Non-Volatile Ternary Content Addressable Memory based on Phase Change Nanoelectromechanical (NEM) Relay

Mohammad Ayaz Masud
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
mmasud@andrew.cmu.edu

Luis Hurtado
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
lhurtado@andrew.cmu.edu

Gianluca Piazza
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
piazza@ece.cmu.edu

***Abstract*— We demonstrate, for the first time, a ternary content-addressable memory (TCAM) architecture based on phase change nanoelectromechanical relays (PCNRs). The non-volatility (NV), high ON-OFF ratio (10^8), and low leakage operation make PCNR an ideal candidate for high density TCAM. Additionally, PCNR devices are back-end-of-the-line (BEOL) compatible, allowing for a very small TCAM cell size of $18F^2$. A TCAM, with only 1 transistor and 2 PCNR devices (1T2P) per cell is simulated and it exhibits 133 ps search latency and 0.721 pJ energy consumption for 64 bits, making it one of the most competitive approaches for TCAM using beyond CMOS technologies.**

Keywords—Phase change material, TCAM, GeTe, NEMS

I. Introduction

The data-driven revolution in modern computational devices has reached a bottleneck where the energy-time cost of memory access is higher than the cost associated with processing. Applications that enable computation at edge devices, such as IoT sensors, suffer the most in this regard due to their limited energy resources. Architectural and device level solution is essential to mitigate this problem. One way of achieving this goal is to utilize innovative hardware designs to enable parallel access to an array of memory cells in one single search cycle. Ternary content addressable memory (TCAM) is one such circuit technology that comes into play during high frequency look-up operations [1]. It searches for a content in a 1D array and returns the address when a match is found, greatly reducing the time and energy requirement of a conventional search architecture. This architecture was originally used in network systems for operations like internet protocols (IP) look-up [2]. However, in recent years, TCAM has been employed for broader data extensive applications such as genome data analysis, natural language processing (NLP), image classification etc [3]. The on-chip circuitry also has the potential to resolve the delay associated with communication between the processor and the memory units.

CMOS-based TCAM is typically based on 16 transistor static random-access memory (16T SRAM) cells, resulting in high energy consumption and large cell area [4]. Implementation with dynamic memory (DRAM) requires frequent refresh cycles, offsetting the improvement of having only 5 transistors [5]. To solve these issues, alternative TCAM approaches utilizing several emerging devices, such as, STT-MRAM, FeFET, RRAM, electromechanical relay etc. have been studied [6-9]. The proposed approaches suffer from various limitations associated with these emerging technologies. For example, resistive switching suffers from high dynamic and static power consumption as well as long latency while having a very low ON-OFF ratio [10]. Devices based on ferroelectric switching can provide higher ON-OFF ratio and can be used in a dense 2T TCAM architecture [11]. However, they also require high energy budget and long latency. Recently, TCAM designs have been presented with nanoelectromechanical (NEM) relay to mitigate some of these issues [12]. The ultra-high ON-OFF ratio, near-zero leakage and abrupt switching mechanism make NEM relays a suitable technology for the TCAM architecture. However, conventional NEM relay designs depend on flexure-based architecture which cannot be scaled down to sub-100 nm node sizes [13]. Moreover, most NEM relay designs are volatile and need periodic refresh cycles to hold data [14].

To solve the scalability and volatility issue of conventional NEM relays, we demonstrate a non-volatile NEM relay based on the volumetric expansion of GeTe phase change material (PCM) and implement it in a TCAM cell to minimize cell area and search energy consumption. The non-volatile nature of the relay eliminates the need for refresh cycles, while the high scalability and the low leakage enable dense TCAM architecture. In a recent work, we have presented a PCM based NEM relay architecture, called Phase Change Nanoelectromechanical Relay (PCNR), which shows very high ON-OFF ratio (10^8) and very low leakage current (30 fA) at the cost of 42 nJ writing energy [15]. Furthermore, PCNR devices are non-volatile and highly scalable. This makes PCNR an exciting new candidate for TCAM architectures. In the following sections of this paper we briefly discuss the operating principle of PCNR and demonstrate our simulation results for a 1T 2 PCNR TCAM cell. We compare the performance of this TCAM design with existing architectures.

