RRAM

111062510 張又仁

- 1. Introduction
- 2. Structure of RRAM
- 3. How does RRAM work ?
- 4. Application
- 5. Why RRAM is not in commercial use now ?
- 6. Conclusion
- 7. Reference

1. Introduction

1.1 An overview of memory

1.1.1 RAM

Random Access Memory (RAM) is a type of memory that we can access any cell randomly. RAM can either be volatile or non-volatile. A volatile memory loses its previous stored data when removing the power supply. Dynamic random-access memory (DRAM) and static random-access memory (SRAM) are volatile RAM. SRAM is often used as cache in the cpu and DRAM is often uses as main memory in the computer.

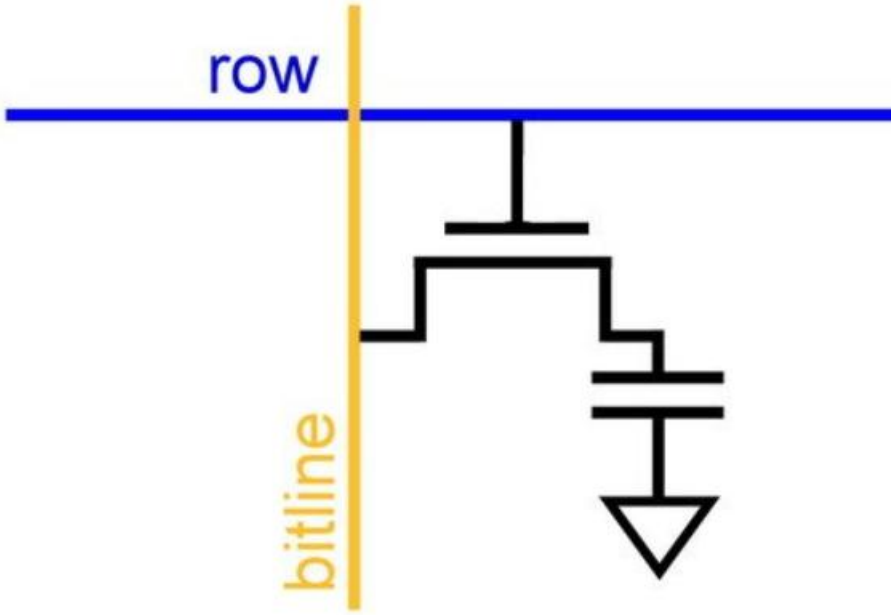

Fig. 1 DRAM

A DRAM contains a transitor and a tiny capacitor. The capacitor can be charged or decharged to store data. Since the structure of the DRAM is

simple, it can easily achieved high density and capacity, and thus the price is lower than SRAM. Currently, a 8GB DDR4-3200 DRAM is less than 800 NT dollors. However, the charge stored in capacitor will leak. Therefore, DRAM needs to be refresh, which means recharge the capacitor, regularly.

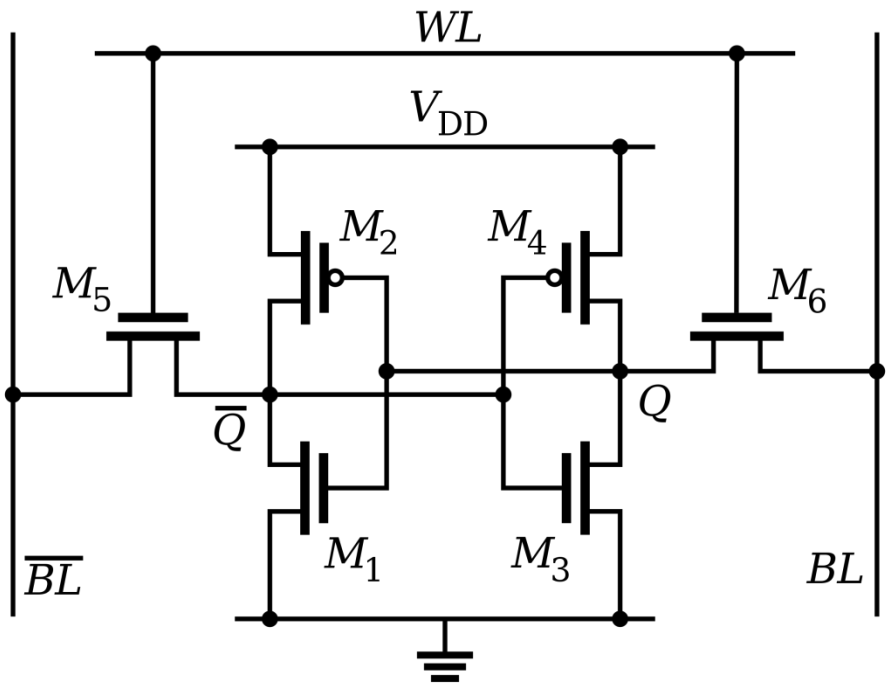

Fig.2 SRAM

A SRAM contains 6 trainsitors, which means that it is “stable”. Since the structire only contains tansistor and on capicitor. SRAM does not need and reflesh, which leads to high speed. However, the density of SRAM is lower than DRAM and therefore the cost of SRAM is much more higher than DRAM.

1.1.2 Flash Memory

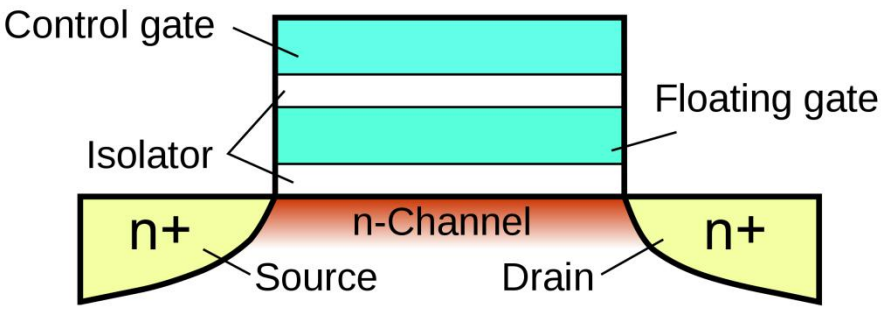

Fig.3 Flash memory

Flash memory consists of Floating-gate MOSFET, which shown as the figure above. The electron can be stored in the floating gate. As a result, flash memory is non-volatile.

