Systematic Analysis on Memory Technology of It Applications and Optimization

Dylan Quanhan Chen[1, *]

[1]Guangzhou ULink International school, Guangzhou, China

*Corresponding author: dylchen2579@ulinkcollege.com

Abstract:

In the era of rapidly advancing technology, the limitations of conventional memory architectures have become a significant bottleneck in meeting the growing demands for data storage and processing. Emerging non-volatile memory technologies are increasingly closing the performance gap between traditional memory and storage solutions, driving innovations in memory cell design. This review provides an overview of conventional memory types, such as static random-access memory (SRAM), dynamic random-access memory (DRAM), and Flash memory, alongside emerging prototype memories like resistive random-access memory (RRAM), magnetoresistive random-access memory (MRAM), and phase-change memory (PCM). Emerging technology nowadays have a wide range of applications. Resistive random-access memory and spin torque transfer random access memory are the most promising memory for in-memory computing which is advanced beyond conventional von Neumann system, neuromorphic computing mimic human's brain structure. Internet of things devices and wireless sensor devices make use of the low power consumption memory, and the security is provided by the stochasticity nature of random-access memory which is not desired and is considered as a challenge for memory cell.

Keywords: Memory technology, in-memory computing, neuromorphic computing, IoT devices

1. Introduction

Nowadays, technologies are developing further and further, the achievement cannot be separated from the development of memory cell technology which allowed the information in the modern device to be stored physically as faster and more capacious with the memory component as smaller as possible. The conventional memory technology is mature now. The SRAM and DRAM are representative volatile memory which will lose their data when the power is off. Flash memory, the non-volatile memory, keep the data even without power supply. These conventional memory cell use the method of keeping the charge in the memory cell to store the data, while emerging technology which is also known as the memristor [1]. The emerging type of memory use the method of controlling the resistance to store the data. For in-

stances, Phase Change Memory (PCM) convert the state of the material, Magneto resistive RAM change the direction of spin of electrons, and Resistive RAM make use of the building up and breaking down of conductive filament. New computation technologies beyond conventional CMOS and von Neumann architecture like in-memory computing and neuromorphic computing. The application of memory cell in our daily life is also talked in this paper such as Internet of Things devices [2]. This paper introduces six basic type of memory cell and delves into future including in-memory computing, neuromorphic computing, and IoT devices, and challenges of memory cell.

2. Basic overview of the conventional and emerging memory technology

Fig.1 shows different category of memory cell. Conventional memory, SRAM, DRAM and flash memory, use charges to store data whereas emerging memory cell including MRAM, RRAM and PCM use resistance.
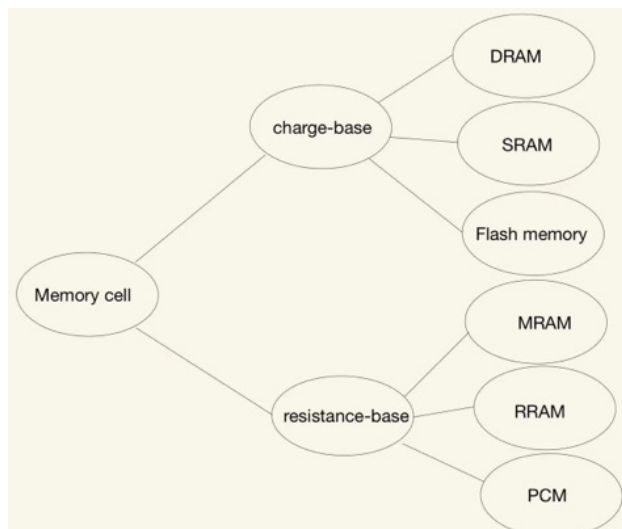

Fig. 1 Different category of memory cell

2.1 Conventional memory cell

Memory cell is widely used in electronics devices and enable devices to store binary digit which represent data in binary systems as "1"and "0". Memories are basically classified as volatile memory, such as dynamic random-access memory (DRAM)and static random-access memory (SRAM), and non-volatile memory such as Flash Memory.

