Multilevel Cell Storage and Resistance Variability in Resistive Random Access Memory

Amit Prakash¹ / Hyunsang Hwang¹

¹ Department of Material Science and Engineering, Pohang University of Science and Technology (POSTECH), 77 Cheongam-Ro, Nam-Gu, Pohang, Gyeongbuk, Korea 790-784, Tel: +82-54-279-5123, Fax: +82-54-279-5122 E-mail: amit.knp02@gmail.com, hwanghs@postech.ac.kr

Abstract:
Multilevel per cell (MLC) storage in resistive random access memory (ReRAM) is attractive in achieving high-density and low-cost memory and will be required in future. In this chapter, MLC storage and resistance variability of multilevel in ReRAM are discussed. Different MLC operation schemes with their physical mechanisms and a comprehensive analysis of resistance variability have been provided. Various factors that can induce variability and their effect on the resistance margin between the multiple resistance levels are assessed. The reliability characteristics and the impact on MLC storage have also been assessed.

Keywords: Emerging memory, non-volatile storage, resistive random access memory (ReRAM), multilevel cell storage (MLC), variability, conductive filament, high density, storage, memory margin, MLC reliability

DOI: 10.1515/psr-2016-0010

1 Introduction
Semiconductor memory, which is an electronic device used for data storage, becomes a key component in today’s electronic systems. The dominant memory technologies of the present time are Dynamic Random Access Memory (DRAM) and Flash. Both these memories store data as electronic charge and are known to suffer from further scaling issues [1]–[2]. The DRAM, which is a volatile memory, offers extremely long endurance (~10¹⁴ program/erase cycles) and fast (<10 ns) switching operation [3]. On the other hand, Flash is a nonvolatile memory, but possesses limited endurance of typically 10⁴–10⁵ cycles and slow program/erase time of (μs up to ms) [2]–[4].

To overcome the scaling and operational limitations of these conventional memories, various next-generation memory technologies have been proposed and are being intensively investigated. Some of these memories are: Ferroelectric Random Access Memory (FeRAM) [4]–[6], Spin Transfer Torque-Magnetic Random Access Memory (STTMRAM) [7]–[9], Phase-change Random Access Memory (PRAM) [10]–[12], and Resistive Random Access Memory (ReRAM) [13]–[22]. One of the basic differences between these next-generation memories and conventional memories is in the way these two types of memories store data; the former stores the data as electronic charge whereas the latter as resistance. Among other next-generation memories, ReRAM is more promising owing to its simple Metal-Insulator-Metal (MIM) structure, excellent scalability (<10 nm), fast switching speed (<10 ns), low power operation, three-dimensional (3D) stackability and good complementary metal-oxide-semiconductor (CMOS) compatibility [13]–[22].

2 Resistive random access memory (ReRAM)

Resistive switching was first observed by Hickmott in 1962 in binary oxides [23], but it was in early 2000s only when the resistive switching effect caught huge interest triggered by the search of an alternative memory technology ([15]24) [27]. Generally, a ReRAM device consists of two electrically conducting metal electrodes and a sandwiched insulating switching layer making it a two-terminal passive device. Such a typical MIM structured ReRAM device is schematically shown in Figure 1(a). It was found that the resistance of such a device can reversely be switched between the high and low states simply by applying external bias across the MIM stack [13, 24, 25]. The electrode on which the external stimulus is applied is referred to as the top electrode (TE) and the other electrode, kept at electrical ground is called as the bottom electrode (BE) [Figure 1(a)]. Usually, prior
to the reversible resistance switching, an electroforming (or forming) process is generally required in which a relatively higher bias is applied across a pristine device. After the forming process, the initial high resistance of the pristine device irreversibly changes to the low resistance [13]–[28]. A more elaborate discussion on the forming process and its mechanism can be found in the references [28]–[29].

Figure 1 (a) Schematic diagram showing typical metal-insulator-metal (MIM) structure of ReRAM with electrical biasing. (b) Schematic illustration of bipolar switching characteristics in ReRAM. For the bipolar switching, “set” and “reset” processes occur at different polarity.