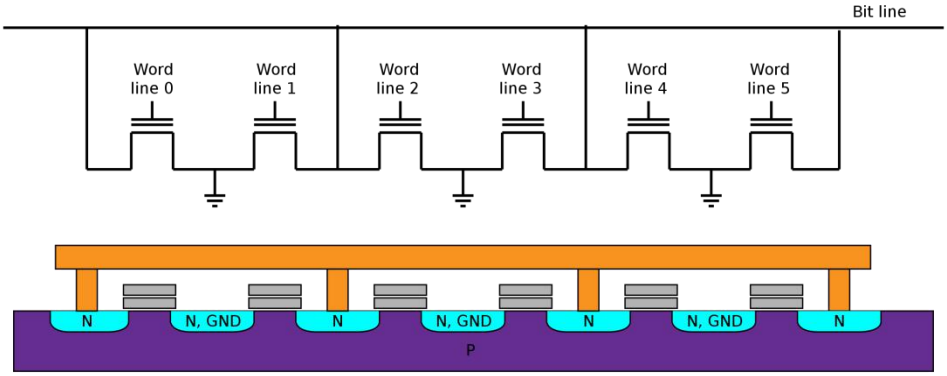

Fig. 4 NOR Flash

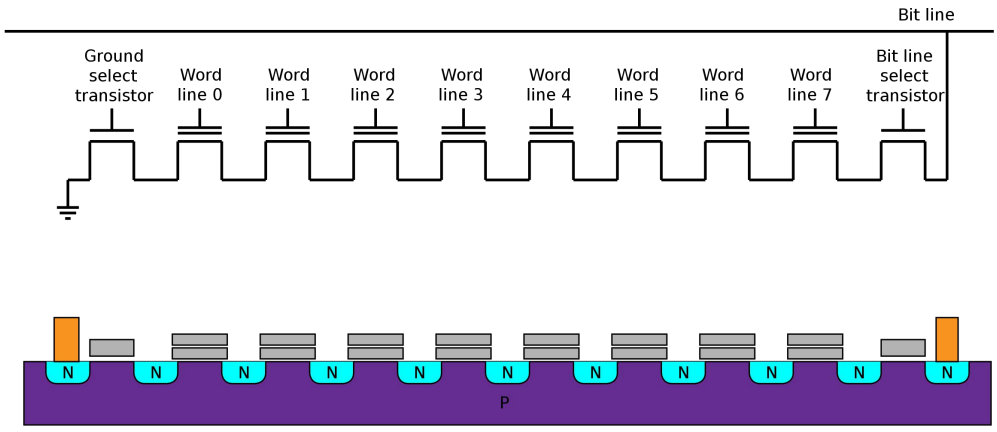

Fig. 5 NAND Flash

Flash memory can be divided into two types, which are NOR-Flash and NAND-Flash. NOR-Flash has better ability of random access. The controler the NOR-Flash can easily access certain part of memory cell. However, the continuous write and read speed are slow for NOR-Flash.

The memory cell of NAND-Flash connected in series. When performing reading or writing, the NAND-Flash controler will read or write a page, which means a large number of cells, at the same time. Thus, it take longer time, but divided the reading or writing data size with the time, the speed is faster than NOR-Flash[8]. As a result, NAND-Flash is used in continuous data storage.

1.2 Why we need new memory, such as Resistive random-access memory ?

As devices are scaling down, DRAM, SRAM and Flash memory are facing the physical limit[7], which is attributed to the loss of stored charge at nanoscale and leads to degradation of the performance, reliability, and noise margin.

Moreover, requirements of large refresh dynamic power for DRAM and leakage power for both SRAM and DRAM pose serious challenges for the

design[5].

Therefore, we need new types of memory, which has the folloing feature[5]：

1. low operating voltage (< 1V)

2. long cycling endurance (> 10^17 cycles)

3. enhanced data retention time (>10 years)

4. low energy consumption (fJ /bit, 1fJ = 10^(-15)J)

5. superior scalability (< 10 nm)

Table 1 Comparison of emerging memory technologies

Memory technology	SRAM	DRAM	NAND Flash	NOR Flash	PCM	STT-MRAM	RRAM
Cell area	$> 100F^2$	$6F^2$	$< 4F^2$(3D)	$10F^2$	$4–20F^2$	$6–20F^2$	$< 4F^2$(3D)
Cell element	6T	1T1C	1T	1T	1T(D)1R	1(2)T1R	1T(D)1R
Voltage	<1 V	<1 V	<10 V	<10 V	<3 V	<2 V	< 3 V
Read time	~1 ns	~10 ns	~ 10 μs	~50 ns	<10 ns	<10 ns	< 10 ns
Write time	~1 ns	~10 ns	100μs–1 ms	10 μs–1 ms	~50 ns	<5 ns	< 10 ns
Write energy (J/bit)	~fJ	~10 fJ	~10 fJ	100 pJ	~10 pJ	~0.1 pJ	~0.1 pJ
Retention	N/A	~64 ms	>10 y	>10 y	>10 y	>10 y	> 10 y
Endurance	$> 10^{16}$	$> 10^{16}$	$> 10^4$	$> 10^5$	$> 10^9$	$> 10^{15}$	$\sim 10^6$– 10^{12}
Multibit capacity	No	No	Yes	Yes	Yes	Yes	Yes
Non-volatility	No	No	Yes	Yes	Yes	Yes	Yes
Scalability	Yes	Yes	Yes	Yes	Yes	Yes	Yes

F: Feature size of lithography

Table 1 compare the different features of memory. SRAM has the best read/write speed and power consumption is good, but it need the larges area. DRAM has good read/write speed, medium area and good power consumption, but it needs to refresh every 64 ms. Both NAND-Flash and NOR flash have small area, but the read/write speed are low and power

consuption are large. In addition, the eudrance are shorter, either. For RRAM, its read/write speed are faster than flash memory and the power consumption are better. The overall performance is close to DRAM in theory, but it is non-volatile and the area is smaller.