2.1.1 Dynamic Random Access Memory (DRAM)

DRAM consist of transistors and capacitors where transistors work as switches and capacitors stores data. When a high voltage is applied at the gate of the transistors current

is able to flow through and charges accumulated on the capacitors and thus the cell represent a binary digit "1" [3]. When the capacitors contain no charge, it represents "0". The 1T1C-based structure shown in Fig. 2 is cheap for manufacturing and mass production.
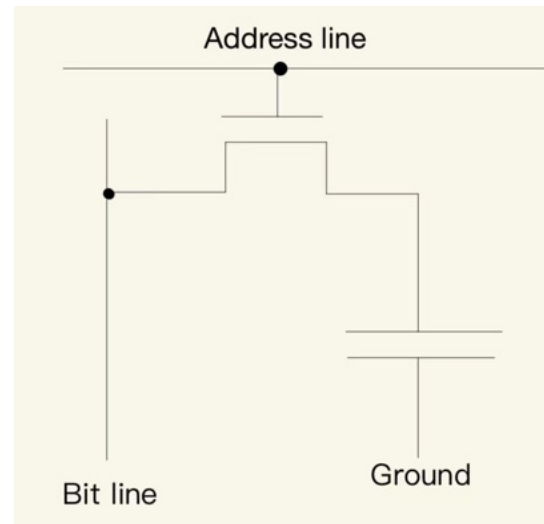

Fig. 2 Fundamental circuit structure of DRAM

DRAM provide higher storage capacity. However, the leaky property of capacitors limited the utility of DRAM in devices that require low power consumption, for example, mobile phone, IoT devices. DRAM need to be constantly recharged to keep the data.

2.1.2 Static Random Access Memory (SRAM)

SRAM with its structure of six transistors or four transistors with two resistors provide advantages beyond DRAM. The structure of SRAM is nearly symmetric, six transistors are separate into two parts, each has a transistor work as a switch. The rest two transistors of different type, PMOS and NMOS, react differently under conditions of voltage is applied through the gate shown in Fig. 3(a). The two components work together as a logic inverter shown in Fig. 3(b) which output "1" which is represented by high voltage when the input voltage is "0" represented by low voltage.
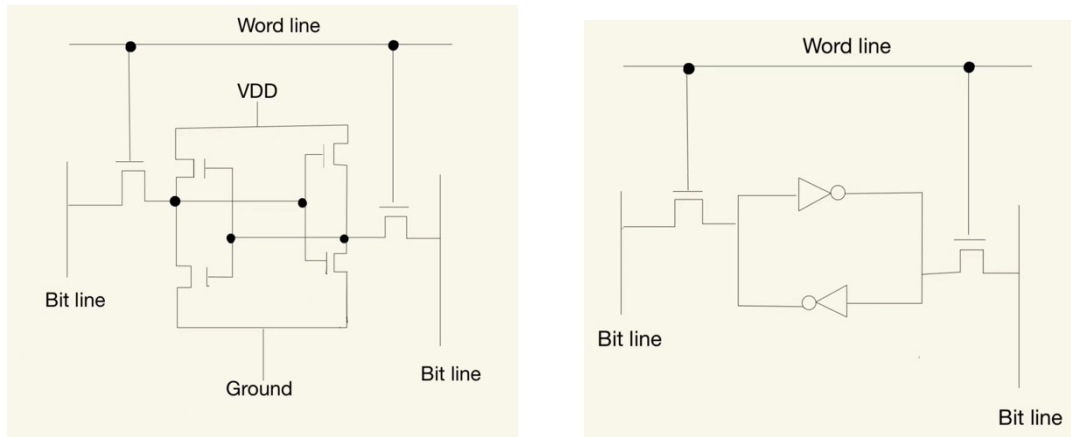
(a) six transistor structure of SRAM   (b) four transistors in can be considered as two logic inverters

Fig. 3 Static Random Access Memory

The data circulated between two inverters to keep the data. The structure provides fastest read/write speed of 100ps, but the 6T structure take up large space which is not cost-effective.

