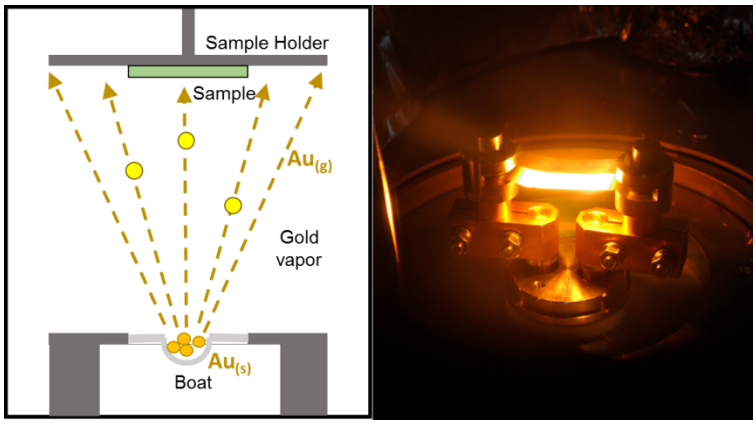


Figure 2.8 - To the left is a schematic of a thermal evaporator under operation. Heated material, held in a boat evaporates and travels towards the sample mounted on a substrate holder. Upon reaching the substrate, the vapor cools and solidifies into a thin film. To the right is a photograph of a thermal evaporator under operation, the heated boat glowing from having reached very high temperatures. (Source: Inmodus, Wikimedia Commons)

2.2.3 Sputtering

Sputtering is another physical deposition technique where the material is removed from the source by momentum transfer. Gas molecules, conventionally argon gas, is ionized and the charged particles are accelerated towards the source. The momentum transferred from the gas to the source "sputters" off atoms of the source material which then travels towards the substrate and settle on it. This process is not carried out in high-vacuum as one needs a steady presence of a secondary gas (i.e. Argon) to sputter the materials. Typically, to reduce contamination, the chamber is maintained at a low pressure (1E-6 to 1E-8 Torr) prior to sputtering to ensure that contaminants are not present in the chamber. During sputtering, the

chamber is filled with pure argon gas. Figure 2.9, below, provides a schematic diagram of the sputtering process. Sputtering is superior to evaporation in certain ways – it yields denser films and better step coverage. However, it suffers from substrate damage due to energetic bombardment of the material, as well as lower deposition rates.

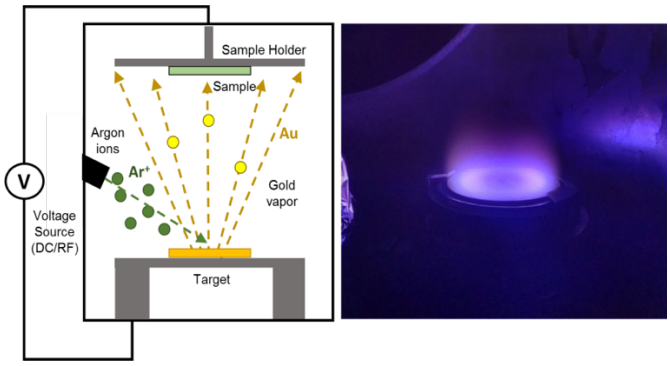


Figure 2.9 - *To the left is a schematic of a sputtering process in operation. Argon ions bombard the target, knocking out material which travels towards the sample mounted on the substrate holder and form into a thin film. A potential difference is maintained in the chamber to accelerate the argon ions formed in a plasma towards the target, allowing it to impart enough momentum into the target to eject material for deposition. To the right is a photograph of a sputtering process underway (image of the target being bombarded with argon ions. The characteristic purple glow results from the plasma employed to form the argon ions. (Source: Inmodus, Wikimedia Commons)*

2.2.4 Chemical Vapor Deposition (CVD)

As the name indicates, this is a chemical process – where the material is deposited on the substrate by means of chemical reactions. The CVD process starts off by introducing reactive gases to the chamber where the substrate resides. The gases are then activated by means of heat or plasma. The activated gases adsorb to the substrate surface and the chemical reaction takes place on the surface itself, forming the material of interest. Here lies the main difference between physical and chemical techniques. In chemical techniques the material is formed on the substrate surface whereas in physical techniques it is simply transferred from one location to another. Volatile by-products of the CVD surface reaction are then carried away from the substrate to allow for more incoming gas molecules to continue the reaction. A variety of enhancements such as the inclusion of a plasma source for activation of the gas, and the reduction of the chamber pressure can be made to this simple description of a CVD system, to achieve better results.

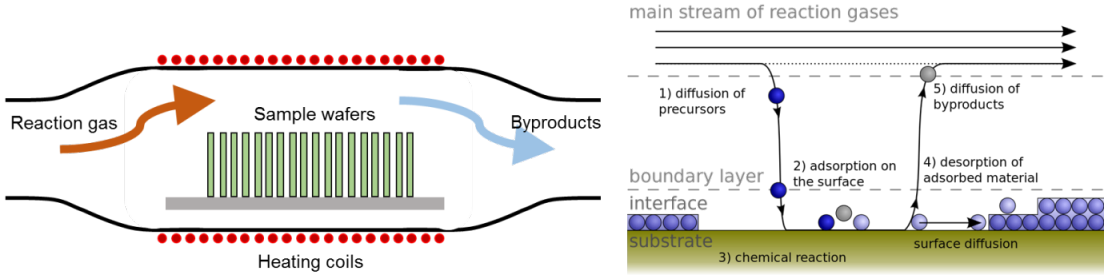


Figure 2.10 - *To the left is a cross-section view of a CVD chamber. Substrates are loaded into a quartz furnace surrounded by heating coils. Reaction gases are fed into the chamber from one side and byproducts collected from the other end. As the reaction gases reach the substrate surface, they undergo a chemical reaction (driven by the heat from the coils), depositing a thin-film of the desired material. Byproducts from the chemical reaction diffuse back into the flow of gases. To the right is a breakdown of each step involved in the CVD process once the chemical reaches the substrate. Mass transport considerations must be taken into account when planning CVD processes as they control the limiting steps in the reaction process. (Source: Cephieden, Wikimedia Commons)*

2.2.5 Atomic Layer Deposition (ALD)

ALD is yet another chemical deposition technique but is distinguished by the fact that unlike CVD where all precursors are introduced together, the reaction is broken into half reactions where the precursors are introduced separately and each half reaction takes place at different stages. One precursor gas is introduced and it forms a monolayer on the substrate. Then the second precursor is introduced and it reacts to form one monolayer of the final material of interest. This process repeats and we achieve atomic height control of the films we deposit – we can count the number of cycles we carry out and relate that to the number of monolayers deposited. Atomic layer deposition also benefits from nearly perfect step coverage, but suffers from a slow deposition rate.

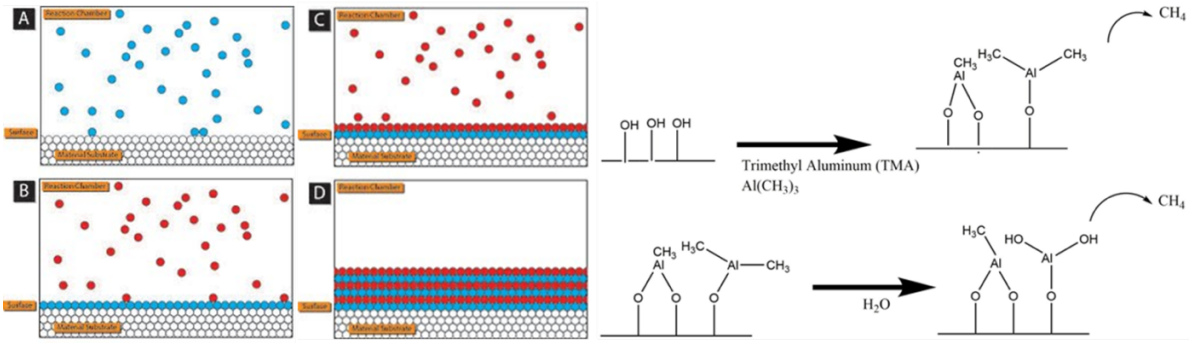


Figure 2.11 - *To the left a schematic of the ALD process. ALD relies of half-reactions reaching saturation to form atomically thin layers. Essentially, each reaction is reduced into two half-reactions. The first reactant is introduced into the system and allowed to saturate the surface of the substrate. Once saturated, the second reactant is introduced, reacting with the previously deposited layer, forming one atomic layer of the desired material. This process is repeated until a desired thickness is reached. To the right is an example of such two half reactions leading to the formation of Alumni (Al_2O_3) films. (Source: Mcat chem446, Wikimedia Commons)*

2.3 Etching

Just as how thin-film deposition addresses the needs for additive techniques, etching processes offer the subtractive tools necessary to remove materials from the surface. In a similar fashion as thin-film deposition techniques, films can be etched, or removed, based on two mechanisms: via physical and/or chemical etching.

Physical etching involves bombardment of a surface with ions launched directly at the film that needs to be etched. This process relies on the transfer of momentum between the incoming ion and the target surface. When the momentum of the ion is transferred to the target, both atoms ideally escape the film. This process is repeated with many ions, thus removing the film. **Chemical etching** occurs when the target film reacts chemically with a reactant in solution or in a gas. This chemical either makes a final product that is soluble (in the case of a solution) or a final product that is a gas (in the case of a gaseous / plasma reaction).

Both of these approaches are used in various etching systems, depending on the critical dimensions and accuracy required for the process or device. Typically, etching is performed with what is known as a **hard mask** - a patterned material (which could be patterned through photolithography or EBL, as described previously) that is relatively resistant to the etchant when compared with the

targeted material. This protects the film underneath the resistant material, leaving only the exposed parts of the film to be etched. To control etching depth, there are two techniques: one is to simply remove the film from the etching environment (i.e. remove the etching chemicals), or to use an **etch stop** layer - a layer of film beneath the target film that is relatively inert to the etching environment. There are many key parameters and types of etching processes that are outlined below.

2.3.1 Selectivity

The selectivity of an etch process is a parameter that describes how much undesired etching is achieved for a given amount of desired etching, or in other words, how well it selects the desired material to be etched. This non-ideal etching is caused because the chemicals used to etch materials (known as etchants) are generally harsh, and can thus react with many materials. Therefore, the etching procedure must consider the unwanted etching to ensure an accurate estimation of the final structure that is possible. Typically, selectivity is defined as: (amount of target atoms / mass etched): (amount of unwanted atoms/mass etched). Since the etchant has an etching rate for every material (to some extent), the ratio between the etching rates of two materials defines its **selectivity** for a given material.

2.3.2 Anisotropy

Anisotropy is defined by the directionality of an etching process i.e. how vertical the etching is. If a process etches exposed regions and the features have straight walls, then this means that it only etches vertically, and not at all horizontally. Such a process is defined as being **anisotropic**. If the etching process erodes material laterally, this means that the material under the mask will to erode away. This process would be called **isotropic.** This can be considered in terms of rates of etching in different directions. In other words, a perfectly anisotropic etch has a high etch rate downwards, and no etching laterally. An isotropic etch has the same etch rate in all directions. Thus, the anisotropy of an etching process is defined by the (vertical etching rate):(horizontal etching rate).

2.3.3 Wet Etching

Wet etching is a term used for any etching process that is carried out in a solution. Typically, the solution is aqueous (in water). This etching scheme relies on chemical mixtures - so most materials have a unique chemical mixture required for optimal etching. This scheme is an exclusively chemical etching process, which means that it is relatively fast and has high etch selectivity. However, it also means that etching is very **isotropic**; sharp, vertical features are difficult to obtain. Wet etching is generally used when the features being etched do not require precision and uniformity, or if isotropy is desired for a specific device parameter. Also, in the case of MEMS, wet etching must be used carefully, since suspended structures (such as floating beams) are prone to collapse due to the surface tension of aqueous solutions.

2.3.4 Dry Etching

Dry etching is a term for any etching process that is carried out in a gas or plasma. This etching scheme can rely on both ion bombardment and reactive chemical gases. This scheme can etch almost any material depending on the specific technique or condition. This etching technique can use heavy, chemically inert molecules or ions for physical etching, chemically active molecules for chemical etching, or a combination of both. This technique typically gives the

best control over the depth of etching and sharpness of features. However, if the etching is physical, it does not exhibit good **selectivity**, since momentum transfer can occur with any compounds. However, heavier compounds are more difficult to etch - and this is the principle used to "mask" features - to protect patterns of a film from etching, while removing the exposed sections.

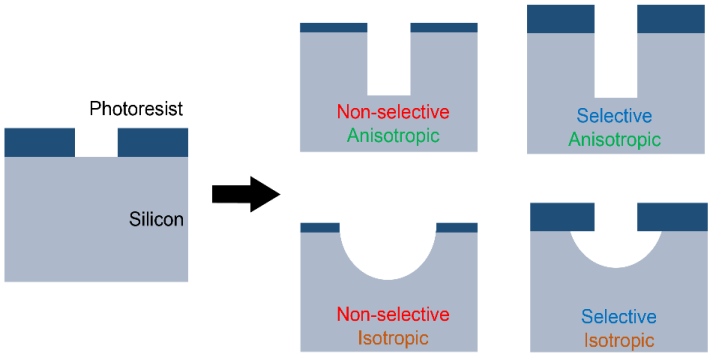


Figure 2.12 - *Anisotropic etches are directional in nature - the etching rate is much faster in one direction than any other directions, allowing for straight sidewalls. Selective etches have differing (many orders) etch rates in different material, thereby allowing one to etch one material without damaging the other.*

2.3.5 Reactive Ion Etching (RIE) and Deep RIE (DRIE)

RIE uses a scheme that allows for anisotropic chemical etching, and is one of the most common dry-etching schemes used in industry for large-scale fabrication. It uses fluorine or fluorine-containing gases which are ionized with different end atomic masses. The resulting plasma is accelerated using a strong electric field, causing the ions to strike the target surface vertically. The resulting etching profile is vertical, despite being driven by chemical process. Different parameters can be tuned to get very high aspect ratio (i.e. very deep) structures.

DRIE uses a similar process with the addition of a **passivation step.** This deposits a protective fluorine layer on all etchant surfaces, followed by an etching step. By alternating the etching and passivation steps, the sidewalls of the etched features are protected from chemical etching. This process also allows the etching step to be more aggressive i.e. faster. DRIE is often used in applications where much faster etching, or very high aspect-ratio structures are desired, and was specifically designed for MEMS applications.

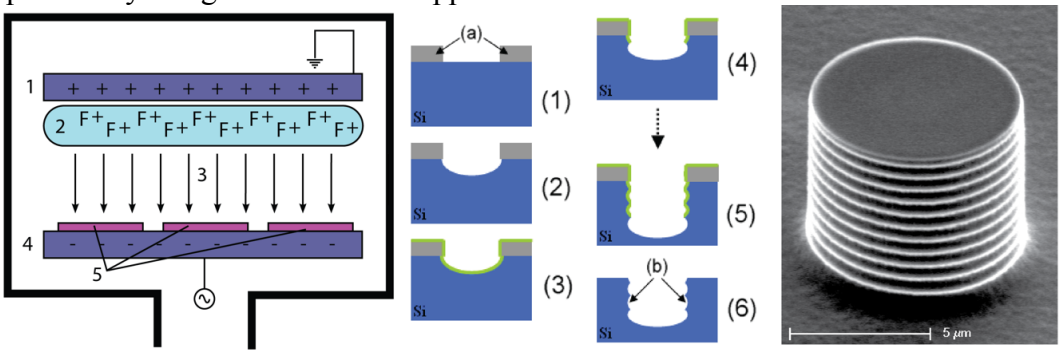


Figure 2.13 *- Reactive Ion Etching (RIE) and Deep RIE (DRIE) are two of the most popular anisotropic etching methods used to date. The reactive ion etching process (left) uses gases like XeF_2 to generate a plasma under a high electric or magnetic field (typically using an oscillating field). This causes fluorine ions to bombard the surface of the wafers. These high-energy ions react with substrate atoms selectively to generate volatile by-products. DRIE uses an additional step of adding an chemically resistant layer called a passivation layer (center, shown in green) to allow for faster anisotropic etching. This DRIE process is called a Bosch Process, and a pillar fabricated through the Bosch process is shown on the right. (Source: Dollhous, Gurgelgonzo, Pgaladja, Wikimedia Commons)*

2.4 Fabrication of Nanostructures

Microfabrication as a technique used with conventional methods are limited by the resolution of the exposing beam. For example, photolithographic features are generally limited by **diffraction**. This diffraction is a function of the wavelength of the incident beam and the numerical aperture of the beam source. As explained above, there exist various top-down approaches to define the nanostructures, which include e-beam lithography and nanoimprint lithography. However, top-down nanofabrication techniques suffer from non-uniformity and surface roughness on the order of a few nanometers which becomes greatly detrimental to devices performance as feature sizes become comparable in size.

Of late, there has been great interest in bottom-up fabrication of nanoscale structure, driven by chemical synthesis of nanomaterials and their guided assembly into more complex structures. Unfortunately, bottom-up techniques are not as scalable as top-down methods, and rely heavily on thermodynamic driving forces which also limit formation of complex three-dimensional active structures. To leverage the atomic precision of bottom-up fabrication and the scalability of top-down methods, there has been vast interest in integrating the two approaches to develop hybrid-techniques which allow for nanoscale precision, uniformity and control while maintaining the scalability. Herein, we elaborate on a few bottom-up techniques used in nanofabrication.

The previous methods mentioned above, notably thin-film deposition and etching methods, have been uniquely leveraged for their nanometer to angstrom-level precision to define features in the few nanometer regime for simple structures. Now, more interest is evolving in using "wet" processing methods, including designed polymers and micro/nanoparticles in solution.

2.4.1 Block-copolymer lithography

Block copolymers are long-chain molecules made of modular units. These can be fabricated using typical polymer processing techniques with different compatible building blocks. If these chains are long enough, they are flexible and are allowed to deform and rearrange. Once they rearrange, they will loop to associate with parts of chains that are alike. For example, a chain with components - or blocks - "A" and "B", which is assembled as: AAAAA-BBBBB-AAAAA-BBBBB- and so on, would wind to arrange with all "A" parts and "B" parts preferring to associate. This causes regions to form, such as the ones shown in the figure below. These regions can selectively be etched to form an etching mask or sacrificial layer for further processing.

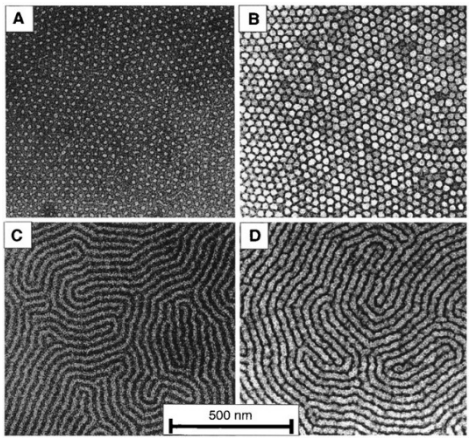


Figure 2.14 - *Example of block-copolymers aggregating to align blocks similar in chemical nature, thereby resulting in nanoscale patterns. (Source: Ref. [1])*

The advantage of these structures is that once the polymer is fabricated, it can be spin-coated on a substrate easily to form nanoscale patterns over large areas. These patterns can be lines or pores, depending on how the polymer is designed. What needs to be further optimized with this technique is the specific process parameters that are required to obtain uniform film

properties across large areas. Moreover, the technique is not fully specified, as it does contain a certain density of irregularities; more studies on these polymers' behavior is required.

2.4.2 Microsphere lithography

Self-assembly is a prevalent and natural method of assembling micro and nanostructures in ordered configurations by leveraging forces that arise through liquid wetting and drying phenomena and the surface interactions between the structures. A popular example of this is the coffee-ring effect, where the higher evaporation rate of water at edges of a "puddle" causes microparticles of coffee to collect and form a stain at only the edges of where the fluid originally was. With rational design, an analogous system can be used to allow colloids, or microspheres, to assemble very uniformly shaped particles into a close-packed pattern. This is shown in the figure below.

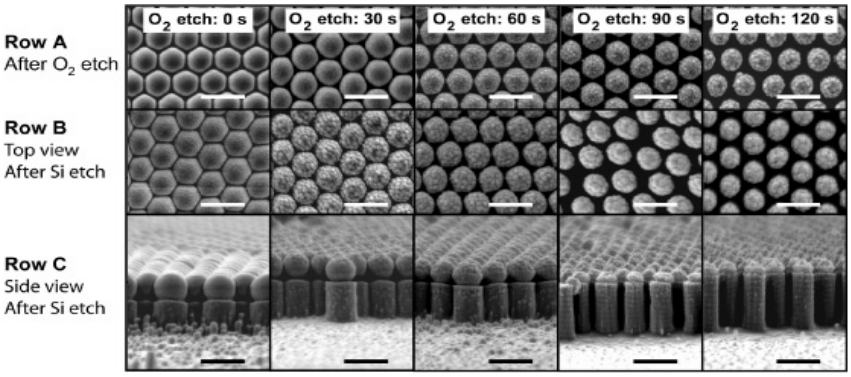


Figure 2.15 *- Colloids, or micro/nanospheres of a material, can be easily fabricated and assembled naturally. These assemblies of colloids can be used as etching masks (as shown in these figures). This can allow for scalable and simple fabrication of nanostructures over large areas. Scale bars are 750 nm. (Source: Ref. [2])*

Once this pattern of spheres is achieved, they can act as either (1) a **mask** for etching, or (2) a **sacrificial layer** for deposition. This allows long-rage patterns of nanoscale (sometimes in the tens of nanometer range) to be fabricated easily over large areas. This method has advantages in battery technologies, electronics, and optoelectronics. As mentioned before, despite being a promising technique, this method requires work in investigating specific process parameters to ensure uniform coverage across large areas, as well as more parameters to make it a scalable and versatile process.

2.4.3 Nanoscale Manipulation - Optical Tweezers, Dielectrophoretic Trapping

One major challenge with bottom-up nanofabrication is the inability to manipulate matter (i.e. single particles) with the same agility we are able to do so in the macro - and even on the micro - scale. The same techniques, do not scale down for nanomanipulation, which is essential in building complex architectures. Methods explained above (sphere lithography and block-copolymers)

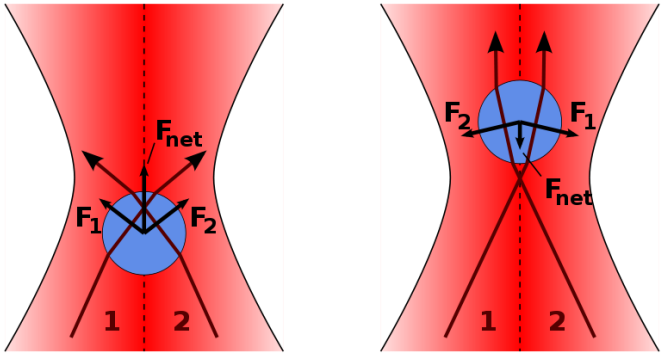


Figure 2.16 *– A schematic of laser trapping. The red region depicts a focussed laser beam. The narrowest region in the beam is known as the beam-waist. Particles (blue) shifted from the equilibrium position (here taken to be the middle of the bean-waist) experience a net restoring force. (Source: Roland Koebler, Wikimedia Commons)*

are not suitable for single particle manipulations, but rather than patterning of repeated structures over large area. Below we discuss two techniques which address the need for single particle manipulation - optical tweezing and dielectrophoretic trapping. The common idea behind both techniques is the ability to make use of nanoscale forces to catch a hold of the particle of interest and then manipulate them into the position of interest.

The general idea behind **optical tweezing** is that a highly focused laser beam can exert forces on a dielectric particle, based on the difference in refractive index between the particle and the beam propagation medium. At the beam-waist (the narrowest part of the laser beam) there exists a point where the net force exerted on the particle is zero. Moving away from the equilibrium position exerts a restoring force on the particle and particle returns to the equilibrium position. Leveraging this, in recent years, there has been great interest in using a highly focussed laser beam to trap, and manipulation nanoparticles. Optical tweezing can be used in nanofabrication, to position particles, one by one at the desired location, thereby allowing for bottom-up assembly of complex structures.

In **dielectrophoretic trapping**, non-uniform electric fields can cause local polarization of a dielectric material (e.g. nanoparticle), which then experience a force along the field lines. Depending on the orientation of the dipole, the force can be attractive or repulsive. A variety of factors contribute to the overall-trapping of the particles - including, shape and electrical properties of the material, the frequency and strength of the electric field, and the polarizability of the trapping medium. Given the dependence on many factors, dielectrophoretic trapping is a highly versatile technique allowing for selective localization (and release) of particles.

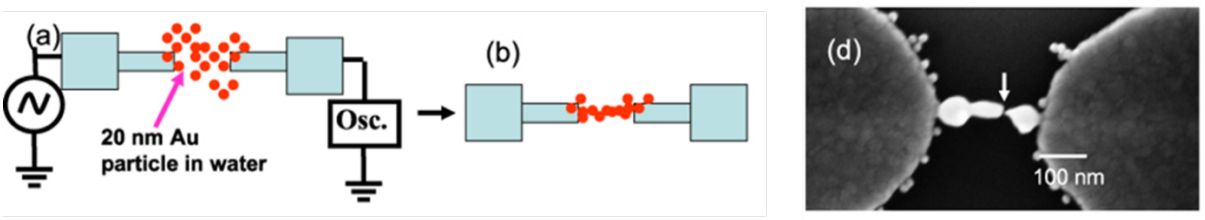


Figure 2.17 - *Dielectrophoretic trapping combines large (microscale) fabrication with nanoscale precision. An oscillating electric field is applied between two electrodes (AC), and this attracts particles between them. Once the particles connect the two electrodes, the field is diminished since the circuit is effectively changed, allowing for precise and controlled assembly of nanoparticles. (Source: Ref. [3])*

2.5 Summary

As demonstrated throughout this chapter, micro/nanofabrication is an important and crucial step in ensuring commercial electronics can be mass produced and continue to excel in performance in the coming years. Over the past 40 years, Moore's law has been consistently followed, which states that the amount of transistors in an integrated circuit doubles every two years. To continue this trend, new and scalable nanofabrication strategies must be realized to ensure mass producibility of technologies that transcend classical fabrication limits. Moreover, the high density of devices that can be achieved has not gained a proportional widespread interest, since the energy efficiency of these systems is not improved over previous systems, and uses more energy per area than predecessors. Therefore, a parallel pursuit of improving device size, as well as **energy efficiency**, is required to power the next generation of computing.

In this chapter, the general process flow of fabrication has been covered: where structures can be patterned with photolithography, material can be deposited with thin-film techniques, and material can be removed using wet or dry-etching schemes. Over several layers of design, these methods can be used to fabricate complex circuit architectures, and are currently being used for fabrication of circuits with critical feature sizes lower than 10 nanometers.

To realize the advent of energy-efficient electronics systems, micro- and nano-electromechanical systems present a unique opportunity to ensure that implemented technologies are scalable for large-scale infrastructure needs, such as data storage for cloud computing. For these systems, unique nanofabrication methods that mitigate defects and fabrication limitations arising from nanoscale surface phenomena must be implemented. Therefore, this chapter also presents some promising possibilities for nanoscale fabrication schemes that may be useful for these pursuits.

References

[1] M. Park, C. Harrison, P. Chaikin, R. Register and D. Adamson, "Block Copolymer Lithography: Periodic Arrays of ~10^11 Holes in 1 Square Centimeter", Science, vol. 276, no. 5317, pp. 1401-1404, 1997.

[2] C. Cheung, R. Nikolić, C. Reinhardt and T. Wang, "Fabrication of nanopillars by nanosphere lithography", Nanotechnology, vol. 17, no. 5, pp. 1339-1343, 2006.

[3] S. Khondaker, K. Luo and Z. Yao, "The fabrication of single-electron transistors using dielectrophoretic trapping of individual gold nanoparticles", Nanotechnology, vol. 21, no. 9, p. 095204, 2010.

Suggested Readings

J. Plummer, M. Deal and P. Griffin, Silicon VLSI technology: Fundamentals, Practice and Modelling, 1st ed. Pearson Education, 2009.

Z. Cui, Nanofabrication - Principles, Capabilities and Limits, 1st ed. Springer US, 2008.

CHAPTER 3

CHAPTER 3: ELECTROMECHANICAL ACTUATORS

Jinchi Han[1] and Alice Ye[2]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[2]Department of Electrical Engineering and Computer Science, University of California, Berkeley, California 94720, USA

3.1 INTRODUCTION

In the past two chapters, we have provided an introduction to MEMS and how mechanical devices are fabricated. This chapter will explain the physical principles guiding electromechanical actuator design. An electromechanical actuator is a structure that moves, or actuates, between different locations due to electricity, or a voltage difference from one side to another.

If one wishes to design a mechanical actuator, how would one go about it? How much voltage would need to be applied to turn a switch on? What dimensions and film thicknesses should be used? The end of this chapter will walk through an example of a MEMS cantilever switch, highlighting key design considerations along the way.

Readers are expected to have some prior background in math (differential equations) and classical mechanics (force, displacement, velocity, and acceleration) to be able to solve some of the equations, however the concepts are open to a general background.

The learning goals of the chapter are as follows:

- Be able to describe the forces governing how a MEMS relay turns on and off
- Explain and calculate the pull-in voltage, hysteresis voltage, and switching delay of a MEMS switch
- Understand spring stiffness and how to calculate it for a simple cantilever structure

3.2 TYPES OF FORCES

Before we begin, an understanding of the forces acting on a typical electromechanical actuator is required. For a typical MEMS switch, let's revisit the simple 3-terminal actuator device. A force needs to be applied between the gate and source in order to actuate the device from the "OFF" state into the "ON" state. This is known as an **actuation force**. Then, when the switch is in the "ON" state, another force is needed to restore it to its original position, known as the **restoring force**.

Forces cause things to move with an acceleration, (*a)*, that is proportional to its mass, (*m)*. Newton's second law states this quite elegantly:

$F_{net} = ma$ ……. Eq. 3.1

where F_{net} is the net force, comprising of the sum of the forces acting on the mass. When the sum of all forces is zero, there is no acceleration, and hence the state is known as "equilibrium". When the object is not moving and there is also no force, then we are in "static equilibrium". If there is

non-zero net force, then the system is no longer in equilibrium, indicating some sort of acceleration or change.

Forces are always acting all around us; however a few main ones tend to dominate the behavior of an electromechanical actuator. These are the spring force, electrostatic force, and adhesive force.