II. Device Design and Operation

PCMs are an exciting class of materials renowned for their rapid and reversible switching between amorphous and crystalline states. This change in nanoscale atomic arrangement alters some of the most useful physical macroscale properties, such as electrical conductivity, optical reflectance, density etc. As the material undergoes an amorphous-to-crystalline transition, its resistivity decreases by ~5 orders of magnitude and optical reflectance may increase by 10-15% [16]. Several Te-based chalcogenides, such as Ge-Sb-Te (GST) and GeTe, have been successfully used in optical and electrical data storage devices [17]. Phase change nano relay (PCNR) utilizes the density difference between two stable phases of a PCM in order to switch a relay device. Fig. 1 shows the structure of a PCNR cell. It is primarily comprised of a phase change mechanical actuator and a pair of metal contacts [18].

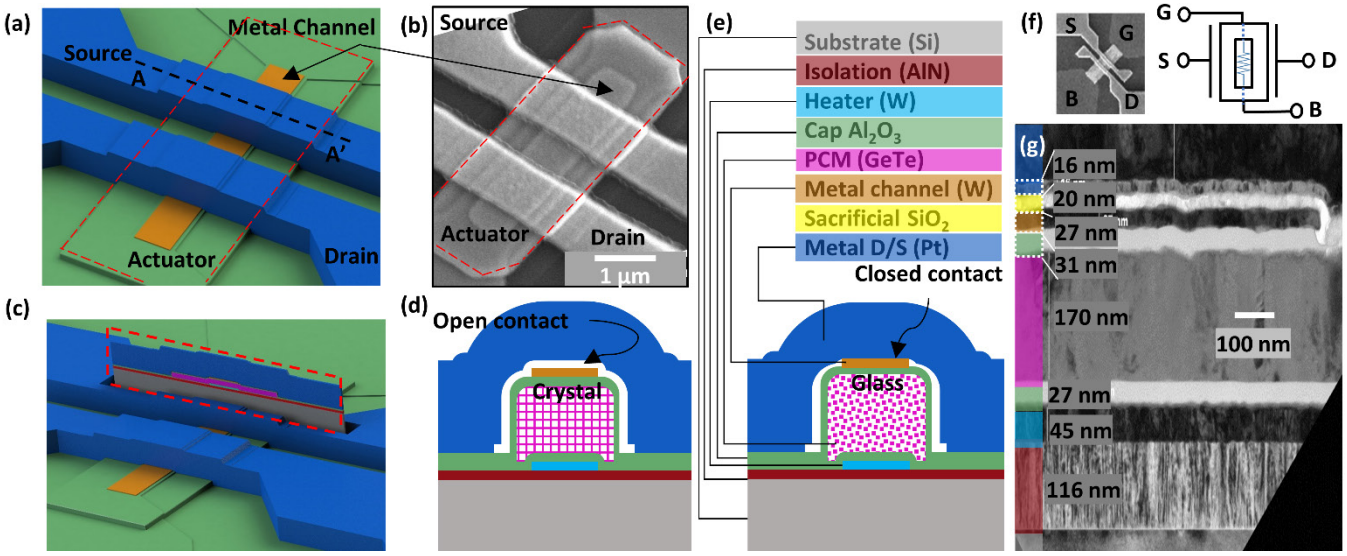


Fig. 1: Overview of PCNR device architecture and operation. (a) 3D view and (b) SEM image of a PCNR. (c) Cross-section taken along AA' line. (d)-(e) schematic cross-section with (d) open and (e) closed contact. (f) Four I/O terminals of PCNR, along with its schematic model, showing the source (S), drain (D), gate (G) and body (B). G and B connects the heater. (g) TEM image of the cross-section before release.