2.1.3 Flash memory

Flash memory is non-volatile memory which will not lose its data when the power of devices is off. It stores data as charges. The single unit of flash memory consist of a general transistor with an extra floating gate which enable storage of data as existence of charges in it as shown in Fig. 4.
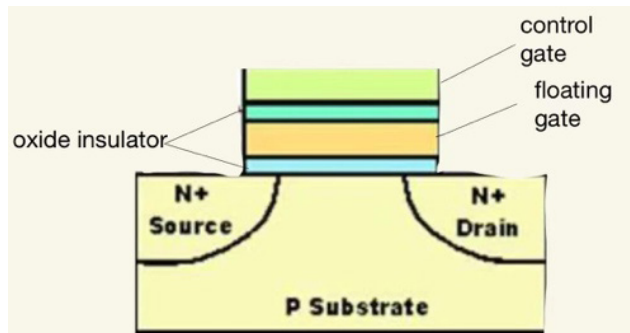

Fig. 4 A basic unit of flash memory

When applying a voltage much higher than the threshold voltage, electrons are able to go through the insulator and be trapped into the floating gate. The electrons raise the threshold voltage [4]. When applying the same voltage through the control gate. Conductive channel is formed between source and drain. Current is able to flow through. While there are electrons in the floating gate, current will be smaller which represent "0" and while there are not electrons in its, the current is interpreted as "1". However, the writing and erasing process have destructive effect on the insulator between the base and the floating gate, and thus have negative effect on endurance.

2.2 Emerging non-volatile memory

Different from conventional memory, emerging prototype make use of changing resistance to hold the data.

2.2.1 Phase change memory (PCM)

Fig. 5 shows the basic structure of PCM. Phase change memory consist of changeable material between top electrode and bottom electrode (often denoted 'heater').
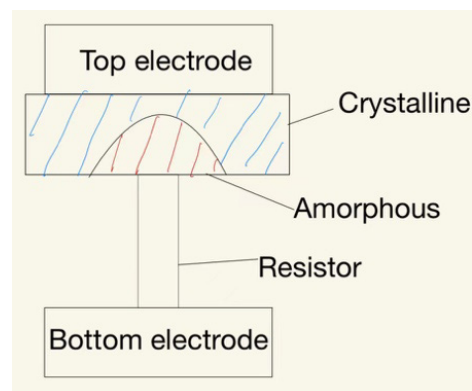

Fig. 5 PCM structure with two electrode, heater, and state changing material

The material changes its state to control its resistance

to store data. Crystalline state provides low-resistance conductive channel for current to flow through which represent logical "1", whereas amorphous state provides high-resistance channel which represent logical "0". The material can change between the different state, and even appear to be half crystalline and half amorphous which provide variation of resistance and thus represent more bits of data beyond volatile memory, for example, '11', '10'. The electrical pulse to change the material to low-resistance crystalline state is known as SET pulse and RE-SET pulse, which is higher than SET pulse, refers to the reverse process. When reading the data inside the PCM, the pulse through it is usually much smaller than SET pulse so that it does not change the data. The advantages of PCM beyond volatile memory is the high retention (data is not losing for more than 10 years at 85°C) [5].

2.2.2 Magnetoresistive Random Access Memory (MRAM)

MRAM consist of Magnetic Tunnel Junction (MTJ) and Complementary Metal Oxide Semiconductor (CMOS) shown. A basic MTJ is constructed simply by two electrode two magnetic layer with a dielectric between them. Data storage is enabled by the mechanism of direction of the polarization of the two magnetic layers. Fig. 6 shows the HRS and LRS of the STT-RAM.
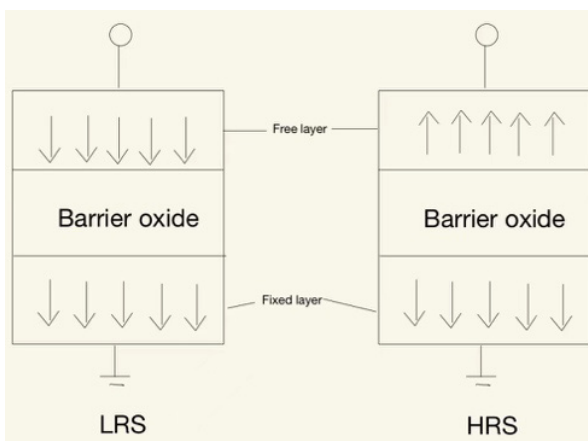

