Functionality Enhanced Memories for Edge-AI Embedded Systems

Alexandre Levisse*, Marco Rios*, W.-A. Simon*, P.-E. Gaillardon†, and D. Atienza*
*ESL, Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
†Laboratory for NanoIntegrated Systems (LNIS), University of Utah, USA

***Abstract*—With the surge in complexity of edge workloads, it appeared in the scientific community that such workloads cannot be anymore overflown to the cloud due to the huge edge device to server communication energy cost and the high energy consumption induced in high end server infrastructure. In this context, edge devices must be able to efficiently process complex data-intensive workloads bringing in the concept of Edge AI. However, current architectures show poor energy efficiency while running data intensive workloads. While the community looks toward the integration of new memory architectures using emerging resistive memories and new specific accelerators, we propose a new concept to boost the energy efficiency of Edge systems running data intensive workloads : Functionality Enhanced Memories (FEM). FEM consist on a memory architecture with new functionalities at a decent area overhead cost. In this work, we demonstrate the feasibility of native transpose access for 1Transistor-1RRAM bitcells leveraging three independent gates transistors. Based on that, we thereby propose a concept of FEM-enabled Edge system embedding the proposed native transpose access RRAM-based memory architecture and an in-SRAM computing architecture (the BLADE).**

***Index Terms*—RRAM, 1T1R, TIGFET, Functionality Enhanced Devices, Functionality Enhanced Memories.**

I. INTRODUCTION

With the foreseen arrival of edge computing devices that utilize complex machine learning algorithms in the consumer market, requirements for embedded devices in terms of memory capacity, processing capability and energy efficiency are skyrocketing. In this context, industry and academia are looking for new computing and memory technologies and architectures that can enable both dense and energy efficient architectures. On one hand, Functionality Enhanced Devices (FED) such as Three Independent Gate Field Effect Transistors (TIGFET) [1] are perceived as a promising opportunity as they (i) are a direct evolution from FinFET technology and (ii) enable dense digital design thanks to their various functionalities such as polarity, sub-threshold slope and threshold voltage control [2]. On the other hand, emerging resistive memory technologies (RRAM) such as filamentary-based RRAM are already penetrating the market as they provide easy technology co-integration with MOS technologies, middle programming voltage and fast switching capabilities [3], [4]. Finally, new breakthrough in-SRAM computing architectures [5], [6], [7] enable new opportunities in computing data-centric workloads in a highly efficient way.

In this work, we propose to extend the concept of FED to the concept of Functionality Enhanced Memories (FEM), which we define as a memory array that provides new functionalities thanks to new technology or architectural innovations. The main motivation comes from the fact that direct scaling does not solve issues associated to data-centric Edge AI workloads and does comes at the cost of increased technology and design costs. We thereby propose to reach a "dense enough"-low cost memory integration density, and then to add it functionalities, making it a FEM and enabling it strong performance and energy efficiency gains. In this context, as presented Figure 1, Edge systems energy efficiency can be improved towards two directions : (i) The introduction of non-volatile memories in the cache hierarchy to mitigate the static leakage during both sleep and active periods. (ii) The implementation of workload specific accelerators (such as in-cache computing) to improve the computing efficiency during active periods. Overall, these enhancements move the compute/store and Volatile/Non-Volatile (VM/NVM) memories limits and makes them closer to each other, opening new perspective for architecture and circuit designers but also opening new questions (reliability for e.g.). The main focus of this paper is to give circuit perspectives for embedded systems Edge AI towards embedding FEM in the two following research directions :

- The functionality enhancement of NVM memories by the integration of transpose access capabilities. We thereby propose a native transposed access memory array enabling both horizontal and vertical data access leveraging TIGFET technology. This new functionality is enabled by the independent use of TIGFET's polarity gates and can be used to enable (i) direct (LSB/MSB) comparisons or (ii) transposed access in the case of 1 word per bitcell (i.e., binarized data or Multiple Level Cells, MLC, RRAM). Finally, SiNWFET-based RRAM array enable strong energy gains [8] and we propose to explore macro-level consideration regarding sense and write amplifiers overhead.
- The functionality enhancement of caches. We thereby highlight the BLADE architecture [5], [6] as a functionality enhanced cache architecture enabling in-memory computing while featuring high density SRAM bitcell, reliable operation and wide voltage range functionality.

