Integrated Circuit Design with NEM Relays

Fred Chen[2], Hei Kam[1], Dejan Marković[3], Tsu-Jae King Liu[1], Vladimir Stojanović[2], Elad Alon[1]
[1]Department of EECS, University of California, Berkeley, CA 94720, USA
[2]Department of EECS, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]EE Department, University of California, Los Angeles, CA 90095, USA

***Abstract*—To overcome the energy-efficiency limitations imposed by finite sub-threshold slope in CMOS transistors, this paper explores the design of integrated circuits based on nano-electro-mechanical (NEM) relays. A dynamical Verilog-A model of the NEM relay is described and correlated to device measurements. Using this model we explore NEM relay design strategies for digital logic and I/O that can significantly improve the energy efficiency of the whole VLSI system. By exploiting the low effective threshold voltage and zero leakage achievable with these relays, we show that NEM relay-based adders can achieve an order of magnitude or more improvement in energy efficiency over CMOS adders with *ns*-range delays and with no area penalty. By applying parallelism, this improvement in energy-efficiency can be achieved at higher throughputs as well, at the cost of increased area. Similar improvements in high-speed I/O energy are also predicted by making use of the relays to implement highly energy-efficient digital-to-analog and analog-to-digital converters.**

I. Introduction

Despite the drastic improvements in performance, cost, and energy efficiency brought by CMOS technology scaling over the last 40 years, integrated circuits today are severely limited by their total power consumption. The issue has been exacerbated recently because threshold voltages have already hit the point at which they optimally balance the leakage and dynamic energy consumption of a design. Thus, further supply scaling comes at the expense of per-core performance, and it is this trend that has forced the move to chip-multiprocessors.

Unfortunately, even this parallelism will eventually become ineffective since the energy efficiency achievable by CMOS transistors is limited by their sub-threshold leakage. This is due to the fact that in the sub-threshold regime of operation, an increase in the threshold voltage (V_{th}) decreases the leakage current by exactly the same amount it increases the delay. This makes the energy per operation of a CMOS functional unit level off to a defined minimum level [1] no matter how slowly the circuit is allowed to run. Thus, if a device with significantly improved leakage characteristics (i.e., steeper sub-threshold slope) were available, major improvements in energy efficiency over CMOS could be achieved.

Many researchers are therefore exploring new switching device concepts [2,3] to achieve sub-threshold slopes steeper than the limit set by k_BT/q in field-effect or bipolar junction transistors. However, many of these devices are based on tunneling and hence have very low *on*-state current at low supply voltages. In order to significantly improve upon the energy-efficiency of CMOS, an alternative switching device which offers extremely low *off*-state current together with relatively high *on*-state current at low supply voltages is needed. In other words, the device should behave as closely as possible to an ideal switch.

In this paper we focus on electrostatically actuated mechanical switches (similar to those described in [4]) because they offer nearly ideal switching characteristics (zero leakage, infinitely steep sub-threshold slope) and are more scalable and compatible with conventional micro/nanofabrication processes than relays that are actuated thermally, magnetically, or piezoelectrically. Specifically, we consider a 4-terminal relay design which guarantees that the state of the switch is determined only by the voltage difference between a movable gate terminal and a fixed body terminal. The time required to mechanically displace such a relay from the *off*- to the *on*-state is in the nanosecond range, while the electrical time constant required to charge and discharge the parasitic capacitance of a single relay is on the order of single picoseconds or less.

Despite the large mechanical delay, we show that NEM relays can be useful for a wide range of VLSI applications by re-examining traditional system- and circuit-level design techniques to take advantage of the electrical properties of the device. Unlike CMOS circuit design, logic functions should be implemented as a single complex gate with minimum-sized relays, resulting in significantly reduced logic complexity. We show that for throughputs in the ~100 MOPS range, relay-based circuits can be over 10x more energy-efficient than CMOS designs without any area penalty. Furthermore, by trading off increased area in order to apply parallelism to relay-based functional blocks, these energy-efficiency benefits can be extended to throughputs greater than 1 GOPS with ~6-25x larger area compared to CMOS. We also show that this improved energy efficiency in digital logic can be extended to the I/O's at the same rate. Although the lack of a "saturated" region of operation at first seems like a significant roadblock to analog/mixed-signal design with NEM relays, we show that highly energy-efficient DAC and ADC structures can be built using the relay purely as a switching element.

To further elucidate the characteristics of these switches and substantiate the subsequent simulation results, in Section II we first describe the structure and operation of the relay design used in this paper, followed by the device model used for the circuit evaluations. Subsequently, since digital logic and I/O are key building blocks required to implement a potential NEM relay VLSI system, we next describe in Sections III and IV relay-based design strategies for these components.

II. Overview of NEM Relays

A. Structure and Operation

Fig. 1 shows cross-sectional and top (layout) views of a NEM relay device suitable for VLSI circuits. The metallic conducting *channel* is attached via an insulating gate dielectric to the cantilever *gate* electrode. In the *off*-state ($|V_{gb}| < V_{th}$), an air gap separates the channel from the metallic *source* and *drain* electrodes so that no current can flow. In the *on*-state ($|V_{gb}| \geq V_{th}$), electrostatic force bends the cantilever beam sufficiently to bring the channel into contact with the source/drain electrodes in the dimple regions, providing a conductive path for current to flow. Since the relay switches on abruptly as $|V_{gb}|$ is increased above V_{th}, the I_d-V_{gb}

characteristic of the relay exhibits an extremely steep (nearly infinite) sub-threshold slope.

The relay is designed so that the spring restoring force of the cantilever beam always overcomes the surface forces that attract the dimple contacts to the source/drain electrodes. Thus, even after being actuated, when $|V_{gb}|$ is reduced sufficiently below V_{th}, the switch returns to the *off*-state.

Since electrostatic attraction is ambipolar, the relay can be turned *on* if a sufficiently large positive or negative gate-to-body voltage is applied. This allows the same switch structure to be operated equivalently as an NMOS transistor or a PMOS transistor by appropriately biasing the body terminal (0V for NMOS operation, V_{dd} for PMOS operation).

