

State-of-the-art flash memory devices and post-flash emerging memories

LU ChihYuan*, LUE HangTing & CHEN YiChou

Macronix International Co., Ltd., 16 Li-Hsin Road, Hsinchu Science Park, Hsinchu, Taiwan, 300, China

Received September 12, 2010; accepted January 5, 2011

Abstract Although conventional Floating gate (FG) flash memory has recently gone into the 2X nm node, the technology challenges are formidable below 20 nm. Charge-trapping (CT) devices are promising to scale beyond 20 nm but below 10 nm both CT and FG devices hold too few electrons for robust MLC (multi-level cell, or more than one bit storage per cell) storage. However, due to the simpler structure and its more robust storage (not sensitive to tunnel oxide defects since charges are stored in deep trap levels), CT is much more desirable than FG in 3D stackable Flash memory. Optimistically, 3D CT Flash memory may allow the density increase to continue for at least another decade beyond the 1Xnm node. In this paper, we review the current status of FG devices, their scaling challenges, and the operation principles of CT devices and several variations such as MANOS and BE-SONOS. We will then discuss various 3D memory architectures, technology challenges and address the poly-silicon thin film transistor (TFT) issues. Devices that do not rely on charge storage are naturally not limited by the number of electrons, thus promise further scaling below 10 nm. Several of the most promising post-flash era devices, their operation principle and critical issues are reviewed. (One of them, phase change memory, will be covered in a separate article thus not included here.) Their potential applications and challenges for 3D stacking are critically examined.

Keywords non-volatile memory, charge-trapping (CT) device, NOR Flash, NAND Flash, BE-SONOS, 3D NAND, VG NAND, FeRAM, MRAM, ReRAM

Citation Lu C Y, Lue H T, Chen Y C. State-of-the-art flash memory devices and post-flash emerging memories. *Sci China Inf Sci*, 2011, 54: 1039–1060, doi: 10.1007/s11432-011-4221-z

1 State-of-the-art floating gate flash memory

Flash memories have become ubiquitous in consumer electronics, mobile devices and PC and enterprise applications [1]. The basic operation principle of flash memory device is to store electrons in a floating gate (FG) or a charge-trapping layer (such as SONOS) and accordingly the V_{th} of the MOSFET devices is changed and identified as the stored logic state. The older EEPROM (electrically erasable and programmable read only memory) uses two transistors to provide RAM-like random access and bit alterable features. Flash memory gives up the bit alterable feature and adopts a block erase (flash) to achieve a 1-transistor (1T) cell that allows much higher density. Over the years, the array architecture has converged into NOR and NAND types (Figure 1) for different applications.

NOR Flash device uses channel hot electron (CHE) injection for programming and has retained fast random access capability suitable for code storage applications. The random access feature requires a

*Corresponding author (email: cylu@mxix.com.tw)

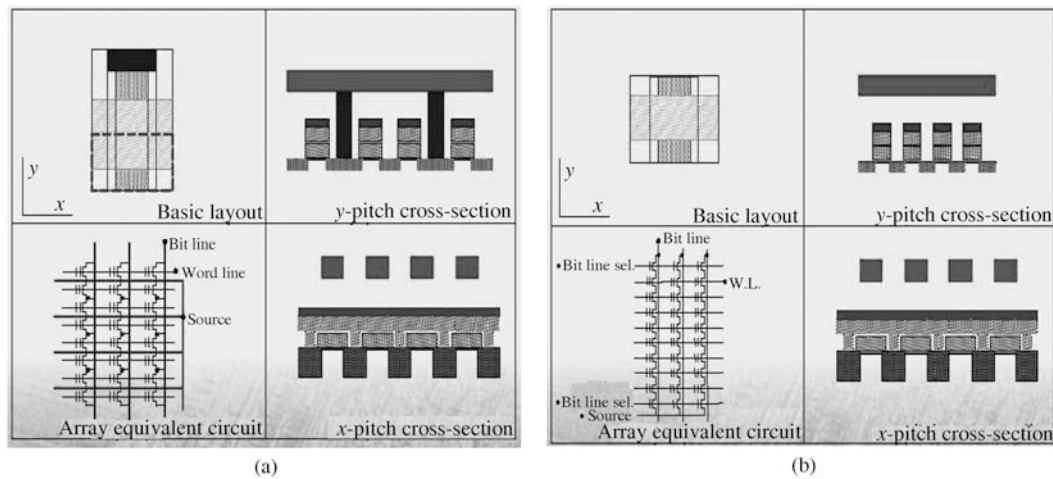


Figure 1 NOR (a) and NAND (b) Flash architectures and cell structures.

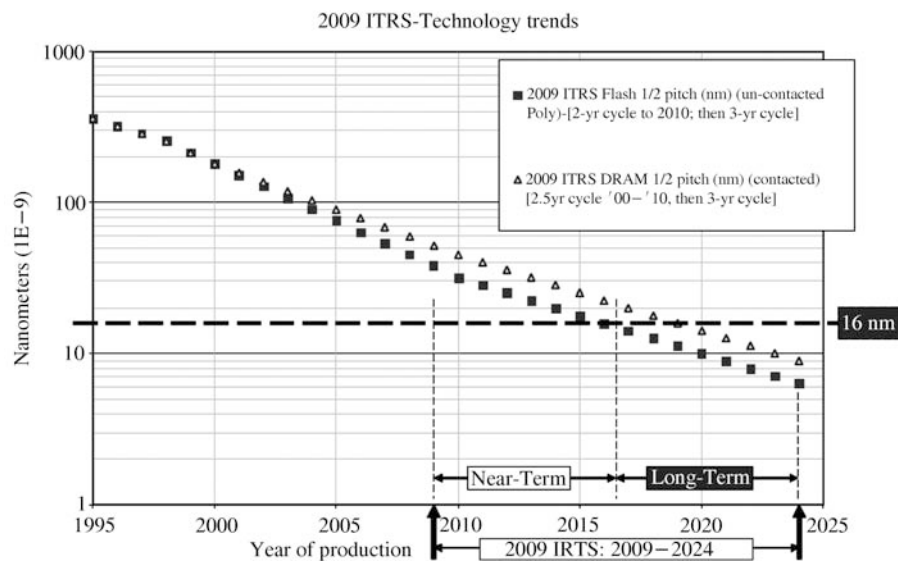


Figure 2 Half-pitch scaling trends for NAND Flash and DRAM (2009 technology roadmap for semiconductors (ITRS)).

contact in the cell thus the cell size is about $10F^2$. In addition, CHE requires deep junction, which limits the scaling capability due to short channel effects and punch-through requirement, and the poor injection efficiency limits the write throughput. Thus NOR Flash has experienced a market erosion by other technologies in the last several years. The lack of high-density demand (due to high bit cost) has limited the state-of-the-art NOR Flash products at about 65 nm node [2] even though 45 nm technology has been demonstrated [3].

NAND Flash completely abandons the random access feature and uses Fowler-Nordheim (FN) tunneling for programming and the low current allows large-page-size (>4 kB) parallel programming and reading. The array architecture requires no contact in the cell and as a result the cell size can be a minimal $4F^2$. The very high density (due to low bit cost) and the fast programming/reading throughput have propelled NAND Flash to a dominating position in mobile, consumer, PC and enterprise applications.

Because the V_{th} window of flash devices is often more than enough for a single-level logic (SLC), the memory window has been optimized to store 2-bit/cell (MLC), 3-bit/cell (TLC), even 4-bit/cell (QLC). The multi-level storage greatly reduces the bit cost that further triggers more applications.

Floating gate NAND Flash has experienced the fastest scaling trend in the IC industry in the last several years (Figure 2), surpassing both DRAM and MPU. The state-of-the-art product has already

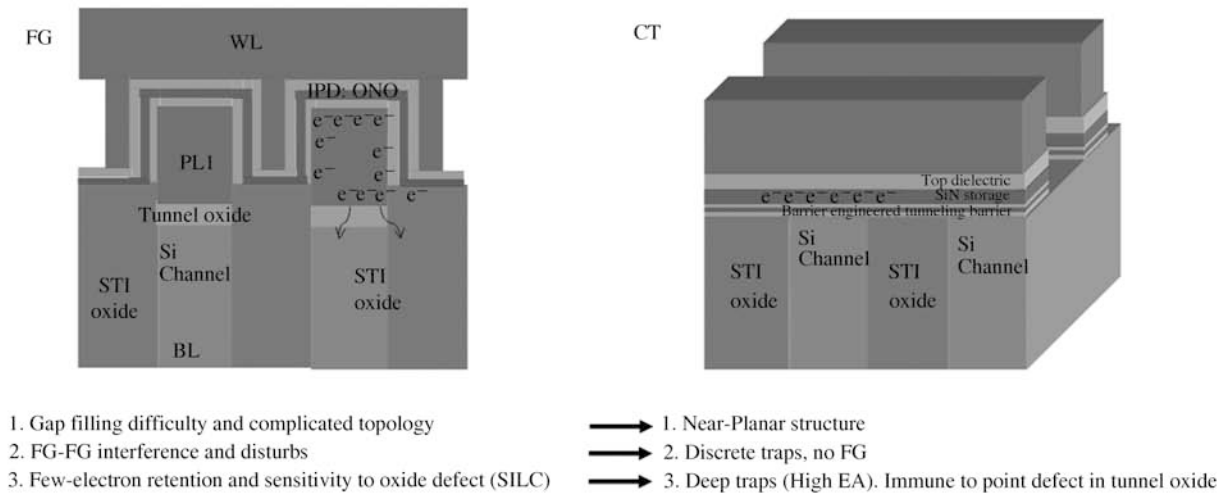


Figure 3 Brief comparison of FG NAND and planar CT NAND Flash. CT NAND has simpler and planar topology that allows better scaling capability.

gone into the 25 nm node [4] with 64 Gb density at MLC. At the time this paper is prepared, 3-bit/cell (TLC) 25 nm NAND Flash has also been announced [5].

Significantly, the rapid progress of NAND Flash is helped by its intrinsic architecture. Due to the large-density page operation, the circuit design can take advantage of error code correction (ECC) techniques and many system-level optimizations are adopted to minimize the reliability learning time.

Going forward, however, NAND Flash faces very difficult challenges: 1) geometric limitations; 2) few electron storage; 3) breakdown between neighboring word lines; and 4) FG-FG interference. These are briefly explained below and in Figure 3. (1) The ONO blocking layer thickness scales very slowly because thinner ONO cannot provide enough blocking function to stop leakage and gate injection during erase operation. At an estimated limit of about 10 nm it is very difficult to provide sufficient space between the floating gates for two ONO layers plus the control gate poly at <20 nm nodes. (2) The number of stored electrons is about 100 at 25 nm node (SLC). For MLC (4 logic levels) it is reduced to about 1/3, or 33. At 18 nm node MLC the number of electrons is near the statistic fluctuation limit. (3) The program voltage is ~15–20 V. At 18 nm (WL spacing = 18 nm) the electric field between selected and unselected word lines is ~10 MV/cm, close to breakdown field of dielectric. (4) FG-FG cross talk increasingly degrades MLC stability as the gap between floating gates shrinks. Unfortunately, there are no known solutions to any of the above issues. Therefore, sooner or later, FG devices will reach its scaling limit.

2 Planar 2D charge-trapping (CT) devices

CT devices have several advantages over the FG gate device, as shown in Figure 3: 1) the simpler and planar structure allows better scaling; 2) the insulator storage layer naturally suppresses the parasitic FG-FG interferences between adjacent cells; and 3) the discretely trapped charge storage has much more tolerance to thin tunnel oxide scaling and point defects.