Fig. 6 Spin of electrons parallel in two layers in low resistance state and anti-parallel in high resistance state in MRAM

The polarization of one of the layers is fixed and data is stored by changing the direction of the polarization of the other magnetic layer. The data is represented by the changing resistance due to the relative magnetization of the two layers. When the polarization in different layer is parallel then the resistance is high and vice versa. To switch the direction of spin of electrons in the free layer, a strong electric field is applied and depending on the way of applying the electric field the spin of the electrons

is either parallel or anti-parallel. To make the fixed layer keeps its polarization under strong electric field, it may be pinned by an anti-magnetic material [6].

2.2.3 Resistive Random Access Memory (ReRAM/ RRAM)

There are two main type of RRAM devices: CBRAM and OxRAM. They are basically same type of things depends on different material. The RRAM take a certain material to form the conductive filament between electrodes the way of applying current through two electrodes shown in Fig. 7, the filament begins to build until it forms a conductive bridge which allow the current to flow through easier and thus appear as low resistance which represent data "1". And if there is no conductive bridge, the cell will have high resistance representing "0".
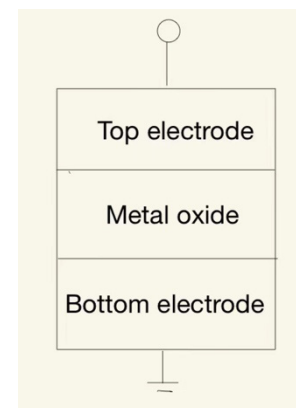
Fig. 7 Two electrons and material for building of conductive filament in RRAM

The forming of conductive bridge allows programmer to write data in it. In addition, depends on the state of the conductive filament, RRAM can achieve multi-level cell which stores more than 1 bit in a single cell [7].

2.3 Comparison between conventional memory and emerging memory technology

The desired universal memory should promise following characteristic: "low operating voltage (<1 V), long cycling endurance (>10^{17} cycles), enhanced data retention time (>10 years), low energy consumption (fJ/bit), and superior scalability (<10 nm)" [8]. Each type of memory cell possesses advantages beyond its proponents, however no single memory satisfies all the requirement of ideal characteristics. For example, while SRAM has the fastest access speed, it lacks data retention (lose data when the power off) and scalability (large 6T/4T2R structure). By contrast, STT-RAM is scalable and keep data for more than ten years. The universal memory that is superior in all aspect of memory cell is pursued to fulfil the gap between storage and memory. This gap may be filled up

by the concept of storage class memory (SCM). "Then emerging non-volatile memory, which can act as SCM, should be highly scalable and ultra-high density, with preferably MLC capacity [9]."

3. Potential applications of memory technology

3.1 In-memory computing

Computer system today mostly depends on conventional von Neumann architecture where concept of stored program is used. The program is a series of instructions that are saved in memory which is separate apart from the central processing unit (CPU) where data is process. The back-and-forth data transmission between memory and CPU cause latency and extra power consumption [10]. The power requirement to processing two values in ALU is less than fetching them from the memory. Researchers are looking for novel technology to overcome the challenge. Early idea mainly focuses on reducing the distance and fetch-and-decode time which is not effective.

The concept of in-memory computing then is considered as a substitution to the conventional computation which remove the fetching process between memory and CPU, instead, data is manage in the memory itself. In conventional computing process, data is requested by the control unit and transmitted through data bus to ALU and the data after arithmetic calculation is then sent back to the memory. In-memory computing, instead of complicated data movement, instruction from the control unit is transmitted to memory and the arithmetical operation logic is sent by the ALU and then the data calculation occurs within the memory. The in-memory computing improves the performance of computation especially with mass calculation because it greatly reduces memory access which save time and energy. These is similar to how brain works, the area for process and the area for memory is not separated clearly, and thus the brain can work under low power consumption.