3.2.1 Spring Force

Oftentimes, springs are used in the construction of MEMS actuators to provide a **restoring force**. Imagine a simple coil spring. When it is pulled or pushed as you exert a force on it, it expands or compresses. However, it always likes to return to its original position if you let it go. Much like the coil spring, when any spring, such as those found in a MEMS actuator, is pulled or pushed, it exerts an equal but opposing force to restore the end of the spring to its original position. This restoring force is known as a spring force (F_s), which depends on the displacement of the spring from equilibrium (x), as well as a proportionality constant known as the spring constant, or spring stiffness (k).

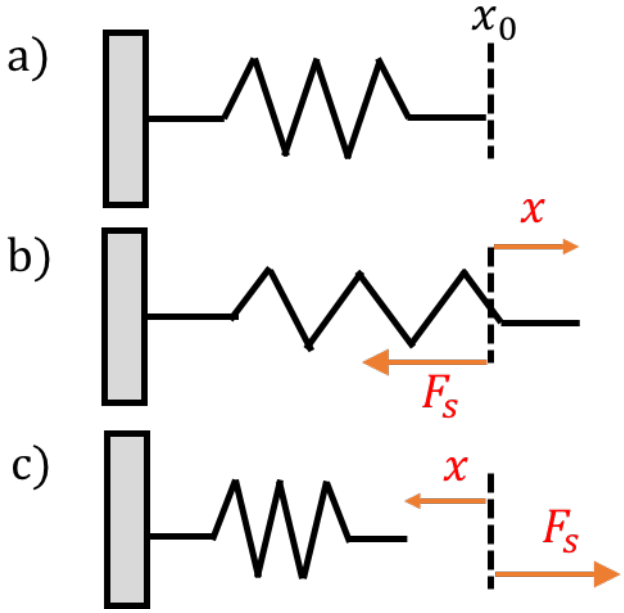


Figure 3. 1 A spring in several different equilibrium states where a) shows a spring with no forces applied, b) shows a spring with rightward displacement, in which the leftward spring restoring force returns the spring to equilibrium, and c) shows a spring with left-going displacement, where the spring restoring force moves to the right.

The spring force for a linear spring is governed by **Hooke's Law**, which states:

$F_s = -kx$ ……. Eq. 3.2

Example:

Q: There is a spring with a spring constant k = 5 N/m. The spring is stretched by 1 cm. Draw the force diagram. How much force has been applied?

A: After stretching the spring, we know that the spring is in static equilibrium since it is no longer extending, as shown in Fig 3-1b.

Therefore, the total force, $F_{net} = F_{applied} + F_s = 0$
In this case, a force is exerted to the right and the spring is exerting restoring force to the left. The spring force is described by Hooke's Law from Eq. (3-2):

$F_s = -kx = 5\ N/m \cdot 0.01\ m$

$F = -F_s = -0.05\ N$

Therefore, 50 mN of force have been applied.

If the applied force is suddenly removed from an actuated spring, the spring will oscillate back and forth, until the energy stored in the spring is dissipated through damping or other energy transfer mechanisms. Eventually, the spring will end up in its original equilibrium position with no net force applied.

3.2.2 Electrostatic Force

Electrostatic force arises from the interacting forces between charged particles. Particles can be positively, negatively, or neutrally charged. When they have opposite charges (positive and negative), they tend to attract [1]. Those particles that are similarly charged tend to repel. Neutrally charged particles neither attract nor repel.

Previously, we saw that springs could be used as a restoring force for a MEMS device to "push it back" from an extended position to its original position. Now, we will see that electrostatic actuators could be used as an attractive force, or an actuator by incorporating regions of opposite charge.

Consider two parallel plates of area (S) separated by a distance (g). The two plates have an opposing charge created by an applied voltage as shown in Figure 3.3. The medium in between the plates has a dielectric permittivity (ε).

The attractive electrostatic force between the two plates can be expressed as:

$F_{es} = Q\mathbb{E}$ ……. Eq. 3.3

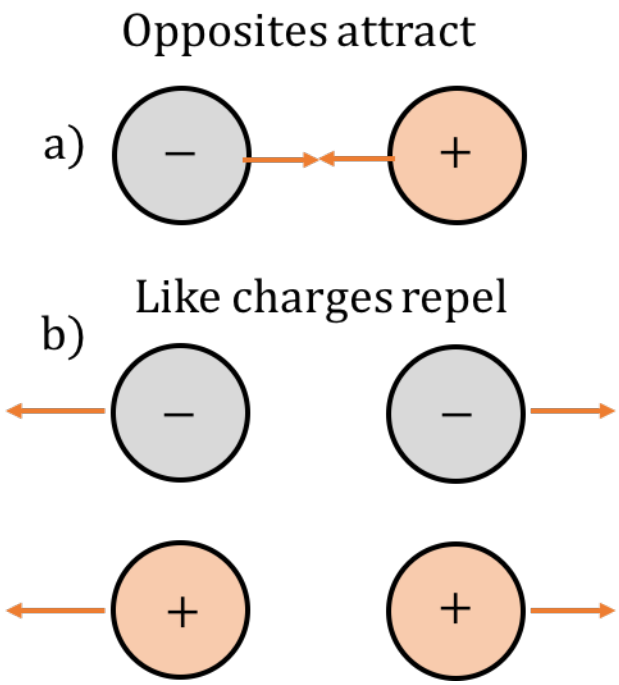


Figure 3. 2 Electrostatic a) attraction and b) repulsion

where Q (in units of Coulombs) is the built-up charge across the two plates and $\mathbb{E}$ is the electric field, defined as the electric force per unit charge (in units of Newtons/Coulomb or Volts/meter).

For a single plate, the force is half of the total electric field intensity, ie.
$F_{+} = Q\mathbb{E}/2$ ……. Eq. 3.4

Here, F_{+} denotes the force applied on the positively charged plate.
Fig 3.3 shows a diagram of two parallel plates, separated by a given gap size (g). For simplicity, we consider that the lower plate is fixed and top plate can be moved. Here, the charge is a function of the applied voltage across the plates (V), the actuation plate area (S), the dielectric permittivity of the medium (ε), and the gap size:

$Q = VS\varepsilon/g$ ……. Eq. 3.5

Under unit inspection, the electric field also can be expressed as the ratio of the applied voltage to the gap size.

$\mathbb{E} = V/2g$ ……. Eq. 3.6

Substituting Eq. 3.4 and 3.5 into 3.3, we arrive at, for the force applied to only the positively charged plate:

$F_+ = Q\mathbb{E} = S\varepsilon V^2/2g^2$..... Eq. 3.7

As one can see, for the parallel plate capacitor case, a smaller gap results in more attractive force, as does a larger area. For an alternative derivation and further reference, please refer to [2].

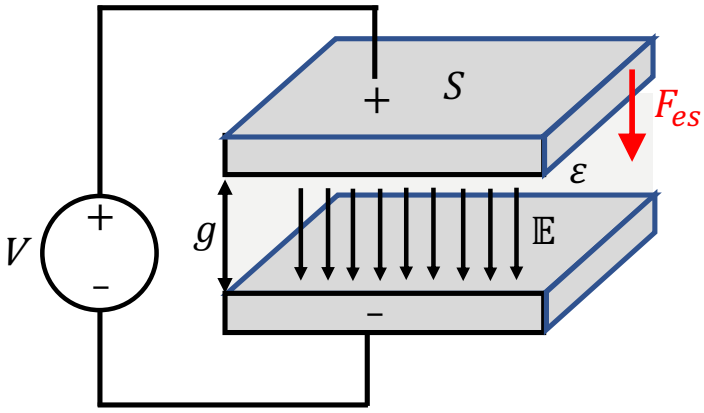


Figure 3. 3 Force on a parallel plate capacitor. An applied voltage (V) is applied to cause an electrostatic attractive force between the top plate and bottom plate of the capacitor. The applied voltage causes positive charge to build up on the top plate and negative charges to build up on the bottom plate. The opposite charges exert an electric field ($\mathbb{E}$) within the gap that is dependent on the gap size (g) and dielectric constant (ε). The electrostatic force depends on the product of the charge and the electric field within the gap between the two plates.

Let us consider intuitively what is happening as well. If no voltage is applied initially, then there is no charge, resulting in neutrally charged plates that no longer attract.

When a voltage is applied across the two plates, the difference in voltage causes a charge to build up of opposite polarity. These charges will now attract each other, pulling the positively charged plate closer to the negatively charged plate.

Interestingly, from a theoretical standpoint, if a voltage difference is applied and there is nothing preventing the parallel plates from moving closer, the electrostatic attraction of the plates could reduce the gap to approach zero, and increase the actuation force towards infinity. In reality, this does not happen, since the contact surface will always have some surface roughness -- even when parts of the surface are in contact, others are not, thus maintaining a small gap between the two plates.

3.2.3 Adhesive Force

When two materials come into very close proximity and physical contact with each other, there are adhesive forces from capillary force, van der Waals Force, and bonding that will act to strongly hold the materials together, causing them to "stick" together. The sticking results from an attractive force (F_{ad}) between the two materials. For the purposes of this chapter, we will only consider the adhesive force due to van der Waals force [3], which depends on the distance between the two materials, g, as follows:

$$F_{ad} = \frac{AS}{g^3} \quad \text{..... Eq. 3.8}$$

Here, A is a constant to describe the adhesive force dependent on materials and surface properties. More information about the nature of adhesive force can be found in Chapter 6.

Now that the forces involved have been introduced, the following sections will explain how these concepts can guide the design of electrostatic actuators.

3.3 BASIC PRINCIPLES OF ELECTROMECHANICAL ACTUATION

3.3.1 A Simple Electrostatic Actuator

In this section, we are going to see how different types of forces, namely electrostatic force (F_{es}) and spring elastic force (F_s), come into balance in an actuator. Then we will derive the typical voltages for this simple actuator to close or open while initially neglecting the adhesion force (F_{ad}).

Consider a simple electromechanical actuator comprising a fixed bottom electrode plate and a movable top electrode plate connected by a spring (Fig. 3-4). Though very simple, many real devices can be modeled by such a structure.

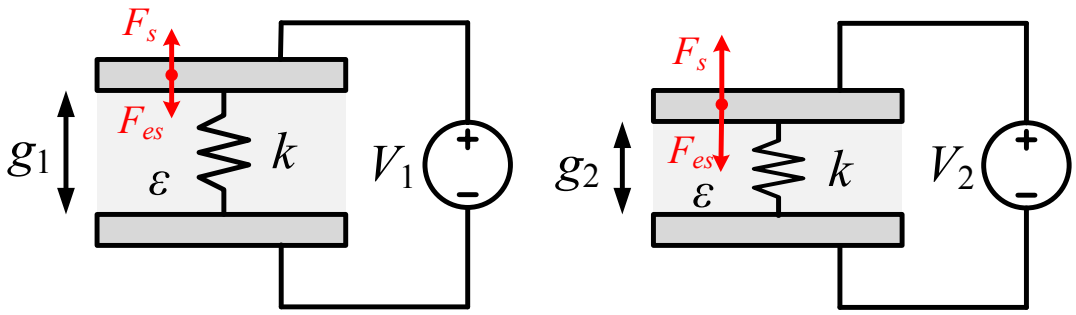


Figure 3. 4 Parallel plate electrostatic actuator without adhesion force

Here, we make a few assumptions to simplify our analysis. The mass of the top electrode is sufficiently small so that the force due to gravity of the top electrode can be neglected compared to other types of forces. We assume that the spring is linear and follows Hooke's Law, as described in 3.3.1. In addition, the spring constant is determined by material selection and geometry, rather than dependent on the distance between two electrodes. Though this electro-elastic system seems rather simplified, one would be surprised at how effective such a model is to analyze many real electromechanical switches!

Without voltage applied across the two electrode plates, there is no electrostatic force and the spring is at its natural length g_0. When we apply a small voltage across the two electrodes, the net force will be electrostatic force, which moves the top electrode down until a new equilibrium is reached at a smaller gap g_1 between electrodes. If we apply a higher voltage V_2 across the electrodes, the increased electrostatic force tends to bring the two electrodes closer and the gap g_2 is smaller than g_1. Such a static equilibrium can be described by the equation below, where we substitute the electrostatic force and elastic spring force by Eq. (3.9) and Eq. (3.2), respectively.

$$\frac{\varepsilon S}{2g^2}V^2 = k(g_0 - g) \quad \text{..... Eq. 3.9}$$

where g_0 is the original length of the spring when no force is applied on it. Note that in order to keep the equation balanced, the decrease of gap g with increasing applied voltage is required.

3.3.2 Pull-in Voltage

Though the spring elastic force is linear to the decrease in gap spacing, the electrostatic force increases quadratically with the decrease of the gap, as observed from Eq. (3.7). As a result, when the gap is getting smaller, the rate of increase of electrostatic force will be larger than the increase of spring elastic force. This results in continuous acceleration of the top electrode until it touches the bottom electrode. Therefore, we can predict a critical point where the rate of change of these two forces are equal, and the voltage at this point is defined as the **pull-in voltage** (V_{pi}). The mechanism of applying a voltage above the threshold value to close a switch or an actuator can be described as operation in the **pull-in mode**. This is one of the most important properties of a MEMS/NEMS switch, for it determines the lowest operation voltage for any circuits made out of these switches without applying a bias voltage. To obtain pull-in mode operation, we need one more equation involving the rate of change for the two forces, obtained by taking the derivative of both sides of Eq. (3.9):

$$-\frac{\varepsilon S}{g^3}V^2 = -k \qquad \text{……} \quad \text{Eq. 3.10}$$

Combining Eqs. (3-9) and (3-10) and solving for g, we find that pull-in occurs when the top plate traverses one-third of the original gap size, ie. $g=2g_0/3$. The corresponding pull-in voltage is:

$$V_{pi} = \sqrt{\frac{8kg_0^3}{27\varepsilon S}} \qquad \text{……} \quad \text{Eq. 3.11}$$

3.3.3 Release Voltage

What happens at the contact interface between the two electrodes at pull-in? As we know, even two adjacent atoms are separated by a distance of several angstroms. Therefore, there is still a tiny gap between two contacting electrodes, which is hard to measure due to the surface roughness of the electrodes. For simplicity, let's assume the distance between the electrodes in contact, denoted as g_c, is identical over the entire electrode surface.

Once the two electrodes make contact, they will reach static equilibrium – the bottom contact will prevent the top contact from displacing further. Therefore, there must be an additional supportive force (F_{su}) from the bottom electrode to balance the resulting force of electrostatic force and spring elastic force. The new force equilibrium for a closed switch becomes $F_s+F_{su}=F_{es}$. When we start to decrease the voltage below V_{pi}, F_s remains constant while both F_{es} and F_{su} decrease. When the applied voltage is decreased to a critical value, the supportive force F_{su} becomes zero, and two electrodes start to separate. Such a threshold voltage is defined as the **release voltage** (V_r).

The release voltage can be determined by manipulating Eq. (3.9) and designating gap g as g_c:

$$V_r = \sqrt{\frac{2kg_c^2(g_0 - g_c)}{\varepsilon S}} \quad \ldots\ldots\ldots \text{ Eq. 3.12}$$

3.3.4 Hysteresis Voltage

Comparing the pull-in voltage and release voltage, we can see that they are not equal. Thus, turning the device on requires a different voltage than to turn it off. The difference between these two values is the **hysteresis voltage** (V_h). It can be observed by plotting the gap between the two electrodes vs. the voltage applied across them as shown in Fig 3-5. Since no adhesion force is considered here, such a hysteresis only results from our assumption that the spring is linear. In reality, this hysteresis can be much smaller considering the fact that nonlinearity of spring is usually significant at tiny gaps.

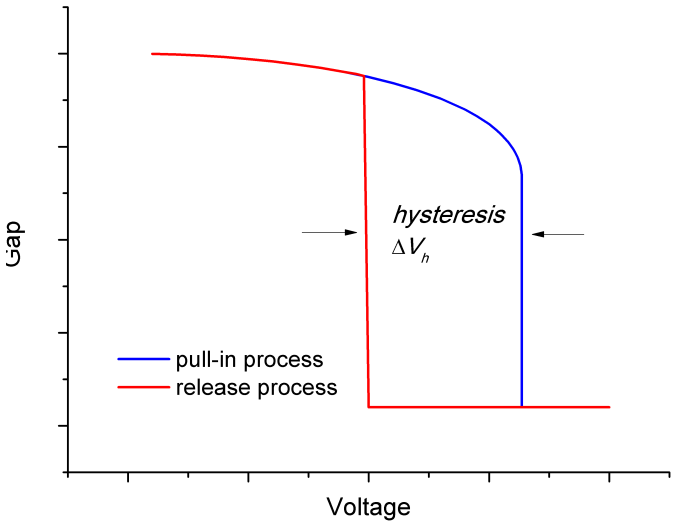


Figure 3. 5 Hysteretic I-V curve. The current is small when the electrodes of the device are not in contact, and then increase dramatically once in contact. In the ascending plot (red), the contacts are initially separate, and then an increasing voltage pulls the electrodes together. The descending curve (black) shows a device in the contacted state, where a decrease in voltage below V_r separates the electrodes. The hysteresis voltage is shown as the difference between pull-in voltage and release voltage.

3.4 ELECTROMECHANICAL ACTUATION WITH ADHESION FORCE

The analysis of the simple electrostatic actuator above does not take into account the adhesion force between the two electrode plates. In fact, the adhesion force can be very large at small gaps, which makes it very important for the design of micro-/nano-electromechanical switches. Compared to the case without adhesion force, we only need to include an additional term to account for the gap-dependent adhesion force. The force balance equation then becomes:

$$F_{es} + F_{ad} = F_s \quad \ldots\ldots\ldots \text{ Eq. 3.13}$$

In general, the adhesion originates from the van der Waals attractive force. More information about adhesion forces can be found in Chapter 6, as well as further reading in reference [3].

Here we represent the adhesion force as a term proportional to g^{-3} as shown in Eq. (3.14) [4].

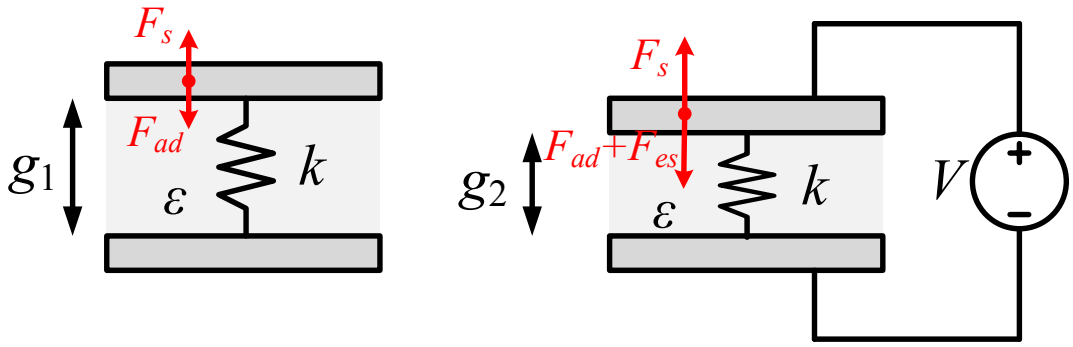


Figure 3. 6 Parallel plate electrostatic actuator with adhesion force.

When no voltage is applied across the two electrode plates, there is no electrostatic force and the spring elastic force balances the adhesive force under equilibrium, corresponding to a compressed length g_1 rather than the natural length g_0 for the case without adhesion force. Now consider when a small voltage is applied across the electrodes. Since both adhesive force and spring elastic force depend solely on the distance between electrodes, the net force will be electrostatic force, which moves the top electrode down until a new equilibrium is reached at a smaller gap between electrodes. If we apply a higher voltage across the electrodes, the increased electrostatic force tends to bring the two electrodes closer and the gap will be decreased. Such a static equilibrium is described in the equation below.

$$\frac{\varepsilon S}{2g^2}V^2 + \frac{AS}{g^3} = k(g_0 - g) \quad \text{.....} \quad \text{Eq. 3.14}$$

where A is a constant to describe the adhesive force dependent on materials and surface properties. In Eq. (3-14), the left-hand side represents the resultant attractive force comprising the electrostatic force and adhesive force, while the right-hand side represents the spring restoring force. The analysis of the release voltage and pull-in voltage are similar to the simple case, except now an additional adhesion force term now needs to be considered. To help understand how distance between electrodes varies with the applied voltage, a general graphical solution method is introduced in Appendix 3.1, which can also be applied to many mechanical actuator systems.

The difference between the pull-in voltage and release voltage is greatly dependent on the adhesion between electrodes in contact, which is proportional to the adhesion coefficient A and contact area S. A stronger adhesion, i.e., a larger coefficient A or large contact area S, leads to a smaller release voltage (in other words, higher hysteresis) of the microelectromechanical switch. An extreme case is when the product AS is so high that the supportive force for electrodes in contact will never disappear even if there is no applied voltage (V=0). As a result, the two electrodes cannot be separated and we call this **stiction failure**, which is a typical failure mode for MEMS/NEMS switch.

In reality, hysteresis voltage and stiction turn out to be significant issues for many MEMS/NEMS switch designs. Further insight into the nature of adhesion and hysteresis is of great importance but not within the scope of this chapter. Readers can find more details on this topic in Chapter 6.

3.5 A NUMERICAL EXAMPLE

Now we can use the model introduced above to calculate a numerical example of the parallel plate actuator from Fig 3-4. The parameters for the actuator are S=0.01 μm^2, g_0=20 nm, ε= 8.854×10^{-12} F/m, g_c= 1nm, and k=10 N/m. For a real electromechanical actuator or switch, the equivalent spring constant can be calculated. Appendix 3.2 presents how to find the spring constant for a cantilever-structured electromechanical switch.

The dependence of the gap on the applied voltage without adhesion force (shown in Fig. 3-7) can be calculated by Eq. (3-9). Note how at one-third of the gap displacement, the device switches shut abruptly at the pull-in voltage. Combining Eqs. (3-9) and (3-10), the pull-in voltage without adhesion force is calculated as V_{pi}=16.36V. For comparison, we added the gap variation with applied voltage in Fig. 3.7 when we consider adhesion force (constant A is chosen as 1×10^{-18} J in this example). We can observe the existence of adhesion will help reduce the gap at small voltages and lower the pull-in voltage. The pull-in voltage is calculated quantitatively to be V_{pi}=15.83V using Eq. (3-16) in Appendix 3.1.

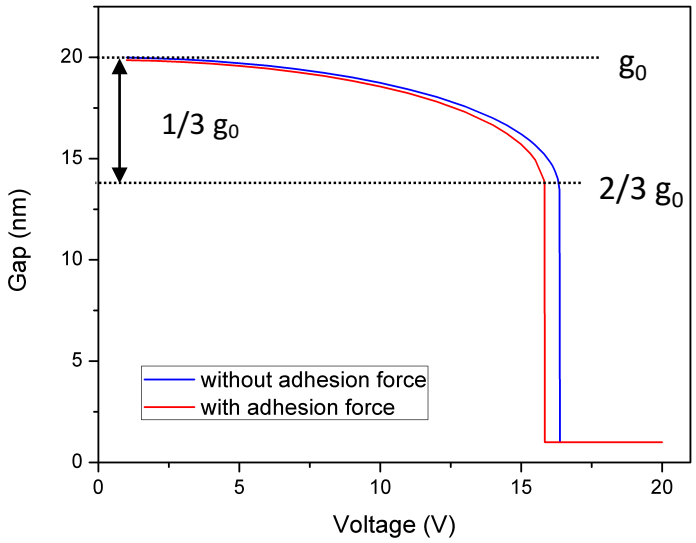


Figure 3. 7 Variation of gap with applied voltage of pull-in process (A = 1×10^{-18} J)

We can also calculate the release voltage with Eq. (3-12) when we assume no adhesion exists, and such a threshold value turns out to be 2.07V, resulting in a hysteresis voltage as 14.29V. If we consider the adhesion as defined above (A=1×10^{-18} J), Eq. (3-17) in Appendix 3.1 has no solution, which means a permanent stiction will occur due to the adhesion. If we manage to reduce the adhesion, namely reduce the contact area S by design optimization or coefficient A by choosing different material or both, we can avoid the stiction failure. For instance, if the coefficient A is reduced to 1×10^{-20} J, the release voltage with adhesion force is calculated to be 1.43 V. We can conclude that with the existence of adhesion, both pull-in voltage and release voltage will decrease, and the actuator will be free from stiction failure.

3.6 ELECTROMECHANICAL ACTUATION WITH BODY BIAS

For ultra-low-voltage MEMS designs, it is desirable to reduce the gate actuation voltage as much as possible. Typical MEMS devices require either a small gap or a fairly large actuation area to reduce the pull-in voltage, which can be difficult or costly to fabricate. A body voltage applied across a parallel plate actuator is another way to reduce the required gate actuation voltage and improve overall power performance of a micro-/nano-electromechanical switch.

Intuitively, an applied body bias voltage will introduce an additional electrostatic force, which draws the two parallel plates closer to each other, reducing the gap between electrodes, as shown in Fig. (3.8a). Fig. (3.8b) illustrates how a smaller actuation voltage (V_g) is then able to modulate the gap and control the closing and opening of the switch. By careful selection of the body voltage, we can bring the release voltage down to 0V, and pull-in voltage to V_h. As a result, the theoretical minimum actuation voltage is limited by the hysteresis voltage, thus providing motivation for minimizing hysteresis in MEMS switch design.

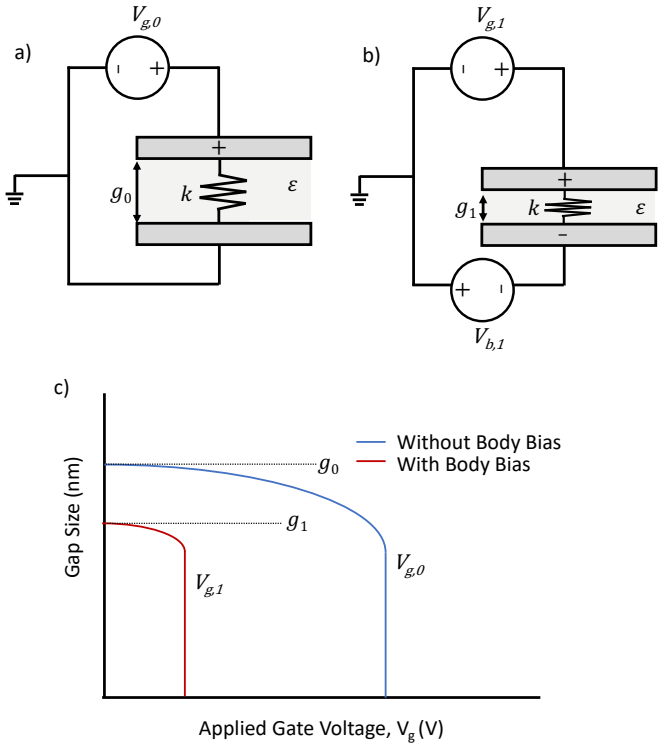


Figure 3. 8 An electromechanical actuator a) without body bias and b) with body bias voltage applied. c) A qualitative illustration of the effect of body bias in reducing the gap size and effective actuation voltage applied to the gate.

A quantitative analysis of body bias is available in [8]. Another method of analysis is to introduce an additional electrostatic force term for the body electrode in Eq. (3-14) and conduct analysis using the graphical method presented in Appendix 3.1 to determine both the pull-in voltage and release voltage.

3.7 SUMMARY

Thorough study of these concepts can lead to a deeper intuition and understanding of the microelectromechanical switch. We found that the physical behavior of an electrostatic actuator can be modelled using Newton's Laws of Motion. By considering the electrostatic force, spring restoring force, and adhesive force, one can now predict the pull-in and release voltage of a mechanical actuator. Empowered with this knowledge, you too can now design your own MEMS switch! For those who are interested, Appendix 3.1 and 3.2 provide further reading on general methods to determine the pull-in and release voltage of different structures of MEMS devices, as well as finding the spring constant of different spring design systems.

APPENDIX 3.1 GRAPHICAL METHOD FOR ANALYSIS OF PULL-IN AND RELEASE VOLTAGE

A graphical method is introduced here for the analysis of the pull-in voltage and release voltage of an electromechanical actuator. In the example presented below, we calculate the actuator assuming an adhesion force. However, it is noteworthy that such a method is general and can be applied to the case without adhesion force, and the case with body bias voltage applied, among others.

We start with the force balance equation for an electromechanical actuator Eqs. (3-13) and (3-14), duplicated below:

$$F_{es} + F_{ad} = F_s \quad \ldots\ldots\ldots \text{ Eq. 3.13}$$

$$\frac{\varepsilon S}{2g^2}V^2 + \frac{AS}{g^3} = k(g_0 - g) \quad \ldots\ldots\ldots \text{ Eq. 3.14}$$

For convenience, we can transform the equation (3-14) by multiplying a g^3 term on both sides and obtain the following:

$$\frac{\varepsilon S}{2}V^2 g + AS = k(g_0 - g)g^3 \quad \ldots\ldots\ldots \text{ Eq. 3.15}$$

Note that the left-hand side and right-hand side of the equations still correspond to the attractive term and repulsive term, respectively. If we plot the two curves of attractive and repulsive forces on the same graph, the intersection between these two curves corresponds to the **force balance point**. The force balance point indicates the displacement, for a given voltage, at which the forces acting in both directions on the device are equal. The attractive force counterpart is linearly dependent on the gap distance, and the slope of the attractive force counterpart has a quadratic dependence on the applied voltage.

Consider how voltage affects the force balance point. We have voltages V_A, V_B, V_C, where V_A=0<V_B<V_C. Points A, B, and C represent the force balance points at each voltage, respectively. We can observe that the line with a greater voltage has a larger slope, and therefore intercepts the restoring force curve at a smaller distance, which agrees with our qualitative analysis.

The static equilibrium as analyzed above holds until the line for attractive term becomes tangential to the repulsive curve (point D). When we further increase the voltage, there no longer exists an intersection between the two curves. In fact, when we increase the voltage a bit beyond V_D, the attractive force will increase more than the repulsive spring force for all gap decrease Δg. In this case, the top electrode will continue accelerating and moving towards the bottom electrode until it comes into contact, once again

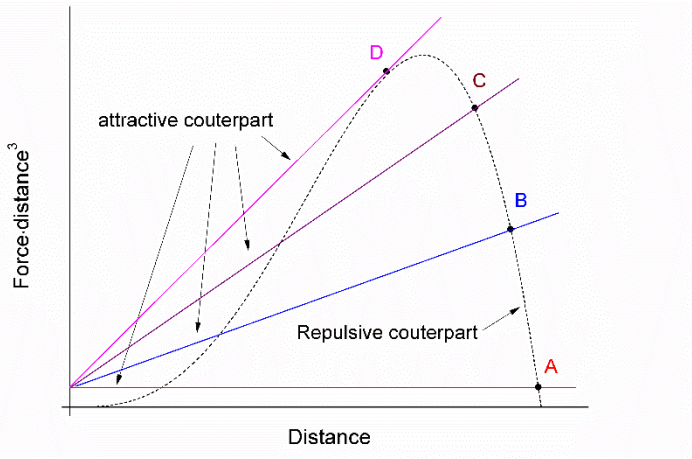


Figure 3. 9 Dependence of static equilibrium on applied voltage.

achieving pull-in mode operation. Therefore, finding the pull-in voltage is simply solving for this tangential point D.