Research studies [6] have already demonstrated the feasibility of 20 nm node CT NAND devices. However, mass production of CT NAND has not realized because the mainstream FG NAND has scaled rapidly. Another reason is that CT NAND often employs new high-K materials that require longer learning time.

Figure 4 summarizes several important CT NAND device structures. The conventional SONOS/MONOS (with a tunnel oxide greater than 3 nm) cannot find a suitable memory window in NAND operation. This is because the deeply trapped electrons in SiN do not easily de-trap during erasing. Although applying a very high electric field may accelerate the de-trapping speed, the gate electron injection

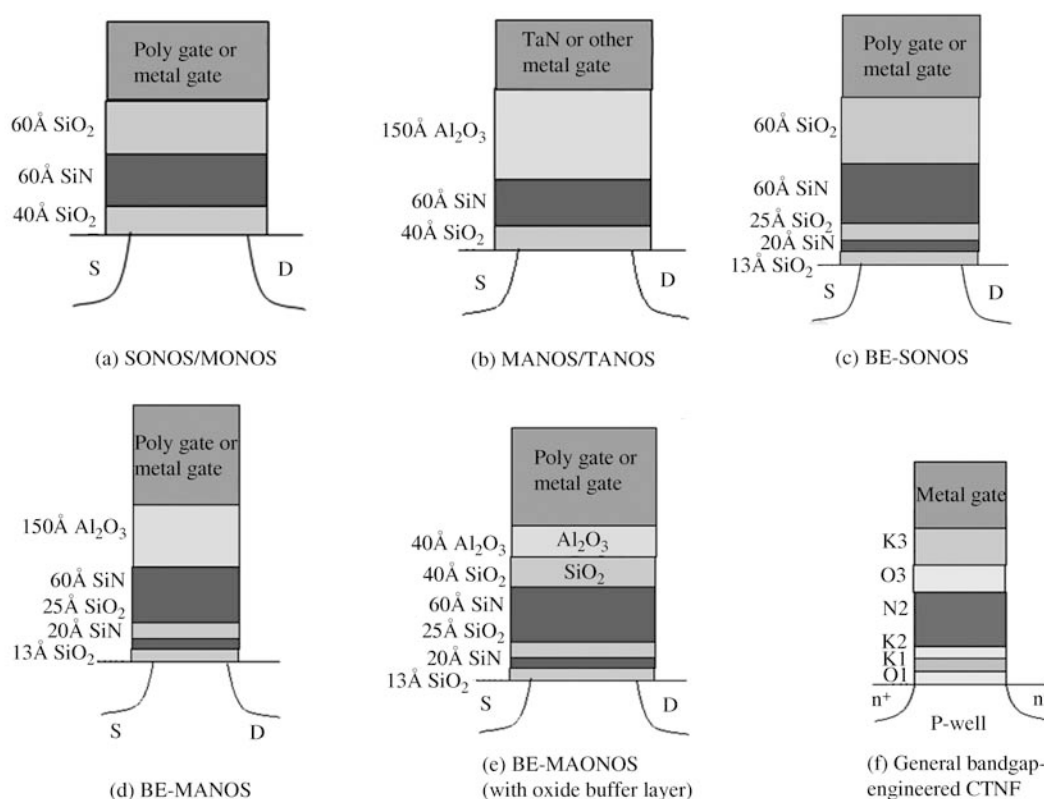


Figure 4 Summary of several important CT NAND device structures. (a) Conventional SONOS/MONOS; (b) MANOS/TANOS with a high-K (such as Al₂O₃) top blocking oxide; (c) BE-SONOS. A barrier engineered ONO tunneling barrier is used to enhance the tunneling efficiency; (d) BE-MANOS. The combination of BE-SONOS and MANOS; (e) BE-MAONOS, where a buffer oxide is inserted in between the SiN trapping layer and high-K Al₂O₃ top blocking oxide; (f) in general, the thin films in BE-MAONOS can be replaced by other materials for the better efficiency.

exceeds the de-trapping resulting in practically an increase in charge and no erasing. Using an ultra-thin (<2 nm) tunnel oxide offers an efficient hole direct tunneling erase and opens a memory window. However, the direct tunneling cannot be turned-off at low electric field, leading to poor retention and read disturb.

MANOS/TANOS [7] solves this problem by using a high-K top dielectric (such as Al₂O₃). It has been clarified [8] that the primary function of high-K top blocking oxide is to suppress the E field and gate electron injection. Thus the erase saturation of MANOS/TANOS is much lower than SONOS/MONOS, allowing a suitable memory window. However, MANOS/TANOS does not solve the fundamental problem of SONOS/MONOS because the deeply trapped electrons are still hard to de-trap. Therefore, MANOS still requires a very high E field (>15 MV/cm) to expel the electrons and this causes strong tunnel oxide degradation. Moreover, the traps in the thick high-K Al₂O₃ inevitably introduce bulk charges with shallowly trapped electrons that degrades the reliability.

A more fundamental way to solve the problem of SONOS/MONOS is to use barrier engineered (BE) tunneling dielectric that offers efficient hole injection for the erase. This eliminates the difficulty in de-trapping the electrons at very high field. A typical example is BE-SONOS [9]. The thin ONO barrier provides very efficient hole injection because the band offset during erase helps to reduce the effective hole barrier. By suitably designing the tunneling barrier thickness, the low-field leakage and read disturb issues can be completely suppressed, thus offering excellent reliability. The thin (<2 nm) nitride in the tunneling barrier is essentially trap free. Moreover, the CT device naturally has excellent immunity to the tunnel oxide defects so that BE-SONOS possesses excellent retention reliability even after strong PE cycling stressing [10].

BE-SONOS may also be combined with high-K top dielectric and form a BE-MANOS device [11].

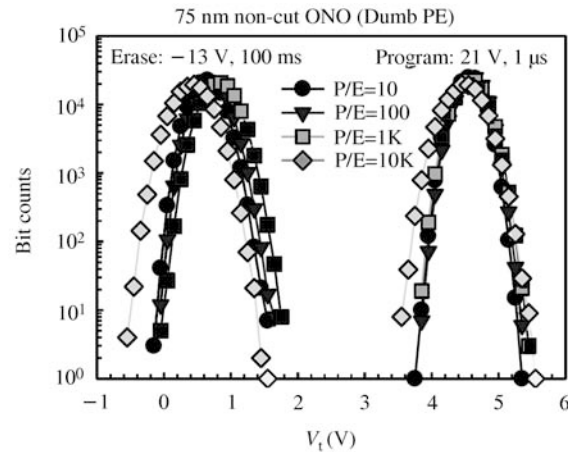


Figure 5 Dumb-mode (single programming/erasing without any P/E verify) V_t distribution of BE-SONOS NAND Flash test chip. The bit density is one block (128 kb).

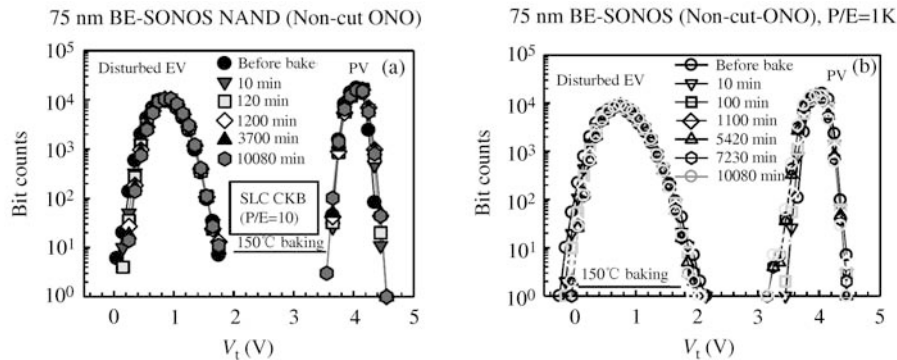


Figure 6 Retention of BE-SONOS NAND test chip after SLC checker board pattern programming and cycling stress. (a) The fresh retention after only 10 cycles. There is no discernable charge loss; (b) post 1 K cycled result.

Again, the thick non-optimized high-K Al_2O_3 also causes additional reliability problems thus we find that the best way to improve the reliability is to insert a sufficiently thick buffer oxide in between Al_2O_3 and SiN, and also minimize the high-K thickness to reduce the bulk trapped charges. The optimized BE-MAONOS [8] shows a large memory window with good reliability.

The theoretical model of the general barrier engineered (BE) CT devices has already been developed [12]. We can well predict various multi-layer CT devices with known material properties.

Among CT devices, the original BE-SONOS device (Figure 4(c)) may be the most unique because it uses only matured materials (poly, oxide, nitride) and no new high-K materials. This greatly minimizes the reliability learning and a highly reliable BE-SONOS NAND Flash test chip has already been demonstrated [13, 14]. In addition, the use of only simple material makes this approach especially important for 3D stacking of devices because it greatly reduces the processing complexity.

In Figure 5, the dumb-mode programming/erasing of BE-SONOS shows very well-controlled behavior. The V_{th} distribution of BE-SONOS NAND is generally narrower than FG NAND at the same technology node, thanks to the simpler planar topology that minimizes the device variations. The typical retention of BE-SONOS is shown in Figure 6. With optimized devices, the intrinsic retention is almost perfect without any discernable lateral or vertical charge loss. The post cycling retention is also excellent.

One remarkable advantage of BE-SONOS is that it does not show single tail bit error after cycling stressing at retention test. This is because the discretely trapped charge is immune to the SILC issue that the conventional FG devices suffer from.

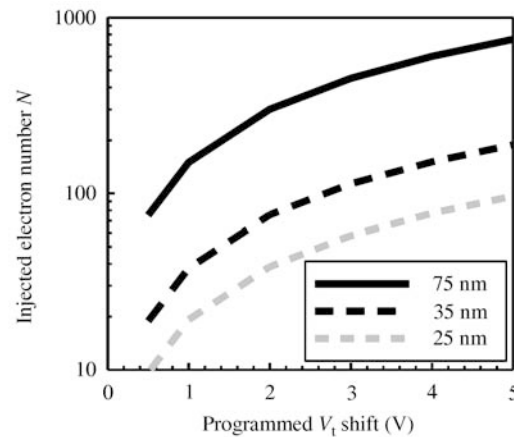


Figure 7 Calculated stored electron number of planar 2D CT NAND device.

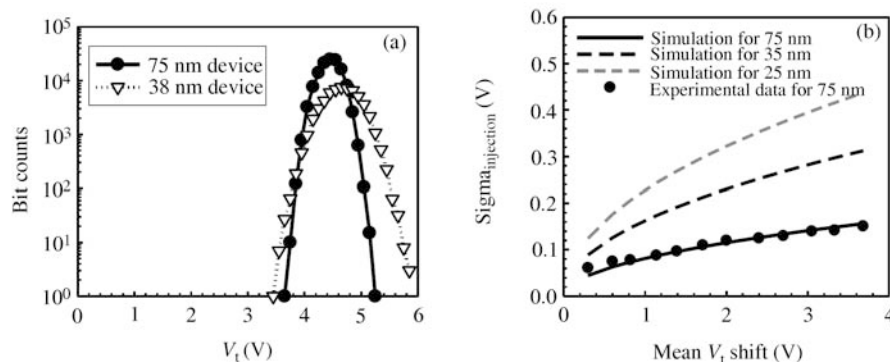


Figure 8 (a) Comparison of the dumb-mode programming distribution of 75 nm and 38 nm half-pitch BE-SONOS NAND; (b) the theoretically calculated standard deviation (sigma) of electron injection statistical spread according to Poisson statistics for CT NAND.

