Embedded Nano-Electro-Mechanical Memory for Energy-Efficient Reconfigurable Logic

Kimihiko Kato, Vladimir Stojanović, and Tsu-Jae King Liu

***Abstract*— A compact, reconfigurable look-up table (LUT) implemented with an array of nano-electro-mechanical memory cells formed using CMOS back-end-of-line metal layers is proposed. The advantages of this LUT architecture over resistive memory- and CMOS-based implementations include faster and more energy-efficient operation. Pre-programmed answers can be read out within 1 ns, consuming less than 10 fJ.**

***Index Terms*— Internet of Things (IoT), look-up table (LUT), nanoelectromechanical systems (NEMS), non-volatile memory.**

I. Introduction

The future Internet of Things (IoT) will require embedded electronics to perform real-time computation on data with high energy-efficiency. For a conventional Von Neumann computer architecture, significant energy and time is spent not only for computing but also for data transfer between a central processing unit (CPU) and off-chip memory [1]. Therefore a novel "in-memory computing" (IMC) architecture that uses a resistive memory (ReRAM) array to implement a look-up table (LUT) was proposed recently to address these challenges [2], [3].

Nanometer-scale electro-mechanical memory (NEMory) cells which leverage contact adhesive forces or trapped charges to achieve bistable operation are ideally suited for IMC applications because they have near-zero leakage through nearly infinite resistance ratio between high-resistance (non-contacting) state and low-resistance (contacting) state [4]–[8], and also because they can be programmed with much lower energy than other non-volatile memory (NVM) devices [5]. It should be noted that a NEMory cell is essentially a reconfigurable interconnection; indeed, the back-end-of-line (BEOL) air-gapped metal wiring layers available in an advanced CMOS process can be used to implement NEMory cells with compact footprint [9]. In this letter, a NEMory-based reconfigurable logic LUT architecture and operating scheme is described and benchmarked against a conventional CMOS LUT architecture (Fig. 1) as well as the ReRAM-based LUT architecture.

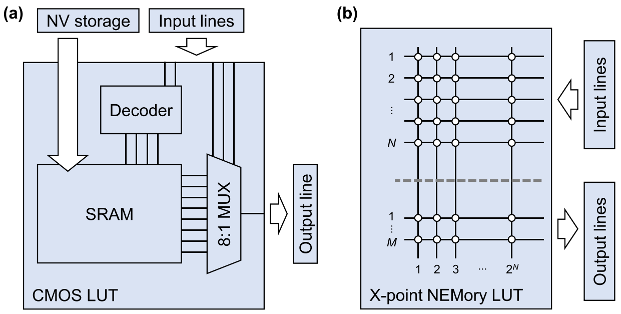


Fig. 1. Circuit block-level comparison of LUT implementations (a) CMOS based, and (b) NEMory based.

II. LUT Architecture and Operating Scheme

The NEMory-based LUT comprises a cross-point memory array with an input (Address) portion and an output (Result) portion, as illustrated in Fig. 2 for a 5-input/2-output LUT; the bit-lines in the input portion (IBLs) are connected to those in the output portion (OBLs) via gated CMOS buffers. The number of columns in the NEMory array is $N + M$, where N is the number of input bits and M is the number of output bits. Each memory cell comprises a vertically oriented movable electrode that is physically anchored at the bottom to the bit-line, and actuation electrodes (PL0 and PL1) on either side of the movable electrode implemented in intermediate metal layers; contacting electrodes (I/O0 and I/O1) on either side of the movable electrode are implemented in a top input/output metal layer. The PL0 and PL1 electrodes (not shown in Fig. 2, for clarity) are shared across the cells within a single column, and are used to set the state of each NEMory cell via electrostatic actuation to bring it into physical contact with either an I/O0 electrode or an I/O1 electrode. A non-linear device is assumed to be integrated into each NEMory cell, either at the bottom (Schottky contact) or at the top (metal-insulator-metal contact), to prevent sneak leakage paths in the cross-point array.