2.1 Resistive switching modes

The resistance of a ReRAM device can be switched between the high resistance state (HRS) and the low resistance state (LRS) by two different ways depending on the applied voltage polarity on the TE. When both the resistance states of HRS and LRS can be obtained by applying either positive voltage or negative voltage on TE, it is called unipolar mode switching [13]. On the other hand, when both positive as well as negative polarity voltage is needed to switch the device between HRS and LRS, it is called bipolar mode switching as schematically shown in Figure 1(b) [13]–[21]. Although unipolar switching is beneficial in terms of requiring only single polarity voltage, which may minimize the circuit complexity, the resistive switching performance of unipolar device is commonly poorer compared to that of bipolar device. Therefore, in this chapter, we will discuss ReRAM with only bipolar mode switching. When a positive (or negative) sweep voltage is applied on TE, the current gradually increases until it reaches to a certain voltage at which the current abruptly increases and attains a predefined compliance current ($I_{C}$) as shown in Figure 1(b). At this stage, the resistance of the device switches from HRS to LRS. This is called set process and the voltage at which the current suddenly increases is termed as set voltage ($V_{set}$). The $I_{C}$ is generally needed to avoid the permanent breakdown of the device. The device remains in the LRS state until applied negative (or positive) sweep voltage reaches beyond a certain value at which the current starts to decrease [Figure 1(b)]. This voltage is called reset voltage ($V_{reset}$) and the process is termed as reset process. After the reset process, the device switches to the HRS. As both the memory states of LRS and HRS retain their respective states even after the removal of applied voltage, ReRAM is a nonvolatile memory.

2.2 Resistive switching mechanism

The switching mechanism of ReRAM can be broadly classified as (i) filamentary type and (ii) interface type and is schematically shown in Figure 2(a) and (b), respectively [21]. In the filamentary model, the resistance switching occurs owing to the formation and rupture of localized conductive filament (CF) in the sandwiched insulating layer also termed as switching layer upon the application of appropriate external electrical stimulus. The filamentary mechanism is schematically shown in Figure 2(a) [13, 14, 21]. As the CF can generate anywhere in the switching layer during bias application, the resistance switching in ReRAM is intrinsically random or stochastic. It is generally required that the switching material, especially the oxide, should have sub-stoichiometric composition in order to obtain stable resistance switching [21]. During forming/set operation, the filamentary paths formed due to the anion (oxygen ion)/cation (Cu$^{+}$ or Ag$^{+}$) migration depend on the electrode and switching materials used and rupture during the reset process [Figure 2(a)]. The device in which switching occurs due to metallic (Cu or Ag) filament formation is called as conductive bridging random access memory (CBRAM) [14]. Electrochemical migration of oxygen ions/cations and redox reaction near the metal/oxide interface is widely accepted as the possible mechanism of filament formation and rupture [14]–[30]. A more detailed discussion on ReRAM switching mechanism can be found in refs. [13, 14, 29, 31]. However,
accurate universal switching model for ReRAM is still to be proposed. In addition, although studies involving high resolution transmission electron microscopy showed the CFs in different systems [32]–[33], their in-situ clear visualization in the switching material have yet to be achieved. The lack of clear understanding of the ReRAM switching mechanism poses one of the biggest hindrances in its advancement towards commercial availability. On the other hand, in the interface type mechanism, the switching occurs at the interface of the metal and switching material owing to the movement of oxygen vacancies/ions or trapping/de-trapping of electrons or holes under the applied electric field [Figure 2(b)] [21].

![Figure 2](image-url) (a) Schematic illustration of filamentary switching mechanism (b) An interface-type switching model [21].