Although the fundamental feasibility of CT NAND (such as BE-SONOS) has been proven, further scaling of planar 2D CT NAND below 20 nm still faces extensive challenges. The most fundamental issue is the small number of stored electrons, as shown in Figure 7. Below 20 nm node, the stored electrons for a 1 V V_{th} shift would be <10 , raising a critical reliability concern. The few-electron storage not only threatens the retention reliability, but also the programming statistics. Figure 8(a) compares the programming V_t distribution of 75 and 38 nm BE-SONOS NAND test chips. The 38 nm device shows a wider distribution than the 75 nm device. This is only partly due to the larger process variation of scaled device and reflects a natural few-electron injection statistical randomness [15]. The theoretical model based on Poisson statistics [15, 16] can well predict the intrinsic sigma (standard deviation) at various technology nodes. This phenomenon indicates that the V_{th} distribution of scaled device naturally gets wider even if we have no device variation.

CT devices are much less prone to FG-FG interference but is not completely immune when the pitch is further scaled. Figure 9 shows that the V_{th} is shifted when we change the pass-gate voltage (V_{pass}). V_{pass} is the applied gate voltage of unselected word lines (WL) to allow “read-through” in NAND operation. Experimentally we find that V_{th} is observably decreased with increasing V_{pass} for the 38 nm node BE-SONOS NAND. This can be explained by the physical model shown in Figure 10. Due to the small WL pitch and relatively thicker ONO, the fringing field by the adjacent WL bias can penetrate into the selected channel [17], leading to the earlier turn-on of the cell. Thus CT NAND is not free from interference and further scaling still require optimizations such as using low-K spacer (or air spacer) to alleviate this effect [18].

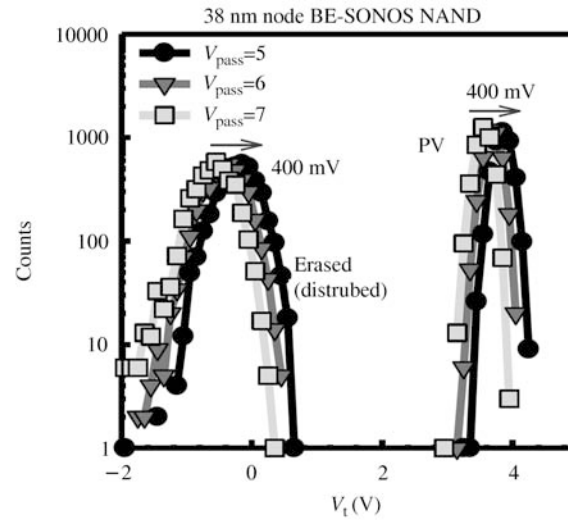


Figure 9 SLC V_t window of the 38 nm half-pitch BE-SONOS NAND device at various pass-gate voltage (V_{pass}). V_{pass} interference is clearly observed.

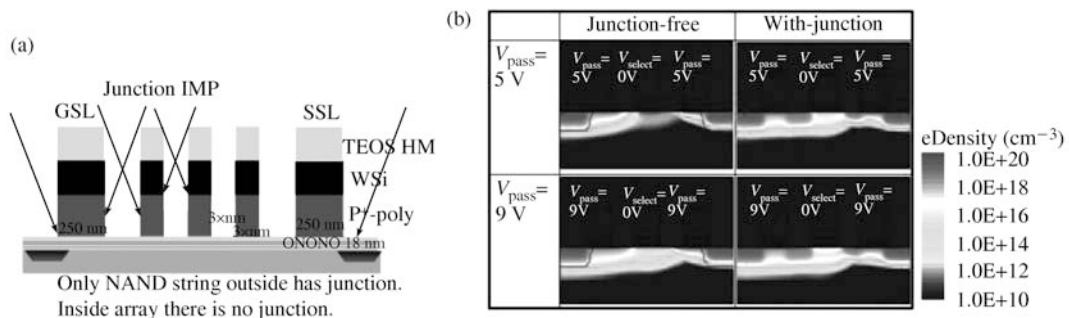


Figure 10 (a) Implantation process for a junction free NAND string; (b) simulated inversion electron density of 38 nm (half-pitch) BE-SONOS NAND device. Larger V_{pass} causes more fringing field that penetrate to the adjacent selected WL, leading to earlier turn-on of the channel.

3 3D NAND Flash devices

3.1 Introduction and current status

3D stacking of memory devices has attracted much attention because it not only alleviates the scaling difficulties in 2D devices but also promises to increase density beyond the Moore's Law. Simple stacking by repeating of CT NAND devices has been successfully demonstrated [19, 20]. Although such a simply stacked NAND is feasible, the cost advantage is limited because the critical lithography and etching steps (such as WL and BL) must be repeated many times, leading to high processing cost and possibly low yield. Moreover, the lower-layer devices experience higher thermal budget because the ONO must be deposited many times, resulting in degradation of device characteristics and modeling difficulties.

In 2007 a breakthrough approach, "BiCS" (bit cost scalable) [21], was proposed. In this approach alternate layers of polysilicon (gate) and insulator are deposited, and a single hole is etched through all memory layers in one single operation. Polysilicon channel is then deposited in the hole to form a vertical-channel NAND array architecture. The most important advantage of this structure is that it greatly simplifies the processing steps for 3D multi-layer memories. Optimistically, the bit cost can be continuously reduced when stacking more memory layers. Moreover, the CT ONO layer is only fabricated once, preserving the thermal budget.

The BiCS concept inspired many research studies of 3D NAND Flash architectures. A pipe-shaped (P-BiCS) [22] was proposed to avoid the tunnel oxide damage and improve the device performance.

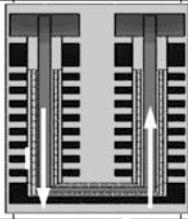
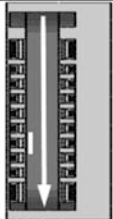

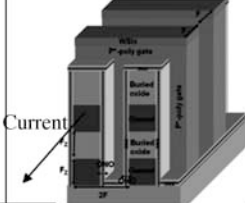



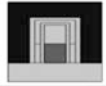
	P-BiCS	TCAT	VSAT	VG
String				
Cell shape				
Cell size in X, Y directions	$6F^2$ ($3F \times 2F$)	$6F^2$ ($3F \times 2F$)	$6F^2$ ($3F \times 2F$)	$4F^2$ ($2F \times 2F$)
Gate process	Gate first	Gate last	Gate first	Gate last
Current flow direction	U-turn	Vertical	Multi-U-turn	Horizontal
Device structure	GAA	GAA	Planar	Double gate
Possible minimal F	~ 50 nm	~ 50 nm	~ 50 nm	$\sim 2X$ nm

Figure 11 Summary of various 3D NAND architectures.

TCAT [23] was proposed as a gate-last process to enable metal gate process. VSAT [24] uses a multi U-turn channel, and vertical gate (VG) [25, 26] uses a horizontal channel but vertical double-gate device.

There are pros and cons for each structure and there is no consensus for the best approach so far. However, all 3D NAND share two common traits: 1) the WL or BL etching must be carried out only once in the multi-layer structure in order to save the processing cost; and 2) the charge-trapping (CT) material must be deposited only once to avoid cumulative thermal processing. Besides, there is no need to cut through the trap layer because it is an insulator. Figure 11 compares the merits and drawbacks for the various 3D proposals.

There are still many difficult challenges. In general the unit cell size of most 3D NAND is quite large compared to the conventional FG NAND. This is because the diameter of the hole going through all layers must accommodate twice the ONO thickness plus the poly channel fill-in, which is essentially the same geometric limitation 2D FG NAND faces. Thus by moving to 3D the advantage of the planar structure of CT device is lost. (Nevertheless, CT is still the only practical way to achieve 3D, since 1) FG structure is much more complex and is very difficult to achieve in 3D; and 2) tunnel oxide grown over polysilicon channel is too defect prone for FG which does not have deep traps and localized storage as in CT.) Moreover, both P-BiCS and TCAT need an additional WL-cut process that must include overlay tolerance and thus enlarge the cell size. It is likely that most 3D NAND must go back a few generations ($F > 50$ nm and $> 4F^2$ cell size) from the current 25 nm mode and stack many layers (> 32) in order to compete with the current 2D NAND Flash.

As the number of stacked layers increases, not only the process becomes very difficult, but also the processing cost inevitably increases and the array efficiency drops due to larger areas used for contacting the various layers. In general, the bit cost gradually saturates when stacking beyond 32 layers. Further stacking beyond 32 layers becomes less beneficial.

Therefore, for a practical 3D NAND Flash solution the cell size for 3D must not lag too far from the current 2D NAND. The best approach for 3D is to start with the same technology node of 2D FG NAND and begin with only a fewer (< 8) stacked layers. VG NAND may be the best approach to realize 3D NAND because of its considerably smaller cell size. An 8-layer VG NAND using TFT BE-SONOS [26] has been demonstrated, as shown in Figure 12, and it shows very successful array performance. Our analysis shows that VG is scalable below 25 nm node theoretically [26].

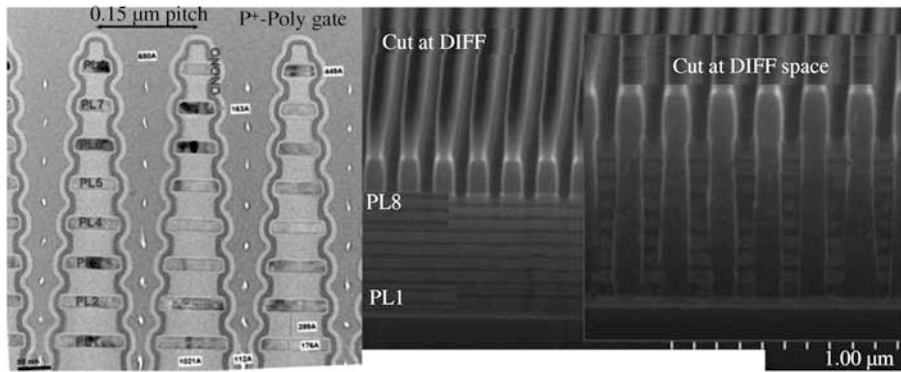


Figure 12 An 8-layer 75 nm half-pitch VG NAND using TFT BE-SONOS devices.

3.2 The challenges for 3D NAND Flash

In addition to the above mentioned issues, there are some specific challenges 3D NAND must address:

(1) Array architecture and decoding method. Either WL's or BL's must be shared in multi-layer structures in order to have an efficient decoding architecture. The decoding architecture is a key issue in 3D NAND Flash design, and it also affects the cell scalability. Furthermore, the WL or BL sharing inevitably introduces more disturb issues. One typical example is that the number of pages in each WL should be proportional to the layer number and this introduces very severe row stress (or “number of programming”, NOP) on the devices during page programming.

(2) Variability of TFT devices. Virtually all 3D NAND must use polycrystalline silicon thin film transistor (TFT) devices. Although a scaled (<30 nm) TFT CT device generally shows excellent device performance [27] for most cells, unavoidably some cells would contain grain boundaries and these cells behave differently. Fortunately, grain boundaries do not affect the FN memory window and those ill-behaved cells can be suitably managed during P/E. Optimistically we hope this variability may be overcome by system-level efforts. An alternative path to avoid the TFT issue is to develop a single-crystal device. One such example is to use epitaxial growth of Si/SiGe sacrificial layers coupled with a selective oxidation technique [28].

(3) High-aspect ratio deep trench etching. The WL or BL etching for the multi layers introduces a very high-aspect ratio trench, which raises concerns about mechanical stress and yield. Moreover, uniformity of multiple heterogeneous layers must be achieved.