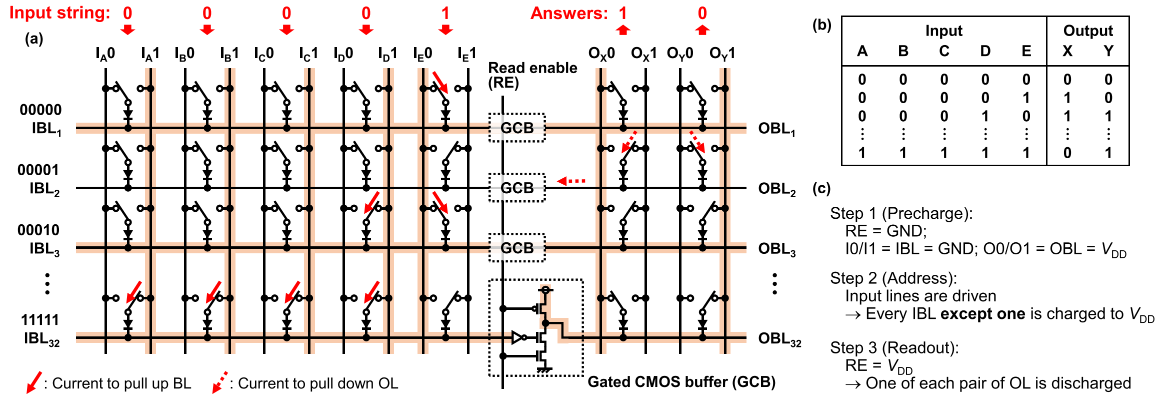
The number of rows in the NEMory array corresponds to the number of possible input bit combinations (up to 2^N); each input bit combination and its corresponding answer is programmed in the input portion and output portion, respectively, as follows: one IBL/OBL is grounded at a time to program the cells one row at a time; a programming voltage (V_{prog}) is applied to the PL0 actuation electrode in each column in which the cell is to be set to the "0" state; then V_{prog} is applied to the PL1 actuation electrode in the other columns to set the remaining cells to the "1" state. Note that the input/output electrodes are electrically floating during a programming

Manuscript received October 4, 2016; accepted October 22, 2016. Date of publication October 25, 2016; date of current version November 22, 2016. This work was supported by the NSF Center for Energy Efficient Electronics Science under Award 0939514. The work of K. Kato was supported by a post-doctoral fellowship from the Japan Society for the Promotion of Science. The review of this letter was arranged by Editor B. Govoreanu.

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: k.kato@eecs.berkeley.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LED.2016.2621187



Input					Output	
A	B	C	D	E	X	Y
0	0	0	0	0	0	0
0	0	0	0	1	1	0
0	0	0	1	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	1	1	1	0	1

Step 1 (Precharge):
RE = GND;
I0/I1 = IBL = GND; O0/O1 = OBL = V_{DD}

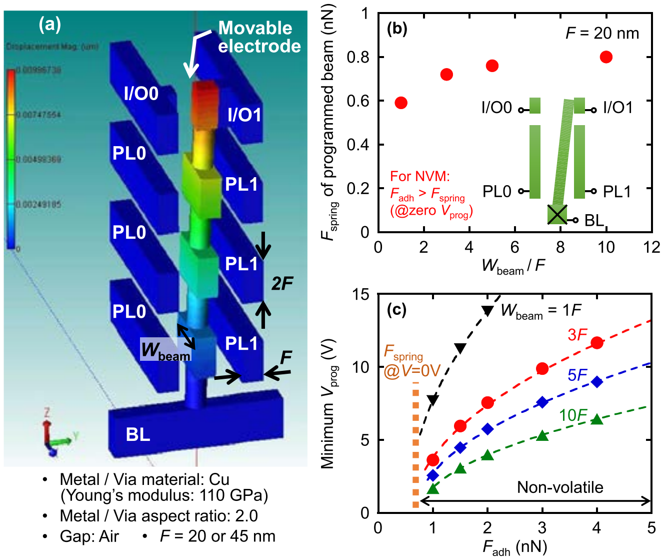
Step 2 (Address):
Input lines are driven
→ Every IBL **except one** is charged to V_{DD}