One of the main differences between the filamentary switching and interface type switching is that the later occurs at the whole area of the device in contrast to the former where switching is highly localized. This can be identified by measuring the device area dependence of both the resistance states of LRS and HRS of a resistive memory device. If, in general, both LRS and HRS vary with device area, the switching is interface type. In contrast, in filamentary switching, though HRS increases with decreasing device area, LRS remains unchanged with device area variation. Due to the highly localized switching, the filamentary ReRAM shows excellent scalability potential and the devices with <10 nm physical size have already been demonstrated [34]–[36]. A comprehensive discussion on ReRAM can be found in the refs. [13, 14, 19, 21, 22, 29, 37].

Although ReRAM shows excellent memory traits, it suffers from some drawbacks, such as its switching mechanism is not fully understood and it shows resistance variability (temporal as well as spatial) in both LRS and HRS [38]. Figure 3(a) shows all the current–voltage (I–V) curves of 20 consecutive direct current switching cycles obtained from a ReRAM device with Ta/TaOx/Pt structure. As can be seen, the resistance of both LRS and HRS varies with the switching cycles. Owing to this variation, the overall memory window (HRS/LRS) decreases, which can be seen more clearly in the cumulative probability plot shown in Figure 3(b).

![Figure 3](image-url) (a) Current–voltage (I–V) curves of all 20 consecutive switching cycles. (b) Cumulative cycle-to-cycle current distribution of Ta/TaOx/Pt ReRAM device. Both LRS and HRS show resistance variability.

The observed variability is attributed to the intrinsic randomness in the CF formation and rupture during the switching events [39]–[40]. In other words, exactly same amount of filament cannot be formed or ruptured due to the stochastic nature of ReRAM switching mechanism, resulting in a cycle-to-cycle resistance variability. A more detailed discussion on resistance variability will be provided later in this chapter.
3 Multilevel per cell (MLC) storage

In order to reduce cost, the density of memory devices should be increased, which in turn minimizes the use of Si substrate area. One of the most obvious ways to increase the storage density is to decrease the physical size of the device to nanoscale dimensions. In this regard, ReRAM is very attractive as it shows excellent scalability potential with the architectural feasibility of the smallest possible cell area of $4F^2$ (F: feature size of the fabrication technology) in two-dimensional layout [41]. ReRAM with <10 nm device size has already been reported with excellent memory performance [34]–[36]. Although the density can be increased by reducing the device size, this method requires complex experimental processes and is limited by patterning techniques as well. The other interesting approach is to stack the devices three-dimensionally (3D), that is, on top of each other vertically. For ReRAM, two kinds of feasible architectures of “crossbar” and “vertical ReRAM” are suggested [41]–[42]. Though both the architectures are appealing for future 3D integration, they also have patterning and integration complexities. Another alternative and simpler way to increase storage density is to use MLC storage technology in which more than one bit per cell can be stored without further decreasing the physical device size [43]–[45]. The dependence of LRS resistance ($R_{LRS}$) and Ireset on $I_{C}$ is shown in Figure 4(b) where $R_{LRS}$ and $I_{reset}$ are plotted as a function of $I_{C}$. Generally, the $R_{LRS}$ shows the universal dependence on $I_{C}$ when reset voltage is constant and can be well-fitted by an equation of the form $R_{LRS} = C/I_{C}$, where $C$ is a constant, with the slope of “−1” [39]–[57]. However, different values of the slope other than the universal value of “−1” are also reported especially in the devices in which the reset voltage also varies with $I_{C}$ as in Figure 4(a) [50]–[52]. In contrast to the inverse dependence of $R_{LRS}$ on $I_{C}$, the $I_{reset}$ depends linearly on $I_{C}$ as can be seen in Figure 4(b).

4 MLC modes in ReRAM

One of the important traits of ReRAM, which makes it useful for high density application, is its MLC behavior. There are mainly three ways to obtain MLC characteristic: (i) changing compliance current, (ii) controlling reset voltage, and (iii) varying pulse width of program/erase operation.

