INTEGRATION, the VLSI journal 88 (2023) 249–261

Contents lists available at ScienceDirect

Integration, the VLSI Journal

journal homepage: www.elsevier.com/locate/vlsi

3-D coarse-grained reconfigurable array using multi-pole NEM relays for programmable routing

Akash Levy [*,1], Michael Oduoza, Akhilesh Balasingam, Roger T. Howe, Priyanka Raina

Department of Electrical Engineering, Stanford University, Stanford, CA, United States of America

ARTICLE INFO

Keywords:
Digital integrated circuits
Programmable logic devices
Coarse-grained reconfigurable array
Microelectromechanical systems
Monolithic 3-D integration

ABSTRACT

We propose the use of multi-pole nanoelectromechanical (NEM) relays for routing multi-bit signals within a coarse-grained reconfigurable array (CGRA). We describe a CMOS-compatible multi-pole relay design that can be integrated in 3-D and improves area utilization by 40% over a prior design. We then demonstrate a method for placing multiple contacts on a relay that can reduce contact resistance variation by 40× over a circular placement strategy. Additionally, we develop static and dynamic SPICE models (calibrated to accurate finite element models) for predicting circuit-level performance with these devices. Finally, we establish a methodology for integrating these relays into an industry-standard digital design flow. Using our multi-pole relay design, we perform post-layout simulation of a hybrid CMOS-NEMS CGRA processing element (PE) tile in 40 nm technology. We achieve up to 19% lower area and 10% lower power at iso-delay, compared to a CMOS-only CGRA PE tile. The results show a way to bridge the performance gap between programmable logic devices (such as CGRAs) and application-specific integrated circuits using NEMS technology.

1. Introduction

Coarse-grained reconfigurable arrays (CGRAs) have been gaining popularity as specialized programmable logic device (PLD) architectures for applications such as image processing and machine learning [1–3]. CGRAs are similar to field-programmable gate arrays (FPGAs), but improve power, performance, and area (PPA) metrics for many applications through the use of coarse-grained datapaths which: (1) route multiple bits together using multi-bit routing switches, and (2) replace fine-grained look-up tables (LUTs, which perform arbitrary Boolean operations) with processing elements (PEs, which perform common arithmetic operations, e.g., addition, multiplication). It was estimated in [3] that a CGRA can provide 1.4× better energy-efficiency and 3.1× better compute density than an FPGA. However, providing general programmability is still costly: compared to an application-specific integrated circuit (ASIC) solution, a CGRA was estimated to have 6–10× worse energy and area efficiency [3]. This wide gap between the performance of PLDs and ASICs motivates the need for techniques to reduce the reconfigurability overhead in PLDs.

Nanoelectromechanical (NEM) relays are nano-scale mechanical switches that can be electrostatically actuated with a gate. Their properties have been widely studied, and they have been proposed as alternatives/complements to CMOS logic for improving PPA in digital circuits [4]. NEM relays have zero static power dissipation and low ON-state resistance (∼1–10 kΩ experimentally, comparable to that of modern NMOS transistors), but they switch much more slowly than NMOS transistors (nanoseconds to microseconds, rather than picoseconds) [4]. However, when used as statically-configured routing switches, the long switching delay usually does not impact PPA negatively, since they only need to be toggled once, during initialization (programming) of the PLD. NEM relays have abrupt switching behavior, meaning they are either OFF or ON, with roughly zero subthreshold current. They can be designed to exhibit hysteresis in ON/OFF state, and this property can enable them to retain their state and behave as memory [4].

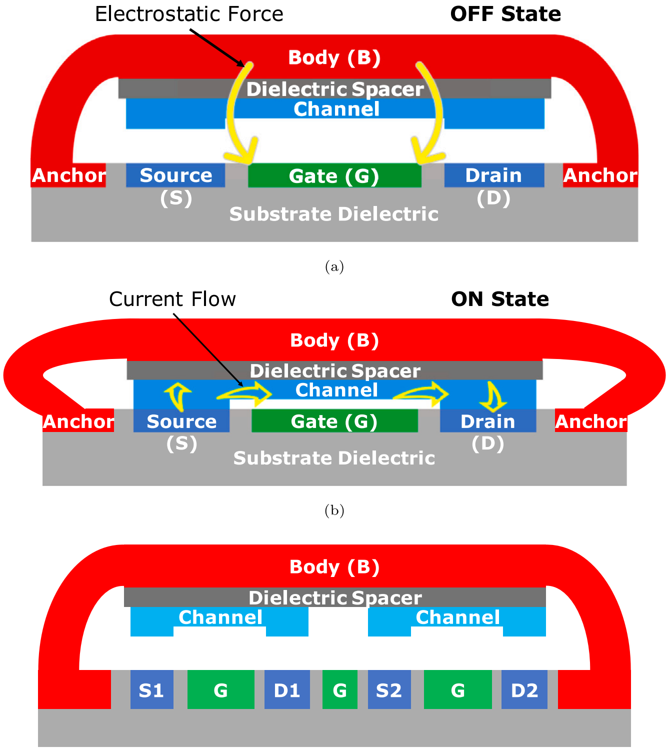
NEM relays may also have multiple poles (we refer to such relays as *multi-pole NEM relays*), which allow multiple signals to be switched by the same gate. Previous work has analyzed the potential benefits of integrating *single-pole* NEM relays into FPGAs as static routing switches and configuration memory and found significant opportunities for improving PPA across several applications [5,6]. However, these studies assumed that *multiple layers of relays* may be monolithically integrated with CMOS, which has not been demonstrated experimentally.

In this paper, we show that integration of *multi-pole* NEM relays into a place-and-routed CGRA design improves PPA, thereby reducing reconfigurability overhead [7]. The major contributions of this paper are:

[*] Corresponding author.
E-mail address: akashl@stanford.edu (A. Levy).

[1] Partially supported by the NSF Graduate Research Fellowship Program, United States.

https://doi.org/10.1016/j.vlsi.2022.10.001
Received 9 June 2022; Received in revised form 29 August 2022; Accepted 3 October 2022
Available online 13 October 2022



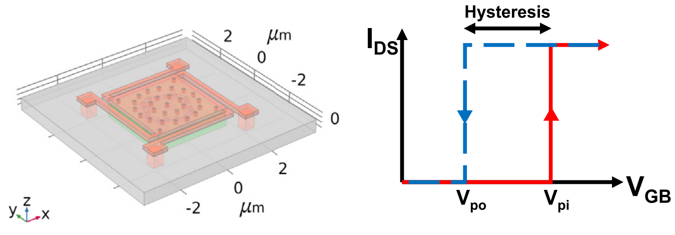
(a)

(b)

(c)

Fig. 1. Schematic illustration of NEM relay side view in (a) OFF state, and (b) ON state. The electric potential between body and gate controls current flow from source to drain. (c) Multi-pole NEM relay side view. The D, S, G, B terminals correspond to drain, source, gate, and body, respectively. Note that in (c) there are multiple source–drain pairs, but only one gate.

1. A multi-pole NEM relay design integrated on top of CMOS back-end-of-line (BEOL) circuitry, featuring *a single layer* of relays which improve area utilization by 40% over a prior design reported in [8].
2. An iterative finite element modeling (FEM) method to optimize multi-pole contact placement, reducing expected contact resistance variation by > 40× over a circular placement.
3. Static and dynamic compact SPICE models for evaluating multi-pole NEM relays in digital and analog circuit designs.
4. Analysis of a body biasing scheme to enable compatibility with typical 40 nm CMOS voltage levels.
5. Methodology to integrate multi-pole NEM relays into a standard digital design flow.
6. Comparison of NEMS multiplexer PPA with CMOS multiplexer PPA when used as a static router, indicating $5.8\times-27.7\times$ lower front-end-of-line (FEOL) area (at the expense of increased BEOL area for vias), 0–27.5% lower switching energy, and delay that improves by up to $4.22\times$ for small output capacitance load (but incurs a penalty at high capacitance load)
7. Design of a hybrid CMOS-NEMS CGRA that integrates multi-pole NEM relays into the PE tile of a CGRA, achieving 19% lower area and 10% lower power at iso-delay.
8. Analysis of multi-pole NEM relay scaling to sub−40 nm technology nodes.



(a)

(b)

Fig. 2. (a) Top view of a NEM relay. The relay moves up and down in the *z*-direction when a voltage is applied to its gate. (b) I_{DS} vs. V_{GB} for a NEM relay.

2. NEM relay background

In this work, we focus on planar vertically-actuated NEM relays [9,10]. These relays can be fabricated at relatively low temperatures (< 450 °C) and use "clean" materials available in foundries today (e.g., poly-SiGe, Al_2O_3, W), making them suitable for inclusion in the back-end-of-line (BEOL) after transistor fabrication. Additionally, these relays have low switching voltage and low contact resistance—both desirable properties for integration with silicon CMOS circuits. We examine the case where relays are laid out in a single layer on top of a CMOS circuit. Having a single layer of relays enables cost-effective fabrication, reducing the number of masks required and eliminating the need to account for any further processing steps. More details on fabrication steps for this type of relay can be found in [9,10].

Fig. 1 shows a 4-terminal (4T) NEM relay. Its four terminals (*source*, *drain*, *body*, *gate*) have similar functions as the four terminals of a MOSFET but make use of different physical mechanisms. The electric potential between the gate (G) and the body (B) (V_{GB}) results in electrostatic attraction between the relay and the gate. When V_{GB} exceeds a critical value, called the pull-in voltage (V_{pi}), the elastic force of the relay can no longer balance the electrostatic force, and the relay body collapses toward the gate. Once this happens, the *conductive channel* touches down on the source (S) and drain (D). This state is referred to as the *ON state* (Fig. 1(b)), and the source and drain are electrically connected, enabling current flow (I_{DS}) between them. As V_{GB} is decreased, the channel disconnects from the drain at another critical voltage, called the pull-out voltage (V_{po}), which is smaller than V_{pi} because the body electrode is closer to the gate [4]. Adhesion or "stiction" between the channel and the source/drain also contributes to $V_{po} < V_{pi}$ [11]. When the channel is disconnected, the relay is in the *OFF state* (Fig. 1(a)), and no current can flow between source and drain. Fig. 2(a) shows the top view of a NEM relay and Fig. 2(b) shows its typical $I_{DS} - V_{GB}$ characteristic. Since $V_{po} < V_{pi}$, a value of V_{GB} such that $V_{po} < V_{GB} < V_{pi}$ allows the relay to retain its state (hysteresis) (Fig. 2(b)).

NEM relays may also have multiple poles that connect independent source–drain pairs (Fig. 1(c)). Working demonstrations of multi-pole relays have been reported [12], but strategies to make use of multiple poles at a system level have not been studied in detail, to our knowledge.

3. Multi-pole NEM relay design and modeling

3.1. Relay layout

We first develop a parametric NEM relay layout, shown in Fig. 3(a), that maximizes the relative size of the relay body (for larger electrostatic force) while maintaining a low enough spring constant to enable CMOS-compatible pull-in/pull-out voltages (< 5 V). Parameters include the various lateral and thickness dimensions of the relay, given in Table 1.