To obtain the pull-in voltage quantitatively, we can take the derivative of the resultant attractive force F_a and repulsive force F_r separately with respect to the gap, and solve the equations $F_a=F_r$ and $dF_a/dg=dF_r/dg$ simultaneously as:

$$\frac{\varepsilon S}{2g_{pi}^2}V_{pi}^2+\frac{AS}{g_{pi}^3}=k(g_0-g_{pi})$$

$$-\frac{\varepsilon S}{g_{pi}^3}V_{pi}^2-\frac{3AS}{g_{pi}^4}=-k \quad \text{……. Eq. 3.16}$$

where g_{pi} is the gap distance at the pull-in point, and V_{pi} is the pull-in voltage. Since the Eq. 3.16 involves an adhesive force term on the leftside, the resulting pull-in voltage will be different from what we arrived at in 3.4.2. Without the adhesive force, Eq. (3.16) reduces to Eq. (3.10) and therefore leads to the same pull-in voltage in Eq. (3.11) as $g_{pi}=2g_0/3$.

After the two electrodes come into contact, another static equilibrium is reached. However, since the resultant attractive force comprising the electrostatic force and adhesive force is significantly higher than the repulsive force, the right-hand side cannot only be represented by the product of the spring constant and amount of compression any more. In general, an additional supportive force from the bottom electrode will add to the right-hand side of Eq. (3.13) to balance the forces, which is typical for cantilever MEMS switches.

Side Note:

Some devices, such as the squitch (See Chapter 7), have compressible molecules between electrodes. At pull-in mode for these devices, instead of a supportive force, there will still be a tiny gap distance between the two contacts. The compressible molecules in between these contacts creates an abruptly increasing stiffness for the spring between electrodes at a tiny gap distance corresponding to the size of the compressible molecules (see Fig. 3-10), and the greatly increased repulsive force from the molecules will balance the attractive force terms when the electrodes are in contact.

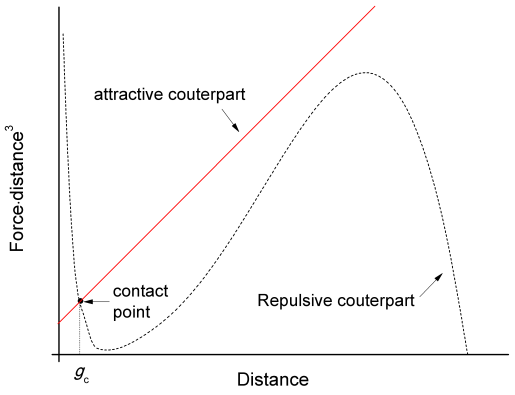


Figure 3. 10 Static equilibrium under contact with the existence of abrupt stiffness.

The resulting force balance equation is therefore:

$$F_{es}(V,g)+F_{ad}(g)=F_{s-sg}(g) \quad \text{…Eq. 3.17}$$

where the F_{es} is the electrostatic force, F_{ad} is the adhesion force, and $F_{s\text{-}sg}$ is spring elastic force at short gaps, respectively. Here, $F_{s\text{-}sg}$ cannot be simply modeled by a linear term $k(g_0\text{-}g)$ anymore, but shows great nonlinear dependence on the short gap length.

Now we know how to turn on a MEMS/NEMS switch by applying the pull-in voltage. However, similar to the case without adhesive force, switching the MEMS device off cannot be achieved by merely reducing the voltage below the pull-in voltage.

Consider a typical cantilever microelectromechanical switch as an example. In the on-state, the two electrodes are in contact and the attractive force, comprising the electrostatic force and the adhesive force, is larger than the elastic restoring force from the deformation of cantilever. As a result, the net force will be balanced by a supportive force provided by the bottom electrode. Here, we are using the same graphical method again to analyze the switch-off process, as shown in Fig. 3-11. During the transition from on-state to off-state, we are reducing the voltage below V_{pi}, which is equivalent to reducing the slope of the line corresponding to attractive forces. When the line intercepts with the black dotted curve representing the repulsive spring force at point g_c, the supportive force from bottom electrode (or compression force between the two electrodes) becomes zero, and the corresponding voltage is called the **release voltage** (V_r). A further reduction on the applied voltage will make the top electrode accelerate and move up until it reaches another static equilibrium point at off-state. To obtain the release voltage quantitatively, we can solve the force balance equation at the critical point of release as

$$\frac{\varepsilon S}{2g_c^2}V_r^2 + \frac{AS}{g_c^3} = k(g_0 - g_c) \quad \ldots\ldots\ldots \quad \text{Eq. 3.18}$$

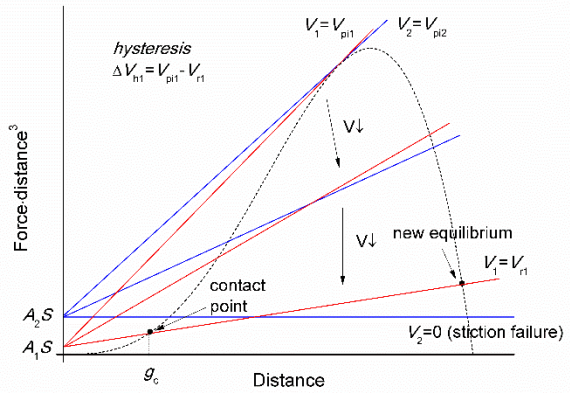


Figure 3. 11 Release process and stiction failure.

APPENDIX 3.2 FINDING THE SPRING CONSTANT

A cantilever consists of a rigid beam with one end anchored and the other end free to move. For a cantilever MEMS switch, if the bottom electrode is right below the tip, we can treat the electrostatic force as a point load on the cantilever, as shown in Fig. 3-12. Here, we just show how to extract the spring constant of a cantilever, and a more detailed discussion can be found in [7]. To quantify the deformation of the cantilever, the concept of curvature is introduced to describe to what extent the cantilever bends. From a geometrical view, when a segment bends to form a circle, the smaller the radius of the circle is, the more the original segment is bent. Therefore, we can use the reciprocal of radius of the circle to evaluate the curvature, namely, $\kappa=1/R=\theta/\theta R$, where θR corresponds to the arc length with respect to an angle of θ. For a general bent structure, we can use the same ratio (angle/arc length, i.e., $d\theta/ds$) to define the curvature of a local point, which equals to the reciprocal of radius ρ of an osculating circle at the same point.

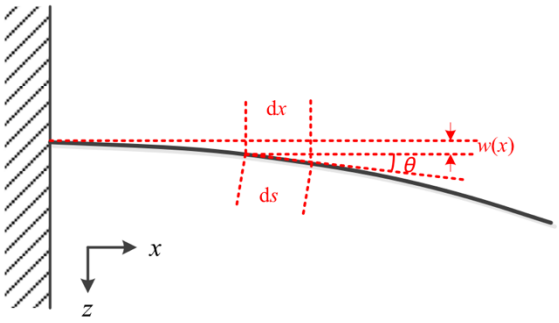


Figure 3. 8 Schematic of a deformed cantilever.

$$\frac{1}{\rho}=\frac{d\theta}{ds}\approx\frac{d\theta}{dx}=\frac{d}{dx}\frac{dw}{dx}=\frac{d^2w}{dx^2} \quad \ldots\ldots\ldots \text{Eq. 3.19}$$

where ρ is the radius of the curvature of the deformed cantilever, $w(x)$ is the deflection at point x.

The relationship between the curvature and applied load for a cantilever is

$$\frac{1}{\rho}=-\frac{M(x)}{EI} \quad \ldots\ldots\ldots \text{Eq. 3.20}$$

where $M(x)$ is the torque on the cantilever at point x. Parameter E, known as the Young's modulus, is an important material property that describes the relationship between the strain and stress. I is the second moment of inertia of the cantilever, which can be calculated based on the beam geometry. The moment of inertia for a beam with a rectangular cross section can be calculated as

$$I=\int_{-h/2}^{h/2}Wz^2dz=\frac{1}{12}Wh^3 \quad \ldots\ldots\ldots \text{Eq. 3.21}$$

where h is the thickness and W is width of the beam, respectively. A more general form that can be used to calculate the momentum of inertia of complicated geometries can be found in Ch 9 of Ref. [5].

For a cantilever deformed by a point force load, the relationship between the force and deflection is represented by

$$\frac{d^2w}{dx^2}=-\frac{M(x)}{EI}=-\frac{F(L-x)}{EI} \quad \ldots\ldots\ldots \text{Eq. 3.22}$$

where F is the point force at the tip, and L is the length of the cantilever. We can solve the deflection at x explicitly by integration Eq. (3-22) over x combined with boundary conditions.

$$w(0) = 0, \quad \left.\frac{dw}{dx}\right|_{x=0} = 0$$

$$w(x) = -\frac{F}{6EI}x^3 + \frac{F}{2EI}x^2 \quad \ldots\ldots\ldots \text{ Eq. 3.23}$$

The maximum deflection, occurring at the tip of the cantilever, is

$$w_{\max} = \frac{F}{3EI}L^3 \quad \ldots\ldots\ldots \text{ Eq. 3.24}$$

The equivalent spring constant for the cantilever MEMS switch is then calculated as

$$k = \frac{F}{w_{\max}} = \frac{3EI}{L^3} \quad \ldots\ldots\ldots \text{ Eq. 3.25}$$

Such an extracted spring constant can be used to model the cantilever switch as spring-damp-mass system actuated by electrostatic force.

For real systems, the spring constant may be comprised of more than one component. Several springs may be connected in series or in parallel to the same moving part, as shown in Fig 3-13. For several springs connected in series, when a force F is placed on the free end of the connected springs, all the spring will experience the same force and as a result, the total displacement is the sum of displacements of each individual spring. The equivalent series spring constant, k_s, is

$$k_S = \frac{F}{\frac{F}{k_1} + \frac{F}{k_2} + \quad + \frac{F}{k_n}} = \left(\frac{1}{k_1} + \frac{1}{k_2} + \quad + \frac{1}{k_n}\right)^{-1} \quad \ldots\ldots\ldots \text{ Eq. 3.26}$$

For springs connected in parallel, the total force is the sum of the force on all springs, but the displacement experience by each individual spring is the same. Therefore, the equivalent parallel spring constant, k_p, is

$$k_P = \frac{F_1 + F_2 + \quad + F_n}{x} = k_1 + k_2 + \quad + k_n \quad \ldots\ldots\ldots \text{ Eq. 3.27}$$

Many micro-/nano-electromechanical switch springs can be described as a combination of springs connected in series and/or in parallel. By combining the above equations, we can calculate the individual spring constant for individual springs and then combine them to determine the effective spring constant.

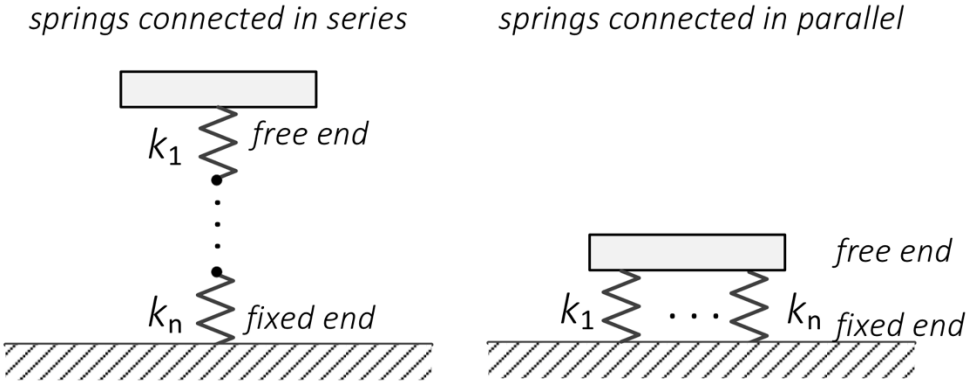


Figure 3. 9 Different types of spring connections.

REFERENCES

1. R. Fitzpatrick, “Coulomb's Law,” 3021 Lecture Notes, 14-Jul-2007. [Online]. Available: http://farside.ph.utexas.edu/teaching/302l/lectures/node16.html. [Accessed: 28-Dec-2017].
2. S. D. Senturia, *Microsystem Design, Appendix B*. Berlin: Springer US, 2000.
3. K. A. Milton, *The Casimir effect: physical manifestations of zero-point energy*. New Jersey: World Scientific, 2005.
4. F. Niroui, A. I. Wang, E. M. Sletten, Y. Song, J. Kong, E. Yablonovitch, T. M. Swager, J. H. Lang, and V. Bulović, “Tunneling Nanoelectromechanical Switches Based on Compressible Molecular Thin Films,” *ACS Nano*, vol. 9, no. 8, pp. 7886–7894, May 2015.
5. E. R. Johnston, F. Beer, and E. Eisenberg, *Vector Mechanics for Engineers: Statics and Dynamics*: McGraw-Hill, 2009.
6. Schneider, John B. "Understanding the finite-difference time-domain method." *School of electrical engineering and computer science Washington State University.–URL: http://www.eecs.wsu.edu/~schneidj/ufdtd/ufdtd.pdf, 2017*.
7. Carol Livermore, course materials for 6.777J / 2.372J Design and Fabrication of Microelectromechanical Devices, Spring 2007. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology.
8. C. Qian. “Electro-Mechanical Devices for Ultra-Low-Power Electronics.” University of California, Berkeley, 2017.

CHAPTER 4

CHAPTER 4: LOGIC RELAY DESIGN

Urmita Sikder[1], Sergio Almeida[1]
[1]Department of EECS, University of California, Berkeley, California 94720, USA

4.1 Introduction

This chapter provides insight into the basic concepts for understanding logic relays. These relays can be used to build digital logic circuits, i.e. hardware implementation of Boolean operations. Of particular importance are the three-terminal and four-terminal relay design. A short overview of the fabrication procedure and operating principle of these relays will be discussed. Next topic of discussion would be how the relay properties depend on dimension scaling. This section will give you an interesting insight into the tremendous benefits of the prevalent trend of scaling. Moreover, a brief outlook into the design and operation of multi-input multi-output relays will be provided. Finally, you will be introduced to some interesting characteristics of four-terminal relays obtained from experiments.

After finishing this chapter, you should be able to answer the following questions:

- How does a logic relay work?
- What electrical and mechanical properties are desirable in a relay?
- How does dimension scaling affect relays?

4.2 Three-terminal relay

A three-terminal relay has the following terminals: gate, drain and source (*Note that the nomenclature is inspired by a MOSFET, which is the basic component of integrated circuits*.). The circuit symbol showing all three terminals is given in Fig. 4.1. The relay can be switched on by applying a voltage difference between the gate and source terminals. The gate-to-source voltage, V_{GS} produces an electrostatic force causing the relay to actuate (Recall the concept of electrostatic actuation from Chapter 3). Upon actuation, the drain and source terminals physically come into contact, hence there is very low resistance between them. In this state, a drain-to-source voltage, V_{DS} can cause current to flow between drain and source. The relay can be switched off by simply setting V_{GS} to zero, which will separate the drain and source terminals physically, making the drain-source resistance very high (Petersen, 1979). At the off-state, applying V_{DS} will not result in a substantial current flow.

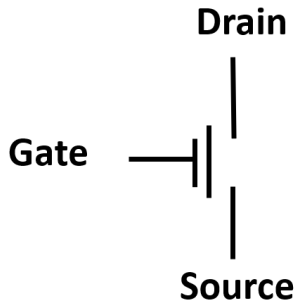


Figure 4.1: *Symbol of a three-terminal relay*

Now we are going to discuss a simple design of a three-terminal relay. We will need to refer to some nanofabrication terminology from Chapter 2. Fig. 4.2(a) demonstrates the top view of a

simple three-terminal relay. The gate, drain and source are patterned by using lithography and etching techniques. Sacrificial etching technique is used to make the suspended movable beam labeled source. For a better understanding of the process, let us examine a cross-section of structure along the line AA' towards the direction of the arrows.

The cross-section in off-state, as shown in Fig 4.2(b), shows that the source is physically separated from the drain by an air-gap. So, no current can flow between drain and source in this state. Note that, the source structure has two parts: the structural part (shown in blue) for mechanical movement and a small contact stud at the bottom (shown in green) for making contact with the drain electrode underneath. The structural material needs to have good mechanical properties, while the contact stud needs to have good electrical contact properties, i.e. low contact resistance, endurance.

Fig 4.2 (c) show the cross-section in on-state. The gate-to-source voltage, V_{GS} creates an electrostatic force, which causes the movable blue structure to bend downwards. When V_{GS} exceeds pull-in voltage, V_{PI}, which is the minimum voltage required to make contact, the relay is actuated into on-state. Hence the contact stud and drain are physically connected, which can cause a current flow between drain and source, I_{DS} depending on the value of V_{DS}.

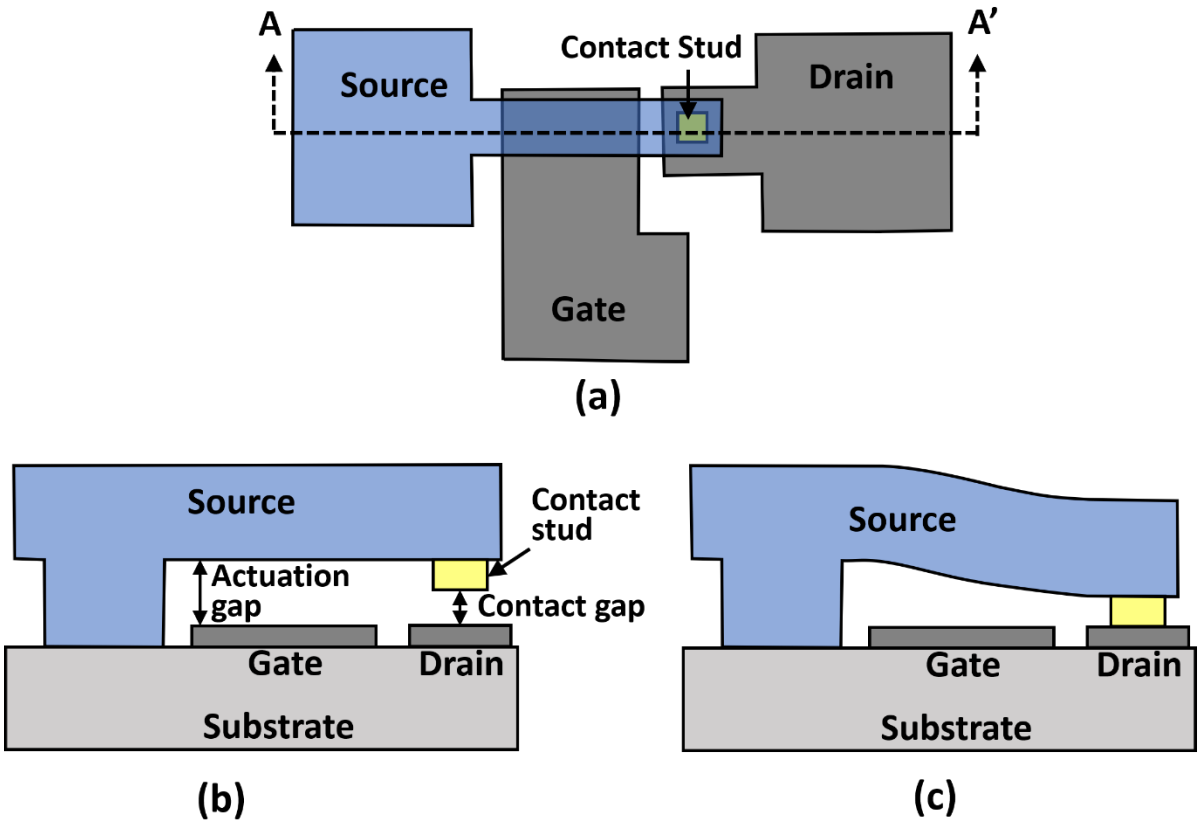


Figure 4.2: a) *Top view of a three-terminal relay*, (b) *cross-section along AA' direction in off-state, and* (c) *cross-section along AA direction' in on-state*

The advantage of the three-terminal relay lies in its simplicity. The design is simple and it is easy to fabricate. The main disadvantage of the three-terminal relay design becomes apparent when it is used in a circuit. In an integrated circuit, multiple relays may be connected in series, as shown in Fig. 4.3. The source voltage, V_S of a three-terminal relay is not fixed. In the figure, the bottom

relay has $V_S = 0$, while the top relay has $V_S \neq 0$. So, the gate switching voltage can vary depending on the position of the relay, which can lead to unreliable circuit behavior.

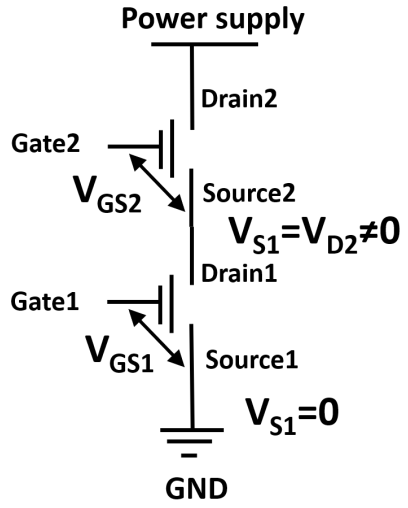


Figure 4-3: *Three-terminal relays connected in series have different source voltages, i.e. different turn-on voltages*

4.3 Four-terminal relay

The four-terminal addresses the disadvantage of the three-terminal relay by introducing a fourth terminal, which will be discussed in details later in this section. This relay comprises of the following terminals: gate, body, drain and source (Nathanael, Pott, Kam, Jeon, & Liu, 2009). Fig. 4.4 shows the circuit symbol with all four terminals. A voltage difference between the gate and body terminals can switch the relay on through electrostatic actuation. In the on-state, are conductive *channel** is established between the drain and source terminals, making the drain-source resistance very low (**Another MOSFET terminology*). Hence current can flow between these terminals, if drain-to-source voltage, V_{DS} is applied. The drain and source terminals are physically separated in off-state, leading to zero I_{DS}.

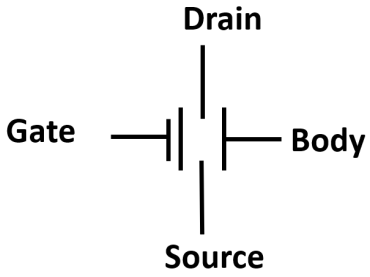


Figure 4.4: *Symbol of a four-terminal relay*

A simple design of a four-terminal relay is shown in Fig. 4.5(a). Here the gate is a movable structure suspended above the body, fabricated via sacrificial etching technique. We are going to examine a cross-section of the structure along the line AA' towards the direction of the arrows. Fig 4.5(b) shows the cross-section in off-state, where the source and drain are physically separated by an air-gap, leading to zero I_{DS}. Note that the gate structure has a metallic channel strip attached underneath via a layer of insulating dielectric. The channel has two studs for making contact with

the source and drain terminals underneath. Fig 4.5 (c) shows the cross-section in the on-state. The gate-to-body voltage, V_{GB} creates electrostatic force, which causes the movable gate structure to move downwards. For $V_{GB} > V_{PI}$ the channel contacts the source and the drain through the studs, so current I_{DS} can flow between drain and source through the channel, depending on the value of V_{DS}. The layer of dielectric insulating the gate and the channel ensures that current flows only from the drain to the source during the on-state, not to the gate.

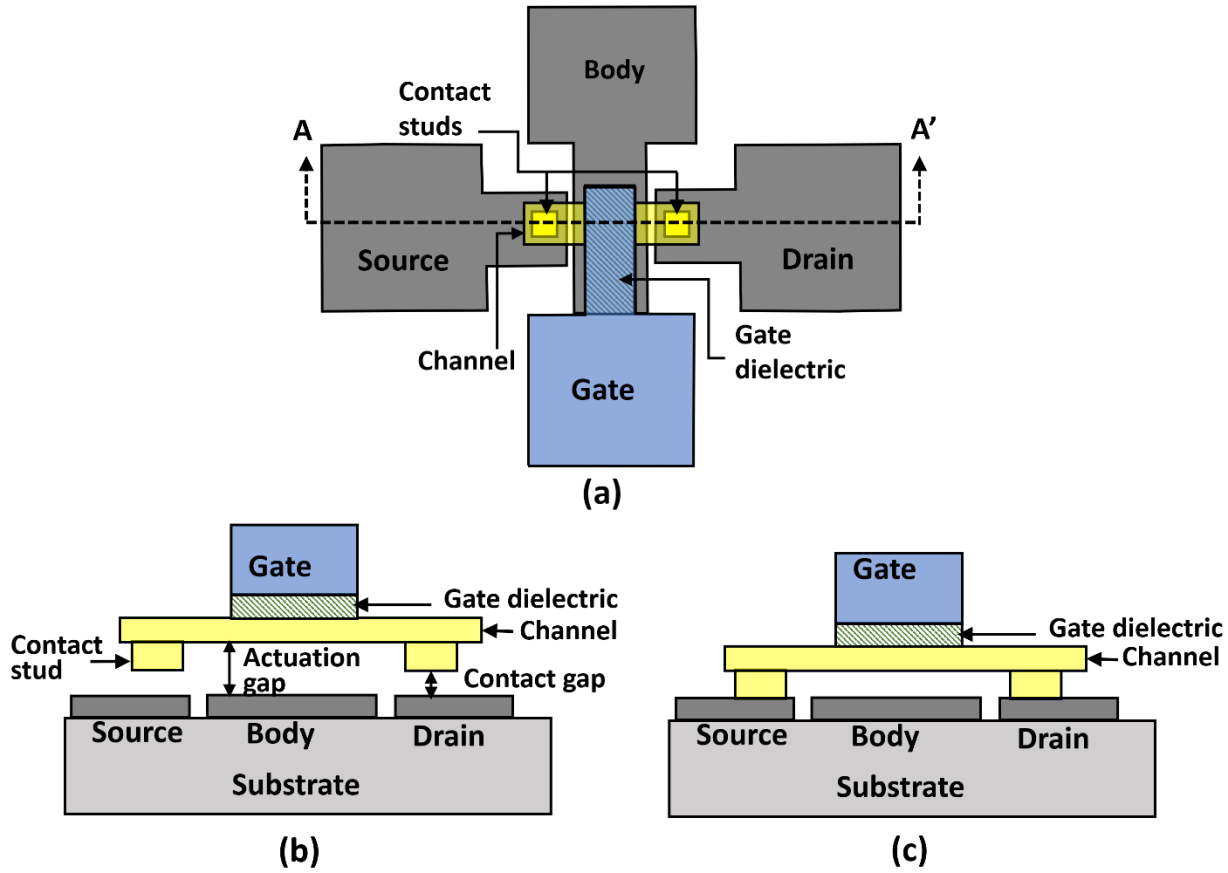


Figure 4.5: (a) *Top view of a four-terminal relay*, (b) *cross-section along AA' direction in off-state, and* (c) *cross-section along AA direction' in on-state*

From the circuit perspective, the four-terminal relay has more advantages over the three-terminal relay, since the gate-to-body voltage, V_{GB} is used to switch the relay, rather than V_{GS}. Hence, four-terminal relays can be connected in series without having variations in switching voltage. The feature solves the unreliable circuit behavior problem associated with the three-terminal relay. Moreover, the body voltage can be tuned in order to change the operating gate voltage of the relay, which can be advantageous for circuit operations. The body-biasing scheme can enable circuit operation at very low voltage, potentially at <100mV, paving the way for extremely energy efficient systems.

4.4 Multi-input relay

In contrast to three-terminal and four-terminal relays, a multi-input relay has two or more gates. In this case, the gate contact is subdivided into multiple contacts, as shown in Figure 4.6(a).

Therefore, each input voltage is a fraction of the total actuation force. With the appropriate voltage values, it is possible to perform logical operations (Nathanael, et al., 2012). Within the most common logical operations are AND and OR which are going to be used in the following section to explain the functionality of the multi-input relay.