B. Relay Modeling

Electrostatically actuated beams have been extensively studied and modeled for RF switching applications [5-8], and thus we will only briefly describe their basic behaviors here. As we describe next, these basic equations were used to develop the Verilog-A model for the 4-terminal relay, and this model was used to enable the subsequent simulation study of NEM relay circuit design and energy-performance characteristics.

1) Beam Dynamics and Relay Threshold Voltage

The gate cantilever beam can be modeled as a linear spring-damper–mass system, as shown in Fig. 2. Although this lumped electro-mechanical model is clearly simplified, it provides useful insight for switch design and has been previously shown to closely match experimental results [9-11]. In order to further corroborate this model, we will present measured results from fabricated cantilever beams in Section II.C.

Using this simplified model, the dynamics of the motion of the gate can be described by the equation:

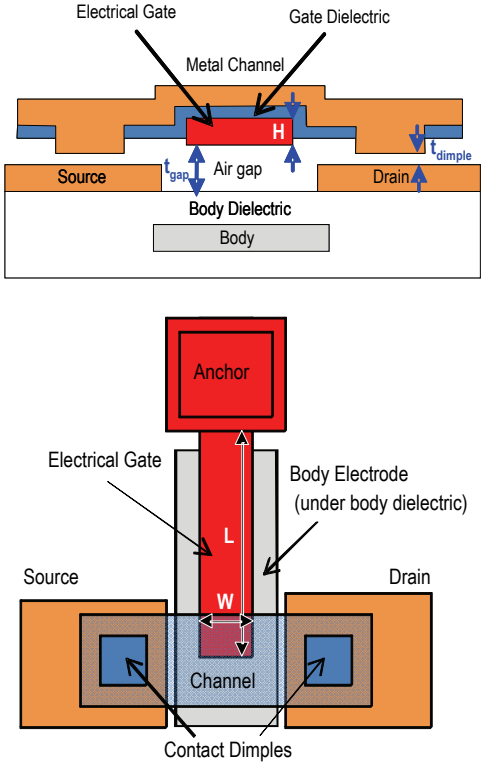

Fig. 1. 4-terminal relay design for VLSI applications: cross-sectional view (top) and plan view (bottom).

$$m\ddot{x} = F_{elec} - b\dot{x} - kx \tag{1}$$

where x is the displacement of the beam from its nominal position, b is the linear damping factor caused by the displacement of air molecules and anchor losses, m is the mass of the beam, and k is the spring constant of the beam.

For a beam of width W, thickness H, and length L (as indicated in Fig. 1), the mass of the beam m is equal to ρWHL, where ρ is the density of the beam material. Similarly, the spring constant is $k=\gamma EW(H/L)^3$, where γ is an empirical constant equal to ~0.25 for a cantilever, and E is the Young's modulus of the beam material.

The relay is actuated by placing a voltage V_{gb} across the gate-to-body capacitance C_{gb}, which can be expressed as:

$$C_{gb} = \frac{\varepsilon_o WL}{d_{eff} - x} \tag{2}$$

where $d_{eff} = t_{gap} + (t_{box}/\kappa_{box})$ is the effective as-fabricated air-gap thickness between the gate and the body, with t_{box} and κ_{box} defined as the physical thickness and relative permittivity of the body dielectric, respectively. In order to reduce the beam travel distance and hence the mechanical delay of the switch, the 4-terminal design employs a smaller air-gap of thickness t_{dimple} in the source/drain contact regions (Fig. 1). The electrostatic force F_{elec} resulting from V_{gb} attracts the gate towards the body, and is equal to:

$$F_{elec} = \frac{\varepsilon_0 (WL) V_{gb}^2}{2(d_{eff} - x)^2} \tag{3}$$

While the electrostatic force increases quadratically with increasing displacement, the spring restoring force $F_{spring} = kx$ (which counteracts the electrostatic force) increases only linearly with displacement. Hence, by setting the dynamic terms in (1) to zero, it can easily be shown that there is a critical displacement equal to $d_{eff}/3$ [12] beyond which F_{elec} is always larger than F_{spring}, causing the gap to close abruptly. This critical displacement has a corresponding value of $|V_{gb}|$ known as the "pull-in" voltage V_{pi}:

$$V_{pi} = \sqrt{\frac{8}{27} \cdot \frac{k d_{eff}^3}{\varepsilon_o (WL)}} = \sqrt{\frac{8}{27} \cdot \frac{\gamma E H^3 d_{eff}^3}{\varepsilon_o L^4}}. \tag{4}$$

As it will impact the I/O circuits described in Section IV, it is important to note here that a relay exhibits hysteretic switching behavior, so that the value of $|V_{gb}|$ required to switch the device *off* can be significantly smaller than the value required to switch it *on*. This is especially the case if the relay is operated in pull-in mode, i.e. if $t_{dimple} > d_{eff}/3$. Simply, once the relay has been pulled-in, the effective gap is smaller than $d_{eff}/3$, making F_{elec} unconditionally larger than F_{spring} at significantly lower $|V_{gb}|$. It is important to note that this hysteresis voltage is also directly impacted by surface forces; since the hysteresis voltage sets a lower limit for V_{dd}, these

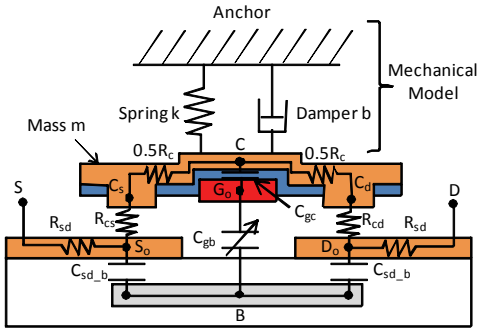

Fig. 2. Relay circuit model implemented in Verilog-A, capturing the device's electro-mechanical behavior as well as its parasitic resistors and capacitors.

surface forces should be minimized.