(4) Process integration with peripheral CMOS circuits. The process integration of 3D TFT and peripheral CMOS circuit is also challenging because of the vastly different topography.

In spite of many technical challenges the fundamental physics of 3D NAND is quite clear and doable in principle. Optimistically, 3D NAND Flash may continue the bit cost scaling below 1X nm node.

4 Post-flash emerging memories

As discussed above, 2D flash memory is near its scaling limit and although 3D flash may extend the density scaling the challenges are formidable. Meanwhile, memories that do not rely on charge storage have shown promise to either replace flash or have performance characteristics suitable for particular applications. In this section we discuss several promising emerging memories, namely FeRAM, MRAM and ReRAM. We will address PCRAM (phase-change memory) only briefly since it will be discussed in detail in another article in this special issue of Science China Information Science. The purpose of including it here is only to compare with other emerging memories.

PCRAM uses the resistance difference between the amorphous-crystalline phases in chalcogenide glass. The basic mechanism of PCRAM is well understood. It is attractive because small devices $3\text{ nm} \times 20\text{ nm}$ have been demonstrated with excellent electrical properties [29] and it is generally thought more scalable than flash memories. The demonstrated density, read/write speed and cycling endurance for PCRAM

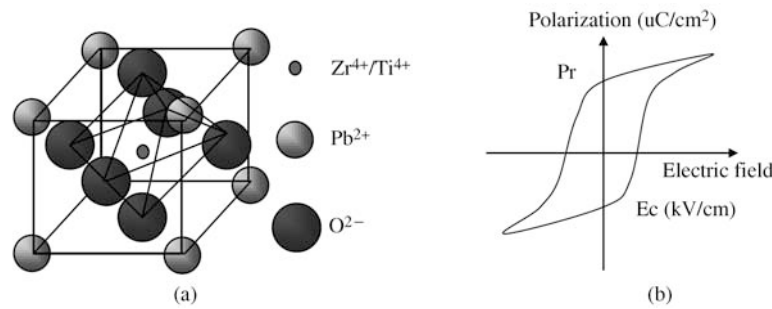


Figure 13 (a) A typical perovskite crystal structure, PZT; (b) the P-E hysteresis loop of a ferroelectric material. P_r is the remnant polarization and E_c is the coercive field.

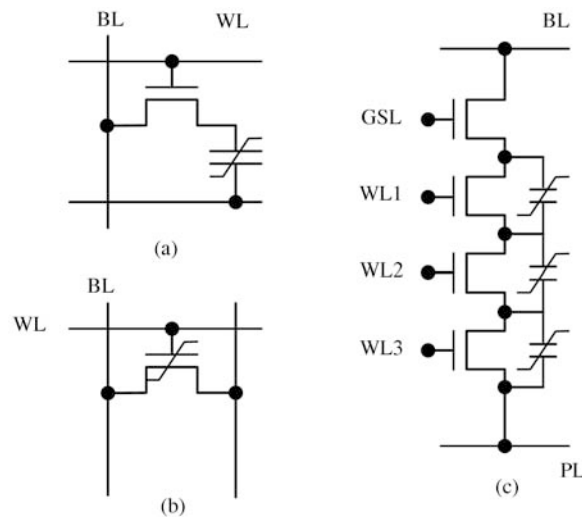


Figure 14 FeRAM array architectures. (a) The 1T-1C FeRAM structure; (b) the 1T FeRAM structure; (c) the NAND type structure.

are either on par or have exceed NOR Flash. However, there are still at least two difficult challenges for PCRAM: 1) it is challenging to realize MLC (multi-level cell) operation because of the drifting of resistance characteristic of the chalcogenide material; and 2) unavoidable recrystallization of the amorphous phase that may forbid achieving good data retention for scaled devices. Despite these challenges, several groups have already developed high-density engineering samples [30, 31] and the technology is at the verge of being adopted for mass application.

4.1 Ferroelectric memory (FeRAM)

FeRAM utilizes the switching of the polarization direction in a ferroelectric material to store data. In a ferroelectric crystal, an electric dipole exists naturally without any external electric field, since the centers of positive charge and negative charge of the crystal do not overlap resulting in spontaneous polarization. Most ferroelectric materials (usually $Pb(Zr, Ti)O_3$, PZT) have a perovskite structure, as in Figure 13(a). The ferroelectric switching is through a displacive transition and the displacement of the ions is very small ($<1 \text{ \AA}$) resulting in a very fast operation speed [32]. The polarization switching caused by the electric field is used to store data (Figure 13(b)). The high speed and voltage driven switching makes FeRAM a candidate for high-speed, low power non-volatile memory.

Several architectures have been proposed for FeRAM (Figure 14): 1) 1T-1C that uses the ferroelectric materials as a capacitor; 2) 1T FET array using a ferroelectric FET; and 3) a NAND-like array (also called chain FeRAM [33, 34]) that ties a ferroelectric capacitor in parallel with each transistor in the bit line string. The 1T-1C array (Figure 14(a)) is similar to a DRAM array and the ferroelectric capacitor's function is also similar. Also similar to DRAM, the cell polarization does not scale with the technology

node since it has to be high enough (at $Pr \sim 20$ uC [35]) not to be overwhelmed by the bit line capacitance. This imposes a stiff challenge to scaling, since even at 90 nm node 3D ferroelectric capacitor is already needed to achieve this polarization. Unlike DRAM capacitor, ferroelectric capacitor fabrication is challenging because the various unusual electrodes and ferroelectric materials are difficult to etch. This leads to large cell size and poor scaling prospective.

The NAND-like chain structure (Figure 14(c)) may reduce the cell size. Although the NAND-like structure has a theoretically $4F^2$ cell size but the ferroelectric capacitors require contacts in each cell to connect and that enlarges the cell. A nested chain structure [36] that shares the bottom electrode for two cells was proposed to reduce the cell size by $\sim 32\%$. An advanced nested chain structure further increases the density by an efficient geographic arrangement has recently been reported [37].

The FET FeRAM places the ferroelectric material directly over the gate oxide of a MOSFET. The cell size can theoretically be $4F^2$ and the read out operation does not switch the polarization direction of the ferroelectric layer and is thus non-destructive. However, the built-in electric field still stresses the polarization of the device resulting in poor data retention. Although several methods have been proposed to increase the retention [33], this is still an open issue. Meanwhile, this structure has been proposed as a possible DRAM replacement for low-power operation [38].

The development of high-density FeRAM has been stagnant in the last several years due to lack of materials breakthrough that may resolve the above issues. If the scaling and reliability issues are solved, FeRAM still promises to be a fast, low-power NVM solution.

4.2 Magnetic memory (MRAM)

MRAM employs the TMR (tunneling magneto-resistance) effect in a MTJ (magnetic tunneling junction) device to store data. A MRAM cell consists of a transistor as the access device along with one MTJ as the storage node (1T-1MTJ). The principle of TMR is shown in Figure 15. When electrons go through the fixed magnetic layer, the spin direction of the electrons is aligned by the magnetic field, and remains the same when going through the ultra thin tunneling layer. If the free layer also has the same polarization there is no force to reverse the spin direction, resulting in lower resistance. If the free layer is of opposite polarity then the spin gets realigned again, resulting in higher resistance.

A quantity, magnetoresistance ratio (MR ratio), is used to measure the resistance change: $MR (\%) = (R_{AP} - R_P)/R_P$. MR ratio as high as 600% has been reported when MgO is used as the tunneling barrier [39].

MRAM has traditionally been using an externally generated magnetic field to switch the free layer. It is difficult to confine the magnetic field, thus two wires, one in the WL direction and one in the BL direction are used to generate the full magnetic field needed to switch the free layer at the intersection, while all other nodes only experience 1/2 of such field (Figure 16). However, polarity switching in a magnetic material is a complex 3-dimensional process and often the creation of a magnetic vertex is the minimum energy path. This results in sensitivity to manufacturing variability (shape and edge sensitive) and memory effects (dependency on history). Thus switching uniformity has prohibited its commercialization for some time.

An innovative solution was published in 2003 [40]. As shown in Figure 17(a), a special MTJ device called Savtchenko device was designed. A synthetic anti-ferromagnetic (SAF) structure was used as the free layer to reduce the switching field and the memory element was placed at 45° angle to allow easier “toggling” (Figure 17(b)). The 45° rotation of the memory element aligns the soft axis of the element to the maximum torque at the crossing point of the WL and BL magnetic fields. Two electric pulses, delayed by a calculated time, were applied to the memory node to toggle the magnetic moment as shown in Figure 17(c). The SAF is a sandwiched structure consisting of two free layers (one hard and one soft) and a thin coupling layer, usually <1 nm of Ru. The two magnetic layers are strongly coupled through the thin Ru such that they are always in antiparallel positions. The SAF allows the finesse to switch a soft magnetic layer at low field but with the effect of switching a much harder magnetic layer on the opposite side of the SAF, which gives good data retention. The invention of the Savtchenko cell and toggling switching greatly improved switching uniformity and commercial SRAM products using this technology

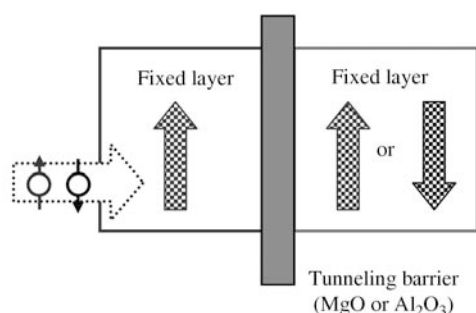


Figure 15 A schematic plot of a MTJ device showing the fixed layer, the tunneling barrier, and the free layer.

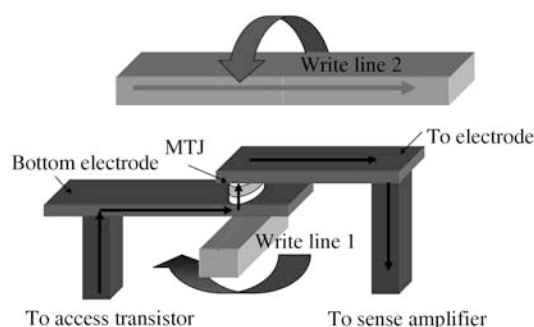


Figure 16 Configuration of writing lines, read lines, and the MTJ in a conventional MRAM memory cell.

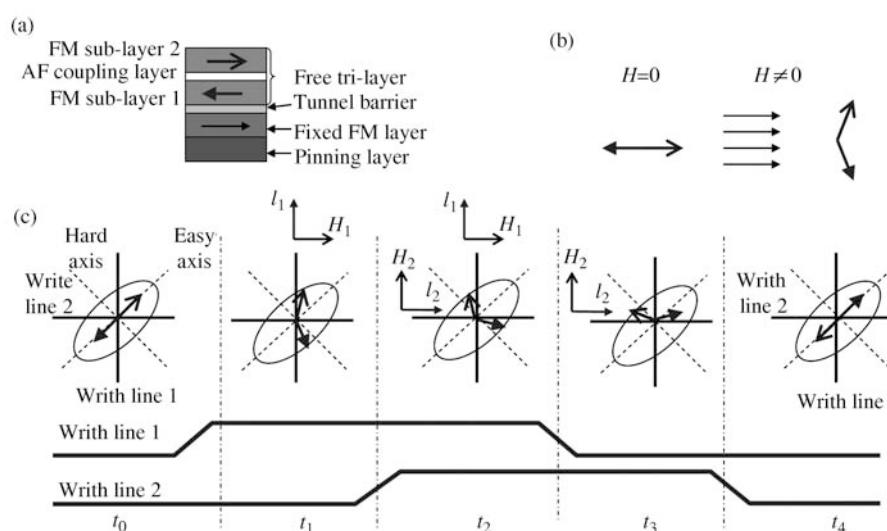


Figure 17 (a) The layer structure of a Savchenko switching device. The synthetic antiferromagnetic (SAF) free layer consists of two ferromagnetic sub-layers and a very thin antiferromagnetic (AF) coupling layer. (b) The memory element is placed at 45° angle to the magnetic field. Under a magnetic field (H) the SAF free layer rotates its magnetic polarization and the net magnetic vector aligns to the external field. (c) The programming operation of the Savchenko switching. Two pulses are applied in sequence. The first pulse produces a magnetic field in the Y direction (in this illustration) and the magnetization vector tries to follow it. When the second pulse is turned on it produces a field along the X direction, and the magnetization vector swings further. The first pulse is then turned off leaving only a field in the X direction, and this “kicks” the magnetization vector over the hard axis. When the second pulse is turned off, the magnetization vector has made an 180° turn.

has been in production for several years now.