2. Structure of RRAM

RRAM works by changing the resistance across a dielectric solid-state material, often referred to as a memristor. Therefore, how to change the resistance is the key.

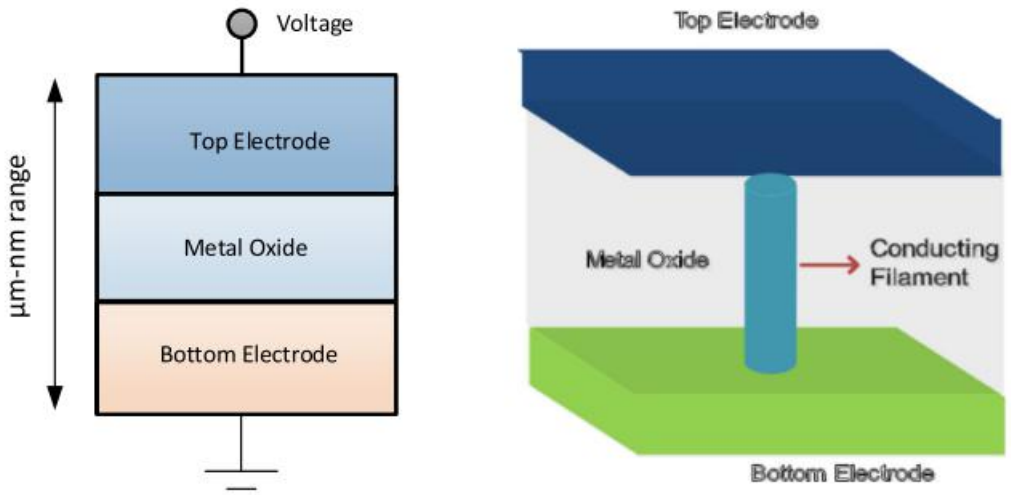

Fig. 6 RRAM Structure

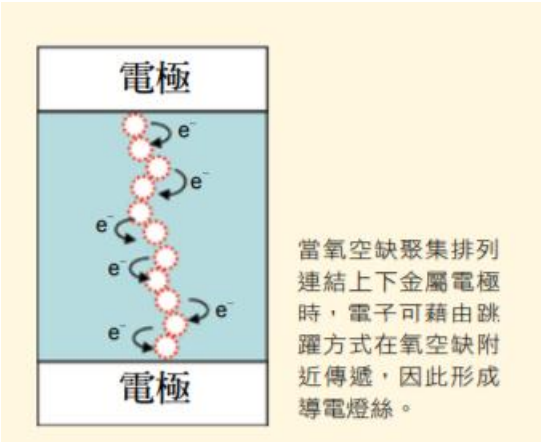

Fig 7 filament

RRAM is a simple Metal-Insulator-Metal Structure. The working mechanism is called Filament Theory. Dielectric, which is normally insulating, can be made to conduct through a filament or conduction path formed after application of a sufficiently high voltage. Filament or conduction path are arised vacancy or metal defect migration.

Table 2 Comparison of various RRAM types

Ref	Year	Top electrode	Oxide material	Bottom electrode	Operation mode	HRS/LRS ratio	Retention	Endurance	V_f	V_{set}	V_{reset}	I_{cc}
[55]	2007	Ti	ZrO_2	Pt	Bipolar	NS	NS	$> 10^4$ cycles	8.8 V	1 V	– 1.5 V	5mA
[46]	2008	Pt	ZnO	Pt	Unipolar	$< 10^5$	NS	$> 10^2$ cycles	3.3 V	– 2 V	– 1 V	NS
[81]	2008	TiN	TiO_x/HfO_x	TiN	Bipolar	$> 10^3$	$\sim 10^5$s	$> 10^6$ cycles	FF	1.5 V	– 1.4 V	25μA
[52]	2008	Pt/Ti	Al_2O_3	Pt	Bipolar	NS	NS	NS	NS	~ 7 V	– 2 V	5mA
[39]	2008	Pt	NiO	Pt/Ti	Unipolar	NS	NS	NS	5 V	~ 1 V	~3.5 V	1 mA
[30]	2009	Cu	TiO_2	Pt	Bipolar	~30	NS	NS	NS	0.8 V	– 1.5 V	300μA
[58]	2009	Ti	ZrO_2	Pt	Bipolar Unipolar	NS	NS	$> 10^2$ cycles	NS	1 V	– 1.5 V	NS
[56]	2009	TiN	ZrO_2	Pt	Bipolar	NS	NS	$> 10^2$ cycles	NS	1 V	– 1.5 V	NS
[49]	2009	Ti	MnO_2	Pt	Bipolar	$\sim 10^2$	$> 10^4$s	$> 10^5$ cycles	NS	~0.7 V	~– 1.1 V	5mA
[51]	2010	Al/Ti	Al_2O_3	Pt	Bipolar Unipolar	>10	10^4s	$> 10^3$ cycles	FF	1.5 V	– 2 V	~1mA
[42]	2010	Pt	ZnO	Pt	Unipolar	$\sim 10^2$	NS	NS	~ 3.5 V	1.1-2.3 V	0.4-1 V	5mA
[57]	2010	Au	ZrO_2	Ag	Bipolar	10^4	$\sim 10^4$s	> 500 cycles	FF	– 0.5 V	0.6 V	1mA
[43]	2011	Au	ZnO	ITO	Bipolar	10^4	$> 10^4$s	10^2 cycles	NS	~- 1.5 V	~0.5V	NS
[82]	2011	TaN	Al_2O_3/ Ru NCs	Pt	Bipolar	$> 10^5$	10^5s	NS	NS	1 V	– 1 V	10mA
[83]	2011	TiN	HfO_x/AlO_x	Pt	Bipolar	NS	NS	10^6 cycles	~ 8 V	2.5 V	– 3 V	300 μA
[44]	2012	Pt	ZnO	Pt	Bipolar	10^6	$> 10^6$s	$> 10^6$ cycles	4 V	1.2 V	– 0.5V	3mA
[84]	2013	Ta	TaO_x/TiO_2	Ti	Bipolar	10^5	$> 10^4$s	$> 10^{12}$ cycles	FF	5 V	– 4 to -6V	NS
[23]	2014	TiN	HfO_2	Pt	Bipolar	10^6	10^4s	NS	FF	– 4.3 V	6 V	NS
[22]	2015	W/Zr	HfO_2	TiN	Bipolar	NS	NS	$> 10^6$ write cycles $> 10^9$ read cycles	2 V	0.5 V	– 1.25 V	50 μA
[32]	2015	Pt	TaO_x	TiN	Bipolar	NS	NS	NS	– 2 V	< 1 V	<– 1 V	50-200 μA
[36]	2015	W	Ta/TaO_x	Pt	Bipolar	$> 10^2$	$> 10^4$s	$> 10^8$ cycles	FF	~0.5 V	~– 1 V	30-300 μA
[85]	2015	Ti	HfO_2	TiN	Bipolar	10^5	10^4s	10^{10} cycles	NS	3 V	– 3.5 V	1mA
[86]	2015	Ti	Ta_2O_5/ TiO_2 NPs	Au	Bipolar	2370	NS	< 40 cycles	FF	0.7 V	– 0.7 V	NS
[24]	2016	ITO	TiO_x/	FTO	Bipolar	10^3	NS	> 300 cycles	FF	<– 1 V	<0.5 V	20μA