The relay consists of four cantilever *beams* that connect to the square-shaped *electrostatic plate*. The electrostatic plate generates the force needed to snap the relay down into the ON state (shown in Fig. 3(d)). The purpose of the beams is to provide a sufficient spring force to restore the relay back to the OFF-state (Fig. 3(c)) when $V_{GB} < V_{po}$. There are also release holes, which allow the sacrificial material

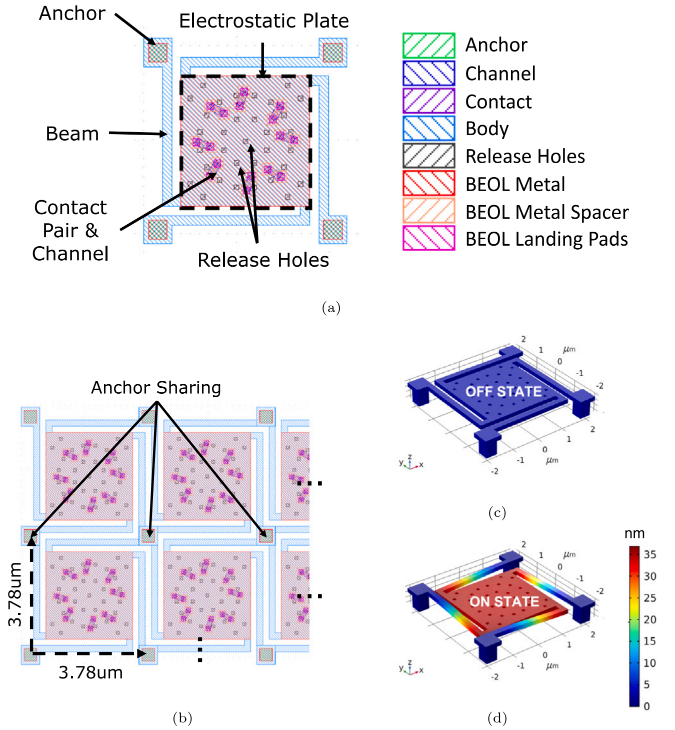


Fig. 3. Relay layout. (a) Top view layout of a single relay, showing the different components and layers. (b) Mosaic of relays, illustrating how anchors may be shared across relays. (c) A top view perspective of a NEM relay in the OFF state, and (d) the ON state. The color indicates the downwards z-displacement of the relay.

Table 1
Important NEM relay dimensions.

Lateral dimensions (x–y plane)		
Length of electrostatic plate (square)	L_{plate}	2780 nm
Side length of contact (square)	L_{cont}	150 nm
Side length of release hole (square)	L_{hole}	100 nm
Side length of anchor (square)	L_{anc}	400 nm
Width of cantilever beam	L_{beam}	200 nm
Gap between landing pad and gate	L_{DG}	200 nm
Gap between cantilever beam and body	L_{Bbeam}	200 nm
Thickness dimensions (z-direction)		
Thickness of electrostatic plate	t_{plate}	120 nm
Thickness of contact	t_{cont}	25 nm
Thickness of dielectric spacer	t_{sp}	30 nm
Thickness of conductive channel	t_{chan}	10 nm
Gap between relay plate and substrate	g_{act}	60 nm

under the relay to be removed during fabrication. These holes also allow gas under the relay to escape and hence affect the relay damping behavior—in our design, these holes are placed in concentric circles. The relay layout is designed to be invariant under 90° rotations in the xy-plane.

To maximize area efficiency of our relays while still using CMOS-compatible voltages (< 5 V), we use the following techniques: (1) We share anchors and tessellate the relays in a mosaic structure, as shown in Fig. 3(b). This reduces the relay pitch and minimizes unused area between relays. (2) We use folded beams to reduce the relay footprint. We define *area utilization* as the ratio of the electrostatic plate area to the total relay area. Under this definition, the area utilization of our relay (shown in Fig. 4, right) is 0.54. For comparison, the NEM relay design from [8] (shown in Fig. 4, left) using a folded flexure (and the same NEMS-on-CMOS integration scheme as this paper) has an area utilization of roughly 0.33 (assuming the top anchors are shared across relays in that design as well). The simple layout used in this work allows reduction of the effective footprint by about 40% over the alternative design in [8], while maintaining a reasonable spring constant—critical to our goal of realizing area-efficient NEMS-on-CMOS routing.

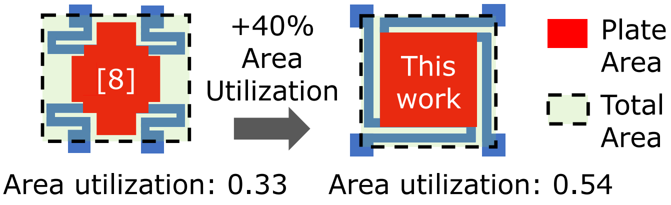


Fig. 4. Area utilization of the NEM relay in [8] (left) versus this work (right).

Table 2
CAD-extracted NEM relay properties.

Param	Value	Description
C_{GB}	1.4 fF	Gate-to-body cap (OFF)
C_{CG}	7.4 aF	Gate-to-channel cap (OFF)
C_{DG}	0.07 fF	Gate-to-drain/source cap (OFF)
C_{CB}	0.15 fF	Body-to-channel cap (OFF)
C_{DB}	1.1 aF	Body-to-drain/source cap (OFF)
C_{DC}	< 1 aF	Channel-to-drain/source cap (OFF)
C_{GB}	2.5 fF	Gate-to-body cap (ON)
C_{CG}	17.3 aF	Gate-to-channel cap (ON)
C_{DG}	0.07 fF	Gate-to-drain/source cap (ON)
C_{CB}	0.15 fF	Body-to-channel cap (ON)
C_{DB}	1.6 aF	Body-to-drain/source cap (ON)
R_{DS}	80 Ω	Total source–drain resistance (ON)

3.2. *Finite element modeling (FEM) and compact SPICE modeling*

We model the relay using a COMSOL 3-D finite element model (FEM) in order to extract key parameters, such as the parasitic resistances/capacitances between different relay terminals. These parameters are given in Table 2. We also extract an estimate for the *contact forces* between the channel contacts and the sources/drains which they touch down upon. These contact forces can be used to predict the expected contact resistance (R_c). The contact resistance determines the source–drain resistance for the NEM relay ($R_{DS} = 2 \times R_c$).

After extracting the parameters of the NEM relay from FEM, we insert these parameters into a compact SPICE model, inspired by the models described in [4,9]. The components of this electromechanical SPICE model are depicted in Fig. 5. The mechanical component of the model is described by a force balance equation (Eq. (1)) that connects the relay position (z) to: (1) the electrostatic force, (2) the spring force, (3) the damping force, and (4) the force from the contacts, i.e., the *contact force*.

$$F_{total} = m\ddot{z} = \underbrace{F_{es}}_{(1)} - \underbrace{kz}_{(2)} - \underbrace{\sqrt{km}/Q}_{(3)} + \underbrace{F_{contact}}_{(4)} \tag{1}$$

For the electrostatic force, we use a parallel plate capacitor model to estimate the Coulomb force between the body and the gate. For the restoring spring force, we use a linear (Hooke) spring model with Q-factor damping. While damping has a significant effect on the speed at which the relay can switch between the OFF and ON states, it mostly affects the transient response. The Q factor is set to 0.5 in our study, which represents the critical damping scenario—it is difficult to estimate the Q factor accurately without knowing the exact process conditions under which the relay will be fabricated. Underdamped systems ($Q > 0.5$) have been reported more widely in the literature [13,14], so a Q factor of 0.5 results in conservative actuation times of around 1.25 µs. Since we use our relays as static switches, the exact transient response is less important for our study.

For the contact force, we use a penalty-based contact force model and neglect contact adhesion [15]. (Contact adhesion is again difficult to predict without knowing the fabrication conditions—it usually

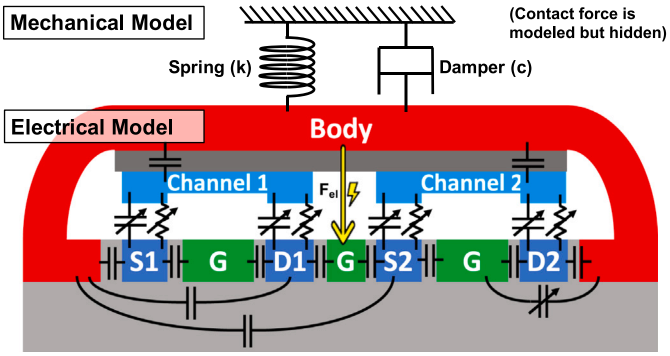


Fig. 5. Electromechanical SPICE model implemented in Verilog-A. Capacitances and resistances to/from a moving component change based on the relay's z-displacement, and are denoted with arrows.

results in slightly lower V_{po}). The FEM solves for the electromechanical forces between each pair of elements in the mesh to determine steady-state displacement.

FEM demonstrates pull-in/pull-out voltages of 4.39 V and 3.43 V, which are within 5% of those predicted by an approximate analytical model based on ideal parallel plates, described in Section 8 (Eq. (10)) [9]. FEM provides the highest accuracy, but is much slower to run than the compact SPICE model. The SPICE model works by solving Eq. (1) for the z-displacement, while modeling parasitic resistances/capacitances in the manner depicted in Fig. 5 (variable capacitances are modeled as simple parallel plates). A sweep of the gate-to-body voltage is given in Fig. 6(a), where the steady-state displacement of the relay is calculated for each point.[2] Transient simulation of pull-in from the SPICE model is shown in Fig. 6(b), displaying a pull-in time of around 1.25 µs, when the gate-to-body voltage is stepped from $V_{GB} = 0$ V to the operating voltage $V_{GB} = V_{oper} = 5$ V at $t = 0$.

3.3. Body biasing scheme for NEM relays

V_{oper} is the gate-to-body voltage (V_{GB}) applied to a relay when we want it to go to the ON state (pull-in operation). In order for pull-in to be possible, we need $V_{GB} > V_{pi}$, so ~ 5 V is a reasonable choice of V_{oper} with our relay. However, this is much higher than the core voltage of most commercial 40 nm technologies. To address this problem, we employ a body biasing strategy, in which each relay's body terminal is tied to a negative voltage [16]. With a body bias of −3.9 V for our relay, an applied gate voltage of 0 V will result in $V_{GB} = 3.9$ V (OFF-state), while a gate voltage of $V_{DD} = 1.1$ V gives $V_{GB} = V_G - V_B = 1.1\text{ V} - (-3.9\text{ V}) = 5$ V (ON-state). Thus, a rising transition on a 1.1 V CMOS signal at the gate terminal of this NEM relay can enable pull-in to the ON state. To pull out the relays, the body bias can be reset to 0 V, thereby restoring all the relays to the OFF state.