Note that if t_{dimple} is less than $d_{eff}/3$ (i.e., the relay is not operated in pull-in), then V_{th} is less than V_{pi}. Nonetheless, V_{pi} sets an upper limit for the gate-to-body voltage required to turn on the relay, and thus in this paper we will use $V_{th} = V_{pi}$ as a conservative estimate.

As can be seen from Equation (4), V_{pi} is a strong function of the beam length. Since the beam length is set lithographically, circuit designers can directly tune the threshold voltage of the device.

Due to the relay's abrupt turn-on behavior and the fact that it does not exhibit any leakage current (as long as the contact dimple gap is greater than ~2 nm to avoid significant direct tunneling), the threshold voltage and hence the supply voltage can both be reduced to minimal levels to improve energy efficiency. The minimum V_{th} for a given relay technology is set by the requirement that the spring restoring force is able to overcome surface forces. For the 90 nm relay devices used in the circuit studies of Sections III and IV and a somewhat conservative estimated surface adhesion energy of ~150 μJ/m^2 (assuming appropriate treatment of the contact surfaces [13]), this V_{th} is less than 200 mV.

2) Parasitic Resistances and Capacitances

Since the relay is electrostatically actuated, its input impedance is largely capacitive in nature (like a MOSFET). In the *off*-state, the input capacitance of the relay is dominated by the gate-to-body capacitance C_{gb}. When the relay is in the *on*-state, the capacitance between the gate and the channel C_{gc} also appears directly as input capacitance:

$$C_{gc} = \frac{\kappa_{gateox}\varepsilon_o (WL_{channel})}{t_{gateox}} \quad (5)$$

where t_{gateox} and κ_{gateox} are the physical thickness and relative permittivity of the gate dielectric, respectively, and $L_{channel}$ is the length of the overlap between the gate and the channel.

The *on*-resistance (R_{on}) of the 4-terminal relay is the sum of the source/drain resistances ($2R_{sd}$), the channel resistance (R_c), and the channel-to-source and channel-to-drain resistances (R_{cs} and R_{cd}, respectively). The electrode resistances can be simply approximated as:

$$R_{sd} = \rho_{sd}\frac{L_{sd}}{H_{sd}W_{sd}} \quad \text{and} \quad R_{ch} = \rho_{ch}\frac{L_{ch}}{H_{ch}W_{ch}} \quad (6)$$

where the symbols ρ, L, H, and W represent the sheet resistance, length, thickness, and width of the electrode. The values of these electrode resistances are obviously highly material dependent. However, since these electrodes are all metallic, and since the electrodes intrinsic to the device can be relatively short, these resistances are typically very small (~1 Ω or less).

On the other hand, the parasitic resistances due to the imperfect contacts (R_{cs} and R_{cd}) [14-15] between the channel and the source/drain electrodes can potentially have a much larger value. The contact resistance is determined by material properties and contact conditions, and can be computed by [16]:

$$R_{cd} = R_{cs} = \frac{4\rho\lambda}{3A_r} \quad (7)$$

where A_r is the effective area of the contact, and ρ and λ are the resistivity and electron mean free path of the contact material, respectively. The effective area of the contact (which is typically dominated by asperities) is a function of the loading force—which is approximately the electrostatic force—the material hardness (H), and the deformation coefficient (ξ) at the contact:

$$A_r \approx \frac{F_{elec}}{\xi H} \quad (8)$$

where ξ measures the plasticity of the contact asperities [17], and is <0.3 for elastic deformation, between 0.3 and 0.75 for elastoplastic deformation, and less than 1 for plastic deformation.

The best-known contact material for minimum contact resistance is gold (Au), which has a hardness of 0.2-0.7GPa, a resistivity of 23 nΩ·m, and an electron mean free path of 36 nm. For the 90 nm equivalent devices used in our circuit study, the total contact resistance of a gold-based elastic-contact relay design is estimated to be on the order of 100 Ω even at a low V_{gb} of 0.3 V.

Unfortunately, gold may not be suitable for fabrication of nano-scale features with fine pitch, and other more suitable materials are typically harder and hence have higher contact resistance. For example, if the physical contacts are instead formed with tungsten (W), which has a hardness of 1.1 GPa, resistivity of 55 nΩ·m, and electron mean free path of 33 nm, the estimated total contact resistance for an elastic contact at a V_{gb} of 0.3 V is ~1 kΩ.

In our circuit study we will present results for both the best-case (gold) and worst-case (tungsten) contact materials, including the reduction in contact resistance as a function of supply voltage[1]. Fortunately, the energy-performance of NEM relay-based circuits is found to be relatively insensitive to even this order-of-magnitude variation in possible contact resistance values.

C. Model Verification

In order to validate the model presented in the previous section, we have fabricated simple 2-terminal switches using doped poly-Si cantilever beams of various dimensions (Table I).

TABLE I. DIMENSIONS OF THE FABRICATED CANTILEVER BEAMS USED TO CONFIRM THE RELAY MODEL.

Technology:	1μm	2μm	5μm
Beam width, W	1μm	2μm	5μm
Beam thickness, H	200nm	400nm	1μm
Gap thickness, g	100nm	200nm	0.5μm
L/W	{4,5,6,…,18,19,20}		

As expected from the model, the measured data plotted in Fig. 3 shows that V_{pi} decreases quadratically with increasing beam length.

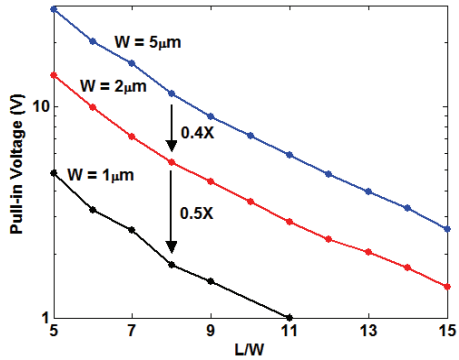

Fig. 3. Measured relay pull-in voltages (V_{pi}) as a function of beam length for various beam widths.

