nature reviews electrical engineering

https://doi.org/10.1038/s44287-023-00002-9

Review article

Check for updates

High-speed emerging memories for AI hardware accelerators

Anni Lu [1,3], Junmo Lee [1,3], Tae-Hyeon Kim [1,3], Muhammed Ahosan Ul Karim [2], Rebecca Sejung Park [2], Harsono Simka [2] & Shimeng Yu [1] ✉

Abstract

Applications of artificial intelligence (AI) necessitate AI hardware accelerators able to efficiently process data-intensive and computation-intensive AI workloads. AI accelerators require two types of memory: the weight memory that stores the parameters of the AI models and the buffer memory that stores the intermediate input or output data when computing a portion of the AI models. In this Review, we present the recent progress in the emerging high-speed memory for AI hardware accelerators and survey the technologies enabling the global buffer memory in digital systolic-array architectures. Beyond conventional static random-access memory (SRAM), we highlight the following device candidates: capacitorless gain cell-based embedded dynamic random-access memories (eDRAMs), ferroelectric memories, spin-transfer torque magnetic random-access memory (STT-MRAM) and spin-orbit torque magnetic random-access memory (SOT-MRAM). We then summarize the research advances in the industrial development and the technological challenges in buffer memory applications. Finally, we present a systematic benchmarking analysis on a tensor processing unit (TPU)-like AI accelerator in the edge and in the cloud and evaluate the use of these emerging memories.

Sections

Introduction

High-speed memory candidates

TPU buffer memory case study and NeuroSim benchmarking

Array-level prototyping landscape

Outlook

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [2]Logic Pathfinding Laboratory, Samsung Semiconductor, Inc., San Jose, CA, USA. [3]These authors contributed equally: Anni Lu, Junmo Lee, Tae-Hyeon Kim. ✉e-mail: shimeng.yu@ece.gatech.edu

Key points

- The global buffer in artificial intelligence (AI) hardware (for example, the tensor processing unit (TPU)) is traditionally based on static random-access memory (SRAM), which is expensive in the silicon footprint and suffers from high stand-by leakage power. Emerging memories with high speed and high endurance could replace SRAM as global buffers.
- A capacitorless two-transistor (2T) gain cell, an implementation of embedded dynamic random-access memory (DRAM), uses amorphous oxide semiconductors as the channel material allowing a high data retention time.
- Ferroelectric memories such as the ferroelectric field effect transistor (FeFET) and magnetic memories such as spin-transfer torque magnetic random-access memory (STT-MRAM) or spin-orbit torque magnetic random-access memory (SOT-MRAM) could be tailored to improve their cycling endurance, making them viable as global buffer candidates.
- Three-dimensional integration that stacks emerging memories and their access transistors all together at the back end of line (BEOL) paves the way for high-density global buffer solutions that are even denser than the leading edge node SRAMs.
- Leading edge node SRAM is still a competitive high-performance technology for AI hardware in the cloud, whereas emerging memories exhibit more advantages in AI hardware at the edge where minimizing the stand-by leakage power is critical.

Introduction

Artificial intelligence (AI) enables a wide range of applications from computer vision to natural language processing. AI hardware accelerators are high-performance parallel computation machines specifically designed for the efficient processing of AI workloads beyond conventional central processing unit (CPU) or graphic processing unit (GPU) platforms (Supplementary Information section 1). Deep neural network (DNN) processing involves heavy multiply and accumulate (MAC) computations and intensive memory access due to the large size of the AI models. Specialized AI hardware can be used to accelerate the DNN training as well as the inference. In systolic array-based accelerators (Fig. 1), for example, each processing element (PE) independently computes a MAC operation of the data received from its upstream neighbours and passes this downstream to minimize the expensive memory access (Fig. 1, right). Digital systolic-array architectures (Supplementary Information section 1) have been gaining commercial success. For instance, Google has deployed the tensor processing unit (TPU) (Supplementary Information section 1), one of the most widely used commercial AI hardware products, in both data centres[1] and edge devices[2]. Since the early 2010s, another spotlighted research topic is the in-memory computing paradigm because of its high energy efficiency that benefits from reduced data transfer between memory and computing units. In compute-in-memory (CIM) engines, the DNN parameters are stored and directly computed inside the memory arrays. Previous reviews in the field focused on either the digital MAC engines[3,4] or the weight memories (Supplementary Information section 1) used in the mixed-signal or analogue CIM engines[5,6]; by contrast, the buffer memories (Supplementary Information section 1) that hold the intermediate input or output activation data are rarely discussed.

Figure 1 shows the generic architecture of the digital MAC engines used in the TPU-like architecture. At each level of the hierarchy, the intermediate data (that is, the input or output feature maps of a DNN) are temporarily stored in the buffer memories. At the top level of the hierarchy, there is a global buffer that has a capacity in the range of 1–100 MB, which is implemented by static random-access memory (SRAM) cache in conventional designs (Supplementary Information section 1). It is known that SRAM is widely used as the mainstream on-chip buffer memory for CPU or GPU thanks to its fast access (less than a few nanoseconds), unlimited endurance (>10^{16} cycles) and superior scalability with the leading edge node logic process (to today's 3 nm node and beyond)[7,8]. However, SRAM is an expensive technology (in terms of the silicon footprint with a relatively low integration density of dozens of megabits per millimetre squared), and also suffers from high stand-by leakage power (tens to hundreds of picowatts per bit)[7,8]. Hence, it is intriguing to explore alternative high-speed memory candidates for the global buffer. Although it is challenging for competing technologies to replace the SRAM technology in the lower-level buffer such as the register file (RF), which may require sub-nanosecond access, opportunities are wide open for the global buffer that is generally slower.

We focus on the following emerging high-speed memory technologies: capacitorless two-transistor (2T) gain cell-based embedded dynamic random-access memory (eDRAM); ferroelectric field effect transistor (FeFET) and ferroelectric random-access memory (FeRAM); and spin-transfer torque (STT) and spin-orbit torque (SOT) magnetic random-access memory (STT-MRAM and SOT-MRAM, respectively). The operation principles, prospects and challenges of each technology will be introduced and discussed in the next section. Benchmarks of these memory candidates and traditional SRAM are then applied for a TPU buffer memory case study. Finally, the state-of-the-art prototype chip demonstrations for some of these memory technologies are surveyed to show the industrial cutting-edge advances and outlooks.

High-speed memory candidates

The criteria to select the memory candidates are writing and reading access speeds (<10 ns) and cycling endurance (>10^{12} cycles)[9] (Supplementary Information section 2). The read access speed is the time to read the stored memory state and the write access speed is the time to write the memory into the desired state. The access speed criteria are assumed according to the last-level cache SRAM speed (~10 ns). The cycling endurance is defined as the number of writings allowed for each memory cell before it becomes unreliable. Assuming 150 training epochs, the total number of write operations for each memory cell is 3.75×10^7 with 50,000 CIFAR-10 training images and 7.5×10^8 with one million ImageNet training images[10]. CIFAR-10 and ImageNet data sets are collections of images commonly used to train machine learning models for image recognition. One TPU chip is able to train a few thousand times on ImageNet in 10 years without any rest, given that 16 TPUv2 chips could finish one training in around 2 h (so one TPUv2 chip needs slightly less than 32 h for one training, which means 2,737.5 trainings in 10 years)[11]. Therefore, we suggest the device endurance criteria (>10^{12} cycles) to support at least thousands of training times during the device lifetime. The training from scratch for challenging tasks (for example, ImageNet) are not on a daily basis even in cloud