The remainder of the paper is organized as follows. Section II presents the background of the paper, presenting the context RRAM technologies, and functionality enhanced devices. Section III presents the concept of transpose access, introduces

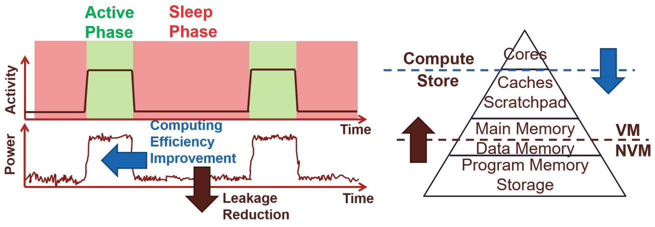

Fig. 1. Edge Devices workload representation with associated circuit and architectural innovations effect.

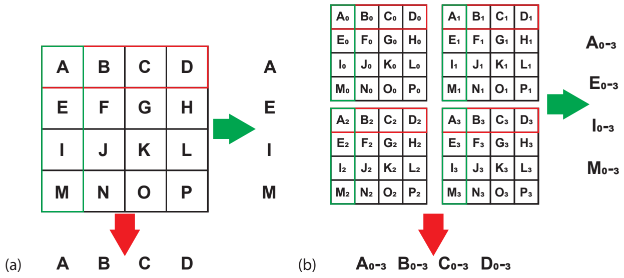

Fig. 2. Transpose access array concept for (a) binarized data and (b) multiple-bit data.

the proposed TIGFET-based architecture and highlights its potential. Section IV discusses the concept of Functionality Enhanced Memories (FEM) for Edge AI system. Finally, Section V concludes the paper.

II. Background

A. RRAM technologies

In order to achieve high density, low cost, high granularity embedded Non-Volatile Memory (eMVM) integration, the scientific community have been rushing into back-End-of-Line (BEoL) integration for eNVM technologies in the last 5 to 10 years. In this context, Resistive Random Access Memory (RRAM) technologies are seen as a future enabler for low-power embedded systems as they could be integrated inside the memory hierarchy, thereby enabling high density and near-zero leakage caches [4], [3]. Furthermore, the introduction of Edge AI triggers the need for high quantities of non-volatile memory to enable local and low energy weight storing. However, RRAM technologies suffer from limited endurance, forbidding their use as computation memories, thereby calling for reasonable usage as program, data or last level cache memory. Popular technologies such as Spin-Transfer Torque Magnetic RRAM (MRAM) [9], Phase Change Memories (PCM) [10] or filamentary-based RRAM (ReRAM) [3] are under intense exploration by the scientific community and some (PCM, ReRAM) are currently being integrated in commercial micro-controllers as a replacement for eflash technologies [11], [12], [13]. However, the jump towards intensive usage has not been achieved yet, and RRAM technologies are still considered as a regular eFlash replacement. From an electrical point of view, RRAM memories technologies can be programmed by applying a voltage across their two electrodes (top and bottom)

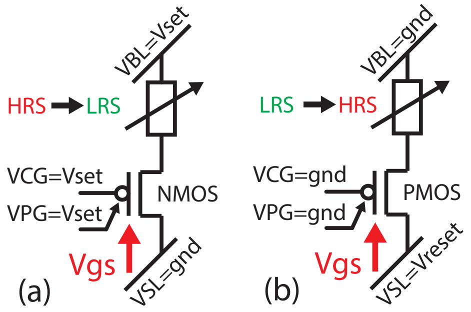

Fig. 3. use of a tigfets for high efficiently programming operations in bipolar rram technology-based arrays.

and controlling the current going through the device (slight different from technologies to technologies can be noted here). The achieved resistance state is non-volatile (it's endurance and retention is highly correlated to the programming energy) [14], [15]. Hence, targeting a RRAM technology for storage forbids its use as a computation element (and vice versa).