Step 3 (Readout):
RE = V_{DD}
→ One of each pair of OL is discharged

Fig. 2. (a) Circuit schematic for (5-input and 2-output) NEMory-based LUT, (b) corresponding truth table, and (c) readout scheme. Highlighted lines indicate driven wires (V_{DD} applied) for an input string of 00001 and answers 1 and 0, as an example.

operation, so that no direct current flows; this "cold-switching" operation provides not only for ultra-low-energy operation but also improved endurance [10].

A lookup operation involves 3 steps as follows: (1) with read enable line (RE) grounded, input lines (I0 and I1) and IBLs are all pre-discharged low (to GND) and output lines (O0 and O1) and OBLs are all pre-charged high (to V_{DD}) through a PMOSFET in the gated CMOS buffer; (2) the input lines are driven, causing all but one bit line – that is, the one corresponding to the input bit combination – to be charged high, as indicated by the arrows in Fig. 2(a); (3) the gated CMOS buffers are enabled (RE $= V_{DD}$), causing one of each pair of output lines to discharge toward GND according to the states of the cells connected to the one bit line that remained low, as indicated by the dotted arrows in Fig. 2(a), so that the result can be detected by the follow-on logic gates. It should be noted that, in principle, the electrical connections in the top electrode layer can be hardwired (*vs.* programmed) in the input portion of the array if there is no need for customization or reconfigurability. Furthermore, the number of output bits can be increased simply by adding output column(s) to the array, with no impact on computational throughput.



- Metal / Via material: Cu (Young's modulus: 110 GPa)
- Metal / Via aspect ratio: 2.0
- Gap: Air • F = 20 or 45 nm

Fig. 3. 3-D NEMory cell simulations: (a) Partial view of simulated structure. (b) Dependence of F_{spring} on W_{beam}. (c) Minimum voltage required to reprogram a NEMory cell (F = 20 nm), as a function of F_{adh}, for various values of W_{beam}.

III. NEMory Cell Design and Simulation

Fig. 3(a) shows a partial view of the three-dimensional (3-D) NEMory cell implemented using multiple BEOL metal interconnect layers. The electrode features in the intermediate metal layers are assumed to have width and spacing (actuation gap size) equal to the minimum lithographically defined feature size (F). The as-fabricated contact gap size in the top metal layer is assumed to be $F/2$, formed using a double-patterning technique such as described in [11], to ensure that contact is made only in the top layer, *i.e.* to avoid catastrophic pull-in of the structure to the actuation electrode. It should be noted that the CMOS tri-state gates (cf. Fig. 2) are fabricated in underlying layers and therefore do not require much extra layout area.

3-D device simulations using Coventor MEMS+ [12] indicate that the spring restoring force (F_{spring}) of the movable electrode is a relatively insensitive function of the electrode width (W_{beam}), as shown in Fig. 3(b). This is because the vias are the more compliant structural components, which have the greatest influence on F_{spring}. Note that the contact adhesive force (F_{adh}) must be greater than F_{spring} so that contact is maintained with no actuation voltage applied, *i.e.* for non-volatile operation.

The voltage required to reprogram a NEMory cell is investigated herein assuming F = 20 nm, cross-sectional aspect ratio (height/width of a layer feature) equal to 2, and copper (Young's modulus = 110 GPa) metal layers and vias. Fig. 3(c) plots the minimum reprogramming voltage, *i.e.* for which the electrostatic force (F_{elec}) plus F_{spring} is equal to F_{adh}. For nanometer-scale contact area, with F_{adh} in the range of a few nN [13], [14], the cell can be reprogrammed with less than 10 V. The catastrophic pull-in voltage is found to be approximately 4 times larger than the programming voltage.

