Review

A Survey of Emerging Memory in a Microcontroller Unit

Longning Qi, Jinqi Fan, Hao Cai * and Ze Fang

School of Integrated Circuits, Southeast University, Nanjing 210096, China; longn_qi@seu.edu.cn (L.Q.); 220236581@seu.edu.cn (J.F.); 220226267@seu.edu.cn (Z.F.)
* Correspondence: hao.cai@seu.edu.cn

Abstract: In the era of widespread edge computing, energy conservation modes like complete power shutdown are crucial for battery-powered devices, but they risk data loss in volatile memory. Energy autonomous systems, relying on ambient energy, face operational challenges due to power losses. Recent advancements in emerging nonvolatile memories (NVMs) like FRAM, RRAM, MRAM, and PCM offer mature solutions to sustain work progress with minimal energy overhead during outages. This paper thoroughly reviews utilizing emerging NVMs in microcontroller units (MCUs), comparing their key attributes to describe unique benefits and potential applications. Furthermore, we discuss the intricate details of NVM circuit design and NVM-driven compute-in-memory (CIM) architectures. In summary, integrating emerging NVMs into MCUs showcases promising prospects for next-generation applications such as Internet of Things and neural networks.

Keywords: embedded NVM; emerging memories; FRAM; RRAM; MRAM; PCM

1. Introduction

In recent years, there has been an extraordinary proliferation in the popularity and adoption of edge computing, reaching unprecedented levels [1]. To prolong the longevity of battery-powered devices, energy conservation modes, such as complete power shutdown (normally-off), can be employed to minimize energy consumption during periods of inactivity [2]. Nevertheless, a significant drawback of this technology is the potential loss of data stored in volatile memory, which can result in substantial performance and responsiveness penalties [3]. An alternative approach is the use of energy autonomous systems, which are battery-less and depend on the energy harvested from ambient sources [4]. Due to the unpredictable nature of ambient sources and the limitations of energy harvesters, these systems often face operational challenges caused by frequent power losses [5]. Energy-efficient technologies are necessary to ensure the preservation of work progress in the event of a power failure, regardless of whether the power outage is intentional or unintended.

The conventional strategy to realize intermittent computing involves utilizing a non-volatile memory (NVM) as a backup for on-chip volatile memory and processor states (flip-flops, latches, and registers) [6–9]. However, the process of sequential long-distance data movement between the volatile parts and the NVM requires many state transitions and significant overheads in terms of the execution time and energy [3,10]. To address this issue, emerging NVMs, such as ferroelectric random-access memory (FRAM), resistive random-access memory (RRAM), magnetoresistive random-access memory (MRAM), and phase-change memory (PCM), are incorporated into microcontroller units (MCUs).

Previous reports have proposed two types of FRAM, one based on ferroelectric capacitors (Fecaps) [11] and the other on ferroelectric transistors (FeFETs) [12]. A Fecap is a nonlinear capacitor with hysteretic behavior, forming the basis of nonvolatility. FeFETs exhibit nonvolatility due to the property of polarization retention in the absence of an externally applied electric field in ferroelectrics [13]. FeFETs may have a lower critical voltage for polarization switching compared to standalone Fecaps, provided they can be operated in the negative polarization-voltage region [14]. An RRAM device is based on a

metal–insulator–metal structure and utilizes voltage pulses to create multi-level conductance levels, including high-resistance state (HRS) and low-resistance state (LRS) [15]. The magnetic tunnel junction (MTJ) device has a free magnetic layer and a pinned magnetic layer, which are separated by a thin insulator layer. It has a variable resistance depending on the direction of the write current [16]. PCM devices can modulate conductance based on the material phase, which can be switched by applying heat or voltage pulses [17]. These memory devices offer faster write times and lower write voltages than Flash memory [18], which makes it possible to provide instant on/off capabilities for MCUs with near-zero power consumption during the inactive phases [2].

Based on the emerging NVMs, nonvolatile static random-access memory (nvSRAM) is suggested as a replacement for the two-macro method (SRAM and Flash) because of its parallel operation and high speed [19]. The store operation of an nvSRAM bitcell involves, in nature, programming the nonvolatile elements according to the data held in the SRAM part. The active power consumption of nvSRAM has received attention and various strategies have been reported. For example, plate-line charge-share and bit-line non-precharge techniques were proposed to mitigate the active power of a 6T-4C nvSRAM [20]. Additionally, ensuring the reliability of real-time storage and restoration data in the nvSRAM is crucial for maintaining a successive workflow and service quality in MCUs. A bitcell-circuit-system co-design was employed to improve the reliability and scalability of a 64 KB RRAM-based nvSRAM, which was integrated into a 32-bit MCU and achieved a sub-0.1% raw bit error rate between the power outages [19]. Prior works have also investigated the possibility of utilizing the emerging NVM as an embedded Flash (eFlash) to enable rapid macro-to-macro backup and recall operations in the event of power loss [12,21,22]. In terms of reliable operations, a software-hardware programming configurable framework was adopted to achieve >95% read accuracy and a 1.3% mean error for all targets in a 3-bit mode of a 1 MB RRAM macro [21]. Despite numerous relevant studies, there is still a lack of universally applicable and significantly effective technologies for utilizing emerging NVMs as a replacement for SRAM and Flash in terms of power consumption, speed, and reliability.

Recently, state-of-the-art MCUs have supported machine learning (ML), allowing real-time data collection, ML model execution, and analysis on low-power devices. This advancement promotes the growth of computational intelligence at the edge by providing benefits such as improved security, privacy, reduced latency, and extended battery life [23–26]. Nevertheless, the pervasive deployment in edge applications is impeded by the substantial computational demands of ML inference, particularly in high-dimensionality matrix-vector multiplications (MVMs) [27]. To tackle this challenge, the concept of compute-in-memory (CIM) was introduced, which integrates high-efficiency computational logic within the memory array to significantly reduce memory and computation energy, thereby enabling the ML implementation on low-power MCUs [28–30]. SRAM-based CIM has gained increasing attention as a promising solution for ML applications [31–39]. Nonetheless, it exhibits inherent drawbacks including a high transistor count, limited data storage capacity, and unsuitability for long-term data retention, which makes it disadvantaged in terms of area, weight density, and event-driven applications [18,28]. In contrast, NVM-based CIM offers low standby power, high density with multi-level cells, and low system power by eliminating initial data writing [40,41]. It also efficiently stores weight-data for neural network models, with its capacity exceeding the mega-bit level. Despite these advantages, NVM-based CIM faces challenges due to memristor nonlinearity, and the need for large write currents and high-precision sense amplifiers, resulting in increased area and power consumption [18,30]. Therefore, this paper reviews existing design techniques for NVM-based CIM, aiming to inspire innovative circuit design strategies to address these challenges.

This article is a comprehensive review that will discuss the role of emerging nonvolatile storage in MCUs. The remainder of the paper is organized as follows. Section 2 discusses the feasibility of replacing SRAM and Flash in MCUs with emerging NVMs in terms of

key parameters and the control scheme. Section 3 explores the design of NVM in an MCU, including the bitcell, read/write circuits, macro structure, and peripheral circuits. Section 4 reviews the design of CIM based on RRAM and MRAM. Conclusions are outlined in Section 5.

2. Feasibility of Replacing Flash and SRAM in MCUs with Emerging NVMs

2.1. Characteristics of Various Storage Types

Figure 1 depicts a prototypical MCU architecture comprising a processing core, Flash memory for code and data storage, SRAM for high-speed data access, and abundant peripheral devices such as PLL, DMA, UART, SPI, TIMER, WDT, GPIO, RTC, and PMU. The memory module primarily determines the power consumption of an MCU due to its intrinsic physical characteristics [42]. Therefore, selecting the appropriate memory for a low-power MCU is crucial.
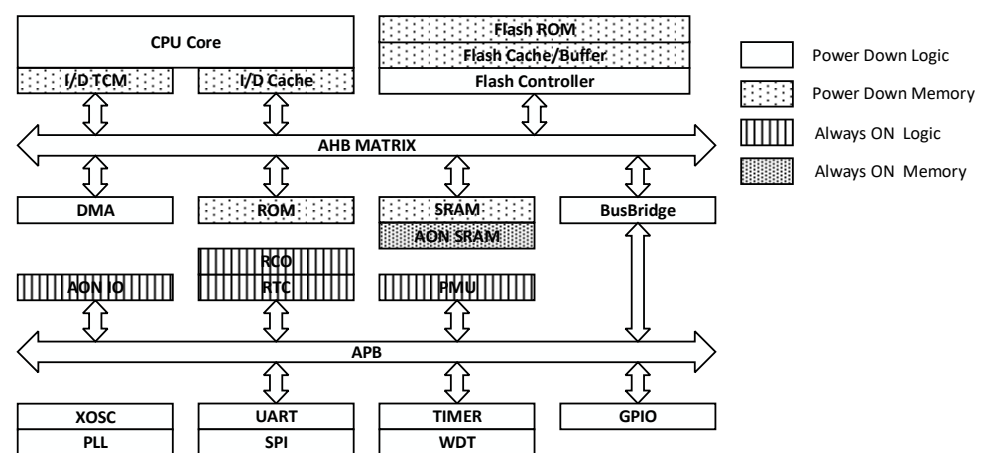

Figure 1. Block diagram of a prototypical MCU.

MCUs integrated with Flash memory are commonly used in the commercial market. Due to its nonvolatile nature, Flash memory significantly reduces standby power consumption compared to volatile memories. However, the active operation of Flash memory requires high power consumption attributed to the need for high-voltage programming and erasing using charge pumps [42]. Moreover, integrating Flash memory into the advanced nodes has become increasingly complex and expensive due to its limited area shrink capability and growing complexity, as highlighted in several publications [43–45]. In response to these challenges, innovative eFlash designs were investigated, as detailed in Table 1 [46–49]. The SG-MONOS cell, combining split-gate and charge-trapping structures, enhances Flash memory performance and reliability by enabling efficient programming through source-side injection (SSI) and preventing column current leakage via series connection [46,50]. The eSTM is a floating gate-based cell, gathering the advantages of a conventional split-gate NVM cell together with a more compact cell area than a typical 1 T cell [49]. By using SG-MONOS and eSTM cell, eFlash macros were successfully fabricated at 40 nm and 28 nm with impressive specifications catering to high-end automotive applications. Nevertheless, these eFlash macros are constrained by low write endurance, making them difficult to use in applications that frequently power down. Furthermore, they tend to encounter elevated manufacturing costs attributable to the intricacy of the fabrication process, and persist in facing reliability challenges as the technology node continues to shrink.

Emerging NVM concepts (such as FRAM, MRAM, PCM, and RRAM) have been extensively researched to address these challenges. These alternatives offer easier integration into CMOS and lower process complexity [45]. Table 1 demonstrates the key features of type-like Flash NVMs based on emerging memory devices [22,51–58]. Type-like Flash NVMs refer to the emerging NVMs that directly store data in nonvolatile devices without the need for backup and restore operations and perform NOR Flash memory operations

by imposing voltages corresponding to block-erase, random program, and random read. Previous reports indicate that FRAM consumes less power than Flash and DRAM, offering fast and high-bandwidth read/write operations [11,15,20]. Nevertheless, from the low clock frequency of FRAM-based MCUs in Table 1, it can be realized that FRAM still suffers from high power dissipation and limited clock frequency constraints when compared with other emerging NVMs. To remedy these constraints, a nonvolatile system-on-chip (NVSoC) integrated an instruction cache and increased the frequency to 30 MHz [10]. As indicated in Table 1, recent years have witnessed notable advancements in MCUs based on RRAM and MRAM, capable of reaching capacities in the megabyte range and working at lower operational voltages. A state-of-the-art MCU utilized four key design techniques to implement a 10.8 MB embedded STT-MRAM macro, which achieved the fastest random read access frequency and write throughput among reported Flash-replacement MRAMs [59]. However, the limited endurance of RRAM and reliability issues of MTJ impede their broader utilization [15,19,54]. PCM featured an attractive cell size of 0.019 F^2 and attained the largest capacity of 21 MB among the embedded NVMs presented in Table 1. Nonetheless, PCM faces challenges due to the crystallization temperature limitations of the standard GST225 material, restricting its applications in consumer temperature ranges [45]. Recent developments have addressed this issue by using an optimized Ge-rich alloy with a higher crystallization temperature and a differential sensing scheme, enabling memory operations and data retention above 150 °C in a 32 KB embedded PCM [22]. In conclusion, each type of emerging NVM has its unique advantages and disadvantages, necessitating the selection of an appropriate storage type based on specific application requirements.

