Design, Optimization, and Scaling of MEM Relays for Ultra-Low-Power Digital Logic

Hei Kam, *Member, IEEE*, Tsu-Jae King Liu, *Fellow, IEEE*, Vladimir Stojanović, *Member, IEEE*, Dejan Marković, *Member, IEEE*, and Elad Alon, *Member, IEEE*

***Abstract*—Microelectromechanical relays have recently been proposed for ultra-low-power digital logic because their nearly ideal switching behavior can potentially enable reductions in supply voltage ($V_{\rm dd}$) and, hence, energy per operation beyond the limits of MOSFETs. Using a calibrated analytical model, a sensitivity-based energy–delay optimization approach is developed in order to establish simple relay design guidelines. It is found that, at the optimal design point, every 2× energy increase can be traded off for a ∼1.5× reduction in relay delay. A contact-gap-to-actuation-gap thickness ratio of 0.7–0.8 is shown to result in the most energy-efficient relay operation, implying that pull-in operation is preferred for an energy-efficient relay design. Based on the analytical model and design guidelines, a scaling theory for relays is presented. A scaled relay technology is projected to provide $>$ 10× energy savings over an equivalent MOSFET technology, for circuits operating at clock frequencies up to ∼100 MHz.**

***Index Terms*—Digital integrated circuits, logic devices, low power circuit, microelectromechanical systems, microswitches, subthreshold slope, 60 mV/dec.**

I. Introduction

OVER the past 40 years, MOSFET feature size scaling has resulted in dramatic improvements in the performance, cost per function, and energy efficiency of integrated circuits. Due to the fact that the OFF-state leakage current ($I_{\rm OFF}$) of a MOSFET increases exponentially with threshold voltage (V_T), V_T can no longer be reduced along with transistor physical dimensions. This has forced the supply voltage ($V_{\rm dd}$) to remain relatively constant across CMOS technology generations, causing power density to increase with transistor density and thus limiting the performance benefits of transistor scaling [1].

The fundamental cause of the CMOS power density issue is that the subthreshold swing of a MOSFET (i.e., the rate at which the transistor turns on/off with increasing/decreasing gate voltage) is directly proportional to the thermal voltage (k_BT/q) which does not scale with transistor dimensions. Therefore, alternative transistor designs that can achieve steeper switching behavior than a MOSFET have been proposed to alleviate this issue [2]–[5]. However, any CMOS or CMOS-like technology will have a lower limit in energy per operation due to $I_{\rm OFF}$ [6]. To overcome this limit, microelectromechanical (MEM) switches have been investigated for digital logic applications [7]–[11] because they ideally offer zero $I_{\rm OFF}$ and perfectly abrupt switching behavior. In principle, then, the threshold voltage of a MEM switch (and therefore $V_{\rm dd}$) can be reduced to be close to 0 V to provide for very low active power consumption.

In terms of device structure and operation, a MEM relay designed for digital logic applications [10] is very similar to relays designed for radio-frequency (RF) signal switching applications [12]–[15]. However, the contact resistance ($R_{\rm on}$) for a logic relay can be as high as 10 kΩ (for a load capacitance of 10–100 fF) because the switching delay of a relay-based circuit is dominated by the mechanical pull-in time ($t_{\rm pi}$, typically 10–100 ns) rather than the electrical charging/discharging delay ($t_{\rm RC}$) [8], [11]. Thus, $R_{\rm on}$ can be traded off for improved endurance, for example, by utilizing TiO_2-coated tungsten contacting electrodes to attain high device yield and endurance $>$ 1 billion switching cycles [10], [11]. This insight enabled the successful demonstration of the first relay-based logic, memory, and clocking integrated circuits [16].

Whereas previous papers mainly focused on relay fabrication process optimization [10], [11] and prototype circuit demonstrations [8], [16], this paper focuses on developing a methodology for an energy–performance optimized relay design. Similarly to MOSFETs, dimensional scaling can be applied to relays to improve device density (for lower cost per function), switching delay (for higher performance), and power consumption (for improved energy efficiency). This paper therefore also examines the implications of scaling relay devices while following the proposed design methodology.

In order to arrive at the proposed methodology, Section II begins this paper by developing a general model for relay logic delay and energy that is calibrated to experimental data. Section III then utilizes this model to establish a sensitivity-based energy-delay optimization methodology as well as a simplified relay device design procedure. Based upon these design guidelines, a scaling theory for relays is presented in Section IV. Conclusions are provided in Section V.

II. Logic Relay Energy–Delay Model

In this section, the principle and analytical models for MEM relay operation are described to provide the necessary

Manuscript received June 29, 2010; revised September 8, 2010; accepted September 23, 2010. Date of publication November 11, 2010; date of current version December 27, 2010. This work was supported in part by the C2S2 and MSD Focus Centers (two of the five research centers funded under the Focus Center Research Program, which is a Semiconductor Research Corporation program) and in part by a DARPA/MTO NEMS program. The work of E. Alon was supported by the Berkeley Wireless Research Center under the National Science Foundation Infrastructure Grant 0403427. The review of this paper was arranged by Editor A. M. Ionescu.

H. Kam was with the University of California at Berkeley, Berkeley, CA 94720-1770 USA. He is now with Intel Corporation, Hillsboro, OR 97124 USA (e-mail: heikam@eecs.berkeley.edu).

T.-J. K. Liu, and E. Alon are with the University of California Berkeley, Berkeley, CA 94720-1770 USA (e-mail: tking@eecs.berkeley.edu; elad@eecs.berkeley.edu).

V. Stojanović is with the Massachusetts Institute of Technology, Cambridge, MA 02139-4307 USA (e-mail: vlada@mit.edu).

D. Marković is with the University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: dejan@ee.ucla.edu).

Digital Object Identifier 10.1109/TED.2010.2082545

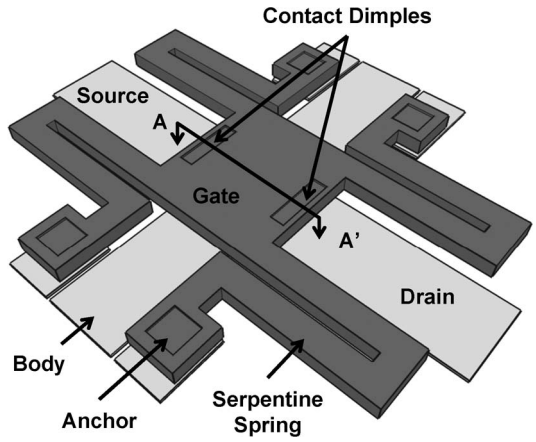

Fig. 1. Schematic 3-D view of the electrostatically actuated 4T relay structure.

background for a subsequent discussion of design optimization and scaling. Fig. 1 shows a schematic 3-D view of an electrostatically actuated four-terminal (4T) relay [10], with key design parameters shown in Fig. 2 and Table I. The voltage applied between the movable gate electrode and the fixed body electrode determines whether current can flow between the source and drain electrodes. The gate electrode is supported by four suspended beams (with an effective spring constant $k_{\rm eff}$) anchored to the substrate at four corners. A folded-flexure beam design is used to reduce the effects of residual/thermal stress and vertical strain gradient. A metallic channel electrode is attached underneath the gate electrode via an intermediary gate dielectric layer.

The position of the gate stack depends on the electric field across the actuation gap between the gate and the body electrode, i.e., the balance between the electrostatic attractive force $(F_{\rm elec})$ and the spring restoring forces of the beams $(F_{\rm spring})$. In the OFF state [Fig. 2(a)], an air gap separates the channel from the coplanar source/drain electrodes so that no current can flow. In the ON state [Fig. 2(b)], the gate stack is "pulled in" by the electrostatic force between the gate and the body so that the channel contacts the source/drain electrodes in the contact dimple regions, allowing current to flow.

The electrical transition between OFF and ON states is abrupt, i.e., current flow between the source and drain electrodes increases abruptly as the gate-to-body voltage $(V_{\rm gb})$ increases above the "pull-in voltage" $(V_{\rm pi})$. Since the electrostatic attractive force is ambipolar, a 4T relay can be turned *on* either by applying a positive gate-to-body voltage (mimicking the operation of an n-channel MOSFET, e.g., with the body biased at the ground) or by applying a negative gate-to-body voltage (mimicking the operation of a p-channel MOSFET, e.g., with the body biased at $V_{\rm dd}$).

The absolute value of the voltage at which a mechanical switch turns *off* is usually smaller than that of the voltage at which it turns *on*, i.e., the I–V characteristics of a mechanical switch exhibit some hysteresis. In the case of a 4T relay, although the device turns *on* when $|V_{\rm gb}|$ increases above $V_{\rm pi}$, it turns *off* when $|V_{\rm gb}|$ is lowered below the "release voltage" $(V_{\rm rl})$. The hysteretic switching behavior $(V_{\rm rl} < V_{\rm pi})$ occurs for "pull-in-mode" operation (wherein $F_{\rm elec}$ remains larger than $F_{\rm spring}$ for values of $V_{\rm rl} < |V_{\rm gb}| < V_{\rm pi}$ when the channel is in contact with the source/drain electrodes) and is exacerbated by surface adhesion force (F_A) in the contact regions. The as-fabricated air gap in the contacting region (g_d) can be made to be smaller than the as-fabricated air gap in the actuation region (g) to reduce the hysteresis voltage. This dimpled contact design is also beneficial to precisely define the area of the contacting region and to reduce the device turn-on delay $(t_{\rm delay})$.

To provide insight into relay operation and guidance for relay scaling, analytical models for relay performance parameters are needed. The most relevant physical properties and equations governing relay behavior such as pull-in voltage, release voltage, and turn-on delay are briefly reviewed here, and measured data[1] for devices with design parameters as shown in Table I are used for model calibration.

A. Switching Voltages

The relay switching voltages can be derived by balancing the electrostatic force against the spring restoring force. As will be derived in the following section, the energy efficiency of a relay is optimized by operating it in pull-in mode, i.e., designing it such that $g_d > g/3$. In this case, hysteretic switching behavior will occur, and the hysteresis is exacerbated by surface adhesion forces. For pull-in-mode operation, the switching voltages are given by [13], [15]

$$V_{\rm pi} = \sqrt{\frac{8k_{\rm eff}g^3}{27\varepsilon_o A}} \qquad V_{\rm rl} = \sqrt{\frac{2(k_{\rm eff}g_d - F_A)(g-g_d)^2}{\varepsilon_o A}} \quad (1)$$

where $k_{\rm eff}$ is the effective spring constant of the movable structure and A is the actuation area ($\approx L_A \times W_A$, ignoring release etch holes). Note that, in (1), non-ideal effects such as fringing electric fields are assumed to be negligible.