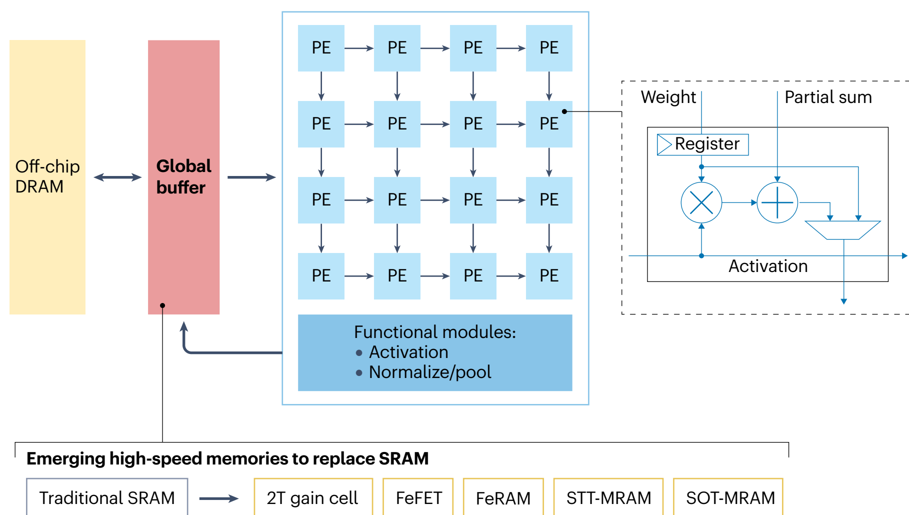


Fig. 1 | Generic diagram of the digital MAC engines used in tensor processing unit (TPU)-like systolic-array architectures. Each processing element (PE) independently computes a multiply and accumulate (MAC) operation of the data received from its upstream neighbour and passes it downstream to minimize the memory access to the global buffer or even to the dynamic random-access memory (DRAM). Weight stationary data flow broadcasts activations and accumulates partial sums spatially across the PEs to maximize the weight data reuse (right-hand side). The data at the beginning or at the end of the stream are fetched from or written back to the global buffer that is typically in the range of 1–100 MB capacity. The global buffer is where the emerging high-speed memories (capacitorless two-transistor (2T) gain cell, ferroelectric field effect transistor (FeFET), ferroelectric random-access memory (FeRAM), spin-transfer torque magnetic random-access memory (STT-MRAM), spin-orbit torque magnetic random-access memory (SOT-MRAM)) discussed in this Review could potentially replace expensive static random-access memory (SRAM).

TPUs, so this endurance criterion is sufficient to sustain the training intensity in most scenarios.

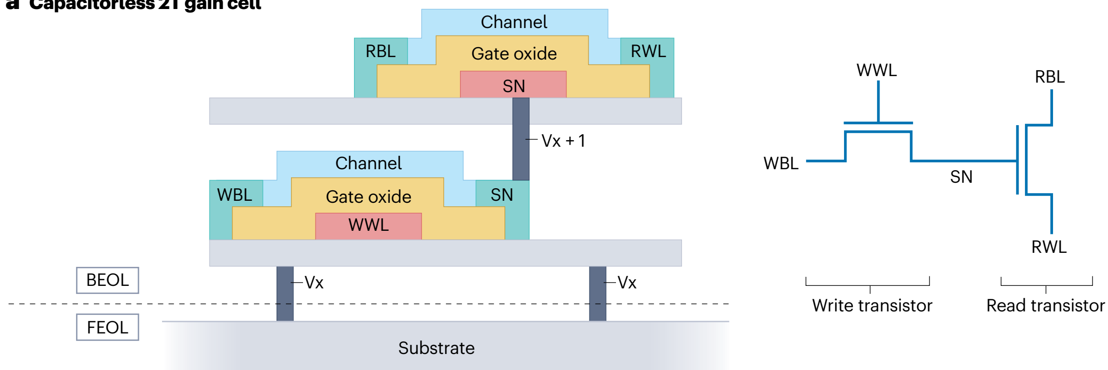
In such a context, some emerging memories such as phase change memory and resistive random-access memory, available on the industrial platforms, do not satisfy these criteria, because of their slow speed (~100 ns), low cycling endurance (10^6–10^9 cycles) and large energy consumption (>1 pJ bit^{-1} for write) (Supplementary Information section 2). Therefore, they are ruled out of this Review. It should be noted that due to variations, reliability and yield issues of emerging memories, an error-correction code scheme must be employed to reliably operate the buffer in the digital domain. Typically, error-correction code encodes original data by adding some redundant parity bits, and then a decoder examines the encoded message to identify and correct errors when reconstructing the original data[12].

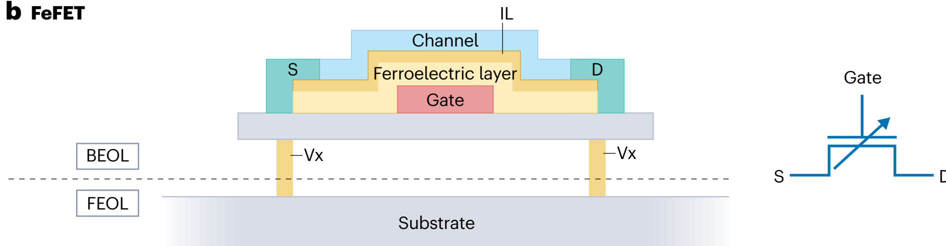
Capacitorless 2T gain cell

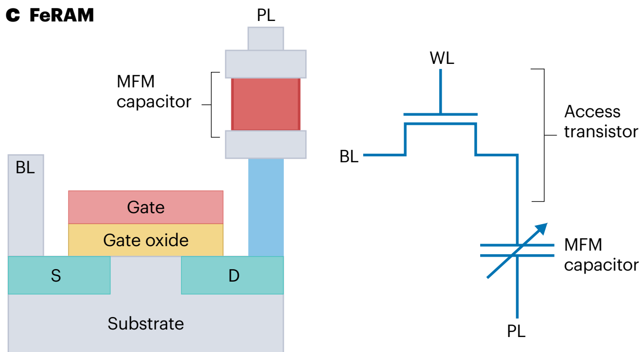
A capacitorless 2T gain cell is essentially eDRAM (Supplementary Information section 1) that could temporarily hold the data. Figure 2a shows the cell architecture in which the drain terminal of the write transistor is connected to the gate terminal of the read transistor. The common node shared by the two transistors acts as a charge storage node (SN) and defines the memory state. The memory state is differentiated by measuring the variation in the drain current of the read transistor when the read bit line (RBL) is enabled. The memory state is written by enabling the write word line (WWL) through the application of the desired voltage (memory state, '1' or '0') to the write bit line (WBL). The operation principle of the 2T gain cell is different from the widely adopted 1T1C (one transistor–one capacitor) stand-alone DRAM or eDRAM (Supplementary Information section 1). Here, the necessity for a large on-chip capacitor is eliminated as the parasitic capacitance formed at the SN replaces the role of the trench or stacked capacitor in the 1T1C DRAM.

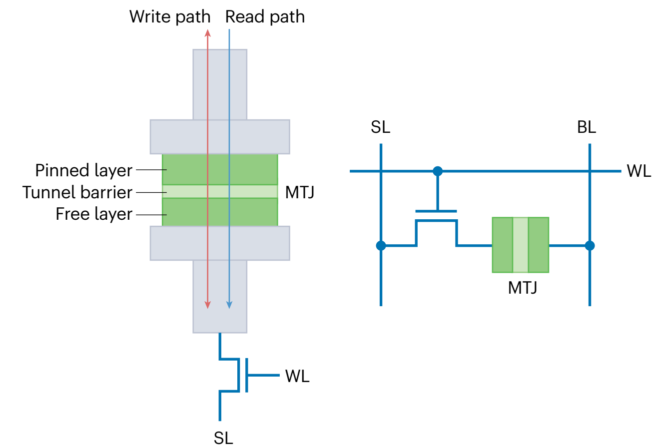
For a 2T gain cell to serve as a practical memory candidate, however, several engineering challenges need to be addressed. First, the parasitic capacitance at the SN should be sufficiently high (10^1–10^3 fF depending on the given operating temperature and transistor leakage current) to provide enough retention (Supplementary Information section 1) time (in the order of a few seconds) for buffer memory[13,14]. The goal is to have a retention time which is longer than the lifetime of the activation data in typical AI workloads; in this way, the explicit refresh operation could be eliminated as the data need to be rewritten before the memory state is lost. Second, considering the high leakage current densities (10^{-4}–10^{-2} μA μm^{-1}) typical of logic transistors[15], achieving a long retention time (in the order of a few seconds) is prohibitive for 2T gain cells. The targeted leakage level should be around 1 fA μm^{-1} to achieve practical ranges of retention time. Third, the charge-injection issue needs to be mitigated during the repeated accesses. The charge-injection problem in 2T gain cells is caused by the capacitive coupling between the WWL and the SN or between the read word line (RWL) and the SN[13]. The voltage transitions happening in the WWL or RWL during read/write processes affect the voltage of the SN, degrading the storage stability.