Table 1. Comparison of eFlash and type-like Flash NVMs in MCUs.

Performance Metrics	JSSC'14 [46]	IEDM'16 [47]	ESSCIRC'13 [48]	IMW'19 [49]	ISSCC'13 [51]	ISSCC'16 [52]	JSSC'22 [53]	JETCAS'16 [54]	ISSCC'21 [55]	VLSI'22 [56]	VLSI'22 [57]	JSSC'18 [22]	VLSI'23 [58]
Architecture	/	/	/	/	Cortex-M0	8051 8 bit	RISC-V	/	RISC-V	Cortex-M0	Cortex-M33	/	/
Technology	40 nm CMOS	28 nm CMOS	65 nm CMOS	40 nm CMOS	130 nm CMOS, FRAM	65 nm CMOS, RRAM	40 nm CMOS, RRAM	180 nm CMOS, RRAM	22 nm FDSOI, MRAM	22 nm FDX, MRAM	22 nm CMOS, MRAM	110 nm BCD, PCM	28 nm FDSOI, PCM
Capacity	2 MB (code) 64 KB (data)	4 MB (code) 64 KB (data)	4 MB (code)	1 MB	64 KB	8 KB	2 MB	256 KB	4 MB	2 MB	2 MB	32 KB	21 MB
Cell structure	SG-MONOS	SG-MONOS	HS3P	eSTM	/	/	1T-1R	1T-1R	/	/	/	1T-1PCM	/
Cell size [F^2]	/	0.053	/	0.049	/	/	0.64 *	20	/	/	/	0.7	0.019
Supply [V]	1.25	1.1	1.3	0.85–1.35	1.5	0.8	1.1	1.6/1.8	0.5–0.8	0.44–1.0	0.5–1.0	1.55–1.95	0.8
Active power [µW/MHz]	/	/	/	<150	112	33	135 mW	/	49.4 mW	387	158 mW	/	/
Standby power [µW]	/	/	/	<10	/	/	/	/	1.7	70	468	/	80 *
Max freq. [MHz]	160	200	81.5	/	8	>100	/	25	450	70	190	10	400
Endurance [cycles]	10 K (code) 10 M (data)	10 K (code) 1 M (data)	/	10 K	/	/	$>3 \times 10^{5*}$ [5*]	2×10^{8} [8]	/	/	/	10^{5} [5]	/
ECC	Yes	/	/	Yes	/	/	Yes	Yes	Yes	/	/	/	/

* It is estimated from figure in source.

Conventional SRAM-based programmable logic has a large area and always consumes static power to maintain the stored data [60,61]. To reduce power consumption, the MCU enters the standby mode and employs multiple strategies to curtail the standby power, which is typically attributed to leakage current dissipation in the always-on domain. For instance, the cutting-edge ultra-low leakage MCU, leveraging 55 nm TFET-CMOS hybrid technology, achieved significant standby power reduction through innovative TFET-Gated-Ground SRAM and voltage-stacking techniques [62]. However, switching between active and standby modes consumes additional power for data transfer between volatile and nonvolatile memories [19,42,63]. Therefore, nvSRAM was proposed as a replacement for the traditional two-macro approach due to its high-speed parallel operation. The NVM elements and SRAM part are integrated in an nvSRAM bitcell by a bit-to-bit connection instead of a macro-to-macro connection [19,64]. By storing commonly used routines in the

nvSRAM, the startup latency and associated energy consumption from data movement can be eliminated, enabling more efficient edge computing applications [65].

A comparison is performed in terms of several key parameters, shown in Table 2, to highlight the performance characteristics of MCUs with volatile SRAM and nvSRAM based on emerging memories [19,20,64,66–70]. Conventional SRAM consumes a much higher retention current than the deep-sleep-mode current required by sensor networks and energy-management systems, particularly as process geometry scales down. This has prompted investigations into leveraging low-leakage transistors and advanced design methodologies to achieve lower standby power levels. By utilizing thick-gate-oxide transistors with source bias control techniques, the retention power of the system was effectively reduced to the order of nanowatts [66,68]. However, these transistors generally cause an increase in memory macro area and active power dissipation, requiring additional techniques for minimizing cell size and active energy. An ultra-low-voltage MCU featuring single-rail SRAM was developed using 22 nm FDX technology and an adaptive reverse body bias scheme, achieving a leakage power of 6.6 μW and active power of 6.3 μW/MHz [67]. Nevertheless, the single rail macro incurs a 20% area overhead compared to the dual rail macro, and a custom bitcell design is needed to ensure stable read operations down to 0.5 V, thereby amplifying the complexity of the design.

Table 2. Comparison of SRAM and nvSRAM in MCUs.

Performance Metrics	ISSCC'14 [66]	COOL CHIPS'20 [67]	A-SSCC'18 [68]	A-SSCC'19 [69]	BioCAS'14 [20]	TCAS-I'23 [19]	VLSI'18 [64]	TCAS-II'21 [70]
Architecture	/	Cortex-M4	/	/	Cortex-M0	/	MSP430	/
Technology	65 nm CMOS	22 nm FDX	40 nm CMOS	110 nm CMOS	130 nm CMOS, FRAM	130 nm CMOS, RRAM	45 nm CMOS, MRAM	90 nm CMOS, RRAM
Non-volatility	N	N	N	N	Y	Y	Y	Y
Capacity	128 KB	256 KB	64 KB	2.5 MB	16 KB	64 KB	32 KB	/
Cell structure	6T	6T	6T	6T	6T-4C	12T-2R	8T-2MTJ	4T-2R
Cell size [F^2]	2.159	0.26 *	2.888	1.84 *	/	16 *	/	2.83
Supply [V]	1.2	0.55	3.3	1.5	1.2	1.1–2.2	1.1/1.6	1.5
Active power [μW/MHz]	25	6.3 (MEP)	174 (Read) 180 (Write)	90 (Read) 105 (Write)	/	/	/	/
Standby power [μW]	/	6.6	0.33	0.73	/	/	2	/
Max freq. [MHz]	/	40	42	147	24	50	25	10 *

* It is estimated from figure in source.

In contrast, nvSRAM can be completely powered off during idle periods, thereby eliminating retention power consumption. The 4T-2MTJ macro, as illustrated in Table 2, exhibits the potential to achieve a smaller footprint compared to 6T SRAM at the 45 nm technology node, while maintaining an almost unchanged operation current over generations compared to the exponentially increasing current of 6T SRAM due to MOSFET off-current degradation with scaling [71]. In addition, innovative plate-line charge-share and bit-line non-precharge techniques effectively mitigate the active power dissipation from the large Fecap, making the 6T-4C macro suitable for an electrocardiograph monitoring SoC [20]. The 12T-2R nvSRAM showcases a raw restore-BER of less than 0.1% between power outages, contributing to the realization of an error-free nvSRAM macro with correction techniques [19]. One drawback of this nvSRAM is the 123% increase in area overhead compared to conventional 6T SRAM implemented on the same technology node.

2.2. Peripheral Circuits of Flash and SRAM

The Flash memory and SRAM exhibit distinct control principles based on their specific functionalities and operational requirements. A typical eFlash system is presented in Figure 2a and can be structured into three levels, including memory cells, peripheral circuits of the hard macro (such as sense amplifiers and high-voltage generators), and various functional blocks on a system level. The design of memory cells and peripheral circuits plays a vital role in determining the electrical characteristics and data reliability, and achieving the target specifications of the Flash memory. To realize higher performance and reliability in Flash memory, more CG or SL stitch regions may be required for faster rise/fall time or noise suppression. Furthermore, finer array division and control are needed to suppress the influence of program disturb on unselected cells, which is caused by sharing nodes among cells during program and erase operations.
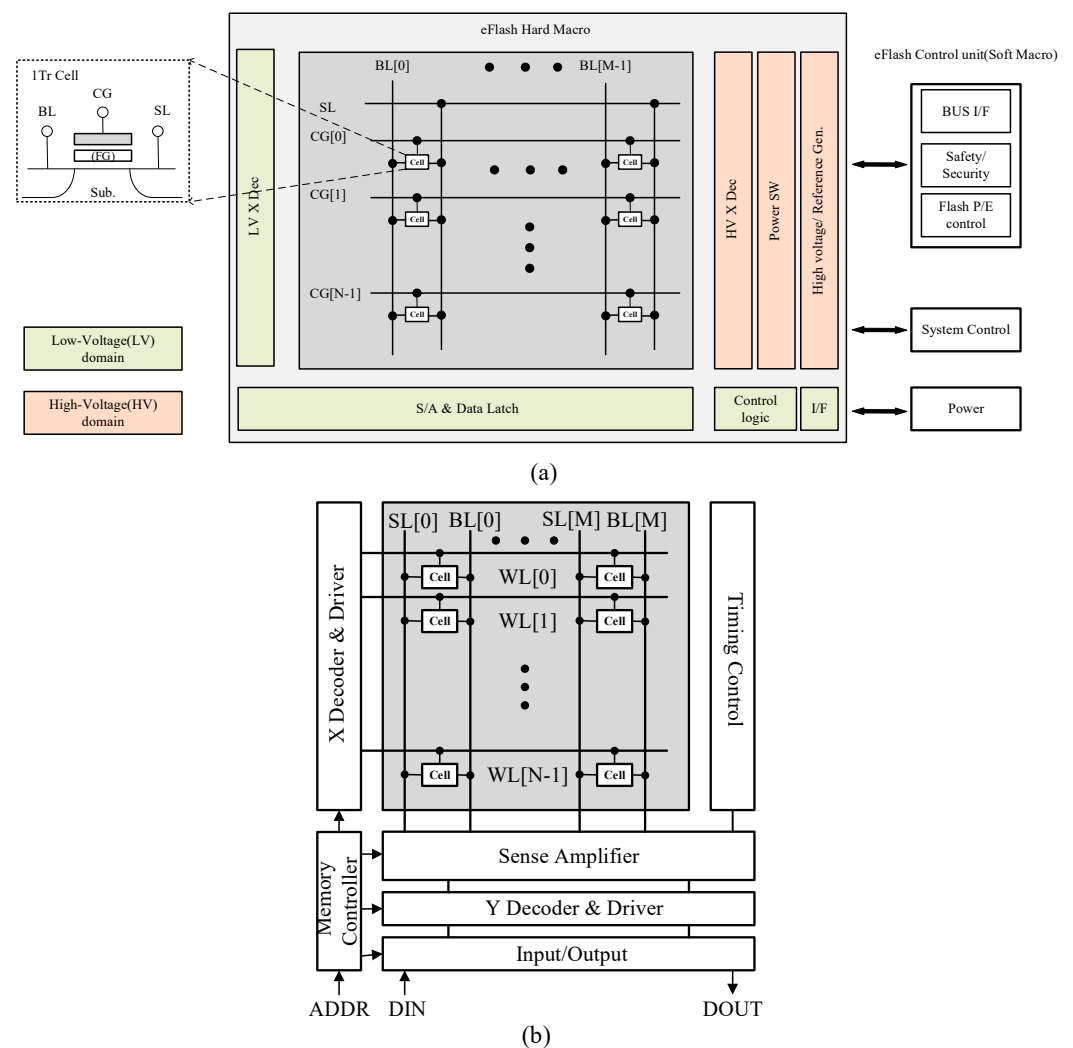

(a)

(b)

Figure 2. Memory floorplan of a typical (**a**) Flash and (**b**) SRAM embedded in MCUs.

On the other hand, the peripheral circuits of a typical SRAM consist of the timing control circuit, the address decoder circuits, the row (X) and column (Y) driver circuits, the sense amplifier circuit, the data input and output circuits, and the memory controller circuit. The timing controller synchronizes signal timing for proper operations of all SRAM components. The X/Y decoders select specific memory cells, drivers provide necessary voltage levels, and sense amplifiers ensure data integrity during read operations. The memory controller is a critical component, which manages data flow, controls read/write operations, and coordinates transfers between the CPU and SRAM. In SRAM design,

ensuring stability across varying temperature and process conditions presents a significant challenge. A proposed SRAM design tackles this challenge by utilizing charge sharing to transfer stored charge from local bit lines (BLs) to global BLs, thereby ensuring a constant charging current for the BLs [66].