[1] At a V_{dd} of 1 V, the estimated contact resistances are ~10 Ω and 100 Ω for Au and W, respectively. Note that although this voltage-dependent change in contact resistance may enable relay circuits based on traditional analog CMOS topologies, relays do not have a saturation region of operation. Thus, voltage gains significantly greater than unity would require a large drain bias voltage, limiting the benefits and practicality of such relay-based analog circuits.

Furthermore, the data confirms that if H, d_{eff}, and L are all scaled by a linear factor S, the pull-in voltage V_{pi} and hence the threshold voltage V_{th} also scale by S. This decrease in V_{th} would allow a corresponding decrease in the supply voltage, and thus relay scaling can achieve very similar benefits to classical constant-electric-field MOSFET scaling [18].

In addition to verifying that V_{pi} behaves as predicted by the model, we also measured the mechanical delay of the fabricated cantilever beams. As with previous studies, these measurements correlated well with the predictions of the lumped electro-mechanical model [11,19]. For example, a 2 μm wide, 400 nm thick, and 12 μm long beam had a measured pull-in time of 150 ns, while the model predicts a value in the range from 188 ns (high Q) to 393 ns (low Q). Despite the fact that it over-estimates the delay, for the results presented in Sections III and IV we have chosen to conservatively base our delay results on the simplified model.[2]

Unless otherwise specified, our subsequent circuit simulations assume the default device dimensions and material values provided in Table II. These values are representative of a potential NEM relay-based technology whose minimum lithographically-defined dimension is 90 nm. Note that the beam length of 2.3 μm was chosen to minimize V_{pi} while ensuring that the spring restoring force overcomes the predicted surface forces.

TABLE II. SUMMARY OF THE RELAY DEVICE MODEL PARAMETERS USED FOR THE VLSI CIRCUIT STUDY OF SECTIONS III AND IV.

Beam width (W)	90 nm
Beam length (L)	2.3 μm
Beam thickness (H)	90 nm
Actuation air-gap thickness (t_{gap})	10 nm
Source-drain metal thickness (t_{sd})	90 nm
Channel thickness ($t_{channel}$)	90 nm
Contact dimple area	90 nm x 90 nm
Contact dimple gap thickness (t_{dimple})	5 nm
Gate oxide thickness (t_{gateox})	10 nm
Body dielectric thickness (t_{bodyox})	2 nm
Young's Modulus (E)	160 GPa
Polysilicon density (p_{si})	2330 kg m^{-3}
Body oxide permittivity (e_{bodyox})	7.9
Pull-in voltage (V_{pi})	194 mV
Max. gate-body capacitance ($C_{gb,max}$)	0.35 fF
Max. *On*-state resistance (Au relay)	100 Ω
Max. *On*-state resistance (W relay)	1 kΩ

III. NEM RELAYS FOR DIGITAL LOGIC

Since NEM relays can be made complementary with the appropriate choice of body voltage, many of the logic styles used in CMOS can be directly extended to relay-based designs. However, the electrical characteristics and behavior of NEM relays are significantly different than that of CMOS transistors. Thus, as we will describe next, an optimized relay-based design will make use of gate and logic network topologies very different from the same logic function optimized for a CMOS implementation.

The delay of a single CMOS transistor is largely set by the time it takes to charge or discharge the output capacitance. In contrast, for a relay, the delay is dominated by the time it takes to mechanically displace the beam. Specifically, the mechanical delay of a relay is typically ~10 ns, while even with 1 kΩ contact resistance, the electrical time constant of a relay driving the gate of another identical relay is less than 1 ps.

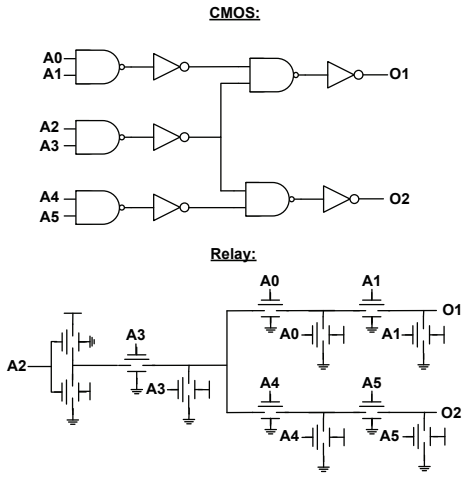

Fig. 4. CMOS to relay logic mapping.

Given this large ratio between the mechanical and electrical delays of the relay, an optimized relay-based design would arrange for all mechanical movement to happen simultaneously – even if this drastically increases the *on*-resistance of the logic gate. In other words, re-buffering the signal would incur an additional mechanical delay that is typically large compared to the electrical delay, so relay-based digital circuits should be comprised of single-stage complex gates, as shown in Fig. 4.

Of course, as the number of series relays in a given gate is increased, the electrical time constant of this gate will increase quadratically. In a gold-based NEM relay with low contact resistance, the number of series relays required for the electrical delay to even approach the mechanical delay is significantly larger than what would be required for any practical function of interest. Even with a tungsten-based relay's worst-case contact resistance of 1 kΩ, the gate could include ~200 device stacks before the electrical delay increases above 10ns. However, if a complex relay gate must drive a significant load capacitance (e.g., due to wiring or a large number of parallel relays) at each one of its nodes, it may become beneficial to split the function into two stages by buffering the outputs. With this buffering, the delay of the entire logic block would be roughly only doubled even for capacitive loads on the order of 1 pF.

A. Energy-Performance Comparisons with CMOS

Unlike circuits built with emerging devices that are still essentially field-effect in nature, relay-based circuits are implemented in a significantly different fashion than CMOS circuits. Thus, any comparisons between the energy-performance characteristics of CMOS designs and NEM relay designs must be made at the circuit level (rather than at the device level). For example, although the mechanical delay of a relay may be significantly larger than that of a CMOS inverter, a complete CMOS logic block will typically require 10-20 gate delays, whereas the relay-based circuit may comprise of only a single stage and hence require only one mechanical delay.

Thus, in order to illustrate the potential benefits of NEM relays, we describe a complete 32-bit relay-based adder, and compare its energy-performance with that of an optimized CMOS implementation [20].