Different approaches have been proposed to overcome these challenges. For instance, hybrid p-type and n-type transistors designed for 2T gain cell configuration can be employed to mitigate the charge-injection problem[13]. The cryogenic temperature operation, proposed in ref. 13, also reduced the leakage level of the write transistor to below 10 fA μm^{-1} and increased the retention time up to 6.5 s (at 4 K), as measured in the 28 nm prototype chip implemented in a pure logic process. However, for room temperature operation, logic transistors


a Capacitorless 2T gain cell
Channel
RBL
Gate oxide
RWL
SN
Vx + 1
Channel
WBL
Gate oxide
SN
WWL
BEOL
FEOL
Vx
Vx
Substrate
WWL
RBL
WBL
SN
RWL
Write transistor
Read transistor


b FeFET
IL
Channel
S
Ferroelectric layer
D
Gate
BEOL
FEOL
Vx
Vx
Substrate
Gate
S
D


c FeRAM
PL
MFM
capacitor
BL
Gate
Gate oxide
S
D
Substrate
WL
BL
Access
transistor
MFM
capacitor
PL


d STT-MRAM
Write path
Read path
Pinned layer
Tunnel barrier
Free layer
MTJ
WL
SL
SL
BL
WL
MTJ

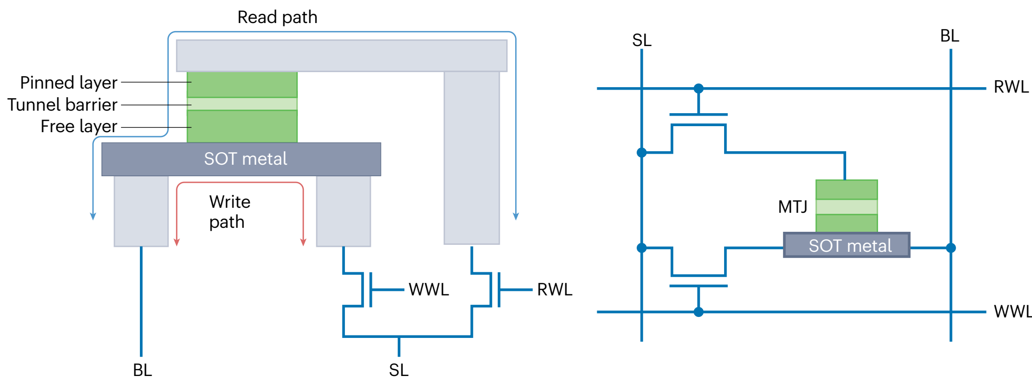

e SOT-MRAM
Read path
Pinned layer
Tunnel barrier
Free layer
SOT metal
Write
path
WWL
RWL
BL
SL
SL
BL
RWL
MTJ
SOT metal
WWL

Fig. 2 | Schematics and circuit diagrams of high-speed memory candidates. **a**, Capacitorless two-transistor (2T) gain cell. **b**, Ferroelectric field effect transistor (FeFET). **c**, Ferroelectric random-access memory (FeRAM). **d**, Spin-transfer torque magnetic random-access memory (STT-MRAM). **e**, Spin-orbit torque magnetic random-access memory (SOT-MRAM). Highlighted are the back end of line (BEOL) integration potential of a 2T gain cell and a FeFET with amorphous oxide semiconductor channel materials. BL, bit line; D, drain; FEOL, front end of line; IL, interfacial layer; MFM, metal–ferroelectric–metal; MTJ, magnetic tunnel junction; PL, plate line; RBL, read bit line; RWL, read word line; S, source; SL, source line; SN, storage node; SOT, spin-orbit torque; WBL, write bit line; WL, word line; WWL, write word line. Vx denotes the *x*th via.

are not the right choice for implementing 2T gain cells because of the high leakage. Alternatively, the use of a 2T gain cell with a different channel material was proposed for the back end of line (BEOL) three-dimensional integration[14] (Supplementary Information section 1). Amorphous oxide semiconductor channel materials, such as the indium oxide (In_2O_3) family, are intensively studied due to their BEOL process compatibility and reasonable field effect mobility[16]. Thanks to its intrinsically wide bandgap (~3 eV), a tungsten-doped In_2O_3 (indium tungsten oxide (IWO))-based BEOL transistor with a short gate length of 20 nm demonstrated a leakage level of approximately 1 fA μm^{-1}; the corresponding 2T gain cell achieved an access time as low as 3 ns (in the fast mode with a write voltage of 2 V) and retention time >1 s at 85 °C[14]. Moreover, BEOL compatibility allows the entire core of the memory array to be placed at the top interconnect level and the peripheral circuits to be hidden underneath. Despite the promises of amorphous oxide semiconductor-based transistors, research is still ongoing to overcome the challenge of threshold voltage instability under voltage stress and elevated temperature and elucidate its origin[17,18]. Through repeated writing and reading in buffer memory, the read and write transistors will experience voltage stress on their gate terminals. Moreover, the power dissipation from the transistors causes the elevation of the operating temperature. These conditions are known to cause a gradual shift of the threshold voltage of the read and write transistors from their original value. This means the key transistor parameters such as gate capacitance, on-current and off-current (Supplementary Information section 1) are also affected by the voltage stress and the elevated temperature. As the SN parasitic capacitance and peripheral circuits (for sensing of the memory state) design parameters are set according to the initial transistor parameters, unwanted threshold voltage shift during buffer operation can have a detrimental effect on the long-term circuit reliability. Thus, it is essential to adopt device and circuit-level techniques to stabilize the threshold voltage of read and write transistors for the implementation of large-scale memory arrays.

FeFET and FeRAM

Advances in ferroelectric memories (Supplementary Information section 1) benefit from the discovery of ferroelectricity in doped hafnium oxide (HfO_2) thin films (<10 nm thickness)[19] in 2011, which is compatible with the current technology for semiconductor manufacturing using atomic-layer deposition. Ferroelectric memories leverage the polarizability of ferroelectric materials to implement memory states. In the FeFET, a ferroelectric layer substitutes the gate dielectric of a conventional transistor[20] (Fig. 2b). The write step is performed by changing the polarization state of the ferroelectric layer through the application of programme or erase pulses through the gate, source or drain terminals, leading to a change in threshold voltage of the channel. The read step is performed by measuring the drain current variation induced by the threshold voltage shift (defined as the memory window). In FeRAM, a metal–ferroelectric–metal (MFM) capacitor is connected to the drain terminal of an access transistor[21] (Fig. 2c). During the read step, the activated word line (WL) turns on the access transistor, and the bit line (BL) current produced after the WL activation depends on the memory state stored in the MFM. State '1' induces a polarization switching current in addition to the discharge current of the state '0'. The BL voltage decays at a different speed depending on the BL current. Thus, by comparing the BL voltage with the reference voltage using a sense amplifier after a certain time after WL activation, two different memory states can be distinguished from each other. Reading of state '1' is a destructive process, as the polarization switching during the reading results in the loss of the original memory state. To compensate for the loss, write-back is typically performed subsequently after the read step.

The main limitation of using FeRAM for buffer memory is caused by the write-back that must be performed after each read step. Due to the write-back, cycling endurance is consumed by both writing and reading operations in FeRAM. State-of-the-art FeRAM has benefited from advanced material and device engineering techniques to achieve high cycling endurance of ~10^{12} cycles[21]. Alternatively, dual-mode operation of FeRAM has been proposed to reduce the frequency of destructive read-out during buffer memory operation[10]. During the dual-mode operation, the frequently updated data are stored in a volatile eDRAM-like mode, whereas the data with longer lifetime are stored in a non-volatile FeRAM mode[10]. As the write-back to recover the original polarization state is not needed during the eDRAM-like mode, the overall cycling endurance consumed by the FeRAM buffer memory could be reduced utilizing dual-mode operation.