Therefore, the key material is the insulator. The table above shows different materials for metal and insulator. Mental oxides are the common used oxide, such as hafnium oxide (HfOx), titanium oxide (TiOx) , tantalum oxide (TaOx) , nickel oxide(NiO) , zinc oxide (ZnO) , zinc

titanate (Zn2TiO4) , manganese oxide (MnOx), magnesium oxide (MgO) , aluminum oxide (AlOx) , and zirconium dioxide (ZrO2). Since during the operation, the Oxygen are likely to escape, the metal parts are plated a layer of Ti or Ta. These two element will store the escape Oxygen so that the filament are turn into oxid again.[8]

Since the structure and the mechanism do not refer to electron. RRAM has radioresistance, which means it can work under radiation.

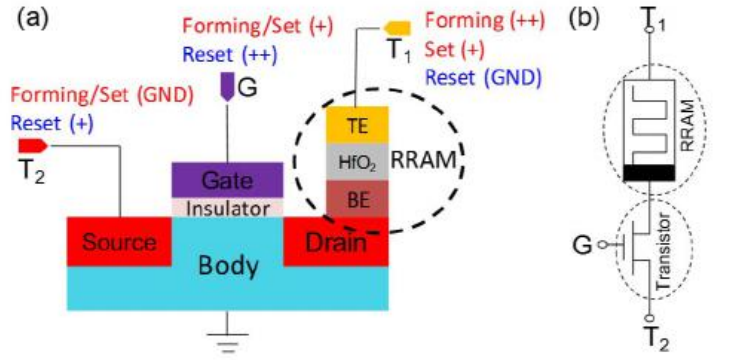

Fig. 1. (a) Schematic of 1T1R RRAM structure and operation conditions during DC sweeping. The RRAM is integrated on the drain-side of an NMOS transistor. Different voltage biases are applied to each terminal during electroforming, set and reset operations. (b) The equivalent circuit of a single 1T1R structure. The top electrode of the RRAM, the source and the gate of the transistor are defined as terminal T_1, T_2, and G, respectively.

Fig. 8 1T1R structure

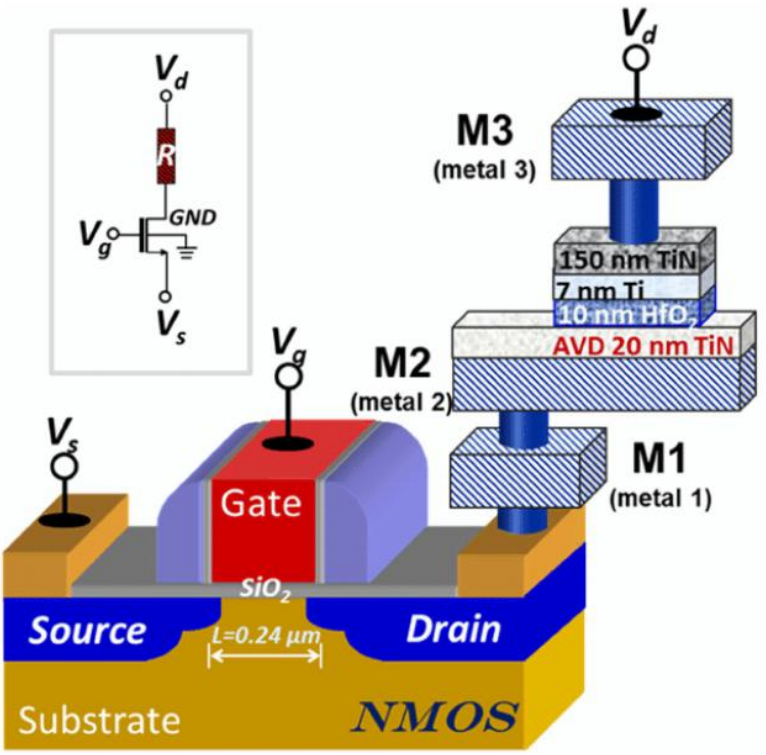

Fig.9 1T1R structure

Typically, a RRAM cell and a transistor form a structure called 1T1R, which 1T means a transitor and 1R means a RRAM cell. The transistor plays an important role in this structure. It control the RRAM and perform set, reset and forming operation. In Fig 7 and 8, the RRAM cell builds above the transistor, which means that RRAM will not occupy the precious area on chip die and is likely to integrate in modern semiconductor fabrication. Fig 7 shows how TSMC plan to build the RRAM looks like. It is built directly on the “drain” of the transistor. Fig 8 shows how Winbond build the RRAM looks like. It is built on mental layer instead of directly on the transistor.[8]