In general, $k_{\rm eff}$ decreases with increasing flexure beam length (L). Thus, $V_{\rm pi}$ and $V_{\rm rl}$ can be adjusted via a lithographic mask design, and different values of $V_{\rm pi}$ and $V_{\rm rl}$ can be achieved for different relays on a single chip. This tunability allows the circuit designer to make direct tradeoffs between layout area, circuit operating speed, and energy efficiency. As was discussed in [11], the beams exhibit both bending and rotational motions when the movable plate is actuated downward. A precise $k_{\rm eff}$ model that accounts for shear displacement and rotational inertia is complex [17]; by sacrificing some degree of accuracy, $k_{\rm eff}$ can be rendered into a more intuitive form consisting of flexural $(\propto 1/L^3)$ and torsional $(\propto 1/L)$ [18] terms

$$\frac{1}{k_{\rm eff}} \cong \left(\gamma_f \frac{EWh^3}{L^3}\right)^{-1} + \left(\gamma_t \frac{GWh^3}{L}\right)^{-1} \quad (2)$$

where γ_f and γ_t are the flexural and torsional constants, respectively, both of which are roughly fixed for a given relay technology. Using ANSYS, γ_f and γ_t are found to be 3.66 and

[1]4T relays have been demonstrated using conventional surface micromachining processes [10]. However, the first prototype 4T relay design reported in [10] suffers from gate dielectric charging issues, leading to unwanted parasitic actuation effects as well as undesirable influence on gate switching voltages and pull-in time. While research toward resolving this issue is on-going, for the purpose of developing an accurate analytical model, it is sufficient to use data for a simpler 3T relay variant (wherein there is no gate dielectric layer, the movable plate together with the attached metallic electrode serves as the source electrode, the underlying electrode serves as the gate electrode, and the coplanar electrodes serve as drain electrodes [11]) to calibrate the analytical models for pull-in voltage, release voltage, and turn-on delay $(t_{\rm delay})$. This is because the mechanical design and models of energy and performance for 4T and 3T relays are essentially identical.

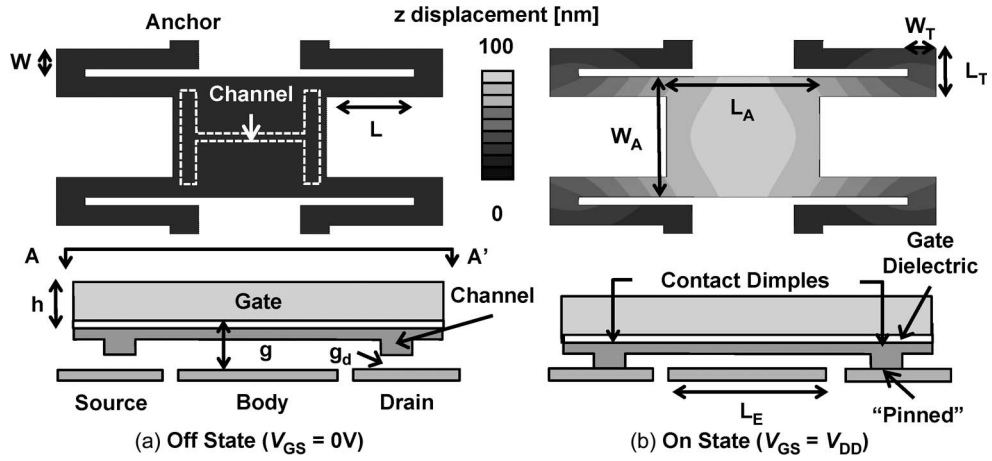

Fig. 2. ANSYS-simulated displacement contours and schematic cross-sectional views of the 4T relay structure in the (a) OFF state ($V_{\rm GS} = 0$ V) and (b) ON state ($V_{\rm GS} = V_{\rm DD}$).

TABLE I
DESIGN PARAMETERS FOR THE FABRICATED RELAYS

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

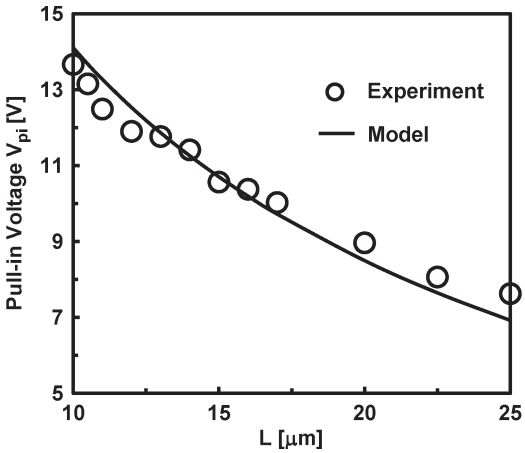

Fig. 3. Measured relay $V_{\rm pi}$ versus L. The analytical model matches the measured data to within 10%.

1.341×10^{10} m^{-2}, respectively. With these values and (2), (1) yields $V_{\rm pi}$ values within 10% of the measured data (Fig. 3).

It is important to note that there is an upper bound for the beam length to ensure that the relay can be turned off. This length corresponds to the requirement that $F_{\rm spring}$ is larger than F_A in the ON state. F_A can be extracted from the measurements of $V_{\rm pi}$ and $V_{\rm rl}$ for devices of various beam lengths, according to the following equation [derived from (1)]:

$$V_{\rm rl}^2 = \frac{27}{4}\frac{g_d}{g}\left(1-\frac{g_d}{g}\right)^2 V_{\rm pi}^2 - \frac{2(g-g_d)^2}{\varepsilon_o A}F_A. \qquad (3)$$

F_A is extracted to be 0.45 μN on average for TiO_2-coated tungsten electrodes with a contact dimple area (A_d) of 2×10 μm^2, as shown in Fig. 4. This is within 2× of the ∼0.2-μN F_A value obtained via atomic force microscopy measurements [19].

It is also important to note that there is an upper bound for $V_{\rm dd}$, not only to avoid gate dielectric breakdown in 4T relays, but also to avoid the undesirable pull-in of the movable plate to the underlying fixed electrode. For the relay structure considered in this work, the lower and upper bounds of this catastrophic pull-in voltage ($V_{\rm cpi}$) can be derived (as shown in Appendix I) by modeling the movable plate as a beam anchored at the two contact dimple regions

$$V_{\rm cpi} \cong \zeta\sqrt{\frac{Eh^3(g-g_d)^3}{\varepsilon_o L_A^4}} \qquad (4)$$

where ζ lies within the range from 1.516 to 3.444. For the switches used in this work, (4) predicts a $V_{\rm cpi}$ value in the range

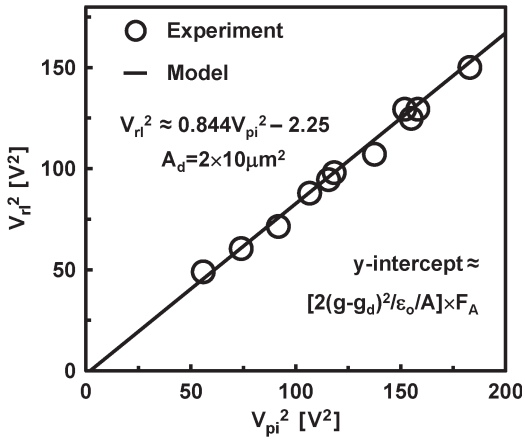

Fig. 4. Measured $V_{\rm rl}^2$ versus $V_{\rm pi}^2$. F_A is extracted to be 0.45 μN.

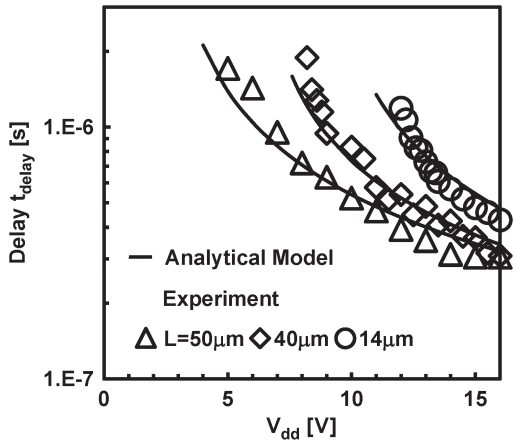

Fig. 5. Measured $t_{\rm delay}$ versus $V_{\rm dd}$ for three different relays.

from 14.6 to 33.2 V, which properly bounds the experimentally measured value of ∼23 V.

B. Turn-On Delay

When an actuation voltage (V) is applied between a movable electrode and a fixed electrode, the motion of the movable electrode is governed by Newton's second law of motion, yielding the following second-order differential equation [20]:

$$m_{\rm eff}\ddot{z} + \frac{\sqrt{k_{\rm eff}m_{\rm eff}}}{Q}\dot{z} + k_{\rm eff}z = \frac{\varepsilon_o A V^2}{2(g-z)^2}. \qquad (5)$$

The right-hand side of the equation represents the electrostatic force, Q is the quality factor, and z is the displacement. Note that non-ideal effects such as contact bounce and settling time are assumed to be negligible. To provide insight for relay design, $t_{\rm delay}$ can be approximated in closed form by (derived in Appendix II)

$$t_{\rm delay} \cong \alpha\sqrt{\frac{m_{\rm eff}}{k_{\rm eff}}}\left(\frac{g_d}{g}\right)^{\gamma}\left(\frac{V_{\rm dd}}{V_{\rm pi}}-\chi\right)^{-\beta}, \quad \text{for } 5V_{\rm pi} \geq V_{\rm dd} > 1.1V_{\rm pi};\quad g_d \geq g/3 \qquad (6)$$

where $\chi \cong 0.8$ and $m_{\rm eff}$ is the effective mass of transport. This effective mass consists of the mass of the movable electrode and the loaded mass of the springs; it can be determined from the total kinetic energy in the relay $(KE_{\rm tot})$ and the velocity of the actuation plate (v_p) [11]

$$m_{\rm eff} = \frac{KE_{\rm tot}}{0.5v_p^2} = \frac{m_p v_p^2 + \int v_b^2 dm_b}{v_p^2} = \alpha_o\rho Ah + \alpha_1\rho WLh \qquad (7)$$

where $\alpha_o = 1.8$ and $\alpha_1 = 2.74$, and m_b and v_b are the mass and velocity of the folded beams, respectively.

Equation (6) shows good agreement with the measured pull-in time, as shown in Fig. 5. As expected intuitively, $t_{\rm delay}$ decreases with increasing resonant frequency $(\sqrt{k_{\rm eff}/m_{\rm eff}})$ and "gate overdrive" $(V_{\rm dd}/V_{\rm pi})$. $t_{\rm delay}$ is also dependent on the parameters α, β, and γ, which, to first order, depend only on Q. These parameters (α, β, and γ) are obtained numerically and are shown in Fig. 6.

It is important to note that β represents the sensitivity of the delay to $V_{\rm dd}$, and as indicated in Fig. 6, β decreases with increasing Q. This is because the relay switching speed becomes mass-transport-limited as Q increases. For RF relays which employ micrometer-scale actuation gaps, Q is limited by squeeze-film damping. In contrast, for scaled logic relays with actuation gaps approaching 10 nm, i.e., less than the mean free path of an air molecule, squeeze-film damping will be negligible. Rather, the quality factor of nanometer-scale logic relays will be dominated by surface-related energy-loss mechanisms, due to their relatively large surface-to-volume ratio. It has been shown that Q decreases linearly with beam thickness and is independent of the beam width and beam length [21], [22]. Therefore, it is reasonable to treat Q as a technology-dependent constant.