The key component of the phase change actuator is GeTe, which expands by 10% during its transformation from crystalline to amorphous phase [19]. The change is induced by Joule-heating from an adjacent heater layer (Gate terminal). A metal channel is then fabricated on top of the actuator. As fabricated, the channel is separated from the drain and source (D/S) metal lines by an air-gap. The air-gap is obtained by depositing a 20 nm sacrificial layer between the channel and the D/S metal layers. The fabrication process of PCNR is significantly different from conventional PCM devices, as we do not need conduction through the PCM.

In the phase-change process, a short intense thermal pulse melts the crystalline material. If the material is cooled slowly, it goes back to the original crystalline state. However, if the material is quenched fast (>10^9 K/s), the atoms get locked in a supercooled state [20]. This sudden quench leaves the material in an amorphous, highly resistive state. The amorphous state can be retained for more than 10 years at a temperature below 400 K [21]. To return to its crystalline phase, a longer pulse of lower intensity is applied to anneal the material at a temperature above the glass temperature which is lower than the melting point. GeTe melts at 1000 K and its glass transition happens at 500 K [22]. This transition phenomenon is also evident in atomically thin layers of PCM [23].

A. *SET and RESET Operation*

The amplitude and duration of the heater pulse determines the state of a PCNR cell during the write operation. SET pulse should be high enough to melt the PCM. Here we apply an 1.1 V input pulse on a 3 µm long heater to reach the melting temperature of GeTe (Fig. 2(a)). An abrupt falling edge in the SET pulse allows rapid quench. The PCM is transformed into the expanded amorphous state by the end of this heat cycle. This expansion switches the device ON by connecting the metal contact with both electrodes, writing a '1'. The RESET pulse is tuned to achieve a temperature between the glass transition temperature and the melting point of the PCM. It has a lower amplitude (0.8 V) compared to the SET pulse (Fig. 2(b)). A slower pulse is used to ensure complete crystallization of the PCM. Fig. 2(c) demonstrates the high ON-OFF ratio of PCNR. We test for leakage current for a range of read voltages. Fig. 2(d) shows that the leakage current

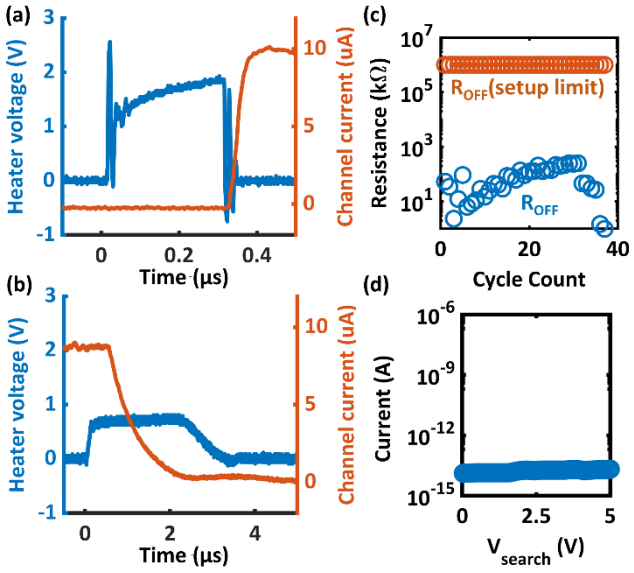


Fig. 2: Electrical characterization of a PCNR device. (a) SET and (b) RESET pulse on PCNR with 3 µm long heater. Channel current is measured from a 100 kΩ resistor in series with the drain (V_{DS}=1V). (c) ON-OFF resistance during cycling and (d) leakage current measured in the channel for a switched OFF device.

remains unchanged even for impractically high search voltages.

B. *Fabrication Process*

Fig. 3 shows the process flow of the PCNR device. At first, 100 nm AlN is deposited on a Si substrate by reactive sputtering. AlN provides electrical isolation between the heater and the substrate while allowing good thermal conduction. To obtain the heater layer, 50 nm W is sputtered at 850°C and etched in a Reactive Ion Etching (RIE) tool. Next, 30 nm Al_2O_3 is deposited in an atomic layer deposition (ALD) tool at 250°C to electrically isolate the PCM from the heater. After this step, 200 nm of crystalline GeTe is deposited by co-sputtering Ge and Te at 400°C, and it is then etched by Ar^+ plasma in an Inductively Coupled Plasma (ICP)-RIE chamber.