IV. Performance Benchmarking

Fig. 4(a) shows the tradeoff between energy and delay for programming a NEMory cell. F_{adh} values of 1.5 and 6.0 nN are assumed for F values of 20 and 45 nm, respectively. It can

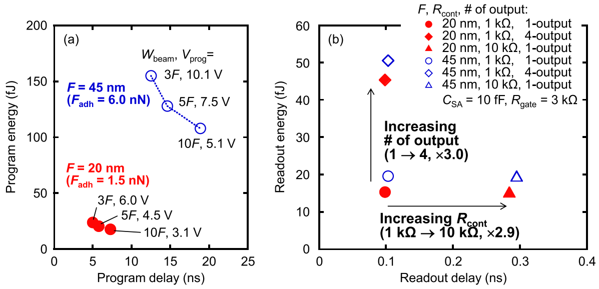


Fig. 4. (a) Calculated energy *vs.* delay for programming a NEMory cell. W_{beam} is varied from $3F$ to $10F$, and the minimum V_{prog} is assumed. (b) Energy and delay for readout operation of a NEMory array. W_{beam} is $3F$ and the contact resistance is 1 or 10 kΩ. 3 kΩ access transistor on-state resistance and 10 fF follow-on logic gate capacitance are assumed.

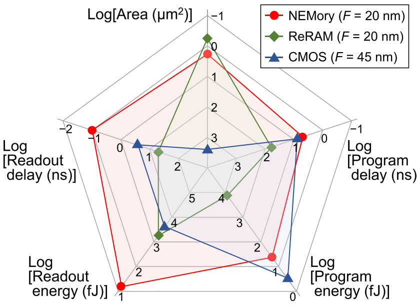


Fig. 5. Radar plot comparing the performance characteristics of NEMory-based LUT against those of ReRAM-based IMC and CMOS-based [15] LUT. For the ReRAM array, circuit layout based on [3], device performance based on [16], and readout time of 20 ns are assumed.

be seen that miniaturization (smaller actuation and contact gap sizes) is beneficial for reducing both the programming energy and delay. Fig. 4(b) shows the impact of increasing the number of output bits on the energy consumed during a readout operation, and the impact of higher contact resistance (R_{cont}) on the readout delay. (R_{cont} limits the rate at which a bit line is charged and the rate at which an output line is discharged.) The stored answers can be looked up in less than 1 ns (no matter the number of output bits) using much less than 1 pJ of energy for 4 output bits. The performance characteristics of NEMory-based LUT for 1 output bit are benchmarked against those of ReRAM-based and CMOS-based LUTs in Fig. 5. Much lower readout energy and delay – as well as zero standby power consumption – are remarkable advantages of the NEMory-based LUT. Practical challenges for implementation include precise control of contacting surface properties (roughness, adhesion energy, *etc*.) and potential reliability issues, which remain to be experimentally investigated.

V. Summary

A reconfigurable LUT architecture utilizing an array of re-programmable non-volatile NEM memory (NEMory) cells and a novel readout scheme is presented. As compared with ReRAM-based and CMOS-based LUTs, NEMory-based LUT is projected to be 10× faster and 100× more energy-efficient, while achieving similar high density to ReRAM, making it a compelling approach to energy-efficient computing in the era of the Internet of Things.

References

[1] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, and K. Yelick, "ExaScale computing study: Technology challenges in achieving exascale systems," in *Proc. Defense Adv. Res. Projects Agency Inf. Process. Tech. Office (DARPA IPTO)*, Sep. 2008, p. 15. [Online]. Available: http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf

[2] M. Sharad, D. Fan, K. Aitken, and K. Roy, "Energy-efficient non-Boolean computing with spin neurons and resistive memory," *IEEE Trans. Nanotechnol.*, vol. 13, no. 1, pp. 23–34, Jan. 2014.

[3] B. Chen, F. Cai, J. Zhou, W. Ma, P. Sheridan, and W. D. Lu, "Efficient in-memory computing architecture based on crossbar arrays," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2015, pp. 17.5.1–17.5.4, doi: 10.1109/IEDM.2015.7409720.

