Physics and Technology of Emerging Non-Volatile Memories

Agostino Pirovano

1 Challenges in Floating-Gate Memory Scaling

Starting from their original introduction at the end of the 1980s, the fundamental properties that enabled flash memory to become first the non-volatile mainstream technology and later the semiconductor technology driver for scalability also enabled flash to follow the well-known Moore's law for semiconductors. This capability was largely demonstrated by the NOR flash-technology evolution that has followed that of the standard CMOS, introducing into the basic process flow many of the materials and modules already developed [1, 3]. In particular, considering the requirements for fast random access time, the supply voltage for mobile application down to 1.8 V and the efficient programming and erasing algorithm execution, starting from the 180-nm technology node, the CMOS structure has mainly followed the high-performance logic roadmap [4–6].

Figure 1 depicts the NOR-flash cross sections for successive generations, from the 0.8 um to the 65 nm, with the materials and the basic modules that have been introduced for each generation. With the device scaling, the front-end process module took advantage of improved salicidation processes introduced in high-performance logic, and the gate material has evolved from WSi_2 to $TiSi_2$ and finally to $CoSi_2$. Similarly in the back-end, the metallization evolved from single aluminum to triple copper layers.

Figure 2 shows the cell-size reduction as a function of the technology node F, where F represents the minimum feature size. Solid points represent the real cell sizes for NOR- and NAND-flash technologies put into production in the last 15 years, clearly showing the capability for both architectures to follow the area scaling suggested by Moore's law. In a flash-NOR cell, the theoretical size is 10 F^2,

A. Pirovano (✉)
Technology Development, Micron Semiconductor Italia,
Via Trento, 26, 20871 Vimercate, MB, Italy
e-mail: apirovan@micron.com

R. Gastaldi and G. Campardo (eds.), *In Search of the Next Memory*,
DOI 10.1007/978-3-319-47724-4_3

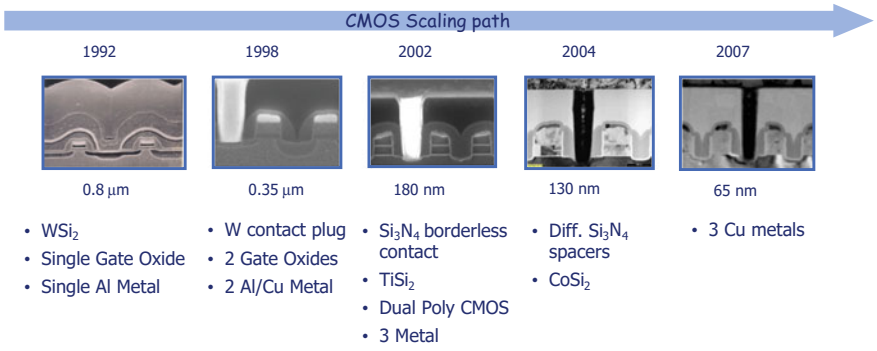

Fig. 1 The NOR flash-cell cross section for successive generations, reported in terms of technology node (year of production) and key materials introduced

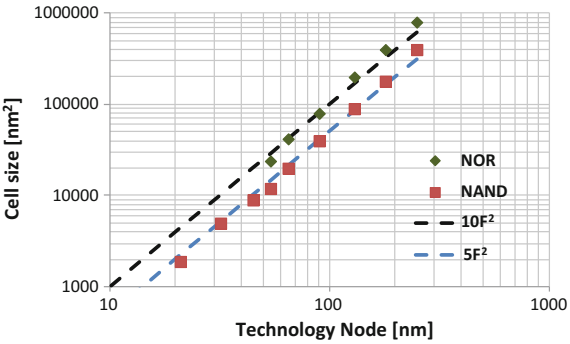

Fig. 2 Evolution of the area of NOR- and NAND-flash cells as a function of the technology node

while, in a NAND cell, the size is around 5 F^2, giving rise to a NAND-memory density higher than the NOR one. The different cell size arises from the array organization and from the cell layout. In a NOR cell, a contact is shared every two cells, and basically this doubles the number of lithography features needed to define the cell contact. Moreover, the CHE programming does not allow an aggressive scaling of the cell-gate channel length, which instead occurs in NAND, where the cell-gate length and space define the technology node.