3. Design Considerations for NVM in MCU: A Focus on Three Metrics

3.1. Bitcell Design: Cell Size Focus

Figure 3a,c,e illustrate the classical configurations of nvSRAM bitcells, which integrate a traditional 6T-SRAM cell with nonvolatile elements and additional access transistors. When continuously powered, the nvSRAM bitcells function equivalently to standard 6T SRAM cells, storing data in cross-coupled inverters with comparable read and write speeds to traditional SRAM. This operational state is commonly called "SRAM mode" or "normal mode". While encountering a power outage, these cells capture the SRAM contents in nonvolatile devices just before power loss and then restore the saved state when receiving power again.
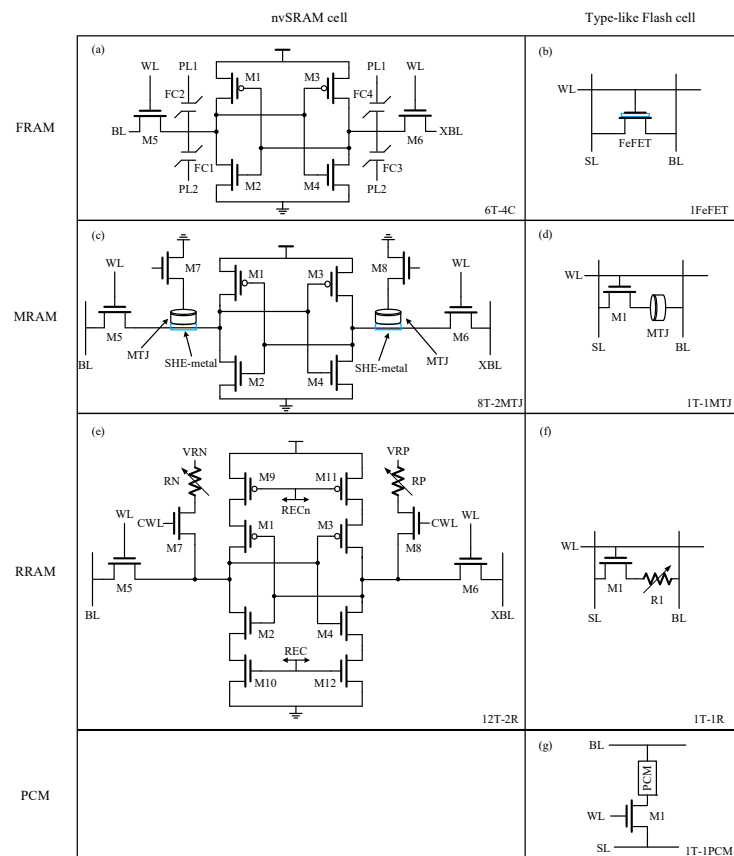

Figure 3. Bitcell structures of (**a**) a 6T-4C nvSRAM [20]; (**b**) a 1FeFET FRAM [72–75]; (**c**) a 8T-2MTJ nvSRAM [64]; (**d**) a 1T-1MTJ MRAM [76,77]; (**e**) a 12T-2R nvSRAM [19]; (**f**) a 1T-1R RRAM [21,54,78]; and (**g**) a 1T-1PCM PCM [22,79].

The 6T-4C FRAM bitcell shown in Figure 3a is based on the Fecap, which is a non-linear capacitor with hysteretic behavior. It has a much higher signal margin than the regular FRAM bitcells which utilize a single ended Fecap and two differential Fecaps, respectively [11]. This is because data are stored in all four capacitors in a complementary fashion, which also causes a higher area and power cost. Figure 3c shows the structure of the nvSRAM bitcell based on the SHE-MTJ, which offers significant advantages over traditional two-terminal MTJ [64]. Conventional MTJs are challenged by high switching currents and the need to balance resistance levels for read/write operations, compromising either writability or read sensitivity [80]. In contrast, the SHE-MTJ capitalizes on the SHE

effect for enhanced spin generation efficiency and offers a low-resistance write path through charge current in the SHE-metal. Moreover, the decoupled read and write terminals facilitate the independent fine-tuning of the MTJ and SHE-metal dimensions, optimizing both readability and writability [64]. In the nvSRAM cell, M7 and M8 serve as extra access transistors that are disabled to separate the SHE-MTJs from the standard 6T SRAM cell for regular SRAM operation, and are activated to realize store and restore operations in the event of a power loss. This approach helps to prevent unnecessary write operations on SHE-MTJs when the data stored in the cross-coupled inverters change during continuous power on, thereby improving energy efficiency. The bitcell depicted in Figure 3e adds four clock-controlled power-gating transistors (M9~M12) to save energy, and it adopts the two-ends nvSRAM scheme [81–83] with two RRAM devices in a bitcell to improve reliability in low-HRS/LRS scenarios through differential sensing [19]. Despite its larger area and energy consumption compared to single-end nvSRAM cells [84–86], this configuration offers a superior sense margin and restoration yield for emerging NVM technology. It is possible to simplify the 12T-2R bitcell by replacing the M9, M11 pair and M10, M12 pair with a single transistor, respectively, but it is less efficient than realizing a compact transistor pair through area-sharing. The 12T-2R bitcell has an area overhead of approximately 123% when compared to a conventional sideway 6T SRAM bitcell at 130 nm [19]. Despite the ability of nvSRAM to combine the fast read/write characteristics of SRAM with the non-volatility of NVM, the cell structure based on 6T SRAM limits the reduction in cell size, resulting in poor scalability. When the extra area overhead outweighs the energy benefits it brings, the significance of this design becomes less evident. Therefore, nvSRAM struggles to achieve high-capacity storage and cannot effectively replace traditional embedded NVM in MCUs.

Figure 3b,d,f,g illustrate the standard type-like Flash bitcells, in which data are directly written into nonvolatile elements to eliminate the requirement for data backup and restore operations across power losses. Figure 3b displays a 1FeFET bitcell that leverages FeFET polarization for storing data and performs NOR flash memory operations in a NOR-type array. Two architectures for 1FeFET-based NOR Flash were introduced in [73]. One architecture offers a high level of scalability, achieving 6 F^2 at a minimum by sharing the source lines in pairs of rows, while the other features more isolated cells, resulting in reduced disturbance but at the expense of scalability. A 1T-1FeFET bitcell with separate read and write paths was reported in [87], achieving non-destructive read and lower write power at iso-write speed compared to 1FeFET FeRAM. However, a slight area penalty is introduced due to the additional MOSFET. Furthermore, a 2T-1FeFET bitcell with separate read/write paths was designed to enhance design flexibility for CIM. Although the 1FeFET/1T-1FeFET bitcells are more compact, they require additional bias circuitry and/or charging of all non-selected WLs and BLs [75,88]. These introduce energy penalties and design complexities owing to the need for multiple voltage levels, rendering them less ideal for intermittently powered systems [12].

Figure 3d,f,g demonstrate a similar bitcell structure consisting of a MOSFET as an access transistor and a nonvolatile device as a storage element. Emerging NVMs based on this bitcell structure commonly serve as a replacement for embedded NOR Flash and show write speed and energy advantages over NOR Flash. The 1T-1MTJ bitcell, shown in Figure 3d, was reported and a local source line (SL) array scheme was implemented to improve write performance [77]. This scheme utilizes a local SL to distribute return current among unselected BLs for preventing select transistor TDDB stress, and it ensures no disturbance occurs in unselected BLs during write operations by connecting a group of MTJs to the local SL. It also enables the concurrent writing of 0 or 1 states without needing to elevate BLs, enhancing write efficiency in MRAM. Figure 3f illustrates a 1T-1R bitcell occupying an area of 20 F^2 [54], which is more compact than earlier FRAM [89,90], STT-MRAM [16,76], and CBRAM [91,92] designs. The RRAM was equipped with a novel sense amplifier and a write-and-verify (WAV) voltage generator to increase the read and write yield. The 1T-1PCM cell, shown in Figure 3g, is fabricated by 0.11 μm BCD technology and covers an area of 0.7 F^2. It utilizes a Ge-rich alloy for a higher crystallization tempera-

ture [93] compared to the conventional Ge$_2$Sb$_2$Te$_5$ alloy, and it adopts differential sensing to mitigate resistance drift [94–96]. These features enable reliable memory operations and data retention at temperatures above 150 °C [22].

3.2. Read/Write Circuit Design: Power Efficiency Focus

As for read schemes for NVM, there are two typical sense amplifiers (SAs): the voltage-mode SA (VSA) and the current-mode SA (CSA). The VSA is used for precharging selected BLs to a target voltage, allowing the reading of both LRS and HRS cells. However, the limited voltage difference between HRS and LRS cells makes it susceptible to BL noise and coupling [97]. The CSA imposes a fixed bias voltage on the BL to induce current in the cell for reading. A current comparator is used to compare the sensed current with a reference current. Compared with VSA, the CSA minimizes the vulnerability to BL noise and coupling. Moreover, it exhibits faster read speeds than VSA when the BL length of a 0.18 μm RRAM macro exceeds 128 rows [97].

The read scheme proposed in [54] uses two types of CSAs to limit the read voltage to 0.3 V and avoid read disturbance, as shown in Figure 4a. By detecting the current of a RRAM cell at 0.3 V, unwanted state transitions, especially in the HRS, can be prevented. Current mirrors are used to provide magnifying power modes and compare the transformed voltage with the reference voltage (VREF) to determine the logic output. A modified VSA was presented for single-ended FeFET NVM, with duplicated read-BL voltages across two cross-coupled, inverter-based SAs connected to VREF-NAND and VREF-NOR during sampling [12]. During read operations, the SA with VREF = 0.95 V is enabled to sense the stored bit, with the inverted NOR output serving as the final READ (or OR) output. A fully configurable offset-tolerant CSA was reported previously [21], which can be tuned by software for precharging, calibrating, and latching. The design includes a dual-mode reference generator that can switch between a compact current-mirror-based mode and an accurate resistance network-based mode, providing a wider range of reference currents for multi-bit programming and read requirements.
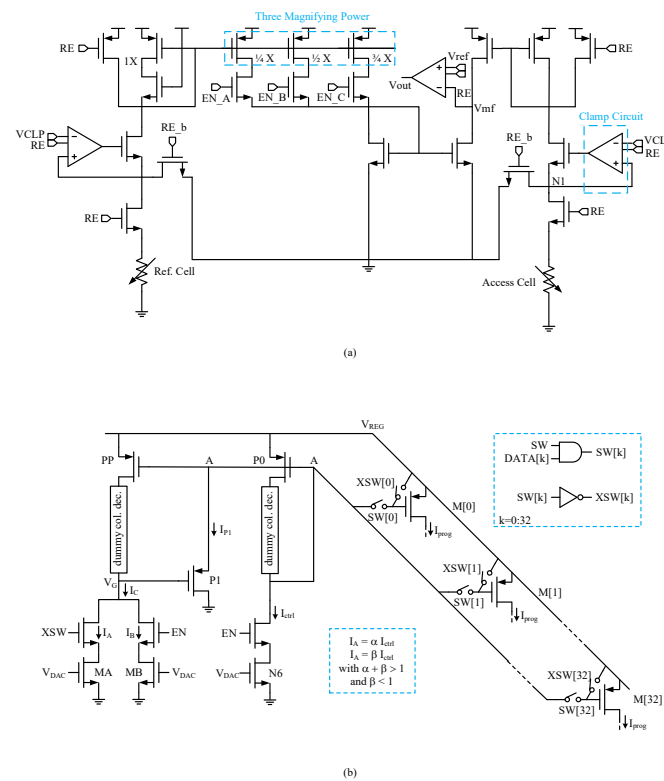

Figure 4. Circuit schematic of (**a**) a sense amplifier [54]; (**b**) a write driver [22].