B. Polarity Controllable Transistors

While fabrication costs of deeply scaled transistors technologies (sub-10nm) becomes hard to sustain for most of historical foundries, the scientific community tends to look toward other kind of devices that could enable to continue performance scaling without making the fabrication bill too big. In this context, new families of transistors called Functionality Enhanced Devices (FED) came in [16]. The principal interest of FED resides in the fact that while being larger than regular CMOS devices, they enable denser logic [1]. In this paper, we focus in Three Independent Gate (All-Around) Field Effect Transistor (TIGFET) which enable several functionalities such as polarity control, sub-threshold steep slope and threshold voltage control. We focus on the TIGFET polarity control capability in the rest of the paper. In [8], we explored the opportunities opened by polarity control to design 1Transistor-1RRAM bitcell enabling low voltage reset operation in the context of 2-terminal bipolar RRAM technologies (i.e., filamentary RRAM or STT-MRAM). In [8], both polarity gates were connected together, enabling the transistor to be configured in n-type during a set operation and p-type during a reset operation, thereby enabling a gate-overdrive-free reset operation. In this work, we propose to couple the polarity control with breakthrough array organization schemes.

III. Native Transpose Access RRAM Array

A. Transpose Access

The main motivation for using transpose access comes from the fact that while computing Edge-level applications, convolutions or matrix multiplications are counting for most of the computation [17]. In this context, the data coming from the data memory (in this case the non-volatile memory) may need to be pre-processed before computing (transpose operation for e.g.). Thereby, being able to perform both regular and transpose access at the sub-array level enables substantial performance gains [18], [19].

Figure 2 presents the concept of transpose read for both binarized and word level data. In a binarized approach, each physical bitcell correspond to the actual data stored in memory. In that context, transpose access is achieved by simply performing a vertical read (as shown Figure 2-a). On the other hand, performing transpose access in regular word level access is not straightforward as regular word access are usually performed along the WordLine. In this context, transpose access would lead to a MSB or LSB read operation on the data stored in memory. While such access could be used to accelerate some parts of a CNN execution (for e.g. rectification or pooling), it does not enable as-is transpose word access. Figure 2-b presents a turnaround as it is proposed in [18]. Bitwise structure is considered and the words are interleaved across several arrays. For each access (for e.g. 32bits word), 32 arrays are accessed and the number of accessed words depends on the array throughput. In [19], the authors explored the usage of the transpose access from an architectural point of view, but the authors assume crosspoint memories and only marginally discuss actual circuit or technology considerations. In [20], the authors propose a local BitLine-based architecture to enable transpose access. However, in this implementation, the periphery area is doubled. Also, the use of local BitLines strengthen data placement constraints from the programmer side as not all the data can be transposed.

B. Proposed Architecture

In this work, we propose to use the previously introduced polarity control TIGFET transistors to enable native transpose access. Thanks to their three gates, a polarity control can be enabled as introduced section II-B. In that sense, we propose here to use the mechanism demonstrated in [8] to perform bidirectional read operations. horizontal real operations are performed by setting up the TIGFET in n-type, and vertical accesses are performed by setting up the TIGFETs in p-type. In that sense, the read margin is kept high in both regular and transpose access, as the transistor Vgs is always precisely controlled. Figure 4 presents the architecture proposed in this work. regular and transposed access are always used to share the peripheral circuitry. In the presented example, during a regular access (in red), two arrays are accessed in transpose mode and two arrays in regular access mode. On the other hand, during a transpose access (in green) the opposite is done. This scheme enables to reuse all the peripheral circuitry for both regular and transpose access.

Figure 5 presents the schematic of the proposed array organization. Two WordLines (WL) are integrated: the Vertical WL (VWL) and the Horizontal WL (HWL). Finally, the Polarity Line (PL) is biased either to vdd or gnd to trigger the TIGFET polarity switch. Peripheral circuitry also relies on single TIGFET transmission gate to optimize the area efficiency. During a regular access, the current flowing though the bitcells is read on the BL periphery while it is read on the SL periphery during a transpose access. The Sense Amplifiers (SA) are connected to the neighboring SLs or BLs and depending on the requested operation, one or the other is selected.

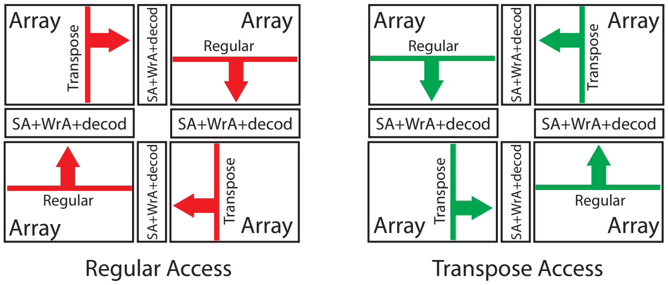

Fig. 4. Proposed Native transpose array organization featuring periphery sharing.