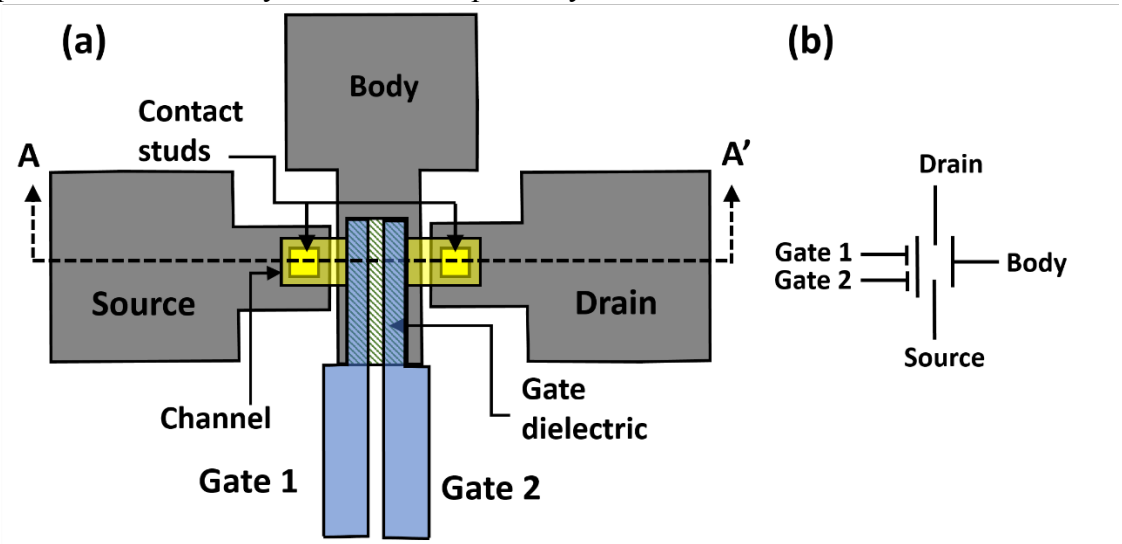


Figure 4.6: (a) *Top view of a multi-input relay*, (b) *Symbol of a five-terminal relay*

The AND logical operation gives a logical 1, a high voltage, output when all the inputs are 1 and a 0, low voltage, when any of the inputs is 0. In this case, the inputs are gate 1 and gate 2 as indicated in Fig. 4.6. At this moment let's consider that the body is connected to the ground and a resistor is connected to the source to provide a low voltage when the relay is not actuated. As explained in Chapter 3, the electrostatic force to actuate the device depends on the actuation area (A_{act}) as follows:

$$F_{elec} = \frac{\varepsilon_r A_{act} V^2}{2g^2} \qquad \text{Eq. 4.1}$$

However, in these relays the area is subdivided equally in two. Now the total force required to close the drain and source contact is the addition of both inputs:

$$F_{elec} = F_{G1} + F_{G2} = \left(\frac{A_{act}}{2}\right)\left(\frac{\varepsilon_r V_{G1}^2}{2g^2}\right) + \left(\frac{A_{act}}{2}\right)\left(\frac{\varepsilon_r V_{G2}^2}{2g^2}\right) \qquad \text{Eq. 4.2}$$

The output is a logic 1 only if the both voltages, gate1 and gate2, are 1, as depicted in Fig. 4.7(a). In the case of the logic operation OR, the output is a logical 1 when any of the inputs is 1 as indicated in Fig. 4.7(b). The relay design is practically the same with the difference that the force of each gate needs to be high enough to make the drain and source contact. This can be achieved by increasing the actuation area or by manipulating the springs making the devices softer. Now the electrostatic force can be indicated as follows:

$$F_{elec} = F_{G1} = F_{G2} = \frac{\varepsilon_r A_{act}/2 V_{G1}^2}{2g^2} = \frac{\varepsilon_r A_{act}/2 V_{G2}^2}{2g^2} \qquad \text{Eq. 4.3}$$

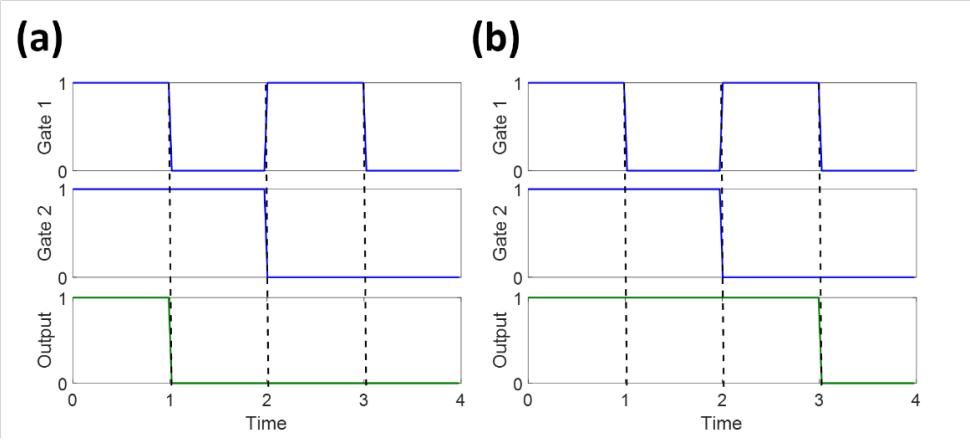


Figure 4.7: (a) *Input and output signals for an AND operation* and (b) *OR operation*

4.5 Scaling

Creating chips more versatile at smaller sizes is a key to drive the progress in electronics. For this reason, it is important that MEMS relays can be reduced in size without sacrificing performance. The best scenario would be the smallest relay operating at the lowest voltage possible. From equation 4.1 it is possible to notice that the electrostatic force can be increase by increasing the actuation area and reducing the actuation gap. Another alternative to operate MEMS at lower voltage is to make the structure, or the effective spring constant k_{eff}, softer such that less electrostatic force is required to actuate the device. However, these options have implications. For instance, the minimum gap size is limited by the fabrication capabilities, it is critical to avoid failures during releasing the device, which is the final step in the MEMS fabrication. Making the MEMS soft may lead to stiction problems during the operation of the device. This last failure is very critical and common among MEMS devices (Kam, Liu, Stojanovic, Markovic, & Alon, 2011). When the MEMS relays are actuated and a contact between two metals is stablish, there are adhesive forces (further explained in chapter 6) that keep the two metals in contact. If the restoring force of the MEMS is not strong enough the relay will never go to off state creating a catastrophic failure. The adhesive force is manifested experimentally as a hysteresis voltage. This means, that the voltage to turn the device on is different from the voltage to turn the device off, as shown in figure 4.8.

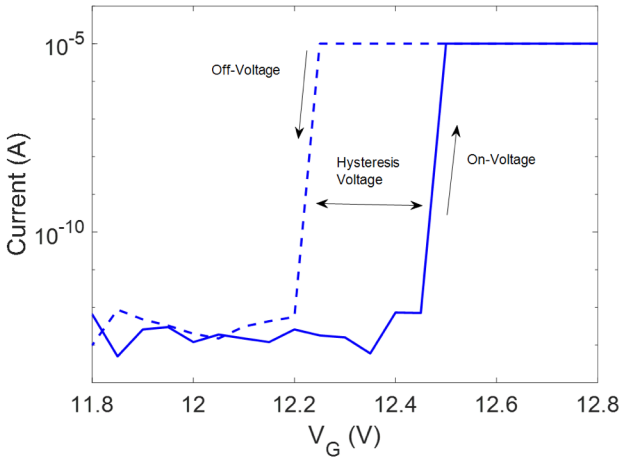


Figure 4.8: *Electrical characteristic of the MEMS relays, current vs voltage*

4.6 Conclusion

In summary, this chapter introduced you to some practical applications of the concepts you learned in chapter 3. The basic topologies for MEM relays were introduced and the operating principles were discussed. Some advanced concepts i.e. multi-input relays and effects of scaling were also discussed briefly. The information presented in this chapter will

help you to understand subsequent chapters discussing advanced relay designs and integrated circuits with relays.

REFERENCES

[1] K. E. Petersen, "Micromechanical Membrane Switches on Silicon," *IBM Journal of Research and Development,* vol. 23, no. 4, pp. 376-385, 1979.

[2] R. Nathanael, V. Pott, H. Kam, J. Jeon and T. K. Liu, "4-terminal relay technology for complementary logic," in *2009 IEEE International Electron Devices Meeting (IEDM)*, Baltimore, 2009.

[3] 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," in *Proceedings of Technical Program of 2012 VLSI Technology, System and Application*, Hsinchu, 2012.

[4] H. Kam, T. K. Liu, V. Stojanovic, D. Markovic and E. Alon, "Design, Optimization, and Scaling of MEM Relays for Ultra-Low-Power Digital Logic," *IEEE Transactions on Electron Devices,* vol. 58, no. 1, pp. 236-250, 2011.

CHAPTER 5

CHAPTER 5: RELIABILITY AND ENDURANCE LIMITS OF N/MES RELAYS

Benjamin Osoba and Tsu-Jae King Liu
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, 94720

Nano/Micro-Electro-Mechanical (N/MEM) relays are especially promising for ubiquitous deployment of electronic devices and sensing technology – widely known as the "Internet of Things" (IoT) – due to zero-leakage current and steep subthreshold slope during switching. For such IoT applications, relays must be able to survive in various environmental conditions for decades at a time. Due to the switching speed of such devices being limited by mechanical delay (~10 ns), their maximum speed of operation can be on the order of 100 MHz. As such, approximately 10^{14} ON/OFF cycles are required for a device lifetime of 20 years if it is operated 1% of the time. For these reasons, it is important to understand the various characteristics of relay reliability and endurance, including the mechanisms of failure and the methods of limiting failure. Fortunately, for each failure mechanism, the relays can be designed and/or operated to circumvent issues in performance, reliability and endurance. In this chapter, 6-terminal vertical MEM relays [1-4] are utilized to discuss relay reliability and endurance limits.

5.1: Relay Failure Mechanisms

Recall the switching mechanism (from Chapter 3) in which electrostatic and mechanical forces are balanced until the pull-in voltage is reached. A brief review of the relay structure and operating principle is presented in Fig. 5.1 below [4-5]. During relay switching operation, especially once physical contact has been made between the S/D electrodes and the channel, there are a number of ways in which the relay performance can be degraded (or perhaps completely prevented).

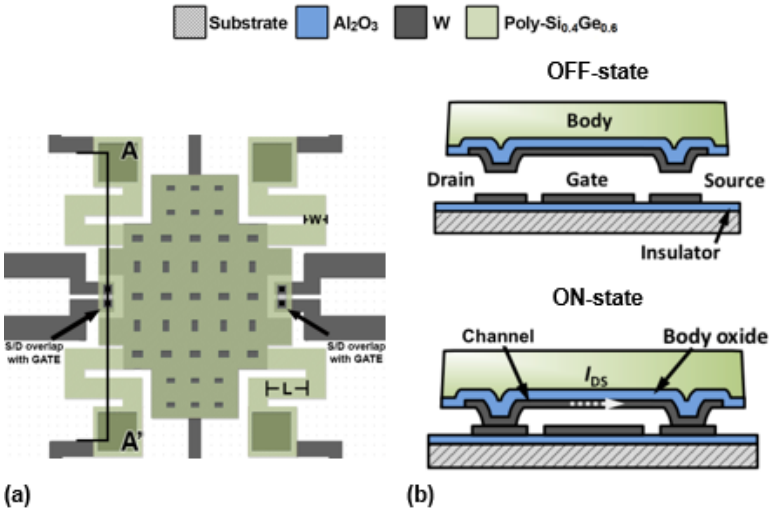


Figure. 5.1. Qualitative diagram illustrating (a) plan-view of 6-terminal NEM relay and (b) A-A' cross-section during operation.

5.1.1: Structural Fatigue and Contact Wear

For the N/MEM relays, thin-film structural failure can occur when strain approaches or exceeds the material's fracture strength [5-6]. Additionally, while the serpentine spring design reduces thermal stress effects [7], high current conducted through the body can still result in structural deformation via joule heating (i.e. resistive heating in the material due to electrical current). Over the course of many ON/OFF switching cycles, such structural fatigue or joule heating can cause cracks to form – or even mechanical deformation, as shown in Fig. 5.2 – within the doped polycrystalline Silicon Germanium (p+ Poly-$Si_{0.4}Ge_{0.6}$) flexural beams. With the ratio of effective suspension beam length to maximum vertical deflection, the maximum induced strain during operation is typically on the order of 0.2%, thereby rendering structural fatigue a relatively minor issue [5]. Structural deformation due to joule heating can be limited simply by artificially limiting the current through the structure.

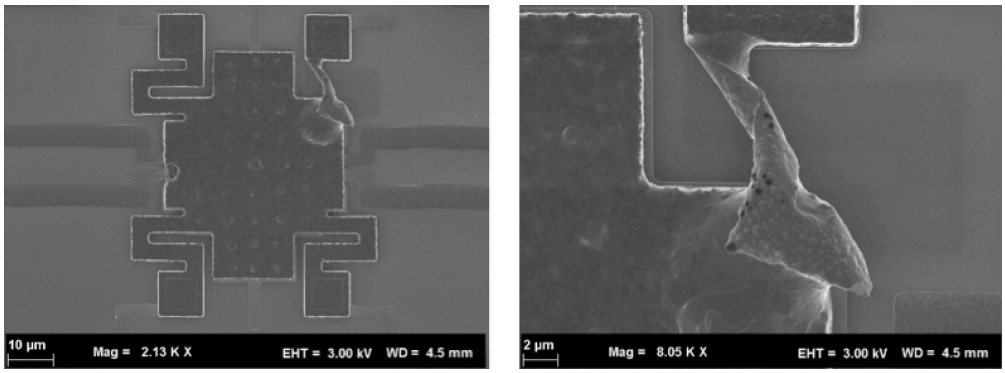


Figure 5.2. Plan-view Scanning Electron Micrograph of 6-terminal relay design, showing the structural deformation of the Poly-$Si_{0.4}Ge_{0.6}$ flexural beam, in this case due to unmitigated joule heating during conduction.

In the case of structural fatigue, the effective spring constant k_{eff} can potentially decrease over time, in turn decreasing the value of V_{PI} in accordance with Eq. 1. Based on Euler-Bernoulli mechanical principles, which cover the mechanical deformation and general elasticity of beams due to a given load, k_{eff} for the vertical 6-terminal is proportional to structural dimensions as indicated in Eq. 2.

$$V_{PI} = \sqrt{\frac{8k_{eff}g^3}{27\varepsilon_0 A}} \quad (1)$$

$$k_{eff} \propto \frac{Ewt^3}{L^3} \quad (2)$$

where A = actuation area, g = actuation gap, E = Young's Modulus, w = beam width, t = structural thickness, and L = beam length [8].

Assuming negligible change in relay geometry, reduction in V_{PI} is primarily caused by reduction in k_{eff} (Eq. (1)), which is itself affected by change in the Young's Modulus E of the structural material (Eq. (2)). As such, reduction in V_{PI} is consistent with the formation of microscale cracks within the Poly-$Si_{0.4}Ge_{0.6}$ structure. In previous experimental work [5], it has been observed that a relatively small change in V_{PI} (approximately 0.4 V) occurs after 10^9 ON/OFF cycles for a non-body-biased, pull-in mode relay operated at f = 300 kHz and positive peak voltage V_{PP} = 18 V. This result suggests that the relays are promising for the aforementioned IoT applications. Further research is currently being conducted to assess the lifetime and efficacy of body-biased, non-pull-in mode relays.

In addition to structural fatigue, contact materials could also potentially deform over many switching cycles, due to the force upon impact during switching (i.e. the contact force) or micro-welding during capacitive discharge into the contact area (as will be discussed in a later section). Since the relay contacts are comprised of Tungsten (W), a very hard material, such deformation is not significant. In fact, previous experimental work indicates that W contacts do not show any signs of physical wear after 10^9 ON/OFF switching cycles, as indicated in the pre-and post-cycling atomic force microscopy (AFM) scans in Fig. 5.3. [9]. By operating the relay with body-bias, the switching impact velocity can be decreased such that softer contact is made [1,2,8].

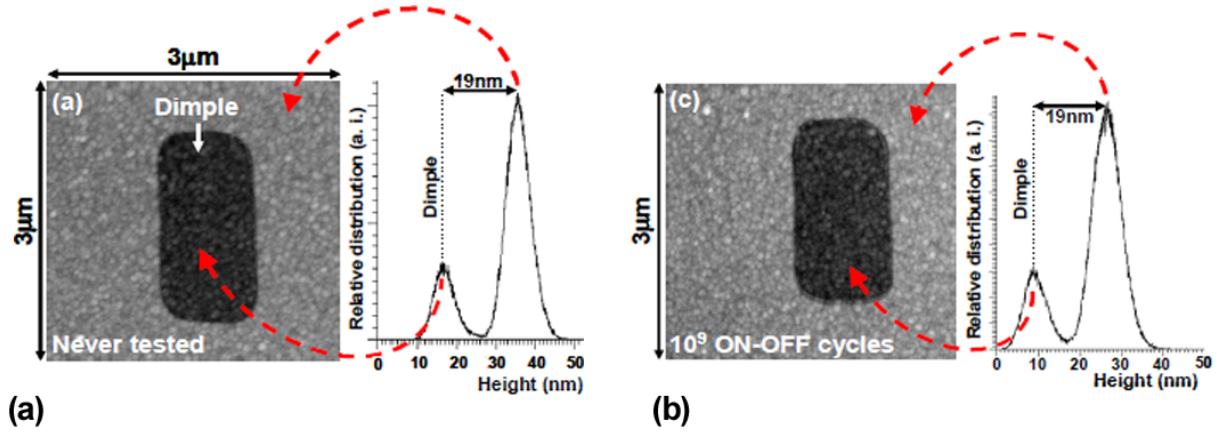


Figure 5.3. Atomic force microscopy (AFM) scans of TiO_2-coated W contacts from (a) a fresh relay and (b) a 10^9 cycled relay, indicating the robustness of such metal contacts for long-term operation [9].

5.1.2: Dielectric Charging

Depending on the quality of the Al_2O_3 insulating gate oxide, electrical charge can be accumulated within the insulating layer (due to energetically favorable "trap states") after many cycles. As such, the electrostatics of the parallel-plate capacitor system could be affected, thereby resulting in instability of the switching voltages, most notably V_{PI} [5]. A previous ON/OFF cyclic study [5], in which a non-body-biased, pull-in mode relay operated at f = 300 kHz and various positive peak voltages V_{PP}, suggests that the effect of such dielectric charging is relatively small over 10^{10} ON/OFF cycles. The result of this experiment is shown in Fig. 5.4.

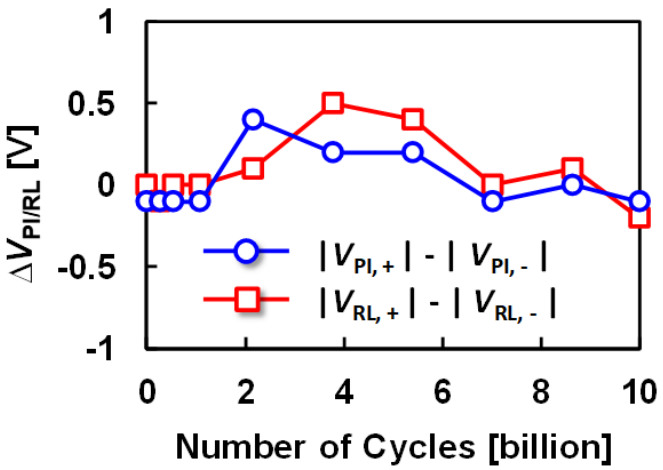


Figure 5.4. Evolution of V_{PI}, V_{RL} and V_H with respect to the number of ON/OFF switching cycles. The relative stability in V_{PI} and V_{RL} suggests that the effect of dielectric charging is negligible [5].

5.1.3: Contact Stiction and Micro-Welding

Contact stiction occurs when operating conditions or other factors prevent the device from operating. For example, contact stiction can occur if a device is exposed to liquid, in which case capillary forces cause the structure to collapse permanently to the substrate. Furthermore, joule heating during operation can lead to the contact terminals being permanently bonded to the channel [3,10], as illustrated in Fig. 5.5.

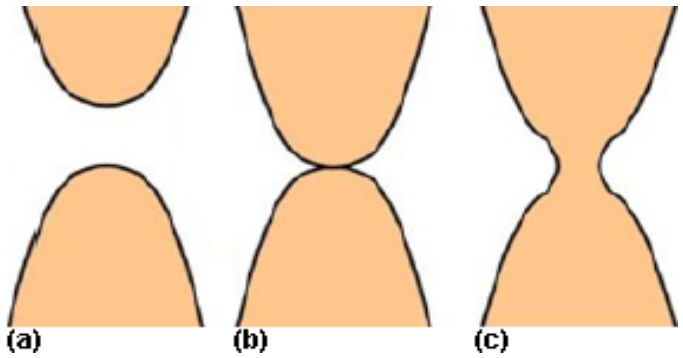


Figure 5.5. Qualitative illustration of relay contacts (a) separated (b) in-contact and (c) micro-welded. Mechanical/electrical stress and joule heat can cause such micro-welding to occur during operation [10].

The latter stiction mechanism, also known as micro-welding, is an issue particularly with low melting point materials, such as Gold. With high enough source-to-drain current, the contacts can weld together during operation, thereby preventing the relay from returning to the OFF-state [6, 10]. The measured number of ON/OFF cycles for a pull-in mode relay operated at 5 μTorr and 200 K (i.e. an environment less conducive to oxidation) is shown in Fig. 5.6. (Operating at an environment less conducive to oxidation, for example at 5 μTorr and 200 K can increase the endurance of the relay, i.e. the number of ON/OFF cycles, as seen from Fig. 5.6.) Additionally,

oxide thin films can be coated atop the contacting surfaces in order to further prevent micro-welding by slightly decreasing the conductivity [11]. Previous work indicates that while the contact resistance slightly increases with the inclusion of thin TiO_2 coating, the behavior of the contacts is still ohmic [9]. Self-assembled molecules have also been shown to reduce stiction (via surface adhesive force reduction) without significantly increasing ON-state resistance [3-4].

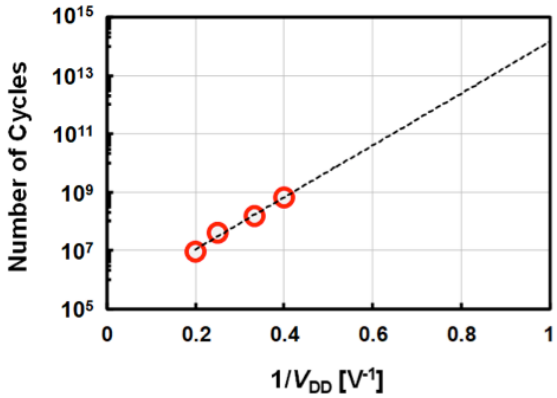


Figure 5.6. Measured and projected relay endurance with respect to supply voltage scaling [5].

During electrical characterization, current is artificially limited in order to prevent such welding. In Fig. 5.7, current is shown to be limited at 10uA by setting a current compliance limit during measurements.

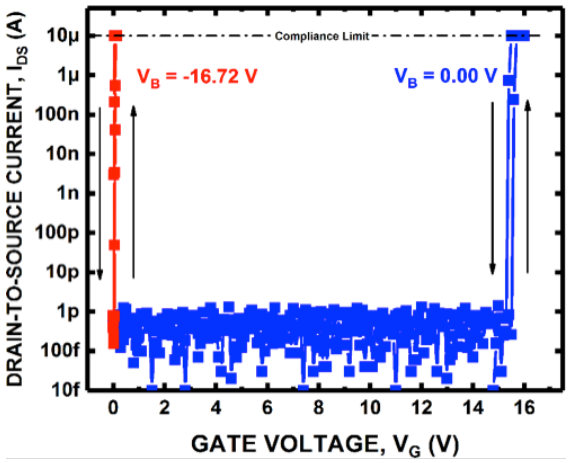
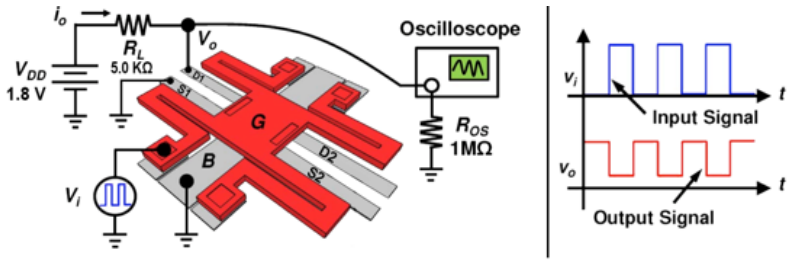


Figure 5.7. Measured current vs. voltage for a 6-terminal NEM relay.

5.1.4: Contact Oxidation

In order for the relays to conduct effectively for integrated circuit applications, the contact ON-state resistance (R_{ON}) must be below 10 kΩ [11-12]. Due to the reactivity of the contact material, Tungsten (W), the contacts oxidize, resulting in increased ON-state resistance. As shown in Fig. 5.8, R_{ON} can be experimentally determined via an inverter circuit, through which a voltage divider calculation bears the result.



$$R_{ON} = \left(\frac{V_O}{V_D - V_O}\right) R_L$$

Figure 5.8. Qualitative illustration of inverter circuit test setup, through which R_{ON} is determined experimentally.

Experimental results show that lower frequency operation (i.e. longer contacting/conduction periods) leads to earlier onset of oxidation [13]. Joule heating caused by conduction provides for higher oxidation rate. Other factors, such as contact force and relay type (e.g. 6-Terminal or 4-Terminal) also affect R_{ON} evolution over many cycles [5]. More explicitly, since the contact force distributed at each contact is higher for relays with less contacting points [14], native oxide is physically broken down more effectively for 4-terminal relays (versus 6-terminal relays), resulting in higher effective endurance shown in Fig. 5.9(c).

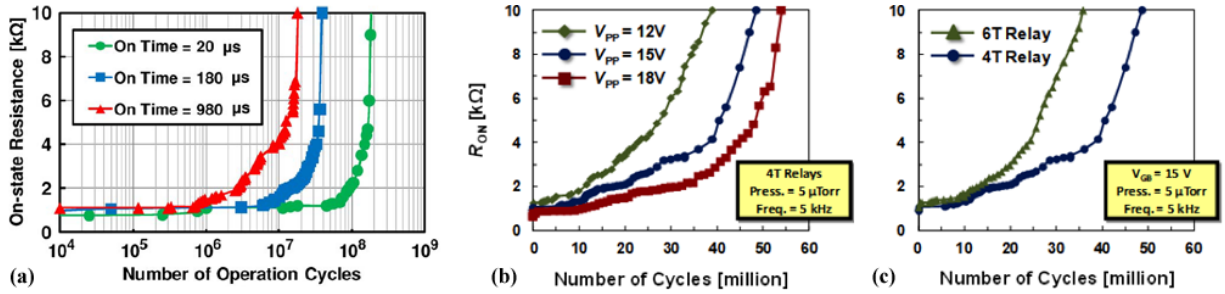


Figure 5.9. Measured R_{ON} over many ON/OFF cycles with respect to (a) switching frequency, (b) V_{PP} and (c) relay contact type [5, 13].

5.2: Reliability/Endurance Benefits of Supply Voltage Scaling

Since the N/MEM relays are primarily proposed as low voltage devices, it is worth noting that voltage scaling has a number of benefits regarding reliability. A major benefit is the reduced electrical fields within the dielectric which lower the chance of dielectric breakdown. Lower voltage across the S/D during contact can provide for lower rate of oxidation. Meanwhile, lower current results in lower chance of micro-welding or deformation due to thermal expansion. Endurance is projected to exceed 10^{15} cycles at [supply voltage] $V_{DD} \leq 1$ V, as indicated in Fig. 5.6 [5,15].

REFERENCES

[1] C. Qian et. al., "Energy-delay performance optimization of NEM logic relays," in *Proc. IEEE IEDM Tech. Dig.*, Dec. 2015.

[2] C. Qian et. al., “Effect of body biasing on energy-delay performance of logic relays,” *IEEE Electron Device Letters*, 2015.
[3] B. Osoba et. al., “Sub-50 mV NEM relay operation enabled by self-assembled molecular coating,” in *Proc. IEEE IEDM Tech. Dig.*, Dec. 2016.
[4] B. Osoba et. al., “Variability study for low-voltage microelectromechanical relay operation,” in *IEEE. Trans. Electron Devices*, Feb. 2018.
[5] Y. Chen et. al., “Reliability of MEM relays for zero leakage logic,” in *SPIE*, 2013.
[6] I. Chen, “Novel material integration for reliable and energy efficient NEM relay technology,” Ph.D. Dissertation, Dept. EECS, University of California, Berkeley, 2014.
[7] H. Kam et. al., “Design and Reliability of a Micro-Relay Technology for Zero-Standby-Power Digital Logic Applications,” in *Proc. IEEE IEDM Tech. Dig.*, Dec. 2009.
[8] C. Qian, “Electro-Mechanical Devices for Ultra-Low-Power Electronics.” Ph.D. Dissertation, University of California, Berkeley, 2017.
[9] R. Nathanael et. al., “4-terminal relay technology for complementary logic,” *IEEE IEDM Tech. Dig.*, pp. 223-226, 2009
[10] T. Ishida et. al., “Degradation mechanisms of contact point during switching operation of MEMS switch,” *IEEE JMEMS*, Vol. 22, No. 4, Aug. 2013.
[11] R. Nathanael, “Nano-Electro-Mechanical (NEM) relay devices and technology for ultra-low energy digital integrated circuits,” Ph.D. Dissertation, Dept. EECS, University of California, Berkeley, 2012.
[12] F. Chen et. al., “Integrated circuit design with NEM relays,” *IEEE ICCAD*, Nov. 2008.
[13] Y. Chen et. al., “Characterization of contact resistance stability in MEM relays with Tungsten electrodes,” in *IEEE JMEMS*, Vol. 21, No. 3, June 2012.
[14] Y.H. Yoon et. al., “4-Terminal MEMS relay with an extremely low contact resistance employing a novel one-contact design,” in *Transducers*, 2017.
[15] H. Kam et. al., “A predictive contact reliability model for MEM logic switches,” in *Proc. IEEE IEDM Tech. Dig.*, 2010.

CHAPTER 6

CHAPTER 6: ADHESION IN NANO-ELECTROMECHANICAL SWITCH (NEMS) DEVICES

Bivas Saha,[1,2] Sara Fathipour[1] and Junqiao Wu[1,2]

[1]Department of Materials Science and Engineering, University of California, Berkeley

[2]Materials Sciences Division and Molecular Foundry, Lawrence Berkeley National Laboratory

6.1 INTRODUCTION

As young students, many of us have often wondered why dew droplets are spherical when they spread on tree leaves or spider webs, or why scotch tape sticks to the surface of paper. These everyday observations of materials sticking on one another have aroused scientific inquisitiveness in many of our young minds, and they are explained broadly as adhesion or cohesion in materials science and solid state physics. Adhesion and cohesion refer to the tendency of dissimilar and similar surfaces to stick to one another respectively[1,2]. It plays an extremely important role in many branches of modern science and engineering, and are of particular interest for nano-electromechanical switches (NEMS)[3,4]. Engineering and controlling adhesive forces are fundamental to the development of energy efficient NEMS switches, and in this chapter, we will address the physical origin of adhesion, adhesion force and energy characterization, anti-adhesive molecules and their effects on the development of energy efficient NEMS devices.

Figure 6. 1 Spherical water droplets on leaves. Image adapted from https://www.designtrends.com/graphic-web/wallpapers/water-drop-wallpapers.html

6.2 ADHESION ENERGY

Adhesion energy ($W_{Ad.}$) per unit area between two materials A and B can be defined as[1]

$$W_{Ad.} = \sigma_A + \sigma_B - \gamma_{AB} \quad \ldots\ldots\ldots \text{Eq. 6.1}$$

where σ_A, σ_B are the surface energies of A and B respectively, and γ_{AB} is the interface energy between the A and B contact.

The origin of Eq. 6.1 can be explained physically with an aid of Fig. 6.2. Initially two materials A and B are shown to be in contact in Fig. 6.2(a). The work of adhesion, or adhesion energy of the contact is defined as the energy required to disintegrate the AB contact and separate the A and B materials to an infinite distance from one another (Fig. 6.2(b)), so that there is no interaction force between them. Therefore, the process involves creation of A and B surfaces. Since the energy required to create free surfaces are described by the surface energy of materials, σ_A and σ_B represent the surface energies of A and B that need to be provided for this process. Interface energy, on the other hand, refers to the extra energy that is stored in the material contacts due to lattice structure, lattice constant, surface energy and other mismatches of material properties that

would aid in the disintegration of AB contact, and therefore, has to be subtracted from the combined surface energies to obtain adhesion energy or the work of adhesion.