The design of write peripheral circuits plays an important role in determining the write error rate and power consumption of NVM, particularly in the case of STT-MRAM. The challenges in write operations arise from the need for both a sufficiently long write time and a high write voltage to avoid errors [98]. However, the exact write time for each STT-MRAM cell differs due to process variations and thermal fluctuations. To ensure reliable writing, the write time is typically kept much longer than the average write time, causing energy wastage as the write current continues to flow even after the MTJ has switched [99].

The enhanced current programming circuitry depicted in Figure 4b improves RESET pulse shaping in high-parallelism programming scenarios by introducing a fast recovery method [22]. By incorporating an additional branch with transistor PP, the circuit ensures the rapid discharge of node A during programming pulses, allowing precise control of the current flowing through transistor P1. This approach eliminates the need for the precise matching of NMOS transistors and enables higher parallelism without increasing static power consumption or area occupation. Additionally, adjusting the values of factors α and β determines the circuit speed and the timing of turning off transistor P1, offering flexibility and efficiency in current programming operations.

Furthermore, the write termination scheme has been extensively discussed in previous research as an effective method for addressing the issue of energy waste during write operations. In this context, the energy waste arises from unnecessary write operations, which occur when the incoming data equal the current value stored in the memory cell. The self-write-termination (SWT) circuit monitors the write operation and prevents redundant writes, thereby improving energy efficiency and reliability in fast-switching cells. As reported in [99], the implementation of an SWT circuit leads to a remarkable 83% reduction in total write energy compared to conventional write circuits. Additionally, this approach utilizes 75% fewer transistors than previously proposed SWT circuits. In another study [52], the integration of SWT into each column of a 1-macro nvSRAM array demonstrates improved clock frequency and substantial reductions in store energy, up to 172$\times$. Although considerable progress has been made, challenges remain in mitigating the extra power consumption and large area overhead associated with the SWT circuits.

3.3. Macro Structure and Peripheral Circuit Design: Area Efficiency Focus

The memory controller serves as a crucial component in managing data transfer between the CPU and memory modules in an MCU. It controls operations such as the reading, writing, and refreshing of memory modules to ensure that data are correctly stored and retrieved. In response to increasing demands for higher speed, lower power consumption, and an enhanced reliability of memory in various applications, previous studies have reported diverse memory controllers which integrate additional circuit modules to support functionalities beyond basic operations, as shown in Figure 5.
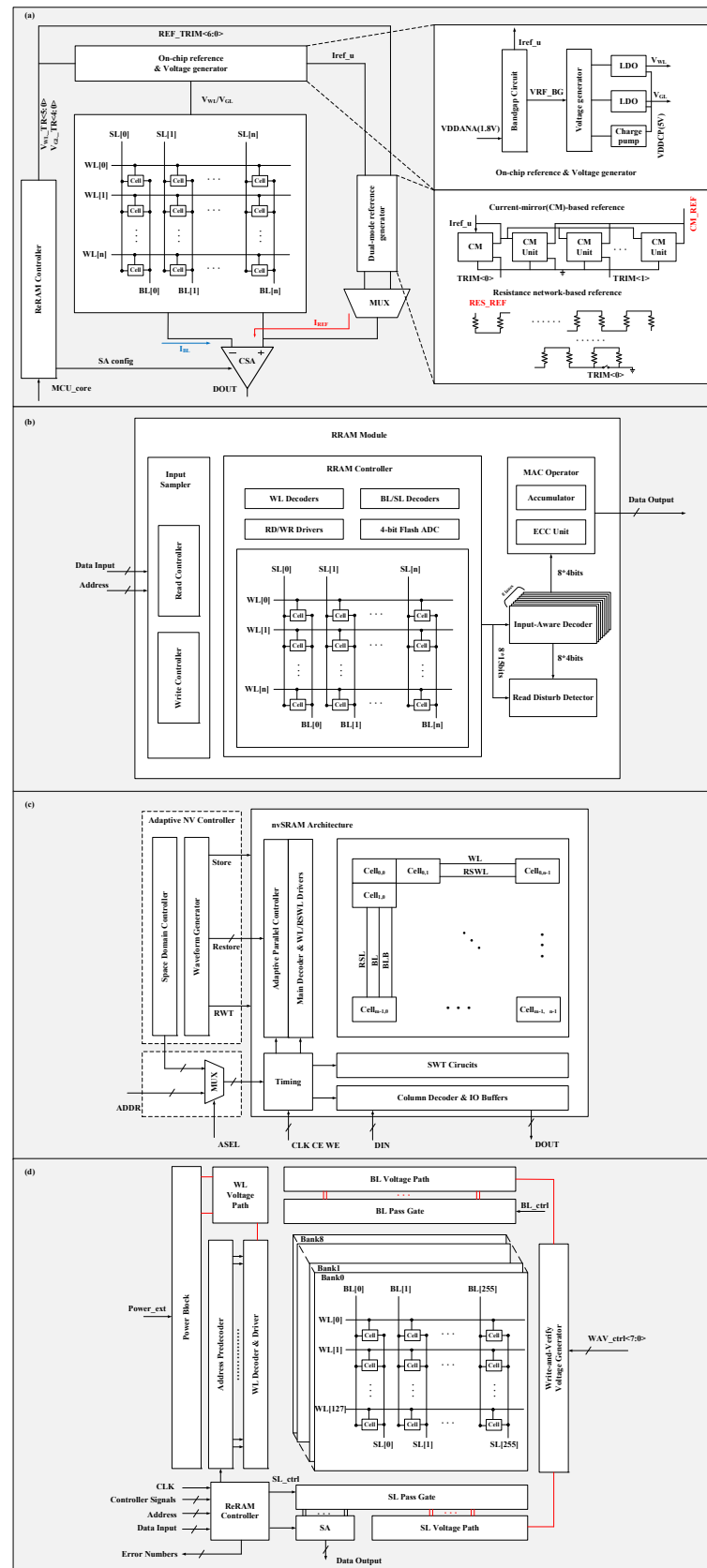
Figure 5. Structure diagrams of (**a**) a 1 MB RRAM macro [21]; (**b**) a 2.25 MB RRAM-based CIM macro [100]; (**c**) an adaptive RRAM-based nvSRAM macro [101]; and (**d**) a 256 KB RRAM macro [54].

In recent years, MCUs with low-power and instant-on features have been highly valued for energy harvesting as well as "normally off" applications. Unfortunately, the conventional data backup strategy tends to store all contents from volatile parts into the NVM, even though most of the data are rarely changed or utilized in practical scenarios. To reduce unnecessary backup operations, a space domain controller was designed, as shown in Figure 5c, to provide the proper store address range of the nvSRAM [101]. The SWT circuit in each column detects bit changes and controls the write driver on the RSL line to terminate the SET or RESET operations as required. This approach eliminates the need to redundantly store or restore unused data, leading to reduced time and energy consumption during storage and restoration processes. A similar strategy was also adopted in [21], as depicted in Figure 5a. The memory controller supports the pre-read function that allows the controller to assess the resistance range of the RRAM device before writing to reduce redundant write operations. For restore operations, an adaptive parallel controller (Figure 5c) was employed to manage different restore parallelism options (1 WL/4 WL/16 WL) based on the maximum tolerant peak current of the power source and to restore the speed requirement, which contributes to the realization of the instant-on operation of the MCU. It is also common to utilize a memory controller for the power management of memory macros. As shown in Figure 5b, the memory controller deactivates the power gates of unselected RRAM modules, and it fully powers down the RRAM after a sufficient period following the last event, leading to an 89.21% reduction in power usage [100].

In memory design, reliability is crucial for ensuring data integrity and system stability, particularly in aerospace systems, medical devices, and autonomous vehicles. As illustrated in Figure 5d, a custom RRAM controller enhances yield to approximately 100% by integrating features such as built-in self-test, built-in self-repair, a shortened Bose–Chaudhuri–Hocquenghem (BCH) error-correlating code (ECC), and asymmetric coding [54]. Significantly, it enables adaptive ECC algorithm selection (Hamming or BCH code), resulting in improved performance with reduced power consumption, minimized parity bit overhead, and increased operation speed. The memory controller shown in Figure 5a is also equipped with an ECC module (ECC encoder and ECC decoder) to rectify erroneous bits utilizing the BCH algorithm [21].

The voltage generator is also a critical component in the memory macro, providing essential and precise voltage references for various operations. For example, two types of voltage generators, illustrated in Figure 5a, serve different purposes. The on-chip reference and voltage generator supplies constant and temperature-compensated references through the bandgap circuit, adjusting to accommodate diverse programming and readout needs [21]. The integrated charge pump produces a high voltage for the forming/set operation of the ReRAM. Additionally, a dual-mode reference voltage generator was designed to meet the diverse requirements of multi-bit storage memory. The current-mirror mode offers a wide range of reference currents for multi-bit programming needs, while the resistance network mode supports high/low-resistance states with high linearity, ensuring accurate voltage references for programming and readout. The integration of these circuits eliminates the need for off-chip references or high voltages, greatly simplifying the system design and connectivity. Furthermore, the design of the voltage generator can also enhance the reliability of memory. As illustrated in Figure 5d, the WAV voltage generator produces eight stepwise voltage levels in the BL or SL path, aiming to enhance write yield by mitigating variations in the transition energy of individual cells [54].

4. Circuit Design for CIM Based on RRAM and MRAM

Among emerging NVMs, RRAM and MRAM are the primary choices for embedded nonvolatile CIM due to their advantageous characteristics. RRAM exhibits a relatively larger on/off ratio than MRAM, less power consumption than PCM, and a higher compatibility with CMOS process than FeFET [30,102]. Moreover, its efficient performance in MVM operation with a crossbar structure has aroused extensive attention as a promising candidate to implement embedded NVM-based CIM [41,87–95]. On the other hand, spintronic

devices provide a superior solution for nonvolatile logic-in-memory (LIM) architecture, enabling the efficient integration of a broad memory bandwidth in logic circuits [103,104]. STT-MRAM, as a representative spintronic memory, stands out for its lower access latency, superior endurance, and better process variation control compared to RRAM and PCM, making it well suited for embedded CIM [103,105,106]. Two silicon-validated examples of CIM will be reviewed later to explore the considerations in CIM design based on RRAM and MRAM technology.

Traditional RRAM-based CIM designs face two major issues in energy harvesting systems [107]: (1) the digital-to-analog (D/A) and analog-to-digital (A/D) conversion circuits between the RRAM array and the CPU significantly reduce energy efficiency and increase chip size; (2) all access transistors have to be turned on during each MVM operation, resulting in high sneak currents and unnecessary energy consumption. To overcome the existing limitations, a redesigned low-power MVM engine (Figure 6) has been introduced, which incorporates a binary interface and input-controlled access transistors [107]. By incorporating the binary interface, a direct link is established between the binary input vector and the WLs, with outputs obtained through the 1-to-3-bit adaptive SAs at the end of the BLs. This cuts out the A/D and D/A overheads, saving 44% in energy and 95% in area. Moreover, a 64% energy reduction is achieved by keeping access transistors off when inputs are zero. The proposed structure provides notable benefits particularly for networks with binary weights and input/output and contributes to the development of a smart processor that attains an energy efficiency of 462 GOPs/J at a clock frequency of 20 MHz.
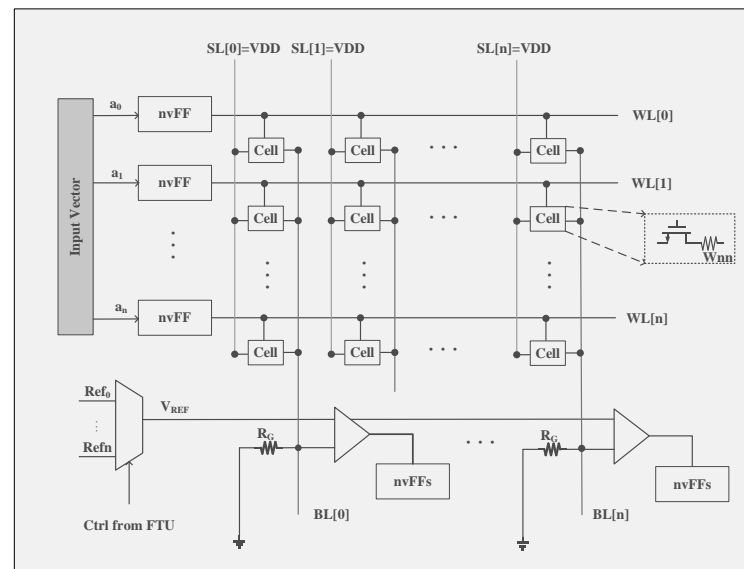

Figure 6. Architecture of an RRAM-based MVM engine for processing-in-memory [107].