[2] There are several possible explanations for the discrepancy in the delay, including that it is largely only the tip of the beam which experiences the full deflection. However, fully modeling such behaviors would significantly impact the run-time of the circuit simulations, motivating our continued use of the simplified model.

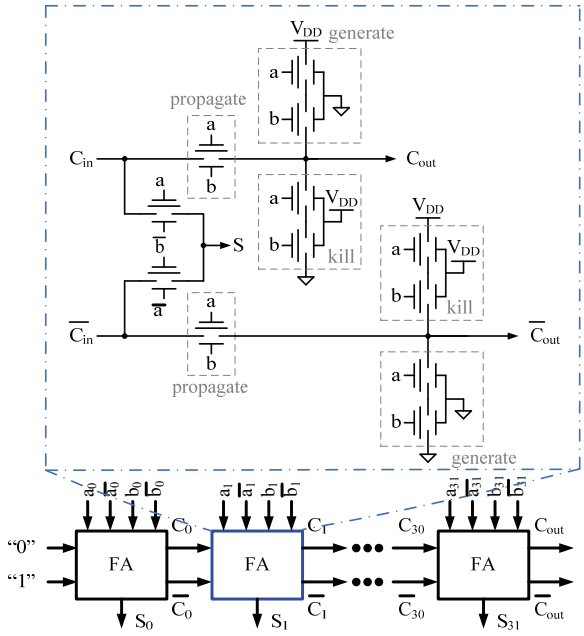

Fig. 5. Schematic of a 32-bit Manchester carry chain adder using a full adder cell implemented as a single compound gate with NEM relays.

1) *Relay-based Adder*

A full adder cell implemented with NEM relays is shown in Fig. 5. Since the relay is actuated whether the gate potential is positive or negative relative to the body (i.e., when the logical polarities are opposite), we exploited the availability of the back-gate as an input to implement the XOR of two signals (for both the *propagate* and *sum* calculations) in a very compact fashion. Sum is implemented as a wired XOR gate, while complementary carry signals are used to avoid an additional mechanical delay that might be required to invert the incoming carry signal. As a result, only 12 relays are used to implement a full adder cell.

To implement the complete adder, this full adder cell is used in a ripple carry configuration also shown in Fig. 5. Overall, the structure implements a complete 32-bit add function in one compound gate.

In order to benchmark the NEM relay-based adder shown in Fig. 5, we have compared it to a 32-bit CMOS Sklansky adder [21]. Specifically, since it was identified as the optimal adder topology across a wide range of energies and delays in [20], we will use a static CMOS Sklansky adder as a reference for all comparisons. The energy per operation of the CMOS adder is based upon the results from [20], and for the relay adder we estimated the average switching activity factor from simulation.

Fig. 6 shows the energy-throughput tradeoffs in CMOS and NEM relay adders, where the adders have been designed to drive a load capacitance of either 25 fF or 100 fF. In order to isolate the energy dissipation of the adder itself and enable comparisons with respect to driving different loads, this plot shows the energy per operation (E_{op}) dissipated in the adder alone[3]. However, the delay

[3] If the load energy is included, the 100 fF area overhead curves in Fig. 7 would shift up to match the 25 fF curves. For a 100 fF load and V_{dd} ranging from 0.5–1 V, the CMOS adder adds 160-640 fJ/op, while the relay adder, with V_{dd} ranging from 0.25–0.9 V, would add 40–520 fJ/op to the E_{op} values in Fig. 6. Thus, for large loads the E_{op} of the relay adder is dominated by the load energy. The impact of the load energy can be reduced by making use of parallelism to lower the supply of the relay adder. For a 100 fF load, this would require an additional area overhead of ~2-3x to maintain the same improvement in energy efficiency.

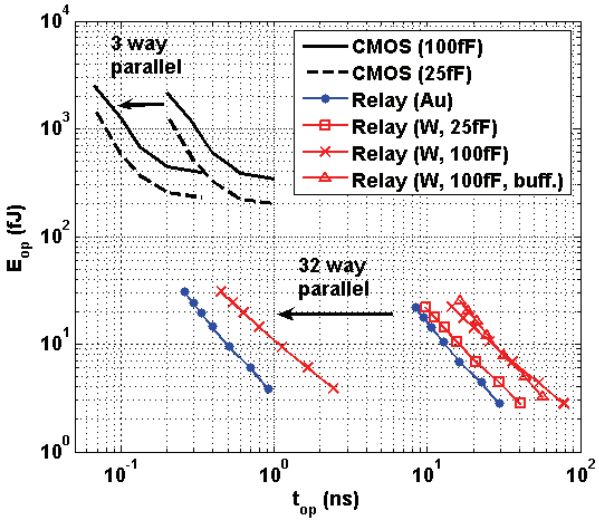

Fig. 6. Energy-throughput comparison of 32-bit adders after sizing and V_{dd} scaling: static CMOS Sklansky design [21] versus NEM relay adder. The area of all of the relay adders is slightly less than 500 μm^2. At per-unit delays of ~1 ns or more, the area of the CMOS adder driving 25fF of load is also ~500 μm^2, while the area of the adder driving 100 fF is ~850μm^2. Applying parallelism to the NEM relay designs shifts their performance to the GOPS range while maintaining their energy advantage over CMOS at a fixed cost in area.

of all of the adders includes the effects of driving the load capacitance at each of their outputs.

The load capacitance has essentially no effect on the delay of gold-based relays with low contact resistance. For the tungsten-based relay adders, driving the load capacitance directly adds at most 12.5 ns and 50 ns to the delay for 25 fF and 100 fF of load, respectively. As shown in Fig. 6, when the contact resistance is large (e.g. at low supply voltages) and a significant load is present at the output, adding a single buffer to drive the adder outputs is the more energy-efficient approach.