It is possible to generate the negative body bias for the NEM relays using an on-chip charge pump scheme [17], as illustrated in Fig. 7(a). A charge pump body bias generator is appropriate for NEM relay body biasing, because it has: (1) no added process complexity (uses only MOS capacitors and MOSFETs), (2) low area overhead (about 500 µm^2), (3) low noise (< 10 mV peak-to-peak ripple noise), and (4) roughly zero power once the body bias is reached at the output. Assuming that a 1.1 V clock already exists in the digital design, this clock signal can be level shifted to the I/O voltage level (2.5 V in our 40 nm technology) and then utilized in a 2× negative voltage multiplier to achieve −5 V. Then, the 1.1 V core voltage can be used to offset the body bias to −3.9 V (= −5 V+1.1 V). We employ a cross-coupled charge pump design, which is described in prior work [17,18]. Our simulations indicate that it would take around 7 µs to reach 95% of this steady state body bias voltage when the total NEM relay body capacitance is conservatively set to 100 pF, as shown in Fig. 7(b).

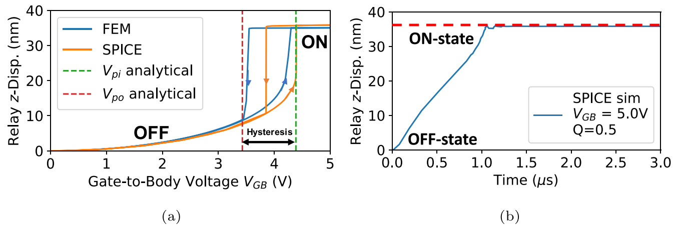
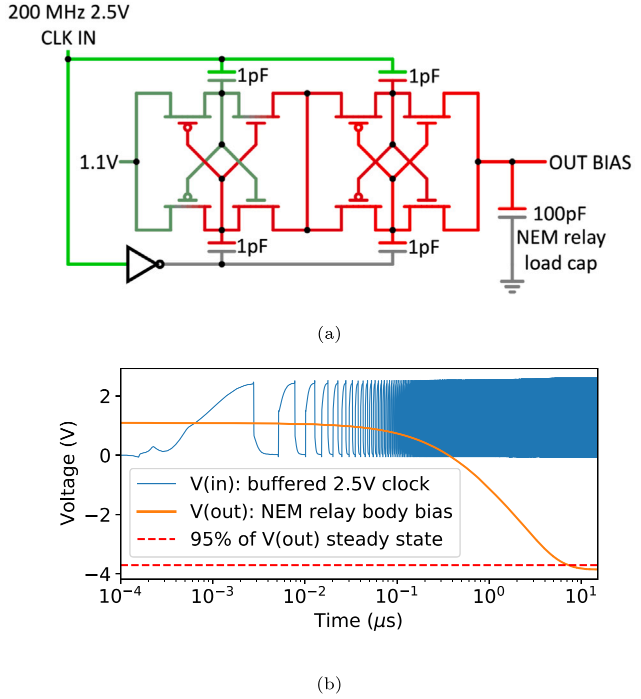


Fig. 6. Relay simulation results. (a) Quasi-static sweep of relay z-displacement versus gate-to-body voltage. Relay pull-in/pull-out phenomena can be observed. (b) SPICE simulation of transient displacement response of relay at $V_{GB} = V_{oper} = 5$ V ON pulse at gate. Pull-in phenomenon can be observed.



(a)

(b)

Fig. 7. Cross-coupled charge pump body biasing scheme. (a) Circuit diagram and specifications for obtaining a −2× voltage multiplier with 40 nm MOSFETs capacitors. Green color indicates positive voltage in steady state, while red color indicates negative voltage in steady state. (b) SPICE simulation results of output voltage vs. time (log scale) starting from relays biased at core voltage (1.1V)

3.4. One-hot multiplexer (OHMux) standard cells

We compose our multi-pole NEM relays into *one-hot multiplexer* (OHMux) standard cells. An OHMux has N selection bits to select from N input signals, with only one selection bit high at a time. To compose an N-input OHMux with NEM relays, we connect N relays by: (1) shorting all of their drains, (2) connecting each source to one of the input signals, and (3) connecting each gate to one of the N selection bits. With multi-pole relays, OHMuxes can multiplex multi-bit inputs. An implementation of an 8-bit-wide 4-input OHmux is shown in Fig. 8.

[2] The SPICE model predicts a higher pull-out voltage than FEM and analytical models. This can likely be attributed to the gear method used in the SPICE simulator, which affects the simulation accuracy when modeling hysteretic behavior. It should not affect the static delay/power behavior when the NEM relays are used as routers.

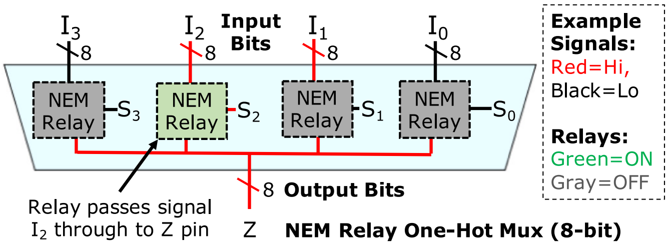


Fig. 8. 4-input 8-bit one-hot multiplexer. Four 8-bit input signals (I_i) are selected by four selection bits (S_i). Only one of the selection bits is hot (S_2 in the example shown).

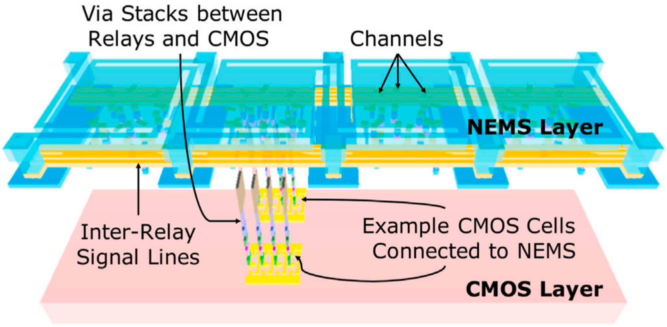

Via Stacks between Relays and CMOS
Channels
NEMS Layer
Inter-Relay Signal Lines
Example CMOS Cells Connected to NEMS
CMOS Layer

Fig. 9. 3-D view of 4-input 8-bit OHMux layout, rendered with GDS3D showing the NEMS-on-CMOS integration scheme and the standard cell design.

We design 3-D standard cell layouts for {2,4,10}-input 8-bit-wide OHMuxes used in different routing components of the CGRA (discussed in Section 5). The 4-input 8-bit-wide OHMux layout is given in Fig. 9, showing vias routing down to a few CMOS standard cells as an example. Inter-relay signal lines short relays' drains (corresponding to the same output bit) to each other. Using parasitic extraction, we conservatively estimate the capacitance of inter-relay signal lines to be $C_{sigline,ohmux} = 0.6\,\text{fF}$.

4. Contact placement for multi-pole relays

Correct placement of NEM relay contacts is critical for multi-pole relay operation. If contacts are placed naively, some of them may not make good connection with the source or drain, resulting in high or variable contact resistance, which can result in reliability problems and increased delay.

To ensure that a multi-pole relay works reliably, its contacts must all touch down simultaneously during the pull-in operation. Therefore, the contacts must be located along a *displacement contour* of the relay. For complex NEM relay designs, it is difficult to analytically find displacement contours. We propose *iterative contact placement (ICP)*, illustrated in Fig. 10, for placing contacts along a displacement contour. Starting with a circular placement, the method works by iteratively finding a displacement contour via FEM, extracting its *connected components* (contiguous sections of the displacement contour), and placing contacts along these components. This process is repeated until the contact forces equalize to within a specified value, yielding a more uniform contact resistance (Fig. 11).

Contact resistance for NEM relays is a function of the contact force, so the problem of balancing contact resistance can be formulated as the problem of placing contacts so that the forces are equal across them. We estimate contact resistance based on the *effective contact area* model [19,20].

$$R_{cont} = \frac{4\rho_W \lambda_W}{3A_r}, \quad A_r \approx \left.\frac{|F_{contact}|}{\xi_W H_W}\right|_{V_{gb}=V_{oper}} \tag{2}$$

Here, ρ_W is the resistivity of the tungsten (W) contact, A_r is the effective area of the contact, and λ_W is the electron mean free path in the W contact. The effective area of the contact (which is typically dominated by asperities) is a function of the contact force, the material hardness (H), and the deformation coefficient (ξ) at the contact. For a tungsten (W) contact:

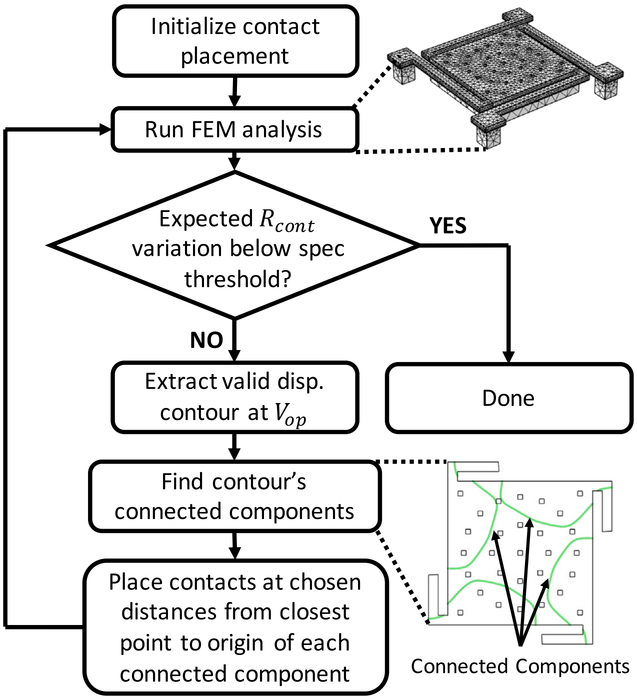

Initialize contact placement
Run FEM analysis
Expected R_{cont} variation below spec threshold?
YES
NO
Extract valid disp. contour at V_{op}
Done
Find contour's connected components
Place contacts at chosen distances from closest point to origin of each connected component
Connected Components

Fig. 10. Iterative contact placement (ICP) algorithm.

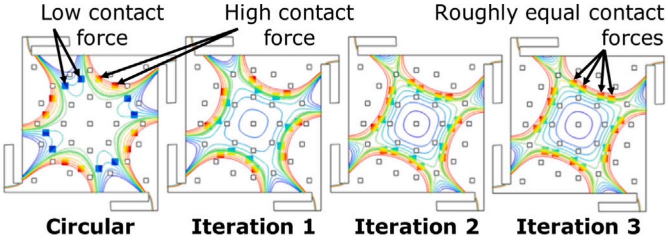


Fig. 11. Iterative contact placement. Contacts are initialized in a circle and displacement contours are extracted using FEM. In each iteration, contacts are placed along the displacement contours from the prior iteration. This is repeated until the contact forces are roughly equalized.

$$H_W = 1.1\,\text{GPa}, \; \lambda_W = 33\,\text{nm}, \; \rho_W = 55\,\text{n}\Omega\,\text{m}, \; \xi_W \approx 0.3 \tag{3}$$

Under this model, ICP yields a contact resistance of around 40 Ω for each contact with a standard deviation of < 1 Ω across contacts ($\sigma < 1\,\Omega$ is the condition we used to terminate ICP). This contact resistance estimate is lower than that reported in the literature for contacts of similar type [12]. This difference may be explained by the fact that our contact forces are much larger, and real contacts include non-idealities in processing conditions, such as oxidation of electrodes or contamination. Regardless, the reduction in contact force variation should improve uniformity across contacts, thereby improving reliability. The predicted contact resistance variation under the model drops by about 40× in just 3 placement iterations (Fig. 12).