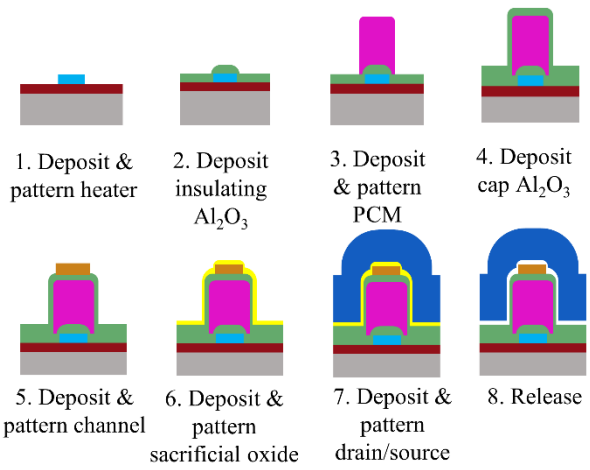
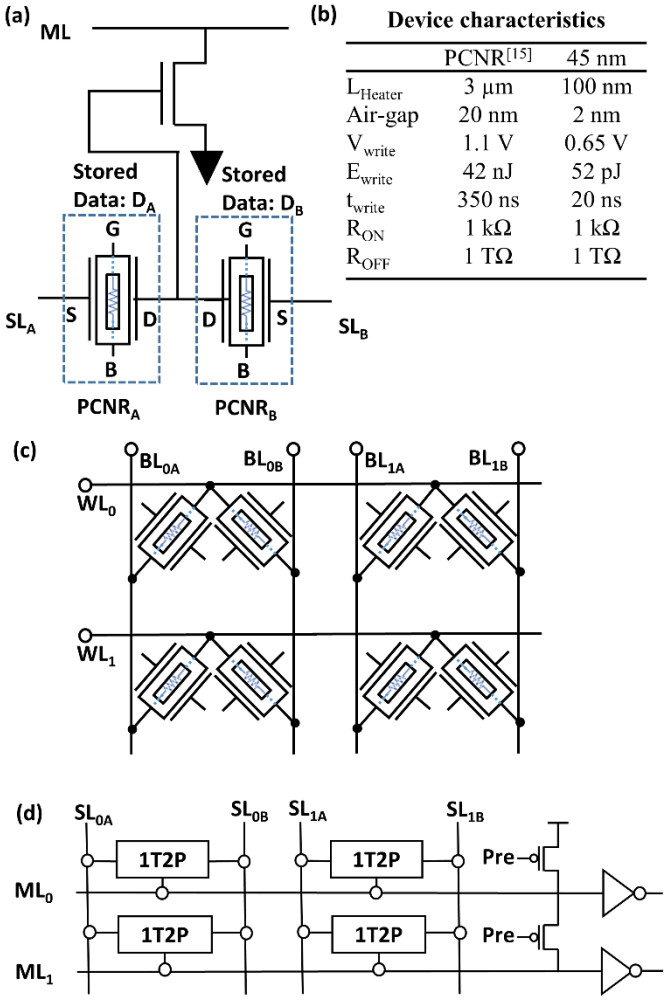


Fig. 3: BEOL compatible fabrication flow of the PCNR device.



	PCNR[15]	45 nm
L_{Heater}	3 µm	100 nm
Air-gap	20 nm	2 nm
V_{write}	1.1 V	0.65 V
E_{write}	42 nJ	52 pJ
t_{write}	350 ns	20 ns
R_{ON}	1 kΩ	1 kΩ
R_{OFF}	1 TΩ	1 TΩ

Fig. 4: PCNR TCAM architecture. (a) A TCAM cell with 1 transistor and 2 PCNR devices. (b) PCNR device characteristics obtained at 45 nm technology node. (c) Write and (d) search circuit of a simple 2×2 array. The two circuits are shown separately to help explain their independent operation. Notice that the four terminals of PCNR can be connected to the search and write circuit on a single layout.