Although flash technology has clearly demonstrated its capability to shrink the cell size according to Moore's law, further reduction of the dimension is facing fundamental physical limits, and it is demanding technological developments that are making the cell scaling less convenient from the economic standpoint. As shown in Fig. 1, in the last two decades the reduction in flash-cell size has been achieved by simply scaling every dimension for both active (flash-cell transistor) and passive elements (interconnections). The technology enabler for such evolution has been the availability of advanced lithographic techniques based on the continuous reduction of the optical-source wavelength and on the reduction of the effective wavelength by interposing a suitable media between the lens and the

silicon substrate. This latter technique, called immersion lithography, has been largely employed in recent years to enable the definition of lithographic features as small as 40–45-nm [7]. Although such equipment is really impressive from the technology and engineering standpoint, further techniques have been developed to shrink the cell size to the deca-nanometer range, while still relying on standard immersion lithography. All of them are based on the usage of sacrificial layers and deposition-etching-deposition techniques that enable doubling or even quadrupling the number of cells that can be defined inside the minimum lithographic pitch. Obviously these techniques come at the expense of much higher cost for cell production, thus creating a fundamental limitation to using them without a further improvement in the lithographic resolution. These additional developments in the lithographic equipment toward the extreme-UV (much shorter wavelengths) represent today the main technological challenge for enabling a further scaling of the planar flash-cell architecture at a reasonable production cost.

Apart from the technological issues for the further downscaling of flash-cell size, there are quite obvious physical limits that prevent further scaling of the cells. Such limitations have been successfully faced and managed over the years, but a few of them are seriously imposing strong compromises, in particular for cell reliability, in the sub-20 nm regime. The most obvious physical limitation is related to the dielectric thickness, namely the tunnel oxide dielectric and the interpoly dielectric. In particular, the minimum thickness of tunnel oxide is 9 nm for maintaining the same reliability specification, thus making the electrostatic control of the cell channel more and more difficult when scaling the transistor length.

Although the cell functionality has been preserved with scaling, cell-reliability degradation represents the main scaling issue in the most recent generations of flash technology. Such worsening is strictly related to the approach to a fundamental physical boundary: the number of electrons stored in the floating gate has been reduced from several thousands to hundreds or even tens of electrons, thus amplifying the impact of the dielectric degradation mechanisms. In fact, the impact of the usual trapping/detrapping mechanism, as well as trap-assisted tunneling phenomenon, is now much larger than in the past, and the loss of just few electrons from the floating gate is enough to delete the stored data, resulting in a significant reduction in overall reliability, particularly for the data-retention capability after cycling [9, 10].

Another side effect of the cell-size reduction is amplification of the noise effect, in particular the random telegraph-noise contribution. In fact, the usual 1/f noise of a MOS transistor has been observed as a giant threshold-voltage leap in sub-45 nm cell devices, thus adding an additional constraint to the overall reliability and noise immunity of this technology.

Finally, one of the most disturbing drawbacks associated with the cell-size reduction is the simultaneous reduction of the distance between adjacent cells. Such lower cell separation is responsible for additional issues in term of breakdown of the dielectric layer that separates adjacent word-lines and that need to maintain very high voltages. Moreover, the closer distance between adjacent cells makes the contribution of lateral capacitances between adjacent floating gates no longer

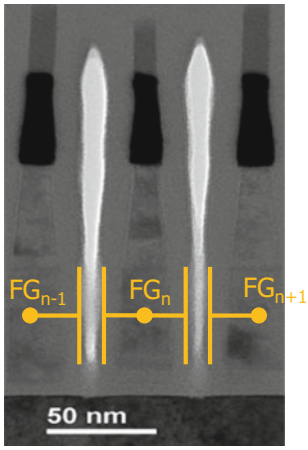

Fig. 3 Cell-proximity interference effect in scaled technologies. The floating-gate potential of each cell is capacitively coupled to the potential of the adjacent cells

negligible, and the capacitive coupling used for determining the floating-gate potential require taking this contribution into account.