5. Hybrid CMOS-NEMS CGRA design

We use an existing open-source CGRA [2] to evaluate our proposed use of multi-pole NEM relays. It has an island-style architecture with

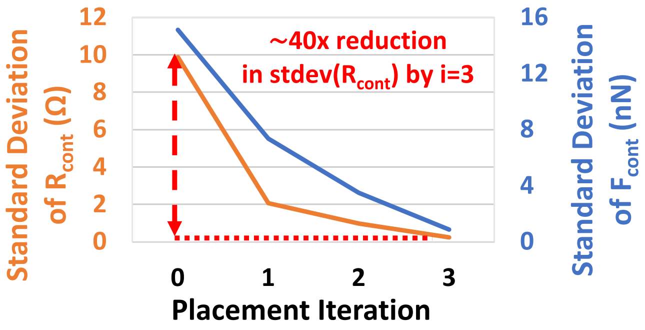


Fig. 12. Reduction in standard deviation of contact force ($|F_{contact}|$) and expected contact resistance (R_{cont}), obtained by using iterative contact placement for an 8-bit relay (with 16 poles).

processing element (PE) tiles, *memory (MEM)* tiles, and a configurable interconnect, illustrated in Fig. 13. PE tiles perform 16-bit arithmetic operations. MEM tiles contain SRAMs that can be used as scratchpad memory. The interconnect contains horizontal and vertical 16-bit routing tracks. Switch boxes (SBs) implement connections between any two tiles. Each SB receives five 16-bit input tracks and drives five 16-bit output tracks on each side (north, south, east, west). Each output track is driven by a multiplexer that selects among the PE/MEM output and the routing tracks coming from the three other sides of the SB (Fig. 13). Each output can also be selectively pipelined. Each connection box (CB) selects a PE/MEM input from among ten 16-bit routing tracks (Fig. 13).

We run a full EDA flow (from RTL to layout) for a single PE tile of the CGRA, both with and without NEM relays, to provide an evaluation of the PPA advantage of NEM relays for programmable routing. Hereafter, we refer to the CMOS-only PE tile as the “CMOS design” and the hybrid-CMOS NEM PE tile as the “NEMS design”. Below, we discuss the modifications to the EDA flow that were necessary to accommodate the integration of NEMS.

The CMOS multiplexers are obtained using logic synthesis, which results in AND-OR-INV (AOI) implementations. To create our hybrid CMOS-NEMS PE tile, we replace all the CMOS multiplexers in the SBs and CBs with NEM OHMuxes (Section 3.4). We add CMOS decoders to the outputs of the configuration registers to generate the one-hot signals that drive the NEM OHMux selection bits. We find that the decoders incur a modest area overhead of $< 100\,\mu\text{m}^2$ per tile (~1% of the PE tile area).

5.1. Physical design flow with NEM relays

Adding the NEM OHMuxes in physical design requires several modifications to the standard place-and-route (P&R) flow. Firstly, the OHMuxes must be declared as *cover cells*, to allow CMOS cells to be placed below them. Additionally, power grids must be constructed to supply the body bias to the NEM relays for proper operation. We create a custom legalization script to align the NEM OHMux cells with these power grids while ensuring no overlap between relays (besides legal anchor sharing). Place-and-route is performed using Cadence Innovus. For fairness of comparison, in both NEMS and CMOS designs, we target the same CMOS standard cell placement density and clock period and use the same design constraints.

5.2. Power and delay analysis

A major challenge in integrating our NEM relays with the rest of the CMOS digital design flow is that without output buffering (i.e., using a CMOS gate to restore the signal at the output of the multiplexer), each OHMux functions similarly to a pass gate—it does not directly drive its outputs, but rather, exposes its downstream capacitance to the driving cell that precedes it. Pass gates have historically not been well supported by commercial EDA tools [21]. So, to produce correct timing and power estimates with standard EDA tools, we employ the following method.

The capacitive load seen by the driving cell of a NEM relay input pin is different depending on whether the relay is OFF or ON. When it is OFF, the capacitance seen by the driver of the source pin is given by:

$$C_{in,relay,off} \approx \underbrace{C_{DB,off} + C_{DG,off} + C_{DC,off}}_{\text{total source cap}} \approx 0.07\,\text{fF} \tag{4}$$

Above, the capitalized subscript letters correspond to the terminals of the relay in Fig. 1(a). When the relay is ON, this effective capacitance increases, as the relay must also charge its channel capacitance and drain capacitance, as well as the downstream load:

$$C_{in,relay,on} \approx \underbrace{C_{CB,on} + C_{CG,on}}_{\text{total channel cap}} + \underbrace{2(C_{DB,on} + C_{DG,on})}_{\text{total source+drain cap}} + \underbrace{C_{load,relay}}_{\text{downstream cap}} \tag{5}$$

The above equations are for a single NEM relay. For an N-input NEM OHMux, the input pin capacitance is the same as the single-relay case when the input is unselected.

$$C_{in,ohmux,off} = C_{in,relay,off} \approx 0.07\,\text{fF} \tag{6}$$

However, when the input pin is selected (i.e., its corresponding select pin is high), two more components are added to the effective pin capacitance: (1) the drain terminals of the other $N - 1$ relays, and (2) the wire capacitance of the signal line connecting the outputs of the NEM relays together. The resulting pin capacitance of a selected pin in an N-input NEM OHMux is given in Eq. (7).

$$C_{in,ohmux,on} = C_{in,relay,on} + \underbrace{(N-1)(C_{DB,on} + C_{DG,on})}_{\text{drain cap of other NEM relays}} + \underbrace{C_{sigline,ohmux}}_{\text{output sig line cap}} \tag{7}$$

To create an EDA flow with NEM relays, we first need to make valid Liberty files for the NEM OHMuxes. Liberty files are the canonical way to provide functional definitions, pin capacitances, timing, and power information for standard cells in a CMOS technology [22]. Liberty files contain lookup tables, which (1) map input transition times and output load capacitances to power and delay estimates, and (2) determine the output transition time to allow the next cell to determine its own power and delay. However, since the NEM relays behave as pass gates, we model them as not having any internal power or delay of their own—rather, they increase the power and delay of the design through the capacitance and resistance they show to their input pin drivers. Therefore, for the NEM OHMuxes, we set all the entries in the power and delay lookup tables to zero. At the same time, we configure the table for the output transition time to simply forward the transition time given at the input pin directly to the output pin. Then we manually adjust the load capacitance seen by the driving cells—specifically, we update the input pin capacitances of the NEM OHMuxes in the Liberty file. We also incorporate the NEM relay drain-to-source resistance by adding R_{DS} onto each of the wires driving the relay inputs. This accounts for the added RC delay through the NEM relay channels/contacts.

For logic synthesis and P&R, we provide a Liberty file containing the *worst-case* estimates for the pin capacitances by setting $C_{in,ohmux} = C_{in,ohmux,on}$ at all input pins, with output load estimates taken from the cells in the CMOS-only PE tile. This enables the P&R tool to conservatively size gates to meet timing, despite the fact that the NEM OHMuxes appear to have zero delay of their own. The same worst-case Liberty file is used for signoff timing analysis. We note that layout parasitics are generated by the physical design tool during place-and-route; hence, to present the driving cells of the NEM OHMuxes with the correct effective load, we also move the parasitic capacitance and resistance at the OHMux’s output net back to the driver output nets before running signoff timing analysis.

Power analysis is performed for each active PE tile in the CGRA and then averaged. We start with *best-case* estimates for the OHMux input

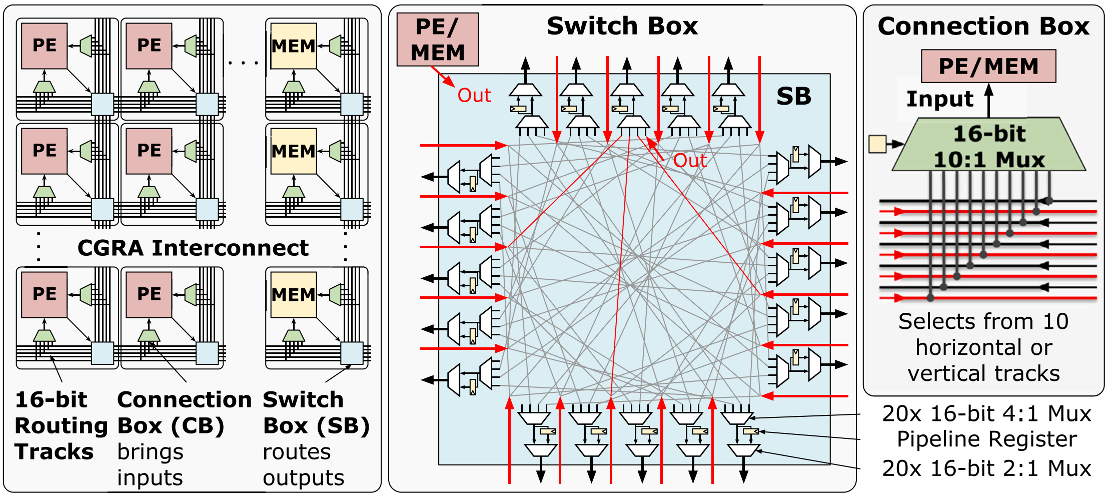


Fig. 13. CGRA with processing element (PE) and memory (MEM) tiles and a configurable interconnect. Also shown are the multiplexers in the switch box (SB) and connection box (CB) that we replace with NEMS-based variants.

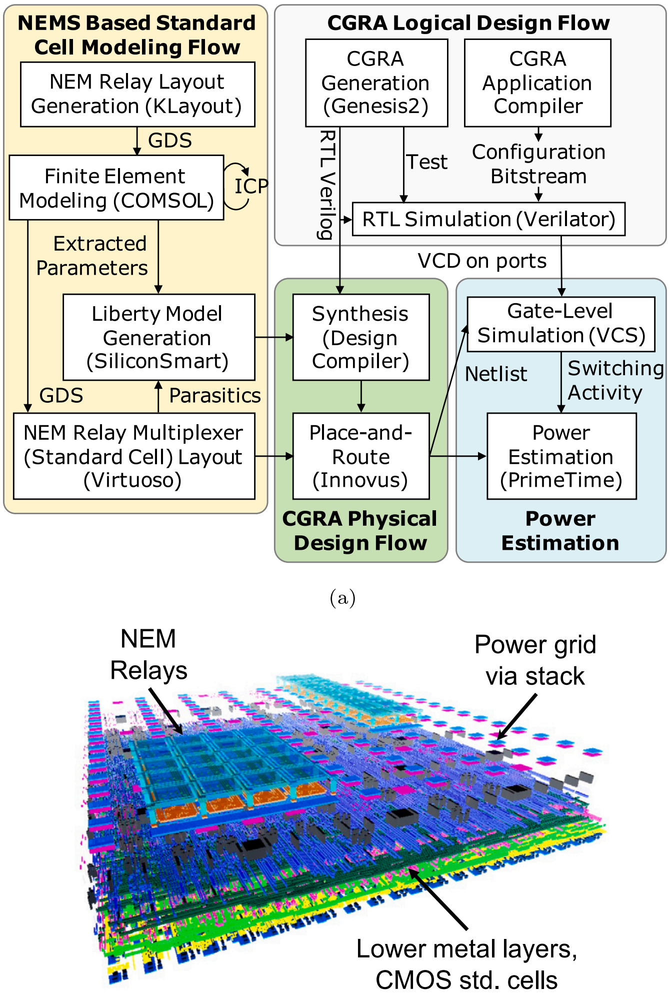
pin capacitances in the Liberty file by setting $C_{in,ohmux} = C_{in,ohmux,off}$. Then, we set the pin capacitances of only the selected OHMux input pins to $C_{in,ohmux,on}$, since these are configured to pass their signal through. We also move the extracted downstream parasitics from the output net of the NEM OHMuxes to the input net, so they are appropriately seen by the drivers. Finally, we run power analysis and zero out the contribution from the NEM OHMuxes that arises due to the charging of its output load (this power would otherwise be double-counted, since we have already included output load charging at the OHMux's input driver).