The FeFET does not suffer from destructive reading processes, but faces other challenges such as endurance degradation. Typical FeFET endurance is around 10^5–10^6 cycles because of the charge traps generated during cycling with high electric fields across the interfacial layer (IL) between the silicon channel and the ferroelectric film[22]. Several solutions have been proposed in the literature. One promising approach is to use a back-gated BEOL FeFET configuration with a channel-last process in which the channel deposition is performed after the bottom gate and ferroelectric layer deposition[23]. The low thermal budget for the channel material and engineering the interface between the gate oxide and the channel enabled a reduction of the IL thickness and, consequently, of the charge trapping effect. In this way, a writing time of 10 ns and cycling endurance of 10^{12} cycles were obtained. In a similar work, IWO films were explored as the channel material for BEOL-compatible FeFETs[24]. Using IWO, the IL between the channel and the gate oxide was completely removed, resulting in a write voltage of 1.6 V and a cycling endurance exceeding 10^{11} cycles. A ferroelectric metal field effect transistor (FeMFET) structure was also proposed to alleviate the IL-related issues[25,26]. By inserting a floating electrode to separate the ferroelectric layer and the metal–oxide–semiconductor (MOS) layer, the ratio between the two areas (A_{FE}/A_{MOS}, where A_{FE} is the MFM capacitor area and A_{MOS} is the MOSFET gate oxide area) could be flexibly adjusted[25]. The smaller area ratio reduces the electric field applied across the IL, decreasing the charge trapping in the ferroelectric layer. The scalability of the FeMFET to advanced nodes (for example, 3 nm) was projected[26], indicating the possibility of further lowering the write voltage with technology scaling.

STT-MRAM and SOT-MRAM

MRAM utilizes the tunnelling magnetoresistance effect in a magnetic tunnel junction (MTJ) to store binary data[27] (Supplementary Information section 1). The MTJ, the fundamental component of magnetic random-access memory (MRAM), consists of a dielectric tunnel barrier sandwiched between two ferromagnetic layers. The resistance of the MTJ depends on the magnetization direction of these two layers: when the magnetization directions are parallel, the MTJ exhibits a low resistance; antiparallel magnetization results in high resistances. In a read operation, the read voltage is applied to both ends of the MTJ cell and a current is measured regardless of the type of MRAM employed (STT-MRAM (Fig. 2d) or SOT-MRAM (Fig. 2e)). The reading current of the MRAM depends on the two resistance states of the MTJ cell. The write operation depends on the physical mechanism behind the STT switching (Fig. 2d) and SOT switching (Fig. 2e). STT-MRAM utilizes the current flow directly through the MTJ stack to reverse the magnetization of the free layer[28]. When electrons flow from the pinned layer to the free layer, those with the same magnetic moment as the pinned layer apply a STT to the free layer, resulting in parallel magnetization and low resistance in the MTJ. Conversely, when electrons flow from the free layer to the pinned layer, those with a parallel spin moment to the pinned layer pass through, whereas others are reflected back to the free layer. These electrons change the magnetization of the free layer to the antiparallel direction, leading to the high resistance in the MTJ. Sharing the read and the write paths in STT-MRAM may degrade the reliability of the MTJ.

SOT-MRAM was introduced to alleviate this issue. In SOT-MRAM, the MTJ cell is positioned above the in-plane electrodes for write operations[29]. The write current flow along the in-plane electrode induces a spin torque through the spin-Hall effect, a phenomenon in which spin currents are generated in the direction perpendicular to the charge current by spin-orbit coupling, and Rashba spin-orbit coupling, interaction between the electron spin of an electron and its orbital motion on a solid surface, altering the magnetization of the free layer. Unlike STT-MRAM, SOT-MRAM isolates the reading from the writing paths, thus improving the endurance and reading reliability, and offering other advantages, such as sub-nanosecond operation speeds. However, the additional metal line and another access transistor required for the writing operation in SOT-MRAM introduce more area penalty.

The industry's initial application target for MRAM is to replace embedded Flash (eFlash) in the microcontroller[30–33] (Supplementary Information section 1). Therefore, the MTJ stack is engineered towards 10-year data retention with sacrificed cycling endurance (~10^6 cycles) and moderate write (around 200 ns) and read (around 50 ns) speeds[33–35]. Furthermore, it is possible to tailor the MTJ stack for the high-speed buffer memory (that is, the last-level cache) on the same MTJ platform as used in the eFlash replacement[36]. However, the major roadblock for the employment of MRAM for buffer memory is the relatively high-power consumption in the write operation. The typical write current density for STT-MRAM is between 4 MA cm^{-2} and 20 MA cm^{-2} (normalized to the MTJ area), and for SOT-MRAMs is between 70 MA cm^{-2} and 150 MA cm^{-2} (normalized to the metal line cross-sectional area)[28,37,38]. Such large current densities require the use of over-sized access transistor, which diminishes the area advantages given by the MRAM architecture, leading to a factor of merely two times or three times improvements over 6T high-density SRAM at the same node. A prior study suggested that the writing current density needs to be reduced three or five times for MRAM to become competitive for buffer memory in terms of the energy efficiency and computing density[37].

TPU buffer memory case study and NeuroSim benchmarking

In this section, we present a quantitative evaluation of the pros and cons of the different high-speed memory candidates to be used as the global buffer memory for TPU-like architectures. Supplementary Information Fig. S1 shows the integrated framework for the benchmark used in this case study. To facilitate the benchmarking, we utilized the open-source NeuroSim simulator[39,40] that captures the recent technological advancements in the emerging non-volatile memories (eNVMs) such as phase change memory, resistive random-access memory, MRAM and FeFET (for a detailed description of the benchmarking process, see Supplementary Information section 2).

A few observations could be made from the benchmarking analysis shown in Fig. 3. The energy and area of different memory technologies are normalized using SRAM as the global buffer baseline. As shown in Fig. 3a, for the 3 nm high-performance cloud server where the compute activity is high – that is, with 100% of the inference queries constantly fed in – SRAM global buffer is still a competitive candidate for high energy efficiency. Thanks to the continued scaling of the SRAM bit cell area (Supplementary Information section 1) down to 0.021 μm^2 at 3 nm node[8], a large capacity of SRAM (up to 32 MB[8]) is an available product from the foundries. The 2T gain cell is also a good candidate with energy efficiency comparable with SRAM. As SRAM and the 2T gain cell are charge-based memories, the energy to move the charge is lower than that in the eNVM counterpart where the ferroelectric or magnetic switching is associated. In Fig. 3a, the global buffer refresh, read, leakage and functional module consume no more than 2% energy in total.

If the expensive silicon area caused by SRAM is a concern, emerging technologies such as the BEOL stacked 2T gain cell, BEOL stacked FeFET, STT-MRAM and SOT-MRAM could lower the global buffer area by 59%, 58%, 54% and 20%, respectively, as shown in Fig. 3b. As a trade-off, the energy efficiency improves and degrades by +0.3%, −13%, −28% and −27%, respectively. Similar trends are found in the 22 nm edge TPU case for energies and chip areas, as shown in Fig. 3c,d. The impact of global buffer technology choice is more significant in the edge TPU because of the reduced compute resources in a smaller number of PEs. Overall, the BEOL stacked 2T gain cell appears to be a promising candidate. The functional module shown in Fig. 3b occupies less than 1% area for all conditions. As shown in Fig. 3c, global buffer refresh and leakage consume no more than 1% energy.

As shown in Fig. 3e, for the low-power edge device, the advantages of emerging technologies-based solutions manifest when the compute activity is low (<1%). When the inference chip is in the stand-by frequent scenario, SRAM global buffer contributes to substantial leakage, whereas the non-volatile memories could power-gate off to save the stand-by leakage. For edge devices, most of the alternative candidates could outperform SRAM-based designs. For a cloud device, the system is expected to be rarely in stand-by and the portion of computing time is close to 100%, so its SRAM-based design is still competitive.