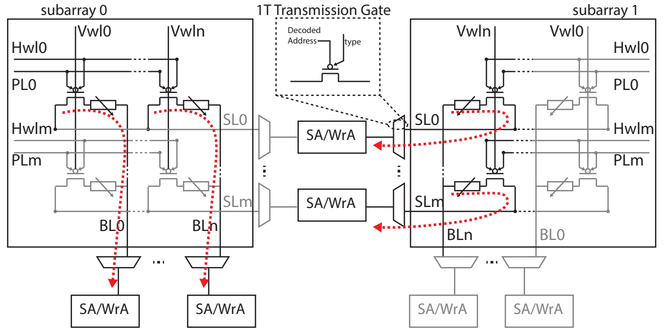

Fig. 5. Array organization with detailed array transpose interleaving and 1-Transistor Transmission gates.

C. Performance evaluation and validation

Figure 6 presents a transient simulation of operation of a 2x2 proposed native transpose memory. Two read operations are performed from each sides. First, a regular horizontal read operation is done. Then, a vertical read operation is performed. As introduced previously, in order to achieve an horizontal operation, the access TIGFETs transistors are setup in n-type configuration. While on the other hand, for the vertical transpose read operation, the access TIGFETs are setup in p-type configuration. Finally the two SA outputs shows a ”11” for the horizontal read (DataOutH) and ”01” for the vertical read (DataOutV), corresponding to the data actually stored in the memory.

IV. FEM for Edge AI

In this paper, we explored the vision proposed Figure 7. The proposed Edge device architecture optimized for data-centric workloads relies on two Functionality Enhanced Memory (FEM) concepts : (i) a native transpose access RRAM-based memory enabling in-situ data pre-shaping. (ii) a SIMD in-SRAM computing architecture, the BLADE which can efficiently perform multiplications on the pre-shaped data. By

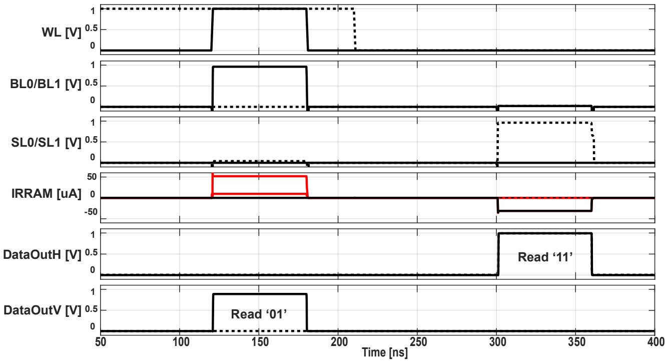

Fig. 6. Waveform showing transient simulation of a proposed 2x2 transpose subarray organization while performing a regular and a transpose access sequentially.

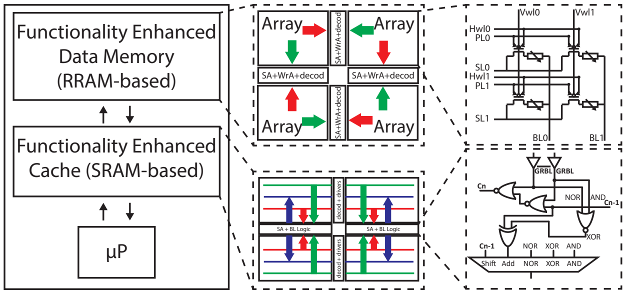

Fig. 7. Proposed vision featuring Functionality Enhanced Memories. Transpose access RRAM and in-SRAM computing architecture are leveraged to improve the energy efficiency of Edge Systems.

doing so, we do not target the most dense possible memory architecture, but we make it worth the price by improving the performances (i.e, in a previous work, we demonstrated 3 to 6x of performances improvements depending on the data-intensive workload thanks to the BLADE in-cache computing architecture in [6]; the authors of [19] demonstrated up to 14.5x performances gains; the authors from [18] showed a 4.7x more efficient use of their SRAM while running AI and filtering workloads thanks to the transpose access). Finally, the integration of non-volatile RRAM technologies in an Edge system is also expected to strongly improve the energy efficiency. As a perspective, in a parallel research path, to compensate for highly unbalanced read/write energy and time cost of RRAM technologies, new circuits [21] and control architectures [22] are under intense investigation and are expected to bring-in 2 to 10x of energy and performances improvements depending on the applications compared to simple SRAM replacement. The rationale behind this work, as primarily discussed in Figure 1, consists in bringing up the non-volatility in the memory hierarchy while bringing down more computing capabilities among the memory. However, for reliability and energy efficiency reasons, it is clear that these two limits may not want to be crossed, or at least not with the current state of RRAM technology developments and understanding.