In the standard, planar flash cell, the basic assumption is that the floating-gate potential is controlled by capacitive-charge sharing where the main contributions come from the control-gate terminal and from the cell-transistor nodes. However, Fig. 3 shows that, in scaled flash technology, the planar-cell approximation is no longer valid and that the short cell separation leads to additional capacitances that must be considered in determining the floating-gate potential. Moreover, the lateral capacitance contribution will depend on the potential of the adjacent floating gate, thus linking the potential of each cell with the surrounding ones and leading to so-called cell-proximity interference. In order to minimize the contribution of such capacitances, a limited filling of the space between adjacent cells has been adopted, thus leaving a void that introduces the smallest possible dielectric constant. Finally, the parasitic capacitive coupling between neighboring floating gates leads to a low coupling ratio with the control gate, which results also in a small stored charge during the programming operation.

Finally, a third level of scaling challenges is related to the suitability of scaled flash cells to meet the requirements of the new generation flash-based products. NAND scaling is indeed increasing the density and lowering costs but reliability is rapidly worsening, thus requiring more complex programming and reading algorithms for managing the higher bit-error-rate (BER). As result, write and read latency is predicted to increase with the device scaling, thus opening the serious question if the NAND-based SSDs price/performance ratio will be competitive with magnetic disks.

Despite the prediction at the end of the last century that the floating-gate concept faced severe technological limits beyond the 32-nm technology node, NAND-memory density has shown the ability to be downscaled to the 16-nm node with the multilevel-cell concept applied [11]. Such an achievement is partially due to the fact that, for NAND products, a significant reliability drop (in particular for

endurance) is acceptable, considering the large ECC available and the processing power of dedicated controllers in SSD applications. A similar trade-off is not acceptable for NOR-flash applications, thus making their scaling more challenging. NOR flash has thus reached its scaling limitation at 45 nm (even considering the constant/declining market demand that is not fostering additional development efforts for this technology) while NAND is available at the 16-nm technology node.

2 The Future of the Floating-Gate Concept

The continuation of the flash scaling following Moore's law has been the main focus of the semiconductor memory industry during recent decades, and several strategies have been proposed throughout the years. These proposals can be schematically summarized into three main branches of developments: system-level management techniques, material engineering, and novel architectures.

2.1 System-Level Management Techniques

An effective workaround to manage the reduced reliability of scaled technology and the correspondingly higher bit-error-rate has been the massive introduction of Error-Correction-Code (ECC) algorithms. The object of the theory of error correction codes is the addition of redundant terms to the message, such that, on reading, it is possible to detect the errors and to recover the message that had most probably been written. Most popular ECC codes that correct more than one error are Reed-Solomon and BCH [12]. BCH and Reed-Solomon codes have a very similar structure, but BCH codes require fewer parity bits, and this is one of the reasons why they were preferred for an ECC embedded in the NAND memory [13]. While the encoding takes few cycles of latency, the decoding phase can require more cycles and visibly reduce read performance, as well as the memory response time at random access. Moreover, ECC requires additional cells to encode the logical data, thus using part of the memory chip just to improve the overall reliability.

Since the main scope of the device scaling is to reduce the cost per bit, a similar result can be obtained by exploiting the Multi-Level-Cell (MLC) capability of floating–gate devices. Planar NAND with MLC capability represents today the most cost-effective solution for an NVM concept. In fact, it can provide a minimum cell size of 4 F^2 combined with MLC capabilities up to 3 bits-per-cell at the leading-edge technology node of 16 nm, resulting in an effective cell size of 1.3 F^2. MLC feasibility is mainly related to the system-level management, and it require a more precise level placement (smarter programming algorithms), as well as the capability to manage worse BER and reliability (a high level of ECC is required).