Another promising field switching MRAM approach is to employ thermal assistance during switching [41]. This may be accomplished by applying a current through the MTJ and thus causes Joule heating of the magnetic element at approximately the same time the magnetic fields are generated. With thermal assistance only the heated element switches and other elements stay unchanged even when they are exposed to the same magnetic field. This approach also simplifies the switching element design, since magnetic field applied to only one WL or BL is required since the writing selection is by the thermal process now. As a result, both the cell size and the magnetic field (or the current required to generate the magnetic field) are reduced. Furthermore, since this approach is insensitive to stray magnetic field there is no need to cumbersomely shield the magnetic fields as in other MRAM approaches.

Using an external magnetic field nevertheless is an unwelcome proposal for IC—it is energy inefficient and cumbersome, resulting in large cell size and power consumption. Recently, a spin transfer torque

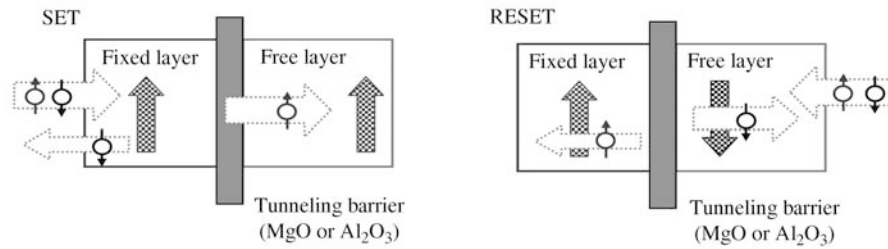


Figure 18 Schematics for operation of a STT device. During SET operation, the electron current is passed from the fixed layer to the free layer. The magnetic material in the fixed layer acts as a spin filter and only allows one type of spin (up spin in this illustration) to pass through. This spin-polarized current in turn interacts with the (unpolarized) electrons in the free layer, where a torque is exerted on electrons with the opposite spin and turns their polarization around. (The SAF structure for the free layer, which makes the free layer easier to reverse its polarity, is not shown for simplicity.) During RESET, the electron current is fed from the opposite direction. Although this current does not go through the fixed layer, a spin-polarized current nevertheless is still generated because electrons with the opposite spin (down spin) are reflected from the fixed layer. Therefore, the electrons in the free layer still experience a torque from the polarized (reflected) current. The torque, however, is now in the opposite direction thus reverts the direction of the spins in the free layer again.

(STT) type of MRAM has attracted much attention [42, 43]. The cell size of the STT MRAM is significantly smaller because wires for generating magnetic fields are not needed. The reading of the STT MRAM is the same as the conventional MRAM but the programming mechanism is completely different—it requires no magnetic field at all.

In an STT device, a torque is generated by the interaction between the spin of the magnetic layer and a spin-polarized current. Special magnetic materials known as half-metals are usually used to generate the spin polarized current. When a normal, unpolarized electric current goes through the fixed magnetic layer it interacts with the spins in the material and becomes spin polarized. The spin polarization is largely preserved when going through the super thin insulator in the MTJ. The polarized current interacts (spin transfer torque) with the free layer and switches the polarity of the free layer, as shown in Figure 18. To switch to the other spin direction, a reverse direction current is passed through the device. This current is also polarized by the fixed magnetic layer because the fixed layer reflects back only one type of spin current. The free layer polarity is switched back because the electrons reflected from the fixed layer generate a torque in the opposite direction.

STT MRAM requires bi-directional current to write both logic states, thus adding complexity to cell array and supporting circuits. Also, the current density required to generate sufficient spin transfer torque is quite high, in the 10^6 – 10^8 A/cm² range, with faster switching speed at higher current density [44]. In addition, the programming current must go through the tunnel oxide in the MTJ, stressing it, and this generates concerns about the reliability of the device. Therefore, how to reduce the current density is one of the top priorities in STT MRAM research.

Since no magnetic field is used the size of the STT MRAM cell can be theoretically as small as $6F^2$. Because of the larger current (so far) required, the cell size is still quite large right now, but is still attractive compared with SRAM. Much research and development effort to further improve and commercialize STT MRAM is on going. Recently, a method to achieve multilevel cell (MLC) was proposed [43].

4.3 Resistive memory (ReRAM)

Resistive memories are based on materials whose resistivity can be electrically switched between high and low conductive states. Although some resistive memory elements have demonstrated very high ON/OFF ratio, a ReRAM cell must still use a transistor or diode access device otherwise the memory array would be shorted along the low-resistance nodes. Therefore, ReRAM cells are either 1T-1R or 1D-1R. Up to now most reported ReRAM's operate in bipolar mode thus must use the 1T-1R cell. The memory element is usually very simple—just one or two layers of semi-insulating material sandwiched between two electrodes. In some proposals only one mask is needed to make the device (including contacts). The

array architecture for the 1T-1R cell is similar to DRAM and thus is not repeated here. The simplicity of the device and the low voltage operation are the principle merits of ReRAM. In addition, there seems so far no physics or geometry imposed limitation to the scalability of this memory. Therefore, it has achieved wide attention in recent years.

There are two major types of resistive memory: conducting bridge (CB-ReRAM) and transition metal oxide (TMO). Both are now believed based on electrochemical principles. CB-ReRAM uses the redox (reduction-oxidation) reaction of cations (e.g. Ag^+ , Cu^+ or Cu^{++}) to form/rupture a conducting bridge in a solid-state electrolyte [45, 46], while TMO relies on moving anions (oxygen) in an oxide. When programming, the oxygen ions move in the oxide to form or rupture a conducting path [47, 48]. The electrochemical nature of the redox reaction dictates a bipolar operation for ReRAM. Although unipolar operation has been reported, the unipolar switching is probably due to thermal effects and not electrochemical in nature.

Figure 19 illustrates the typical I-V characteristics for a bipolarly switched ReRAM element. The switching direction depends on the type of ReRAM. For an anion type ReRAM (e.g. metal oxide) a positive bias causes a sudden switching from low resistance state (LRS, or SET state) to a high resistance state (HRS, or RESET state). When a negative bias is applied the device switches from HRS to LRS. For a cation type device (e.g. CB-ReRAM) the operation is reversed.

It should be noted that ReRAM is also sometimes referred to as memristor (memory resistor), a name first coined by Chau in 1971 [49]. Based on the possibility of symmetry for passive devices, he proposed that there should exist an element beyond R, L, C , as the counterpart of inductor (Figure 20). As shown in Figure 20, memristor was thought should bear the same relationship to inductor the way capacitor to resistor. Based on this, memristor should have novel properties beyond variable resistance and in principle should not be equated to ReRAM. Indeed, memristor devices have been shown to be more than just resistors, for example, for logic gate applications [50, 51]. However, in memory application, memristor and ReRAM now seem to mean the same device—at least both are realized by the same materials and structures and share the same characteristics. Thus for memory applications it is safe to treat both as ReRAM.

4.3.1 CB-ReRAM

A typical CB-ReRAM uses Ag^+ or Cu^+ to form the conducting bridge and several electrolytes such as GeS [52], GeSe [45], GdO_x [46], and $\text{TiO}_x/\text{TaO}_x$ [53] have been proposed. The electrolyte should be a reasonably good electrical insulator that however allows a degree of ionic conduction.

Figure 21 illustrates a simple memory element sandwiched between two inert electrodes. The memory element is consisted of an ion source (anode layer) and the solid-state electrolyte.

As shown in Figure 22, under a positive biasing voltage, the metal atoms in the ion source layer (e.g., Cu) go through an oxidation process (in the electrochemical sense when the Cu loses an electron and becomes Cu^+ , no oxygen is involved in the process) and becomes Cu^+ and the cation drifts to the cathode and there it is reduced back to metal (Figure 22(b)). Atom-by-atom, a low resistance conducting bridge is grown from the cathode (Figure 22(c)). During RESET a negative bias is applied and the reduction process in the redox becomes operative. The Cu atom at the tip of the bridge loses an electron to the cathode and becomes Cu^+ and drifts toward the anode (Figure 22(d)). Atom-by-atom, the conducting bridge is ruptured and a high resistance RESET state is achieved (Figure 22(e)).

One weakness of CB-ReRAM is the frailness of the conducting filament in the SET state. As shown in Figure 23, metal atoms may easily diffuse back into the ion source layer from the feeble conducting bridge. This is because naturally the metal species used for the conducting bridge must be dissolvable in the ion source layer otherwise it cannot diffuse through. Therefore, during data retention these atoms have a tendency to diffuse back, since now there is no electric force holding them to the filament.

A solution to this problem has recently been reported [54]. In this work, a buffer layer is added between the electrolyte (SiO_x) and the ion source (Cu doped GeSbTe). A TiTe_x compound is chosen as the buffer layer. Without the buffer layer, the data retention is only a few minutes at room temperature because the reaction $\text{Cu} + x\text{Te} \rightleftharpoons \text{CuTe}_x$ has a low enthalpy of only -10 kJ/mol [55], thus Cu can easily react with

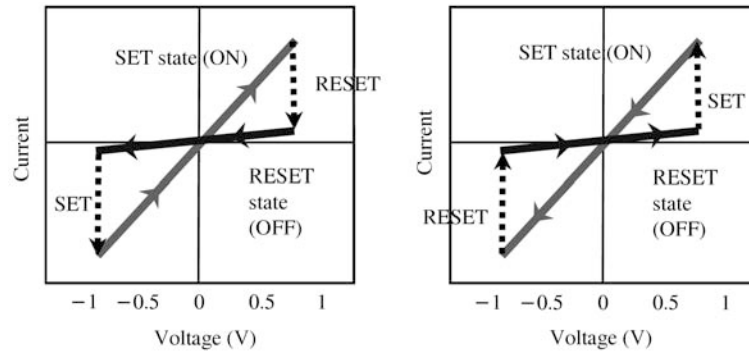


Figure 19 Schematics illustrating the bipolar switching of ReRAM elements. The left figure depicts an anion-based device (e.g. TMO, where oxygen ions move). A positive bias induces a LRS→HRS switch (RESET operation) while a negative bias causes a HRS→LRS transition (SET operation). For a cation-based device (e.g. CB-ReRAM) the positive ions move under bias and the switching operation is reversed.