The energy of the CMOS adder reaches its minimum point [1,20] for delays above ~1 ns. Thus, even with high contact resistance, at delays of ~10-20 ns, a single relay adder would offer an improvement of ~10x or more in energy efficiency within the same area (~500 μm^2) as the CMOS adder. Clearly, applications requiring throughputs of ~50-100 MOPS or less would immediately benefit from such a relay technology.

The improved energy-efficiency offered by relays can also be

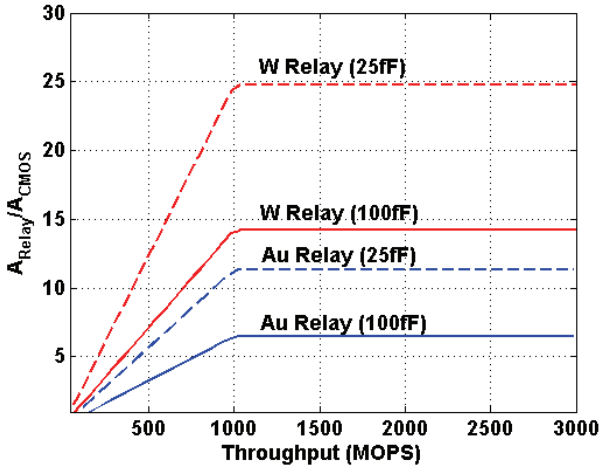

Fig. 7. Area-throughput tradeoffs in relay adders targeting an E_{op} of 20 fJ in comparison to CMOS adders with 25 fF or 100 fF of load.

extended to higher throughputs by trading-off increased area to make use of parallelism (also indicated in Fig. 6). As shown in Fig. 7, at a constant energy-efficiency improvement of ~10x, an area overhead of 5x over the 100 fF CMOS adder brings the throughput of the relay adders up to ~770 MOPS (for gold contacts) and 350 MOPS (for tungsten contacts). For this same 5x increase in area relative to the 25 fF CMOS adder, the relay adders achieve ~440 MOPS (Au) and ~200 MOPS (W). Note that as previously stated, the energy/op shown in Fig. 6 and used to calculate the area overheads in Fig. 7 excludes the energy needed to drive the external load.

Further improvements in throughput at the same CMOS-to-relay E_{op} ratio come at a near-linear scaling in area overhead until the GOPS range. Beyond 1 GOPS, the CMOS adder would then also need to be parallelized in order for each adder to maintain its peak energy efficiency. Thus, at throughputs of 1 GOPS or above, the area overhead of the NEM relay adder is limited to as low as 6.5x for gold relays compared to a CMOS adder loaded with 100 fF, and at most 25x for tungsten-based relays versus a CMOS adder loaded with 25 fF.

In summary, NEM relay adders can improve the energy-efficiency by an order of magnitude across a relatively wide range of delays/throughputs. At low throughputs (~50-100 MOPS), the areas of the relay adders are identical to or even lower than the CMOS designs, and higher throughputs (over 1 GOPS) can be achieved at ~6-25x area overhead while maintaining essentially the same energy efficiency.

IV. NEM Relays for High-Speed I/O

Having shown the potentially significant benefits of NEM relays for digital computation, ensuring that the energy-efficiency of the complete system will enjoy similar improvements requires examination of other key components, like memory and I/O. While challenges to implementing dense memories out of NEM relays clearly exist, given the abrupt switching behavior of these devices, the implementation of efficient analog and mixed-signal building blocks critical to I/O's is significantly less clear. Thus, in this paper we focus on how to implement digital-to-analog (DAC) and analog-to-digital converters (ADC) for the purposes of signal generation and reception.

Many of the ADC and DAC architectures used in CMOS rely on the ability to operate transistors in their saturation region, providing high output impedance and the ability to generate relatively linear voltage gain with low drain-source voltage. For NEM relays, there is no such mode of operation, and thus the challenge is to build mixed-mode circuits using these relays despite their ineffectiveness in performing traditional analog processing.

A. DAC Design

Fig. 8 shows one implementation of a DAC inspired by an existing CMOS design [22] that is suitable for use as an I/O transmitter. In this example, each inverter/driver is driven by a thermometer encoded input where $N=2^k-1$ and k is the bit resolution of the DAC. Each inverter is composed of a NEM relay-based inverter followed by a resistor; the resistor is necessary to provide both a constant controlled termination (R/N) and a means for intermediate voltage generation. It is also important to observe that the inputs of the NEM relay DAC can operate at a core voltage that is independent of the I/O voltage.

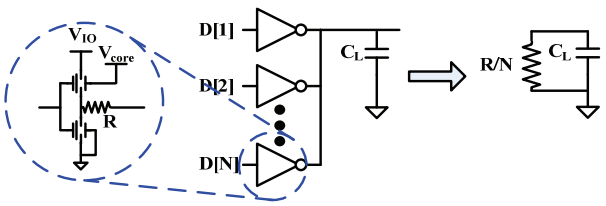

Fig. 8. DAC topology, schematic and equivalent circuit

Assuming that the contact resistance of the relays is unpredictable and varies over a wide range, for the DAC to operate as described, R should be at least an order of magnitude greater than the worst case expected contact resistance. For gold-based contacts this does not present a stringent constraint, but for tungsten-based contacts R would have to be at least 10 kΩ to ensure at worst a 10% error in any single DAC element.

The DAC power is dominated by the current needed to drive the output, and the worst-case power can be related to the resolution and bandwidth by (9):

$$P_{DAC} = \frac{V_{IO}^2 N}{4R} = \frac{V_{IO}^2}{4} BW \cdot C_L,\ BW = \frac{N}{RC_L} \tag{9}$$

The power for the DAC is largely independent of the DAC's resolution, but rather is determined by specifying the termination resistance (or equivalently, the bandwidth) for a given output load. For example, a DAC with a 1 pF load, a voltage swing (V_{IO}) of 200 mV, and a desired bandwidth of 1 GHz will require about 62.8 µW of power. Similarly, a 50-Ohm termination will cost 200 µW. Assuming that we always reduce the output bandwidth to our desired signaling rate, or else use parallelism to signal on the channel up to its bandwidth, the energy per bit is simply:

$$E_{BIT} = \frac{\pi}{2} V_{IO}^2 C_L \tag{10}$$

For the example given above, this translates to 62.8 fJ/bit. This is more than an order of magnitude smaller than the transmitter power reported in [23]. While this energy estimate does not include the switching energy of the NEM relays, Section II showed that the switching capacitance per relay is roughly 0.35 fF, which would require nearly 3000 relays switching at the I/O voltage to have the equivalent switching capacitance as a 1 pF load.