5.3. Hybrid CMOS-NEMS CGRA flow summary

The full CMOS-NEMS design flow is shown in Fig. 14(a), and a 3-D view of a portion of the resulting 3-D PE tile layout is shown in Fig. 14(b). The full design flow consists of four major sub-flows: (1) NEMS-based standard cell modeling, (2) CGRA logical design, (3) CGRA physical design, and (4) application power estimation. The standard cell modeling procedure has been described in Sections 3 and 4. The CGRA logic design procedure involves generating a register transfer level (RTL) description of the array from a specified configuration file. Applications can then be compiled and mapped to the CGRA, after which RTL simulation can be performed to get switching activities for power estimation. More details on the logical design flow for the CGRA can be found in [2,23]. The physical design flow involves synthesizing the design and performing 3-D PnR as described in Section 5.1. The power and delay analysis procedures are described in Section 5.2. Altogether, the four sub-flows compose a rigorous methodology for the development and evaluation of a 3-D CMOS-NEMS CGRA.

6. Evaluation methodology

We first evaluate the PPA improvement between {2,4,10}-input 8-bit-wide NEM multiplexers and {2,4,10}-input 8-bit-wide CMOS multiplexers. When comparing area, we include one-hot decoding circuitry as part of the NEM multiplexers. The CMOS multiplexers and the NEM one-hot decoders are both synthesized for minimal transistor area, i.e., front-end-of-line (FEOL) area. For an apples-to-apples comparison of delay and switching energy, we use a fixed buffer standard cell (with low drive strength) to drive the multiplexers. This buffer cell's delay and switching energy are included in the multiplexer delay/energy calculations, to duly account for the increase in capacitance that NEM multiplexers expose to their input drivers. We vary the output load being driven by the multiplexers to evaluate how delay/energy are


NEMS Based Standard Cell Modeling Flow
NEM Relay Layout Generation (KLayout)
GDS
Finite Element Modeling (COMSOL)
ICP
Extracted Parameters
Liberty Model Generation (SiliconSmart)
GDS
Parasitics
NEM Relay Multiplexer (Standard Cell) Layout (Virtuoso)
CGRA Logical Design Flow
CGRA Generation (Genesis2)
CGRA Application Compiler
RTL Verilog
Test
Configuration Bitstream
RTL Simulation (Verilator)
VCD on ports
Synthesis (Design Compiler)
Place-and-Route (Innovus)
CGRA Physical Design Flow
Gate-Level Simulation (VCS)
Netlist
Switching Activity
Power Estimation (PrimeTime)
Power Estimation
NEM Relays
Power grid via stack
Lower metal layers, CMOS std. cells

(a)

(b)

Fig. 14. (a) Start-to-finish hybrid CMOS-NEMS CGRA flow. (b) A 3-D view of a portion of the place-and-routed hybrid CMOS-NEMS CGRA PE tile, displaying several NEM muxes and the metal stack below them. Rendered with GDS3D.

affected in different contexts. All delay and energy characterizations are performed using Synopsys PrimeTime PX, along with non-linear delay models (NLDM) from Liberty files/SPICE models.

We then evaluate the PPA improvement from NEM multiplexers at the system level for several applications on the CGRA. To compile applications to the CGRA, we use the toolchain from [2], which is similar to a high-level synthesis toolchain for a commercial FPGA. It takes as input an application written in a high-level domain-specific language called Halide [23], compiles it to a dataflow graph, and then maps, places and routes the graph on the CGRA. The output of the toolchain is a configuration bitstream that contains both the information on what the PE and MEM tiles should do, as well as the routing information (in the form of values of select bits for each multiplexer in the SBs and CBs). We then run a Verilog simulation of the configured CGRA to obtain the switching activity at all the nodes in the PE tiles that are active. We feed this into PrimeTime PX, along with the post-P&R PE tile netlist and parasitics. After modifying the capacitances/resistances, as detailed in Section 5.2, PrimeTime is able to generate accurate delay/power numbers for each application. This is performed for three different values of R_{DS}, starting from $80\,\Omega$ (predicted by effective contact area model), up to $5\,\mathrm{k}\Omega$ (observed experimentally in [12]).

We evaluate the baseline and our hybrid PE tile on three image processing applications (Table 3). *Conv. 3 × 3* is a 2-D convolution with a 3 × 3 kernel. *Cascade* has two back-to-back 3 × 3 convolutions. *Harris* is a corner detector. We measure the power after the CGRA bitstream has been loaded, since the cost of configuration is amortized over the long-running operation.

7. Results

7.1. NEM multiplexer power/performance/area

The PPA results for a single NEM multiplexer vs. a minimally-sized CMOS multiplexer with the same functionality are summarized in Fig. 15. The results indicate NEM relay multiplexers have $5.8\times-27.7\times$ lower front-end-of-line (FEOL) area, and, under the right driver/load conditions, up to 27.5% lower switching energy as well as up to $4.22\times$ lower delay.

7.1.1. Area

The NEM multiplexers yield a significant improvement in FEOL area compared to CMOS multiplexers, due to the fact that they only require area for the decoding circuit. Fig. 15(a) illustrates that the NEMS area benefit diminishes with greater number of multiplexer inputs, as the area of the decoding logic starts to approach that of the multiplexers. It should be noted that NEM multiplexers still occupy a large amount of area in the NEMS layer, and may become the area bottleneck if the total NEMS layer area exceeds the total FEOL area. The NEMS multiplexers also may increase routing congestion, since vias/wires are needed to traverse the entire metal stack between the CMOS layer (at the bottom) and NEMS layer (at the top). The end result is that some of the area benefits are lost in practice, as reflected by our “Area Figure of Merit (FOM)” metric presented in Section 7.2.

7.1.2. Switching energy

Fig. 15(b) illustrates the energy for an input signal (I) to propagate to the output (Z) through a NEMS/CMOS multiplexer as a function of the capacitive load being driven (see Fig. 8). The largest $I \rightarrow Z$ energy benefit (a 27.5% decrease) is achieved for a 2-input 8b NEMS multiplexer when compared to a 2-input CMOS multiplexer at very small load capacitance ($0.21\,\mathrm{fF}$)—a smaller relative benefit is achieved for more inputs and larger load capacitance.

7.1.3. Switching delay

Fig. 15(c) illustrates the time required for an input signal (I) to propagate to the output (Z) through a NEM/CMOS multiplexer as a function of the capacitive load being driven (see Fig. 8). The largest $I \rightarrow Z$ delay benefit (a $4.22\times$ decrease) is achieved for a 10-input 8b NEMS multiplexer when compared to a 10-input CMOS multiplexer at very small load capacitance ($0.21\,\mathrm{fF}$)—a smaller relative benefit is achieved for fewer inputs and larger load capacitance. The intuition behind the switching delay improvement is that there are fewer logic levels for the signal to traverse in a NEMS multiplexer, therefore less gate delay is incurred. For larger load capacitances, the CMOS multiplexer starts to quickly outperform the NEMS multiplexer, given its improved handling of output load capacitance.

7.1.4. Discussion

Here, we have analyzed minimally-sized CMOS multiplexers. NEM relays provide significant area and $I \rightarrow Z$ switching energy benefits even when the CMOS multiplexers are sized up. However, for $I \rightarrow Z$ switching delay, NEM multiplexers begin to incur a significant delay penalty when compared to CMOS multiplexers with larger drive strengths. Next, we evaluate how NEMS multiplexers impact the PPA of a CGRA PE tile and find that in order to meet timing at high clock frequencies, the surrounding logic driving the NEM multiplexers must be sized up, which attenuates the overall benefit of introducing NEMS.

7.2. NEMS CGRA power/performance/area

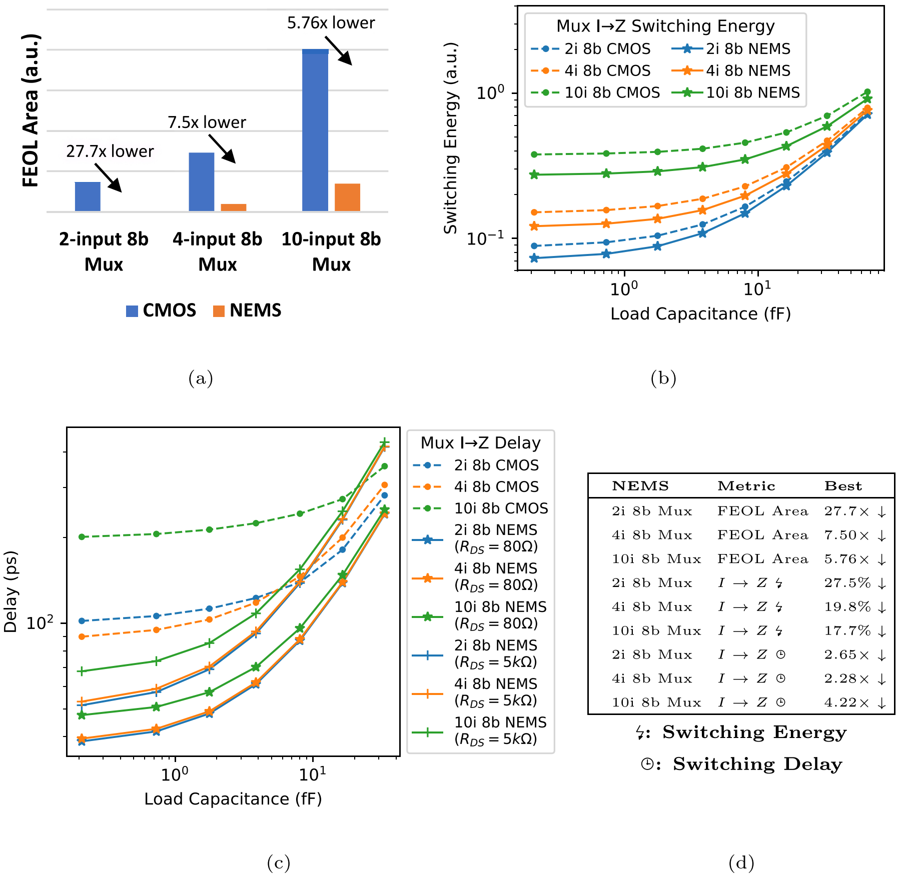
The power and area results are summarized in Fig. 16, with more detail on the application-level breakdown and the effect of R_{DS} in Table 3. Our design shows 19% better area and 10% better power at iso-critical path delay (clock period = 5 ns) across the applications. The power reduction mainly occurs in the switch boxes, where having fewer CMOS standard cells results in less leakage and smaller dynamic power dissipation. The area reduction comes from direct mapping of the CMOS multiplexers in the SB/CB to NEM OHMuxes integrated in 3-D. In order to evaluate how efficiently we integrated NEMS with CMOS, we examine the following area figure of merit (FOM):

$$\text{Area FOM} = \frac{\text{CMOS Design Stdcell Area} - \text{NEMS Design Stdcell Area}}{\text{SB/CB Mux Area}} \tag{8}$$

This FOM indicates what percentage of the maximum possible area savings (from moving SB/CB multiplexers to the NEM relay layer) was realized. We achieve an area FOM of **73.3%**. The reason this is less than 100% is due to the area overhead from decoders for the one-hot select signals, and increased area from logic surrounding the NEM routers to meet timing.