Another 30 nm layer of Al_2O_3 is deposited by ALD to encapsulate the patterned PCM. This encapsulation layer contains the molten PCM during phase transformation. At this point, the devices are ready to be used and tested as stand-alone actuators.

Next, the channel is made by sputtering 30 nm W at room temperature. It is then patterned and etched following the same W-etch process. The airgap is formed by depositing 20 nm sacrificial SiO_2 in the ALD and patterning it in RIE. Finally, a thick layer of W is lifted-off as the D/S pair. The wafer is then etched in vapor HF to isotropically remove the sacrificial oxide.

III. PCNR TCAM Architecture

Fig. 4 shows a PCNR TCAM cell and a simple 2x2 array during write and search operation. In the TCAM cell, PCNR devices are modeled as resistors with either low resistive state (~1kΩ) representing a logic 1 or a high resistive state (~1TΩ) representing a logic 0. The TCAM architecture assumes that both PCNR devices will not be at simultaneous low resistive state. The PCNR are BEOL compatible, leading to a cell area of $18F^2$, only 10% of the 16T TCAM cell area.

The writing portion of the PCNR devices is decoupled from the read and search operations and the cross-bar array configuration, shown in Fig. 4(c), offers minimal cell area. The non-volatile nature of the PCNR device allows writing operation at the heater crossbar array without the need of select transistors [24]. Specific word line (WL) and bit line (BL) are selected to apply a heater current on a desired device. Adjacent heaters are exposed to a fraction of the input write current due to sneak path formation. Typically, this current level is not high enough to cause inadvertent switching in adjacent cells. However, write energy increases significantly due to the undesired heat dissipation in adjacent cells. This poses a design trade-off between high cell density in selector-free architecture and low write energy consumption. Writing energy required for a single PCNR cell at 45 nm technology node is 52 pJ.

The search circuit includes the PCNR TCAM cell, pre-charging circuit, and the inverter sensing amplifier (SA). The search lines are connected to the source end of the PCNR channel. TCAMs are benchmarked based on energy consumption and delay during the search operation [25]. The search operation of the PCNR TCAM is shown in Fig. 5 with corresponding simulation set-up identical to Fig. 4(d), based on 45nm CMOS technology node.

During the search operation, the match-line (ML) is first charged to V_{DD} and is then left floating. If there is a mismatch

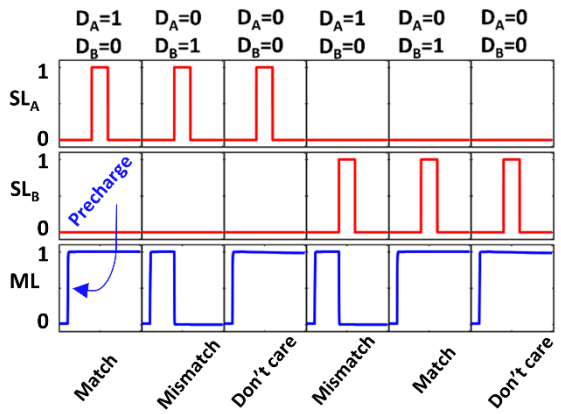


Fig. 5: Logic of the search operation in PCNR TCAM. Match line (ML) is precharged at the beginning of every search cycle. Search lines SL_A and SL_B are symmetric. For a mismatch between the SL and corresponding stored PCNR data, ML discharges within 133 ps for a 64 bit word line.

detected between the search-line (SL) input and the stored data (D) then the ML is pulled to ground via the transistor, otherwise the ML maintains the charge and results in a match. A capacitance of 0.2 fF/μm is used to represent the parasitics of the ML in the simulation setup. The worst case is denoted by discharge through a single transistor in a long array of data. This happens when only one bit has a mismatch. The search latency is calculated for the worst case of only 1-bit mismatch in a 64 bit word row and search energy is calculated based on the pre-charging of the ML for all the entries. Our simulation results exhibit 133 ps search latency and 0.721 pJ energy consumption for 64 bit word row. An overview of TCAM technology benchmark is presented in Table 1.