Array-level prototyping landscape

In the past decade, industry has invested heavily in emerging memories, and many of these efforts could have a profound impact when being tailored towards high-speed buffer memory engineering. Here, we have surveyed the state-of-the-art prototype chip demonstrations for some

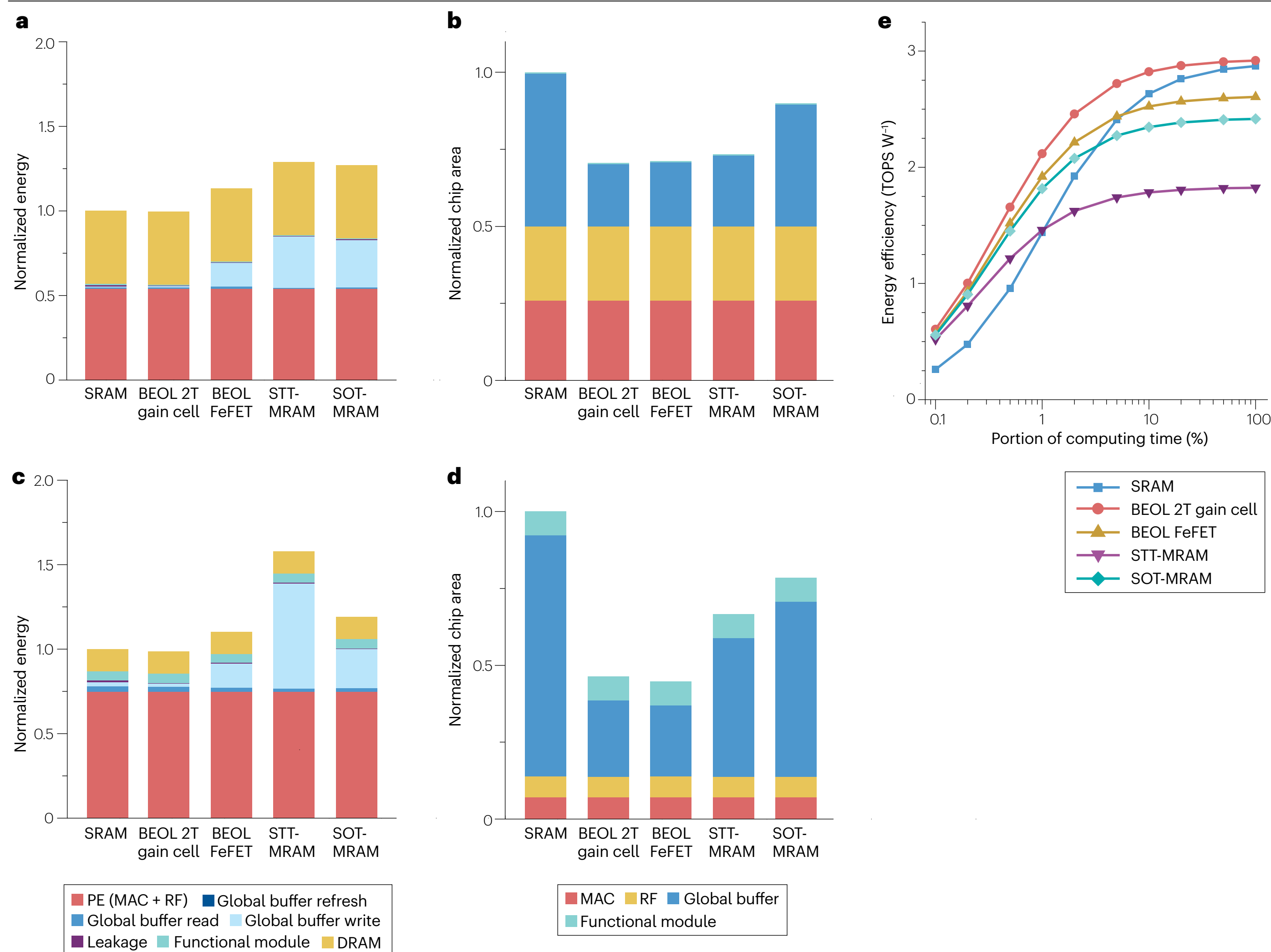


Fig. 3 | Benchmarking analysis for high-speed memory candidates on cloud and edge TPUs. a–d, Benchmarking analysis results for energy breakdown (**a**) and area breakdown (**b**) of a cloud tensor processing unit (TPU), and for energy breakdown (**c**) and area breakdown (**d**) of an edge TPU. Energy levels and areas are normalized using static random-access memory (SRAM) as the global buffer baseline. **e**, Energy efficiency of the edge TPU with respect to a portion of activated computing time. Emerging memories outperform SRAM when the compute activity is low (<1%). Note that the functional module includes activation, batch normalization and pooling units. BEOL, back end of line; DRAM, dynamic random-access memory; FeFET, ferroelectric field effect transistor; MAC, multiply and accumulate; PE, processing element; RF, register file; SOT-MRAM, spin-orbit torque magnetic random-access memory; STT-MRAM, spin-transfer torque magnetic random-access memory; 2T gain cell, two-transistor gain cell.

of the technologies even though their initial application target might be different. Because the amorphous semiconductor oxide-based 2T gain cells are still in their early research stage, there are rarely any array-level demonstrations. Therefore, in the following we focus on the ferroelectric and magnetic memories where prototype chips have been reported.

FeFET and FeRAM prototype chips

Table 1 presents a comprehensive overview of recently reported prototype chips featuring FeFET and FeRAM architectures. The first FeFET macro (a functional memory block) in 28 nm node was launched in 2016 by GlobalFoundries[41] in collaboration with German research organizations (NaMLab, Fraunhofer and so on), which were the first to report doped HfO_2-based ferroelectric films in 2011 (ref. 19). Since then, although various FeFET devices continue to be actively optimized, GlobalFoundries has emerged as a forerunner towards risk manufacturing with the design house partner FMC[42]. It should be noted that GlobalFoundries' effort has focused on the development of a front end of line (FEOL) FeFET (Supplementary Information section 1). The overall research trend for FeFET macros centres around enhancing reliability, including addressing limited endurance and device variations, while achieving large capacity to ensure commercial viability. Unfortunately, most of the efforts here are towards eNVM applications rather than to buffer memory. As discussed earlier, the BEOL FeFET holds more potential towards high-speed memory applications with density advantages.

Table 1 | Overview of recently reported FeFET and FeRAM prototype chips

Technical node	FMC/GF FeFET VLSI 2021 (ref. 42)	Sony FeRAM VLSI 2020 (ref. 43)	Leti FeRAM IEDM 2019 (ref. 44)	Leti FeRAM IEDM 2021 (ref. 45)	ITRI FeRAM IEDM 2022 (ref. 46)	CAS FeRAM ISSCC 2023 (ref. 47)
Target application	eNVM	eNVM	eNVM	eNVM	eNVM	eNVM
Technology node (nm)	28	130	130	130	NA	130
Capacity	32 MB	64 kB	16 kB	16 kB	4 kB	9 MB
Macro size (mm^2)	4.77	NA	NA	NA	NA	NA
Cell size (μm^2)	0.076	0.4–1.0 (cap area)	0.28–1.13 (cap area)	0.16 (cap area)	0.36 (cap area)	NA
$2P_r$ ($\mu C cm^{-2}$)	NA	NA	>40	NA	NA	NA
Memory window (V)	1.53	NA	NA	NA	NA	NA
Operating voltage (V)	4	2.5	<4	2.5	2.6	<3
Programme/erase time (ns)	$10^4/10^5$	1.4	<100	4 at 4.8 V	40	7 at 3.3 V/20 at 2.5 V
Read pulse width (ns)	25	8	NA	NA	60	5 at 3.3 V/30 at 2.5 V
Write endurance (cycles)	$>10^5$	$>10^{11}$	$>10^{11}$	$>10^7$	$>10^{12}$	$>10^{12}$
Retention	NA	100 min at 85 °C	1,000 min at 125 °C	10^4 s at 125 °C	5×10^4 s at 120 °C	10 years at 85 °C

eNVM, emerging non-volatile memory; FeFET, ferroelectric field effect transistor; FeRAM, ferroelectric random-access memory; NA, not available; Pr, remnant polarization.