The obvious advantage of a 2 bit/cell implementation (MLC) with respect to a 1 bit/cell device (SLC) is that the area occupied by the matrix is half as large; on the other hand, the area of the periphery circuits, both analog and digital, increases. This is mainly due to the fact that the multilevel approach requires higher voltages for programming (and therefore bigger charge pumps), higher precision and better performance in the generation of both the analog signals and the timings, and an increase in the complexity of the algorithms. Driven by cost, flash manufacturers are now developing 3 bit/cell (eight threshold voltage levels) and 4 bit/cell (16 levels). Three- and four-bits per cell are usually referred to as XLC (8 and 16 LC, respectively).

The capability to mitigate the scaling issues with the adoption of system-level management techniques is well represented by the adoption of a sophisticated memory controller to manage the NAND flash inside products. The original aim of the memory controller was to provide the most suitable interface and protocol suitable to both the host and the flash memories. However, with the intrinsic reliability and performance degradations observed in sub-45 nm flash devices, the role of the memory controller has been extended, and now it is employed to efficiently handle data, maximizing transfer speed, data integrity, and information retention.

For example, in a typical consumer application, not all the information stored within the same memory location changes at the same frequency, and some data are often updated while others remain the same for a very long time. It's clear that the cells containing frequently-updated information are stressed with the large number of write/erase cycles, while the cells containing information updated very rarely are much less stressed. In order to mitigate the reliability issues, it is important to keep the aging of each block of cells at a minimum and as uniform as possible, and to monitor the maximum number of allowed program/erase cycles. To this aim, wear leveling techniques are employed to dynamically map the logical data onto different cells, keeping track of the mapping. In this way, all the physical cells are evenly used, thus keeping the aging under a reasonable value.

2.2 Dielectric-Materials Engineering

The second scaling strategy explored during the last 15 years has been the improvement of flash-cell active materials, in particular the dielectrics employed in the floating-gate definition. One of the most important attempts made to mitigate scaling limitations, while retaining the very high integration density of NAND-flash architecture, has been the attempt to replace the conventional floating gate with a charge-trapping layer. Silicon nano-crystal trapping layers have been investigated in the past [14], but they present a few drawbacks, like reduced threshold shift and the presence of percolation paths between source and drain that become more severe with the scaling of the cell size. The silicon nano-crystal technology requires a careful control of the nano-dots size, dimension, shape, and density, because these parameters significantly impact device performance and reliability. Moreover the

down-scaling of this technology was expected to be difficult beyond the 32-nm technology node due to the minimum nano-crystal size that has so far been achieved in a reproducible way.

Other alternatives include the use of a continuous trapping layer (charge-trap memories, also called CT memories), like silicon nitride in the SONOS-device architecture [15]. This approach promises to solve several of the scalability issues: the charge is trapped in a thin dielectric layer, and therefore there is no problem of capacitive interference between neighboring cells; since the charge is stored in electrically insulated traps, the device is also immune to SILC, the parasitic leakage current caused by single defects in the dielectric layer, while, in conventional floating-gate devices, even a single defect can discharge the whole floating gate, which is a conductive storage medium; the replacement of the floating gate with a trapping layer reduces the overall thickness of the gate stack, and allows for easier integration of the cell in the CMOS process. Even if alternatives are available, silicon nitride is probably the best storage material since it is characterized by a high trap density and by a very long lifetime of the charged state that ensures large threshold windows and excellent data retention in memory applications [16].

In this architecture, the charge is trapped in a silicon nitride layer, inserted between two silicon-oxide layers, which act as a tunnel dielectric and blocking layer to prevent charge injection from the control gate. Although this cell architecture is known since the 1980s and in spite of its better compatibility with standard CMOS process flow and lower costs, it lost ground in favor of the floating-gate one, because of several fundamental problems: first, cell programming is limited by the erase saturation, which takes places because of the parasitic electron injection from the control gate through the top oxide, balancing the hole injection from the substrate. Second, he thinning of the tunnel oxide (<2.5 nm) improves the threshold window and the programming speed, but results in poor retention, even at room temperature, because of direct tunnelling through the tunnel oxide, and charge mobility in the nitride layer. Finally, increasing the tunnel thickness improves the retention, but requires larger programming voltages; it also reduces the speed and activates the tunnelling through the top oxide.