B. ADC Design

As the energy-efficiency of NEM relay-based circuits can be maximized in topologies which leverage parallelism, it makes sense to pursue a Flash ADC topology for signal reception. Thus, we next present a Flash ADC design that is compatible with the NEM relay and shows significant potential energy reduction over CMOS-based ADCs.

Two distinct characteristics of relays (both of which were pointed out in Section II) must be taken into account when designing a relay-based Flash ADC: 1) NEM relays exhibit hysteresis with a larger turn-*on* threshold (V_{pi}) than turn-*off* threshold (V_{po}); 2) The relay will be actuated in response to voltage inputs that are *either* above *or* below its body voltage. This second

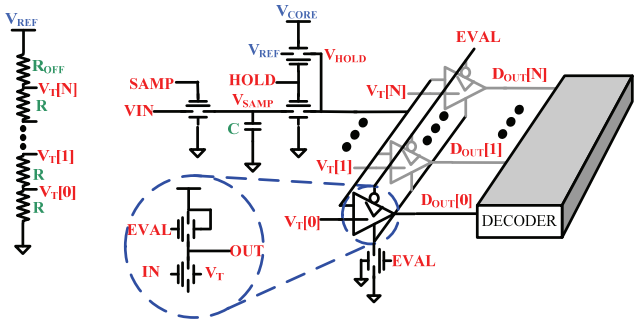

Fig. 9. Proposed ADC circuit diagram

characteristic presents a decoding challenge, as it implies that the response of relay-based "comparators" to the input voltage will be non-monotonic.

Fig. 9 shows the resulting ADC architecture that addresses the NEM relay characteristics described above. The front-end of the ADC is a sample-and-hold circuit similar to those used in CMOS, and the Flash converter consists of a bank of dynamic inverters (or buffers) with varying body-bias thresholds generated by a resistor string. These dynamic inverters act as (absolute value) comparators and their outputs are decoded to produce the final digital code.

Fig. 10 shows the timing waveforms from a Spectre® simulation of the ADC design during a single conversion. The conversion begins with the sampling relay tracking the input voltage one turn-*on* delay, $t_{D,ON}$ (~40 ns in this case), after *SAMP* goes high. The sampling relay then "samples" the input one turn-*off* delay, $t_{D,OFF}$ (~5 ns), after *SAMP* falls. The holding relay then turns on, driving the input to the comparator bank with the sampled input voltage.

Shortly afterwards, *EVAL* rises to activate the shared footer relay for the comparator bank. Prior to activating the comparators for evaluation, the comparator outputs are all initially pre-charged and the state of each relay is reset to be on to avoid data dependent offsets due to hysteresis in the comparators. When the footer relay is activated, only the relays whose V_T is within one V_{po} of the sampled input voltage will remain pre-charged. This is because all voltages either above or below the back-gate bias by more than one V_{po} will cause the relay to remain on (if its previous state was *on*). We use an offset resistor (R_{OFF}) to set the input range of the conversion and to ensure that the reset voltage, V_{REF}, is more than one V_{pi} above all of the comparator V_T's. Note that there is no single reset voltage that can reset the state of all of the comparators to be *off*.

In the interest of reducing latency, the *HOLD* signal can be driven high as early as $t_{D,ON}-t_{D,OFF}$ before the falling edge of *SAMP* to account for the long turn-on delay in the relay. This allows the hold NEM relay to turn on immediately after the sampling relay turns off. A similar strategy can be employed for *EVAL* in relation to *HOLD*. However, this strategy only decreases the latency of the conversion and not the throughput. Any clock signals that drive complementary gates (e.g. *EVAL* and *HOLD*), will be limited to a minimum period of $2*(t_{D,ON} + t_{D,OFF})$ which then sets the maximum throughput for a single ADC.

Fig. 11 shows a plot of the resulting conversion code versus input voltage for a ramped input for a 6-bit version of the relay-based ADC. As mentioned earlier, for the same input voltage, we will have a range of comparators in the conversion that do not turn *on*. The three lines shown in the plot reflect the upper, lower, and average of this range. Depending on the decoder chosen, any one of these three curves can be used as the output.

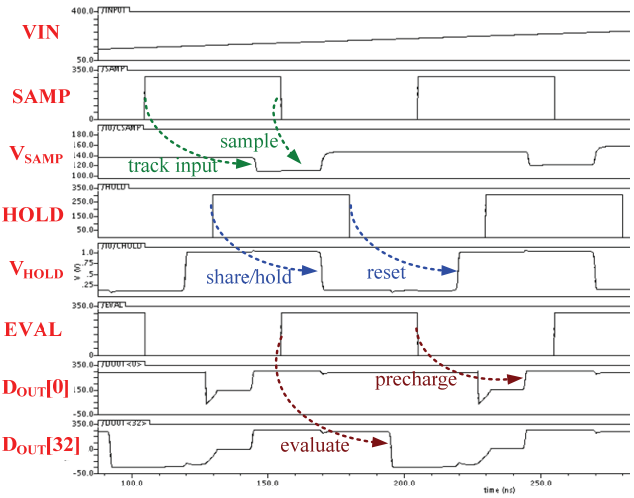

Fig. 10. Timing diagrams for an ADC conversion

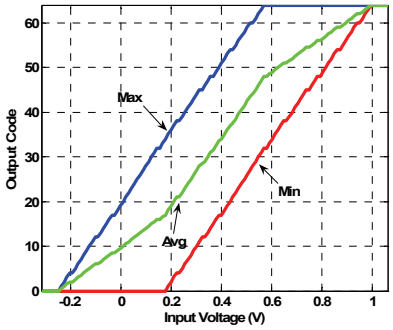

Fig. 11. ADC comparator output range vs. input voltage

The energy consumed by this ADC design is dominated by the reference generation. For the example given, where V_{CORE} = 300 mV, V_{REF} = 1 V, C = 500 fF, R = 4 kΩ, and R_{OFF} = 74 kΩ, the energy consumption in a single cycle is ~350 fJ, out of which 320 fJ is dissipated by the reference supply for threshold generation and reset. This translates to 5.5 fJ/conversion step for a 6-bit 10 MS/s converter, which is over an order of magnitude better than modern CMOS based converters [24]. The energy for this converter, which is quadratically dependent on supply voltage, can be further reduced if the input dynamic range is reduced by scaling down V_{REF}. Additional resolution in the converter comes at a relatively low penalty in energy since much of the energy is consumed in the reference generation.