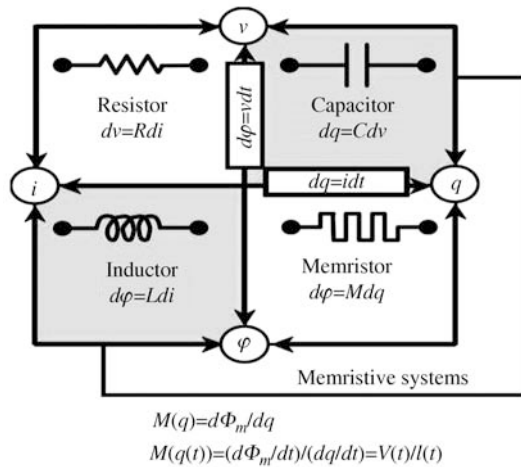


Figure 20 Memristor as a 2-terminal passive device [49]. Defined as the derivative of flux by charge, the impedance of a memristor has the same unit of resistance (V/I). The dynamic behavior of a memristor could be much more complex than a variable resistor and may possess novel properties. However, in purely memory applications it may be equated to ReRAM.

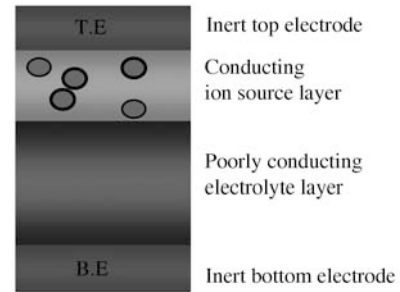


Figure 21 Schematic of a CB-ReRAM device using a Cu (or Ag) impregnated anode as the Cu (Ag) ion source.

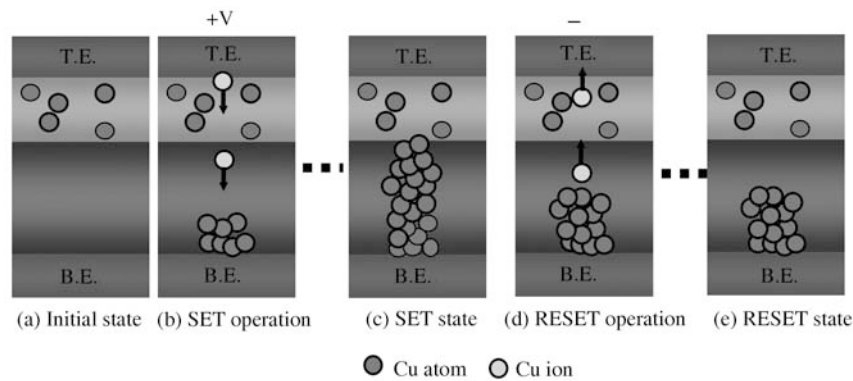


Figure 22 Illustration of the SET/RESET operation of a CB-ReRAM device using a Cu impregnated anode as the Cu ion source. A Cu bridge is formed under positive bias (SET) and the bridge dissolves under negative bias (RESET).

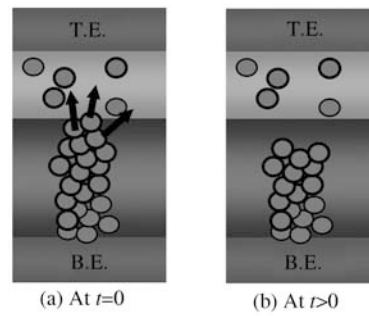


Figure 23 Mechanism for the dissolution of a conducting bridge. (a) Atoms diffuse into the ion source layer driven by high solubility; (b) filament ruptured.

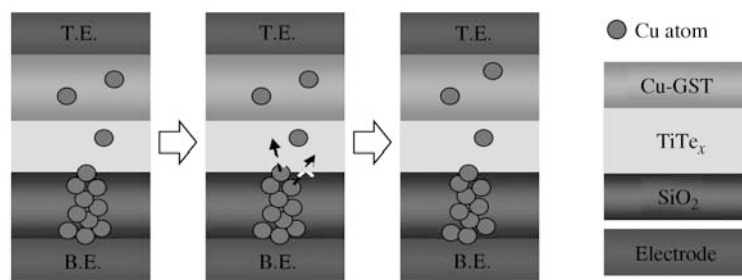


Figure 24 Time resolved schematics of SET state of a device with an ion buffer layer (TiTe_x). Cu at the interface of SiO_x and TiTe_x does not dissolve or diffuse into the TiTe_x layer, thus the bridge stays intact. Without the buffer layer, the bridge dissolves within a few minutes due to Cu dissolution into GST.

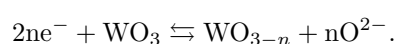
Te even at room temperature and the Cu conducting bridge is easily spontaneously ruptured. By inserting a buffer layer between the ion supply layer and the SiO_x electrolyte, the data retention is greatly improved. TiTe_x is selected as the buffer layer because the enthalpy of TiTe_x is -76.8 kJ/mol thus the structure is stable and does not react with Cu [56]. Figure 24 illustrates the effect of the TiTe_x buffer layer. The stability of this buffer layer prevents the dissolution of the Cu bridge back into the Cu source (GST) thus stabilizes the SET state. The device structure and process flow are shown in Figure 25(a) and (b), and the improved data retention in Figure 25(c). The data retention is dramatically improved from a few minutes at room temperature to several hundred hours at 150°C .

CB-ReRAM has been proposed for programmable logic device (PLD) and SoC applications [52]. The device can significantly reduce the size of PLD and also increase its performance. This device is also suitable for stand-alone applications because of its high scalability and performance.

4.3.2 TMO ReRAM

The structure of TMO ReRAM is even simpler (than CB-ReRAM) since no ion source is needed, with typically only an MIM (metal/insulator/metal) structure. In order to be compatible with the CMOS process many groups have focused on the oxides of metals already used in mass production, such as CuO_x [57], CuSiO_x [58], $\text{TaO}_x/\text{TiO}_x$ [59], NiO_x [60], HfO_x [61], and WO_x [62]. Here we use WO_x as an example to discuss the switching characteristics and mechanisms. WO_x is especially interesting because it is universally used in CMOS and a ReRAM cell can be easily fabricated by oxidizing the surface of the W plug and the entire process only adds one mask.

There are several oxidation states of WO_x , such as WO_2 , W_2O_5 , and WO_3 . Among these states only WO_3 has high resistivity. The resistive switching of WO_x ReRAM is basically a redox reaction that converts the WO_x between WO_3 and WO_{3-n} . The reaction may be written as



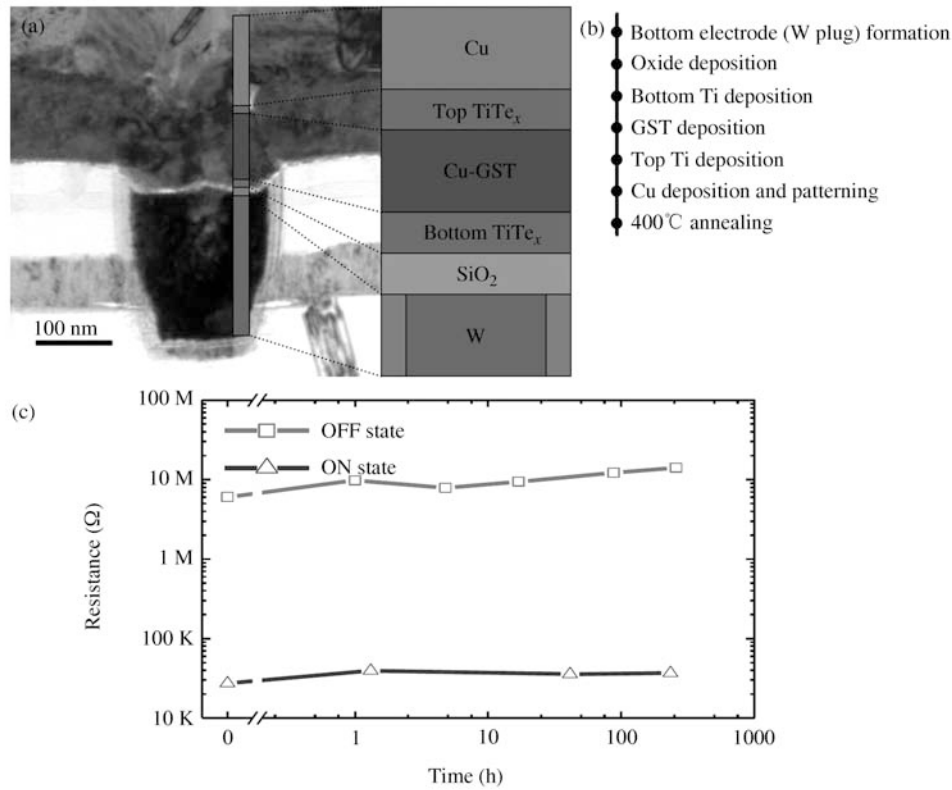


Figure 25 (a) Transmission electron microscopy (TEM) picture of the TiTe_x buffered device; and (b) the process flow for making the device. (c) Data retention of the CB-ReRAM with TiTe_x buffer layer. The baking temperature is 150°C .

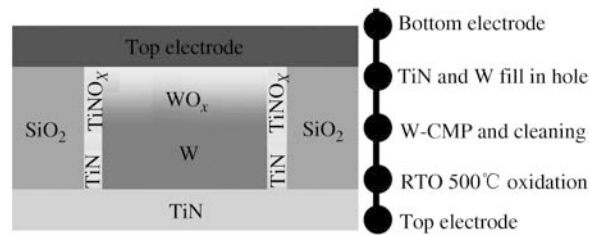


Figure 26 Schematics of a WO_x ReRAM storage device and the simple process flow [63].

A WO_x ReRAM storage element is shown in Figure 26. The top portion of a W contact plug is oxidized by either a plasma process or by RTO (rapid thermal oxidation). Fortunately, the metallic TiN liner is also oxidized and the resulting TiNO_x has high resistance and does not short out the WO_x device.

Right after fabrication, the device is in a low resistance state. This is due to numerous conducting paths, as shown by the dark spots in Figure 27(a) by C-AFM (conductive atomic force microscopy) [64]. A “forming” process is then used to seal the conducting paths, as shown in Figure 27(b). The mechanism for the forming process is briefly explained in Figure 28. By applying a positive bias the negatively charged oxygen ions from the inner (deeper) layer of the WO_x are driven to the interface between WO_x and the top electrode. The WO_x near the interface is thus converted to WO_3 and the leakage paths are largely sealed. The SET and RESET operations are just reconnecting and rupturing of the filaments in the WO_3 layer through the redox reaction depending on the polarity of the biasing voltage [65].

In addition to the WO_x layer itself, the top electrode plays an important role too. When the oxygen ion is attracted to the surface of WO_x by a positive bias, the top electrode has to block the further penetration of the oxygen (into the top electrode) otherwise the low resistance WO_3 cannot be formed.

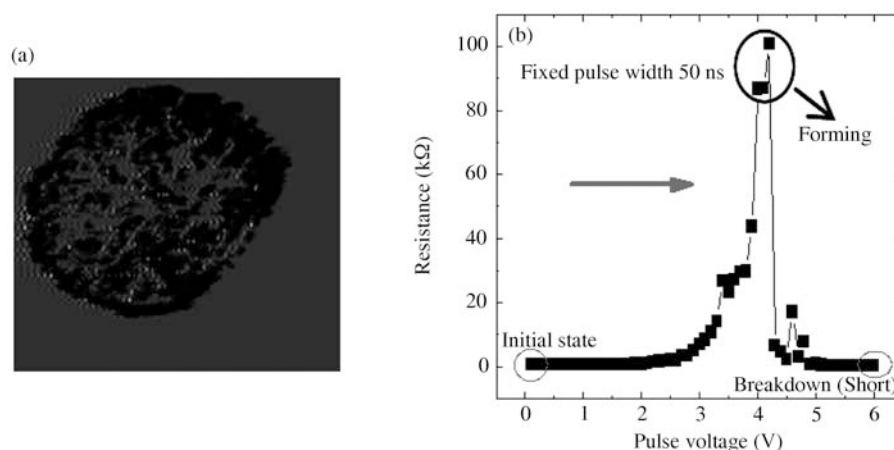


Figure 27 (a) C-AFM image of a pristine WO_x device; dark spots (high current) indicate numerous leakage paths. (b) Forming procedure of the WO_x device. The resistance of the device can be increased by applying a positive pulse. At high voltage, the device becomes irreversibly shorted.