It is also important to note that since α, β, and γ saturate quickly for Q values greater than ∼1, Q values significantly higher than 1 do not significantly improve relay performance. In fact, it may be preferable to avoid high-Q relay designs to minimize non-ideal switching effects such as contact bounce and long settling time. Thus, additional processing/manufacturing steps to support vacuum packaging may not be necessary. Subsequently, in this paper, relays with $Q \geq 1$ will be referred to as "high-Q relays," while relays with $Q < 1$ will be referred to as "low-Q relays."

C. Switching Energy

The energy consumed in switching a relay is supplied by the voltage source which charges/discharges capacitances. These include the capacitance associated with the actuation air gap, $\varepsilon_o A/(g-g_d)$, and the fixed parasitic capacitance C. In addition, there will be an actuation-area-dependent extrinsic capacitance (e.g., due to wire routing). Assuming an area proportionality factor r and an average relay activity factor a, the switching energy of a relay can be modeled by the following equation[2]:

$$E_{\rm switch} \cong a \times \left[\frac{\varepsilon A}{g-g_d}(1+r) + C\right] V_{\rm dd}^2. \qquad (8)$$

[2]Note that the capacitance associated with the wires used to interconnect a relay-based circuit would likely scale with $\sqrt{A}$ rather than A. Furthermore, parasitic interconnect capacitance would likely increase with beam length L. However, using the linear dependence on A greatly simplifies the calculations, and the overall findings are relatively unaffected by these simplifications.

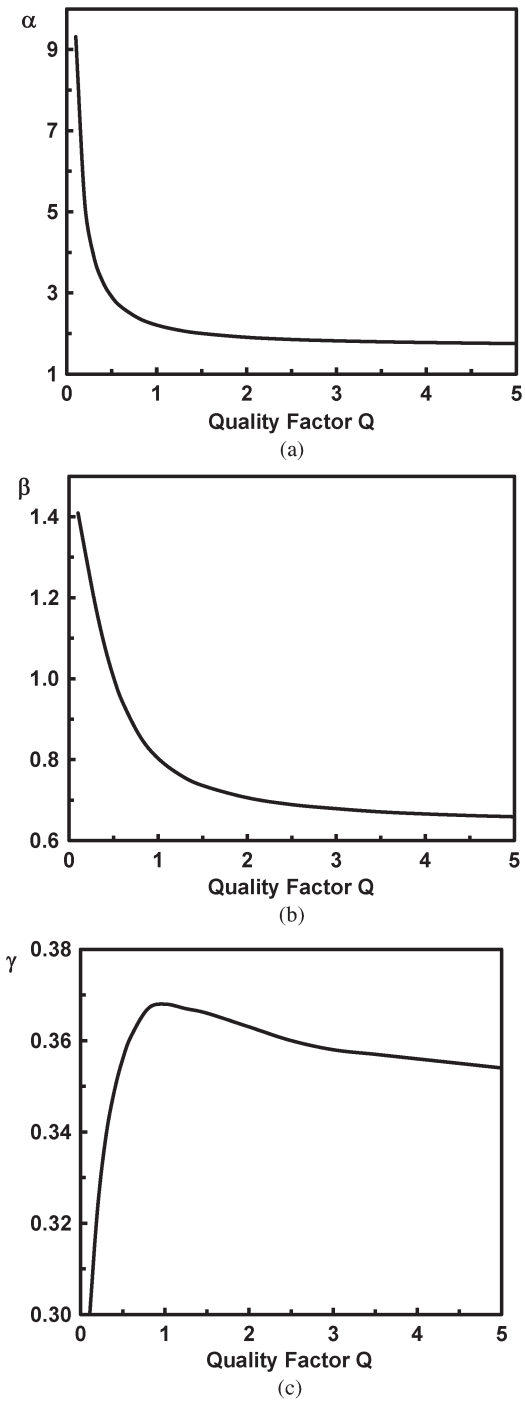

Fig. 6. (a) Dependence of α on the quality factor. (b) Dependence of β on the quality factor. (c) Dependence of γ on the quality factor.

By using the calibrated delay and energy models, the switching voltages, delay, and energy for relays with various design parameters can be predicted. These, in turn, can be used for relay circuit design optimization. For most relay designs, the beam width W will be the minimum feature size set by photolithography limitations. The switching delay and energy are always minimized by utilizing the smallest achievable beam thickness h and contact dimple gap thickness g_d, and hence these two design parameters will be set by process technology constraints. Once the beam thickness is set, the quality factor Q [21], [22] and therefore the α, β, and γ values are known.

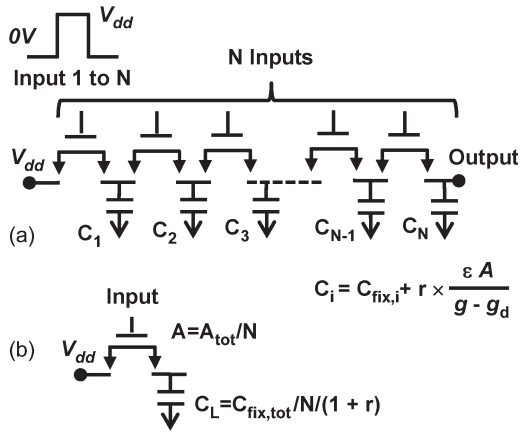

Fig. 7. (a) Optimized circuit topology for a relay-based logic gate [8]. For simplicity, only the pull-up network is drawn here. Load capacitance C_i consists of fixed capacitance $C_{\mathrm{fix},i}$ and an area-dependent load capacitance. (b) Energy-delay optimization problem for a relay-based circuit can be simplified into optimizing a single relay driving an average fixed capacitance.

Therefore, the supply voltage, actuation area, as-fabricated gap thickness, and the beam length are the remaining variables for design optimization.

III. Relay Energy–Delay Optimization

Since the switching delay of a relay is dominated by its mechanical delay rather than its electrical delay, an optimized relay-based circuit design should utilize complex gates [8] such that all the relays move simultaneously and only one mechanical delay is incurred per operation. Thus, relay circuits with stacked devices are modeled herein, as shown in Fig. 7.

In optimizing a relay circuit design, it is important to note that, as for a CMOS circuit design [23]–[28], energy and delay are traded off by adjusting various device design parameters. The energy-delay tradeoff is optimized essentially by solving the following constrained optimization problem for the N-relay complex gate:

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

$$\text{Subject to}: \quad E_{\mathrm{tot}} = \sum_{i=1}^{N} a_i \left(\frac{\varepsilon A_i}{g - g_d}(1 + r) + C_i\right) V_{\mathrm{dd}}^2 \quad (9)$$

where E_{tot} is the total energy and A_i, a_i, and C_i are the area, average activity factor, and the load capacitance of the ith relay in the complex gate, respectively. In general, all relays have the same g, g_d, and r. Furthermore, if we generalize the relay circuit with an average[3] activity factor a, the total energy consumed by the circuit can be expressed as

$$E_{\mathrm{tot}} = a V_{\mathrm{dd}}^2 \left(\frac{\varepsilon A_{\mathrm{tot}}}{g - g_d}(1 + r) + C_{\mathrm{fix,tot}}\right) \quad (10)$$

where A_{tot} and $C_{\mathrm{fix,tot}}$ are total actuation area and the total fixed capacitance, respectively. Dividing both sides of (10) by $a \cdot N \cdot (1 + r)$, one finds that the energy consumption of relay circuit can be represented by a generalized relay with

[3]In a complex gate, not all the nodes (i.e., capacitances) switch with equal probability; therefore we can generalize the circuit to have an average activity.

an average activity factor a and actuation area A driving an average fixed capacitance C_L

$$E_s = \frac{E_{\rm tot}}{aN(1+r)} = \left(\frac{\varepsilon A}{g-g_d} + C_L\right) V_{\rm dd}^2,$$
$$A = A_{\rm tot}/N; \quad C_L = C_{\rm fix,tot}/\left(N\times(1+r)\right). \quad (11)$$

Note that (11) is very similar to (8) for the energy dissipated by a single relay, i.e., the energy-delay optimization problem is equivalent to optimizing the energy-delay of a generalized relay driving an average fixed capacitive load. To solve this constrained optimization problem, a sensitivity-based analysis is performed next to explore the relay energy-delay tradeoff.

A. Sensitivity Analysis

CMOS integrated circuit optimization utilizing energy-delay sensitivity analysis has been described extensively in [23]–[28], so only the key concepts are briefly reviewed here. The energy-delay sensitivity to a design variable var is defined as

$$S_{\rm var} \equiv \frac{\partial t_{\rm delay}/\partial var}{\partial E_s/\partial var} \quad (12)$$

which is interpreted as the delay reduction per energy cost by adjusting the value of the variable var. As mentioned previously, the supply voltage $(V_{\rm dd})$, actuation area (A), as-fabricated gap thickness (g), and the beam length (L) are the variables available for design optimization. The optimal relay design is reached when the sensitivities to all tuning variables are balanced [23]–[28].[4] The analytical formulas for the sensitivities to each of these variables and their implications on relay design are derived below, using the prototype relay design in [10] and [11] as a baseline: The nominal values for the actuation area, as-fabricated gap thickness, and beam length are 15×30 μm^2, 200 nm, and 40 μm, respectively, yielding a nominal pull-in voltage of 4.15 V. For a nominal supply voltage of 5 V and an average fixed capacitance per relay of 50 fF, the nominal delay and switching energy are 2.86 μs (at $Q = 0.5$) and 2 pJ, respectively.

1) Sensitivity to Supply Voltage $(V_{\rm dd})$*:* As $V_{\rm dd}$ increases, the switching delay decreases because the electrostatic force increases, but the switching energy increases. The negative sensitivity to $V_{\rm dd}$ is shown in Fig. 8 and given by

$$\frac{\partial t_{\rm delay}/\partial V_{\rm dd}}{\partial E_s/\partial V_{\rm dd}} = -\frac{\beta t_{\rm delay}}{2E_s} \times V_{\rm dd}/(V_{\rm dd} - \chi V_{\rm pi}) \equiv -\frac{V_{\rm norm}}{2}\frac{t_{\rm delay}}{E_s} \quad (13)$$

[4]If the sensitivities are not equal, e.g., if $S_A < S_B < 0$ at a given operating point (A, B), one can always decrease the switching delay without paying any energy penalty by tuning the parameters A and B (within the constraints for these parameters) as follows:

1) Adjust A to decrease the switching delay by $\Delta t_{\rm delay} = S_A \cdot (\Delta E_s)$. Switching energy is increased by ΔE_s as a result.
2) Adjust B to recover the increased energy ΔE_s. Switching delay is increased by $\Delta t_{\rm delay} = S_B \cdot (-\Delta E_s)$ as a result.
3) Since $S_A < S_B < 0$, an overall delay reduction $\Delta t_{\rm delay} = (S_A - S_B) \cdot (\Delta E_s)$ is achieved without paying any energy penalty.

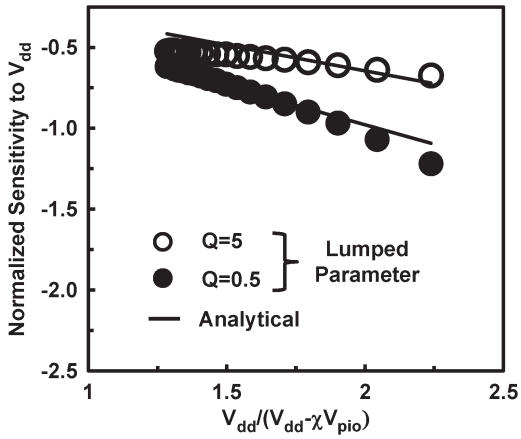

Fig. 8. Normalized sensitivity to supply voltage $(V_{\rm dd})$. The nominal pull-in voltage $(V_{\rm pio})$ is 4.15 V.