Previous research has mainly focused on small-scale primitive logic-circuit elements, memory-like structures like FPGAs and filters, or simulation-based assessments, due to the lack of a well-defined design process tailored to MTJ technology in the chip fabrication environment. From this point of view, design flows for MTJ/MOS-hybrid logic circuits have been presented to realize practical-scale logic LSI based on a nonvolatile LIM architecture [103,108]. Utilizing the design flow, Figure 7 showcases a compact MTJ-based full adder (FA) with nonvolatile LIM architecture, enabling efficient, fully parallel processing for high-speed motion-vector extraction. The proposed FA exhibits a dynamic power consumption of 16.3 µW at 500 MHz, significantly lower than CMOS-only-based FA designs. A motion vector prediction unit was developed, comprising twenty-five processing elements (PEs) equipped with the reported FAs. It maintains intermediate data in nonvolatile memory, enabling precise power control during each operation cycle, which further diminishes leakage power and total power consumption [61].
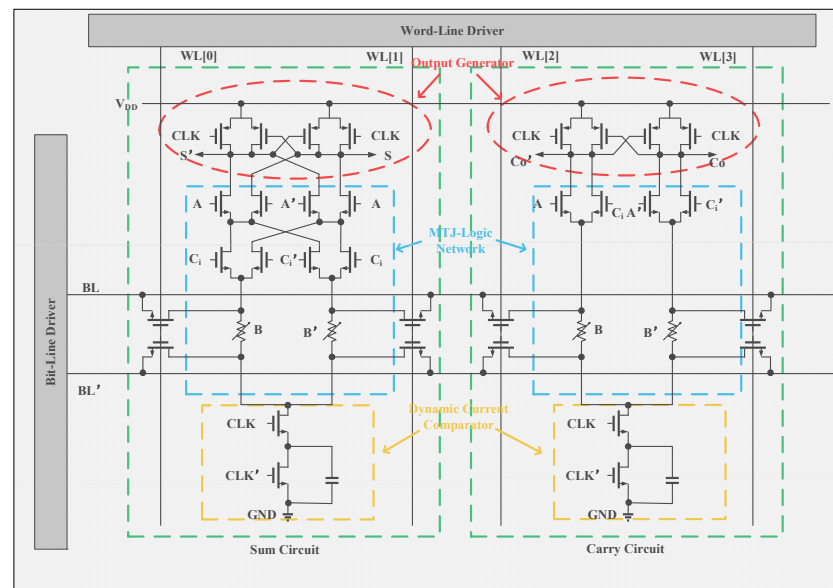
Figure 7. Circuit diagram of an MTJ-based nonvolatile full adder [61].

Table 3 shows a comparison of state-of-the-art CIM designs based on volatile SRAM [36,38,39,109] and emerging nonvolatile memories [100,110–114]. For SRAM-based CIM design, various techniques were proposed to efficiently perform multiply-accumulate (MAC) operations, such as bit-serial multiplication and parallel adder trees [109], a segmented-BL charge-sharing scheme [38], and a time-domain incremental-accumulation scheme [39]. Additionally, innovative circuit structures and schemes, including 6T local-computing cells [36], source-injection local multiplication cells, prioritized-hybrid-ADC [38], and dynamic differential-reference time-to-digital converter [39], were incorporated to reduce energy consumption. Consequently, these SRAM-based CIMs achieved superior output accuracy, fast operation speeds, and high energy efficiency. Nevertheless, the limited capacity, volatility, and large leakage current impeded their deployment in intricate neural network architectures. As depicted in Table 3, a mass storage capacity of 2.25 MB was attained utilizing RRAM, leveraging its compact footprint enabled by the 1T1R structure [100]. On the other hand, MRAM achieved accelerated computational speeds compared to RRAM, with a latency of 5 ns for 1-bit input and 1–8-bit weight configurations [114]. Additionally, the nonvolatility of emerging memories allowed the complete power-down of unselected cells, leading to substantial reductions in standby power and thereby enhancing energy efficiency when compared to SRAM-based CIMs.

Table 3. Comparison of state-of-the-art CIM based on SRAM, RRAM, and MRAM.

Performance Metrics	ISSCC'20 [36]	ISSCC'21 [109]	ISSCC'21 [38]	ISSCC'22 [39]	ISSCC'20 [111]	ISSCC'21 [110]	ISSCC'22 [100]	VLSI'19 [112]	VLSI'20 [113]	ISCAS'23 [114]
Technology	28 nm CMOS	22 nm CMOS	28 nm CMOS	28 nm CMOS	22 nm CMOS, RRAM	40 nm CMOS, RRAM	40 nm CMOS, RRAM	22 nm CMOS, MRAM	22 nm CMOS, MRAM	28 nm CMOS, MRAM
Capacity	64 KB	64 KB	384 KB	1 MB	256 KB	8 KB	2.25 MB	/	/	/
Supply [V]	0.7–0.9	0.72	0.7–0.9	0.65–0.9	0.7–0.9	0.9	0.9	/	/	1
Input Precision [bit]	4/4/8	1–8	4/8	4/8	1–4	1–8	1–8	/	/	1–16
Weight Precision [bit]	4/8/8	4/8/12/16	4/8	4/8	2–4	1–8	1–8	/	1.7	1–8
Output Precision [bit]	12/16/20	16 (4b/4b) 24 (8b/8b)	12/20	14/22	6–11	20	32	4	4	8–16 (1b IN) 24–32 (1b IN)

Table 3. *Cont*.

Performance Metrics	ISSCC'20 [36]	ISSCC'21 [109]	ISSCC'21 [38]	ISSCC'22 [39]	ISSCC'20 [111]	ISSCC'21 [110]	ISSCC'22 [100]	VLSI'19 [112]	VLSI'20 [113]	ISCAS'23 [114]
Energy efficiency [TOPS/W]	47.85–68.44/ 23.26–33.52/ 11.54–16.63	24.7 (8/8/24b) 89 (4/4/16b)	60.28– 94.31/ 15.02–22.75	84.45– 112.6/ 21.19–27.75	121.38	56.67	60.64	9.2	19.6	25.43 (1/8/15b) 129.83 (1/1/8b)

5. Conclusions

The rapid progress of edge computing has led to a growing need for emerging NVM technologies with low power consumption, high speed, and long-term durability. This paper explores the potential of four emerging NVMs in replacing conventional MCU memories and demonstrates their unique advantages. The discussion on NVM circuit design focuses on bitcell structures, read and write circuits, and macro structures, summarizing existing strategies for optimizing area, energy, and reliability. Moreover, previous works indicate that RRAM and MRAM offer notable benefits in CIM applications. Novel circuit designs, such as a binary interface structure and spintronic device integration, are effective in improving area and energy efficiency in CIM macros. As a future prospect, there is a need for universally applicable and energy-efficient strategies to leverage emerging NVMs in both MCUs and CIMs.

Author Contributions: Methodology, L.Q.; investigation, J.F.; writing—original draft preparation, J.F.; writing—review and editing, L.Q. and H.C.; visualization, Z.F. and J.F.; supervision, L.Q. and H.C.; funding acquisition, L.Q. and H.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alioto, M.; Sánchez-Sinencio, E.; Sangiovanni-Vincentelli, A. Guest Editorial Special Issue on Circuits and Systems for the Internet of Things—From Sensing to Sensemaking. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2017**, *64*, 2221–2225. [CrossRef]
2. Soriano, T.; Novo, D.; Prenat, G.; Pendina, G.D.; Benoit, P. MemCork: Exploration of Hybrid Memory Architectures for Intermittent Computing at the Edge. In Proceedings of the 2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC), Patras, Greece, 3–5 October 2022; pp. 1–6.
3. Kamei, A.; Kojima, T.; Amano, H.; Yokoyama, D.; Miyauchi, H.; Usami, K.; Hiraga, K.; Suzuki, K.; Bessho, K. Energy Saving in a Multi-Context Coarse Grained Reconfigurable Array with Non-Volatile Flip-Flops. In Proceedings of the 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Singapore, 20–23 December 2021; pp. 273–280.
4. Kroener, M. Energy Harvesting Technologies: Energy Sources, Generators and Management for Wireless Autonomous Applications. In Proceedings of the International Multi-Conference on Systems, Signals & Devices, Chemnitz, Germany, 20–23 March 2012; pp. 1–4.
5. Fu, S.; Narayanan, V.; Wymore, M.L.; Deep, V.; Duwe, H.; Qiao, D. No Battery, No Problem: Challenges and Opportunities in Batteryless Intermittent Networks. *J. Commun. Netw.* **2023**, *25*, 806–813. [CrossRef]
6. Sliper, S.T.; Wang, W.; Nikoleris, N.; Weddell, A.S.; Savanth, A.; Prabhat, P.; Merrett, G.V. Pragmatic Memory-System Support for Intermittent Computing Using Emerging Nonvolatile Memory. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2023**, *42*, 95–108. [CrossRef]
7. Nakamura, H.; Nakada, T.; Miwa, S. Normally-off Computing Project: Challenges and Opportunities. In Proceedings of the 2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC), Singapore, 20–23 January 2014; pp. 1–5.
8. Balsamo, D.; Das, A.; Weddell, A.S.; Brunelli, D.; Al-Hashimi, B.M.; Merrett, G.V.; Benini, L. Graceful Performance Modulation for Power-Neutral Transient Computing Systems. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2016**, *35*, 738–749. [CrossRef]
9. Jayakumar, H.; Raha, A.; Raghunathan, V. QUICKRECALL: A Low Overhead HW/SW Approach for Enabling Computations across Power Cycles in Transiently Powered Computers. In Proceedings of the 2014 27th International Conference on VLSI Design and 2014 13th International Conference on Embedded Systems, Mumbai, India, 5–9 January 2014. Available online: https://ieeexplore.ieee.org/document/6733152 (accessed on 22 February 2024).

10. Liu, Y.; Su, F.; Yang, Y.; Wang, Z.; Wang, Y.; Li, Z.; Li, X.; Yoshimura, R.; Naiki, T.; Tsuwa, T.; et al. A 130-Nm Ferroelectric Nonvolatile System-on-Chip with Direct Peripheral Restore Architecture for Transient Computing System. *IEEE J. Solid-State Circuits* **2019**, *54*, 885–895. [CrossRef]

11. Khanna, S.; Bartling, S.; Clinton, M.; Summerfelt, S.; Rodriguez, J.; McAdams, H. An FRAM-Based Nonvolatile Logic MCU SoC Exhibiting 100% Digital State Retention at VDD = 0 V Achieving Zero Leakage with <400-Ns Wakeup Time for ULP Applications. *IEEE J. Solid-State Circuits* **2013**, *49*, 95–106.

12. Thirumala, S.K.; Raha, A.; Raghunathan, V.; Gupta, S.K. IPS-CiM: Enhancing Energy Efficiency of Intermittently-Powered Systems with Compute-in-Memory. In Proceedings of the 2020 IEEE 38th International Conference on Computer Design (ICCD), Hartford, CT, USA, 18–21 October 2020; pp. 368–376.

13. Müller, J.; Yurchuk, E.; Schlösser, T.; Paul, J.; Hoffmann, R.; Müller, S.; Martin, D.; Slesazeck, S.; Polakowski, P.; Sundqvist, J.; et al. Ferroelectricity in HfO$_2$ Enables Nonvolatile Data Storage in 28 Nm HKMG. In Proceedings of the 2012 Symposium on VLSI Technology (VLSIT), Honolulu, HI, USA, 12–14 June 2012. Available online: https://ieeexplore.ieee.org/document/6242443 (accessed on 22 February 2024).