Despite the scaling issue of the floating-gate concept, the actual drawbacks of the CT memories made conventional planar flash a preferred choice for mainstream technologies, thus limiting the usage of CT layers in planar flash production.

2.3 Novel Flash Architectures

In this scenario and considering the growing difficulties that must be faced for planar NAND scaling, there is a strong interest in the so-called 3D NVM technologies. This is the third scaling strategy employed by the semiconductor memory industry, where novel cell architectures are proposed to further extend the scaling. Among the possible architectures, the most promising are the vertical ones, i.e., all the architectures that try to exploit the vertical direction to increase the cell density,

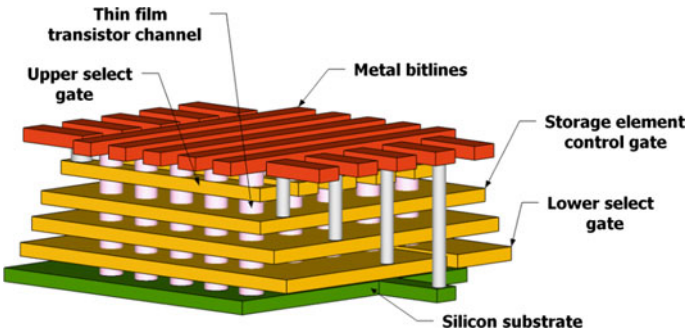

Fig. 4 Schematic representation of a vertical 3D-NAND architecture

thus moving from the conventional planar cell construction to a three dimensional (3D) approach.

There are currently two main trends in the efforts to develop 3D structures that can provide a higher integration density: cross-point memory array, where several memory layers can be stacked, and the so called 3D NAND, where the standard NAND strings are integrated along the vertical dimension (Fig. 4).

3D NAND was proposed as a cost-effective solution for vertical NAND fabrication. The process is intended to use a minimum number of masks, thus reducing the fabrication costs. However much effort is required to develop the suitable integration modules in order to have at least 32 layers along the vertical dimension. Such a huge number of layers is needed to make it possible for a 3D NAND with a relaxed pitch to be cost-effective with respect to the existing planar NAND.

Although several approaches have been proposed to fabricate a 3D NAND chip, the most effective solution employs a vertical channel with a horizontal gate. In this case, the number of critical masks is low, since the entire stack is etched at the same time. The limited dependence of wafer cost on the number of levels results in a fortunate economic scenario for these arrays, even if the typical cell size is relatively large and many stacked layers are necessary to reach a small equivalent cell area (i.e., a single cell area divided by the number of memory layers). The typical 3D NAND structure is illustrated in Fig. 4, the quantity of cells inside a string is defined by the number of vertical wordlines layers stacked in the array. Bitlines and drain selector lines run horizontally and are used to select string in the two directions. Three architectures with vertical channels and horizontal gates are: BiCS [20], VRAT [21], and TCAT [22]. It has been announced that the 3D-NAND technology will be commercialized starting with the SSD application, with 24 layers and MLC in a 128-Gb monolithic module based on CT technology [18] and with 32 layers 256-Gb MLC and 384-Gb triple-level cell (TLC) 3D NAND based on floating-gate technology [19].

3 Alternative Storage Concepts

To improve the performance and scalability with respect to floating-gate devices, innovative concepts for alternative NVM have been proposed in the past and are under investigation today, as we dream of find the ideal memory that combines fast read, fast write, non-volatility, low-power, and unlimited endurance, and obviously at a cost comparable to flash or DRAM.

Table 1 reports a schematic grouping of the alternative NVM concepts based on the decoding technique and on the selected architecture. Generally speaking, electronic memories can be divided into two main classes: solid-state-device memory, where each cell is placed at the intersection of two orthogonal metal lines defined through lithographic technique (e.g., DRAM, Flash); and mechanically decoded memories in which a mechanical positioning of a programming/reading equipment is adopted to address data on a flat substrate (e.g., magnetic disc and optical disc).