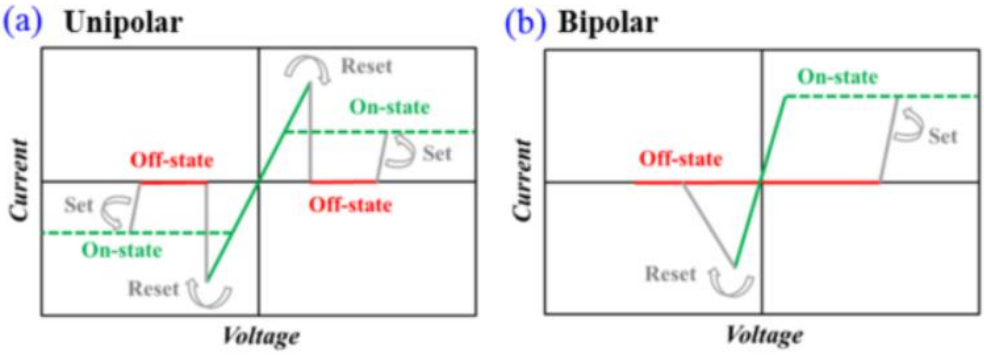

Fig. 10 bipolar and unipolar

The RRAM can also be divivded into two types, which are bipolar and unipolar. As the Fig. 9 shows, unipolar RRAM can be set or reset on positive voltage or negative voltage. On the Contrary, for bipolar RRAM, set should be operated under positive voltage and reset should be operate under negative voltage. In reality, bipolar RRAM is the common one. The voltage gap between set and reset of unipolar RRAM is smaller than bipolar RRAM, which means that it will be more difficult to control unipolar RRAM and unipolar RRAM is more unstable.

3. How does RRAM works ?

3.1 How to write data ?

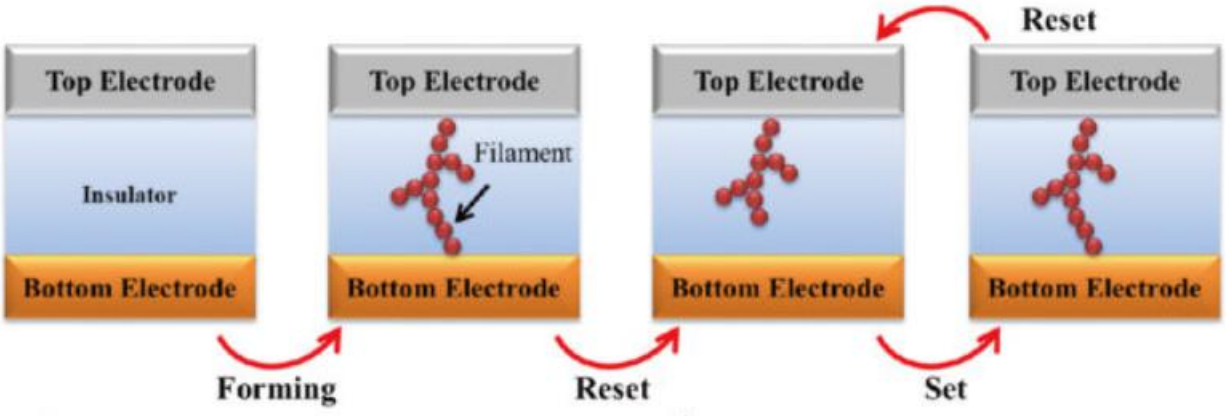

Fig 11

The working process of RRAM can be divided into three parts :

(1) Forming :

Apply a large current, about 100 μ A to create the filament in oxide. In this part, the large current heat the inuslator so that the defect part in insultator start to form the filament.

(2) Set :

Apply positive voltage so that the oxide turns into filament. Here the RRAM is in Low Resistance State (LRS).

(3) Reset:

Apply the opposite and higher voltage so that some of the filament turn into oxide, which is high resistance material. Here the RRAM is in High Resistance State (HRS)

3.2 How to read data ?

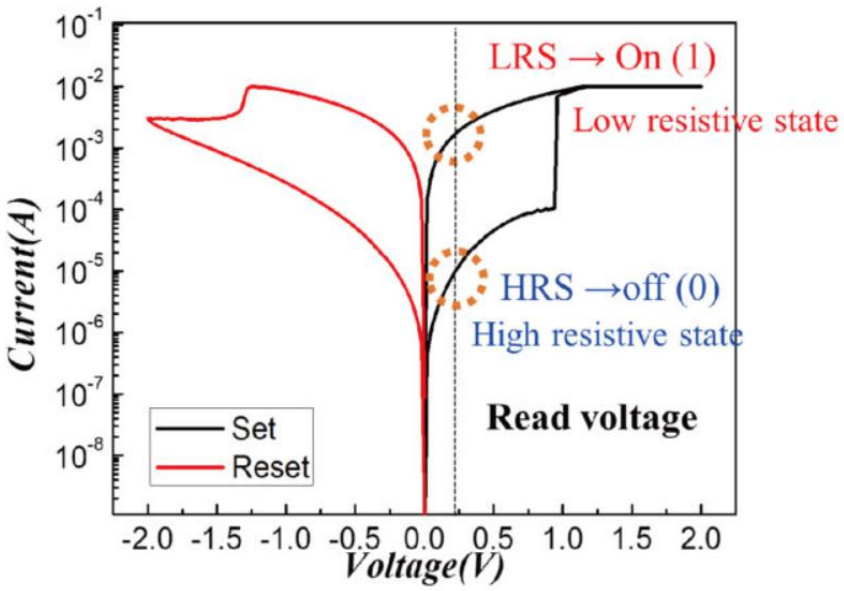

Fig. 12

When reading the data, all we needs to do is applying a low voltage on RRAM. Since the RRAM can be either in Low Resistance State or High Resistance State, the output current will be small if the resistance is large or the current will be larger if the resistance is small. Therefore, by measuring the output current, we can get the data in the RRAM.