On the FeRAM front, Sony reported the development of a 64 kbit FeRAM macro at 130 nm node comprising sub-500 °C TiN/$HfZrO_x$ (HZO)/TiN stack MFM capacitors[43]. At the array level, this FeRAM macro exhibited 100% bit functionality, an operating voltage as low as 2.5 V, a write latency of 14 ns and an endurance exceeding 10^{11} cycles. CEA-Leti reported similar FeRAM macros based on HZO[44] and silicon-doped HfO_2 (ref. 45) stacks in 130 nm node. ITRI's recent work focused on enhancing the reliability of FeRAMs[46] and proposed the use of TiON as a barrier metal in the electrodes of ferroelectric capacitors to mitigate capacitor fatigue. Additionally, post-metal annealing was found to further suppress fatigue. The reported 4 kb 1T1C FeRAM macro exhibited a high yield (>98%) and demonstrated wake-up free characteristics, achieving an endurance surpassing 10^{12} cycles. CAS researchers presented a 9 Mb HZO-based non-volatile FeRAM macro and introduced circuit designs aimed at improving chip performance for mass production[47]. The design incorporated a temperature-aware write-voltage driver with error-correction code-assisted refresh to enhance endurance and reduce bit error rates. Furthermore, an offset-cancelled sense amplifier was introduced to improve the sensing margin. As discussed earlier, the read-destructive nature of FeRAM is the major roadblock for its application towards the buffer memories.

STT-MRAM and SOT-MRAM prototype chips

Table 2 presents an overview of recently reported STT-MRAM and SOT-MRAM macros. STT-MRAM macros have been under development at advanced technology nodes (28 nm or below) in major foundries (TSMC, GlobalFoundries, Samsung and Intel) since 2019, whereas SOT-MRAM macros have mainly been reported by IMEC, with a recent report from TSMC/ITRI. The general research trend involves engineering STT-MRAM and SOT-MRAM macros with various specifications, such as long retention (>10 years), high temperature reflow compatibility (230–260 °C)[48–50], magnetic shielding capability (0.5–4,000 Oe)[34,48,50] and high speed (<50 ns), to cater to industrial applications as well as automotive and aerospace applications.

Sony presented the first report of STT-MRAM macros as buffer memories for the CMOS image sensor[51]. The study demonstrated that smaller MTJ cells could be employed as buffer memory, whereas larger MTJ cells could be utilized for eNVM applications due to their increased data retention capabilities. Furthermore, the integration of both buffer memory and eNVM-type MTJ cells on a single chip was achieved through MTJ size modulation. However, most efforts for STT-MRAM development are targeting eNVM applications such as eFlash replacement for microcontrollers. Renesas reported STT-MRAM macros operated at high temperatures (up to 150 °C) for automotive applications[52]. Unlike consumer semiconductors characterized up to 85 °C, automotive semiconductors require guaranteed performance at high temperatures of 125–150 °C[53,54]. This work achieved random read access times of 5.1 ns and 5.9 ns at 125 °C and 150 °C, respectively. Additionally, proposed write algorithms reduced the number of write pulses (Supplementary Information section 1) through optimized bit line voltage, resulting in improved write throughput.

Samsung reported STT-MRAM macros designed for energy-efficient stand-alone memories[55]. Fabricated on 28 nm and 14 nm technology nodes, the macros demonstrated almost unlimited endurance ($>10^{14}$ cycles) with a write power of 27 mW and a read power of 14 mW. To overcome the cell size limitations of conventional 1T1M (one transistor–one MTJ) structures, SK Hynix and Kioxia presented a 1S1M (one selector–one MTJ) cell design for high-density memory applications[56]. By utilizing optimized arsenic-doped SiO_2 as the selector material, the 1S1R cell achieved a footprint of 4F (ref. 2), a cell pitch of 45 nm and an MTJ size of 20 nm.

Everspin proposed an STT-MRAM macro focused on industrial applications[57]. The macro exhibited a high 400 MB s^{-1} writing throughput across the industrial temperature range of −40 °C to +80 °C, while maintaining compliant data retention and endurance characteristics. TSMC presented an STT-MRAM macro in the 16 nm FinFET platform tailored for automotive applications[58]. The work introduced several chip-level algorithms and schemes, including sensing schemes to enhance small read margins and mitigate external magnetic field interference.

SOT-MRAM macros have predominantly been reported for cache[59,60] and CIM weight cell applications[61,62]. In IMEC's work,

solutions for high-density and high-performance SOT-MRAM macros were provided through design technology co-optimization[59]. PPA analysis at the 5 nm node identified SOT efficiency and BL resistance as the two main parameters limiting SOT-MRAM macro performance. TSMC/ITRI reported an 8 kb SOT-MRAM macro with the fast-switching speed (around 1 ns)[60]. The study addressed challenges hindering mass production of SOT-MRAM, such as identifying a high spin-Hall conductivity SOT channel material capable of withstanding 400 °C post-annealing, etch stop in the SOT channel and ensuring non-contaminated MTJ sidewalls. The issues were resolved through the application of a unique tungsten-based SOT channel material and optimized etching processes. IMEC presented SOT-MRAM macros for CIM applications[61,62]. As parallel operations are performed in mixed-signal and analogue MAC computations, weight memory cells (Supplementary Information section 1) require high cell resistance (greater than megaohms) and low resistance variability. In IMEC's initial design[61], ternary weights for DNN inference are experimentally validated using SOT-MRAM macros. Differential pairs of MRAM cells are employed to implement ternary weights, and cell variations depending on MTJ cell dimensions are characterized through a fast write speed of 0.5 ns. DNNs for MNIST and CIFAR-100 data sets were tested with design technology co-optimization, and target values of weight variation were proposed to achieve acceptable error rates. Based on the initial design, IMEC introduced perpendicular-SOT devices with a new free layer design[62]. The new free layer design exhibited improved data retention and endurance characteristics at 125 °C, even after undergoing 400 °C post-metal annealing. It should be noted that the engineering directions for cache and CIM weight memory cell are quite diverged, and the buffer memory requires similar techniques as for the cache applications. Reduction of the switching current density is necessary before the SOT-MRAM becomes competitive against the SRAM-based baselines.

Table 2 | Overview of recently reported STT-MRAM and SOT-MRAM prototype chips

Technical node	Sony STT VLSI 2021 (ref. 51)	Renesas STT VLSI 2022 (ref. 52)	Samsung STT IEDM 2022 (ref. 55)	SK Hynix/ Kioxia STT IEDM 2022 (ref. 56)	Everspin STT IEDM 2022 (ref. 57)	TSMC STT ISSCC 2023 (ref. 58)	IMEC SOT IEDM 2020 (ref. 59)	TSMC/ ITRI SOT VLSI 2022 (ref. 60)	IMEC SOT VLSI 2020 (ref. 61)	IMEC SOT VLSI 2021 (ref. 62)
Target application	Buffer, eNVM	eNVM	Stand-alone	Stand-alone	eNVM	eNVM	Cache	Cache	CIM weight cell	CIM weight cell
Technology node (nm)	40	22	28/14	45	28	16	5 (simulation)	NA	22	NA
R_{ON}/R_{OFF} (Ω)	NA	NA	NA	NA	NA	NA	4,970/NA	900/2,100	6 million/ 15 million	55,000/114,000
On–off ratio	NA	NA	NA	NA	NA	NA	NA	2.3	2.5	2.1
MTJ size (nm^2)	NA	NA	NA	20×20	NA	NA	D=32 (circle)	75×230	D=80 (circle)	D=80 (circle)
Resistance–area product (Ω μm^2)	NA	NA	NA	NA	NA	NA	4	10	5,000–50,000	2,000
Capacity	30 MB	32 MB	16 MB	4 GB	64 MB	32 MB	NA	8 kB	NA	NA
Macro size (mm^2)	NA	NA	30	NA	NA	2.5	NA	NA	NA	NA
Cell size (F^2)	NA	NA	NA	4 (1S1M)	NA	NA	540 (high density)	NA	78	NA
Cell size (μm^2)	0.061	0.0456	0.0242	0.002025	NA	0.033	0.0135	NA	0.038	1.18
Write voltage (V)	0.89–1.16	NA	1.8, 3.3	NA	1.65–2	0.72–0.88	0.9	0.8	<1	NA
Write current density (MA cm^2)	NA	NA	NA	NA	NA	NA	NA	68	23	NA
Write pulse width (ns)	26 at 30 °C	NA	100/50	30–200	NA	NA	1.4	1	0.5	1
Read pulse width (ns)	NA	5.0–5.9	40/15	>30	NA	<6	0.9	NA	NA	NA
Write endurance (cycles)	>10^{10} at 105 °C	NA	>10^{14} at 25 °C	>10^6	NA	>10^6	NA	7×10^{12}	NA	NA
Retention	>1 s at 85 °C, 45.1 years at 85 °C	NA	10 years at 89 °C	10 years at 90 °C	10 years at 105 °C	20 years at 150 °C	NA	24 h at 160 °C	NA	NA

CIM, compute-in-memory; eNVM, emerging non-volatile memory; MTJ, magnetic tunnel junction; NA, not available; 1S1M, one selector–one MTJ; SOT, spin-orbit torque; SOT-MRAM, spin-orbit torque magnetic random-access memory; STT, spin-transfer torque; STT-MRAM, spin-transfer torque magnetic random-access memory.