7.3. Reconfiguration delay penalty

It should be noted that the NEM relays altogether add $1 \times t_{\text{pull-in}}$ to the total reconfiguration time (where $t_{\text{pull-in}}$ is the time to go from the OFF state to the ON state). This is because when the last configuration register is programmed, it takes time $t_{\text{pull-in}}$ for that change to take effect in the last NEM multiplexer. Each application may require a different number of clock cycles to configure—Cascade takes 510 cycles ($510 \times 5\,\mathrm{ns} = 2550\,\mathrm{ns}$) to configure, Harris takes 1097 cycles ($1097 \times 5\,\mathrm{ns} = 5485\,\mathrm{ns}$), and Conv 3 × 3 takes 150 cycles ($150 \times 5\,\mathrm{ns} = 750\,\mathrm{ns}$). The switching delay for the NEM relay we developed is conservatively estimated to be around $1250\,\mathrm{ns}$ in Section 3.2, so the reconfiguration process will take $1250\,\mathrm{ns}$ longer than in the CMOS-only design (18%–167% increase). However, in a typical setting, a CGRA application runs without reconfiguration for a period of time much longer than $1250\,\mathrm{ns}$. Therefore, this delay penalty is amortized over the run time of the application.



(a)

(b)

(c)

NEMS	Metric	Best
2i 8b Mux	FEOL Area	27.7× ↓
4i 8b Mux	FEOL Area	7.50× ↓
10i 8b Mux	FEOL Area	5.76× ↓
2i 8b Mux	$I \to Z$ ϟ	27.5% ↓
4i 8b Mux	$I \to Z$ ϟ	19.8% ↓
10i 8b Mux	$I \to Z$ ϟ	17.7% ↓
2i 8b Mux	$I \to Z$ ⊙	2.65× ↓
4i 8b Mux	$I \to Z$ ⊙	2.28× ↓
10i 8b Mux	$I \to Z$ ⊙	4.22× ↓

ϟ: Switching Energy

⊙: Switching Delay

(d)

Fig. 15. Power, performance, and area metrics of NEM multiplexers vs. CMOS multiplexers. (a) Comparison of CMOS/NEMS transistor area (FEOL). (b) Comparison of CMOS/NEMS $I \to Z$ switching energy in static router configuration. (c) Comparison of CMOS/NEMS $I \to Z$ switching delay in static router configuration. (d) Summary of best-case improvement for NEMS multiplexers over CMOS multiplexers in each metric.

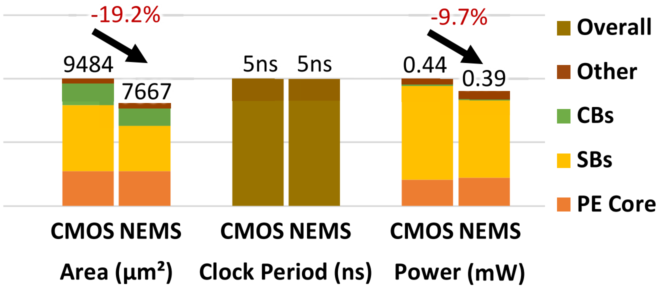


Fig. 16. Signoff power, performance, and area, for one CGRA PE tile of CMOS and NEMS designs ($R_{DS} = 80\,\Omega$).

Table 3
CGRA PE tile power and area (per tile) of NEMS-based vs. the CMOS-only design. All entries at iso-delay (clock period = 5 ns).

PPA metrics summary			CMOS	NEMS	
		R_{DS}	Total	Total	% Improvement
Power (mW)	Conv 3 × 3 (20 tiles)	80 Ω	0.510	0.460	9.8
		1 kΩ		0.465	8.8
		5 kΩ		0.475	6.9
	Cascade (71 tiles)	80 Ω	0.401	0.362	9.8
		1 kΩ		0.366	8.8
		5 kΩ		0.373	7.0
	Harris (154 tiles)	80 Ω	0.396	0.357	9.9
		1 kΩ		0.360	8.9
		5 kΩ		0.368	7.1
Area (μm^2)	Overall	80 Ω	9484	7667	19.2
		1 kΩ		7673	19.1
		5 kΩ		7808	17.7

7.4. NEMS layer area considerations

In this paper, we have analyzed a 20-track CGRA in a 40 nm technology, where the total area of the relays in the NEMS layer ($\approx 3675\,\mu m^2$) is significantly less than that of the total area in the CMOS layer ($\approx 7667\,\mu m^2$). However, when a larger number of routing tracks is used in the CGRA, the occupied area in the NEMS layer increases, as more multiplexers are needed for routing. At some critical number of tracks, it is clear that the NEMS layer area starts to surpass that of the CMOS layer. When this happens, the NEMS layer becomes the area bottleneck.

A lower bound on the critical number of I/O tracks, T, can be estimated by: (1) finding a formula for the total area of the relays as a function of I/O tracks, then (2) constraining that to be less than the CMOS area when 20 tracks are used. In reality, the CMOS area for $T > 20$ will increase somewhat because of extra decoding logic and configuration registers, but this area may be conservatively estimated to be constant. Assuming a 16-bit datapath with 8-bit relays, the inequality to consider is:

$$\text{Total NEM Relay Area} = 13T \times (3.76\,\mu m)^2 \leq 7667\,\mu m^2 \qquad (9)$$

Above, $3.76\,\mu m$ is the relay pitch. The number of relays required is linearly proportional to the number of tracks since a one-hot multiplexing scheme is used. Solving the inequality for T yields: $T \leq 40$. Thus, we see that even a 40-track CGRA would have the CMOS layer remain the area bottleneck.

Besides considering the number of tracks, one could consider adding more channels and contacts per relay. More contacts per relay would enable wider signals to be routed, which reduces the number of relays required. However, this only reduces the area in the NEMS layer—the CMOS layer area would actually increase. As the number of contacts per relay increases, the contact force becomes divided across a greater number of contacts. Therefore, the contact force per contact decreases, which leads to a higher effective channel resistance (see Section 8). Higher resistance results in larger delay, which means larger standard cells are needed to meet timing requirements. The larger cells have higher power, so both CMOS area and power increase as more channels are added.

8. Scalability of multi-pole NEMS to future technology nodes

Up to this point, we have analyzed multi-pole NEM relays in 40 nm CMOS technology. Scalability of NEMS technology to sub−40 nm CMOS processes will be key to its real-world adoption and is desirable for improved density and performance. In this section, we estimate how NEM relay scaling may impact: (1) operating voltages, (2) node capacitances, and (3) channel resistance. We also discuss how these parameters will likely affect PPA and reliability. The analysis here builds off of prior scaling studies in [24] and provides insights on the benefits/challenges associated with scaling multi-pole relays to the most recent CMOS technology nodes.

8.1. Formulas for estimating NEM relay parameters

Prior work [24] has derived approximate formulas for the pull-in and pull-out voltages V_{pi}, V_{po}:

$$V_{pi} \approx \sqrt{\frac{8k_{total}g_{act}^3}{27\epsilon_0 A_{plate}}} \qquad V_{po} \approx \sqrt{\frac{2(k_{total}t_{cont} - F_A)(g_{act} - t_{cont})^2}{\epsilon_0 A_{plate}}} \tag{10}$$

Above, k_{total} is the total spring constant of the NEM relay, g_{act} is the actuation gap between the relay body and the substrate, A_{plate} is the area of the NEM relay electrostatic plate, t_{cont} is the thickness of the contact, F_A is the adhesive force at the contacts, and ϵ_0 is the vacuum permittivity constant. In the analysis here, we continue to ignore adhesion ($F_A \approx 0$). We can estimate the spring constant k_{total} of our relays using a clamped-guided beam approximation (see Eq. (9).55 in [25]) for each of the four cantilever beams based on the dimensions and material parameters:

$$k_{beam} = \frac{E_{SiGe}W_{beam}t_{body}^3}{4L_{beam}^3} \tag{11}$$

Above, E_{SiGe} is the Young's modulus of the poly-SiGe, W_{beam} is the width of the cantilever beam, t_{body} is the thickness of the poly-SiGe body, and L_{beam} is the length of the cantilever beams providing the restoring spring force. The total spring constant k_{total} is found by multiplying by four, since the beams act in parallel:

$$k_{total} \approx 4 \times k_{beam} \approx \frac{E_{SiGe}W_{beam}t_{body}^3}{L_{beam}^3} \tag{12}$$

We can approximate the most important parasitic capacitances using parallel plate capacitor models between different NEM relay nodes (see Table 2 for description). These models are illustrated in Fig. 17.