4.1 MLC by changing compliance current

MLC characteristic in a ReRAM device with 1-ReRAM (1R) cell configuration can be obtained by changing the current compliance ($I_{C}$) during “set” operation [44, 50–55]. In 1-Transistor 1-ReRAM (1T-1R) cell configuration, which is more practically viable, the current during “set” operation can be controlled by varying the applied voltage at the gate of the transistor [50]–[54]. The typical MLC I–V curves of a ReRAM device in TiN/Ti/HfO$_x$/TiN structure in 1R cell configuration are shown in Figure 4(a) [56]. When the $I_{C}$ is increased sequentially from 100 to 250 and 500 μA, the respective current of LRS ($I_{LRS}$) also increases resulting in three different levels of LRS whereas the HRS current ($I_{HRS}$) remains same for all the LRS levels. These three distinct LRS levels with the same HRS, leading to a total of 4-resistance levels can be used in 2-bit per cell storage and can enhance the storage density up to 2x higher as compared to a single level cell with the same Si die area. Recently, Prakash et al. have demonstrated 3-bit per cell storage by stack engineering in a TaO$_x$-based ReRAM [52]. Furthermore, it is observed that the maximum reset current ($I_{reset}$) also increases with $I_{C}$ whereas the set voltage almost remains unchanged [44]–[52]. The dependence of LRS resistance ($R_{LRS}$) and $I_{reset}$ on $I_{C}$ is shown in Figure 4(b) where $R_{LRS}$ and $I_{reset}$ were plotted as a function of $I_{C}$. Generally, the $R_{LRS}$ shows the universal dependence on $I_{C}$ when reset voltage is constant and can be well-fitted by an equation of the form $R_{LRS} = C/I_{C}$, where $C$ is a constant, with the slope of “−1” [39]–[57]. However, different values of the slope other than the universal value of “−1” are also reported especially in the devices in which the reset voltage also varies with $I_{C}$ as in Figure 4(a) [50]–[52]. In contrast to the inverse dependence of $R_{LRS}$ on $I_{C}$, the $I_{reset}$ depends linearly on $I_{C}$ as can be seen in Figure 4(b).
The mechanism of MLC in $I_C$ control mode can be attributed to the formation and subsequent lateral widening of the CF with increasing $I_C$ as schematically shown in Figure 5 [52]–[53]. As the CF size (or diameter) increases, its resistance becomes smaller and hence, results multiple LRS levels. This argument is well-supported by the fact that the $I_{\text{reset}}$ also increases with $I_C$ because it needs higher power to break the CF having larger diameter. Although MLC behavior in $I_C$ control mode is relatively easier to be obtained, it may have limited applications especially in passive cross-point array architecture owing to the difficulty in limiting the current by the “on chip” circuit.

![Figure 5 Schematic illustration of switching mechanism of MLC operation by varying $I_C$. The increase in the size (diameter) of the conductive filament (CF) owing to accumulation of more oxygen vacancies with increasing $I_C$ results in multiple resistance levels of LRS.](image)

### 4.2 MLC by controlling reset voltage

The MLC characteristics can also be obtained by controlling the applied maximum reset voltage for “reset” operation [43, 44, 55, 58–61]. Figure 6(a) shows the typically observed I–V curves in semilog plot obtained from a HfO$_x$-based ReRAM for three different reset voltages of $-0.9$, $-1.1$, and $-1.3$ V. As the applied voltage for the “reset” operation increases, the HRS current ($I_{\text{HRS}}$) decreases, leading to different resistance levels of HRS with same LRS resistance as shown in Figure 6(b). Further, in addition to HRS resistance, the “set” voltage ($V_{\text{set}}$) also increases with increasing reset voltage whereas $I_{\text{reset}}$ remains almost constant [44]–[58]. Some studies have shown the possibility of 3-bit per cell storage in this mode of MLC operation by using modified program-verify technique already being used in Flash memory [59]–[61].

![Figure 6 (a) Typical MLC current–voltage curves obtained by limiting the current compliance ($I_C$). (b) Dependence of LRS resistance ($R_{\text{LRS}}$) and reset current ($I_{\text{reset}}$) on ($I_C$) [56].](image)
Figure 6 (a) Typical MLC current-voltage curves obtained by varying the reset voltage. (b) Resistance of all the three distinct levels of HRS and LRS. $R_{\text{HRS}}$ increases with reset voltage.