Outlook

In this Review, a guide for future research efforts dedicated to the development of novel buffer memories for AI hardware accelerators, which may have ground-breaking impact on many cloud and edge AI applications, is reported and discussed. In particular, we survey those emerging devices that could replace the global buffer in AI hardware, traditionally based on SRAM. The memories are required to have fast access speed (<10 ns) and high endurance (>10^{12} cycles), and should be fabricated on BEOL or scaled down with logic technology.

Among the reported candidates, the continued industrial investment in ferroelectric (FeFETs and FeRAMs) and magnetic (STT-MRAMs and SOT-MRAMs) memories creates opportunities to tailor their specifications in terms of cycling endurance. Meanwhile, with the technological advancements in BEOL-compatible amorphous oxide semiconductor channel materials, the 2T gain cell is another promising candidate that would allow high data retention time.

These memory devices are suitable for AI hardware at the edge where stand-by scenarios benefit from low leakage. However, their high write energy should be improved to be more competitive for AI hardware in the cloud. Beyond the benchmarking and analysis presented here, the community is encouraged to perform the silicon implementation and prototyping of the proposed AI hardware designs with the emerging technologies.

Published online: 11 January 2024

References

1. Jouppi, N. et al. TPU v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Proc. 50th Annual Int. Symp. Computer Architecture* (ed. Solihin, Y.) 1–14 (Association for Computing Machinery, 2023).
2. Cass, S. Taking AI to the edge: Google's TPU now comes in a maker-friendly package. *IEEE Spectr.* **56**, 16–17 (2019).
3. Deng, L., Li, G., Han, S., Shi, L. & Xie, Y. Model compression and hardware acceleration for neural networks: a comprehensive survey. *Proc. IEEE* **108**, 485–532 (2020). **This work provides a survey of hardware acceleration for neural networks**.
4. Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. How to evaluate deep neural network processors: TOPS/W (alone) considered harmful. *IEEE Solid-State Circuits Mag.* **12**, 28–41 (2020).
5. Zhang, W. et al. Neuro-inspired computing chips. *Nat. Electron.* **3**, 371–382 (2020).
6. 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.* **21**, 31–56 (2021). **This work provides recent trends in weight memory in CIM engines as background for this Review**.
7. Aoyagi, Y. et al. A 3-nm 27.6-Mbit/mm2 self-timed SRAM enabling 0.48 - 1.2 V wide operating range with far-end pre-charge and weak-bit tracking. In *IEEE Symp. VLSI Technology and Circuits* (eds Miyashita, K. & Oike, Y.) (IEEE, 2023).
8. Chang, J. et al. A 3nm 256Mb SRAM in FinFET technology with new array banking architecture and write-assist circuitry scheme for high-density and low-VMIN applications. In *IEEE Symp. VLSI Technology and Circuits* (eds Miyashita, K. & Oike, Y.) (IEEE, 2023).
9. Yu, S. *Semiconductor Memory Devices and Circuits* 1-4 (CRC, 2022). **This work provides criteria for high-speed memory candidates for global buffer memory**.
10. Luo, Y., Luo, Y.-C. & Yu, S. A ferroelectric-based volatile/non-volatile dual-mode buffer memory for deep neural network accelerators. *IEEE Trans. Comput.* **71**, 2088–2101 (2021).
11. Coleman, C. A. et al. DAWNBench: An end-to-end deep learning benchmark and competition. In *Conference on Neural Information Processing Systems, Machine Learning for Systems Workshop* (eds Guyon, I. & von Luxburg, U.) (NeurIPS, 2017).
12. Cai, Y., Ghose, S., Haratsch, E. F., Luo, Y. & Mutlu, O. Error characterization, mitigation, and recovery in flash-memory-based solid-state drives. *Proc. IEEE.* **105**, 1666–1704 (2017).
13. Saligram, R., Datta, S. & Raychowdhury, A. CryoMem: a 4K-300K 1.3GHz eDRAM macro with hybrid 2T-gain-cell in a 28nm logic process for cryogenic applications. In *IEEE Custom Integrated Circuits Conf. (CICC)* (ed. Raychowdhury, A.) (IEEE, 2021).
14. Ye, H. et al. Double-gate W-doped amorphous indium oxide transistors for monolithic 3D capacitorless gain cell eDRAM. In *2020 IEEE Int. Electron Devices Meeting (IEDM)* (ed. Datta, S.) 613–614 (IEEE, 2020).
15. International Roadmap of Devices and Systems 2022 Edition, More Moore; https://irds.ieee.org/editions/2022/more-moore (accessed 24 November 2022).
16. On, N. et al. Boosting carrier mobility and stability in indium–zinc–tin oxide thin-film transistors through controlled crystallization. *Sci. Rep.* **10**, 18868 (2020).
17. Hu, Y., Chakraborty, W., Ye, H., Datta, S. & Cho, K. First-principles investigation of amorphous n-type In2 O3 for BEOL transistor. In *Int. Conf. Simulation of Semiconductor Processes and Devices (SISPAD)* (ed. Vandenberghe, W.) 116–119 (IEEE, 2021).
18. Shiah, Y.-S. et al. Mobility–stability trade-off in oxide thin-film transistors. *Nat. Electron.* **4**, 800–807 (2021).
19. Böscke, T., Müller, J., Bräuhaus, D., Schröder, U. & Böttger, U. Ferroelectricity in hafnium oxide thin films. *Appl. Phys. Lett.* **99**, 102903 (2011).
20. Mulaosmanovic, H. et al. Ferroelectric field-effect transistors based on HfO_2: a review. *Nanotechnology* **32**, 502002 (2021).
21. Haratipour, N. et al. Hafnia-based FeRAM: a path toward ultra-high density for next-generation high-speed embedded memory. In *International Electron Devices Meeting (IEDM)* (ed. De Salvo, B.) 138–141 (IEEE, 2022).
22. Salahuddin, S., Ni, K. & Datta, S. The era of hyper-scaling in electronics. *Nat. Electron.* **1**, 442–450 (2018).
23. Sharma, A. A. et al. High speed memory operation in channel-last, back-gated ferroelectric transistors. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Datta, S.) 391–394 (IEEE, 2020).
24. Dutta, S. et al. Logic compatible high-performance ferroelectric transistor memory. *IEEE Electron. Device Lett.* **43**, 382–385 (2022).
25. Ni, K. et al. SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Rim, K.) 296–299 (IEEE, 2018).
26. Choe, G. & Yu, S. Multigate ferroelectric transistor design toward 3-nm technology node. *IEEE Trans. Electron. Devices* **68**, 5908–5911 (2021).
27. Baibich, M. N. et al. Giant magnetoresistance of (001)Fe/(001)Cr magnetic superlattices. *Phys. Rev. Lett.* **61**, 2472 (1988).
28. Apalkov, D. et al. Spin-transfer torque magnetic random access memory (STT-MRAM). *ACM J. Emerg. Technol. Comput. Syst.* **9**, 1–35 (2013).
29. Shao, Q. et al. Roadmap of spin-orbit torques. *IEEE Trans. Magn.* **57**, 1–39 (2021).
30. Shum, D. et al. CMOS-embedded STT-MRAM arrays in 2x nm nodes for GP-MCU applications. In *2017 Symp. VLSI Technology* (ed. Inaba, S.) T208–T209 (IEEE, 2017).
31. Lee, K. & Kang, S. H. Development of embedded STT-MRAM for mobile system-on-chips. *IEEE Trans. Magn.* **47**, 131–136 (2010).
32. Lee, K. et al. 22-nm FD-SOI embedded MRAM with full solder reflow compatibility and enhanced magnetic immunity. In *IEEE Symp. VLSI Technology* (ed. Khare, M.) 183–184 (IEEE, 2018).
33. Antonyan, A., Pyo, S., Jung, H. & Song, T. Embedded MRAM macro for eFlash replacement. In *2018 IEEE Int. Symp. Circuits and Systems (ISCAS)* (eds Maloberti, F. & Setti, G.) (IEEE, 2018).
34. Naik, V. et al. JEDEC-qualified highly reliable 22nm FD-SOI embedded MRAM for low-power industrial-grade, and extended performance towards automotive-grade-1 applications. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Datta, S.) 219–222 (IEEE, 2020).
35. Naik, V. et al. Manufacturable 22nm FD-SOI embedded MRAM technology for industrial-grade MCU and IOT applications. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Takayanagi, M.) 26–29 (IEEE, 2019).
36. Alzate, J. et al. 2 MB array-level demonstration of STT-MRAM process and performance towards L4 Cache applications. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Takayanagi, M.) 30–33 (IEEE, 2019).
37. Luo, Y. et al. Performance benchmarking of spin-orbit torque magnetic RAM (SOT-MRAM) for deep neural network (DNN) accelerators. In *2022 IEEE Int. Memory Workshop (IMW)* (ed. Wouters, D.) (IEEE, 2022).
38. Garello, K. et al. SOT-MRAM 300mm integration for low power and ultrafast embedded memories. In *2018 IEEE Symp. VLSI Circuits* (ed. Lehmann, G.) 81–82 (IEEE, 2018).
39. Peng, X., Huang, S., Luo, Y., Sun, X. & Yu, S. DNN+NeuroSim: an end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Takayanagi, M.) 771–774 (IEEE, 2019).
40. Lu, A., Peng, X., Li, W., Jiang, H. & Yu, S. NeuroSim simulator for compute-in-memory hardware accelerator: validation and benchmark. *Front. Artif. Intell.* **4**, 659060 (2021).
41. Trentzsch, M. et al. A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Fay, P.) 294–297 (IEEE, 2016).
42. Müller, S. et al. Development status of gate-first FeFET technology. In *2021 Symp. VLSI Technology* (ed. Yamakawa, S.) TFS1–5 (IEEE, 2021).
43. Okuno, J. et al. SoC compatible 1T1C FeRAM memory array based on ferroelectric Hf0.5Zr0.5O2. In *IEEE Symp. VLSI Technology* (eds Chang, C.-P. & Chang, K.) TF2.1 (IEEE, 2020).
44. Francois, T. et al. Demonstration of BEOL-compatible ferroelectric Hf0.5Zr0.5O2 scaled FeRAM co-integrated with 130nm CMOS for embedded NVM applications. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Takayanagi, M.) 362–365 (IEEE, 2019).
45. Francois, T. et al. 16kbit HfO2:Si-based 1T-1C FeRAM arrays demonstrating high performance operation and solder reflow compatibility. In *IEEE Int. Electron Devices Meeting (IEDM)* (ed. Grasser, T.) 697–700 (IEEE, 2021).
46. Lin, Y.-D. et al. Highly reliable, scalable, and high-yield HfZrOx FRAM by barrier layer engineering and post-metal annealing. In *Int. Electron Devices Meeting (IEDM)* (ed. De Salvo, B.) 747–750 (IEEE, 2022).