14. Thirumala, S.; Raha, A.; Gupta, S.; Raghunathan, V. Exploring the Design of Energy-Efficient Intermittently Powered Systems Using Reconfigurable Ferroelectric Transistors. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2022**, *30*, 365–378. [CrossRef]

15. Aswathy, N.; Sivamangai, N.M. Future Nonvolatile Memory Technologies: Challenges and Applications. In Proceedings of the 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Ernakulam, India, 2–4 September 2021; pp. 308–312.

16. Chun, K.C.; Zhao, H.; Harms, J.D.; Kim, T.-H.; Wang, J.-P.; Kim, C.H. A Scaling Roadmap and Performance Evaluation of In-Plane and Perpendicular MTJ Based STT-MRAMs for High-Density Cache Memory. *IEEE J. Solid-State Circuits* **2013**, *48*, 598–610. [CrossRef]

17. Roy, K.; Chakraborty, I.; Ali, M.; Ankit, A.; Agrawal, A. In-Memory Computing in Emerging Memory Technologies for Machine Learning: An Overview. In Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 20–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.

18. Chen, W.-H.; Khwa, W.-S.; Li, J.-Y.; Lin, W.-Y.; Lin, H.-T.; Liu, Y.; Wang, Y.; Wu, H.; Yang, H.; Chang, M.-F. Circuit Design for beyond von Neumann Applications Using Emerging Memory: From Nonvolatile Logics to Neuromorphic Computing. In Proceedings of the 2017 18th International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, USA, 14–15 March 2017; pp. 23–28.

19. Gong, H.; He, H.; Pan, L.; Gao, B.; Tang, J.; Pan, S.; Li, J.; Yao, P.; Wu, D.; Qian, H.; et al. An Error-Free 64KB ReRAM-Based nvSRAM Integrated to a Microcontroller Unit Supporting Real-Time Program Storage and Restoration. *IEEE Trans. Circuits Syst. I* **2023**, *70*, 5339–5351. [CrossRef]

20. Izumi, S.; Yamashita, K.; Nakano, M.; Nakagawa, T.; Kitahara, Y.; Yanagida, K.; Yoshimoto, S.; Kawaguchi, H.; Kimura, H.; Marumoto, K.; et al. Normally off ECG SoC with Non-Volatile MCU and Noise Tolerant Heartbeat Detector. In Proceedings of the 2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings, Lausanne, Switzerland, 22–24 October 2014; pp. 280–283.

21. Gong, H.; He, H.; Gao, B.; Tang, J.; Yu, J.; Wu, D.; Chen, J.; Zhang, Q.; Mou, X.; Qian, H.; et al. A 1-Mb Programming Configurable ReRAM Fully Integrating into a 32-Bit Microcontroller Unit. *IEEE Trans. Circuits Syst. II* **2023**, *70*, 2734–2738. [CrossRef]

22. Pasotti, M.; Zurla, R.; Carissimi, M.; Auricchio, C.; Brambilla, D.; Calvetti, E.; Capecchi, L.; Croce, L.; Gallinari, D.; Mazzaglia, C.; et al. A 32-KB ePCM for Real-Time Data Processing in Automotive and Smart Power Applications. *IEEE J. Solid-State Circuits* **2018**, *53*, 2114–2125. [CrossRef]

23. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.

24. Lin, C.-T.; Huang, P.X.; Oh, J.; Wang, D.; Seok, M. iMCU: A 28-nm Digital In-Memory Computing-Based Microcontroller Unit for TinyML. *IEEE J. Solid-State Circuits* **2024**, 1–10. [CrossRef]

25. Han, H.; Siebert, J. TinyML: A Systematic Review and Synthesis of Existing Research. In Proceedings of the 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 21–24 February 2022; pp. 269–274.

26. Tsoukas, V.; Gkogkidis, A.; Kakarountas, A. Internet of Things Challenges and the Emerging Technology of TinyML. In Proceedings of the 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Pafos, Cyprus, 19–21 June 2023; pp. 491–495.

27. Jia, H.; Valavi, H.; Tang, Y.; Zhang, J.; Verma, N. A Programmable Heterogeneous Microprocessor Based on Bit-Scalable In-Memory Computing. *IEEE J. Solid-State Circuits* **2020**, *55*, 2609–2621. [CrossRef]

28. Kim, S.; Yoo, H.-J. An Overview of Computing-in-Memory Circuits with DRAM and NVM. *IEEE Trans. Circuits Syst. II Express Briefs* **2024**, *71*, 1626–1631. [CrossRef]

29. Shanbhag, N.R.; Roy, S.K. Benchmarking In-Memory Computing Architectures. *IEEE Open J. Solid-State Circuits Soc.* **2022**, *2*, 288–300. [CrossRef]

30. Yu, S.; Jiang, H.; Huang, S.; Peng, X.; Lu, A. Compute-in-Memory Chips for Deep Learning: Recent Trends and Prospects. *IEEE Circuits Syst. Mag.* **2021**, *21*, 31–56. [CrossRef]

31. Biswas, A.; Chandrakasan, A.P. Conv-RAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-Based Machine Learning Applications. In Proceedings of the 2018 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, CA, USA, 11–15 February 2018; pp. 488–490.

32. Wang, J.; Wang, X.; Eckert, C.; Subramaniyan, A.; Das, R.; Blaauw, D.; Sylvester, D. 14.2 A Compute SRAM with Bit-Serial Integer/Floating-Point Operations for Programmable In-Memory Vector Acceleration. In Proceedings of the 2019 IEEE International Solid- State Circuits Conference-(ISSCC), San Francisco, CA, USA, 17–21 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 224–226.

33. Jiang, Z.; Yin, S.; Seok, M.; Seo, J. XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks. In Proceedings of the 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 18–22 June 2018; pp. 173–174.

34. Si, X.; Chen, J.-J.; Tu, Y.-N.; Huang, W.-H.; Wang, J.-H.; Chiu, Y.-C.; Wei, W.-C.; Wu, S.-Y.; Sun, X.; Liu, R.; et al. 24.5 A Twin-8T SRAM Computation-In-Memory Macro for Multiple-Bit CNN-Based Machine Learning. In Proceedings of the 2019 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, CA, USA, 17–21 February 2019; pp. 396–398.

35. Jiang, Z.; Yin, S.; Seo, J.-S.; Seok, M. C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism. *IEEE J. Solid-State Circuits* **2020**, *55*, 1888–1897. [CrossRef]

36. Si, X.; Tu, Y.-N.; Huang, W.-H.; Su, J.-W.; Lu, P.-J.; Wang, J.-H.; Liu, T.-W.; Wu, S.-Y.; Liu, R.; Chou, Y.-C.; et al. 15.5 A 28nm 64Kb 6T SRAM Computing-in-Memory Macro with 8b MAC Operation for AI Edge Chips. In Proceedings of the 2020 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, CA, USA, 16–20 February 2020; pp. 246–248.

37. Lee, J.; Valavi, H.; Tang, Y.; Verma, N. Fully Row/Column-Parallel In-Memory Computing SRAM Macro Employing Capacitor-Based Mixed-Signal Computation with 5-b Inputs. In Proceedings of the 2021 Symposium on VLSI Circuits, Kyoto, Japan, 13–19 June 2021; pp. 1–2.

38. Su, J.-W.; Chou, Y.-C.; Liu, R.; Liu, T.-W.; Lu, P.-J.; Wu, P.-C.; Chung, Y.-L.; Hung, L.-Y.; Ren, J.-S.; Pan, T.; et al. 16.3 A 28nm 384kb 6T-SRAM Computation-in-Memory Macro with 8b Precision for AI Edge Chips. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; Volume 64, pp. 250–252.

39. Wu, P.-C.; Su, J.-W.; Chung, Y.-L.; Hong, L.-Y.; Ren, J.-S.; Chang, F.-C.; Wu, Y.; Chen, H.-Y.; Lin, C.-H.; Hsiao, H.-M.; et al. A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices. In Proceedings of the 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2022; Volume 65, pp. 1–3.

40. Hung, J.-M.; Jhang, C.-J.; Wu, P.-C.; Chiu, Y.-C.; Chang, M.-F. Challenges and Trends of Nonvolatile In-Memory-Computation Circuits for AI Edge Devices. *IEEE Open J. Solid-State Circuits Soc.* **2021**, *1*, 171–183. [CrossRef]

41. Huang, W.-H.; Wen, T.-H.; Hung, J.-M.; Khwa, W.-S.; Lo, Y.-C.; Jhang, C.-J.; Hsu, H.-H.; Chin, Y.-H.; Chen, Y.-C.; Lo, C.-C.; et al. A Nonvolatile Al-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W. In Proceedings of the 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 19–23 February 2023; pp. 15–17.

42. Joo, S.; An, Y.-J.; Oh, T.W.; Jung, S.-O. Comparative Analysis of MCU Memory for IoT Application. In Proceedings of the 2018 International Conference on Electronics, Information, and Communication (ICEIC), Honolulu, HI, USA, 24–27 January 2018; pp. 1–3.

43. Strenz, R. Embedded Flash Technologies and Their Applications: Status & Outlook. In Proceedings of the 2011 International Electron Devices Meeting, Washington, DC, USA, 5–7 December 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 9.4.1–9.4.4.

44. Maurelli, A. Status and Perspectives of Embedded Non-Volatile Memories. In Proceedings of the 2013 International Conference on IC Design & Technology (ICICDT), Pavia, Italy, 29–31 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 77–80.

45. Strenz, R. Review and Outlook on Embedded NVM Technologies–From Evolution to Revolution. In Proceedings of the 2020 IEEE International Memory Workshop (IMW), Dresden, Germany, 17–20 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–4.

46. Kono, T.; Ito, T.; Tsuruda, T.; Nishiyama, T.; Nagasawa, T.; Ogawa, T.; Kawashima, Y.; Hidaka, H.; Yamauchi, T. 40-Nm Embedded Split-Gate MONOS (SG-MONOS) Flash Macros for Automotive with 160-MHz Random Access for Code and Endurance Over 10 M Cycles for Data at the Junction Temperature of 170 °C. *IEEE J. Solid-State Circuits* **2014**, *49*, 154–166. [CrossRef]

47. Yamauchi, T.; Yamaguchi, Y.; Kono, T.; Hidaka, H. Embedded Flash Technology for Automotive Applications. In Proceedings of the 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 3–7 December 2016; pp. 28.6.1–28.6.4.

48. Jefremow, M.; Kern, T.; Backhausen, U.; Elbs, J.; Rousseau, B.; Roll, C.; Castro, L.; Roehr, T.; Paparisto, E.; Herfurth, K.; et al. A 65nm 4MB Embedded Flash Macro for Automotive Achieving a Read Throughput of 5.7GB/s and a Write Throughput of 1.4MB/s. In Proceedings of the 2013 Proceedings of the (ESSCIRC), Bucharest, Romania, 16–20 September 2013; pp. 193–196.

49. Rosa, F.L.; Niel, S.; Regnier, A.; Maugain, F.; Mantelli, M.; Conte, A. 40nm Embedded Select in Trench Memory (eSTM) Technology Overview. In Proceedings of the 2019 IEEE 11th International Memory Workshop (IMW), Monterey, CA, USA, 12–15 May 2019; pp. 1–4.

50. Nakano, M.; Kaneda, Y.; Nakanishi, S.; Murai, Y.; Tashiro, Y.; Taito, Y.; Ogawa, T.; Mitani, H.; Ito, T.; Kono, T. A 40-Nm Embedded SG-MONOS Flash Macro for High-End MCU Achieving 200-MHz Random Read Operation and 7.91-Mb/Mm^2 Density with Charge-Assisted Offset Cancellation Sense Amplifier. *IEEE J. Solid-State Circuits* **2022**, *57*, 3094–3102. [CrossRef]

51. Bartling, S.C.; Khanna, S.; Clinton, M.P.; Summerfelt, S.R.; Rodriguez, J.A.; McAdams, H.P. An 8MHz 75μA/MHz Zero-Leakage Non-Volatile Logic-Based Cortex-M0 MCU SoC Exhibiting 100% Digital State Retention at VDD=0V with <400ns Wakeup and Sleep Transitions. In Proceedings of the 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 17–21 February 2013; pp. 432–433.