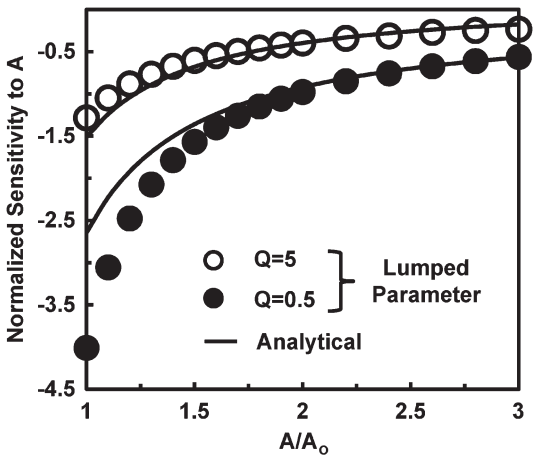

Fig. 9. Normalized sensitivity to fabricated actuation area (A). The nominal actuation area (A_o) is 15×30 μm^2.

where $t_{\rm delay}$ and E_s are the delay and energy, respectively, at a given $V_{\rm dd}$ and $V_{\rm norm} = \beta V_{\rm dd}/(V_{\rm dd} - \chi V_{\rm pi})$. As will be discussed later, $V_{\rm norm}$ typically lies within the range of 1.17–1.29. Therefore, the normalized sensitivity [29], which is defined as $E_s/t_{\rm delay} \times (\partial t_{\rm delay}/\partial V_{\rm dd})/(\partial E_s/\partial V_{\rm dd})$, is roughly $-V_{\rm norm}/2$ or $-$(0.585–0.645). As a rough rule of thumb, then, every $2\times$ energy increase by $V_{\rm dd}$ adjustment will yield a $\sim$1.5$\times$ reduction in turn-on delay.

2) Sensitivity to Actuation Area (A)*:* The sensitivity to actuation area is shown in Fig. 9 and given by

$$\frac{\partial t_{\rm delay}/\partial A}{\partial E_s/\partial A} = \frac{t_{\rm delay}}{2E_s}(m_{\rm norm} - V_{\rm norm})(1 + C_{\rm norm}) \quad (14)$$

where $m_{\rm norm} = \alpha_o phA/m_{\rm eff}$ is the actuation-plate-to-total mass ratio and $C_{\rm norm} = C_L/(\varepsilon_o A/(g-g_d))$ is the fixed-to-area-dependent capacitance ratio. Since $m_{\rm norm} < 1$ and $V_{\rm norm} > 1$, the sensitivity is always negative. Intuitively, as A increases, the ON-state capacitance and therefore the switching energy increase, while the switching time decreases due to the increase in gate overdrive.

3) Sensitivity to As-Fabricated Gap Thickness (g)*:* Fig. 10 shows the sensitivity as a function of the as-fabricated gap,

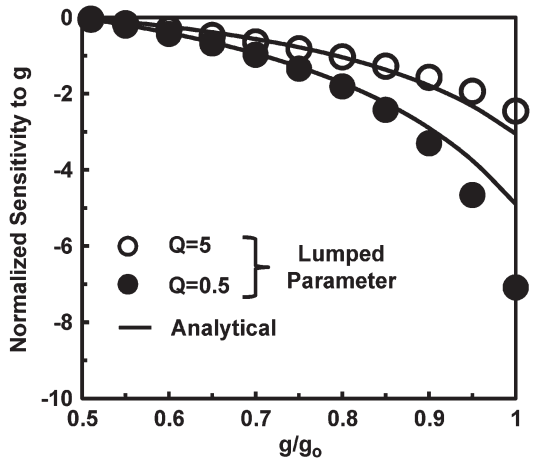

Fig. 10. Normalized sensitivity to fabricated gap thickness. The nominal gap thickness (g_o) is 200 nm.

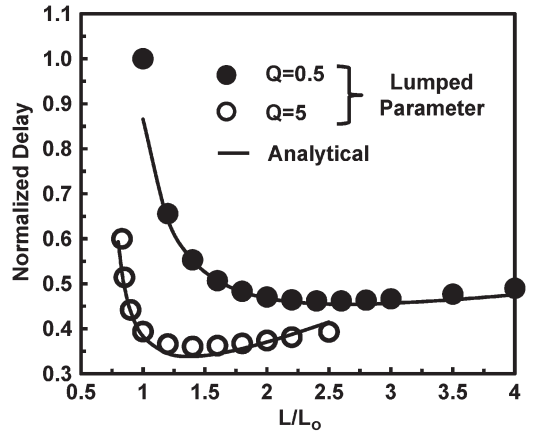

Fig. 11. Normalized delay as a function of the beam length. The nominal beam length (L_o) is 40 μm.

which is also given by

$$\frac{\partial t_{\rm delay}/\partial g}{\partial E_s/\partial g} = -\frac{t_{\rm delay}}{E_s}\left(1-\frac{g_d}{g}\right)(1+C_{\rm norm})\left(-\gamma+\frac{3}{2}V_{\rm norm}\right). \quad (15)$$

Since $g_d/g < 1$ and $\gamma < 1 < V_{\rm norm}$, (15) shows that the sensitivity to g is always negative. This is expected since a small actuation gap (g) is desirable for increasing the electrostatic actuation force, but comes at the expense of increasing the ON-state capacitance and, hence, the switching energy. Ideally, the smallest contact dimple gap should be utilized to minimize the displacement of the actuation electrode.

4) Sensitivity to Beam Length (L): If the suspension beams do not contribute substantial capacitance, then to first order and as indicated by (11), the relay switching energy is independent of the beam length. As the beam length is increased, $V_{\rm pi}$ decreases and hence the gate overdrive increases, leading to improved switching speed. However, the increase in loaded mass counteracts this effect. As a result, as shown in Fig. 11, there exists an optimal beam length that balances the gate overdrive and the resonant frequency such that

$$\frac{dt_{\rm delay}}{dL} = 0 \Rightarrow \left[\left(\frac{1}{\omega_o}\frac{\partial \omega_o}{\partial L}\right) / \left(\frac{1}{V_{\rm pi}}\frac{\partial V_{\rm pi}}{\partial L}\right)\right] = V_{\rm norm} \quad (16)$$

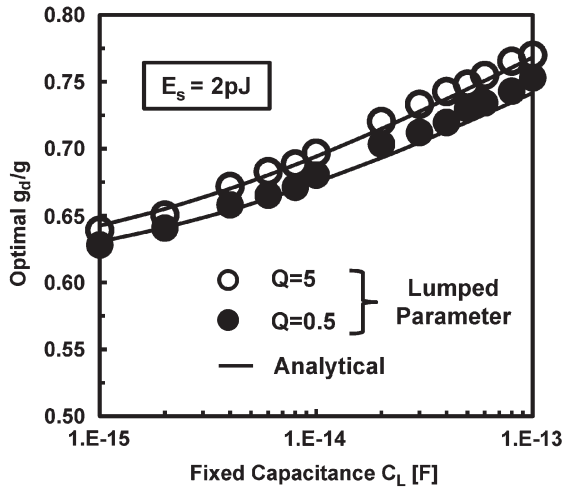

Fig. 12. Optimal relay g_d/g ratio (with an arbitrary energy constraint E_s of 2 pJ and other design parameters optimized as well) as a function of the fixed capacitance.

which can further be simplified to

$$V_{\rm norm} = 1 - \kappa(1 - m_{\rm norm}) \quad (17)$$

where $\kappa = (k_{\rm eff}/L)/(dk_{\rm eff}/dL)$, which is related to the normalized sensitivity of spring constant to beam length, and is roughly -0.4 or -0.34 for relays with short or long beam length, respectively. Thus, $V_{\rm norm}$ is directly proportional to $m_{\rm norm}$, and proper design of the relay entails the selection of the beam length and actuation area that balance $m_{\rm norm}$ and $V_{\rm norm}$.

B. Relay Design Optimization

An optimized relay design is reached by balancing the various sensitivities to the design variables. Based on this balance, simple guidelines for an energy-efficient relay design are derived in this section.

1) Optimal Gap Thickness Ratio (g_d/g): Balancing the sensitivities to $V_{\rm dd}$ and to g results in the following design equation:

$$-\frac{1}{2}V_{\rm norm} = -\left(1-\frac{g_d}{g}\right)(1+C_{\rm norm})\left(-\gamma+\frac{3}{2}V_{\rm norm}\right). \quad (18)$$

Since $\gamma < 1 < V_{\rm norm}$, $-\gamma + 3V_{\rm norm}/2$ is approximately equal to $3V_{\rm norm}/2$, and so, the optimal g_d/g ratio can be approximated by

$$\frac{g_d}{g} \approx \frac{2+3C_{\rm norm}}{3+3C_{\rm norm}}. \quad (19)$$

Therefore, as shown in Fig. 12, the optimal g_d/g ratio is roughly 0.66–0.75 across a large fixed-to-area-dependent capacitance ratio range, and the optimal gap thickness ratio is largely independent of the quality factor. Since $g_d/g > 2/3$, this means that relay operation in pull-in mode provides for the optimum energy efficiency.

2) Optimal $V_{\rm dd}/V_{\rm pi}$: To find the optimal $V_{\rm norm}, m_{\rm norm}$ first must be found by balancing the sensitivities to actuation area

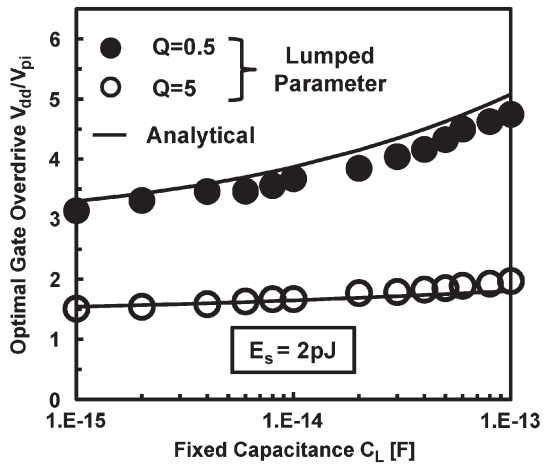

Fig. 13. Optimal gate overdrive ($V_{\rm dd}/V_{\rm pi}$) values (with an arbitrary energy constraint E_s of 2 pJ and other design parameters optimized as well) for high- and low-Q relays versus fixed capacitance.