[4] K. Kato, V. Stojanović, and T.-J. K. Liu, "Non-volatile nano-electro-mechanical memory for energy-efficient data searching," *IEEE Electron Device Lett.*, vol. 37, no. 1, pp. 31–34, Jan. 2016, doi: 10.1109/LED.2015.2504955.

[5] N. Xu, J. Sun, I.-R. Chen, L. Hutin, Y. Chen, J. Fujiki, C. Qian, and T.-J. K. Liu, "Hybrid CMOS/BEOL-NEMS technology for ultra-low-power IC applications," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2014, pp. 28.8.1–28.8.4, doi: 10.1109/IEDM.2014.7047130.

[6] B. W. Soon, E. J. Ng, Y. Qian, N. Singh, M. J. Tsai, and C. Lee, "A bi-stable nanoelectromechanical non-volatile memory based on van der Waals force," *Appl. Phys. Lett.*, vol. 103, no. 5, p. 053122, Jul. 2013, doi: 10.1063/1.2360143.

[7] W. Y. Choi, H. Kam, D. Lee, J. Lai, and T.-J. K. Liu, "Compact nano-electro-mechanical non-volatile memory (NEMory) for 3D integration," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2007, pp. 603–606, doi: 10.1109/IEDM.2007.4419011.

[8] Y. Tsuchiya, K. Takai, N. Momo, T. Nagami, H. Mizuta, S. Oda, S. Yamaguchi, and T. Shimada, "Nanoelectromechanical nonvolatile memory device incorporating nanocrystalline Si dots," *J. Appl. Phys.*, vol. 100, no. 9, p. 094306, Nov. 2006, doi: 10.1063/1.2360143.

[9] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, A. Dasgupta, K. Fischer, Q. Fu, T. Ghani, M. Giles, S. Govindaraju, R. Grover, W. Han, D. Hanken, E. Haralson, M. Haran, M. Heckscher, R. Heussner, P. Jain, R. James, R. Jhaveri, I. Jin, H. Kam, E. Karl, C. Kenyon, M. Liu, Y. Luo, R. Mehandru, S. Morarka, L. Neiberg, P. Packan, A. Paliwal, C. Parker, P. Patel, R. Patel, C. Pelto, L. Pipes, P. Plekhanov, M. Prince, S. Rajamani, J. Sandford, B. Sell, S. Sivakumar, P. Smith, B. Song, K. Tone, T. Troeger, J. Wiedemer, M. Yang, and K. Zhang, "A 14 nm logic technology featuring $2^{n}d$-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2014, pp. 3.7.1–3.7.3, doi: 10.1109/IEDM.2014.7046976.

[10] Y. Chen, "Reliability studies of micro-relays for digital logic applications," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California at Berkeley, Berkeley, CA, USA, 2015.

[11] S.-W. Kim, P. Zheng, K. Kato, L. Rubin, and T.-J. King Liu, "Tilted ion implantation as a cost-efficient sublithographic patterning technique," *J. Vac. Sci. Tech. B*, vol. 34, no. 4, p. 040608, Jul./Aug. 2016, doi: 10.1116/1.4953085.

[12] *Coventor MEMS+ User Guide*, Coventor, Singapore, 2013.

[13] J. Yaung, L. Hutin, J. Jeon, and T.-J. K. Liu, "Adhesive force characterization for MEM logic relays with sub-micron contacting regions," *IEEE/ASME J. Microelectromech. Syst.*, vol. 23, no. 1, pp. 198–203, Feb. 2014. doi: 10.1109/JMEMS.2013.2269995

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

[15] S. Paul and S. Bhunia, *Computing with Memory for Energy-Efficient Robust Systems*. New York, NY, USA: Springer, 2014.

[16] B. Govoreanu, G. S. Kar, Y.-Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. J. Wouters, J. A. Kittl, and M. Jurczak, "10×10 nm² Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2011, pp. 31.6.1–31.6.4, doi: 10.1109/IEDM.2011.6131652.