The decrease in $I_{\text{HRS}}$ current is attributed to the increase in the gap between CF tip and metal electrode when higher reset voltage is applied as schematically illustrated in Figure 7 [43, 58, 61, 62]. It is also argued that the observed change in the resistance of HRS can be due to CF thinning (decrease in CF size) with increasing reset voltage, especially in the devices that show gradual reset current change, rather than abrupt during “reset” operation [63]. This mode of MLC operation is more practically feasible especially for passive cross-point array in terms of comparatively less circuit complexity.

Figure 7 Schematic illustration of switching mechanism of MLC operation by controlling reset voltage. The increase in the gap between CF tip and BE with increasing reset voltage results in multiple resistance levels of HRS.

4.3 MLC by changing program/erase pulse width

Another suggested method to obtain MLC characteristics is to vary the program/erase pulse width while keeping the pulse amplitude constant [58]–[64]. Three distinct HRS resistance levels were successfully obtained by exponentially changing the reset pulse width from 50 ns to 5 $\mu$s in a HfO$_x$-based ReRAM [58]. The origin of the HRS resistance modulation and its equivalence with the reset voltage control scheme may be attributed to the universal voltage-time relationship in the switching dynamics of ReRAM where the ions migration velocity exhibit the hyper-sinusoidal dependence on the applied electric field [58]. Although it is relatively easier to generate the program/erase pulses with varying widths, this scheme has the disadvantage of being energy inefficient [58]. The comparison of the transient responses between reset pulse amplitude and pulse width control confirmed the higher energy consumption for the later scheme owing to the unwanted energy dissipation in the switching material as thermal energy [58].

5 Resistance variability and MLC operation

Although the multiple resistance levels can be easily obtained in ReRAM by the above-mentioned methods, the successful implementation of the MLC technology mainly depends on the ability to precisely control the resistance margin between the two resistance levels. There exist various factors that can hinder the MLC operation in ReRAM by degrading the resistance margin and eventually leading to its failure [48]–[60]. Therefore, enhanced understanding of the origin of these factors is required along with the search of the feasible approaches.
to minimize the effects. We will discuss about these factors and their effect on MLC margin more elaborately in the following subsections.

5.1 Cycle-to-cycle variability

One of the most critical factors that can degrade the margin between the two resistance levels and that can result in an erroneous MLC operation, especially in ReRAM, is the cycle-to-cycle (or temporal) switching variability [31, 48, 65–67]. Figure 8(a) shows the multiple switching I–V curves obtained from a HfO\textsubscript{x} based device when operated in MLC mode by changing I\textsubscript{C} [56]. The resistance distribution of all the 100 consecutive switching cycles for all the four resistance levels (three LRS and one HRS levels) are also shown in Figure 8(b) [56]. The resistance margin is reduced owing to the cycle-to-cycle switching variability. The origin of the cycle-to-cycle variability at a particular switching current is mainly attributed to the dynamic change in the number of constituent oxygen vacancy defects present in the CF with each switching event because of the intrinsic stochastic nature of CF formation and rupture during switching in ReRAM [47, 66, 67]. This compositional variation in the conductive channel leads to the conductivity fluctuation. More fundamentally, it can be understood in terms of having different activation energy of defect migration due to the random location as well as different surroundings of the defects within the conductive channel [66]. In addition, the variability becomes more severe when the I\textsubscript{C} is reduced [65]–[66]. It is observed that not only the variation in the LRS resistance increases, but also the variations in I\textsubscript{reset} and reset voltage increase when the I\textsubscript{C} is decreased as shown in Figure 9 where the statistical analysis of R\textsubscript{LRS}, I\textsubscript{reset}, and reset voltage as a function of I\textsubscript{C} is presented [56]. The I\textsubscript{C} dependent resistance variability behavior in ReRAM can be related to the change in the total number of oxygen vacancy defects (N) contributing to CF formation with varying I\textsubscript{C} [39]. Figure 10 shows the resistance–voltage (R–V) curves of LRS corresponding to the I–V curves of positive bias shown in Figure 8 and the schematic illustration of the CF defects change with I\textsubscript{C}. The smaller relative LRS resistance spread (ratio of standard deviation (σ) and average resistance (μ)) at higher I\textsubscript{C} (see Figure 9 and Figure 10) is attributed to the increased number of defects present in the CF rendering well-defined path for current conduction. Balatti et al. have theoretically shown a quantitative relation between relative LRS resistance spread and I\textsubscript{C} from Poisson statistics, assuming ohmic conduction in CF [39]. It is demonstrated that (σ/μ)\textsubscript{LRS} = (N)^{-0.5} is proportional to (I\textsubscript{C})^{-0.5} [39]. Various methods including materials engineering, different electrical methods and device operation conditions modification have been suggested to improve the cycle-to-cycle resistance variability in ReRAM [20, 62, 68–72].