The resulting expressions for the important parasitic capacitances using these models are given in the equations below:

$$C_{GB,off} \approx \frac{\epsilon_0 A_{plate}}{g_{act}} \parallel \frac{\epsilon_{sp}A_{plate}}{t_{sp}} = \frac{A_{plate}}{g_{act}/\epsilon_0 + t_{sp}/\epsilon_{sp}} \tag{13}$$

$$C_{GB,on} \approx \frac{\epsilon_0 A_{plate}}{t_{cont}} \parallel \frac{\epsilon_{sp}A_{plate}}{t_{sp}} = \frac{A_{plate}}{t_{cont}/\epsilon_0 + t_{sp}/\epsilon_{sp}} \tag{14}$$

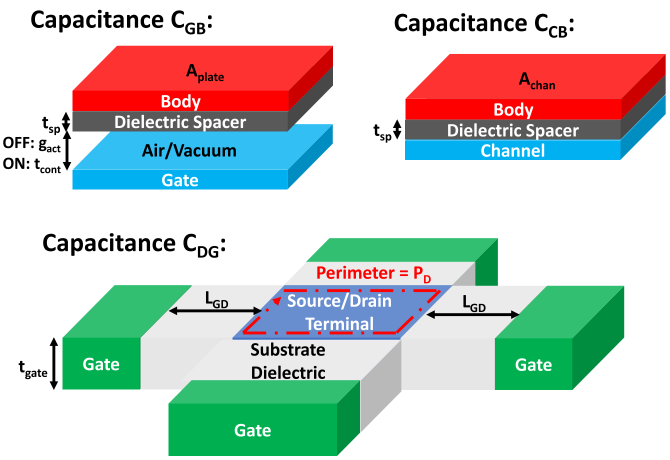


Fig. 17. Illustration of parallel plate capacitor models used in scaling analysis of parasitic capacitances.

$$C_{CB} \approx \frac{\epsilon_{sp}A_{chan}}{t_{sp}} \qquad C_{DG} \approx \frac{\epsilon_{sd}P_D t_{gate}}{L_{GD}} \tag{15}$$

Above, t_{sp} is the thickness of the dielectric spacer between the body and the channels, ϵ_{sp} is the dielectric constant of the spacer between the body and the channels, ϵ_{sd} is the dielectric constant of the substrate dielectric material, A_{chan} is the planar area of a single channel, P_D is the perimeter of the source/drain contact, t_{gate} is the thickness of the gate metal, L_{GD} is the lateral spacing between the gate and the drain, and ϵ_0 is the vacuum permittivity. Finally, the formula for contact resistance is given earlier in Eq. (3). The contact force can be approximated by rearranging the force balance equation Eq. (1) and ignoring damping force to get:

$$|F_{contact}| \approx F_{es} - k_{total}(g_{act} - t_{cont}) = \underbrace{\frac{\epsilon_0 A_{plate}V_{oper}^2}{2t_{cont}^2}}_{\text{Coulomb force}} - \underbrace{k_{total}(g_{act} - t_{cont})}_{\text{spring force}} \tag{16}$$

8.2. Scaling analysis

We consider three types of dimensional scaling: (1) *lateral scaling*, in which the planar relay dimensions are shrunk by a factor λ while the vertical dimensions are not scaled, (2) *constant field scaling* [26], in which all three dimensions are shrunk by a factor λ, and (3) a *hybrid scaling* approach, in which lateral scaling is applied while the vertical dimensions are selectively scaled to achieve desirable electrical properties. Under each scaling strategy, we can derive the proportionality between the relay's electrical parameters and λ to determine how they will evolve in future technology nodes.

8.2.1. Lateral scaling

In lateral scaling, only the lateral dimensions (see Table 1) are scaled down by λ. This means:

$$\{A_{plate}, A_{chan}\} \propto 1/\lambda^2 \qquad \{W_{beam}, L_{beam}, P_D, L_{GD}\} \propto 1/\lambda \tag{17}$$

We can then derive the resulting impact on the electrical parameters:

$$k_{total} \propto \frac{W_{beam}}{L_{beam}^3} \propto \lambda^2 \tag{18}$$

$$V_{pi} \propto \sqrt{\frac{k_{total}}{A_{plate}}} \propto \lambda^2 \qquad V_{po} \propto \sqrt{\frac{k_{total}}{A_{plate}}} \propto \lambda^2 \tag{19}$$

$$C_{GB,off} \propto A_{plate} \propto 1/\lambda^2 \qquad C_{GB,on} \propto A_{plate} \propto 1/\lambda^2 \tag{20}$$

$$C_{DG} \propto \frac{P_D}{L_{GD}} \propto \lambda^0 \qquad C_{CB} \propto A_{chan} \propto 1/\lambda^2 \tag{21}$$

$$V_{oper} \approx V_{pi} \qquad \frac{A_{plate}V_{oper}^2}{t_{cont}^2} \propto \lambda^2 \qquad k_{total}(g_{act} - t_{cont}) \propto \lambda^2 \tag{22}$$

$$|F_{contact}| \propto \lambda^2 \qquad R_{cont} \propto 1/|F_{contact}| \propto 1/\lambda^2 \tag{23}$$

The takeaways are that when scaling down laterally by a factor of λ: (1) the pull-in and pull-out voltages increase by a factor of λ^2, (2) the parasitic capacitances to the body scale down by a factor of λ^2, (3) the parasitic capacitance between the drain and gate terminals does not change, and (4) the contact resistance decreases by a factor of λ^2. While the improvements in the parasitic capacitances and resistances are evident, the quadratic increase in the pull-in and pull-out voltages poses a major challenge in practice, since large voltages pose compatibility issues with deep sub-micron CMOS technology.

8.2.2. Constant field scaling

In constant field scaling, all relay dimensions (see Table 1) are scaled down by λ. This means in addition to Eq. (17):

$$\{g_{act}, t_{cont}, t_{sp}, t_{body}\} \propto 1/\lambda \tag{24}$$

Using these proportionality relations, we can derive the impact of scaling on the electrical parameters:

$$k_{total} \propto \frac{W_{beam}t_{body}^3}{L_{beam}^3} \propto 1/\lambda \tag{25}$$

$$V_{pi} \propto \sqrt{\frac{k_{total}g_{act}^3}{A_{plate}}} \propto 1/\lambda \qquad V_{po} \propto \sqrt{\frac{k_{total}t_{cont}(g_{act} - t_{cont})^2}{A_{plate}}} \propto 1/\lambda \tag{26}$$

$$g_{act}/\epsilon_0 + t_{sp}/\epsilon_{sp} \propto 1/\lambda \qquad t_{cont}/\epsilon_0 + t_{sp}/\epsilon_{sp} \propto 1/\lambda \tag{27}$$

$$C_{GB,off} \propto \frac{A_{plate}}{1/\lambda} \propto 1/\lambda \qquad C_{GB,on} \propto \frac{A_{plate}}{1/\lambda} \propto 1/\lambda \tag{28}$$

$$C_{DG} \propto \frac{P_D t_{gate}}{L_{GD}} \propto 1/\lambda \qquad C_{CB} \propto \frac{A_{chan}}{t_{sp}} \propto 1/\lambda \tag{29}$$

$$V_{oper} \approx V_{pi} \qquad \frac{A_{plate}V_{oper}^2}{t_{cont}^2} \propto 1/\lambda^2 \qquad k_{total}(g_{act} - t_{cont}) \propto 1/\lambda^2 \tag{30}$$

$$|F_{contact}| \propto 1/\lambda^2 \qquad R_{cont} \propto 1/|F_{contact}| \propto \lambda^2 \tag{31}$$

The takeaways are that when applying constant field scaling with a factor of λ: (1) the pull-in and pull-out voltages decrease by a factor of λ, (2) the parasitic capacitances scale down by a factor of λ, and (3) the contact resistance increases by a factor of λ^2. While the decrease in pull-in and pull-out voltages is desirable, the parasitic capacitance does not scale as well as in the lateral scaling case (except C_{DG}, which scales better). The quadratic increase in contact resistance poses a major challenge, as large path resistance causes increased delay.

8.2.3. Hybrid scaling strategy

To address the issues with the above two scaling strategies, a hybrid scaling strategy may be adopted. Rather than scaling down *all* of the vertical dimensions (as in constant field scaling), we can notice that lateral scaling in conjunction with scaling of *only the body thickness* solves the problem of quadratically increasing operating voltages/contact resistances. In the proposed hybrid scaling strategy:

$$\{A_{plate}, A_{chan}\} \propto 1/\lambda^2 \qquad \{W_{beam}, L_{beam}, P_D, L_{GD}, t_{body}\} \propto 1/\lambda \tag{32}$$

Deriving the scaling of electrical parameters as before:

$$k_{total} \propto \frac{W_{beam}t_{body}^3}{L_{beam}^3} \propto 1/\lambda \tag{33}$$

$$V_{pi} \propto \sqrt{\frac{k_{total}}{A_{plate}}} \propto \sqrt{\lambda} \qquad V_{po} \propto \sqrt{\frac{k_{total}}{A_{plate}}} \propto \sqrt{\lambda} \tag{34}$$

$$C_{GB,off} \propto A_{plate} \propto 1/\lambda^2 \qquad C_{GB,on} \propto A_{plate} \propto 1/\lambda^2 \tag{35}$$

$$C_{DG} \propto \frac{P_D}{L_{GD}} \propto \lambda^0 \qquad C_{CB} \propto A_{chan} \propto 1/\lambda^2 \tag{36}$$

$$V_{oper} \approx V_{pi} \qquad \frac{A_{plate}V_{oper}^2}{2t_{cont}^2} \propto 1/\lambda \qquad k_{total}(g_{act} - t_{cont}) \propto 1/\lambda \tag{37}$$

$$|F_{contact}| \propto 1/\lambda \qquad R_{cont} \propto 1/|F_{contact}| \propto \lambda \tag{38}$$

The takeaways are that when applying this hybrid strategy with scaling factor λ: (1) the pull-in and pull-out voltages increase by a factor of $\sqrt{\lambda}$, (2) the parasitic capacitances to the body scale down by a factor of λ^2, (3) the parasitic capacitance between the drain and gate terminals does not change, and (4) the contact resistance increases by a factor of λ. The pull-in/pull-out voltages scale in a much more manageable fashion than the lateral scaling scenario and we maintain the excellent scaling of parasitic capacitance. While the contact resistance increases by a factor of λ, this is more reasonable than the quadratic increase precipitated by constant field scaling.

8.3. Variability and other challenges associated with scaling

We anticipate several other challenges associated with scaling. As mechanical structures are brought to nanometer-scale dimensions, effects that can otherwise be ignored begin to take center stage. For example, it has been observed that stress and strain in scaled devices lead to higher likelihood of mechanical failure [12,27]. Leakage may also become a significant problem if the relay's actuation gap is small enough that tunneling current can flow.

At highly scaled dimensions, variations in electromechanical parameters become rather prominent. If process variation induces differences in a relay's layer thicknesses, the relay's contact resistance can change drastically. This is because when multiple contacts are positioned at different vertical distances from the substrate, it is probable that some contacts end up touching down before others. The contacts that do not touch down fully have higher contact resistance, causing increased RC delay. Process variation may also affect the mechanical properties of the relays, for example the stiffness of the cantilevers and electrostatic plate. The altered structural and mechanical properties then lead to uncertainty in the pull-in voltage.

We can use our prior approximations (Eqs. (10) and (11)) for the pull-in voltage and spring constant to determine roughly that:

$$V_{pi} \propto \sqrt{k_{total}g_{act}^3} \propto \sqrt{t_{body}^3 g_{act}^3} \tag{39}$$

Thus, for a ±2% thickness variation across the chip, the pull-in voltage would vary by about ±6% (between 4.13 V and 4.66 V for the relay design presented in Section 3). Either a conservatively large pull-in voltage may be adopted, or if pull-in voltages can be measured per die, the operating voltages may be adjusted on a chip-by-chip basis—this is possible by storing the operating voltage in a non-volatile memory to increase the yield, a standard practice in modern MEMS manufacturing [28]. Additionally, there are ways to improve the uniformity of deposited films by ion-beam trimming [29]. The thickness of the sacrificial and structural layers can be measured across a wafer and adjusted using an ion beam, reducing variation by more than an order of magnitude.