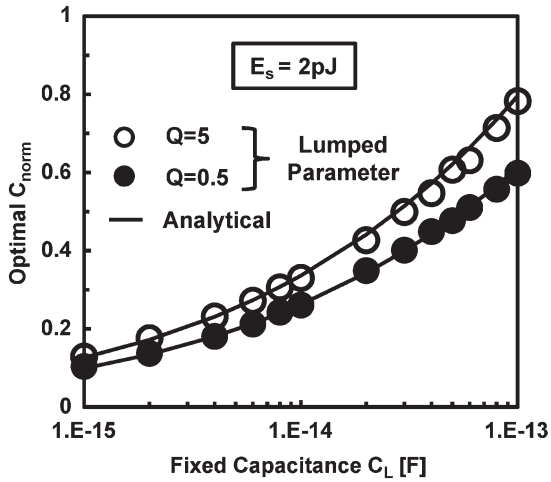

Fig. 14. Optimal fixed-to-area-dependent capacitance ratio for a relay (with an arbitrary energy constraint E_s of 2 pJ and other design parameters optimized) versus fixed capacitance.

and to fabricated gap thickness. The result is shown in

$$m_{\rm norm} \cong -2\left(1-\frac{g_d}{g}\right)\left(-\gamma+\frac{3}{2}V_{\rm norm}\right)+V_{\rm norm}. \quad (20)$$

Substituting (20) into (17) and using the result that the optimal g_d/g value is roughly 0.68–0.76, the optimal $V_{\rm norm}$ can be shown to be largely fixed

$$\begin{aligned}&\beta V_{\rm dd}/(V_{\rm dd}-\chi V_{\rm pi})\\ &\quad = V_{\rm norm} = \left(2\gamma-1+\frac{1}{\kappa}-2\gamma\frac{g_d}{g}\right)/\left(2+\frac{1}{\kappa}-3\frac{g_d}{g}\right)\\ &\quad \approx 1.17-1.29. \end{aligned} \quad (21)$$

Using this result, the optimal gate overdrive can be obtained and is shown in Fig. 13. For low-Q relays with $\beta > 1$, the optimal value of $V_{\rm dd}/V_{\rm pi}$ is ~4.5. For high-Q relays with $\beta \approx 0.7$, the optimal value of $V_{\rm dd}/V_{\rm pi}$ lies within the range of 1.6–2.[5] As will be discussed in the next section, these results are very useful for selecting the optimal actuation area, supply voltage, and beam length.

3) Optimal Actuation Area (A) and Supply Voltage ($V_{\rm dd}$): Since the optimal g_d/g ratio is roughly constant at 0.7, the relay can be sized if the optimal fixed-to-area-dependent capacitance ratio is known. By balancing the sensitivities to the actuation area and to the supply voltage, the optimal fixed-to-area-dependent ratio can be obtained

$$C_{\rm norm} = \frac{m_{\rm norm}}{V_{\rm norm}-m_{\rm norm}}. \quad (22)$$

Substituting (17) into (22), the optimal fixed-to-area-dependent capacitance ratio can be expressed as a function of $V_{\rm norm}$

$$C_{\rm norm} = \frac{\kappa-1+V_{\rm norm}}{(1-V_{\rm norm})(1-\kappa)}. \quad (23)$$

[5]Due to the hysteretic switching behavior, the optimal $V_{\rm dd}$ and $V_{\rm pi}$ derived herein do not necessarily achieve the largest noise margin. Similarly as for CMOSFETs, the energy-delay optimal design for a relay is therefore not the same as a design optimized for a static noise margin [30].

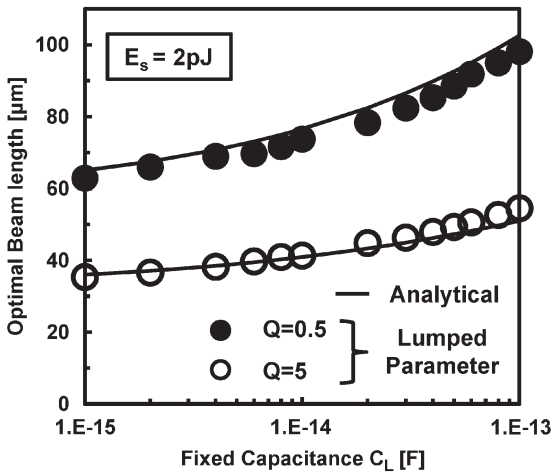

Fig. 15. Optimal relay beam length (with an arbitrary energy constraint E_s of 2 pJ and other design parameters optimized) as a function of the fixed capacitance. Low-Q relays have higher gate overdrive and therefore longer beams are preferred.

If the relay must drive a fixed capacitance under a given energy constraint, it is preferable to upsize the relay (i.e., increase the actuation area) and lower $V_{\rm dd}$ to reduce the energy spent on driving the fixed capacitance. As a result, the optimal fixed-to-area-dependent capacitance ratio is less than one, as shown in Fig. 14.[6] The optimal $C_{\rm norm}$ increases as the fixed capacitance increases, since an increase in gate overdrive (with relay upsizing) decreases $V_{\rm norm}$. From (23), for a low-Q relay with $\kappa \approx -0.34$ and $V_{\rm norm} \approx 1.17-1.29$, the optimal $C_{\rm norm}$ ranges from 0.74 to 0.13; for a high-Q relay with $\kappa \approx -0.4$ and $V_{\rm norm} \approx 1.17-1.29$, the optimal $C_{\rm norm}$ ranges from 0.97 to 0.27.[7] These calculations are qualitatively consistent with the simulation and modeling results shown in Fig. 14. Once the optimal values for the as-fabricated gap thickness and actuation

[6]Note that for a CMOS transistor driving a fixed capacitive load, one would also obtain similar optimal fixed-to-area-dependent capacitance ratio.

[7]Variations in $C_{\rm norm}$ are much larger than those in $V_{\rm norm}$ since $V_{\rm norm}$ is close to one, and therefore, $C_{\rm norm}$, which is proportional to $1/(1-V_{\rm norm})$, is a large number.

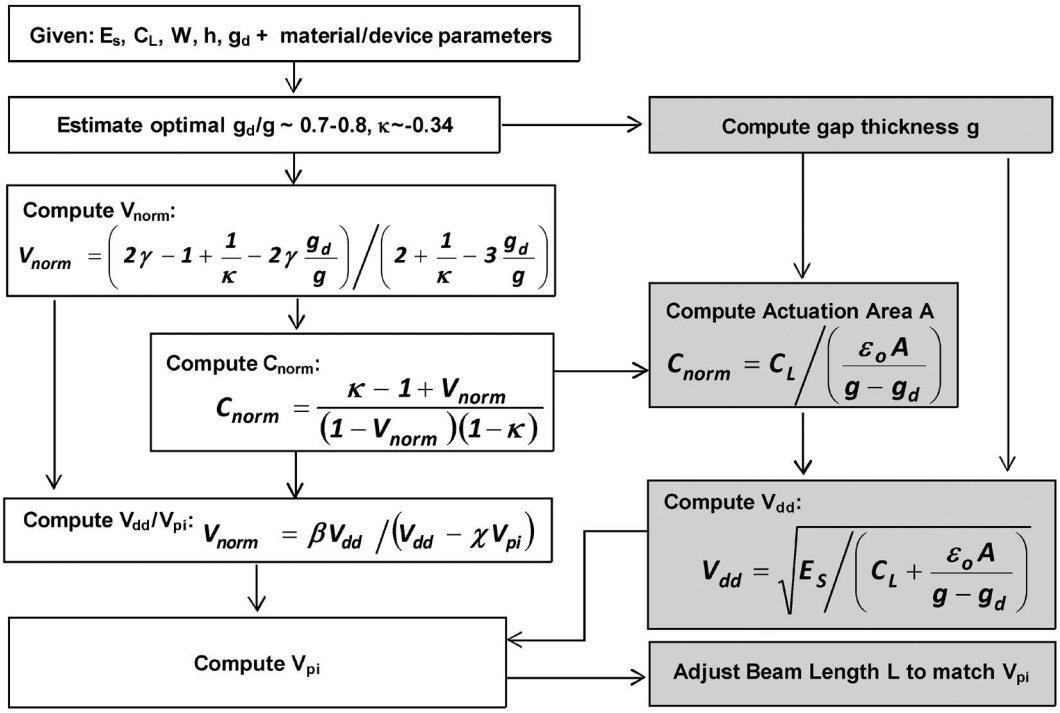

Fig. 16. Flowchart illustrating a simple relay design optimization procedure.

area are obtained, the optimal $V_{\rm dd}$ is simply set by the energy constraint, with the upper bound for $V_{\rm dd}$ set by the catastrophic pull-in voltage.

4) Optimal Beam Length (L): With the optimal values of $V_{\rm norm}$ and $V_{\rm dd}$ known, the optimal beam length can be selected to achieve the optimal $V_{\rm pi}$. Fig. 15 shows the optimal beam length for low- and high-Q relays. For low-Q relays with $\beta > 1$, the optimal value of $V_{\rm dd}/V_{\rm pi}$ is $\sim$4.5 so that longer beams are preferred. In practice, the longest beam length will be set by the surface adhesion energy, or by layout area constraints. For high-Q relays, the optimal value of $V_{\rm dd}/V_{\rm pi}$ lies within the range of 1.6–2 so that shorter beams are preferred.

5) Relay Design Optimization Procedure: From the results of the sensitivity analysis, the following simple relay design optimization procedure (illustrated in Fig. 16) can be established.

1) Since both g_d/g and κ are roughly fixed, start with an estimated optimal g_d/g ratio of 0.7, and assume that $\kappa = -0.34$. The dimple gap thickness g_d is fixed by the limits of the fabrication process technology; thus, the optimal gap thickness g can be calculated. Furthermore, using (21) and a knowledge of the expected Q of the relay (which sets γ), the optimal $V_{\rm norm}$ can be estimated.
2) Next, with κ and $V_{\rm norm}$ known, the optimal fixed-to-area-dependent capacitance ratio $C_{\rm norm} = (\kappa - 1 + V_{\rm norm})/[(1 - V_{\rm norm})(1 - \kappa)]$ and hence the optimal actuation area can be calculated, given the fixed load capacitance.
3) Using $C_{\rm norm}$, the fixed load capacitance, and the fact that $V_{\rm norm} = \beta(V_{\rm dd}/V_{\rm pi})/((V_{\rm dd}/V_{\rm pi}) - \chi)$, $V_{\rm dd}$ can be calculated for the given energy budget. Then, the optimal $V_{\rm pi}$ can be calculated.
4) Finally, the optimal beam length is calculated to result in the optimal $V_{\rm pi}$.

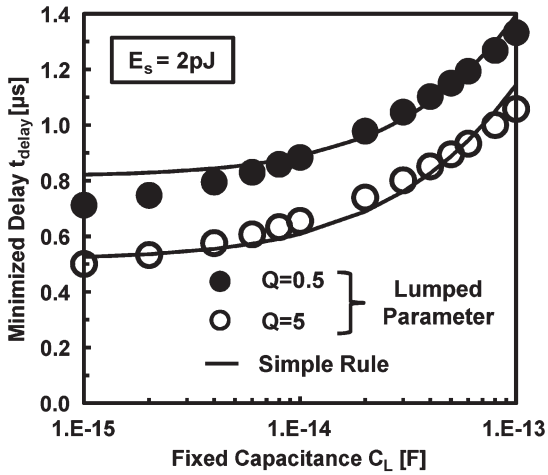

Fig. 17. Optimized relay delay as a function of the fixed capacitance. Optimally designed low-quality-factor relays can achieve a comparable switching delay as their high-quality-factor counterparts, to within 30%.

Note that the switching delay will be determined once all of the relay design parameters are set.