![Figure 8](image-url)

**Figure 8** (a) Multiple MLC I–V curves of HfO\textsubscript{x}-based ReRAM device. (b) Resistance distribution of 100 switching cycles of all the three distinct LRS levels with same HRS. The cycle-to-cycle resistance variability increases with reducing I\textsubscript{C} [56].
5.2 Device-to-device variability

In addition to the cycle-to-cycle variability, ReRAM also exhibits device-to-device or spatial non uniformity in the switching parameters which additionally degrades the usable resistance margin between the memory states and hence, can hinder the MLC operation [15, 16, 38, 69, 73–76]. The device-to-device variability can originate from the nonuniformities in the fabrication processes such as variation in the switching film thickness, surface roughness of the electrodes, etching damages, etc., as well as the lack of precise control over the defect generation and filament formation during the electroformation step of a pristine device [73, 74, 77, 78]. Lee et al. have demonstrated significant improvement in the resistance distribution in an array by adopting chemical mechanical polish to smoothen the bottom electrode in HfO₂-based device [73]. Chen et al. suggested the application of ramping voltage instead of constant voltage across the gate in 1T–1R cell to improve device-to-device variations [15]. Recently, Hayakawa et al. have reported improved resistance distribution in a nanoscale TaOₓ-based ReRAM by developing the fabrication process technologies, such as low-damage etching, cell side oxidation and proper cell encapsulation [78]. On the other hand, one of the main origins of the electroformation-induced resistance variation is attributed to the overshoot current due to the parasitic effects[69, 76, 77]. The overshoot current that can be different for different devices leads to the uncontrolled defects generation and subsequent CF formation in the switching layer resulting in device-to-device resistance variation. To minimize this, various methods, such as constant voltage forming [77] and forming at higher temperature, so-called hot forming [69], are advocated, and improvement in the resistance distribution is reported.
5.3 Effect of operation temperature

In ReRAM, it was observed that depending on the programmed resistance state (LRS or HRS) and the operation current (low or high), the resistance of the different multilevel varies differently with changing the operation (ambient) temperature [45, 63, 79, 80]. The resistance of HRS generally decreases when the operation temperature is increased, whereas the LRS resistance can increase or decrease depending on the programming current [45, 63, 80]. Goux et al. have shown that the LRS resistance programmed at a high current (>1 mA) in an HfO$_2$-based device increases with temperature, showing metallic-like behavior with negative activation energy [63], whereas Wei et al. observed that the LRS resistance of a TaO$_x$-based device programmed at a relatively lower current (<100 μA) decreased with increasing temperature [79]. Although the resistance of HRS decreases with increasing operation temperature, the degree (or slope) of the resistance change is found to be different for different HRS levels [45–63]. Owing to the different behavior of the multiple resistance levels on operation temperature, the resistance margin between the resistance levels deteriorates (with respect to room temperature) when the operation temperature is raised as shown in Figure 11, which may limit the temperature range of MLC operation. In addition, as the ambient temperature of the device can be increased during its operation, this can result in an erroneous MLC operation.