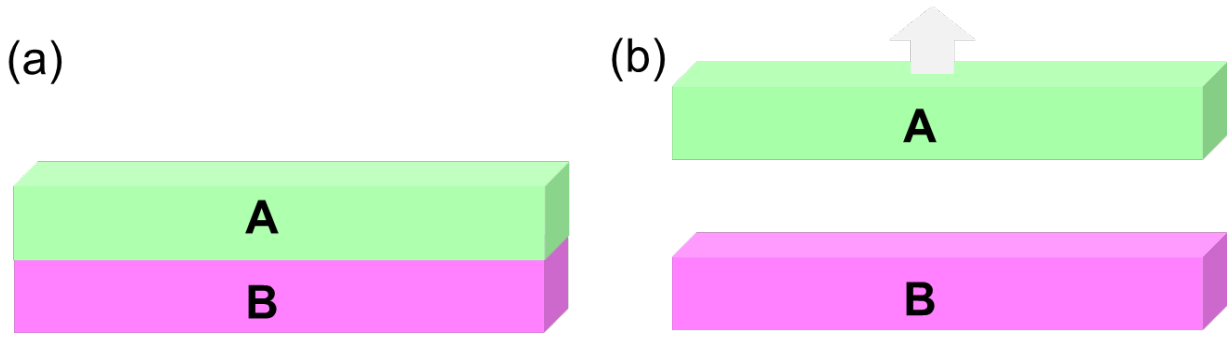


Figure 6.2: (a) Two materials A and B are in contact with one another. (b) Adhesion energy is referred as the energy required to separate A and B materials from their contact to an infinite distance. As shown in this process two surfaces A and B are created.

For all practical purposes (such as work function mismatch and others that are described in Chapter 5), contacting electrodes are usually fabricated with the same materials (i.e. A=B) in NEMS switches. This ensures $\gamma_{AB} = 0$, as the interface between same materials would not contain any extra interface energy. Equation 6.1 can be represented as

$W_{Ad.} = 2\sigma_A$ ……. Eq. 6.2

Eq. 6.2 suggests that for the same materials in contact, the adhesion energy to separate them is twice that of the surface energy. Therefore, knowledge of the surface energy of materials serves as an important metric to determine its adhesion properties.

For NEMS applications, electrode materials should not only have low surface and adhesion energies, they are also required to have good electrical conduction properties[5]. The simultaneous requirement of low adhesion energy and high electrical conductivity is one of the fundamental challenges in NEMS switch technology, and would be discussed in details in following sections.

6.3 ATOMISTIC ORIGIN OF ADHESION

Several factors contribute to adhesion between materials at the atomic scale. While some of the contributors are extrinsic in nature, which means that with proper care their effects can be reduced or even eliminated, several contributing factors are intrinsic to the materials and their properties, and as a result, they can only be engineered and cannot be avoided. Adhesion force can be obtained by summing over all the short-range and long-range forces that act between two materials when they are separated from each other. For most practical investigations[6], adhesion force $F_{Ad.}$ is a combination of van der Waals force ($F_{vdW.}$), capillary force ($F_{cap.}$) electrostatic force ($F_{el.}$), hydrogen bonding force ($F_{H.B.}$), and forces due to the chemical bonding or acid-base interactions (F_{Chem}).

$F_{Ad.} = F_{vdW.} + F_{cap.} + F_{el.} + F_{H.B.} + F_{Chem}$ ……. Eq. 6.3

6.3.1 van der Waals Force

The van der Waals (vdW) forces, named after Dutch scientist Johannes Diderik van der Waals, are distance-dependent weak interactions between atoms, molecules or solids. Unlike ionic or covalent bonds, these attractions are not a result of any chemical electronic bond, and are inherently weak. Van der Waals forces have no directional characteristic, and quickly vanish at longer distances between interacting molecules. If no other forces are present, the point at which the force becomes repulsive rather than attractive as two atoms near one another is called the van der Waals contact distance. This results from the electron clouds of two atoms unfavorably coming into contact.

Discussed in the form of inter-molecular forces, vdW forces have the following four major contributions. Except for the first one, they are all attractive and play a critical role in the adhesion forces in NEM switches.

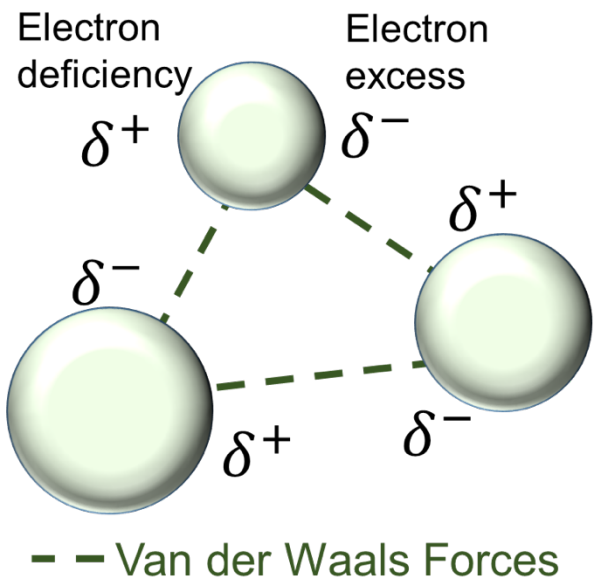


Figure 6.3 Van der Waals force is schematically presented.

A strongly repulsive component of the vdW forces results from a quantum mechanical effect known as the Pauli Exclusion Principle that prevents the collapse of molecules. This occurs only when the two molecules or surfaces are extremely close to each other, and is not the case for the NEM switches.

Attractive or repulsive electrostatic interactions between permanent charges (in the case of molecular ions), permanent dipoles (in the case of polar or asymmetric molecules), and in general permanent multipoles. This could exist when the two contact surfaces of the NEM switch contain trapped charges or molecules with permanent dipoles that are not screened by mobile charges.

Polarization force, which is the attractive interaction between a permanent multipole on one molecule and an induced multipole on another.

Dispersion force, which is the attractive interaction between any pair of molecules, including non-polar atoms, arises from the interactions of instantaneous multipoles. This force exists as attraction between the two surfaces of any closely placed objects, and is one of the key forces that needs to be alleviated in NEM switch technologies.

6.3.2 Capillary Force

Capillary forces are meniscus forces resulting from the presence of water or other liquid like contaminants on materials surfaces. Surfaces containing water and such other liquid contaminants will increase the adhesion energy through the capillary forces, and may cause significant problems. Capillary forces are particularly detrimental for the NEMS relay switch performance, as they increase the adhesion energy between the contacting electrodes, which as a result increase the

hysteresis voltage and limits energy efficiency. Previous implementations of the NEMS relay switch fabrication process used SiO_2 as sacrificial layers, which was typically etched with liquid Hydrofluoric Acid (HF). HF acid etching invariably makes the contacting electrode surfaces highly hydrophilic and increases the capillary forces. Recent implementations[5] of the NEMS switch by-pass this problem by using vapor HF, which mitigate capillary forces significantly. Yet capillary forces can be present in NEMS switches after the HF vapor release process due to exposure in humid conditions. Capillary force per unit area is expressed in the following equation,

$$F = \frac{2\gamma_l \cos\theta}{d} \qquad \text{..... Eq. 6.4}$$

where γ_l is the surface tension of the liquid, d is the separation between two surfaces and θ is the contact angle between the liquid and solid surface.

6.3.3 Hydrogen Bonding

Hydrogen bonding is an electrostatic attraction between two polar groups of molecules that occurs when a hydrogen atom covalently bonding to a highly electronegative atom (such as nitrogen, oxygen, or fluorine) experiences the electrostatic field of another highly electronegative atom nearby. Depending on the donor and acceptor atoms participating in bonding, the strength of the hydrogen bonds could range[6] from 5 to 50 mJ/mol. Hydrogen bonding is stronger than that of the van der Waals interaction, but weaker than covalent or ionic bonds. Presence of hydroxyl groups on NEMS contacting electrode surfaces could lead to strong hydrogen bonds when the separation between the contacting electrodes are very small.

6.3.4 Electrostatic Force

Electrostatic forces could play a dominant role in the adhesion interactions between contacting electrodes, especially on insulating surfaces in gaseous environments. Such interactions are caused by the inefficient charge dissipations from insulating surfaces. In aqueous solutions, most surfaces become charged due to dissociation of surface functional groups, and electrostatic forces become increasingly important. As conventional NEMS switch contact electrodes are usually made of conductive hard metals, electrostatic forces may not be of great concern. However, for functionalized switch surfaces, charge buildup may cause significant electrostatic interactions that would adversely impact the relay performance.

6.3.5 Chemical Bond Formation or Acid-Base Interactions

Presence of various chemical functional groups on the surface of contacting materials or constituent materials themselves (if they are reactive) may form chemical bonds that would cause increased adhesion. Along with the chemical bond formation, specific chemical interactions such as acid-base, receptor-ligand interactions may cause enhanced adhesion.

For energy efficient NEMS switch applications, where the adhesion between contacting electrodes need to be minimized, proper choice of materials and efficient device fabrication methodologies could eliminate most of the adhesive force mechanisms, except for van der Waals forces that are intrinsic to the adhesion.

6.4 ATOMIC FORCE MICROSCOPY (AFM) CHARACTERIZATION OF ADHESIVE FORCE.

Atomic force microscopy (AFM) is an efficient method for the measurement of adhesive force between materials. The basic operating principle of an AFM is presented in Fig. 6.4(a). An AFM consists primarily of a piezoelectric stage, a tip attached on a long cantilever, a laser, a photodetector, and feedback electronics as shown in Fig. 6.4(a). The laser light reflects from the back of the cantilever onto the photodetector. The movement of the laser light on the photodetector gives a measure of the deflection of the cantilever. The photodetector converts the deflection into an electrical signal. The AFM technique is widely used for topographic information such as height variations, surface roughness, etc., of materials. For adhesive force measurements, AFMs are used in the force spectroscopic mode.

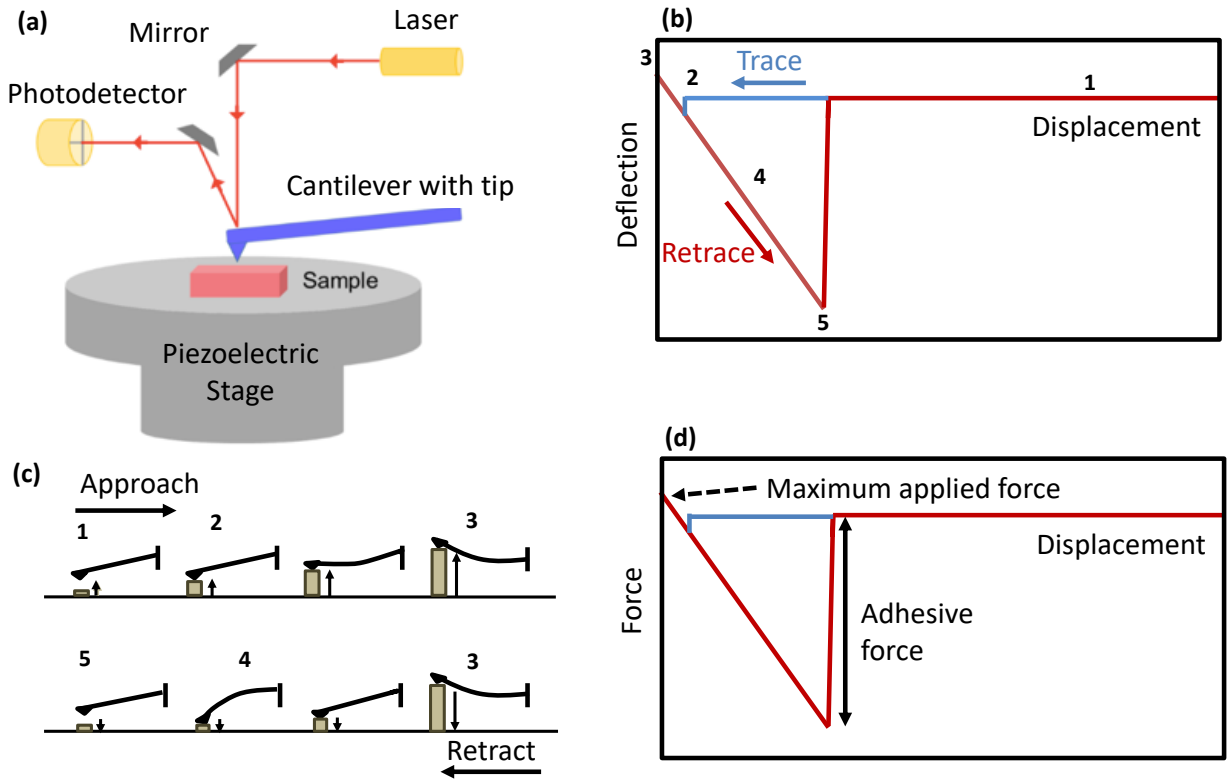


Figure 6.4. (a) Schematic diagram of an atomic force microscope (AFM) is presented. When operated in force-curve mode, AFM measurements can be used to determine adhesive interaction between materials. (b) Deflection-displacement curve of an AFM is presented with the trace and retrace components. (c)Schematic explaining the AFM deflection-displacement curve in (b). (d) Force-displacement curve of an AFM. Adhesive force can be determined from the retrace part as shown in the figure.

Fig. 6.4(b) shows the typical deflection-displacement curve of an AFM and Fig. 6.4(c) shows schematically how this curve is generated. Before the contact between the sample and the tip is established in section (1) of the curve, there is no deflection in the cantilever. At point (2) tip jumps into contact with sample. As the stage moves further toward the sample, the cantilever bends and its deflection is detected by the photodetector (3). As the piezo starts to move away from the sample in section (4), the cantilever bends in the opposite direction and the deflection is negative. There is always a hysteresis of force between approach and retrace curves, which is due to the adhesion between tip and the sample. At some tip-sample distance, the force from cantilever will be enough to overcome adhesion and the tip will be pulled off at point (5).

The deflection-displacement curve in Fig. 6.4(b) can be converted to force-displacement curve. For that first the voltage from photodetector should be correlated to the distance that the cantilever deflects. The deflection distance of the cantilever can be calculated by multiplying the voltage from photodetector in to sensitivity. Sensitivity is the reciprocal of the slope of force-deflection curve in region (4) of Fig. 6.4(b). When the deflection distance is calculated, it is straight forward to convert it to force by using the Hooke's law and by knowing the stiffness of the cantilever.

The typical force-displacement curve of an AFM is presented in Fig. 6.4(d). The small dip in the trace curve is known as snap-in and is due to the electrostatic attraction between the tip and sample surfaces when their separation is very small. The large jump in the retrace curve gives the amount of adhesive force.

6.5 THEORETICAL MODELS FOR CALCULATING ADHESION ENERGY

While an AFM is used to measure the adhesive force between materials as shown above, the knowledge of only adhesive force is not sufficient for the quantification of adhesion properties. Adhesive force depends on the tip-sample contact area during the AFM measurements, and scales with such contact areas between materials. Determination of the contact area during or after experiments, however, is extremely challenging and may not be feasible even with most advanced techniques available nowadays. Adhesion energy per unit area, on the other hand, is an absolute gauge for the determination of adhesion properties. Several theoretical models have been developed to quantify adhesion energy per unit area from adhesive force measurements, along with the knowledge of experimental conditions which, in some ways, approximates the contact area. Derjaguin used continuum mechanics model and suggested that the adhesion energy per unit area ($W_{Ad.}$) between a material with flat surface and a smooth AFM tip with diameter (R) is given by,

$$W_{Ad.} = \frac{F_{Ad.}}{\lambda \pi R} \quad \text{..... Eq. 6.5}$$

where $F_{Ad.}$ is the adhesive force, and λ is a parameter. Later work showed that based on the physical properties of the sample and tip, λ can vary from 1.5 to 2. In the simplest case when $\lambda = 2$, Eq. 6.5 is referred as Derjaguin-Muller-Toporov (DMT) model, while it is called Johnson–Kendall–Roberts (JKR) model when $\lambda = 1.5$.

Surface roughness of the sample and tip, however, can significantly impact the adhesion energy densities, and Eq. 6.5 would require corrections to incorporate the roughness information (see Ref. 6) for proper determination of adhesion properties of materials.

Surface energy per unit area of several common metals and polymers are presented in Table. 6.1. Most elemental metals such as W, Cu and others have extremely large adhesion energy in several Joules per meter square due to their large surface energies. Large surface energies on metals arise due to unsaturated dangling bonds. Soft materials such as polymers, on the other hand, usually possess low surface energies due to their saturated dangling bonds, in the range of a few milli-Joules per meter squared.

Table 6.1. Surface energy per unit area of several metals and

	σ (mJ/m^2)
W	~ 2300
Cu	~ 1650
Si	~ 1240
PMMA	~ 41
Teflon	~ 20

6.6 ANTI-STICTION MOLECULAR COATING

NEM relay switch contacts are usually made with conducting, hard metals that can survive mechanical wear and tear, and at the same time conduct electricity with low on-state resistance. However, metallic electrodes result in large adhesion energies, and hence, larger hysteresis voltage which increases the relay operating voltage. To mitigate this challenge, thin layers of anti-adhesive molecular coatings can be deposited on the metal electrode surfaces to reduce adhesion energy.

Several self-assembled molecular (SAM) coatings have been developed over the years for NEMS applications. Most of these SAMs consist of a head group (such as a -Silane or -Thiol functional groups) that help them attach on metal surfaces, and a tail group (such as $(–CH_2)_n$, $(-CF_2)_n$ and others) which acts as anti-adhesive regions. Silane functional groups are effective in attaching on Si, oxides, and some other metal surfaces (ideally having –OH bonds), while Thiols are used to attach on metals such as Au, Ag and Pt, to name a few.

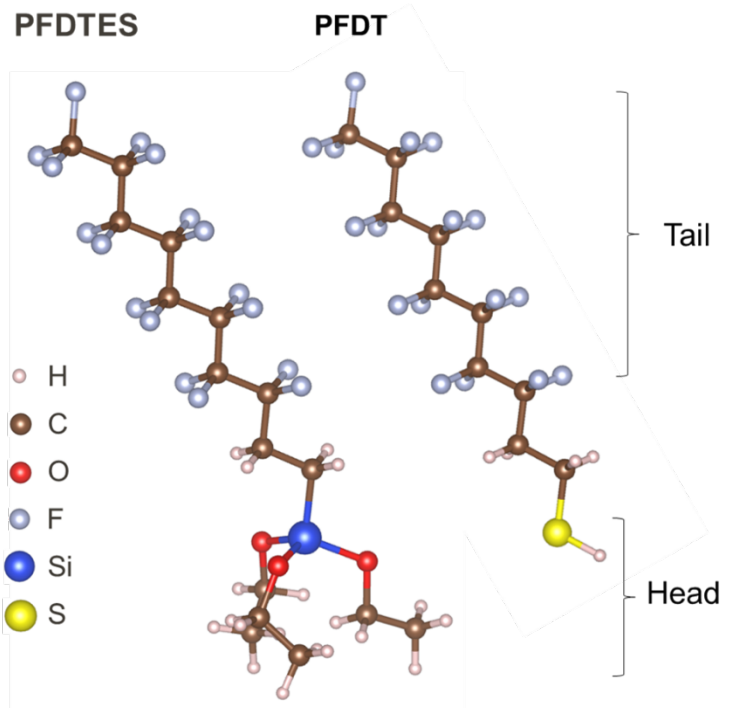


Figure 6. 5 Schematic diagram of 1H,1H,2H,2H-Perflurodecyltriethoxysilane (PFDTES) and 1H,1H,2H,2H-Perfluorodecanethiol (PFDT). The head group of the molecules consists of silane and thiol functional groups, respectively. The tail groups are made of CF_2.

A schematic diagram of anti-adhesive molecules is presented in Fig. 6.5.

For NEMS applications, anti-adhesive molecules are usually deposited on metal electrodes via vapor phase assembly, and characterized by techniques such as AFM, X-ray Photoelectron Spectroscopy (XPS) and others.

AFM measurements have showed that anti-adhesive SAMs such as *1H,1H,2H,2H-Perflurodecyltriethoxysilane (PFDTES)* based molecules have an extremely small adhesive energy of ~25 mJ/m^2, which is significantly smaller than the adhesive energies of metals such as W and Cu.

6.7 SUB-50 mV SWITCH ENABLED BY MOLECULAR COATING

We have seen in Chapter 4 that for a four terminal body-biased relay switch, the gate actuation voltage (amount of voltage required to switch a relay from ON to OFF state and vice versa) can be lowered to as small as ~150 mV by applying a suitable body-bias. Such low voltage operations were achieved primarily by the reduced hysteresis voltages that result from lower impact velocities of the electrodes due to body-biasing[7]. Reducing the gate actuation voltage, and hence the operating energy to even lower values, would require further reduction of the hysteresis voltage, which can be achieved by self-assembled anti-adhesive molecular coatings on W electrodes. In Fig. 6.6, the current-*vs.*-voltage (IV) characteristics for a PFDTES molecule coated NEM relay is presented along with the IV characteristics of an uncoated relay. The results show that the hysteresis voltage decreases from ~120 mV in uncoated relay to about 12 mV for the coated ones, which results in a reduction of gate actuation voltage from ~140 mV to ~60 mV. Repeated measurements with 100s of operating cycles have shown similar IV behavior, which demonstrates the benefit of anti-adhesive molecular coatings for lowering the adhesion energy of the contacts and hysteresis voltage in NEMS switches.

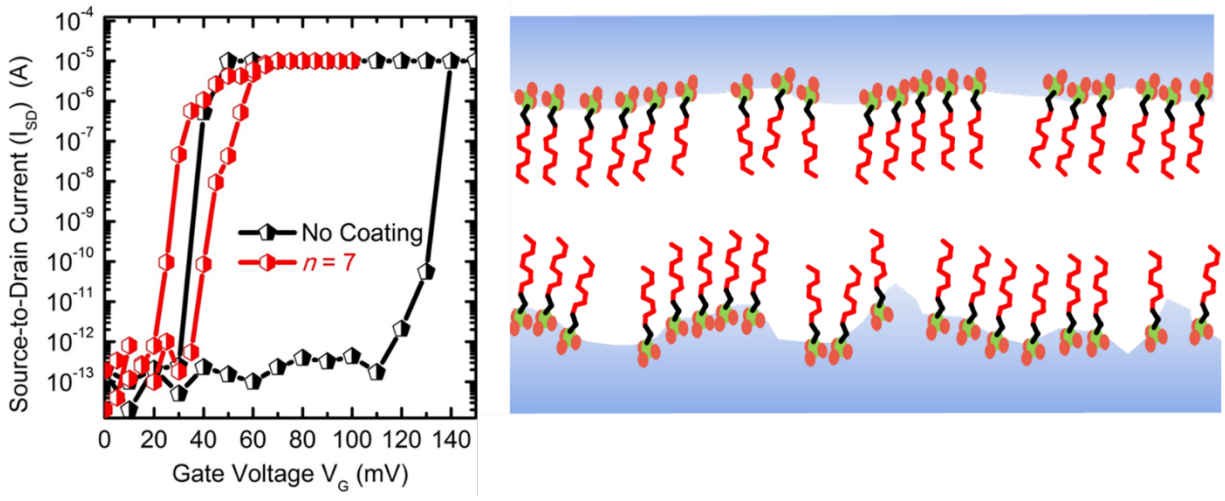


Figure 6.6. (a) Current-vs.-voltage (IV) characteristics of a relay before and after coating with PFDTES molecule (shown as n=7 here). Hysteresis voltage reduces from ~100 mV to ~10 mV for this relay, which help achieve an overall gate actuation voltage reduction from ~140 mV to ~60 mV. (b) Schematic diagram of relay contact surface coated with PFDTES molecule is shown.

Closer inspection of the IV curve reveals that although the hysteresis voltage of the molecules coated relays range between ~10 mV and ~20 mV, the gate actuation voltages range from 60 mV to 80 mV. This is because while the molecular coating reduced hysteresis voltage, the insulating nature of the molecules degrades the sharpness of the increase in current as a function of the applied voltage (sub-threshold swing). This increase in the sub-threshold swing was the reason why full benefits of the hysteresis voltage reduction could not be leveraged in reducing the overall gate actuation voltage. On average, the sub-threshold swing of uncoated relays ranges from 1-2 mV/decade due to the sharp increase in current at metal-metal contacts, while molecular coating increases the sub-threshold swing to ~10 mV/decade due to the metal/molecule-molecule/metal contact. Further analysis shows that molecules with shorter chain lengths improve the sub-threshold swing at the expense of increased hysteresis voltage, hence, the insulating nature of the molecules present a fundamental limit of how small operating voltages can be reduced with such approach.

6.8 WHY CONDUCTIVE MOLECULES WOULD BE IDEAL FOR LOW VOLTAGE NEMS SWITCH

Due to the trade-off between hysteresis voltage and sub-threshold swing in insulating anti-adhesive molecules coated NEM relays as presented above, the gate actuation voltages cannot be reduced to smaller values. The best results so far have showed that for body-biased relays with 5-6 orders of magnitude changes in current, the minimum gate-actuation voltage is ~25-40 mV. Therefore, further research is required and effective strategies need to be developed, whereby, the hysteresis voltage of the electrodes can be reduced without affecting the sub-threshold swing characteristics.

One approach to address this challenge will be to develop anti-adhesive molecules that are also electrically conductive. In such molecules, lower adhesion energy would reduce the hysteresis voltage, while their metallic electrical conductivity will result in sharp current increase with the increase in voltages. Unfortunately, no such molecules are readily available today that may help in this pursuit, and collaborative research involving scientists and engineers from Chemistry, Materials Science and Electrical Engineering will be necessary.

6.9 SUMMARY

In summary, contact adhesive force plays an extremely important role on nano-electromechanical (NEM) switch device performance. Engineering adhesion will be essential and important not only for NEM-based switch technology, but also for NEM based non-volatile memory and other emerging applications. Identification and characterization of materials with appropriate adhesive properties will greatly help the entire NEMS research field, and will result in energy-efficient device technologies with unprecedented social and economic impacts.

REFERENCES

1. L. Lee, Fundamentals of Adhesion (Springer Science Press, 1991 and 2001).
2. J. N. Israelachvili, Intermolecular and Surface Forces (Academic Press, 2011).

3. C. Qian, A. Peschot, I. Chen, Y. Chen, N. Xu, and T. J. K. Liu, IEEE Electron Device Lett. 36(8), 862 (2015).
4. J. Yaung, L. Hutin, J. Jeon, and T.-J. K. Liu, J. Microelectromech. Syst. 23(1), 198–203 (2014).
5. J. Yaung "NEM Relay Scaling for Ultra-low Power Digital Logic" Ph.D. dissertation (University of California, Berkeley, 2014).
6. H. Butt, B. Cappela, and M. Kappl, Surf. Sci. Rep. 59, 1–152 (2005).
7. C. Qian, A. Peschot, I. Chen, Y. Chen, N. Xu, and T. J. K. Liu, IEEE Electron Device Lett. 36(8), 862 (2015).

CHAPTER 7

CHAPTER 7: Tunneling Nanoelectromechanical Switches "SQUITCHES"

Farnaz Niroui[1,2], Timothy Swager[3], Jeffrey H. Lang, Vladimir Bulovic

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Massachusetts 02139, USA

[2]Department of Chemistry, Massachusetts Institute of Technology, Massachusetts 02139, USA

[3]Miller Research Institute, University of California Berkeley, California 94720, USA

7.1 INTRODUCTION

As we discussed in the previous chapters, to achieve an energy-efficient nanoelectromechanical (NEM) switch, we need a design that enables low-voltage and low-hysteresis actuation. The most successful implementation requires simultaneous miniaturization of the switching gap and elimination of stiction by controlling the surface adhesive forces. An example approach was introduced in the previous chapters where a combination of body-bias and molecular coatings enables sub-100 mV switching operation. In this design the body-bias helps lowering the effective actuation voltage while the molecular coating modifies the surfaces to achieve a lower surface energy resulting in a smaller hysteresis. In this chapter we will introduce an alternative design utilizing molecular layers. Here, beyond leveraging the chemical properties of the molecules, their structural and mechanical properties are used to create a platform for low-voltage and low-stiction switches that work based on modulation of quantum tunneling current.

7.2 Tunneling Nanoelectromechanical Switches - "Squitches"

Conventionally, electromechanical switches operate by electrostatically actuating a movable electrode to come into direct contact with an opposing stationary electrode. As the surfaces come into contact, they experience a large amount of surface adhesive forces leading to stiction. Stiction can be minimized if direct contact between the electrodes is eliminated while at the same time a mechanism for force control is introduced. To achieve this, we propose a design which operates based on electromechanical modulation of tunneling current. Unlike conventional switches, these devices, which we refer to as squeezable switches or squitches, consist of a switching gap made out of a compressible insulating molecular layer sandwiched between two conductive surfaces as see in Figure 1. This

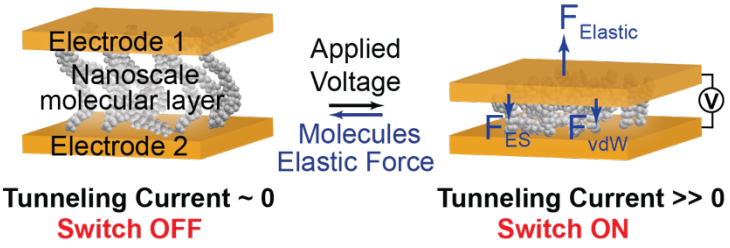


Figure 7.1 – Switching mechanism of a squitch: Molecular layers are used as nanoscale springs to provide nanoscale force control while avoiding direct contact between the two electrodes. An applied electrostatic force (F_{ES}) compresses the molecules connecting two approaching electrodes. The elastic force ($F_{elastic}$) stored in compressed molecules overcomes the van der Waals forces (F_{vdW}) to make the process controlled and reversible.

metal-molecule-metal junction with metal-metal separation of < 3 nm can allow conduction of electrons through the process of quantum tunneling despite no direct contact with the two conductive surfaces. Before discussing the switching mechanism in more detail, let's first define what is quantum tunneling.

7.2.1 What is Quantum Tunneling?

In classical terms, electrons are considered to be particles. When approaching a barrier, these particles will not be able to get across the barrier if they do not have enough energy to move over it. In quantum mechanics though the electrons are considered to have both particle- and wave-like properties. The wave nature of electrons allows them to be able to get through a barrier that they would classically not be able to overcome. These waves don't end abruptly at the barrier, but taper off quickly. If the barrier is thin enough, then the probability function can extend into the next region through the barrier. This process of electrons passing through the barrier is referred to as tunneling. In the case of metal-molecule-metal junction used in the proposed design of squitches, the insulating molecular layer creates a barrier to electron flow. If the molecular film is thin enough, that is less than <3 nm, then one can expect a tunneling current to flow in response to an applied voltage across the electrodes. The tunneling probability though increases as the barrier width decreases, this leads to an exponential increase in the current with decreasing tunneling distance (barrier width) (Figure 3). This current modulation helps define the squitch switching mechanism.