4. Application

4.1 In-Memory Computing

Modern computer is base on von Neumann architecture or Havard architecture. These two types of architecture contains the following components :

(1) A processing unit with both an arithmetic logic unit and processor registers.

(2) A control unit that includes an instruction register and a program counter.

(3) Memory that stores data and instructions

(4) External mass storage

(5) Input and output mechanisms

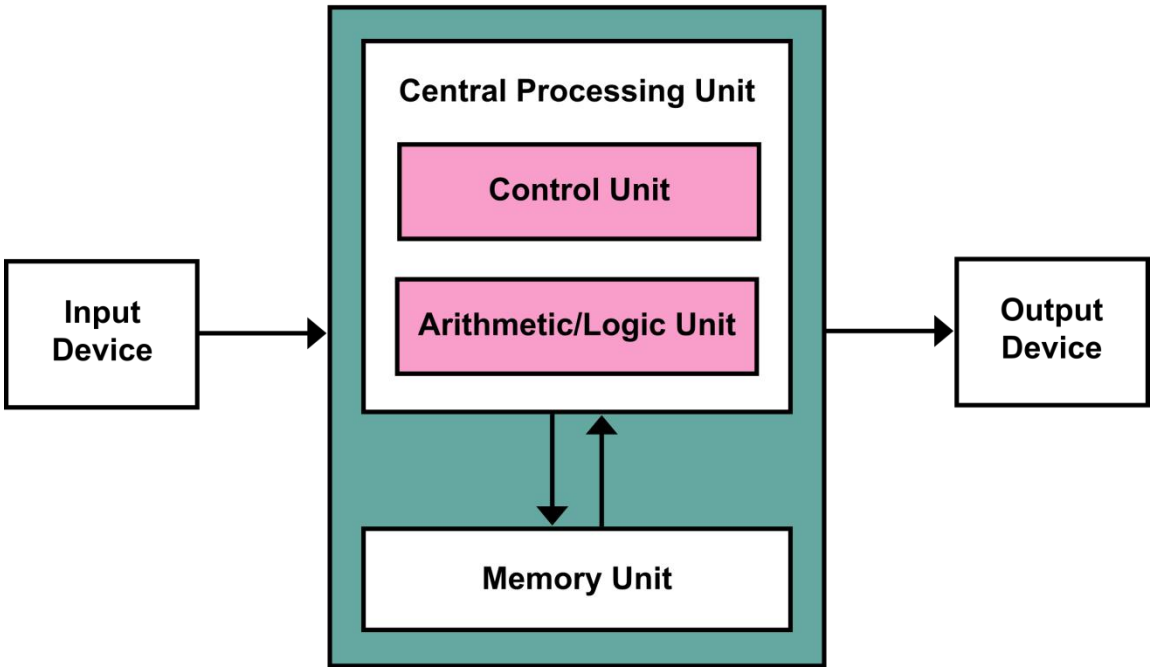

Fig. 13 von Neumann architecture[11]

The shared bus between the program memory and data memory leads to the von Neumann bottleneck, which means the limited throughput between CPU and memory compared to the amount of memory[11]. Throughput is lower than the rate at which the CPU can work because the single bus can only access one of the two classes of memory at a time. When CPU needs to perform some simple instructions on large amount of data, this limitation seriously limits the effective processing speed.

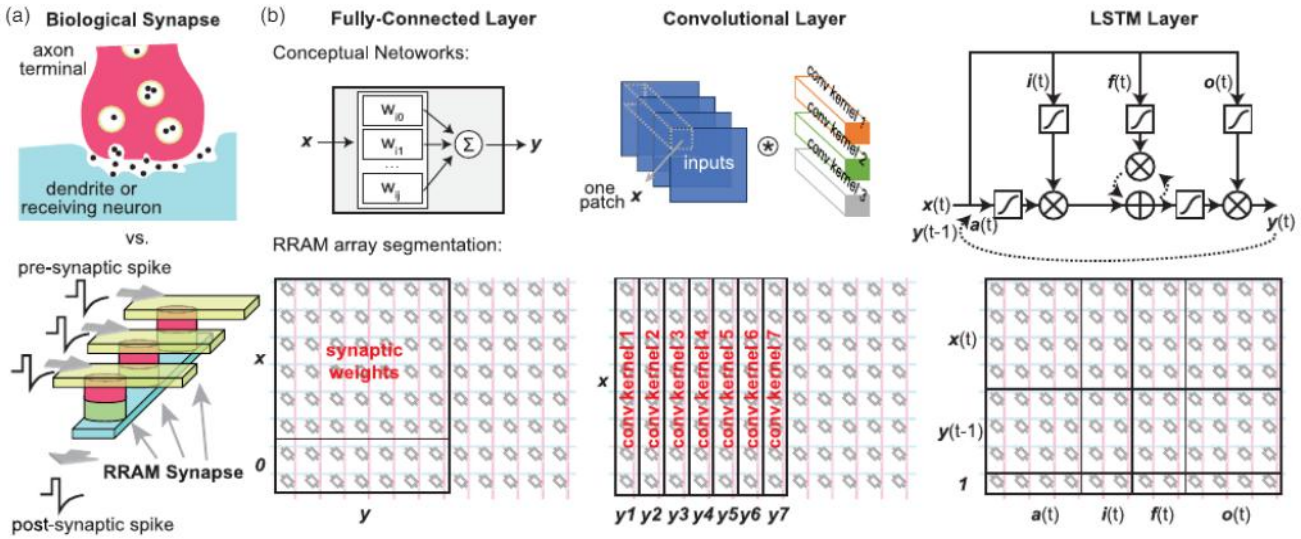

Figure 2. a) Biological synapse versus RRAM synapse.[5] b) Different segmentation schemes for fully connected layers (DNN), convolutional layers (CNN), and LSTM layers (LSTM network), all of which are converted into VMM using the RRAM array.