![Figure 11](image1.png)

**Figure 11** Multilevel resistance dependence on temperature. The HRS resistances decrease with increasing temperature. The slope of resistance changes with temperature, as it is different for different HRS levels (state 01, 10, and 11), leading to smaller resistance margin at higher operation temperature [45].

5.4 Effect of random telegraph noise (RTN)

Random telegraph noise (RTN) that can cause severe fluctuations in the read current during read operation of a programmed resistance level in a ReRAM cell leads to a reduction in the detectable margin between the resistance levels [43, 81–84]. The origin of the RTN in ReRAM is attributed to the capture and emission (trap and detraps) of electrons in the trap (more probably oxygen vacancy trap) near the filament in LRS state and in the tunneling gap in the HRS state [43, 81, 83]. It is also reported that the RTN can be governed by the oxygen vacancy perturbation in the conducting channel during switching [84]. A typical two-step RTN observed in a Ta/ TaO$_x$/Pt stacked device is shown in Figure 12. Ielmini et al. have shown a universal relation between the signal-to-noise ratio or relative amplitude ($\Delta R/R$) of the RTN and the average resistance (R) of the memory state based on the size-dependent depletion of the carriers [81]. The relative amplitude ($\Delta R/R$) increases with average resistance of the programmed memory state. In other words, the RTN amplitude increases with decreasing the operation current and is higher for those multiple resistance levels that are programmed to the higher resistance. In addition, the HRS levels show higher RTN amplitude than the LRS levels [43]. Therefore, additional resistance margin should be ensured between the two resistance levels especially with higher resistance in order to avoid the RTN effect and to achieve a reliable MLC operation.
5.5 Interstate switching variability

It was observed that the resistance of the memory levels can also vary when a device operating in the MLC mode is switched from one resistance level to the other especially from the highest resistance level to the lowest and vice versa. This can further reduce the margin between the resistance levels for the MLC operation. Figure 13 shows that the two values of LRS resistance corresponding to $I_C$ of 100 $\mu$A before and after increasing the $I_C$ to 200 $\mu$A are not the same [43]. The variation can be due to the failure to form the same amount of filament to that of the previous switching event of the same resistance level in case of $I_C$-controlled MLC mode or to generate exactly the same gap between the CF tip and the electrode in case of reset voltage control MLC scheme.

As the above discussed factors can degrade the usable resistance margin among the various multiple resistance levels as schematically shown in Figure 14 and hence, can limit the MLC implementation, proper care should be taken while designing the MLC storage in ReRAM.
6 Reliability of multiple resistance levels

6.1 Retention characteristics

ReRAM is a nonvolatile memory because it can retain the memory state (LRS or HRS) after being programmed or erased. The nonvolatile behavior is due to the fact that in ReRAM, once the filament is formed or ruptured, it remains as it is even when the electric field is removed. The retention time of LRS and HRS can be evaluated by monitoring the resistance of the respective resistance state with time by applying a small constant read voltage pulse with much smaller amplitude than that of the program/erase voltage pulse as shown in Figure 15. The retention loss is mainly due to the out diffusion/annihilation of the oxygen vacancies from the CF resulting CF dissolution in LRS while in HRS, the retention loss can be attributed to the oxygen vacancy accumulation/generation within the gap between CF and electrode [85]–[90]. The driving force for vacancy diffusion is the concentration gradient between the filament and bulk; and vacancy generation can be due to the electrical/thermal stress [85, 86, 88, 89]. The argument is supported by the fact that retention loss accelerates when ambient temperature is increased due to enhanced diffusion [85–87, 90]. In a previous study, the LRS retention loss was remarkably improved in a TaOx-based device operating at low current by generating the CF having higher vacancy concentration or denser filament using a two-step forming technique [65]–[91]. In another study on HfOx-based ReRAM, significant improvement in the LRS retention at small current has been achieved by limiting the oxygen diffusion with post-annealing the ReRAM stack [90]. It was also observed that stronger forming operation resulted in better retention [90]. In addition, it is commonly observed that the retention time especially for LRS is a strong function of programming current and the retention loss can be severe when operating at a small current (<50 μA)[65, 90, 91]. This retention degradation dependence on programming current can be related to the total number of vacancies present in the CF and their diffusion/annihilation. As at higher Ic, the CF is thicker and hence has higher number of vacancies, the retention time is longer. This means that in case of MLC operation, the resistance levels will have a different rate of retention loss depending on the Ic with the worst data retention for the resistance level programmed at the lowest current level. This can limit the MLC operation in a way that the MLC retention time should be determined according to the retention time of the highest resistance LRS level, which may be significantly lower than the other resistance levels. This is also valid for reset voltage controlled MLC mode with the only exception that the lowest HRS level (obtained from the smallest rest voltage) exhibits the worst data retention [87].
6.2 Endurance