52. Liu, Y.; Wang, Z.; Lee, A.; Su, F.; Lo, C.-P.; Yuan, Z.; Lin, C.-C.; Wei, Q.; Wang, Y.; King, Y.-C.; et al. 4.7 A 65nm ReRAM-Enabled Nonvolatile Processor with 6× Reduction in Restore Time and 4× Higher Clock Frequency Using Adaptive Data Retention and Self-Write-Termination Nonvolatile Logic. In Proceedings of the 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 31 January–4 February 2016; pp. 84–86.

53. Giordano, M.; Prabhu, K.; Koul, K.; Radway, R.M.; Gural, A.; Doshi, R.; Khan, Z.F.; Kustin, J.W.; Liu, T.; Lopes, G.B.; et al. CHIMERA: A 0.92 TOPS, 2.2 TOPS/W Edge AI Accelerator with 2 MByte On-Chip Foundry Resistive RAM for Efficient Training and Inference. In Proceedings of the 2021 Symposium on VLSI Circuits, Kyoto, Japan, 13–19 June 2021; pp. 1–2.

54. Chien, T.-K.; Chiou, L.-Y.; Sheu, S.-S.; Lin, J.-C.; Lee, C.-C.; Ku, T.-K.; Tsai, M.-J.; Wu, C.-I. Low-Power MCU with Embedded ReRAM Buffers as Sensor Hub for IoT Applications. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2016**, *6*, 247–257. [CrossRef]

55. Rossi, D.; Conti, F.; Eggiman, M.; Mach, S.; Mauro, A.D.; Guermandi, M.; Tagliavini, G.; Pullini, A.; Loi, I.; Chen, J.; et al. 4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7μW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; Volume 64, pp. 60–62.

56. Fan, Z.; An, H.; Zhang, Q.; Xu, B.; Xu, L.; Tseng, C.-W.; Peng, Y.; Cao, A.; Liu, B.; Lee, C.; et al. Audio and Image Cross-Modal Intelligence via a 10TOPS/W 22nm SoC with Back-Propagation and Dynamic Power Gating. In Proceedings of the 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 13–17 June 2022; pp. 18–19.

57. Zhang, Q.; An, H.; Fan, Z.; Wang, Z.; Li, Z.; Wang, G.; Kim, H.-S.; Blaauw, D.; Sylvester, D. A 22nm 3.5TOPS/W Flexible Micro-Robotic Vision SoC with 2MB eMRAM for Fully-on-Chip Intelligence. In Proceedings of the 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 13–17 June 2022; pp. 72–73.

58. Grossier, N.; Disegni, F.; Ventre, A.; Barcella, A.; Mariani, R.; Marino, V.; Mazzara, S.; Scavuzzo, A.; Bansal, M.; Soni, B.; et al. ASIL-D Automotive-Grade Microcontroller in 28nm FD-SOI with Full-OTA Capable 21MB Embedded PCM Memory and Highly Scalable Power Management. In Proceedings of the 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Kyoto, Japan, 11 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–2.

59. Ogawa, T.; Matsubara, K.; Taito, Y.; Saito, T.; Izuna, M.; Takeda, K.; Kaneda, Y.; Shimoi, T.; Mitani, H.; Ito, T.; et al. 15.8 A 22nm 10.8Mb Embedded STT-MRAM Macro Achieving over 200MHz Random-Read Access and a 10.4MB/s Write Throughput with an In-Field Programmable 0.3Mb MTJ-OTP for High-End MCUs. In Proceedings of the 2024 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 18–22 February 2024; Volume 67, pp. 290–292.

60. Tsuji, Y.; Bai, X.; Miyamura, M.; Sakamoto, T.; Tada, M.; Banno, N.; Okamoto, K.; Iguchi, N.; Sugii, N.; Hada, H. Sub-μW Standby Power, <18 μW/DMIPS@25MHz MCU with Embedded Atom-Switch Programmable Logic and ROM. In Proceedings of the 2015 Symposium on VLSI Technology (VLSI Technology), Kyoto, Japan, 16–18 June 2015; pp. T86–T87.

61. Hanyu, T.; Endoh, T.; Suzuki, D.; Koike, H.; Ma, Y.; Onizawa, N.; Natsui, M.; Ikeda, S.; Ohno, H. Standby-Power-Free Integrated Circuits Using MTJ-Based VLSI Computing. *Proc. IEEE* **2016**, *104*, 1844–1863. [CrossRef]

62. Hou, Y.; Wang, K.; Liu-Sun, C.; Hang, J.; Tong, X.; Peng, C.; Wu, Y.; Ren, Y.; Bu, W.; Si, X.; et al. A Sub-100nA Ultra-Low Leakage MCU Embedding Always-on Domain Hybrid Tunnel FET-CMOS on 300mm Foundry Platform. In Proceedings of the 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 9–13 December 2023; pp. 1–4.

63. Natsui, M.; Suzuki, D.; Tamakoshi, A.; Watanabe, T.; Honjo, H.; Koike, H.; Nasuno, T.; Ma, Y.; Tanigawa, T.; Noguchi, Y.; et al. A 47.14-μW 200-MHz MOS/MTJ-Hybrid Nonvolatile Microcontroller Unit Embedding STT-MRAM and FPGA for IoT Applications. *IEEE J. Solid-State Circuits* **2019**, *54*, 2991–3004. [CrossRef]

64. Raha, A.; Jaiswal, A.; Sarwar, S.S.; Jayakumar, H.; Raghunathan, V.; Roy, K. Designing Energy-Efficient Intermittently Powered Systems Using Spin-Hall-Effect-Based Nonvolatile SRAM. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2018**, *26*, 294–307. [CrossRef]

65. Jew, T. MRAM in Microcontroller and Microprocessor Product Applications. In Proceedings of the 2020 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 12–18 December 2020; pp. 11.1.1–11.1.4.

66. Fukuda, T.; Kohara, K.; Dozaka, T.; Takeyama, Y.; Midorikawa, T.; Hashimoto, K.; Wakiyama, I.; Miyano, S.; Hojo, T. 13.4 A 7ns-Access-Time 25μW/MHz 128kb SRAM for Low-Power Fast Wake-up MCU in 65nm CMOS with 27fA/b Retention Current. In Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, 9–13 February 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 236–237.

67. Walter, D.; Scharfe, A.; Oefelein, A.; Schraut, F.; Bauer, H.; Csaszar, F.; Niebsch, R.; Schreiter, J.; Eisenreich, H.; Höppner, S. A 0.55V 6.3uW/MHz Arm Cortex-M4 MCU with Adaptive Reverse Body Bias and Single Rail SRAM. In Proceedings of the 2020 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS), Tokyo, Japan, 15–17 April 2020; pp. 1–3.

68. Yokoyama, Y.; Miura, T.; Ouchi, Y.; Nakamura, D.; Ishikawa, J.; Nagata, S. 40-Nm 64-Kbit Buffer/Backup SRAM with 330 nW Standby Power at 65 °C Using 3.3 V 10 MOSs for PMIC Less MCU in IoT Applications. In Proceedings of the 2018 IEEE Asian Solid-State Circuits Conference (A-SSCC), Tainan, Taiwan, 5–7 November 2018.

69. Yokoyama, Y.; Goto, K.; Miura, T.; Ouchi, Y.; Nakamura, D.; Ishikawa, J.; Nagata, S.; Tsujihashi, Y.; Ishii, Y. A Cost Effective Test Screening Circuit for Embedded SRAM with Resume Standby on 110-Nm SoC/MCU. In Proceedings of the 2019 IEEE Asian Solid-State Circuits Conference (A-SSCC), Macau, China, 4–6 November 2019; pp. 17–20.

70. Majumdar, S. Single Bit-Line Differential Sensing Based Real-Time NVSRAM for Low Power Applications. *IEEE Trans. Circuits Syst. II Express Briefs* **2021**, *68*, 2623–2627. [CrossRef]

71. Ohsawa, T.; Koike, H.; Miura, S.; Honjo, H.; Kinoshita, K.; Ikeda, S.; Hanyu, T.; Ohno, H.; Endoh, T. A 1 Mb Nonvolatile Embedded Memory Using 4T2MTJ Cell with 32 b Fine-Grained Power Gating Scheme. *IEEE J. Solid-State Circuits* **2013**, *48*, 1511–1520. [CrossRef]

72. Kuk, S.-H.; Han, J.-H.; Kim, B.H.; Kim, J.; Kim, S.-H. Proposal of P-Channel FE NAND with High Drain Current and Feasible Disturbance for Next Generation 3D NAND. In Proceedings of the 2023 IEEE International Memory Workshop (IMW), Monterey, CA, USA, 21–24 May 2023; pp. 1–4.

73. Takahashi, M.; Zhang, W.; Sakai, S. High-Endurance Ferroelectric NOR Flash Memory Using (Ca,Sr)Bi2Ta2O9 FeFETs. In Proceedings of the 2018 IEEE International Memory Workshop (IMW), Kyoto, Japan, 13–16 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.

74. Sharma, A.; Roy, K. 1T Non-Volatile Memory Design Using Sub-10nm Ferroelectric FETs. *IEEE Electron Device Lett.* **2018**, *39*, 359–362. [CrossRef]

75. Ni, K.; Li, X.; Smith, J.A.; Jerry, M.; Datta, S. Write Disturb in Ferroelectric FETs and Its Implication for 1T-FeFET AND Memory Arrays. *IEEE Electron Device Lett.* **2018**, *39*, 1656–1659. [CrossRef]

76. Yu, H.-C.; Lin, K.-C.; Lin, K.-F.; Huang, C.-Y.; Chih, Y.-D.; Ong, T.-C.; Chang, J.; Natarajan, S.; Tran, L.C. Cycling Endurance Optimization Scheme for 1Mb STT-MRAM in 40nm Technology. In Proceedings of the 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, USA, 17–21 February 2013; pp. 224–225.

77. Alam, S.M.; Houssameddine, D.; Neumeyer, F.; Rahman, I.; DeHerrera, M.; Ikegawa, S.; Sanchez, P.; Zhang, X.; Wang, Y.; Williams, J.; et al. Persistent xSPI STT-MRAM with up to 400MB/s Read and Write Throughput. In Proceedings of the 2022 IEEE International Memory Workshop (IMW), Monterey, CA, USA, 21–24 May 2022; pp. 1–4.

78. Yang, J.; Xue, X.; Xu, X.; Wang, Q.; Jiang, H.; Yu, J.; Dong, D.; Zhang, F.; Lv, H.; Liu, M. 24.2 A 14nm-FinFET 1Mb Embedded 1T1R RRAM with a 0.022μm^2 Cell Size Using Self-Adaptive Delayed Termination and Multi-Cell Reference. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; Volume 64, pp. 336–338.

79. Shao, Z.; Chang, N.; Dutt, N. PTL: PCM Translation Layer. In Proceedings of the 2012 IEEE Computer Society Annual Symposium on VLSI, Amherst, MA, USA, 19–21 August 2012; pp. 380–385.

80. Jaiswal, A.; Fong, X.; Roy, K. Comprehensive Scaling Analysis of Current Induced Switching in Magnetic Memories Based on In-Plane and Perpendicular Anisotropies. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2016**, *6*, 120–133. [CrossRef]

81. Sheu, S.-S.; Kuo, C.-C.; Chang, M.-F.; Tseng, P.-L.; Chih-Sheng, L.; Wang, M.-C.; Lin, C.-H.; Lin, W.-P.; Chien, T.-K.; Lee, S.-H.; et al. A ReRAM Integrated 7T2R Non-Volatile SRAM for Normally-off Computing Application. In Proceedings of the 2013 IEEE Asian Solid-State Circuits Conference (A-SSCC), Singapore, 11–13 November 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 245–248.

82. Chiu, P.-F.; Chang, M.-F.; Wu, C.-W.; Chuang, C.-H.; Sheu, S.-S.; Chen, Y.-S.; Tsai, M.-J. Low Store Energy, Low VDDmin, 8T2R Nonvolatile Latch and SRAM with Vertical-Stacked Resistive Memory (Memristor) Devices for Low Power Mobile Applications. *IEEE J. Solid-State Circuits* **2012**, *47*, 1483–1496. [CrossRef]

83. Dai, S.; Zhang, Y.; Zhang, H.; Li, J.; Lin, Y. A ReRAM-Based 10T2R SRAM Using Power-off Recovery Function for Reducing Power. In Proceedings of the 2021 IEEE 14th International Conference on ASIC (ASICON), Kunming, China, 26 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4.