Fig. 14 In memory computing

In memory computing can some how solve this problem. The storage size of memory is much larger than CPU. Therefore, if the memory can perform some simple computation, than it can easy deal with large input data in Parallel instead of sending to CPU and wait for output. As a result, in memory computing saves time, consumes lower power and is high efficiency.

For RRAM, the 1T1R structure gives it the ability to perform in memory computing. Fig. 14 shows how the in-memory computing feature of RRAM helps speed up AI computation.

4.2 Synapse operation

Human brain is considered as the most complex machine and a high effitiency and low power consumption system. It not only can finish complex work in short time, such as classification, speech recognition and graph recognition but also learn and remeber new knowledge.[1]

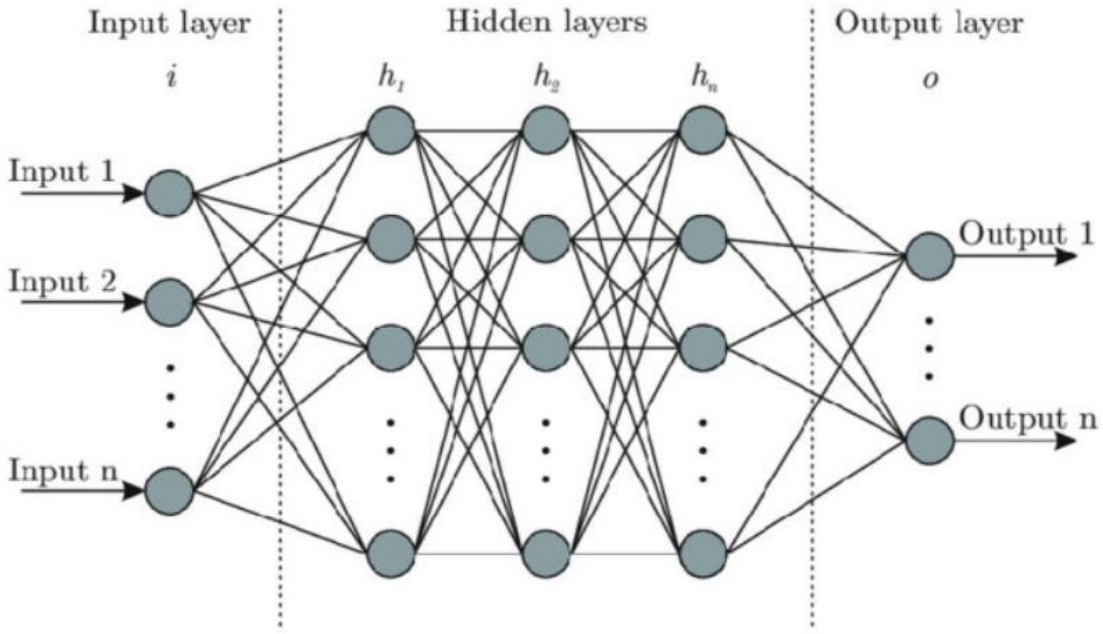

Fig. 15 Artifitial Neural Network

As a result, we human use Artifitial Neural Network (ANN) to simualte this kind of operation. ANN contains lots of nodes, shown as in Fig 15. These nodes connected with each other and transfer signals to simulate humnas brain. However, modern computer works under digital signals, which only contains 0 and 1, but humans brains work under analog. Therefore the simutlation of analog by digital system takes more time.

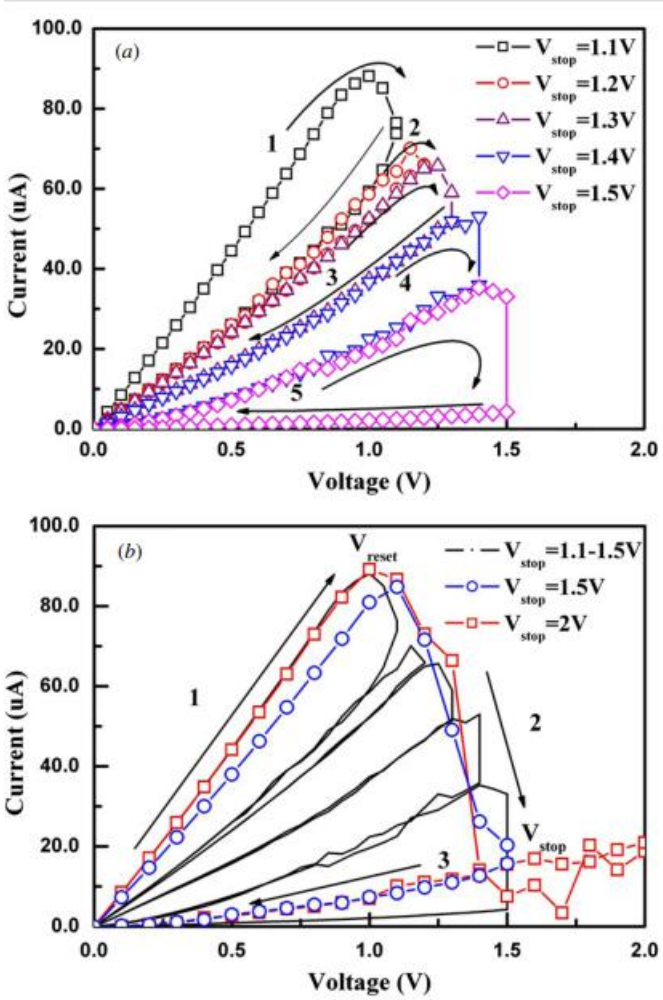

Figure 5. (*a*) Multilevel reset switching behaviors by controlling the V_{stop} ranging from 1.1 to 1.5 V. (*b*) Reset process behavior with the sweeping V_{stop} up to 2 V.