3.2 Brain-inspired computing

Brain-inspired computing, namely neuromorphic computing, just as its name implies, is the memory technology that illuminated by the information operating mode in the brain. The brain performs favorably in the tasks of recognition and inference with minimal power consumption beyond the conventional von Neumann architecture [9].

Researchers studied on constructing transistors simulating the structure of the integration of neurons and synapses in human's brain. Neurons are specialized cell which is the main component of nervous system. Axon, a part of neurons, send signal impulse transmitting through neurons. Synapse, where signal impulse is sent from one neuron and receive by the second neuron, is the junction between the pre- and post-synaptic neurons. When signal impulse arrives at the end of the presynaptic neuron, it causes the vesicles contain the neurotransmitter to move towards the presynaptic membrane. Then the neurotransmitter is release into the synapses, postsynaptic neuron at the opposite site of the synapse receives the neurotransmitter to trigger a new signal. Neuromorphic computing mimics the structure of neurons and synapses to make computer hardware perform learning and thinking skills like human being.

'Neurons' in computer are connected parallel and transmit data through 'synapses'. Each neurons perform data processing itself [10]. In this case, the memory needed to perform the job of the neurons and synapses should be scalable and has low power consumption. the emerging non-volatile memory will be a suitable choice because they fulfil the requirement and can achieve the same level of density of neurons and synapses in brain($\sim 10^{10}/cm^2$) [9]. PCM and RRAM may be the best choice for this kind of function which require high scalability, low power consumption, multilevel cell (about 5 bits/).

To emulate the behavior of the synapse, CBRAM, a category of RRAM, perform well to hit the requirement of synapse function [11]. The two terminals structure CBRAM consists of reactive electrode and inert electrode which act as pre- and post-synaptic neuron and electrolyte between the two electrodes act as synaptic gap. Ions are the neurotransmitter which release by the reactive electrode. In the brain, the synaptic strength greatly affects the performance of fundamental function of the brain, namely learning and memory. The strength, also known as synaptic plasticity can be modified by the neurons [12]. CBRAM perform this function by building the conductive filament between the electrodes.

Neuron, another important component of nervous system in the brain, is emulated by integrate-and-fire (IF) model and leaky integrate-and-fire (LIF) model [13]. The signal transfer into the pre-neuron is processed and the neuron determine whether the signal should be 'fire' into the synaptic gap. The IF model keeps the signal until the spike is fired whereas signal in LIF model will leak out though the signal is not fired.

3.3 Memory in IoT devices

Internet of Things aim to provide communication channel between physical things and the internet. This enables the interaction between the physical applications and cloud

servers [14]. Data is allowed to be transmitted from the microprocessor embedded into the smart devices in daily life.

Emerging non-volatile memory perform an important job in this kind of application. IoT devices, especially the wireless type, is limited by the restricted power supplying, which means the power consumption of the memory devices in the system is required to be low so that the IoT devices will function correctly and enduringly without the physical electric connection [9]. In this case, STT-RAM and RRAM will be the best choice which possesses the advantages of low power consumption and the long endurance of the emerging non-volatile covers the shortage of the inconvenience of repairing the fault in the small piece of cell. The longer endurance will lower the times of wasting smart devices.

Security of the IoT devices is another important area where emerging non-volatile memory play an important role. Due to the connection between the IoT devices, a poor protected devices can be a potential back door for hacking [15]. The internet as a whole may suffer from the weakest devices A non-ideal property for variation of the memory are very functional to the security function of the IoT devices. RRAM is a useful memory for the construction of Physical Unclonable Function (PUF) and True Random Number Generator (TRNG). The properties of stochasticity work as a random key generator for the encryption work [9]. This greatly ensure the safety of the IoT connection.