To provide a concrete example, this procedure is applied to optimize the design of a 5-μm-wide relay (with the parameters shown in Table I) for a nominal switching energy $E_s = 2$ pJ and an average fixed capacitance C_L ranging from 10 to 100 fF. The results are shown in Fig. 17 and match the predictions using the lumped parameter model to within 10%. Note that the delay of the optimized low-Q relay is within 30% of the optimized high-Q relay.

To obtain the energy-performance tradeoff curve, this optimization process can be repeated for different values of E_s. An easier approach is to note that the optimal normalized sensitivity ($\approx -V_{\rm norm}/2$) is related to the slope of the tradeoff

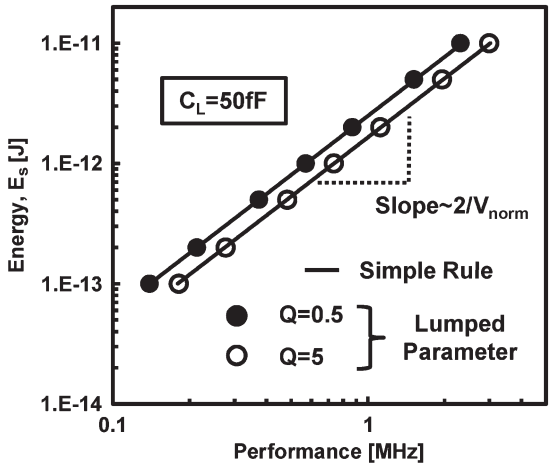

Fig. 18. Relay energy versus performance. Note that the tradeoff curve is a straight line with slope $\sim 2/V_{\rm norm}$.

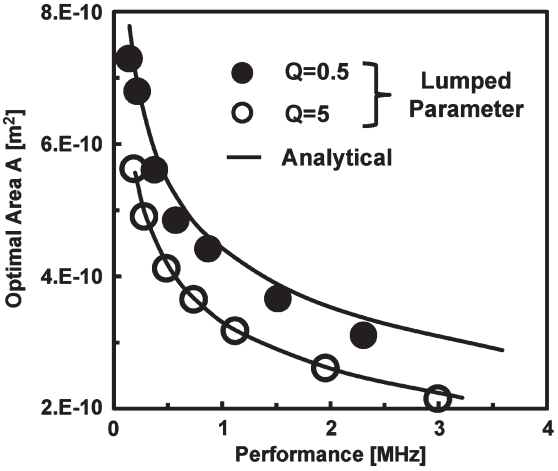

Fig. 19. Optimal actuation area as a function of the relay performance. For the same performance, low-quality-factor relays are roughly 30%–50% larger than their high-quality-factor counterparts.

curve on a log–log plot [31]. Since $V_{\rm norm}$ is relatively constant, the energy-delay tradeoff curve is roughly a straight line with a slope of $1/(V_{\rm norm}/2) \sim 1.64$, as shown in Fig. 18. Note that for high-performance relays, it is preferable to downsize the relay (i.e., decrease the actuation area A) and increase $V_{\rm dd}$ to reduce the mass of transport and increase the electrostatic force, respectively. Therefore, the optimal actuation area decreases with increasing relay performance, as shown in Fig. 19.

C. Energy-Efficiency Limit

For ultra-low-power electronics applications such as wireless sensor networks, switching energy [32] rather than speed is the primary concern. The minimum switching energy for relays is dictated by the need for the spring restoring force to overcome the surface adhesion energy (Γ) in the ON state, in order to break physical contact

$$0.5 k_{\rm eff} g_d^2 \geq \Gamma. \tag{24}$$

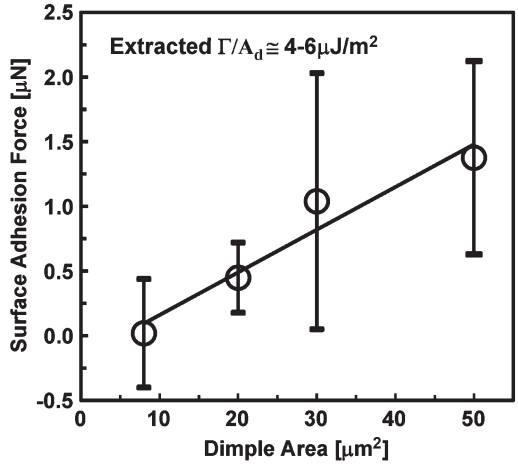

Fig. 20. Extracted average F_A (with standard deviation indicated) versus A_d. The surface adhesion energy per unit dimple area (Γ/A_d) is extracted from the surface force [11]. Each data point is obtained from measurements for more than ten relays with different L values.

In other words, the minimum spring stiffness is $k_{\rm eff} = 2\Gamma/g_d^2$. Therefore, the minimum pull-in and supply voltages are

$$V_{\rm dd,min} = V_{\rm pi,min} = \sqrt{\frac{8k_{\rm eff}g^3}{27\varepsilon_o A}} = \sqrt{\frac{16\Gamma g^3}{27\varepsilon_o A g_d^2}}. \tag{25}$$

If the fixed load capacitance is ignored, the minimum relay switching energy is

$$E_s = \left(\frac{\varepsilon A}{g - g_d}\right) V_{\rm dd,min}^2 = \left(\frac{16\Gamma}{27}\right) \frac{1}{(1 - g_d/g)(g_d/g)^2} \tag{26}$$

which has a minimum value of 4Γ at $g_d/g = 2/3$. Of course, the switching energy will be higher for any realistic relay design since operation with $V_{\rm rl} = 0$ V and $V_{\rm dd} = V_{\rm pi}$ provides for no noise margin and thus would be highly impractical.

IV. Scaling Implications

Relay miniaturization is desirable for improved device density and reduced operating voltage. Analogous to the classic scaling methodology developed for MOSFETs [33], a constant-field scaling methodology (by which the electric field across the actuation gap is maintained at a constant value while each of the device dimensions is scaled by a factor S) for microelectromechanical systems (MEMS) has been reported [34]. Although this simple methodology provides useful insight into the benefits of relay scaling, it may not provide for the optimal relay design. To remedy this, the implications of the relay design optimization methodology for scaling to improve switching speed, energy, and layout area are presented herein and summarized in Table III.

As previously discussed, the energy-delay tradeoff curve is approximately a straight line with a slope of $\sim$1.64. Furthermore, relay energy efficiency is ultimately limited by the surface adhesion energy. Thus, one only needs to focus on how surface adhesion/minimum switching energy scales with contact dimple area to understand how the relay energy-performance tradeoff changes with device scaling. As shown in Fig. 20, the surface adhesion force (which consists of

TABLE II
COMPARISON OF RELAY SCALING METHODOLOGIES

Parameter	Constant Scaling Factor	Variable Scaling Factor
(W, h, g_d, C_L)	S	(S_W,S_h,S_{gd},S_{CL})
Dimple Area A_d, Surface Adhesion Energy E_s	S^2	S_{Ad}^2
Actuation Gap Thickness g,	S	S_{gd}
Gate Capacitance	S	S_{CL}
Actuation Area A	S^2	S_{gd},S_{CL}
Supply and Pull-in Voltages V_{dd}, V_{pi}	$S^{0.5}$	S_{Ad},$S_{CL}^{-0.5}$
Beam Length L	$S^{4/3}$	$S_{Ad}^{-2/3}$,$S_W^{1/3}$,S_h,$S_{gd}^{2/3}$
Speed	$\sim S^{1.5}$	$\sim(S_{Ad}^{-2}$,S_{gd}^3,S_{CL},$S_h)^{0.5}$

van der Waals forces, capillary forces, and hydrogen bonds [35]) and, hence the surface adhesion energy reduce with A_d. This means that relay designs with lower beam stiffness, smaller contact dimple gap thickness, and therefore lower actuation area and supply voltage are feasible if a smaller contact dimple area is utilized.

With this in mind, suppose that A_d is reduced by a factor S^2 and that W, h, g_d, and C_L each are reduced by a factor S. To maintain the same optimal g_d/g ratio of $\sim$0.7, the as-fabricated gap thickness g must be reduced by S. As a consequence, the actuation area must be reduced by S^2 to achieve the same optimal fixed-to-area-dependent capacitance ratio. Surface adhesion energy and, therefore, the minimum relay switching energy improve by S^2. Since the total capacitance is reduced by S and the switching energy improves by S^2, the supply voltage can be scaled down by $S^{0.5}$. Finally, to maintain the same optimal gate overdrive, $V_{\rm pi}$ is also reduced by $S^{0.5}$; to achieve this goal, the beam length must be reduced by $S^{4/3}$. As a consequence, the switching speed improves by $\sim S^{1.5}$.

Ultimately, one or more of the variables W, h, g_d, and C_L will reach a lower limit and may not be scaled as readily than the other parameters. For example, g_d may be limited by nanogap-formation technology. In that case, suppose that A_d is still reduced by the factor S^2 and that each of W, h, and C_L is reduced by a factor S. If g_d is not scaled, the as-fabricated gap thickness g must also be fixed to maintain the optimal g_d/g ratio of $\sim$0.7. The actuation area must now be reduced by S (instead of S^2) to achieve the same optimal fixed-to-area-dependent capacitance ratio. Surface adhesion energy (S^2), the minimum relay switching energy (S^2), total capacitance (S), the switching energy (S^2), supply voltage ($S^{0.5}$), and the pull-in voltage ($S^{0.5}$) still improve by the same factors, but the beam length must now be reduced by $S^{2/3}$ to maintain the same gate overdrive. As a consequence, switching speed is improved by $\sim S^{1/3}$.

To generalize the scaling methodology, suppose that A_d, W, h, g_d, and C_L are respectively reduced by the factors $S_{\rm Ad}^2$, S_W, S_h, $S_{\rm gd}$, and $S_{\rm CL}$; using the same procedure as described previously, the minimum relay switching energy is improved by $S_{\rm Ad}^2$, and the switching speed is improved by $\sim (S_{\rm Ad}^{-2} \cdot S_{\rm gd}^3 \cdot S_{\rm CL} \cdot S_h)^{0.5}$. These results are summarized in Table II.

It should be noted that, for a very small contact dimple area, Γ will reach a lower limit set by the degree of bonding needed to meet the contact resistance requirement [35], [36], with the associated energy typically in the 0.2-aJ/bond range [36], [37]. For example, with five metal–metal bonds, $E_{\rm min} \cong 4$ aJ ($> 10\times$ lower than CMOS) would be achievable. This sets the ultimate energy scaling limit for relays.

Using the calibrated analytical relay model with scaled device dimensions as shown in Table III, and following the scaling methodology established herein with $\Gamma = 1$ aJ, the energy-delay performance of a relay in a 65-nm equivalent technology is compared against that of a MOSFET at the 65-nm technology node in Fig. 21. Relay technology is projected to provide for $> 10\times$ energy savings as compared against an equivalent MOSFET technology for circuits operating at clock frequencies up to $\sim$100 MHz. Note that since relays have a relatively low gate capacitance, their performance is very sensitive to load/wiring capacitance.