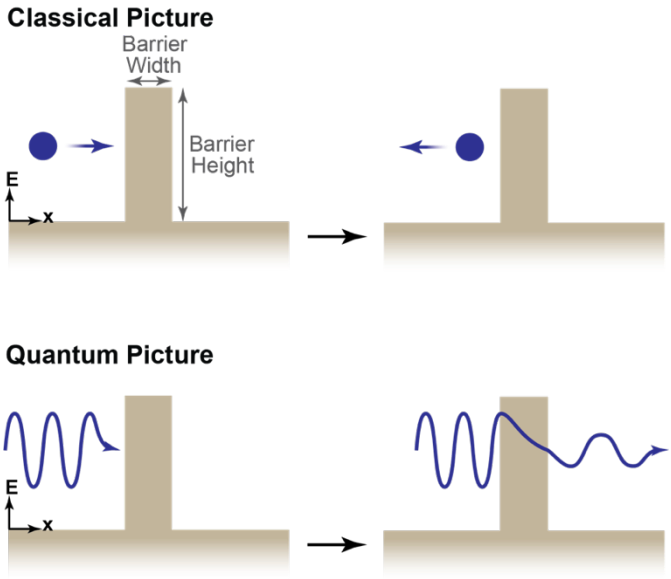


Figure 7.2 – In classical terms, an electron acting like a particle will not overcome a barrier without sufficient energy. In quantum terms, an electron with wave-like properties is able to get through a barrier to the other side through a process referred to as tunneling.

7.2.2 Switching mechanism in Squitches

In a squitch, the molecular layer serves as a structural support to help define the tunneling gap with thickness much smaller than feasible in absence of the molecules. In absence of the molecules, the van der Waals forces can cause collapse of the electrodes onto each other. The molecules provide an elastic resorting force though to balance out the van der Waals forces to allow for a stable structure. These compressible molecules also serve as nanoscale springs to help precisely modulate the tunneling gap. When a voltage is applied across the electrodes, an electrostatic force is induced which attracts the top electrode towards the bottom. In the process, the molecular layer which has a low Young's modulus will be compressed. As the molecules compress, the gap between the electrode is reduced and as a result the tunneling current through the junction exponentially increases. This drastic increase in current corresponds to the device turning on. When the applied

voltage is removed, the elastic restoring force stored in the compressed molecules help overcome the surface adhesive forces to allow the top electrode to transition back to its original position to turn off the device.

7.3 Squitch Fabrication

The main component of a squitch is the metal-molecule-metal switching gap. The molecular layer is formed through a self-assembly process. In this approach, the molecules are selected with a desired functional group which allows spontaneous and selective assembly onto the desired surface. For example, a S-H end group would allow specific assembly onto gold. The molecules are also selected to have a low Young's modulus such that the layer is compressible enough to be modulated electrostatically at a low voltage. An example molecule that can be used to form the molecular gap is poly(ethylene glycol)dithiol (PEG-dithiol) which gives the desired thickness (~ 3 nm in the off-state) and low Young's modulus (Figure 4a). The transmission electron microscope (TEM) images in Figure 4b shows a 3 nm thin layer of PEG-dithiol assembled onto Au surface.

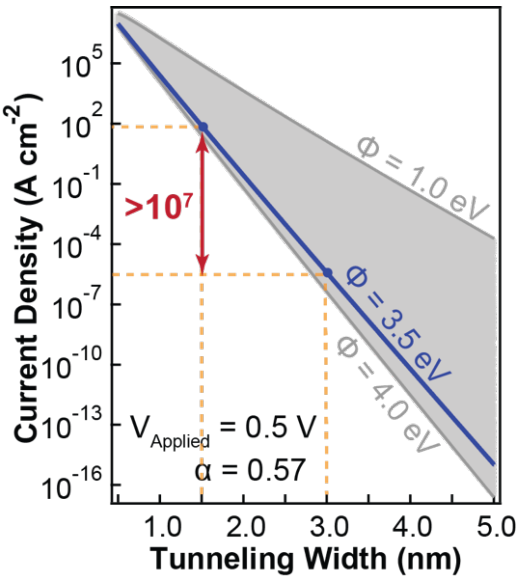


Figure 7.3 - Exponential increase in tunneling current as a function of decrease in tunneling gap due to electrostatically-induced compression of molecules in metal-molecule-metal tunnel junctions serves as a switching mechanism for the squitches. Here, Φ is the tunneling barrier height and α is a parameter accounting for barrier shape and electron effective mass.

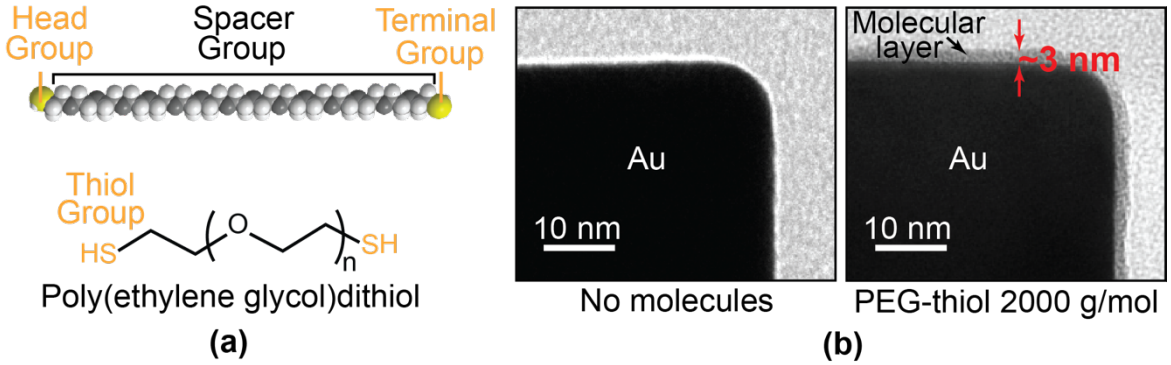


Figure 7.2 – (a) Molecules which can be precisely engineered through chemical synthesis with the desired head, terminal and spacer groups can serve as nanoscale building blocks to promote development of squitches. An example molecule used in the design is poly(ethylene glycol)dithiol. (b) The thiol end group in PEG-dithiol allows specific self-assembly onto Au film as shown in the TEM images.

The next step in fabrication is to ensure that the molecular layer is sandwiched between two conductive contacts. Placement of the top electrode on the molecular layer through conventional deposition techniques without inducing damage to the molecules is challenging. Often, the metal directly deposited onto the molecular layer can penetrate through the film and form a nonuniform molecular gap and consequently lead to electrical shorting during operation. Such nonuniform gap is compared to an ideal design is shown in Figure 5. This leads to a low device yield and an unreliable performance. To avoid this problem, an approach is necessary to introduce the top

contact on the molecular film without inducing damage to the self-assembled layer. As an example, we can use an atomically smooth graphene layer for the top contact. The graphene can be grown through chemical vapor deposition and transferred onto the molecular layer after self-assembly. This additive process which prevents direct deposition of metal over the molecular film avoids damage to the molecular layer while maintaining a uniform thickness because of the smooth surface of graphene. The fabrication scheme for an example squitch formed through this approach is shown in Figure 6.

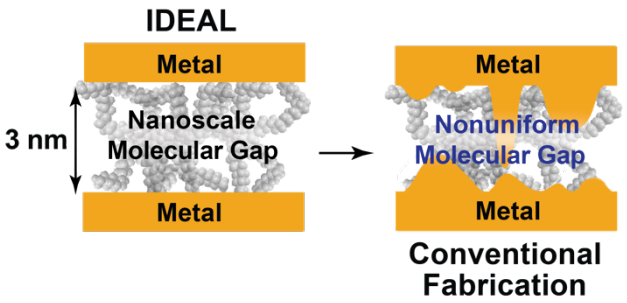


Figure 7.5 – Conventional fabrication techniques lead to rough surfaces which prevents formation of uniform molecular gaps as ideally desired and leads to non-uniform device performance and low device yield.

7.4 Example Squitch Performance

In a squitch with the design shown in Figure 7 and fabricated using the scheme shown in Figure 6, an applied voltage between the bottom two electrodes produces an electrostatic force to attract the graphene top electrode towards the underlying electrodes [1]. During this process the molecular layer is compressed and tunneling current modulated. An example current-voltage characteristic of such a device is shown in Figure 7. Initially, there is an exponential increase in the current corresponding to the decrease in the tunneling gap. This region is followed by an abrupt jump in current. This rapid increase occurs at the pull-in point- this is the point at which the combined electrostatic and van der Waals forces overwhelm the elastic restoring force of the molecules such that the top electrode rapidly accelerates towards the bottom to cause a rapid decrease in the gap width and a rapid increase in the current. In this device, electromechanical modulation of the PEG nanogap with few MPa Young's modulus leads to greater than 4 orders of magnitude change in the current conduction with sub-2 V operating voltage.

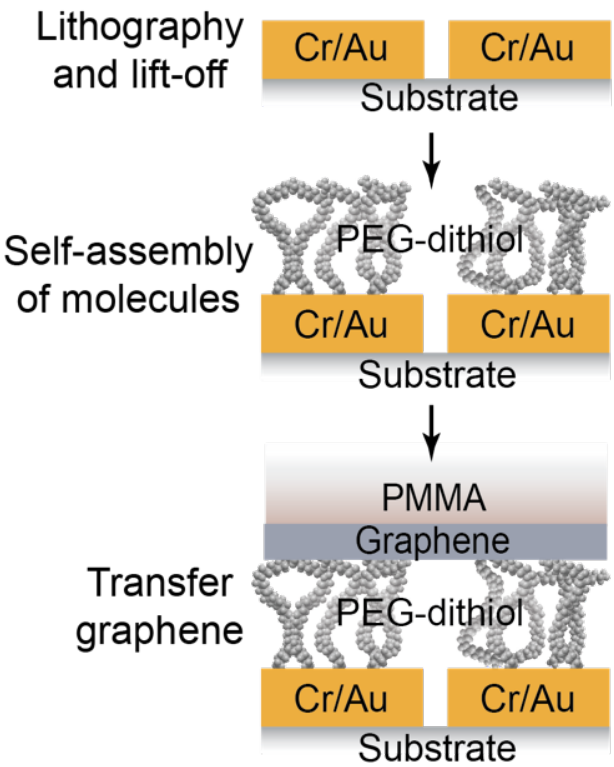


Figure 7.6 – Squitch fabrication scheme based on lithographically patterned Au bottom electrodes, self-assembled PEG-dithiol molecular layer and graphene top electrode.

7.5 Conclusions

Electromechanical modulation of the tunneling current in a metal-molecule-metal switching gap can serve as a switching mechanism in squitches. In these devices, the molecular layer helps in defining miniaturized switching gaps and nanoscale force control to allow for lowering of the actuation voltage and possibility of stiction. Optimization of the device design and mechanical properties of the molecular film can lead to further improvements in the switching performance, making these devices a promising platform to develop stiction-free and low-voltage NEM squitches.

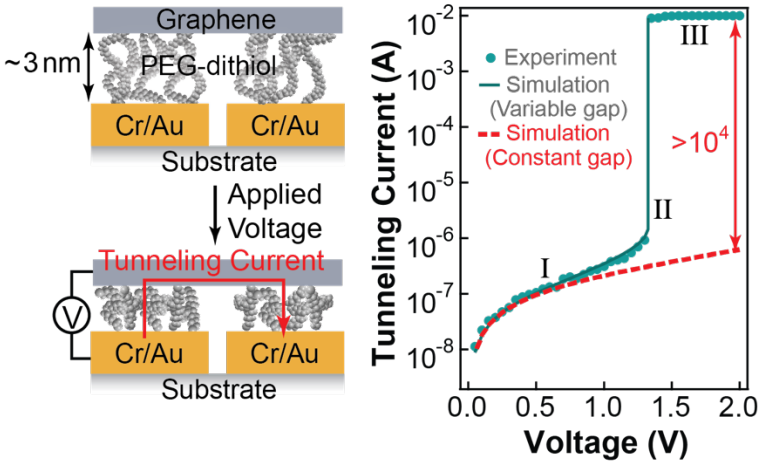


Figure 7.7 – Current-voltage characteristics of a tunneling squitch based on a gold bottom electrode, self-assembled PEG-dithiol molecular layer, and graphene top electrode. The device shows an on-off ratio more than 10^4 and actuation voltage of ~ 1.4V.

REFERENCES

1. F. Niroui, A.I. Wang, E. M. Sletten, Y. Song, J. Kong, E. Yablonovitch, T. M. Swager, J. H. Lang, and V. Bulović, "Tunneling Nanoelectromechanical switches based on compressible molecular thin films," ACS Nano, vol. 9, pp. 7886-7894, 2015.
2. F. Niroui, E.M. Sletten, P.B. Deotare, A.I. Wang, T.M. Swager, J.H. Lang, and V. Bulović, "Controlled fabrication of nanoscale gaps using stiction," in *Proc. 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS)*, 2015, pp. 85-88.
3. F. Niroui, P.B. Deotare, E.M. Sletten, A.I. Wang, E. Yablonovitch, T.M. Swager, J.H. Lang, and V. Bulović, "Nanoelectromechanical tunneling switches based on self-assembled molecular layers," in *Proc. 27th IEEE International Conference on Micro Electro Mechanical Systems (MEMS)*, 2014, pp. 1103-1106.

CHAPTER 8

CHAPTER 8: THE STRITCH-SWITCH DEVICE

Aldo Vidana[1], Sergio Almeida[2], Mariana Martinez[1], Edgar Acosta[1], Jose Mireles[1], David Zubia[1]
[1]Electrical and Computer Engineering, University of Texas at El Paso
[2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

8.1 INTRODUCTION

Excellent switches can be made with NEMS devices. However, as mentioned in previous Chapters, the NEMS switch has the problem of not being reliable enough for electronic applications. Another problem is that the voltage needed to operate a NEMS switch is too high and consumes an excessive amount of energy. These are the main problems the squitch device, discussed in the previous Chapter, is hoping to solve. However, another approach to making highly reliable and low-voltage switches with NEMS is to stretch a thin material so that the electrical conductivity of the thin material changes by several orders of magnitude. This makes the thin material work like a switch.

8.2 THE STRITCH-SWITCH CONCEPT

Diagrams of this Stretch-Switch, or "Stritch", are shown in Figure 8.1. The idea of the stritch device is that a NEMS stretches a thin film to cause it to dramatically change its conductivity. In Figure 8.1(a), a NEMS shown in blue acts as a cantilever actuator. The actuator has three electrical terminals and several fixed and movable parts. The three terminals are the source, gate and drain and they are fixed or cannot move. In contrast, the cantilever is able to move and is attached to the source through a serpentine spring. When the voltage between the source and the gate is zero, the cantilever rests equidistant between the gate and drain as shown in Figure 8.1(b). This is called the "off" state and the thin film which is attached to the cantilever and drain will have a very low electrical conductance. In this "off" state, the switch is turned off. However, when a voltage is applied between the source and gate, an electrostatic force of attraction is created between the cantilever and gate, which causes the cantilever to move towards the gate (and away from the drain) according to how much voltage is applied, as shown in Figure 8.1(c). In this case, the thin film is stretched causing its conductivity to dramatically increase by many orders of magnitude. This is the "on" state of the device, because the thin film is able to conduct electric current between the drain and source.

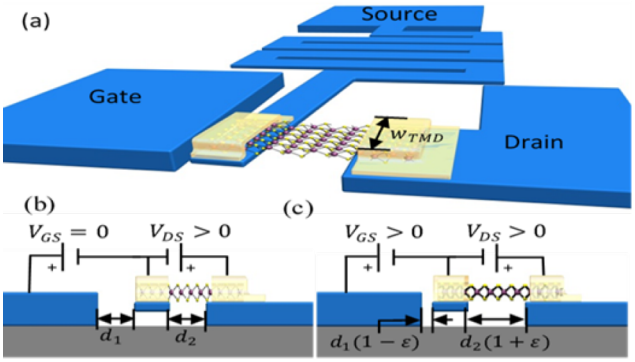


Figure 8.1 - (a) Perspective view of the stretch-switch with a stretched thin film. Cross-sectional views of the device in the (b) "off" and (c) "on" states."

8.3 CONDUCTIVITY VERSUS STRAIN OF THIN FILMS

A key to the device functioning very well as a switch is that the conductivity of the thin film needs to increase by several orders of magnitude when it is stretched. Luckily, there is a type of thin film which has this important property. The thin films are made of transition-metal and chalcogen atoms from the periodic table and have the general chemical formula MX_2, where M represents the transition metal and X_2 represents two chalcogen atoms. Some examples of transition metals are molybdenum, tungsten, titanium, zirconium and hafnium. Examples of chalcogens are sulfur, selenium and tellurium. Because there are two chalcogen atoms for every transition-metal atom, the material is called "transition- metal di-chalcogenides" or TMDs for short. Another interesting property of TMDs is that they can be made very thin. Actually, they can be made into films that are only 3 atoms thick as shown in Figure 8.2. Because TMDs can be only 3 atoms thick they are often referred to as two-dimensional (2D) layers.

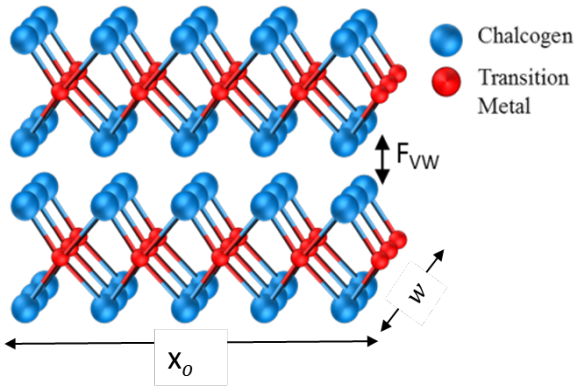


Figure 8.2 - Schematic of a double-layer TMD thin-film, where each layer consists of three atomic layers in which the transition metal is covalently bonded by two chalcogen atoms in a sandwich-like structure. The two TMD layers are held together by weak van der Waals forces (F_{VW}).

Figure 8.3 plots the reduced conductivity of TMD layers as a function of strain or how much they are stretched. Strain (ε) is a scientific way to define how much a material is stretched using the formula $\varepsilon = (x_f - x_o)/x_o$, where x_o is the original length of the material and x_f is the length when it is stretched. If a material is not stretched, then $\varepsilon = 0$. If a material is stretched to twice its original size, then $\varepsilon = 1$.

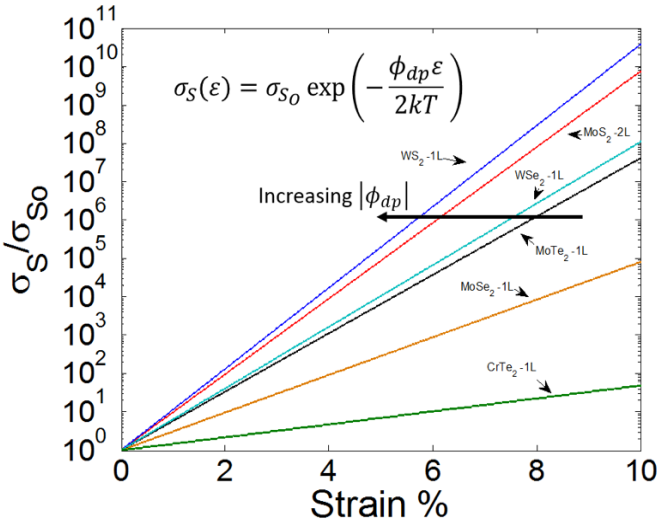


Figure 8.3 – Calculated conductivity versus strain of several TMD thin-films.

The conductivity (σ_s) of a material specifies how easily electric current can flow through it. Low conductivity means that it is difficult for electric current to flow through the material. High conductivity means that electric current can flow easily. A good switch used for digital circuits can change its conductivity by at least six orders of magnitude when it is switched between on and off ($\sigma_{s\text{-}on}/\sigma_{s\text{-}off} = 10^6$). When the conductivity is low, the switch is "off". When the conductivity is high, the switch is "on".

Figure 8.3 shows that the conductivity increases with increasing strain of several TMDs including; MoS_2, WSe_2, WS_2, $MoSe_2$, $MoTe_2$, and $CrTe_2$. In this Chapter we will focus on MoS_2. For example, Figure 8.3 shows that a bi-layer of MoS_2 will increase its conductivity, relative to when

is it not stretched (σ_s/σ_{so}), by six orders of magnitude when it is stretched to only approximately $\varepsilon = 0.06$. This means that MoS_2 can be used as a switch but it needs to be stretched to 6% longer than its original length.

8.4 MODEL OF THE MEMS ACTUATOR

A mechanical device or apparatus is needed to physically stretch a material. For example, a person can stretch a rubber band using hands. In the stritch, a NEMS is used to stretch a TMD as described in Figure 8.1. However, in order to be able to calculate the forces that are needed to stretch the TMD, a formula is needed to mathematically describe the forces in the different parts of the NEMS. The mathematical formula is called the "model" of the NEMS.

Figure 8.4 is a schematic model showing the forces active in the NEMS. The blue component is the cantilever which moves according to the forces pulling on it. The cantilever is the movable part in Figure 8.1 that can move either to the gate or drain. Although the cantilever can move, it is attached to a fixed electrical terminal via a serpentine spring as shown in Figure 8.1. In Figure 8.4, the serpentine spring is represented as a simple spring which will create the force called F_{MEMS}. Similarly, the TMD (which is attached to the cantilever at one end and to the drain at the other) is also shown as a simple spring and will create the force labeled, F_{TMD}, when it is stretched. F_{MEMS} and F_{TMD} act together to oppose motion of the cantilever towards the gate. On the other side of the cantilever is an electrostatic force, F_E, which acts to move the cantilever towards the gate and therefore stretch the TMD. F_E is created by applying a voltage, V_{GS}, between the gate and source electrodes as shown in Figure 8.1.

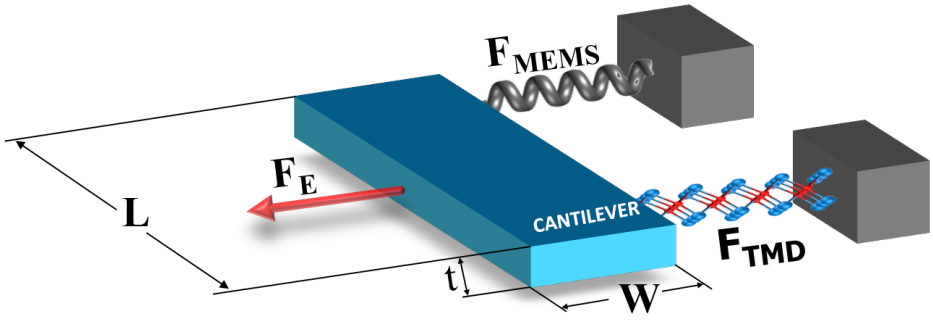


Figure 8.4 - Schematic model of the NEMS showing the active forces.

Since F_E acts to stretch the TMD, and F_{MEMS} and F_{TMD} act to oppose stretching the TMD, this can be described by a simple equation called the "force balance model", $\boldsymbol{F_E = F_{MEMS} + F_{TMD}}$. As can be seen in the force balance model, F_E is on one side of the equation, while F_{MEMS} and F_{TMD} are on the other side.

The actual value for each of the forces depends on the physical size and mechanical properties of the NEMS and TMD. While in general the forces decrease as the thickness of a material is reduced, it is essential to also take into consideration the stiffness of the materials. For example, it is easier to stretch a skinny rubber band than a thick one. However, it is harder to stretch a thin metal wire compared to a rubber band because metal is stiffer than rubber. Therefore, to accurately calculate

the forces in the stritch device, both the size and stiffness of the materials need to be taken into account.

8.5 STRAIN VERSUS VOLTAGE

While there are many choices of materials for the NEMS and TMD, the analysis will focus on using silicon for the NEMS and MoS_2 for the TMD. This specification will allow calculation of the forces and determination of how much voltage is needed to strain the stritch device. Figure 8.5 plots the amount of strain in the TMD as a function of voltage, V_{GS}, assuming the length and width of the TMD equal each other (X_o=W_{TMD}). It is observed that the amount of voltage needed to strain the MoS_2 is reduced with decreasing lengths/widths of the TMD. For example, a voltage of V_{GS}= 0.055 Volts is needed obtain a strain of $\varepsilon = 0.06$ when $x_o = W_{TMD} = 10$ nm. It is desirable to reduce the voltage because this also reduces the energy needed to switch the device. This analysis indicates that reducing the length/width of TMD will decrease the switching energy.

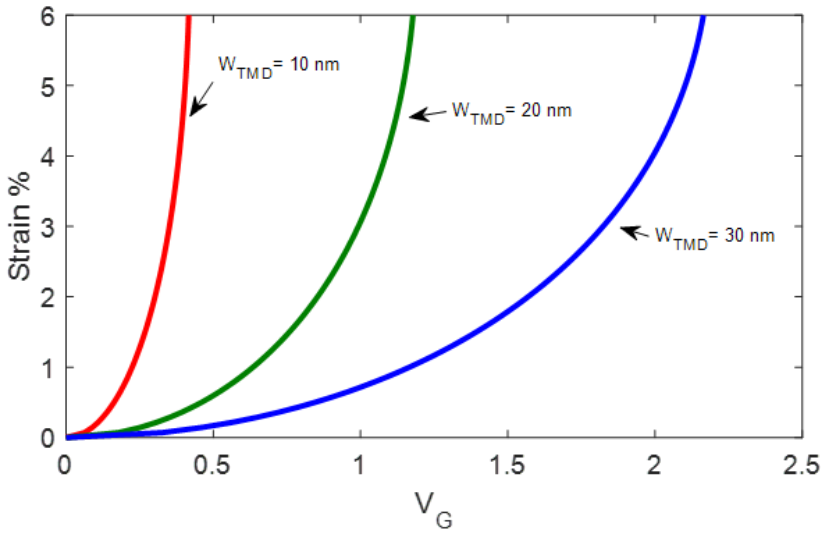


Figure 8.5 - The calculated strain versus applied voltage for three lengths/widths of the TMD.

8.6 CONDUCTIVITY VS. VOLTAGE SWITCHING CHARACTERISTIC OF THE DEVICE

By combining the information on conductivity versus strain (Figure 8.3) with information on strain versus voltage (Figure 8.5), a plot of the conductivity versus the voltage can be obtained as shown in Figure 8.6. This plot is important because it shows the relationship between the output (conductivity) versus the input (voltage) of the device. Figure 8.6 takes into consideration the characteristics of both the NEMS and TMD aspects of the stritch. It shows that the amount of voltage needed to switch the device decreases with decreasing length/width of the TMD from x_o =W_{TMD} = 30 nm to 20 nm to 10 nm.

For comparison, Figure 8.6 also shows the theoretical maximum subthreshold swing of a standard CMOS device which has a slope 60 mV/decade. It is important to note that in theory, the stritch

device can have a comparable or better switching characteristic to CMOS if its length/width is 10 nm or less.

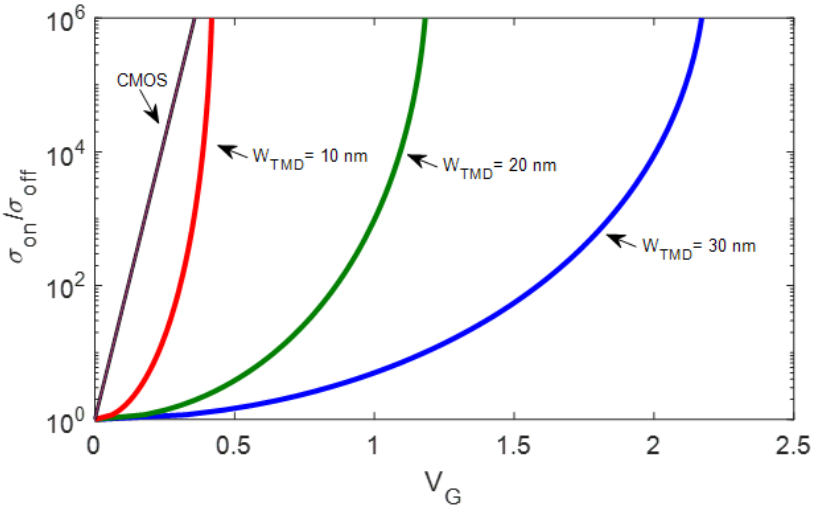


Figure 8.6 - The change in the conductivity versus the applied voltage for three lengths of the TMD. The theoretical subthreshold swing of a CMOS device is also shown as reference.

8.7 SWITCHING ENERGY

Figure 8.7 shows the energy needed to switch the stritch device as a function of the TMD length, x_o. The energy decreases as the length of the TMD is reduced. At x_o=W_{TMD} = 10 nm, only 23 attojoules are needed to switch the device. The energy is comparable to the energy needed for CMOS.

Although, the switching energy is very low, the voltage is not as low as desired. This is a challenge with the stritch device. However, there is a possibility to use other forces within the NEMS to lower the voltage when the distance between the gate and source is smaller than approximately 15 nm.

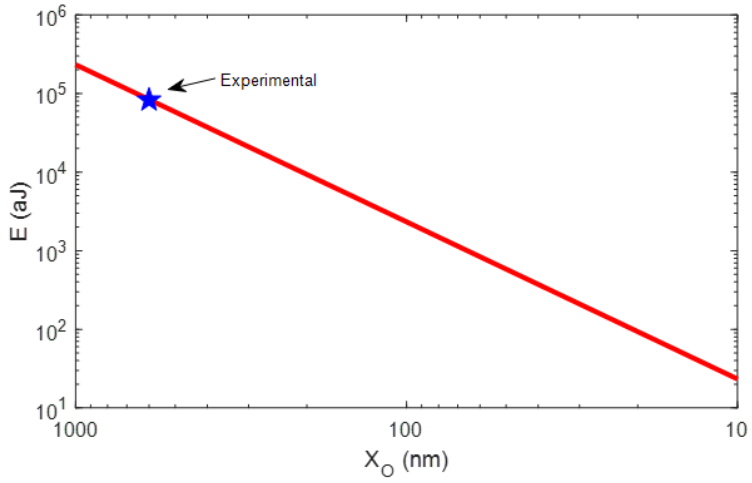


Figure 8.7 - The energy needed to switch the stritch device as a function of the TMD length.