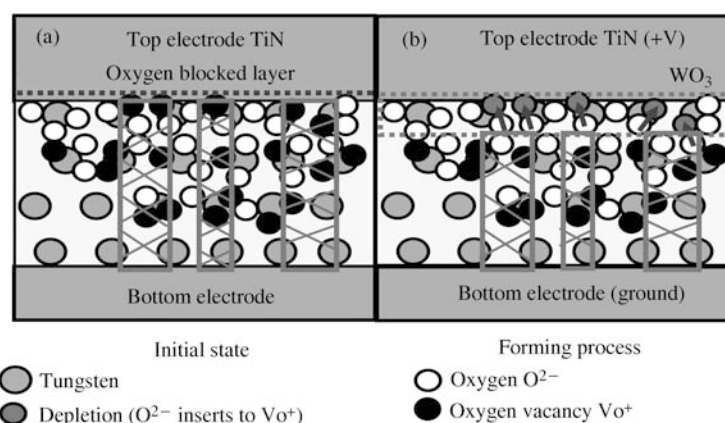


Figure 28 Proposed WO_x ReRAM forming mechanism. (a) After RTO many oxygen vacancies are present in the WO_x film and coagulate into leakage paths; (b) during the forming process, the positive voltage drives oxygen ions into vacancies to form an insulating WO_3 near the top electrode, sealing the leakage paths.

Furthermore, the work function of the top electrode affects the operation current very significantly, as shown in Figure 29 [65]. When the TE work function is high, the device is limited by thermionic emission resulting in a low operation current. When the work function is lower, for example TiN, the device operates in the space charge limited current (SCLC) range with higher current. Thus by simply replacing the top electrode from TiN to Ni, the RESET current is significantly reduced. Furthermore, excellent retention has been demonstrated. From the Arrhenius plot shown in Figure 30, with an E_a of 1.34 eV, the retention at 85°C is estimated to be 300 years.

TMO ReRAM's have demonstrated good cycling endurance using various oxides. Figure 31 illustrates good cycling endurance (1×10^9 cycles) for a WO_x ReRAM, typical for TMO devices [66, 67].

4.3.3 Unipolar operation and 3D stacking of ReRAM

ReRAM primarily operates in bipolar mode, but unipolar operation of TMO devices has been multiply reported [59, 63]. Unipolar operation may be achieved by two methods: 1) by using two different program voltages; or 2) by using different pulse durations. The mechanism for unipolar operation is not well understood. The RESET operation usually uses the same program conditions as in the bipolar mode and the mechanism for reset is believed to be the same too. In general, the SET operation conditions

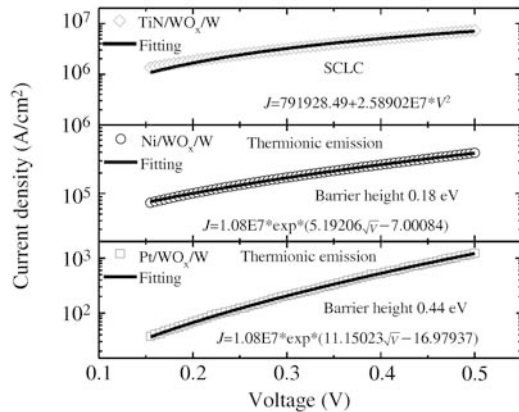


Figure 29 The influence of top electrode work function on conduction mechanism and operation current. (Top) SCLC mechanism fits the J-V curve for the initial state of a TiN/WO_x/W device. Thermionic emission fits the J-V for Ni/WO_x/W (middle), and Pt/WO_x/W (bottom) [65].

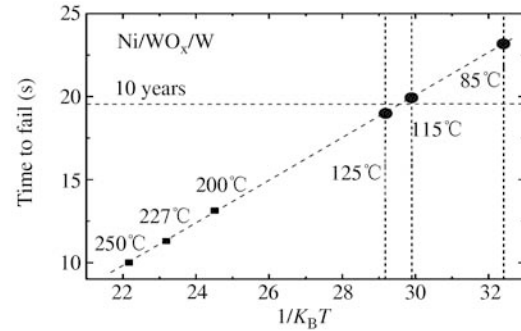


Figure 30 Arrhenius plot for the Ni/WO_x/W cell. E_a is ~ 1.34 eV. The retention time is 10 years @115°C, and 300 years @ 85°C [63].

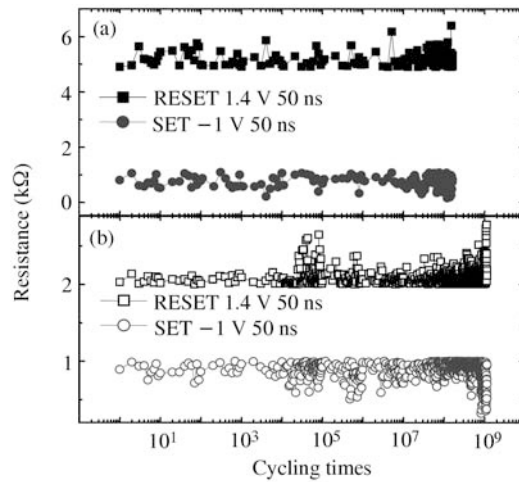


Figure 31 Typical cycling endurance for TiN/WO_x/W device. Note that the cycling endurance depends on the RESET/SET resistance ratio. Larger resistance ratio allows shorter cycling life, possibly due to longer conducting filaments.

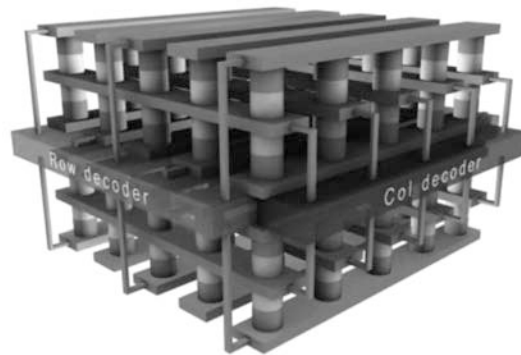


Figure 32 A schematic for a concept to build a stacked ReRAM 3D array. In this proposal a diode is connected in series with each ReRAM element and TFT devices are used for column and row decoding [68].

are considerably more severe. More energy is delivered to the device during SET operation. Therefore, it is speculated that the SET operation is thermal in nature.

It is difficult to build 3D stacked devices using bipolar ReRAM that requires a 1T-1R structure. Unipolar is much more suitable for 3D stacking since a diode may be built in series with each ReRAM element. Figure 32 shows a proposal to build a 1D-1R 3D array [68].

ReRAM has seen substantial progress in the past several years. Apart from the WO_x discussed above, HfO_x has been demonstrated with 30 nm devices [61, 66]. A 1 Mb test chip using Cu_xSi_yO has been reported recently [58].

It should be mentioned that it is possible to achieve 3D stacking using a bipolar device. In such an approach the switching current must be sufficiently low so that a bidirectional access device that can only deliver limited current density (e.g. a tunneling diode) may be used. Bi-directional 3D operation is necessarily more complex but a 4-layer 64 Mb 3D stacked circuit has been demonstrated recently [69].

At this time it is too early to predict the success of 3D stacked ReRAM. However, even without 3D, the simple structure and processing, and the outlook of unlimited scalability have already given ReRAM an important role for the next generation memory. Much work, however, is still needed to fully release its potential.

5 Summary

Flash memory has undergone unprecedented scaling speed in the last decade. Despite many challenges and fundamental limitations the momentum seems unstoppable. However, physics and statistical limitations will eventually arrest further scaling in 2D perhaps in the next several years. 3D NAND Flash using CT devices seems poised to continue the path to further increase density for at least another decade beyond the 2D limit.

New emerging devices that do not rely on charge storage promise to carry the torch even further. FeRAM, while facing the most difficulty in scaling, possesses the desirable properties of being fast and energy efficient and should find increasing high-speed/low-power applications. Thermal assisted and STT MRAM's are promising approaches for the next generation SRAM. ReRAM's, being believed the most scalable, are showing promise as long term solutions to both memory and storage.

References

- 1 Lu C Y, Kuan H. Nonvolatile semiconductor memory revolutionizing information storage. *IEEE Nanotech Mag*, 2009, 3: 4–9
- 2 Servalli G, Brazzelli D, Camerlenghi E, et al. A 65 nm NOR flash technology with $0.042 \mu\text{m}^2$ cell size for high performance multilevel application. In: *International Electron Device Meeting (IEDM)*, session 35-1, 2005. 869–872
- 3 Fastow R, Banerjee R, Bjeletich P, et al. A 45 nm NOR flash technology with self-aligned contacts and $0.024 \mu\text{m}^2$ cell size for multi-level applications. In: *VLSI Technology, System and Applications (VLSI-TSA)*, 2008. 81–82
- 4 Kuchibhatla R. IMFT 25-nm MLC NAND: technology scaling barrier broken. In: *EE Times News and Analysis*, March 22, 2010
- 5 Darling P. Intel, Micron First to Sample 3-Bit-Per-Cell NAND Flash Memory on Industry-Leading 25-Nanometer Silicon Processing Technology. Micron Technology Inc. Press Release, August 17, 2010
- 6 Lue H T, Hsu T H, Lai S C, et al. Scaling evaluation of BE-SONOS NAND flash beyond 20 nm. In: *VLSI Symposia on Technology*, session 12-1, 2008. 116–117
- 7 Lee C H, Choi K I, Cho M K, et al. A novel SONOS structure of $\text{SiO}_2/\text{SiN}/\text{Al}_2\text{O}_3$ with TaN metal gate for multi-giga bit flash memories. In: *International Electron Device Meeting (IEDM)*, session 26-5, 2003. 613–616
- 8 Lai S C, Lue H T, Liao C W, et al. An oxide-buffered BE-MANOS charge-trapping device and the role of Al_2O_3 . In: *NVSMW-ICMTD*, 2008. 101–102
- 9 Lue H T, Wang S Y, Lai E K, et al. BE-SONOS: A bandgap engineered SONOS with excellent performance and reliability. In: *International Electron Device Meeting (IEDM)*, session 22-3, 2005. 555–558,
- 10 Lue H T, Wang S Y, Hsiao Y H, et al. Reliability model of bandgap engineered SONOS (BE-SONOS). In: *International Electron Device Meeting (IEDM)*, session 18-5, 2006. 495–498
- 11 Lai S C, Lue H T, Yang M J, et al. MA BE-SONOS: A bandgap engineered SONOS using metal gate and Al_2O_3 blocking layer to overcome erase saturation. In: *IEEE Non-Volatile Semiconductor Memory Workshop (NVSMW)*, 2007. 88–89
- 12 Lue H T, Lai S C, Hsu T H, et al. Modeling of barrier engineered charge-trapping NAND flash (BE CTNF) devices. *IEEE Trans Device Mater Reliab (TDMR)*, 2010, 10: 222–232
- 13 Lue H T, Pan J F, Wang S Y, et al. Chip-level reliability study of barrier engineered (BE) floating gate (FG) flash memory devices. In: *International Reliability Physics Symposium (IRPS)*, session 5D-2, 2010. 627–633
- 14 Hsieh C C, Lue H T, Chang K P, et al. A novel BE-SONOS NAND flash using non-cut trapping layer with superb reliability. In: *International Electron Device Meeting (IEDM)*, session 5, 2010. to be published
- 15 Lue H T, Hsu T H, Lai S C, et al. Study of electron and hole injection statistics of BE-SONOS NAND flash. In: *International Memory Workshop (IMW)*, 2010. 92–95
- 16 Compagnoni C M, Spinelli A S, Gusmeroli R, et al. Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics. In: *IEEE T-ED*, 2008. 2695–2702
- 17 Hsiao Y H, Lue H T, Hsieh K Y, et al. A study of stored charge interference and fringing field effects in sub-30 nm charge-trapping NAND flash. In: *International Memory Workshop (IMW)*, 2009. 34–35
- 18 Prall K, Parat K. 25 nm 64 Gb MLC NAND technology and scaling challenges. In: *International Electron Device Meeting (IEDM)*, session 5-5, 2010. 98–101
- 19 Lai E K, Lue H T, Hsiao Y H, et al. A multi-layer stackable thin-film transistor (TFT) NAND-type flash memory. In: *International Electron Device Meeting (IEDM)*, session 2-4, 2006. 41–44
- 20 Jung S M, Jang J, Cho W, et al. Three dimensionally stacked NAND flash memory technology using stacking single crystal Si layers on ILD and TANOS structure for beyond 30 nm node. In: *International Electron Device Meeting (IEDM)*, 2006. 37–40