47. Yang, J. et al. A 9Mb HZO-based embedded FeRAM with 10^{12}-cycle endurance and 5/7ns read/write using ECC-assisted data refresh and offset-canceled sense amplifier. In *2023 IEEE Int. Solid-State Circuits Conference (ISSCC)* (ed. Cantatore, E.) 498–500 (IEEE, 2023).
48. Wang, C.-Y. et al. Reliability demonstration of reflow qualified 22nm STT-MRAM for embedded memory applications. In *2020 IEEE Symp. on VLSI Technology* (eds Chang, C.-P. & Chang, K.) TM3.2 (IEEE, 2020).
49. Lee, K. et al. 28nm CIS-compatible embedded STT-MRAM for frame buffer memory. In *2021 IEEE Int. Electron Devices Meeting (IEDM)* (ed. Grasser, T.) 27–30 (IEEE, 2021).
50. Chen, C.-H. et al. Reliability and magnetic immunity of reflow-capable embedded STT-MRAM in 16nm FinFET CMOS process. In *2021 Symp. VLSI Technology* (ed. Yamakawa, S.) T12–1 (IEEE, 2021).
51. Oka, M. et al. 3D stacked CIS compatible 40nm embedded STT-MRAM for buffer memory. In *2021 Symp. VLSI Technology* (ed. Yamakawa, S.) T2–5 (IEEE, 2021).
52. Shimoi, T. et al. A 22nm 32Mb embedded STT-MRAM macro achieving 5.9ns random read access and 5.8MB/s write throughput at up to Tj of 150 °C. In *IEEE Symp. VLSI Technology and Circuits* (eds. Palacios, T. & Ginsburg, B.) 134–135 (IEEE, 2022).
53. Johnson, R. W., Evans, J. L., Jacobsen, P., Thompson, J. R. & Christopher, M. The changing automotive environment: high-temperature electronics. *IEEE Trans. Electron. Packag. Manuf.* **27**, 164–176 (2004).
54. Watson, J. & Castro, G. A review of high-temperature electronics technology and applications. *J. Mater. Sci. Mater. Electron.* **26**, 9226–9235 (2015).
55. Lee, T. et al. World-most energy-efficient MRAM technology for non-volatile RAM applications. In *2022 Int. Electron Devices Meeting (IEDM)* (ed. De Salvo, B.) 242–245 (IEEE, 2022).
56. Seo, S. M. et al. First demonstration of full integration and characterization of $4F^2$ 1S1M cells with 45 nm of pitch and 20 nm of MTJ size. In *2022 Int. Electron Devices Meeting (IEDM)* (ed. De Salvo, B.) 218–221 (IEEE, 2022).
57. Ikegawa, S. et al. High-speed (400MB/s) and low-BER STT-MRAM technology for industrial applications. In *2022 Int. Electron Devices Meeting (IEDM)* (ed. De Salvo, B.) 230–233 (IEEE, 2022).
58. Lee, P.-H. et al. 33.1 A 16nm 32Mb embedded STT-MRAM with a 6ns read-access time, a 1M-cycle write endurance, 20-year retention at 150°C and MTJ-OTP solutions for magnetic immunity. In *IEEE Int. Solid-State Circuits Conf. (ISSCC)* (ed. Cantatore, E.) 494–496 (IEEE, 2023).
59. Gupta, M. et al. High-density SOT-MRAM technology and design specifications for the embedded domain at 5nm node. In *2020 IEEE Int. Electron Devices Meeting (IEDM)* (ed. Datta, S.) 513–516 (IEEE, 2020).
60. Song, M. et al. High speed (1ns) and low voltage (1.5V) demonstration of 8Kb SOT-MRAM array. In *2022 IEEE Symp. VLSI Technology and Circuits* (eds Palacios, T. & Ginsburg, B.) 377–378 (IEEE, 2022).
61. Doevenspeck, J. et al. SOT-MRAM based analog in-memory computing for DNN inference. In *IEEE Symp. VLSI Technology* (eds Chang, C.-P. & Chang, K.) JFS4.1 (IEEE, 2020).
62. Couet, S. et al. BEOL compatible high retention perpendicular SOT-MRAM device for SRAM replacement and machine learning. In *2021 Symp. VLSI Technology* (ed. Yamakawa, S.) T11–1 (IEEE, 2021).

Acknowledgements

This work is supported in part by PRISM, one of the SRC/DARPA JUMP 2.0 Centers.

Author contributions

S.Y., A.L., J.L. and T.-H.K. researched data for the article and contributed to discussion of content and writing. All authors reviewed and edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44287-023-00002-9.

Peer review information *Nature Reviews Electrical Engineering* thanks Can Li, Zhefan Li, Giacomo Predetti and the other anonymous reviewer for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2024