Given that relay circuits would be used for applications with clock frequencies up to $\sim$100 MHz, it is of practical interest to compare them against subthreshold-CMOS circuits (operated with $V_{\rm dd} < V_T$), which are designed to operate with very low power consumption [6], [32]. The minimum energy for subthreshold-CMOS circuits is reached by properly balancing the dynamic and the leakage energies. As derived in [6] and [32], the optimum supply voltage is proportional to the thermal voltage

$$V_{\rm ddopt,MOS} \propto n \times k_B T/q \tag{27}$$

where $n \approx 1.2$ is the subthreshold factor [6], [32]. Hence, the minimum energy is

$$E_{\rm min,MOS} \propto C_{\rm MOS} V_{\rm ddopt,MOS}^2. \tag{28}$$

TABLE III
RELAY DESIGN PARAMETERS FOR A 65-nm EQUIVALENT TECHNOLOGY

Parameter	Value[8]
Beam Width, W	65nm
Beam Thickness, h	15nm
Fabricated Dimple Gap Thickness, g_d	10nm
Dimple Area, A_d	$50\times50\text{nm}^2$
Truss Width, W_T	65nm
Truss Length, L_T	156nm
(γ_f, γ_t)	$(2.15, 5.13\times10^{13}\ \text{m}^{-2})$
(α_0, α_1)	(1.11 , 0.5)

[8]A 10nm dimple gap thickness is assumed because it is thus far the thinnest gap that has been successfully realized for MEMS [38]; a $50\times50\text{nm}^2$ contact dimple area is reasonable for a 65nm technology.

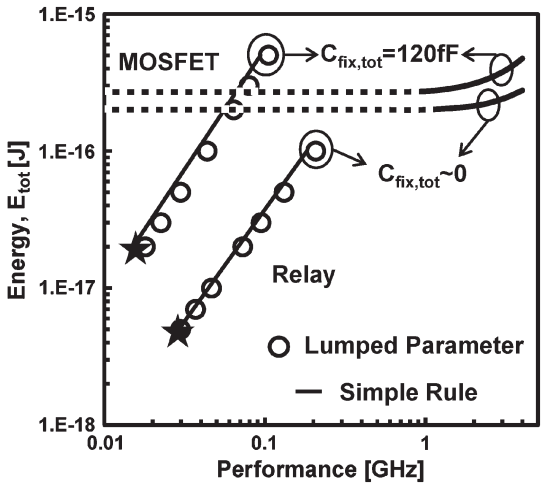

Fig. 21. Simulated energy-performance comparison for a 30-stage fan-out-4 inverter chain versus a relay chain (average transition probability $=0.01$, $r=1$, and total fixed capacitance $C_{\text{fix,tot}}$) [9]. $C_{\text{fix,tot}} = 120$ fF corresponds to an average fixed capacitance per relay C_L of 0.5 fF. The MOSFET parameters are taken from the ITRS, for the 65-nm LSTP technology node. The inverter chain is optimized by gate sizing and supply and threshold voltage adjustment [23]–[28]. The relay parameters are summarized in Table III. The minimum energy (indicated as stars in the figure) is set by the surface adhesion energy Γ. Notice that, due to low gate capacitance, relay performance is more sensitive to fixed capacitance than MOSFET performance [9].

Assuming ideal MOSFET scaling and constant operating temperature, $V_{\text{ddopt,MOS}}$ remains relatively constant, and hence, the minimum energy of CMOS scales linearly by the factor S. Based on Fig. 21, the physical gate length would need to be scaled down by approximately 20 times (i.e., to below 5 nm) to match the minimum energy potentially achievable with relays. However, it is unlikely that CMOS scaling will proceed in this manner for several reasons. Due to worsening short-channel effects, the subthreshold factor n typically increases with transistor scaling, leading to an increase in $V_{\text{ddopt,MOS}}$. As discussed in [39], due to this effect, the optimal supply voltage of a 32-bit adder increases from $\sim$0.25 to $\sim$0.33 V from the 65- to 32-nm technology nodes. Furthermore, due to increasing variability, the minimum device width has been scaling relatively slowly and leading to minimal reduction in C_{MOS}. In fact, the International Technology Roadmap for Semiconductors (ITRS) [40] predicts that the gate capacitance will only decrease by $\sim2.5\times$ as the transistor physical gate length is scaled from 38 to 7.4 nm. Therefore, the minimum energy for CMOS is expected to saturate (and perhaps even increase), and hence, relays will likely retain (or perhaps even increase) their energy-efficiency benefits.

V. CONCLUSION

Circuit-level energy-performance analysis is necessary to evaluate the promise of any new device technology for potentially overcoming the energy-efficiency limitations of CMOS technology. In this work, a circuit-driven energy-delay sensitivity analysis is performed to establish guidelines for relay design optimization. Closed-form analytical models for relay switching delay and energy were first derived, and are found to fit well with experimental data. These models were then used to establish a sensitivity-based energy-delay design optimization procedure for relays. The optimal normalized sensitivity of delay to energy is roughly constant at -0.6, implying that every $2\times$ energy increase will yield a $\sim1.5\times$ reduction in relay delay. Using this fact, simple rules for optimal relay design are developed. For a given contact dimple gap thickness, the optimal dimple-gap thickness to actuation-gap thickness ratio is roughly 0.7, meaning that pull-in operation is preferred for energy-efficient relay design. Interestingly, an optimally designed low-Q relay can achieve a switching delay within 30% of its high-Q counterpart. This implies that vacuum packaging might not be necessary for logic relays from a performance perspective.

Much like transistor scaling, relay miniaturization leads to dramatic improvements in density (for lower cost per

function), switching delay (for higher performance), and energy efficiency. A scaled relay technology is projected to provide for $> 10\times$ energy savings as compared against an equivalent MOSFET technology for circuits operating at clock frequencies up to $\sim$100 MHz.

Appendix I
Derivation of the Catastrophic Pull-In Voltage

When the relay is in the ON state, the movable electrode anchored at the two contact dimple regions, as shown in Fig. 2(b). In general, the bending of the movable electrode can be estimated by the Euler–Bernoulli equation [18]

$$\left(\frac{E}{1-\upsilon^2}\right) I_A \frac{d^4 z^*}{dx^4} = \begin{cases} -\frac{\varepsilon_o W_A V^2}{2(g-g_d-z^*)^2}, & \text{in the actuation region} \\ 0, & \text{everywhere else} \end{cases} \quad \text{(A1.1)}$$

where $E/(1-\upsilon^2)$ instead of the Young's modulus E is used to account for the plate effects, I_A is the moment of inertia $I_A = W_A h^3/12$, and $z^* = z - g_d$.

An exact $V_{\rm cpi}$ model that accounts for anchor stability is complex. However, a simplified model readily provides both lower and upper bounds for $V_{\rm cpi}$. Specifically, a lower bound for $V_{\rm cpi}$ can be derived by modeling the movable plate as a beam pinned at the two contact dimple regions. Therefore, (A1.1) is solved with the appropriate boundary conditions at both dimples

$$z^*(x=0)=z^*(x=L_A)=0 \qquad \left.\frac{d^2 z^*}{dx^2}\right|_{x=0} = \left.\frac{d^2 z^*}{dx^2}\right|_{x=L_A} = 0 \quad \text{(A1.2)}$$

which gives the analytical formulation for the lower bound of $V_{\rm cpi}$

$$V_{\rm cpi} \cong 1.516\sqrt{\frac{Eh^3(g-g_d)^3}{\varepsilon_o L_A^4}}. \quad \text{(A1.3)}$$

Similarly, an upper bound for $V_{\rm cpi}$ can be derived by modeling the movable plate as a beam solidly anchored at the two contact dimple regions, which has the following boundary conditions:

$$z^*(x=0)=z^*(x=L_A)=0 \qquad \left.\frac{dz^*}{dx}\right|_{x=0} = \left.\frac{dz^*}{dx}\right|_{x=L_A} = 0 \quad \text{(A1.4)}$$

resulting in the following analytical formulation for the upper bound of $V_{\rm cpi}$:

$$V_{\rm cpi} \cong 3.444\sqrt{\frac{Eh^3(g-g_d)^3}{\varepsilon_o L_A^4}}. \quad \text{(A1.5)}$$

Appendix II
Origin of the Analytical Delay Equation

Newton's second law of motion for a relay can be normalized into

$$\frac{d^2 z^*}{dt^{*2}} + \frac{1}{Q}\frac{dz^*}{dt^*} + z^* = \frac{4}{27}\frac{V^{*2}}{(1-z^*)^2} \quad \text{(A2.1)}$$

where $z^* = z/g$, $t^* = \sqrt{k_{\rm eff}/m_{\rm eff}} \times t$, and $V^* = V/V_{\rm pi}$. Therefore, the normalized delay $t^*_{\rm delay} = \sqrt{k_{\rm eff}/m_{\rm eff}} \times t_{\rm delay}$ is a function of the gate overdrive $V_{\rm dd}/V_{\rm pi}$, the normalized dimple gap thickness g_d/g, and the quality factor Q. In fact, as was discussed in [41], the pull-in delay is roughly proportional to $t_{\rm delay} \sim \sqrt{k_{\rm eff}/m_{\rm eff}} \times (V_{\rm pi}/V_{\rm dd})$; the accuracy of this approximate solution is improved herein by the following signomial expression:

$$t_{\rm delay} \cong \alpha\sqrt{\frac{m_{\rm eff}}{k_{\rm eff}}}\left(\frac{g_d}{g}\right)^{\gamma}\left(\frac{V_{\rm dd}}{V_{\rm pi}}-\chi\right)^{-\beta}, \quad \text{for } 5V_{\rm pi} \geq V_{\rm dd} > 1.1V_{\rm pi};\quad g_d \geq g/3 \quad \text{(A2.2)}$$

where $\chi \approx 0.8$ accounts for the dramatical increase in $t_{\rm delay}$ as $V_{\rm dd}$ approaches $V_{\rm pi}$. α, β, and γ depend on Q and can be found numerically using (A2.1). Equation (A2.2) predicts $t_{\rm delay}$ values to within 20% of (A2.1) over the range of interest, with the accuracy improving with increasing $V_{\rm dd}/V_{\rm pi}$.

Acknowledgment

The authors would like to thank R. Nathanael, V. Pott, J. Jeon, and M. Spencer of UC Berkeley; F. Chen and H. Fariborzi of MIT; and C. Wang and K. Dwan of UCLA for the helpful discussions. The MEM relays were fabricated in the UC Berkeley Microfabrication Laboratory.

References

[1] M. A. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein, "Scaling, power, and the future of CMOS," in *IEDM Tech. Dig.*, Dec. 2005, pp. 9–15.

[2] T. Baba, "Proposal for surface tunnel transistor," *Jpn. J. Appl. Phys.*, vol. 31, no. 4B, pp. L455–L457, Apr. 1992.

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

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

[5] S. Salahuddin and S. Datta, "Use of negative capacitance to provide a sub-threshold slope lower than 60 mV/decade," *Nanoletters*, vol. 8, no. 2, pp. 405–410, 2008.

[6] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. K. Das, W. Haensch, E. J. Nowak, and D. M. Sylvester, "Ultralow-voltage, minimum-energy CMOS," *IBM J. Res. Develop.*, vol. 50, no. 4/5, pp. 469–490, Jul. 2006.