- 21 Tanaka H, Kido M, Yahashi K, et al. Bit cost scalable technology with punch and plug process for ultra high density flash memory. In: VLSI Symposia on Technology, 2007. 14–15
- 22 Katsumata R, Kito M, Fukuzumi Y, et al. Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices. In: VLSI Symposia on Technology, 2009. 136–137
- 23 Jang J, Kim H S, Cho W, et al. Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND flash memory. In: VLSI Symposia on Technology, 2009. 192–193
- 24 Kim J, Hong A J, Kim S M, et al. Novel vertical-stacked-array-transistor (VSAT) for ultra-high-density and cost-effective NAND flash memory devices and SSD (solid state drive). In: VLSI Symposia on Technology, 2009. 186–187
- 25 Kim W, Choi S, Sung J, et al. Multi-layered vertical gate NAND flash overcoming stacking limit for terabit density storage. In: VLSI Symposia on Technology, 2009. 188–189
- 26 Lue H T, Hsu T H, Hsiao Y H, et al. A highly scalable 8-layer 3D vertical-gate (VG) TFT NAND flash using junction-free buried channel BE-SONOS device. In: VLSI Symposia on Technology, 2010. 131–132
- 27 Hsu T H, Lue H T, Hsieh C C, et al. Study of sub-30 nm thin film transistor (TFT) charge-trapping (CT) devices for 3D NAND flash application. In: International Electron Device Meeting (IEDM), session 27-4, 2009. 629–632
- 28 Hubert A, Nowak E, Tachi K, et al. A stacked SONOS technology, up to 4 levels and 6 nm crystalline nanowires, with gate-all-around or independent gates (Φ -Flash), suitable for full 3D integration. In: International Electron Device Meeting (IEDM), 2009. 637–640
- 29 Chen Y C, Rettner C T, Raoux S, et al. Ultra-thin phase-change bridge memory device using GeSb. In: International Electron Device Meeting (IEDM), session 30-3, 2006. 777–780
- 30 Oh J H, Park J H, Lim Y S, et al. Full integration of highly manufacturable 512 Mb PRAM based on 90 nm technology. In: International Electron Device Meeting (IEDM), session 2-6, 2006. 49–52
- 31 Servalli G. A 45 nm generation phase change memory technology. In: International Electron Device Meeting (IEDM), session 5-7, 2009. 113–116
- 32 Kittel C. Introduction to Solid State Physics. 7th ed. New York: Wiley, 1996. 393–398
- 33 Ishiwara H. Current status and prospects of ferroelectric memories. In: International Electron Device Meeting (IEDM), session 33-1, 2001. 725–728
- 34 Takashima D, Kunishima I. High-density chain ferroelectric random access memory (chain FRAM). *J Solid-State Circ*, 1998. 33: 787–792
- 35 Koo J M, Seo B S, Kim S, et al. Fabrication of 3D trench PZT capacitors for 256 Mbit FRAM device application. In: International Electron Device Meeting (IEDM), session 14-2, 2005. 351–354
- 36 Kanaya H, Tomioka K, Matsushita T, et al. A $0.602 \mu\text{m}^2$ nestled chain cell structure formed by one mask etching process for 64 Mbit FeRAM. In: VLSI Symposia on Technology, session 15-3, 2004. 150–151
- 37 Shimojo Y, Konno A, Nishimura J, et al. High-density and high-speed 128 Mb chain FeRAMTM with SDRAM-compatible DDR2 interface. In: VLSI Symposia on Technology, session 11B-1, 2009. 218–219
- 38 Han J P, Ma T P. Ferroelectric-gate transistor as a capacitor-less DRAM cell. *Integrat Ferroelectr*, 1999, 27: 1053–1062
- 39 Yuasa S, Nagahama T, Fukushima A, et al. Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. *Nat Mater*, 2004, 3: 868–871
- 40 Durlam M, Addie D, Akerman J, et al. A $0.18 \mu\text{m}$ 4 Mb toggling MRAM. In: International Electron Device Meeting (IEDM), session 34-6, 2003. 995–997
- 41 Prejbeanu I L, Kula W, Ounadjela K, et al. Thermally assisted switching in exchange-biased storage layer magnetic tunnel junctions. *IEEE Trans Magnet*, 2004, 40: 2625–2627
- 42 Lee Y M, Yoshida C, Tsunoda K, et al. Highly scalable STT-MRAM with MTJs of top-pinned structure in 1T/1MTJ cell. In: VLSI Symposia on Technology, session 5-2, 2010. 49–50
- 43 Ishigaki T, Kawahara T, Takemura R, et al. A multi-level-cell spin-transfer torque memory with series-stacked magnetotunnel junctions. In: VLSI Symposia on Technology, session 5-1, 2010. 47–48
- 44 Huai Y. Spin-transfer torque MRAM (STT-MRAM): challenges and prospects. *AAPPS Bull*, 2008, 18: 33–40
- 45 Hönigschmid H, Angerbauer M, Dietrich S, et al. A non-volatile 2 Mbit CBRAM memory core featuring advanced read and program control. In: VLSI Symposia on Circuits, session 13-2, 2006. 110–111
- 46 Aratani K, Ohba K, Mizuguchi T, et al. A novel resistance memory with high scalability and nanosecond switching. In: International Electron Device Meeting (IEDM), session 30-5, 2007. 783–786
- 47 Waser R. Electrochemical and thermochemical memories. In: International Electron Device Meeting (IEDM), session 12-1, 2008. 289–292
- 48 Xu N, Gao B, Liu L F, et al. A unified physical model of switching behavior in oxide-based RRAM. In: VLSI Symposia on Technology, session 10-3, 2008. 100–101
- 49 Chua L O. Memristor-the missing circuit element. *IEEE Trans Circ Theory*, 1971, 18: 507–519
- 50 Strukov D B, Snider G S, Stewart D R, et al. The missing memristor found. *Nature*, 2008, 453: 80–83
- 51 Borghetti J, Snider G S, Kuekes P J, et al. ‘Memristive’ switches enable ‘stateful’ logic operations via material implication. *Nature*, 2010, 464: 873–876

- 52 Kozicki M N, Balakrishnan M, Gopalan C, et al. Programmable metallization cell memory based on Ag-Ge-S and Cu-Ge-S solid electrolytes. In: 2005 Non-volatile Memory Technology Symposium, 2005. 83–89
- 53 Sakamoto T, Tada M, Banno N, et al. Nonvolatile solid-electrolyte switch embedded into Cu interconnect. In: VLSI Symposia on Technology, session 6B-4, 2009. 130–131
- 54 Lin Y Y, Lee F M, Chen Y C, et al. A novel TiTe buffered Cu-GeSbTe/SiO₂ electrochemical resistive memory (ReRAM). In: VLSI Symposia on Technology, session 8-4, 2010. 91–92
- 55 Da Silva J L F, Wei S H, Zhou J, et al. Stability and electronic structures of Cu_xTe. *Appl Phys Lett*, 2007, 91: 091902–091904
- 56 Cabral C, Chen K N, Krusin-Elbaum L, et al. Irreversible modification of Ge₂Sb₂Te₅ phase change material by nanometer-thin Ti adhesion layers in a device-compatible stack. *Appl Phys Lett*, 2007, 90: 051908–051910
- 57 Chen A, Haddad S, Wu Y C J, et al. Non-volatile resistive switching for advanced memory applications. In: International Electron Device Meeting (IEDM), session 5-5, 2005. 746–749
- 58 Wang M, Luo W J, Wang Y L, et al. A novel Cu_xSi_yO resistive memory in logic technology with excellent data retention and resistance distribution for embedded applications. In: VLSI Symposia on Technology, session 8-3, 2010. 89–90
- 59 Sakotsubo Y, Terai M, Kotsuji S, et al. A new approach for improving operating margin of unipolar ReRAM. In: VLSI Symposia on Technology, session 8-2, 2010. 87–88
- 60 Baek I G, Lee M S, Seo S, et al. Highly scalable non-volatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses. In: International Electron Device Meeting (IEDM), session 23-6, 2004. 587–590
- 61 Lee H Y, Chen P S, Wu T Y, et al. Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO₂ based RRAM. In: International Electron Device Meeting (IEDM), session 12-3, 2008. 297–300
- 62 Ho C H, Lai E K, Lee M D, et al. A highly reliable self-aligned graded oxide WO_x resistance memory: conduction mechanisms and reliability. In: VLSI Symposia on Technology, session 12B-2, 2007. 228–229
- 63 Chien W C, Chen Y C, Lai E K, et al. Unipolar switching behaviors of RTO WO_x RRAM. *IEEE Electr Device Lett*, 2010, 31: 126–128
- 64 Chen Y C, Chien W C, Lin Y Y, et al. Cu-based and WO_x-based resistive switching memories (ReRAMs) for embedded and stand-alone applications. In: International Conference on Solid-State and Integrated-Circuit Technology, 2010. invited
- 65 Chien W C, Chen Y C, Lee F M, et al. A novel Ni/WOX/W ReRAM with excellent retention and low switching current. In: 2010 International Conference on Solid State Devices and Materials (SSDM), 2010. 1104–1105
- 66 Lai E K, Chien W C, Chen Y C, et al. Tungsten oxide resistive memory using rapid thermal oxidation of tungsten plugs. *Japan J Appl Phys*, 2010, 49: 04DD17–04DD17-4
- 67 Chien W C, Chen Y C, Lai E K, et al. High-speed multilevel resistive RAM using RTO WO_x. In: 2009 International Conference on Solid State Devices and Materials (SSDM), session G-7-3, 2009. 1206–1207
- 68 Lee M J, Lee C B, Kim S, et al. Stack friendly all-oxide 3D RRAM using GaInZnO peripheral TFT realized over glass substrates. In: International Electron Device Meeting (IEDM), session 4.4, 2008. 85–88
- 69 Chevallier C J, Chang H S, Lim S F, et al. A 0.13 μm 64 Mb multi-layered conductive metal-oxide memory. In: 2010 International Solid-State Circuit Conference Digest (ISSCC), session 14-3, 2010. 260–261