4. Conclusion

The stochasticity of RRAM performs an important job in the security for IoT devices, however, it is not a desired properties for researchers in the development of the memory cell. Especially in the devices for high precision data computation, a controllable switching mechanism in memory is required. The filamentary switching mechanism become the major challenges for researchers. Maikap et al. introduce a possible solution by applying a dual nanostructure converting the filamentary switching which is not desired into non-filamentary switching which is desired [16]. By comparison with the filamentary switching, non-filamentary one can control the AC and DC switching without strong fluctuation at different conditions with different level of resistance. In filamentary switching devices, RRAM change the state between low resistance state to high resistance state by forming conductive bridge (CBRAM) and oxygen vacancy (OxRAM). The filamentary-length-dependent high resistance state is the main contributor of variation and randomness challenge. The non-filamentary conduction provides stable

resistance change.

Researchers nowadays work hard to achieve the universal memory technology which fulfil all the requirement for developing technology. By changing the material of the cell or the architecture of the cell, memory cell has the probability of promising access speed, scalability, low power consumption and long retention.

Emerging non-volatile memory, as the potential substitution of conventional baseline memory, may be able to fulfil the requirement of universal memory. The advanced control of data storage by resistance are promising. They play important roles in in-memory computing and neuromorphic computing beyond the confine of the CMOS and von Neumann architecture. The low power consumption operation and stochasticity in favor of the use of IoT devices and its security. Though challenges still exist as manufacturing process and application optimization, the future of the memory cell is worth expecting.

5. References

[1] Yu S, Chen P Y. Emerging memory technologies: Recent trends and prospects. IEEE Solid-State Circuits Magazine, 2016, 8(2): 43-56.

[2] Makarov A, Sverdlov V, Selberherr S. Emerging memory technologies: Trends, challenges, and modeling methods. Microelectronics Reliability, 2012, 52(4): 628-634.

[3] Keeth B, Baker R J, Johnson B, et al. DRAM circuit design: fundamental and high-speed topics. John Wiley & Sons, 2007.

[4] Bez R, Camerlenghi E, Modelli A, et al. Introduction to flash memory. Proceedings of the IEEE, 2003, 91(4): 489-502.

[5] Kim T, Lee S. Evolution of phase-change memory for the storage-class memory and beyond. IEEE Transactions on Electron Devices, 2020, 67(4): 1394-1406.

[6] Slaughter J M, Dave R W, DeHerrera M, et al. Fundamentals of MRAM technology. Journal of superconductivity, 2002, 15(1): 19-25.

[7] Chen Y. ReRAM: History, status, and future. IEEE Transactions on Electron Devices, 2020, 67(4): 1420-1433.

[8] Yu S. Resistive random access memory (RRAM). Morgan & Claypool Publishers, 2016.

[9] Banerjee W. Challenges and applications of emerging nonvolatile memory devices. Electronics, 2020, 9(6): 1029.

[10] Sebastian A, Le Gallo M, Khaddam-Aljameh R, et al. Memory devices and applications for in-memory computing. Nature nanotechnology, 2020, 15(7): 529-544.

[11] Laghari A A, Wu K, Laghari R A, et al. Retraction note: A review and state of art of internet of things (IoT). Arch Computat Methods, 2023.

[12] Tang Y P, Shimizu E, Dube G R, et al. Genetic enhancement of learning and memory in mice. Nature, 1999, 401(6748): 63-69.

[13] Wang R, Yang J Q, Mao J Y, et al. Recent advances of volatile memristors: Devices, mechanisms, and applications. Advanced Intelligent Systems, 2020, 2(9): 2000055.

[14] Kopetz H, Steiner W. Internet of things. Real-time systems: design principles for distributed embedded applications. Cham: Springer International Publishing, 2022: 325-341.

[15] Gatsis K, Pappas G J. Wireless control for the IOT: Power, spectrum, and security challenges. Proceedings of the Second International Conference on Internet-of-Things Design and Implementation. 2017: 341-342.

[16] Maikap S, Banerjee W. In quest of nonfilamentary switching: a synergistic approach of dual nanostructure engineering to improve the variability and reliability of resistive random-access-memory devices. Advanced Electronic Materials, 2020, 6(6): 2000209.