TABLE I. PERFORMANCE SUMMARY OF TCAM DESIGNS

	16T [4]	2T2R [5]	2FeFET [7]	3T2N [9]	**1T2P**
Node (nm)	45	90	45	45	**45**
Non-Volatility	No	Yes	Yes	No	**Yes**
Search Voltage (V)	1	1.2	1	1	**1**
Write Voltage (V)	1	2.5	4	1	**1.1**
Search delay (ps)	582	1900	~400	106	**133**
Search energy (fJ)	1600	-	~1700	693	**721**
Area	$171F^2$	$50F^2$	$22F^2$	$32F^2$	**$18F^2$**

IV. CONCLUSION

We present the PCNR TCAM design and simulation results, exhibiting the lowest energy-delay product per unit area in comparison to available TCAM designs. The device itself offers a high ON-OFF ratio, non-volatility, and low leakage. The prototype PCNR device requires high write energy. However, scaling analysis at lower technology nodes promises pJ energy consumption during writing. Further investigation in scaling PCNR and increasing its endurance will pave the way for most energy efficient TCAM architecture.

ACKNOWLEDGMENT

The authors acknowledge the Kavcic-Moura Endowment Fund and the National Science Foundation, USA (Award 1854702) for supporting this work. The authors are grateful to the staff members of the cleanroom facility at Claire & John Bertucci Nanotechnology Laboratory in Carnegie Mellon University. We also acknowledge the use of the CMU Materials Characterization Facility at Carnegie Mellon University supported by grant MCF-677785.

REFERENCES

[1] R. Yang *et al.*, "Ternary content-addressable memory with MoS2 transistors for massively parallel data search," *Nat Electron*, vol. 2, no. 3, pp. 108–114, Mar. 2019.

[2] R. Panigrahy and S. Sharma, "Reducing TCAM power consumption and increasing throughput," in *Proceedings 10th Symposium on High Performance Interconnects*, Stanford, CA, USA, 2002, pp. 107–112.

[3] W. Jiang, Q. Wang and V. K. Prasanna, "Beyond TCAMs: An SRAM-Based Parallel Multi-Pipeline Architecture for Terabit IP Lookup," IEEE INFOCOM 2008 - The 27th Conference on Computer Communications, 2008, pp. 1786-1794.

[4] K. Pagiamtzis and A. Sheikholeslami, "Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, Mar. 2006.

[5] J. Li, R. K. Montoye, M. Ishii, and L. Chang, "1 Mb 0.41 μm^2 2T-2R Cell Nonvolatile TCAM With Two-Bit Encoding and Clocked Self-Referenced Sensing," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 896–907, Apr. 2014.

[6] B. Yan, Z. Li, Y. Chen, and H. Li, "RAM and TCAM designs by using STT-MRAM," in *2016 16th Non-Volatile Memory Technology Symposium (NVMTS)*, Pittsburgh, PA, USA, Oct. 2016, pp. 1–5.

[7] X. Yin, X. Chen, M. Niemier, and X. S. Hu, "Ferroelectric FETs-Based Nonvolatile Logic-in-Memory Circuits," *IEEE Trans. VLSI Syst.*, vol. 27, no. 1, pp. 159–172, Jan. 2019.

[8] L. Zheng, S. Shin, S. Lloyd, M. Gokhale, K. Kim, and S.-M. Kang, "RRAM-based TCAMs for pattern search," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, Montréal, QC, Canada, May 2016, pp. 1382–1385.

[9] H. Zhong *et al.*, "DyTAN: Dynamic Ternary Content Addressable Memory Using Nanoelectromechanical Relays," *IEEE Trans. VLSI Syst.*, vol. 29, no. 11, pp. 1981–1993, Nov. 2021.