Fig.16 Multi-level Resistance Characteristics[7]

RRAM has a unique feature called Multi-level Resistance Characteristics, which means the RRAM has various resistance value instead of only 0 and 1 stage, shown in Fig. 16. With this feature, RRAM can help speeding up Synapse operation.

5. Why RRAM is not in commercial use now ?

5.1 Endurance is not as good as the research

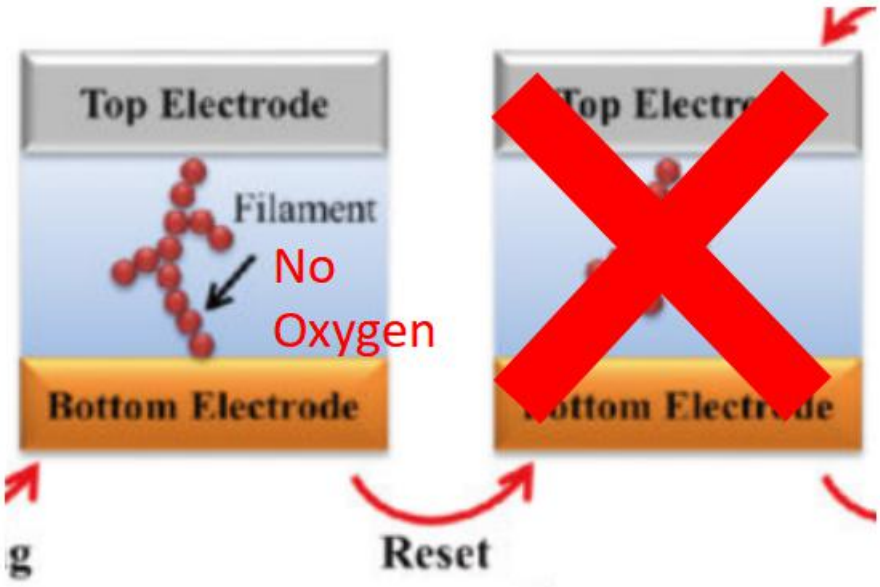

Fig. 17

The average lifetime of RRAM is about 1000 times set/reset. For every set/reset operation, the oxygen will perform reduction or oxidation on metal oxide. Therefore, the oxygen is not stable and is likely to escape or go around. The reset operation fails since there is no enough oxygen to form the metal oxide and the RRAM stays at LRS.

5.2 Power efficiency is also not good

As we describe early, forming and reset operation require large current

and voltage, which leads to high power consumption and temperature. If we try to reduce the power consumption, we must lower current. However, the filament will be unstable and the data are likely to loss since the low current can not produce enough metal oxide or filament. The oxygen is unstable as we said in 5.1. Therefore, when performing reset, the oxide sometimes turn into filament and the data 0 will turns into 1. On the other hand, when performing set, the filaments sometimes turn into oxide and the data 1 will turn to 0.

In order to make sure that set and reset poerform correctly, using larger current (100uA) makes the filament or oxide structure stabler. This leads to higher power cost.

5.3 The resistance is not stable

Section 2 introduce filament theory. It is obvious that the filament grows randomly. Therefore, for every set, the differnt structures of filament leads to different resistance. Resistance LRS will be vary for each set. For example, current set operation create 1kΩ and next time is 10kΩ. The resistance of LRS is still lower than HRS, but the value can be different. As a result, the Multi-level Resistance Characteristics is unstable and cause error when performing analog operation. In this case, RRAM cannot use

it in analog representation.

5.4 Speed is not as fast as DRAM

The set/reset operation of RRAM takes time to turn the oxide into filament of filament into oxide. It is slower than DRAM but faster than flash memory. If we apply larger current to try to speed up the process, the RRAM cell may be damage and the power consumption may be highter. In addition, the set process can not apply large current. As a rersult, the RRAM is not fast enough to replace DRAM.

6. Conclusion

RRAM is an type of emerging memory and form by Metal-Insulator-Metal structure. It works under filament theory so that RRAM can change from HRS to LRS or the opposite way by apply certain amount of current and voltage on it and store the data. However, its endurance, low energy efficiency, unstable resistance and medium speed make it unable to replace DRAM and use in commercial purpose.

7. Reference

[1] https://technews.tw/2022/04/25/ma-tek-popular-science-rram/

[2] https://technews.tw/2022/05/04/ma-tek-popular-science-rram-2/

[3] Wang, Zhuo-Rui, et al. “Functionally complete Boolean logic in 1T1R resistive random access memory." IEEE Electron Device Letters 38.2 (2017): 179-182.

[4] Yan, Bonan, et al. “Resistive Memory‐Based In‐Memory Computing: From Device and Large‐Scale Integration System Perspectives." Advanced Intelligent Systems 1.7 (2019): 1900068.

[5] Furqan Zahoor, Tun Zainal Azni Zulkifli, et al. “Resistive Random Access Memory (RRAM): an Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications.” Nanoscale Reaserch Letters.

[6] 鍾裕隆、陳貞夙，“離開歐姆定律—電阻式記憶體材料”，科學發展 2013 年 6 月 486 期 34-38

[7] Ming-Chi Wu, Wen-Yueh Jang, et al. “A study on low-power, nanosecond operation and multilevel bipolar resistance switching in Ti/ZrO2/Pt nonvolatile memory with 1T1R architecture”, Semiconductor Science and Technology, 2012.

[8] Interview with Wen-Yueh Jang, PSMC

[9] https://en.wikipedia.org/wiki/Flash_memory

[10] Gang Niu, et al.“Material insights of HfO2-based integrated 1-transistor-1-resistor resistive random access memory devices processed by batch atomic layer deposition”, Scientific Report, 2016

[11] https://en.wikipedia.org/wiki/Von_Neumann_architecture