[7] K. Akarvardar, D. Elata, R. Parsa, G. C. Wan, K. Yoo, J. Provine, P. Peumans, R. T. Howe, and H.-S. P. Wong, "Design considerations for complementary nanoelectromechanical logic gates," in *IEDM Tech. Dig.*, Dec. 2007, pp. 299–302.

[8] F. Chen, H. Kam, D. Markovic, T.-J. K. Liu, V. Stojanovic, and E. Alon, "Integrated circuit design with NEM relays," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Des.*, 2008, pp. 750–757.

[9] H. Kam, T.-J. K. Liu, E. Alon, and M. Horowitz, "Circuit level requirements for MOSFET replacement devices," in *IEDM Tech. Dig.*, Dec. 2008, p. 427.

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

[11] H. Kam, V. Pott, R. Nathanael, J. Jeon, E. Alon, and T.-J. K. Liu, "Design and reliability of a MEM relay technology for zero-standby-power digital logic applications," in *IEDM Tech. Dig.*, Dec. 2009, pp. 809–812.

[12] G.-L. Tan and G. M. Rebeiz, "A DC-contact MEMS shunt switch," *IEEE Microw. Wireless Compon. Lett.*, vol. 12, no. 6, pp. 212–214, Jun. 2002.

[13] P. M. Zavracky, S. Majumder, and N. E. McGruer, "Micromechanical switches fabricated using nickel surface micromachining," *J. Microelectromech. Syst.*, vol. 6, no. 1, pp. 3–9, Mar. 1997.
[14] C. Goldsmith, J. Randall, S. Eshelman, T. H. Lin, D. Denniston, S. Chen, and B. Norvell, "Characteristics of micromachined switches at microwave frequencies," in *Proc. IEEE MTT-S Int. Microw. Symp. Dig.*, San Francisco, CA, Jun. 1996, pp. 1141–1144.
[15] A. Q. Liu, M. Tang, A. Agarwal, and A. Alphones, "Low-loss lateral micromachined switches for high frequency applications," *J. Micromech. Microeng.*, vol. 15, no. 1, pp. 157–167, Jan. 2005.
[16] F. Chen, M. Spencer, R. Nathanael, C. Wang, H. Fariborzi, A. Gupta, H. Kam, J. Jeon, T.-J. K. Liu, D. Markovic, V. Stojanovic, and E. Alon, "Demonstration of integrated micro-electro-mechanical (MEM) switch circuits for VLSI applications," in *Proc. IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, Feb. 2010, pp. 150–151.
[17] W. Weaver, Jr., S. P. Timoshenko, and D. H. Young, *Vibration Problems in Engineering*, 5th ed. New York: Wiley, 1990.
[18] S. P. Timoshenko and J. M. Gere, *Mechanics of Materials*. Pacific Grove, CA: Brooks/Cole, 2001.
[19] D. Lee, V. Pott, H. Kam, R. Nathanael, and T.-J. K. Liu, "AFM characterization of adhesion force in MEM relays," in *Proc. Int. Conf. MEMS*, 2010, pp. 232–235.
[20] R. K. Gupta and S. D. Senturia, "Pull-in time dynamics as a measure of absolute pressure," in *Proc. Int. Conf. MEMS*, 1997, pp. 290–294.
[21] K. Y. Yasumura, T. D. Stowe, E. M. Chow, T. Pfafman, T. W. Kenny, B. C. Stipe, and D. Rugar, "Quality factors in micron- and submicron-thick cantilevers," *J. Microelectromech. Syst.*, vol. 9, no. 1, pp. 117–125, Mar. 2000.
[22] D. W. Carr, S. Evoy, L. Sekaric, H. G. Craighead, and J. M. Parpia, "Measurement of mechanical resonance and losses in nanometer scale silicon wires," *Appl. Phys. Lett.*, vol. 75, no. 7, pp. 920–922, Aug. 1999.
[23] D. Markovic, V. Stojanovic, B. Nikolic, M. A. Horowitz, and R. W. Brodersen, "Methods for true energy–performance optimization," *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1282–1293, Aug. 2004.
[24] B. Nikolic, "Design in the power-limited scaling regime," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 71–83, Jan. 2008.
[25] D. Marković, "A power/area optimal approach to VLSI signal processing," Ph.D. dissertation, UC Berkeley, Berkeley, CA, May 2006.
[26] V. Stojanovic, D. Markovic, B. Nikolic, M. A. Horowitz, and R. W. Brodersen, "Energy–delay tradeoffs in combinational logic using gate sizing and supply voltage optimization," in *Proc. 28th ESSCIRC*, Sep. 2002, pp. 211–214.
[27] V. Zyuban, D. Brrok, V. Srinivasan, M. Gschwind, P. Bose, P. N. Strenski, and P. G. Emma, "Integrated analysis of power and performance for pipelined microprocessor," *IEEE Trans. Comput.*, vol. 53, no. 8, pp. 1004–1016, Aug. 2004.
[28] R. Broderson, M. Horowitz, D. Markovic, B. Nikolic, and V. Stojanovic, "Methods for true power minimization," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2002, pp. 35–42.
[29] V. Zyuban and P. Strenski, "Unified methodology for resolving power–performance tradeoffs at the microarchitectural and circuit levels," in *Proc. ISLPED*, Aug. 2002, pp. 166–171.
[30] R. Nathanael, V. Pott, H. Kam, J. Jeon, E. Alon, and T.-J. K. Liu, "Four-terminal-relay body-biasing schemes for complementary logic circuits," *IEEE Electron Device Lett.*, vol. 31, no. 8, pp. 890–892, Aug. 2010.
[31] S. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi, "A tutorial on geometric programming," *Optim. Eng.*, vol. 8, no. 1, pp. 67–127, 2007.
[32] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE J. Solid-State Circuits*, vol. 50, no. 9, pp. 1778–1786, Sep. 2005.
[33] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SSC-9, no. 5, pp. 256–268, Oct. 1974.
[34] M. L. Roukes, "Nanoelectromechanical systems," in *Proc. Solid-State Sens. Actuator Workshop, Tech. Dig.*, Jun. 2000, pp. 367–376.
[35] B. D. Jensen, K. Huang, L. L. W. Chow, and K. Kurabayashi, "Adhesion effects on contact opening dynamics in micromachined switches," *J. Appl. Phys.*, vol. 97, no. 10, p. 103 535, May 2005.
[36] R. Holm and E. Holm, *Electric Contacts; Theory and Application*, 4th ed. Berlin, Germany: Springer-Verlag, 1967.
[37] G. Rubio-Bollinger, S. R. Bahn, N. Agraït, K. W. Jacobsen, and S. Vieira, "Mechanical properties and formation mechanisms of a wire of single gold atoms," *Phys. Rev. Lett.*, vol. 87, no. 2, p. 026 101, Jul. 2001.
[38] T. J. Cheng and S. A. Bhave, "High-Q, low impedance polysilicon resonators with 10 nm air gaps," in *Proc. Int. Conf. MEMS*, 2010, pp. 695–698.
[39] A. P. Chandrakasan, D. C. Daly, D. F. Finchelstein, J. Kwong, Y. K. Ramadass, M. E. Sinangil, V. Sze, and N. Verma, "Technologies for ultradynamic voltage scaling," *Proc. IEEE*, vol. 98, no. 2, pp. 191–214, Feb. 2010.
[40] International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://public.itrs.net
[41] G. M. Rebeiz, *RF MEMS, Theory, Design and Technology*. New York: Wiley, 2003.

Hei Kam (S'04–M'10) received the B.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley (UCB), in 2004 and 2009, respectively. His doctoral research focused on the research and development of reliable micro/nanorelays for ultra-low-power digital logic.

From December 2009 to June 2010, he was a Postdoctoral Research Fellow with UCB. He is currently with Intel Corporation, Hillsboro, OR. His research interests include novel logic devices with subthreshold swing steeper than 60 mV/dec, microelectromechanical systems, and low-temperature fabrication processes.

Dr. Kam has been the recipient of the Ford Motor Scholarship and the Graduate Assistance in Areas of National Need fellowship.

Tsu-Jae King Liu (SM'00–F'07) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1984, 1986, and 1994, respectively.

In 1992, she joined the Xerox Palo Alto Research Center, Palo Alto, CA, as a Member of the research staff, where she worked on the research and development of polycrystalline-silicon thin-film transistor technologies for high-performance flat-panel displays. In August 1996, she joined the faculty of the University of California, Berkeley, where she is currently a Conexant Systems Distinguished Professor of electrical engineering and computer sciences and the Associate Dean for Research in the College of Engineering. Her current research activities are in nanometer-scale integrated-circuit devices and technology, as well as materials, processes, and devices for integrated microsystems.

Dr. Liu was the recipient of the DARPA Significant Technical Achievement Award (in 2000) for the development of FinFET and the IEEE Kiyo Tomiyasu Award (in 2010) for her contributions to nanoscale MOS transistors, memory devices, and MEMS devices. She has served on several committees for many technical conferences, including the International Electron Devices Meeting and the Symposium on VLSI Technology. She was the Editor for the IEEE Electron Devices Letters from 1999 to 2004.

Vladimir Stojanović (S'96–M'05) received the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 2000 and 2005, respectively, and the Dipl.Ing. degree from the University of Belgrade, Belgrade, Serbia, in 1998.

He was a Visiting Scholar at the Advanced Computer Systems Engineering Laboratory, Department of Electrical and Computer Engineering, University of California, Davis, in 1997–1998. He was also with Rambus, Inc., Los Altos, CA, from 2001 to 2004. He is currently an Emmanuel E. Landsman Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge. His current research interests include the design, modeling, and optimization of integrated systems, from novel switching and interconnect devices (such as NEM relays and silicon photonics) to standard CMOS circuits.

Dr. Stojanović was the recipient of the 2009 NSF CAREER Award.

Dejan Marković (S'96–M'06) received the Dipl.Ing. degree in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 1998 and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley (UC Berkeley), in 2000 and 2006, respectively.

Since 2006, he has been an Assistant Professor with the Electrical Engineering Department, University of California, Los Angeles. His current research is focused on integrated circuits for emerging radio and healthcare systems, design with post-CMOS devices, optimization methods, and CAD flows.

Dr. Marković was the recipient of the CalVIEW Fellow Award in 2001 and 2002 for excellence in teaching and mentoring of industry engineers through the UC Berkeley distance learning program and the 2007 David J. Sakrison Memorial Prize from the Department of EECS, UC Berkeley. In 2004, he was the corecipient of the Best Paper Award from the IEEE International Symposium on Quality Electronic Design.

Elad Alon (S'02–M'06) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 2001, 2002, and 2006, respectively.

He has held positions at Sun Labs, Intel, AMD, Rambus, Hewlett Packard, and IBM Research, where he worked on digital, analog, and mixed-signal integrated circuits for computing, test and measurement, and high-speed communications. In January 2007, he joined the University of California, Berkeley (UC Berkeley), as an Assistant Professor of electrical engineering and computer sciences, where he is currently the Codirector of the Berkeley Wireless Research Center. His research focuses on energy-efficient integrated systems, including the circuit, device, communications, and optimization techniques used to design them.

Dr. Alon was the recipient of the IBM Faculty Award in 2008, the 2009 Hellman Family Faculty Fund Award, and the 2010 UC Berkeley Electrical Engineering Outstanding Teaching Award.