84. Lee, A.; Chang, M.-F.; Lin, C.-C.; Chen, C.-F.; Ho, M.-S.; Kuo, C.-C.; Tseng, P.-L.; Sheu, S.-S.; Ku, T.-K. RRAM-Based 7T1R Non-volatile SRAM with 2x Reduction in Store Energy and 94x Reduction in Restore Energy for Frequent-off Instant-on Applications. In Proceedings of the 2015 Symposium on VLSI Circuits (VLSI Circuits), Kyoto, Japan, 16–18 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. C76–C77.

85. Abdelwahed, A.M.S.T.; Neale, A.; Anis, M.; Wei, L. 8T1R: A Novel Low-Power High-Speed RRAM-Based Non-Volatile SRAM Design. In Proceedings of the Proceedings of the 26th Edition on Great Lakes Symposium on VLSI, Boston, MA, USA, 18–20 May 2016; ACM: Boston, MA, USA, 2016; pp. 239–244.

86. Wei, W.; Namba, K.; Han, J.; Lombardi, F. Design of a Nonvolatile 7T1R SRAM Cell for Instant-on Operation. *IEEE Trans. Nanotechnol.* **2014**, *13*, 905–916. [CrossRef]

87. George, S.; Ma, K.; Aziz, A.; Li, X.; Khan, A.; Salahuddin, S.; Chang, M.-F.; Datta, S.; Sampson, J.; Gupta, S.; et al. Nonvolatile Memory Design Based on Ferroelectric FETs. In Proceedings of the 53rd Annual Design Automation Conference, Austin, TX, USA, 5–9 June 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1–6.

88. Li, X.; Wu, J.; Ni, K.; George, S.; Ma, K.; Sampson, J.; Gupta, S.K.; Liu, Y.; Yang, H.; Datta, S.; et al. Design of 2T/Cell and 3T/Cell Nonvolatile Memories with Emerging Ferroelectric FETs. *IEEE Des. Test* **2019**, *36*, 39–45. [CrossRef]

89. Hoya, K.; Takashima, D.; Shiratake, S.; Ogiwara, R.; Miyakawa, T.; Shiga, H.; Doumae, S.M.; Ohtsuki, S.; Kumura, Y.; Shuto, S.; et al. A 64-Mb Chain FeRAM with Quad BL Architecture and 200 MB/s Burst Mode. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2010**, *18*, 1745–1752. [CrossRef]

90. Takashima, D.; Nagadomi, Y.; Hatsuda, K.; Watanabe, Y.; Fujii, S. A 128 Mb Chain FeRAM and System Design for HDD Application and Enhanced HDD Performance. *IEEE J. Solid-State Circuits* **2011**, *46*, 530–536. [CrossRef]

91. Belmonte, A.; Degraeve, R.; Fantini, A.; Kim, W.; Houssa, M.; Jurczak, M.; Goux, L. Origin of the Deep Reset and Low Variability of Pulse-Programmed W\Al$_2$O$_3$\TiW\Cu CBRAM Device. In Proceedings of the 2014 IEEE 6th International Memory Workshop (IMW), Taipei, Taiwan, 18–21 May 2014; pp. 1–4.

92. Belmonte, A.; Kim, W.; Chan, B.; Heylen, N.; Fantini, A.; Houssa, M.; Jurczak, M.; Goux, L. 90nm W\Al$_2$O$_3$\TiW\Cu 1T1R CBRAM Cell Showing Low-Power, Fast and Disturb-Free Operation. In Proceedings of the 2013 5th IEEE International Memory Workshop, Monterey, CA, USA, 26–29 May 2013; pp. 26–29.

93. Zuliani, P.; Varesi, E.; Palumbo, E.; Borghi, M.; Tortorelli, I.; Erbetta, D.; Libera, G.D.; Pessina, N.; Gandolfo, A.; Prelini, C.; et al. Overcoming Temperature Limitations in Phase Change Memories with Optimized Ge$_x$Sb$_y$Te$_z$. *IEEE Trans. Electron Devices* **2013**, *60*, 4020–4026. [CrossRef]

94. Close, G.F.; Frey, U.; Morrish, J.; Jordan, R.; Lewis, S.C.; Maffitt, T.; BrightSky, M.J.; Hagleitner, C.; Lam, C.H.; Eleftheriou, E. A 256-Mcell Phase-Change Memory Chip Operating at 2+ Bit/Cell. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2013**, *60*, 1521–1533. [CrossRef]

95. Ciocchini, N.; Palumbo, E.; Borghi, M.; Zuliani, P.; Annunziata, R.; Ielmini, D. Modeling Resistance Instabilities of Set and Reset States in Phase Change Memory with Ge-Rich GeSbTe. *IEEE Trans. Electron Devices* **2014**, *61*, 2136–2144. [CrossRef]

96. Athmanathan, A.; Stanisavljevic, M.; Papandreou, N.; Pozidis, H.; Eleftheriou, E. Multilevel-Cell Phase-Change Memory: A Viable Technology. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2016**, *6*, 87–100. [CrossRef]

97. Chang, M.-F.; Lin, K.-F.; Chuang, C.-H.; Huang, L.-Y.; Chien, T.-F.; Sheu, S.-S.; Su, K.-L.; Lee, H.-Y.; Chen, F.T.; Lien, C.-H.; et al. Circuit Design Challenges and Trends in Read Sensing Schemes for Resistive-Type Emerging Nonvolatile Memory. In Proceedings of the 2012 IEEE 11th International Conference on Solid-State and Integrated Circuit Technology, Xi'an, China, 29 October–1 November 2012; pp. 1–4.

98. Xue, C.; Zhang, Y.; Chen, P.; Zhu, M.; Wu, T.; Wu, M.; He, Y.; Ye, L. Reliability-Improved Read Circuit and Self-Terminating Write Circuit for STT-MRAM in 16 Nm FinFET. In Proceedings of the 2022 IEEE International Symposium on Circuits and Systems (ISCAS), Austin, TX, USA, 27 May–1 June 2022; pp. 595–599.

99. Rajpoot, J.; Verma, S. Area-Efficient Auto-Write-Terminate Circuit for NV Latch and Logic-in-Memory Applications. *IEEE Trans. Circuits Syst. II Express Briefs* **2023**, *70*, 2630–2634. [CrossRef]

100. Chang, M.; Spetalnick, S.D.; Crafton, B.; Khwa, W.-S.; Chih, Y.-D.; Chang, M.-F.; Raychowdhury, A. A 40nm 60.64TOPS/W ECC-Capable Compute-in-Memory/Digital 2.25MB/768KB RRAM/SRAM System with Embedded Cortex M3 Microprocessor for Edge Recommendation Systems. In Proceedings of the 2022 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 18–22 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–3.

101. Wang, Z.; Liu, Y.; Lee, A.; Su, F.; Lo, C.-P.; Yuan, Z.; Li, J.; Lin, C.-C.; Chen, W.-H.; Chiu, H.-Y.; et al. A 65-Nm ReRAM-Enabled Nonvolatile Processor with Time-Space Domain Adaption and Self-Write-Termination Achieving > 4\times Faster Clock Frequency and > 6\times Higher Restore Speed. *IEEE J. Solid-State Circuits* **2017**, *52*, 2769–2785. [CrossRef]

102. Wang, L.; Ye, W.; An, J.; Dou, C.; Liu, Q.; Chang, M.-F.; Liu, M. Sparsity-Aware Clamping Readout Scheme for High Parallelism and Low Power Nonvolatile Computing-in-Memory Based on Resistive Memory. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Republic of Korea, 22–28 May 2021; pp. 1–4.

103. Natsui, M.; Hanyu, T.; Sakimura, N.; Sugibayashi, T. MTJ/MOS-Hybrid Logic-Circuit Design Flow for Nonvolatile Logic-in-Memory LSI. In Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, 19–23 May 2013; pp. 105–109.

104. Sakimura, N.; Nebashi, R.; Tsuji, Y.; Honjo, H.; Sugibayashi, T.; Koike, H.; Ohsawa, T.; Fukami, S.; Hanyu, T.; Ohno, H.; et al. High-Speed Simulator Including Accurate MTJ Models for Spintronics Integrated Circuit Design. In Proceedings of the 2012 IEEE International Symposium on Circuits and Systems, Seoul, Republic of Korea, 20–23 May 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1971–1974.

105. Wang, C.; Wang, Z.; Zhang, Y.; Zhao, W. Computing-in-Memory Paradigm Based on STT-MRAM with Synergetic Read/Write-like Modes. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Republic of Korea, 22–28 May 2021; pp. 1–5.

106. Wang, S.; Cai, H. Computing-in-Memory with Enhanced STT-MRAM Readout Margin. *IEEE Trans. Magn.* **2023**, *59*, 3401705. [CrossRef]

107. Su, F.; Chen, W.-H.; Xia, L.; Lo, C.-P.; Tang, T.; Wang, Z.; Hsu, K.-H.; Cheng, M.; Li, J.-Y.; Xie, Y.; et al. A 462GOPs/J RRAM-Based Nonvolatile Intelligent Processor for Energy Harvesting IoE System Featuring Nonvolatile Logics and Processing-in-Memory. In Proceedings of the 2017 Symposium on VLSI Technology, Kyoto, Japan, 5–8 June 2017; pp. T260–T261.

108. Natsui, M.; Suzuki, D.; Sakimura, N.; Nebashi, R.; Tsuji, Y.; Morioka, A.; Sugibayashi, T.; Miura, S.; Honjo, H.; Kinoshita, K.; et al. Nonvolatile Logic-in-Memory LSI Using Cycle-Based Power Gating and Its Application to Motion-Vector Prediction. *IEEE J. Solid-State Circuits* **2015**, *50*, 476–489. [CrossRef]

109. Chih, Y.-D.; Lee, P.-H.; Fujiwara, H.; Shih, Y.-C.; Lee, C.-F.; Naous, R.; Chen, Y.-L.; Lo, C.-P.; Lu, C.-H.; Mori, H.; et al. 16.4 An 89TOPS/W and 16.3TOPS/Mm^2 All-Digital SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; Volume 64, pp. 252–254.

110. Yoon, J.-H.; Chang, M.; Khwa, W.-S.; Chih, Y.-D.; Chang, M.-F.; Raychowdhury, A. 29.1 A 40nm 64Kb 56.67TOPS/W Read-Disturb-Tolerant Compute-in-Memory/Digital RRAM Macro with Active-Feedback-Based Read and In-Situ Write Verification. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; Volume 64, pp. 404–406.
111. Xue, C.-X.; Huang, T.-Y.; Liu, J.-S.; Chang, T.-W.; Kao, H.-Y.; Wang, J.-H.; Liu, T.-W.; Wei, S.-Y.; Huang, S.-P.; Wei, W.-C.; et al. 15.4 A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices. In Proceedings of the 2020 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, CA, USA, 16–20 February 2020; pp. 244–246.
112. Garello, K.; Yasin, F.; Hody, H.; Couet, S.; Souriau, L.; Sharifi, S.H.; Swerts, J.; Carpenter, R.; Rao, S.; Kim, W.; et al. Manufacturable 300mm Platform Solution for Field-Free Switching SOT-MRAM. In Proceedings of the 2019 Symposium on VLSI Circuits, Kyoto, Japan, 10–14 June 2019; pp. T194–T195.
113. Doevenspeck, J.; Garello, K.; Verhoef, B.; Degraeve, R.; Van Beek, S.; Crotti, D.; Yasin, F.; Couet, S.; Jayakumar, G.; Papistas, I.A.; et al. SOT-MRAM Based Analog in-Memory Computing for DNN Inference. In Proceedings of the 2020 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 16–19 June 2020; pp. 1–2.
114. Lu, L.; Mani, A.; Do, A.T. A 129.83 TOPS/W Area Efficient Digital SOT/STT MRAM-Based Computing-In-Memory for Advanced Edge AI Chips. In Proceedings of the 2023 IEEE International Symposium on Circuits and Systems (ISCAS), Monterey, CA, USA, 21–25 May 2023; pp. 1–5.