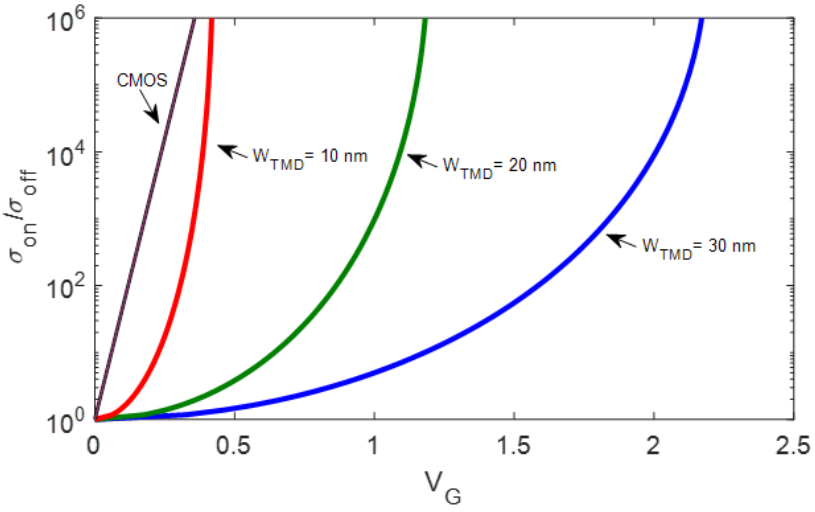


Figure 8.6 - The change in the conductivity versus the applied voltage for three lengths of the TMD. The theoretical subthreshold swing of a CMOS device is also shown as reference.

8.8 SWITCHING ENERGY

Figure 8.7 shows the energy needed to switch the stritch device as a function of the TMD length, x_o. The energy decreases as the length of the TMD is reduced. At x_o=W_{TMD} = 10 nm, only 23 attojoules are needed to switch the device. The energy is comparable to the energy needed for CMOS.

Although, the switching energy is very low, the voltage is not as low as desired. This is a challenge with the stritch device. However, there is a possibility to use other forces within the NEMS to lower the voltage when the distance between the gate and source is smaller than approximately 15 nm.

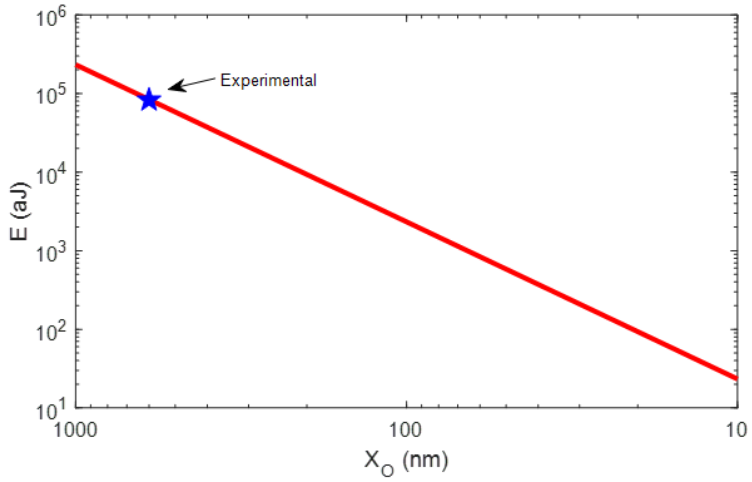


Figure 8.7 - The energy needed to switch the stritch device as a function of the TMD length.

CHAPTER 9

CHAPTER 9: Relay-based Integrated Circuits

Rawan Naous and Vladimir Stojanovic

Department of Electrical Engineering and Computer Science, University of California, Berkeley

9.1 Introduction

The widespread use of smart devices, cell phones, sensors, and electronics [9.1] have triggered the need for advancement in electronics design. Innovations have been introduced at different levels – from the top system level down to the architecture, circuit, and device level. However, these innovations all hold the common goal of integrating memory and processing in the same medium, allowing for faster operation, reduced area, and further energy savings.

In this chapter, system level design with nano-electro-mechanical (NEM) relays is investigated. System design principles are described to take advantage of the intriguing features of NEM relays, including zero leakage and sharp sub-threshold switching. Applications in logic, arithmetic and advanced computing architectures of data searching are explored. The savings achieved with respect to the conventional CMOS and emerging alternatives are addressed, highlighting the potential improvements attained with NEM-based systems.

9.2 Logic Design Considerations

Conventional integrated circuit design relies on the complementary behavior of NMOS and PMOS transistors, which are the nanoscale devices that comprise the basic building blocks of CMOS circuit elements. In this regard, circuit design with NEM relays can be analogous to that of CMOS transistors, since the 4-terminal NEM relay can also be configured to behave in a complementary manner by setting the body voltage accordingly [9.2]. As shown in Figure 9.1, connecting the body to ground allows the device to operate as an N-relay, while connecting the body to V_{dd} converts the operation into a P-relay.

Aside from correct functionality, any novel integrated circuit alternative must also consider other metrics, primarily the ***switching delay***, ***area***, and ***energy*** of the underlying functional block. The following subsections illustrate the challenges faced with system-level design of this novel device, and how innovative design paradigms can overcome the limitations to establish a set of competitive circuit applications.

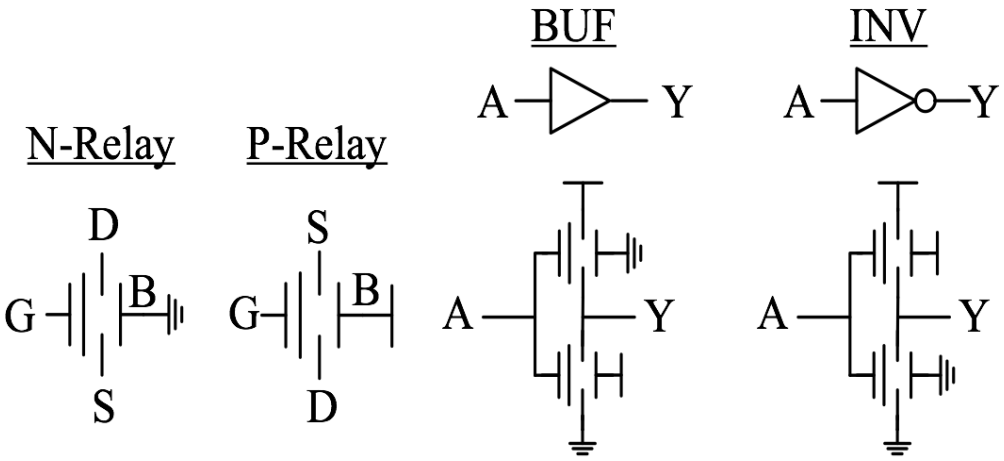


Figure 9.1 - MEM Relay as a logic element [9.5].

9.2.1 Delay

The delay of a particular logic circuit is primarily governed by the overall equivalent resistance (R) and capacitance (C) of the underlying gates [9.3, 9.4]. The electric time constant or circuit delay (τ) for N stacked devices is calculated by the quadratic Elmore delay ($\tau_{RC} = \sum_i^N R_i C_i$). Hence, in CMOS-based designs, buffers are usually inserted between logic stages in order to have simpler blocks with shorter transistor chains and consequently lower delay.

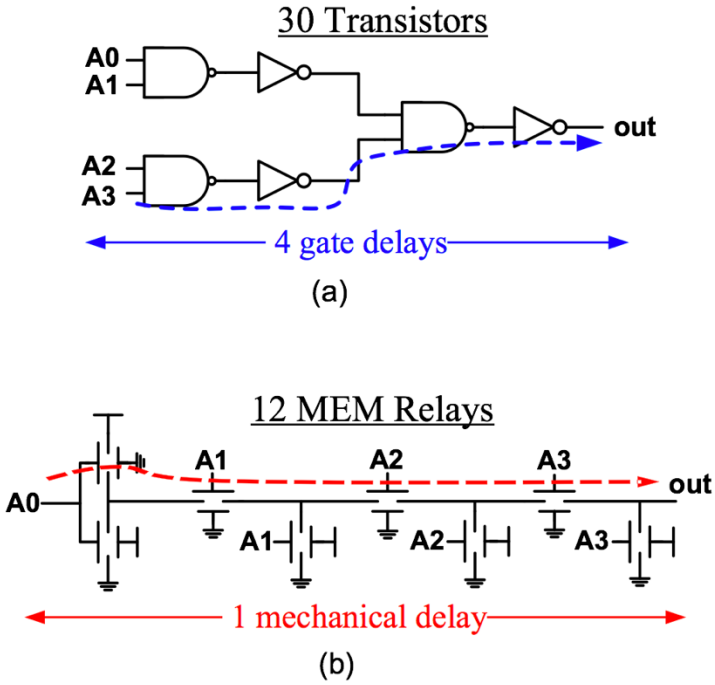


Figure 9.2 - Comparison between (a) CMOS and (b) NEM relay logic design and delay. The CMOS delay is governed by the electrical/gate delay, whereas the NEM design is dominated by the mechanical delay [9.5].

On the other hand, with NEMS relays, the dominant factor is actually the mechanical delay, that is, the time needed to move the beam and completing the switching operation. Thus, circuit design for NEM-based relays is improved by using pass-logic, where all relays are simultaneously switched and thus incur only a single mechanical delay for the complete block. The way to implement such a model is to restrict the connection of the body and gate terminals of the NEM devices to either V_{dd} or *Gnd*. Figure 9.2a and 9.2b illustrate the difference in the design process and delay for the CMOS and NEM circuits respectively [9.5].

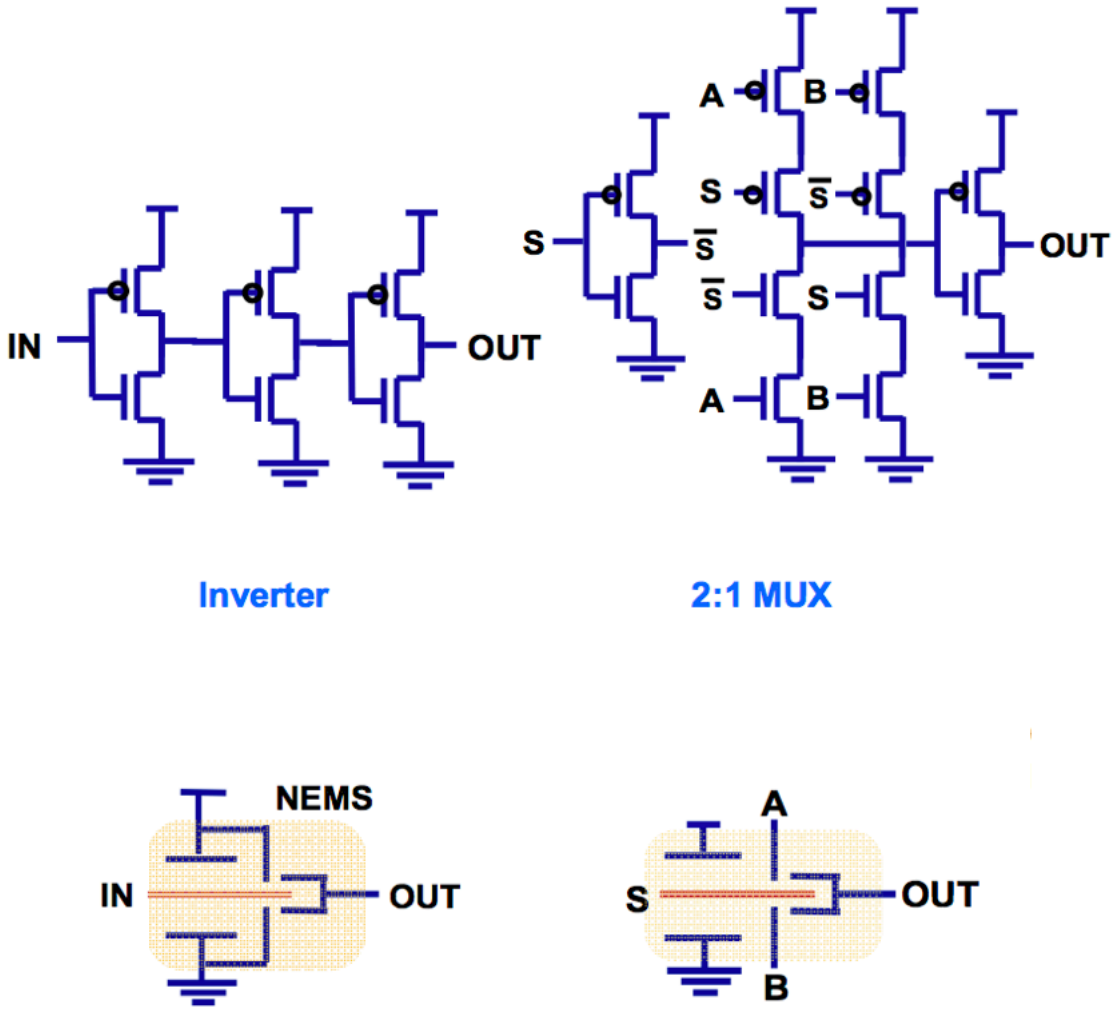


Figure 9.3 - Inverter and Multiplexer, a circuit that selects from a set of inputs and forwards it to the output, circuits in CMOS (top row) and NEM relays respectively (bottom row). The structure of the NEM device allows for a more compact design for these logic blocks [9.7].

9.2. 2 Area

At a first glance, the dimensions of a stand-alone NEM relay is considered quite large compared to its CMOS counterpart. However, the comparison among the different alternatives needs to be performed at the block level rather than the device itself [9.6]. Since NEM-based logic does not

directly map to the same design in CMOS, the number of NEM relays used to realize a particular logic function is much smaller than the number of transistors for the same functionality. Figure 9.3 illustrates the underlying components of an inverter and a mux with CMOS and NEM designs respectively [9.7]. As shown, the number of devices is drastically reduced with the use of NEM switches, particularly with the Back-end-of-the-line (BEOL) NEM model in which the NEM device are integrated within the standard CMOS fabrication process.

9.2.3 Energy

The overall energy (E_{Total}) consumed in a system is divided into two main components: ***dynamic energy*** ($E_{Dynamic}$) and ***leakage energy*** ($E_{Leakage}$).

$$E_{Total} = E_{Dynamic} + E_{Leakage}$$

The dynamic energy is the energy consumed during the operation of the circuit, mainly for the charging/discharging of a load capacitance. The dynamic energy is calculated as

$$E_{Dynamic} = \alpha C_L V_{dd}$$

with α being the switching factor of the circuit, C_L as the load capacitance, and V_{dd} as the input voltage. On the other hand, the leakage energy is the amount of sub-threshold leakage and it is represented as

$$E_{Leakage} = \alpha K C_L V_{dd}^2 e^{-\frac{V_{dd}}{nV_T}}$$

with K being the Boltzmann's constant, V_T being the thermal voltage, and *e* being Euler's number. To that end, leveraging the main characteristics of the NEM relays in having a zero-leakage current and a steep sub-threshold switching allows for an energy-efficient design, since V_{dd} can be reduced. However, the level of energy savings attained is contingent upon optimizing the NEM design space according to the principles illustrated earlier, where savings up to 10x are achieved in digital applications [9.8].

9.3 Arithmetic Applications

Beyond basic logic operations, NEM relays can be integrated into more complex circuit architectures that are tailored towards leveraging the characteristic features of the underlying device. The main blocks in arithmetic operations, such as adder and multiplier, are illustrated to highlight the potential advantage of incorporating NEM relays into digital architectures. Further illustration on the basic operation principles of logic design and circuits is available on the following link from Khan Academy. Circuits & Logic

9.3.1 Relay-Based Adder

The function of an adder is to calculate the Sum (S) and Carry (C) of two input binary numbers (A, B). For each digit, or bit, of the input, the adder structure has both a sum and carry generation block. The per bit formulas are

$$S_i = A_i \; XOR \; B_i \; XOR \; C_i$$
$$C_{out} = C_{i+1} = Majority\ (A_i, B_i, C_i)$$

where A_i , B_i are the input bits and C_i the input carry bit to the corresponding 1-bit Full adder. The design for the carry is crucial to the performance of the overall adder as each of the carry bits is dependent on the availability of the previous bits. Hence, sub-blocks of generate, propagate, and kill functionality are usually used in the carry generation to optimize the operation. Generate (G) = A AND B is set to 1 if both input signals are 1. Propagate (P) = A XOR B propagates the carry in (C_{in}) to the output carry (C_{out}) in case any of the input bits (A or B) are set to 1. The Kill (K) = !A AND !B is set to 1 when both inputs are 0 thus the carry out is set to 0. These signals are quite beneficial in the design of faster and more energy efficient adders. The design of a NEM-relay-based adder builds on the same concepts while incorporating pass-gate-logic and a parallelized design with single mechanical delays.

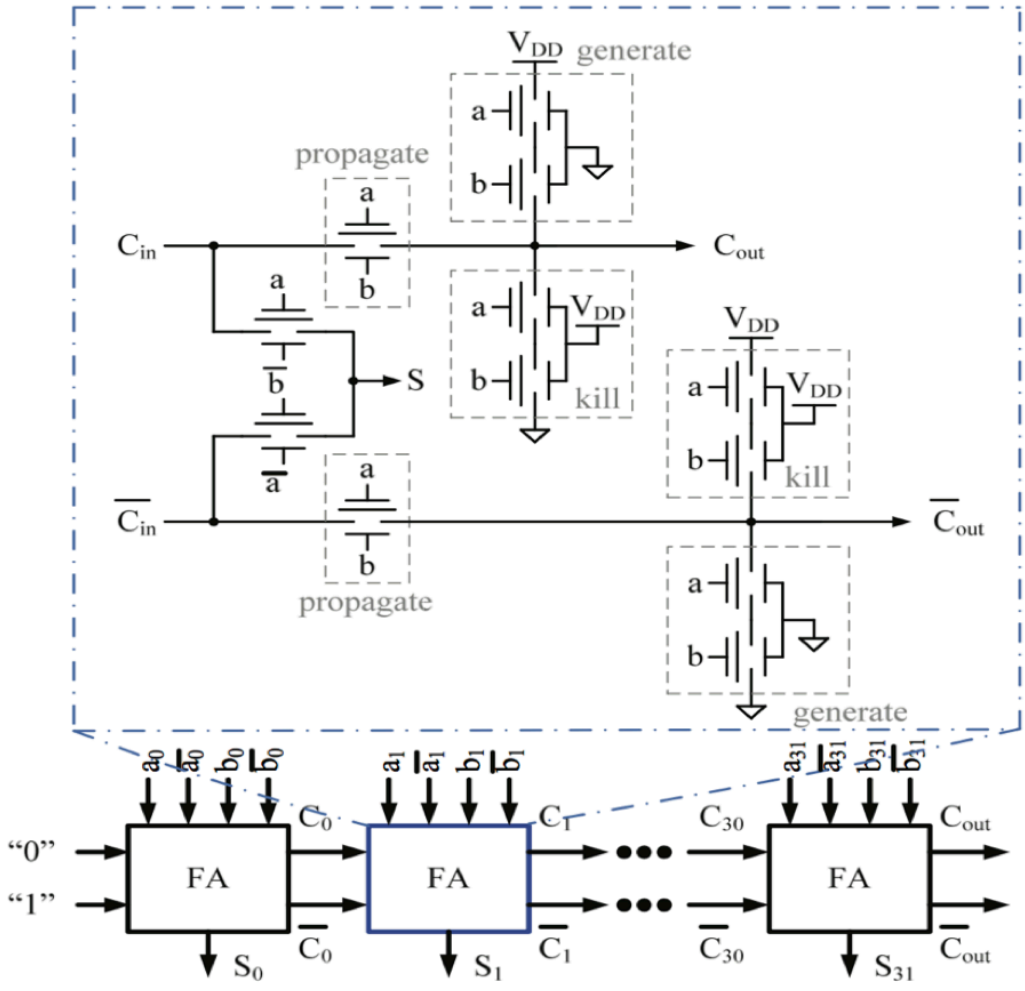


Figure 9.4 - NEM relay-based Full adder cell as part of a 32-bit Manchester carry chain adder [9.3, 9.5].

Figure 9.5 shows the energy per operation E_{op} and delay t_{op} tradeoff for the CMOS and NEM-based adders. With the same area for both adders, the NEM-based adder offers around 10x more energy efficient operation with lower device capacitance and supply voltage but incurs around 9x larger delay. This characteristic allows the NEMs adder to be more optimized towards low-latency applications.

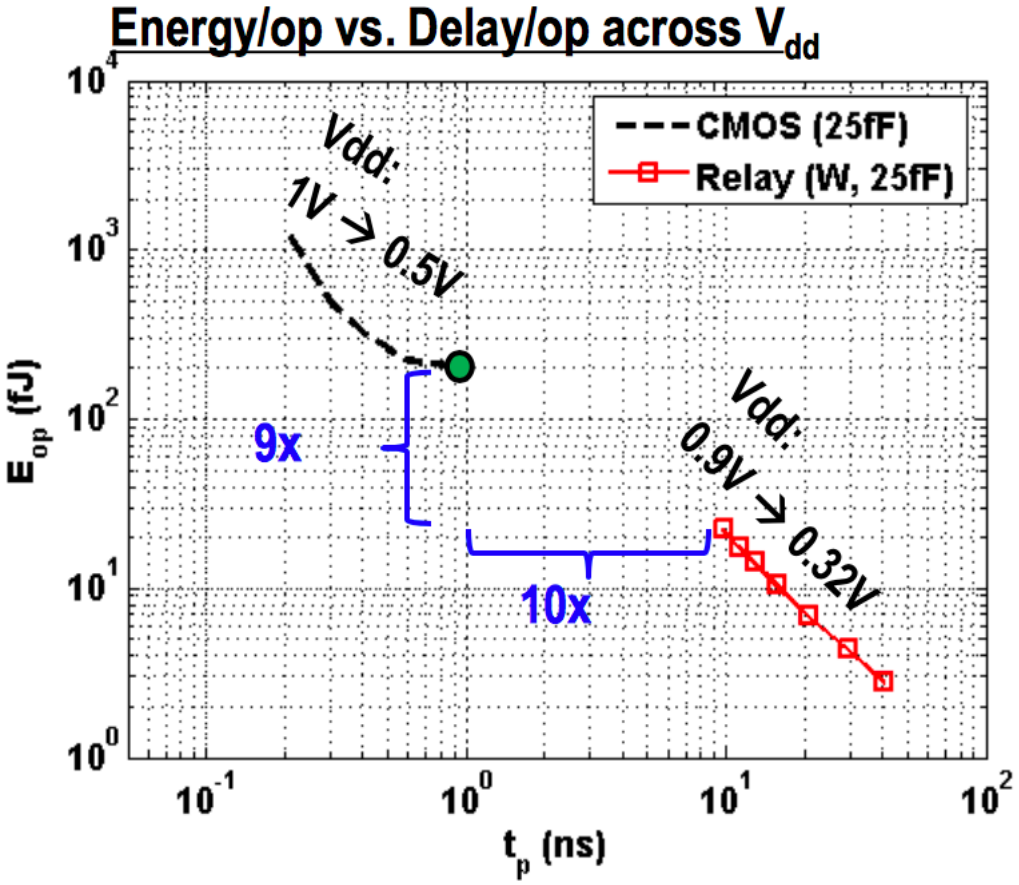


Figure 9.5 - Energy-delay tradeoff for CMOS and NEM relay adders with scaling of the supply voltage. The NEM-based adder achieved 10x more energy efficiency but at an expense of 9x larger delay in comparison to the CMOS adder [9.3].

The energy efficient NEM relay adder could be extended to perform Giga operations per second (GOP/s) by implementing multiple adders in parallel, thus reducing the delay of operation to a nanosecond. The load capacitance together with the ON resistance (R_{ON}) of the NEM device also plays a role in setting the delay of the overall addition operation. Fig 9.6 shows the role of these factors by plotting two different load capacitances (25fF and 100 fF) along with two designs for the NEM relays that incorporate tungsten (W) or gold (Au) as the contact material. The change in the contact mainly affects the R_{ON} of the NEM switch. As depicted in Figure 9.6, the low ON resistance in the gold-coated NEM devices has negligible impact on the delay of the adder. However, with larger R_{ON}, the delay starts increasing, as shown in the case of the 100fF capacitance.

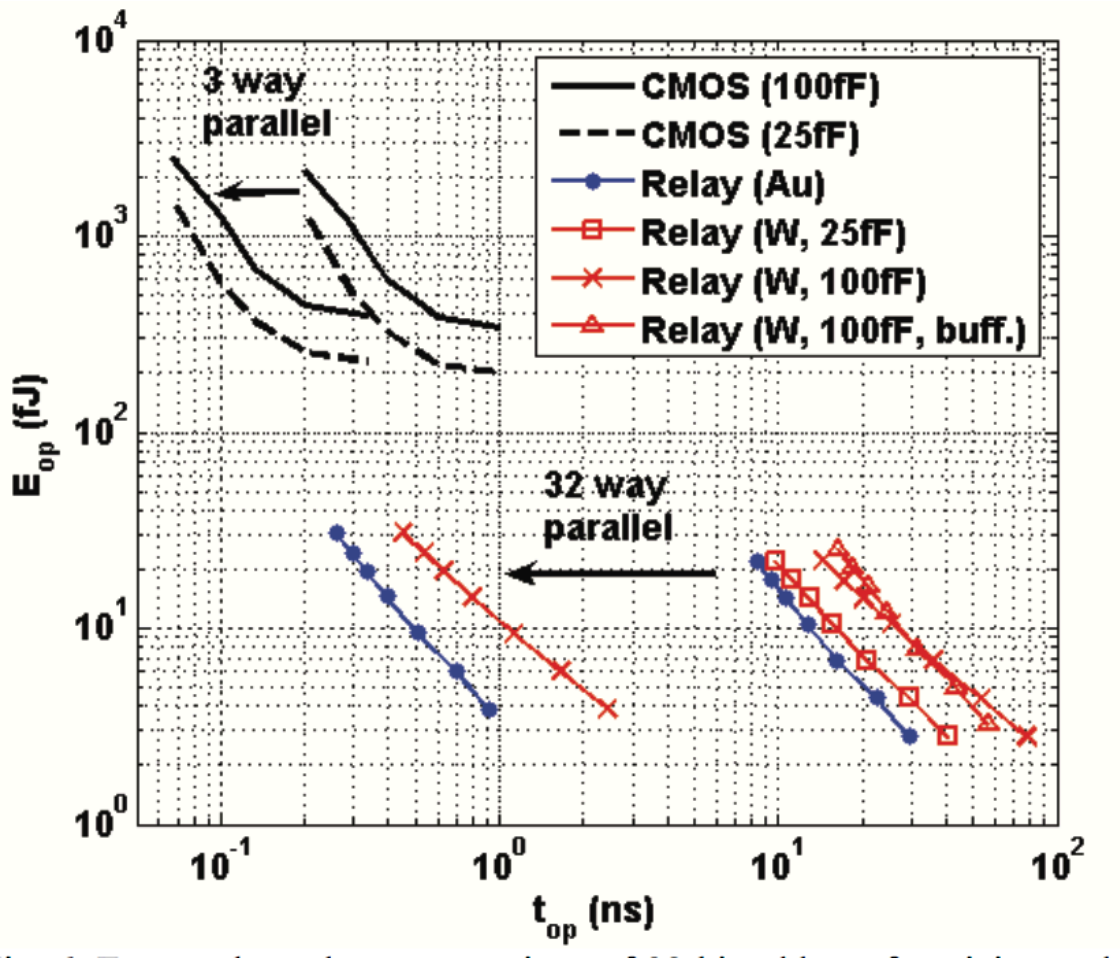


Figure 9.6 - The impact of added parallelism on the throughput of the NEMs adder with a reduced delay allowing the adder to operate in GOP/s regime. The effect of the load capacitance and the ON resistance is also depicted, as the gold material has a low ON resistance and no additional delay, while the higher R_{ON} of the tungsten coating imposes a larger delay [9.3].

9.3.2 Relay-based Multiplier

The multiplier is considered the most complex arithmetic circuit [9.10]. The design flow for the NEM relay multiplier follows the same principles as that of the NEM relay adder discussed earlier. The optimization of the circuit structure is focused towards minimizing the delay as much as possible by designing larger compressor circuits with a single mechanical delay and minimal number of gates. Hence, the area of innovation is in the design of the partial product generation matrix that is then fed to the corresponding addition blocks. Figure 9.7 illustrates an example of a 6-bit NEM relay-based multiplier using half and full adder blocks.

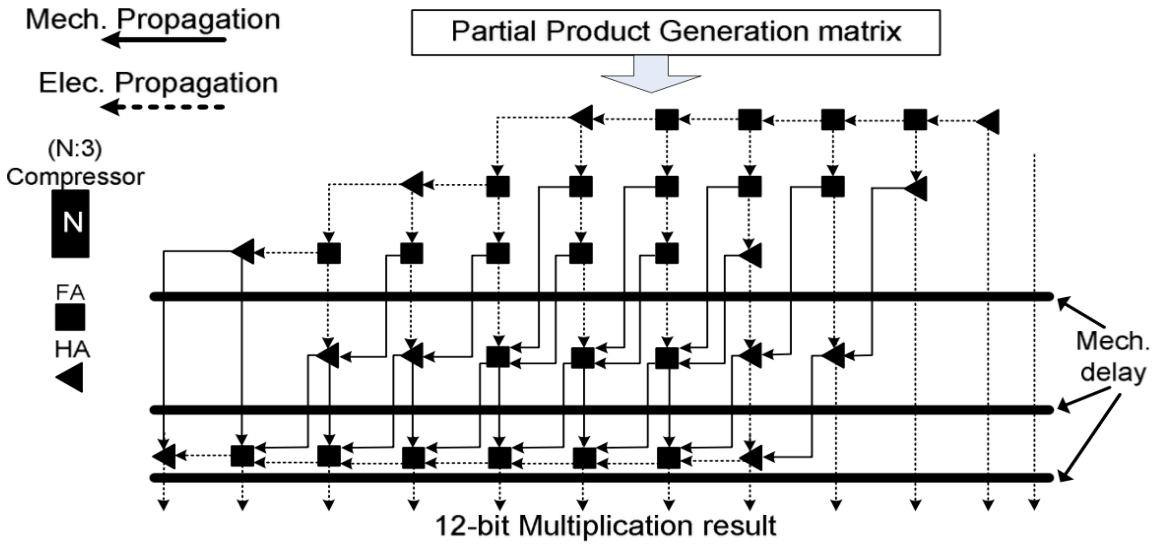


Figure 9.7 - A 6-bit NEM relay multiplier that is built using half and full adder blocks [9.10].

9.3.2.1 Partial Product Generation

Typically, when multiplying two numbers, the partial products are generated by a series of shift operations. For each column in the multiplier, a corresponding partial product is generated by first shifting the value of the multiplicand to the left by a number of columns. and then multiplying it with the value of the bit in the multiplicand column. The overall product uses half and full adder blocks in the generation of the output results. The half adder takes the two input operands and provides the sum and carry outputs. The full adder does the function of a half adder, and also accommodates the carry input to provide two outputs (the sum and carry) for the three-input operation, acting as a 3:2 compressor. The operation is illustrated in Figure 9.8. In this format, the number of partial products generated is the same as the number of bits in the multiplier which results in a large delay.