V. CONCLUSIONS

In this paper we evaluate the use of NEM relays in VLSI applications to extend the energy efficiency of digital systems beyond the limitations imposed by sub-threshold leakage in CMOS technology. To facilitate this evaluation, a simplified Verilog-A model suitable for circuit simulation has been developed and verified with measurements of micron-scale switches.

Unlike CMOS circuits, making optimal use of NEM relays leads to implementing logic functions out of compound gates whose delays are bounded by a single mechanical switching time. Combined with the low parasitic capacitance of the relays and the elimination of *off*-state leakage, this design style allows NEM adders to achieve an order of magnitude lower energy than CMOS adders within the same area at delays of 10-20 ns. Furthermore, at the cost of increased area, parallelism allows these energy-efficiency benefits to be extended to higher throughputs as well.

To maintain this level of energy efficiency across the whole VLSI system, we have also described design techniques enabling NEM relays to implement I/O circuits (DACs and ADCs) again at nearly order of magnitude improvement in energy over currently reported CMOS designs. While many issues such as device fabrication and reliability clearly remain to be solved, taken together these digital and mixed-signal design techniques show that NEM relays may offer a compelling future alternative to CMOS for a wide range of VLSI applications.

ACKNOWLEDGEMENTS

H. Kam, T.-J. King Liu, and E. Alon would like to acknowledge the support of the students, faculty, and sponsors of the Berkeley Wireless Research Center, the National Science Foundation

Infrastructure Grant No. 0403427, the Defense Advanced Research Projects Agency Contract #N66001-07-1-2044, and the Focus Center for Circuit and System Solutions (C2S2), one of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation Program. F. Chen would like to acknowledge the support of the Intel Graduate Fellowship.

REFERENCES

[1] B.H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," IEEE J. Solid-State Circuits, vol. 40, Sept. 2005, pp. 1778-1786.

[2] K. Gopalakrishnan et al., "I-MOS: a novel semiconductor device with a subthreshold slope lower than kT/q," in IEDM Tech. Dig., 2002, pp. 289-292.

[3] W.-Y. Choi et al., "Tunneling Field-Effect Transistors (TFETs) With Subthreshold Swing (SS) Less Than 60 mV/dec," IEEE Electron Device Letters, pp. 743-745, Aug. 2007.

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

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

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

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

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

[9] K. Wang, A.-C. Wong, and C.T.-C. Nguyen, "VHF free-free beam high-Q micromechanical resonators," J. Microelectromechanical Systems, vol. 9, no. 3, Sep. 2000, pp. 347-360.

[10] G. K. Fedder, "Simulation of microelectromechanical systems," Ph.D. dissertation, Department of Electrical Engineering and Computer Sciences, Univ. of California, Berkeley, Sept. 1994.

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

[12] S. D. Senturia, Microsystem Design. Boston: Kluwer Academic, 2001.

[13] R. Maboudian and R. T. Howe, "Critical Review: Adhesion in surface micromechanical structures," AVS Journal of Vacuum Science and Technology B, vol 15, no. 1, Jan. 2007, pp. 1-20.

[14] D. Hyman and M, Mehregany, "Contact Physics of Gold Microcontacts for MEMS Switches," IEEE Trans. on Components and Packaging Technology, vol. 22, no. 3, Sept. 1999, pp. 357-364.

[15] K. Akarvardar et al, "Design Considerations for Complementary Nanoelectromechanical Logic Gates," IEEE International Electron Devices Meeting, pp. 299-302, Dec. 2007.

[16] R.Holm: Electric Contact, Springer Verlag, pp.7-26, 1967

[17] S. C. Bromley and B. J. Nelson, "Performance of microcontacts tested with a novel MEMS device," in Proc. 47th IEEE Holm Conf. Elect. Contacts, 2001, pp. 122–127.

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

[19] L. Castaner, A. Rodriguez, J. Pons, and S. D. Senturia, "Pull-in time-energy product of electrostatic actuators: comparison of experiments with simulation," Sensors and Actuators A: vol. 83, no. 1-3, 22 May 2000, pp. 263-269.

[20] D. Patil et. al., "Robust Energy-Efficient Adder Topologies," in Proc. 18th IEEE Symp. on Computer Arithmetic (ARITH'07).

[21] J. Sklansky. "Conditional-sum addition logic". IRE Trans. on Electronic Computers, EC-9:226–231, 1960.

[22] K.-L.J. Wong, H. Hatamkhani, M. Mansuri, and C.-K.K. Yang, "A 27-mW 3.6Gb/s I/O Transceiver," IEEE Journal of Solid-State Circuits, vol. 39, no. 4, Apr. 2004, pp. .602-612.

[23] R. Palmer et al., "A 14mW 6.25Gb/s Transceiver in 90nm CMOS for Serial Chip-to-Chip Communication," IEEE International Solid-State Circuits Conference, pp. 440-441, Feb. 2007.

[24] B.P. Ginsburg and A.P. Chandrakasan, "500-MS/s 5-bit ADC in 65-nm CMOS With Split Capacitor Array DAC", IEEE J. Solid-State Circuits, vol. 42, no. 4, April 2007, pp. 739 – 747.