[10] M.-F. Chang *et al.*, "17.5 A 3T1R nonvolatile TCAM using MLC ReRAM with Sub-1ns search time," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, San Francisco, CA, USA, Feb. 2015, pp. 1–3.

[11] M. Yabuuchi, M. Morimoto, Y. Tsukamoto, and S. Tanaka, "A 7nm Fin-FET 4.04-Mb/mm2 TCAM with Improved Electromigration Reliability Using Far-Side Driving Scheme and Self-Adjust Reference Match-Line Amplifier," in *2020 IEEE Symposium on VLSI Circuits*, Honolulu, HI, USA, Jun. 2020, pp. 1–2.

[12] H. Zhong, S. Cao, H. Yang, and X. Li, "Dynamic Ternary Content-Addressable Memory Is Indeed Promising: Design and Benchmarking Using Nanoelectromechanical Relays," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Grenoble, France, Feb. 2021, pp. 1100–1103.

[13] A. Peschot, C. Qian, and T.-J. Liu, "Nanoelectromechanical Switches for Low-Power Digital Computing," *Micromachines*, vol. 6, no. 8, pp. 1046–1065, Aug. 2015.

[14] T. Qin, S. J. Bleiker, S. Rana, F. Niklaus, and D. Pamunuwa, "Performance Analysis of Nanoelectromechanical Relay-Based Field-Programmable Gate Arrays," *IEEE Access*, vol. 6, pp. 15997–16009, 2018.

[15] J. T. Best, M. A. Masud, M. P. Boer, and G. Piazza, "Phase Change Nanoelectromechanical Relay for Nonvolatile Low Leakage Switching," *Adv Elect Materials*, p. 2200085, Apr. 2022.

[16] S. Raoux, W. Wełnic, and D. Ielmini, "Phase Change Materials and Their Application to Nonvolatile Memories," *Chem. Rev.*, vol. 110, no. 1, pp. 240–267, Jan. 2010.

[17] R. Jeyasingh *et al.*, "Ultrafast Characterization of Phase-Change Material Crystallization Properties in the Melt-Quenched Amorphous Phase," *Nano Lett.*, vol. 14, no. 6, pp. 3419–3426, Jun. 2014.

[18] J. T. Best, M. A. Masud, M. P. de Boer, and G. Piazza, "Phase Change NEMS Relay," in *2019 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, Dec. 2019, p. 34.1.1-34.1.4.

[19] J. Best and G. Piazza, "High Work Density Gete Mechanical Phase Change Actuator," in *2019 IEEE 32nd International Conference on Micro Electro Mechanical Systems (MEMS)*, Seoul, Korea (South), Jan. 2019, pp. 962–965.

[20] J. Pries, S. Wei, M. Wuttig, and P. Lucas, "Switching between Crystallization from the Glassy and the Undercooled Liquid Phase in Phase Change Material $Ge_2Sb_2Te_5$," *Adv. Mater.*, vol. 31, no. 39, p. 1900784, Sep. 2019.

[21] A. Fantini *et al.*, "Comparative Assessment of GST and GeTe Materials for Application to Embedded Phase-Change Memory Devices," in *2009 IEEE International Memory Workshop*, Monterey, CA, USA, May 2009, pp. 1–2.

[22] J. Pries *et al.*, "Approaching the Glass Transition Temperature of GeTe by Crystallizing $Ge_{15}Te_{85}$," *Physica Rapid Research Ltrs*, vol. 15, no. 3, p. 2000478, Mar. 2021.

[23] F. Xiong *et al.*, "Towards ultimate scaling limits of phase-change memory," in *2016 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, Dec. 2016, p. 4.1.1-4.1.4.

[24] A. Chen, "Analysis of Partial Bias Schemes for the Writing of Crossbar Memory Arrays," *IEEE Trans. Electron Devices*, vol. 62, no. 9, pp. 2845–2849, Sep. 2015.

[25] B. Agrawal and T. Sherwood, "Modeling TCAM power for next generation network devices," in *2006 IEEE International Symposium on Performance Analysis of Systems and Software*, Austin, TX, USA, 2006, pp. 120–129.