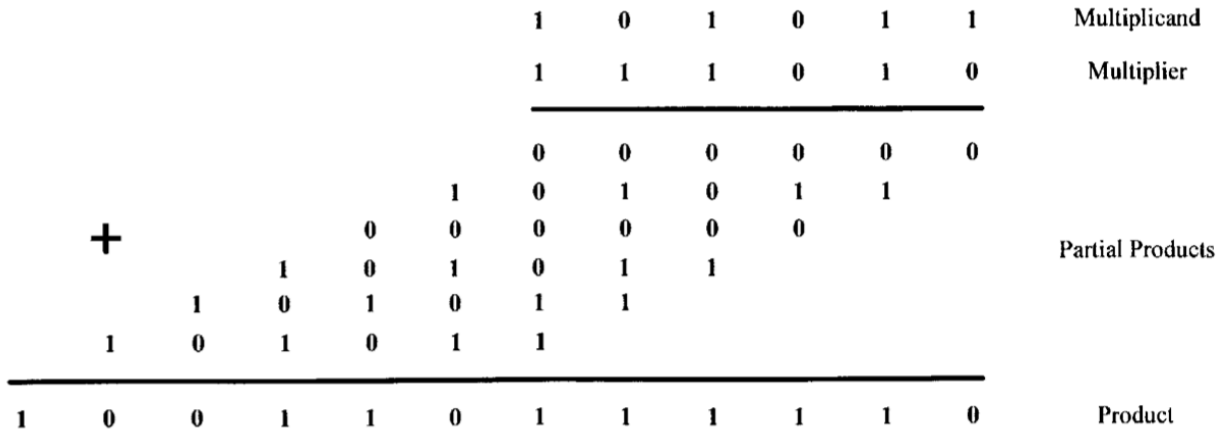
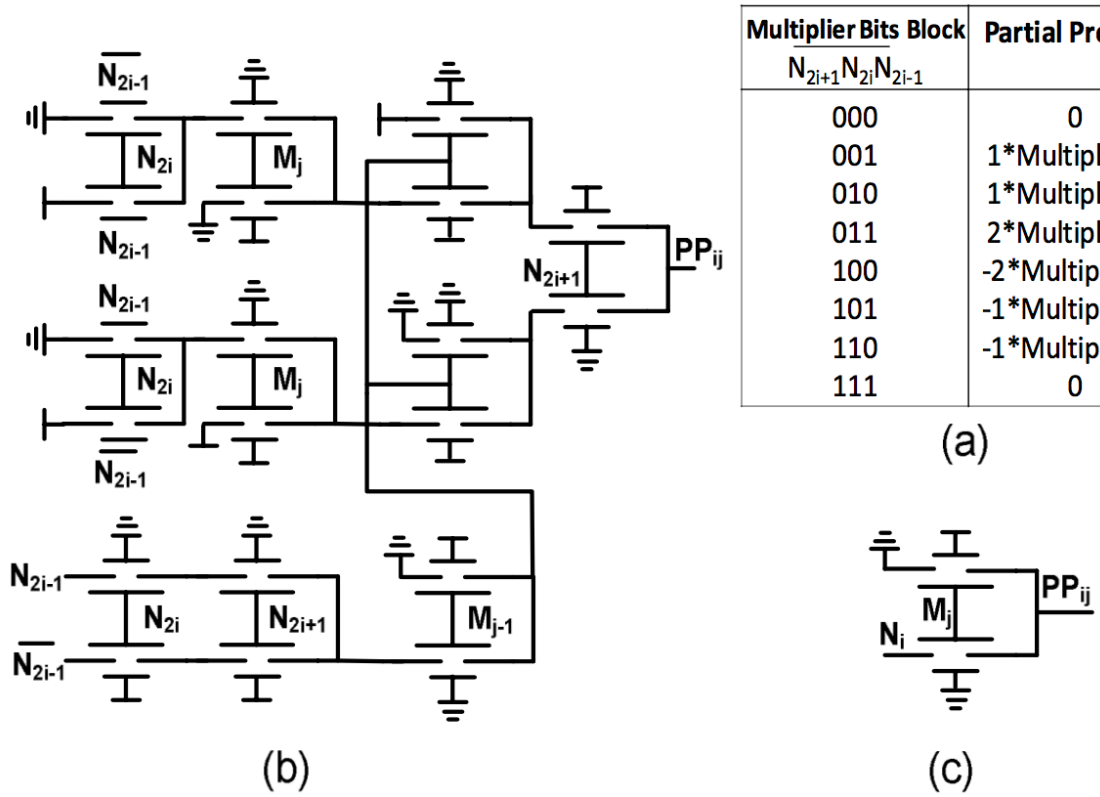


Figure 9.8 - Partial product generation in the standard format.

Thus, a simplifying technique is used in order to reduce the number of partial products to halve that of the standard approach. A redox 4 Booth encoding is adopted, where every other column in the multiplier is used and the multiplication is done by $\pm 1, \pm 2, or\ 0$ to obtain the same result. Figure 9.9 shows the Booth encoding table along with the NEM relay-based circuit for the partial product generation circuit.



Multiplier Bits Block $N_{2i+1}N_{2i}N_{2i-1}$	Partial Product
000	0
001	1*Multiplicand
010	1*Multiplicand
011	2*Multiplicand
100	-2*Multiplicand
101	-1*Multiplicand
110	-1*Multiplicand
111	0

(a)

Figure 9.9 - (a) Booth encoding table. (b) Partial Product generation circuit for Booth encoded generation, (c) Simple AND gate partial product generation [9.10].

9.3.2.2 Energy/Delay Estimation

The NEM relay-based multiplier design was assessed in terms of energy and delay and compared to its CMOS alternative. Two optimized 16-bit CMOS multipliers are used for the comparison. The energy and delay outputs for all three multipliers are shown in Figure 9.10 for supply voltage range from 1.4 – 0.7 V. As depicted, the NEM relay-based multiplier reaches 10x better energy efficiency than both CMOS multipliers. With respect to the delay, by increasing the area of the NEM relay multiplier with applied parallelism and allowing the 16 multiplications to occur simultaneously, a throughput of around 100 Mega Operations Per Second (MOP/s) is achievable while preserving energy efficiency.

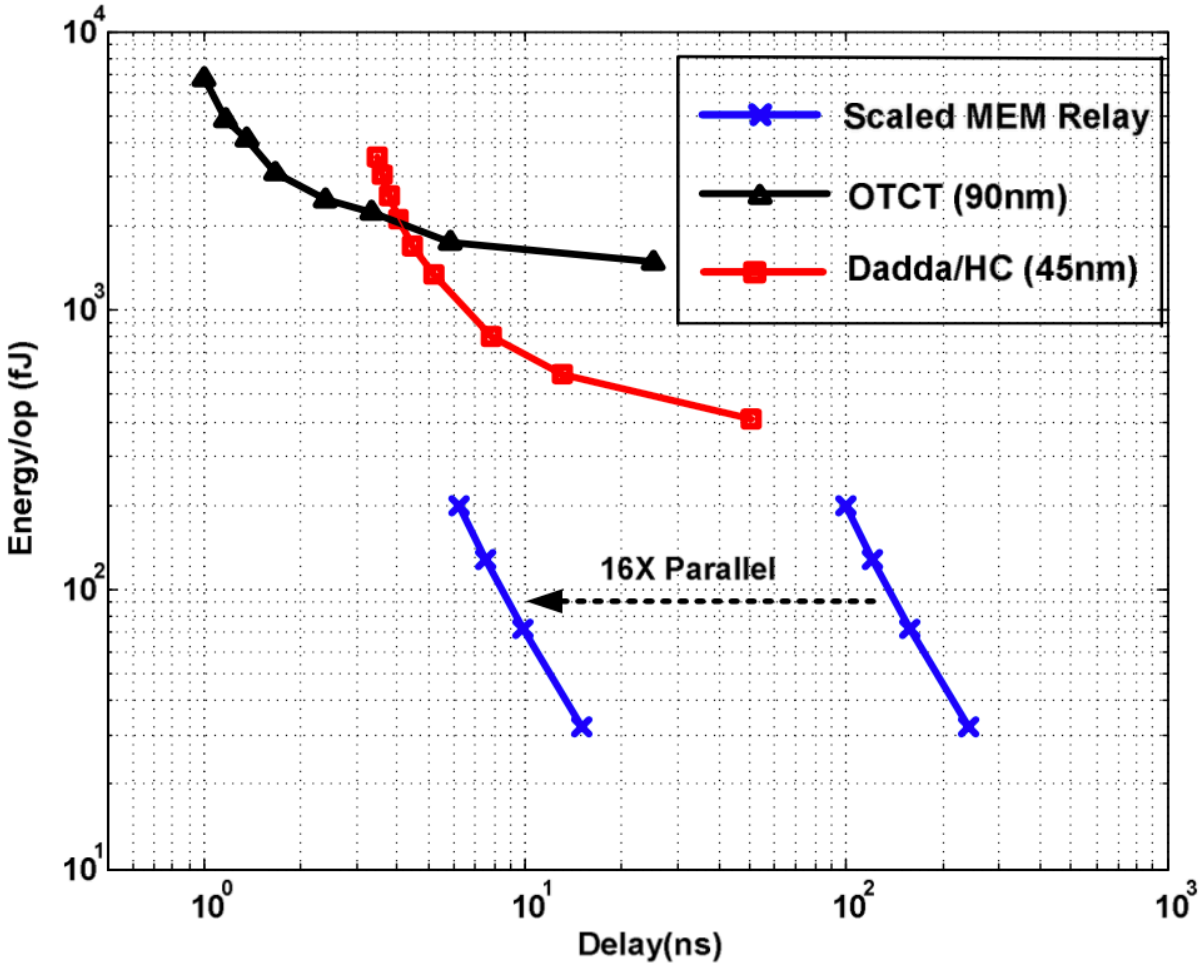


Figure 9.10 - Energy-Delay comparison for two CMOS multiplier designs (OTCT and Dadda/HC) and NEM-relay 16-bit multipliers.

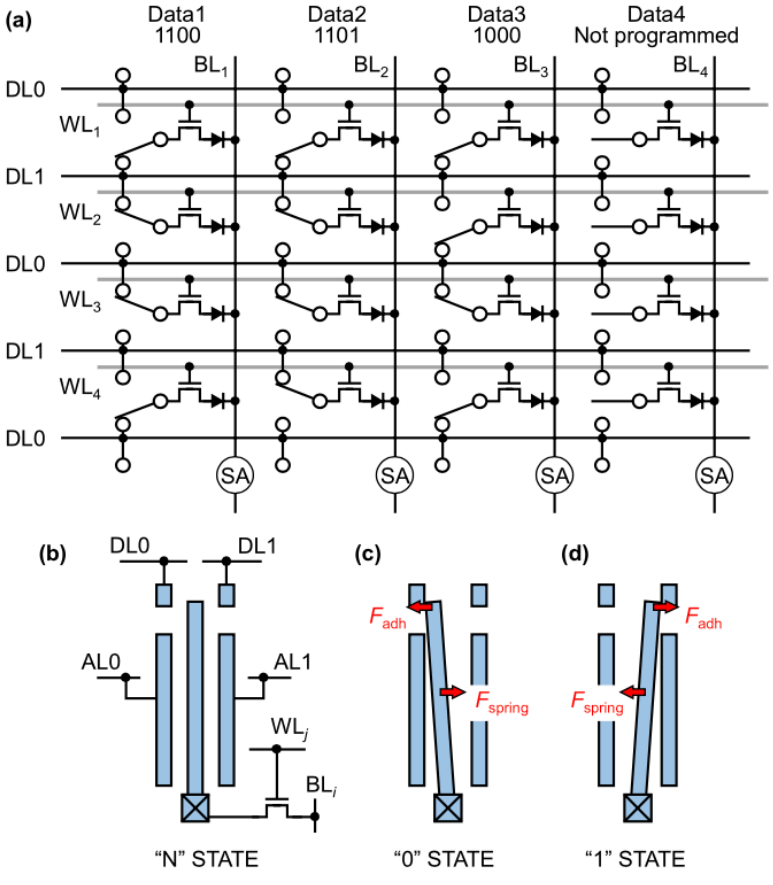
9.4 Advanced Computing Architectures

Big data analytics and machine learning are computation-intensive applications that impose further constrains on the design of VLSI circuits, especially if deployed on hand-held devices. For efficient operation of the overall computer system, the energy and delay requirements for the underlying circuits become crucial. In this domain, the non-volatility and the low energy of the NEM relays are leveraged to build new computing systems for **data searching** and **reconfigurable computing** [9.11].

9.4.1 Data Searching

Data in computers is represented in bits, where each **bit** is a unit of memory, which represents either a "1" or "0". More complex data, for example larger numbers or letters, can be represented using several bits (often 32 or 64 bits) combined together, i.e. a **word**.

In conventional memory, the data is stored into cells with word line and bit line addresses used to fetch a specific data from the memory. In Fig 9.11, each word is stored in a separate row, with a word line to access a specific row, and a bit line to access a bit on the row. A data search operation will pattern-match for a particular bit pattern within the memory. To do this, the entire memory space needs to be screened in order to find a match [9.12]. Traditionally, a combination of a central processing unit (CPU) chip and a memory chip, such as dynamic access memory (DRAM), is used to perform the operation. An alternative approach to data searching is presented in [9.13] where the non-volatility of the NEM relays allows for a high-speed and energy-efficient solution, fabricated on a single chip. Figure 9.11a shows the corresponding memory structure with embedded NEM relays and an access transistor per cell. The 6T NEM device allows for a fast-parallel readout operation for all the NEM devices in this crossbar structure.



(e)

Operation	Input voltage						Output current	
	AL0	AL1	DL0	DL1	BL_i	WL_j	"0" state	"1" state
Program "0"	V_{prog}	Float	Float	Float	0	V_{dd}	N/A	N/A
Program "1"	Float	V_{prog}	Float	Float	0	V_{dd}	N/A	N/A
Store	X	X	X	X	X	X	N/A	N/A
Match "0"	X	X	Float	V_{read}	SA	V_{dd}	zero	> 0
Match "1"	X	X	V_{read}	Float	SA	V_{dd}	> 0	zero

SA: Sense amplifier, X: Any state acceptable

Figure 9.11 - a) Circuit Diagram of the Non-volatile Memory with NEM relays with one access transistor per cell. [9.13] (b-d) The structure and operation of the NEM relay in the ON and OFF states [9.14]. (e) The input voltage and output current for the different operating conditions of the memory.

9.4.1.1 Memory Design and Operation

The NEM relay-based memory cell [9.13] depicted in Figure 9.11b is composed of two actuation terminals (AL0 and AL1). The application of a programming voltage at one of the terminals moves the electrode to be in contact with one of the data lines (DL0 or DL1) to save the value '0' or '1' in the NEM device, respectively. For example, in order to program a cell (j,i) to '0' , the actuation line AL0 will be set to V_{prog} , the word line j (WL_j) will be set to V_{dd} in order to turn on the access transistor, and the bit line i (BL_i) will be set to 0V. With this setup, the electrode becomes in contact with the DL0 allowing the data bit '0' to be saved in the memory (Figure 9.11c). Alternatively, to save a value of '1', the actuation line AL1 is activated with V_{prog} instead of AL0 while all other terminals remain intact (Figure 9.11d).

In this structure, the data search process requires only two read operations, regardless of the size of the array. The first step looks for a match in the columns that contains a '0'. In this step, all the data lines DL1 are set to the read voltage (V_{read}) and the word lines that contain a 0 bit in the corresponding search string are set to 0V. In case a particular cell contains '1', a current will pass through the bit line and a mismatch occurs. Alternatively, in case of a mismatch, a zero current will pass through the bit line and the sense amplifier connected at the terminal of the bit line will sense the match condition. The second search step includes looking for a '1' match. In a similar manner to the '0' match, the DL0 lines will be set to V_{read} and the word lines corresponding to '1' in the search string will be set to 0V. The sense amplifiers will detect the match or mismatch conditions. A perfect match occurs if a match is found on a particular bit line in both read operations. The details of the programing and the match sequences are highlighted in Figure 9.11e with respect to the input voltage setting and the expected output current.

9.4.1.2 Data Searching Architecture Performance Metrics

The energy and delay performance of the array can be measured from the programming (write) and search (read) operations. When writing into the array, no current flows into the NEM relay as the data lines are floating. The programming energy is estimated to be around 50 aJ for a programming voltage of 2V and a delay of less than 10 ns. With respect to the searching operation, the delay is that of two reads reaching around 0.1 ps. The energy consumption varies according to the size of the array as shown in Table 9.1. The NEM relay-based memory array with access transistors offer a highly efficient and fast solution to data search operation. This feature is apparent in comparison to the current approach where conventional CPU and memory chips would require around 90 mJ and 80 ms to find the address of a match on an 8 Gb chip. However, the NEM relay-based chip would only need 300nJ and less than 0.5 ns to perform the same task [9.11].

	Energy			Delay
Cells involved:	**1 column × 1 row**	**1 column × 256 rows**	**256 columns × 256 rows**	
Program (V_{prog} = 2.5 V)	**15 fJ**	**2.0 pJ**	**N/A**	**< 10 ns**
Match "0" or Match "1"	**N/A**	**N/A**	**1.2 pJ**	**< 0.2 ns**

Table 9.1 - Energy-Delay for 256x256 NEM array for Data Search.

9.4.2 Reconfigurable Computing

Conventional computing architectures transfer data and instructions between the memory chip and the central processing unit (CPU) chip. This requires a lot of energy and causes delay penalties on the operation of overall system. In the move towards more optimized edge devices with longer battery life and added processing capabilities, a paradigm shift is thus required in the computational execution. The concept of **in-memory computing** emerges**,** where the memory and CPU are integrated on a single chip. Reconfigurable logic and memory indexing using NEM devices can thus replace the actual computation. A Lookup table design with NEM relays is described further below.

A lookup table (LUT) is a functional mapping of a set of input bits to a particular set of outputs. The input bits are considered to be an address to a specific entry in the table that saves the desired value. In other words, instead of doing computations, the results are pre-calculated and saved in the lookup table, which is much more cost-effective in terms of energy and delay. Conventionally, CMOS-based LUTs require a non-volatile storage, a decoder, an SRAM, and multiplexer to provide the output. An innovative design with emerging non-volatile resistive memory is demonstrated in [9.15] that is based on a crossbar structure. An alternative to resistive memory is proposed in [9.16] leveraging the low programming energy of the NEM relay devices along with high speed readout potential to attain an enhanced energy and delay performance over the afore proposed designs. Figure 9.12 shows the circuit diagrams for the corresponding CMOS-based LUT and the NEM memory (NEMory) LUT design.

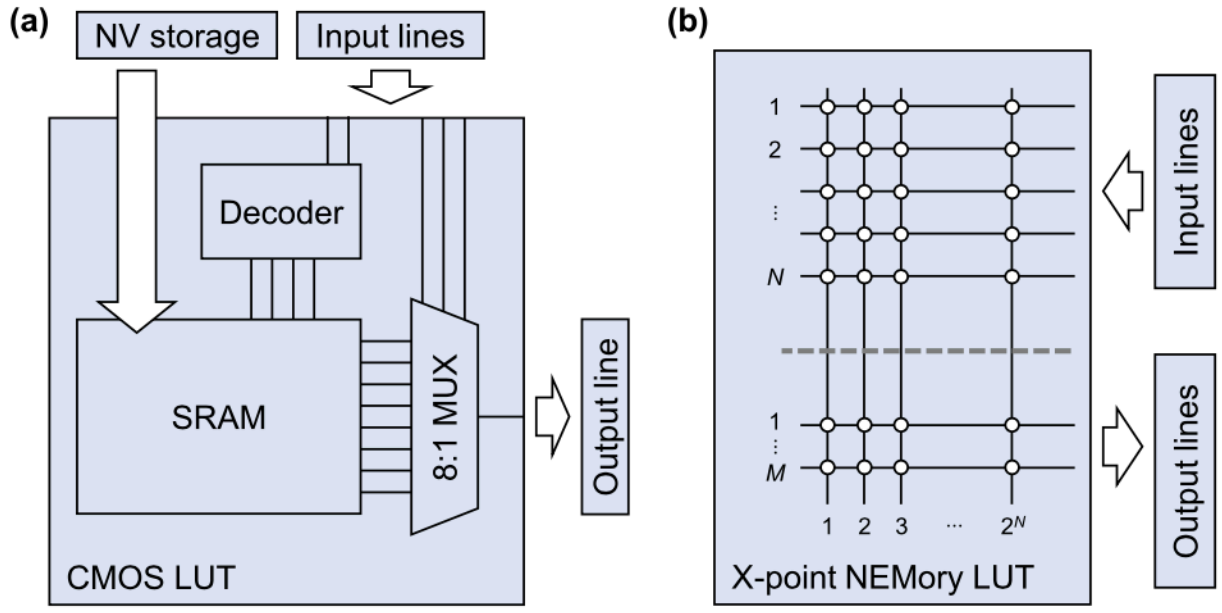


Figure 9.12 - Lookup table circuit structures (a) CMOS and (b) NEM relay-based.

The operation of the NEMory-based LUT utilizes a crossbar structure with N inputs and M outputs, leading to N+M columns array. The inputs are the memory addresses whereas the outputs are the desired results. The total number of rows in the array is proportional to the size of the inputs 2^N.

Figure 9.13a and 9.13b shows the circuit details of a 5:2 NEMory-based LUT and the truth table for a (5:2) LUT example.

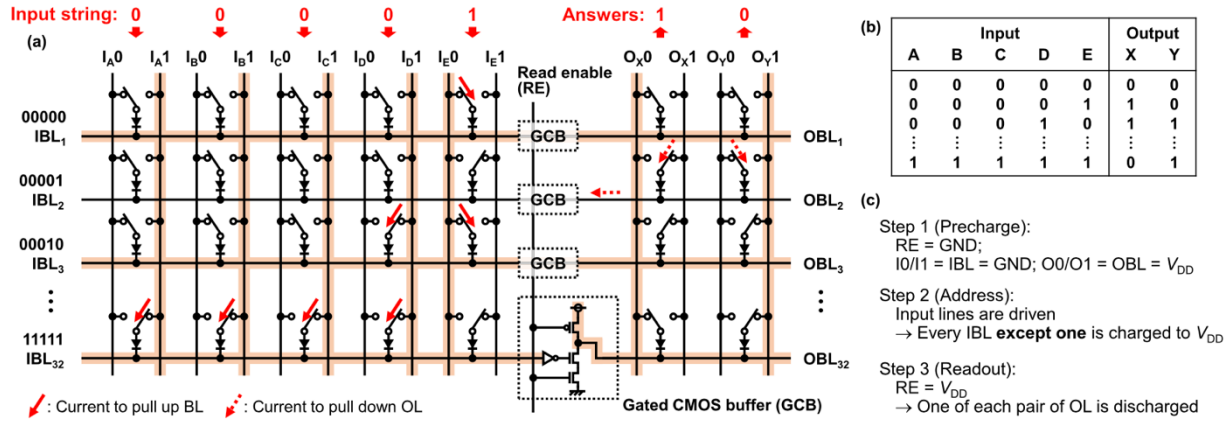


Figure 9.13 - a) circuit structure for the NEMory-based LUT, (b) the corresponding truth table for the (5:2) LUT example illustrated. (c) The steps applied for the readout operation [9.16]. The highlighted lines on the crossbar structure correspond to the driven lines on the input and output portion.

Programming the crossbar array is performed one row at a time with grounding of a particular input/output bit line (IBL/OBL) and setting a programming voltage (V_{prog}) on the actuation terminals PL1/PL0 for setting the NEM device at a particular column to 1 or 0 respectively.

Reading out a value from the LUT requires a three-step process:

Step 1: The Read enable line (RE) and the input side (IBL_i and I_0/I_1) are pre-charged to ground (GND), and the output side (OBL_i and O_0/O_1) are set to high voltage V_{dd}.

Step 2: The input columns (I_0/I_1) are driven to a high voltage setting then all the bit lines except for the ones that corresponds to the input string

Step 3: The read enable (RE) is set to high voltage (V_{dd}), activating then the gated CMOS buffer (GCB), and driving the output lines that correspond to the input string to GND. Hence, the following logic can detect the correct output bits.

The performance of the NEMory-based LUT is quantified with respect to area, delay, and energy in comparison with CMOS and resistive RAM (ReRAM) alternatives. Figure 9.14 shows a radar plot for the corresponding analysis. As depicted, the readout energy and delay for the NEMory based LUT are orders of magnitude more energy efficient than their alternative implementations.

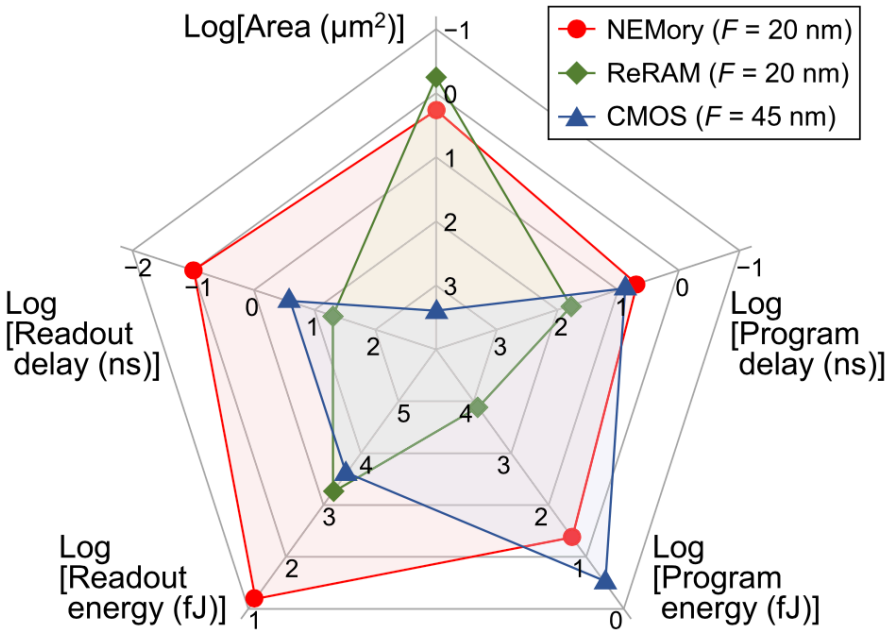


Figure 9.14 - Radar plot comparing the performance metrics of area, energy and delay for the NEMory-based LUT, CMOS LUT and the ReRAM in-memory computing scheme [9.16].

9.5 Conclusion

In electronics, the system level metrics are directly impacted by the underlying characteristics of the physical devices. A study of the design requirements and considerations is presented with systems based on NEM relays. Enhancements are achieved in comparison to the conventional CMOS alternative, particularly in terms of energy and delay. Designs in logic, arithmetic and advanced computing architectures are presented with applications suitable for the ever-increasing trend of having more compact and energy-efficient end devices.

REFERENCES

[1] ITRS Technology Roadmap for Semiconductors 2015. [Online]. Available: http://www.itrs2.net

[2] F. Chen, H. Kam, D. Markovic, T. King Liu, V. Stojanovic, and E. Alon. 2008. Integrated circuit design with NEM relays. In Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design (ICCAD '08).

[3] H. Kam, F. Chen, Micro-Relay Technology for Energy-Efficient Integrated Circuits, New York, NY:Springer, 2015.

[4] R. Nathanael, V. Pott, H. Kam, J. Jeon, E. Alon, T.-J. King Liu, "Four-terminal-relay body-biasing schemes for complementary logic circuits", IEEE Electron Device Lett., vol. 31, no. 8, pp. 890-892, Aug. 2010.

[5] F. Chen, "Energy-efficient Wireless Sensors: Fewer Bits More MEMS", Dept. Electr. Eng. Comp. Sc., Massachusetts Inst. Technol., Sep. 2011.

[6] H. Fariborzi, "Design and demonstration of integrated micro-electro-mechanical relay circuits for VLSI applications," PhD thesis, Dept. EE&CS, MIT, Cambridge, MA, USA, Jun. 2013. Available http://dspace.mit.edu/handle/1721.1/82348

[7] N. Xu, J. Sun, I.-R. Chen, L. Hutin, Y. Chen, J. Fujiki, C. Qian, T.-J. K. Liu, "Hybrid CMOS/BEOL-NEMS technology for ultra-low-power IC applications", Proc. IEEE Int. Electron Devices Meeting (IEDM), pp. 28.8.1-28.8.4, Dec. 2014.
[8] H. Kam, T.-J. King Liu, V. Stojanovic, D. Markovic, 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, Jan. 2011.
[9] D. Patil et. al., "Robust Energy-Efficient Adder Topologies," in Proc. 18th IEEE Symp. on Computer Arithmetic (ARITH'07).
[10] H. Fariborzi, F. Chen, R. Nathanael, J. Jeon, T.-J. K. Liu, V. Stojanovic, "Design and demonstration of micro-electro-mechanical relay multipliers", Proc. Asian Solid-State Circuits Conf., 2011-Nov.
[11] T.-J. King Liu, U. Sikder, K. Kato, V. Stojanovic, "There's Plenty of Room at the Top", IEEE MEMS, 2017.
[12] K. Pagiamtzis, A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: A tutorial and survey", IEEE J. Solid-State Circuits, vol. 41, no. 3, pp. 712-727, Mar. 2006.
[13] K. Kato, V. Stojanovic, T.-J. K. Liu, "Non-volatile nano-electro-mechanical memory for energy-efficient data searching", IEEE Electron Device Lett., vol. 37, no. 1, pp. 31-34, Jan. 2016.
[14] W. Y. Choi, H. Kam, D. Lee, J. Lai, T.-J. K. Liu, "Compact nano-electro-mechanical non-volatile memory (NEMory) for 3D integration", Proc. IEEE Int. Electron Devices Meeting (IEDM), pp. 603-606, Dec. 2007.
[15] S. Paul and S. Bhunia, Computing with Memory for Energy-Efficient Robust Systems. New York, NY, USA: Springer, 2014.
[16] K. Kato, V. Stojanović and T.-J. K. Liu, "Embedded nano-electro-mechanical memory for energy-efficient reconfigurable look-up tables," IEEE Electron Device Letters, vol. 37, no. 12, pp. 1563-1565, 2016.
[17] S. Han , X. Liu , H. Mao , J. Pu , A. Pedram , M. Horowitz , W. Dally, EIE: efficient inference engine on compressed deep neural network, Proceedings of the 43rd International Symposium on Computer Architecture, June 18-22, 2016, Seoul, Republic of Korea
[18] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," CoRR, vol. abs/1704.04760, pp. 1–17, Apr. 2017.
[19] M. Rastegari, V. Ordonez, J. Redmon, A. Farhadi "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," arXiv:1603.05279
[20] M. Horowitz, "Computing's energy problem (and what we can do about it)", ISSCC Digest of Technical Papers, pp. 10-14, Feb. 2014.