The switching endurance, which tells how many times a memory device can be switched between the resistance states without degradation is one of the most important performance metrics of a memory device. In ReRAM, an endurance of over $10^{12}$ cycles in a bilayer TaO$_x$-based device has been demonstrated [92]. The endurance is known to be highly dependent on the program/erase conditions of a resistive memory device [93]–[95] and various types of endurance failure mechanisms are discussed [95]–[97]. The existence of a trade-off relation between the endurance and retention characteristics in ReRAM has been reported [98]. Figure 16 shows reset voltage-dependent endurance of a HfSiO$_x$-based device with constant pulse width of 1 μs [93]. As the reset voltage is increased, the endurance degrades sharply due to the initiation of filament growth (negative set) from the bottom electrode [93]–[94]. In light of the above discussion, it can be said that the endurance of multiresistance levels obtained by controlling reset voltage may not be the same for all the resistance levels and hence the endurance of MLC operation will be limited by the endurance of the HRS level erased with the largest reset voltage. Similarly, it may also be possible that the endurance of the LRS state programmed with the largest set voltage will be the shortest due to the increased stress in the switching material. Therefore, the endurance of a single level cell is much higher than that of the multilevel cell.

Figure 15 Typical retention characteristics of multiple resistance levels obtained by controlling $I_c$ [56]. No noticeable change in the resistance with time confirms the nonvolatile behavior.

Figure 16 Effect of reset voltage amplitude on pulse endurance of an HfSiO$_x$-based ReRAM. The endurance significantly decreased at higher reset voltage [93].
7 Summary and future work

ReRAM is particularly appealing for ultrahigh density and low-cost, nonvolatile memory applications owing to its excellent cell scalability, MLC storage capability and 3D stackability. MLC technology is an alternative way to increase the storage density without the requirement of decreasing the physical device size or 3D stacking. The MLC storage in ReRAM along with the resistance variability and reliability of multiple resistance levels were discussed in detail in this chapter. Then, various MLC operation schemes and their physical mechanisms have been explored. A comprehensive analysis of variability of multiple resistance levels by identifying various factors such as cycle-to-cycle and device-to-device nonuniformity, RTN, effect of ambient temperature and interstate switching variability have been provided, which can help in estimating the safe resistance margin to ensure the successful MLC operation. The temporal and spatial switching variability contribute most in resistance margin degradation. Finally, the reliability characteristics such as retention and endurance of MLC levels and the impact on the MLC operation have been assessed. The retention and endurance of the MLC operation is limited by that resistance level which has the worst endurance/retention time.

Although comprehensive analysis on MLC storage is provided for a single ReRAM cell, a similar and practically useful analysis on the MLC characteristics and resistance variability can be performed in a cross-point array in 1selector-1ReRAM cell configuration. Furthermore, ways to combine the ReRAM cell physical salability, MLC characteristics, and 3D stackability should be researched in order to fully exploit the benefits of ReRAM for ultrahigh density and inexpensive nonvolatile memory applications.

References

Chien WC, Chen YR, Chen YC, et al. A forming-free WO


Park SG, Yang MK, Ju H, et al. A non-linear ReRAM cell with sub-1