V. Conclusion

In this work, we have proposed a new concept that we called Functionality Enhanced Memories (FEM) in order to improve the energy efficiency of embedded systems running Edge-level AI applications. In that sense, we proposed a new native transpose access memory using TIGFET transistors and validated its functionality through circuit simulations. Finally, we discussed the integration of FEM memories in order to answer the new questions opened by data intensive Edge AI workloads.

Acknowledgment

This work has been supported by the ERC Consolidator Grant COMPUSAPIEN (GA No. 725657), the NSF Career Award number 1751064 and ReRouting, LLC. P.-E. Gaillardon has financial interest in ReRouting.

References

[1] M. De Marchi et al., "Polarity control in double-gate, gate-all-around vertically stacked silicon nanowire fets," in *IEEE IEDM*, 2012.

[2] J. Romero-González et al., "Bcb evaluation of high-performance and low-leakage three-independent-gate field-effect transistors," in *IEEE JXCDC*, 2018.

[3] H.-S. P. Wong et al., "Metal–oxide rram," 2012.

[4] E. Vianello et al., "Resistive memories for ultra-low-power embedded computing design," in *IEEE IEDM*, 2014.

[5] A.-W. Simon et al., "A fast, reliable and wide-voltage-range in-memory computing architecture," in *IEEE/ACM DAC*, 2019.

[6] ——, "Blade: A bitline accelerator for devices on the edge," in *ACM GLSVLSI*, 2019.

[7] M. Rios et al., "An associativity-agnostic in-cache computing architecture optimized for multiplication," in *IEEE VLSI-SoC*, 2019.

[8] A. Levisse et al., "Resistive switching memory architecture based on polarity controllable selectors," in *IEEE TNANO*, 2018.

[9] D. Apalkov et al., "Magnetoresistive random access memory," in *Proc. of the IEEE*, 2016.

[10] H.-S. P. Wong et al., "Phase change memory," in *Proc. of the IEEE*, 2010.

[11] F. Disegni et al., "Embedded pcm macro for automotive-grade microcontroller in 28nm fd-soi," in *VLSI symp.*, 2019.

[12] "Reram embedded super low-power consumption mcu mn101l." [Online]. Available: https://industrial.panasonic.com/ww/products/semiconductors/microcomputers/mn101l

[13] A. Kawahara et al., "Filament scaling forming technique and level-verify-write scheme with endurance over 107 cycles in reram," in *IEEE ISSCC*, 2013.

[14] C. Nail et al., "Understanding rram endurance, retention and window margin trade-off using experimental results and simulations," in *IEEE IEDM*, 2016.

[15] C. Y. Chen et al., "Programming-conditions solutions towards suppression of retention tails of scaled oxide-based rram," in *IEEE IEDM*, 2015.

[16] P.-E. Gaillardon (editor), "Functionality-enhanced devices an alternative to moore's law," 2019.

[17] M. Chang et al., "Hardware accelerator for boosting convolution computation in image classification applications," in *IEEE GCCE*, 2017.

[18] K. Bong et al., "14.6 a 0.62 mw ultra-low-power convolutional-neural-network face-recognition processor and a cis integrated with always-on haar-like face detector," in *IEEE ISSCC*, 2017.

[19] S. Li et al., "Rc-nvm: Dual-addressing non-volatile memory architecture supporting both row and column memory accesses," 2019.

[20] L. Chang et al., "Multi-port 1r1w transpose magnetic random access memory by hierarchical bit-line switching," 2019.

[21] M. Alayan et al., "Switching event detection and self-termination programming circuit for energy efficient reram memory arrays," in *IEEE TCASII*, 2019.

[22] S. Tuli et al., "Rram-vac: A variability-aware controller for rram-based memory architectures," in *IEEE/ACM ASP-DAC*, 2020.