9. Conclusion

In this paper, we have explored the use of multi-pole NEM relays as multi-bit routers in CGRAs. We have designed and modeled an area-efficient multi-pole relay that can share anchors and be fabricated on top of CMOS. We have demonstrated techniques for overcoming design challenges including generation of negative body bias and mitigation of contact resistance variation. Finally, we have shown that integration of these multi-pole relays into the PE tiles of a hybrid CMOS-NEMS CGRA can yield 19% lower area and 10% lower power at iso-delay.

There are several directions of future work. Avenues for improving the power/area savings include: (1) smarter integration with EDA tools—fixing over-sized cells caused by pass gate logic insertion, (2) using NEM relay hysteresis to eliminate the need for CMOS configuration registers as proposed in [5], and (3) introducing NEM multiplexing into the PE core rather than just the SBs/CBs. More detailed NEM relay FEM/SPICE modeling could be explored; in particular, analysis of damping behavior, stiction, strain, and cross-talk between closely placed NEMS devices could be studied. It would also be interesting to consider other applications of multi-pole relays in digital design, for example network-on-chip routing. Optimization of individual multi-pole NEM relay devices can be taken further; in particular, the design dimensions, perforations (placement/size of release holes), and relay materials could be optimized. Finally, and perhaps most importantly, experimental demonstration of a multiplexer built with multi-pole relays (integrated on top of CMOS) would confirm that the approaches described here are indeed practical.

The tools and models used in this paper are all available on GitHub,[3] and we hope they will be useful to the community. We believe 3-D integration of multi-pole NEM relays presents a promising approach for bridging the power-performance-area gap between programmable logic devices (such as CGRAs) and ASICs.

List of Abbreviations

AOI—AND Gate-OR Gate-Inverter (AND-OR-INV)
ASIC—Application-Specific Integrated Circuit
BEOL—Back End Of Line
CB—Connection Box
CGRA—Coarse-Grained Reconfigurable Array
CMOS—Complementary Metal–Oxide-Semiconductor
FEOL—Front End Of Line
FEM—Finite Element Model
FOM—Figure of Merit
ICP—Iterative Contact Placement
LUT—Look-Up Table
MEM—Memory
MOS—Metal–Oxide-Semiconductor
MOSFET—Metal–Oxide-Semiconductor Field-Effect Transistor
NEM—Nanoelectromechanical
NMOS—N-channel Metal–Oxide-Semiconductor
OHMux—One-Hot Multiplexer
P&R—Place-and-Route
PE—Processing Element
PLD—Programmable Logic Device
PPA—Power Performance Area
RTL—Register Transfer Level
SB—Switch Box

[3] https://github.com/users/akashlevy/projects/4

CRediT authorship contribution statement

Akash Levy: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Visualization. **Michael Oduoza:** Methodology, Investigation, Data curation, Writing – original draft. **Akhilesh Balasingam:** Investigation, Writing – review & editing. **Roger T. Howe:** Resources, Writing – review & editing. **Priyanka Raina:** Supervision, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] M. Wijtvliet, L. Waeijen, H. Corporaal, Coarse grained reconfigurable architectures in the past 25 years: Overview and classification, in: 2016 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation, SAMOS, IEEE, 2016, pp. 235–244.

[2] R. Bahr, C. Barrett, N. Bhagdikar, A. Carsello, R. Daly, C. Donovick, D. Durst, K. Fatahalian, K. Feng, P. Hanrahan, T. Hofstee, M. Horowitz, D. Huff, F. Kjolstad, T. Kong, Q. Liu, M. Mann, J. Melchert, A. Nayak, A. Niemetz, G. Nyengele, P. Raina, S. Richardson, R. Setaluri, J. Setter, K. Sreedhar, M. Strange, J. Thomas, C. Torng, L. Truong, N. Tsiskaridze, K. Zhang, Creating an agile hardware design flow, in: 2020 57th ACM/IEEE Design Automation Conference, DAC, IEEE, 2020, pp. 1–6.

[3] A. Vasilyev, N. Bhagdikar, A. Pedram, S. Richardson, S. Kvatinsky, M. Horowitz, Evaluating programmable architectures for imaging and vision applications, in: 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO, IEEE, 2016, pp. 1–13.

[4] F. Chen, H. Kam, D. Markovic, T.-J.K. Liu, V. Stojanovic, E. Alon, Integrated circuit design with NEM relays, in: 2008 IEEE/ACM International Conference on Computer-Aided Design, IEEE Press, 2008, pp. 750–757.

[5] C. Chen, R. Parsa, N. Patil, S. Chong, K. Akarvardar, J. Provine, D. Lewis, J. Watt, R.T. Howe, H.-S.P. Wong, S. Mitra, Efficient FPGAs using nanoelectromechanical relays, in: 18th ACM/SIGDA International Symposium on Field Programmable Gate Arrays, ACM, 2010, pp. 273–282.

[6] C. Chen, W.S. Lee, R. Parsa, S. Chong, J. Provine, J. Watt, R.T. Howe, H.-S.P. Wong, S. Mitra, Nano-electro-mechanical relays for FPGA routing: Experimental demonstration and a design technique, in: Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012, IEEE, 2012, pp. 1361–1366.

[7] A. Levy, M. Oduoza, A. Balasingam, R.T. Howe, P. Raina, Efficient routing in coarse-grained reconfigurable arrays using multi-pole NEM relays, in: 2022 27th Asia and South Pacific Design Automation Conference, ASP-DAC, 2022, pp. 472–478, http://dx.doi.org/10.1109/ASP-DAC52403.2022.9712515.

[8] B. Osoba, B. Saha, L. Dougherty, J. Edgington, C. Qian, F. Niroui, J.H. Lang, V. Bulovic, J. Wu, T.-J.K. Liu, D. Marković, E. Alon, V. Stojanović, Sub-50 mv NEM relay operation enabled by self-assembled molecular coating, in: 2016 IEEE International Electron Devices Meeting, IEDM, IEEE, 2016, pp. 26–28.

[9] M. Spencer, F. Chen, C.C. Wang, R. Nathanael, H. Fariborzi, A. Gupta, H. Kam, V. Pott, J. Jeon, T.-J.K. Liu, et al., Demonstration of integrated micro-electro-mechanical relay circuits for VLSI applications, IEEE J. Solid-State Circuits 46 (1) (2010) 308–320.

[10] C.W. Low, S.F. Almeida, E.P. Quévy, R.T. Howe, Poly-sige surface micromachining, in: 3D and Circuit Integration of MEMS, John Wiley & Sons, Ltd, ISBN: 9783527823239, 2021, pp. 69–97, http://dx.doi.org/10.1002/9783527823239.ch5.

[11] R. Maboudian, R.T. Howe, Critical review: Adhesion in surface micromechanical structures, J. Vacuum Sci. Technol. B 15 (1) (1997) 1–20.

[12] R. Nathanael, Nano-electro-mechanical (NEM) relay devices and technology for ultra-low energy digital integrated circuits, Ph.D. diss., University of California Berkeley, Department of Electrical Engineering and Computer Science, 2013.

[13] K. Yasumura, T. Stowe, E. Chow, T. Pfafman, T. Kenny, B. Stipe, D. Rugar, Quality factors in micron- and submicron-thick cantilevers, J. Microelectromech. Syst. 9 (1) (2000) 117–125, http://dx.doi.org/10.1109/84.825786.

[14] D.W. Carr, S. Evoy, L. Sekaric, H.G. Craighead, J.M. Parpia, Measurement of mechanical resonance and losses in nanometer scale silicon wires, Appl. Phys. Lett. 75 (7) (1999) 920–922, http://dx.doi.org/10.1063/1.124554.

[15] Setting up a contact problem, 2019, https://doc.comsol.com/5.5/doc/com.comsol.help.sme/sme_ug_modeling.05.100.html.

[16] R. Nathanael, V. Pott, H. Kam, J. Jeon, T.-J.K. Liu, 4-terminal relay technology for complementary logic, in: 2009 IEEE International Electron Devices Meeting, IEDM, IEEE, 2009, pp. 1–4.

[17] A. Ballo, A.D. Grasso, G. Palumbo, A review of charge pump topologies for the power management of IoT nodes, Electronics 8 (5) (2019) 480.

[18] A. Kumar, C. Debnath, P.N. Singh, V. Bhatia, S. Chaudhary, V. Jain, S. Le Tual, R. Malik, A 0.065-mm 2 19.8-mw single-channel calibration-free 12-b 600-MS/s ADC in 28-nm UTBB FD-SOI using FBB, IEEE J. Solid-State Circuits 52 (7) (2017) 1927–1939.

[19] R. Holm, Electric Contacts: Theory and Application, Springer Science & Business Media, 2013.

[20] S.C. Bromley, B.J. Nelson, Performance of microcontacts tested with a novel MEMS device, in: Proceedings of the Forth-Seventh IEEE Holm Conference on Electrical Contacts (IEEE Cat. No. 01CH37192), IEEE, 2001, pp. 122–127.

[21] S.-F. Hsiao, J.-S. Yeh, D.-Y. Chen, High-performance multiplexer-based logic synthesis using pass-transistor logic, in: 2000 IEEE International Symposium on Circuits and Systems, vol. 2, ISCAS, 2000, pp. 325–328 vol.2, http://dx.doi.org/10.1109/ISCAS.2000.856327.

[22] Open source liberty specification, 2017, http://www.opensourceliberty.org/.

[23] J. Ragan-Kelley, C. Barnes, A. Adams, S. Paris, F. Durand, S. Amarasinghe, Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines, ACM SIGPLAN Notices 48 (6) (2013) 519–530.

[24] H. Kam, V. Pott, R. Nathanael, J. Jeon, E. Alon, T.-J.K. Liu, Design and reliability of a micro-relay technology for zero-standby-power digital logic applications, in: 2009 IEEE International Electron Devices Meeting, IEDM, IEEE, 2009, pp. 1–4.

[25] S.D. Senturia, Microsystem Design, Springer Science & Business Media, 2007.

[26] R. Dennard, F. Gaensslen, H.-N. Yu, V. Rideout, E. Bassous, A. LeBlanc, Design of ion-implanted MOSFET's with very small physical dimensions, IEEE J. Solid-State Circuits 9 (5) (1974) 256–268, http://dx.doi.org/10.1109/JSSC.1974.1050511.

[27] M. Ramezani, S. Severi, H.A.C. Tilmans, K.D. Meyer, Study of electrical breakdown and secondary pull-in failure modes for NEM relays, J. Micromech. Microeng. 27 (1) (2016) 015030, http://dx.doi.org/10.1088/1361-6439/27/1/015030.

[28] A. Partridge, H.-C. Lee, P. Hagelin, V. Menon, We know that MEMS is replacing quartz. But why? And why now? in: 2013 Joint European Frequency and Time Forum & International Frequency Control Symposium (EFTF/IFC), IEEE, 2013, pp. 411–416.

[29] S. Mishin, Y. Oshmyansky, F. Bi, Thickness control by ion beam milling in acoustic resonator devices, in: 2010 IEEE International Frequency Control Symposium, IEEE, 2010, pp. 642–645.