In 2000, IBM proposed an alternative memory device [23], basically a sort of miniaturized hard disk system based on a micro-electro-mechanical system (MEMS) that actuates thousands of tips capable of decoding the information stored on a flat media. Several media were proposed at that time to store data with a bit size in the range of ten nanometers, but the overall complexity and cost of this system did not allow it to compete with the fast-growing NAND technology.

On the other side, the more traditional electronically decoded memories tried to exploit the properties of novel materials to create self-selecting (cross-point) and transistor-selected NVM memories. The main class of emerging NVM technologies so far investigated is based on inorganic materials, and it includes the alternative memory concept with the highest maturity level, namely Ferroelectric Memories

Table 1 Schematic grouping of the alternative NVM concepts

Electronic decoded, lithography depend (Moore's law follower)	Mechanical decoded, lithography independent (beyond Moore's law)
• Transistor selected (like DRAM or Flash) – Ferroelectric memory (FERAM) – Magnetoresistive memory (MRAM and STT-MRAM) – Resistive RAM (RRAM) – Phase-Change Memory (PCM)	• Probe storage (Seek and scan, like Hard Disk or CD) – Polymers – Chalcogenide – Ferroelectric
• Cross-point memories (Passive arrays) – Ferroelectric polymers (PFRAM or TFEM) – Organic charge-transfer complex (conductive polymers) – Resistive switching	

(FeRAM), Magnetoresistive Memories (MRAM), Phase Change Memories (PCM) and Resistive RAM (ReRAM). These NVM alternative concepts will be described in detail in the next sections.

3.1 Ferroelectric Memories

FeRAM is one of the few alternative NVM that has been commercialized so far, even if at a technology node much more relaxed than the one used for flash memories and with several challenging technological problems, mainly related to new materials and new manufacturing technologies. Two classes of ferroelectric materials are currently used for FeRAM memories: perovskite structures and layered structures. Actually, the most-used perovskite material for ferroelectric memories is $PbZr_xTi_{1-x}O_3$, also called PZT, while the layered ferroelectric choices for FeRAM memories are either strontium-bismuth-tantalate $Sr_{1-y}Bi_{2+x}Ta_2O_9$, also called SBT, or lanthanum substituted-bismuth-titanate $Bi_{4-x}La_xTi_3O_{12}$, also called BLT. Among them, the commercially preferred option is represented by the PZT, usually deposited with the MOCVD technique at temperatures higher than 600 °C. Ferroelectric materials can be polarized spontaneously by an electric field. The polarization occurs as a lattice deformation of the cubic form below the Curie point, the temperature above which the material becomes paraelectric. For example, in PZT the titanium atom can be moved by an electric field to two stable positions that are above and below the oxygen plane of the structure. An important property of ferroelectric materials is therefore their residual permanent polarization, typically in the range of 10–30 $\mu C/cm^2$. The voltage required to switch the permanent polarization is in the range of 1.5–3 V for typical deposited-layer thicknesses of 70–100 nm. It follows that ferroelectric memories can be a valuable solution for low-power and very low-voltage application, like a battery-operated embedded system, smartcards, and RFID applications.

One of the most challenging features of this technology is presented by the integration of the ferroelectric layer into the standard CMOS process. The bottom and top electrodes of the ferroelectric layer must be realized with a specific alloys usually constituted by iridium and platinum. Hydrogen contamination of the ferroelectric material must be avoided in order to prevent the reduction of the permanent-polarization capability, thus requiring a specific barrier layer all around the ferroelectric capacitor. At the same time, oxygen can diffuse during the high-temperature treatment required for ferroelectric alloy deposition, oxidizing the underlying metal layers. Finally, special care must be devoted to the definition of the capacitor shape through a dry etching, to the final dielectric layer used to seal the capacitor from the surrounding environment and to the effect of the plasma damage due to the CMOS back-end process, resulting in a possible discharge through the capacitor that destroys its capability to store data.

Endurance, also called electric fatigue, is an important reliability characteristic, and it is related to the decrease of the ability to switch the memory cell into the

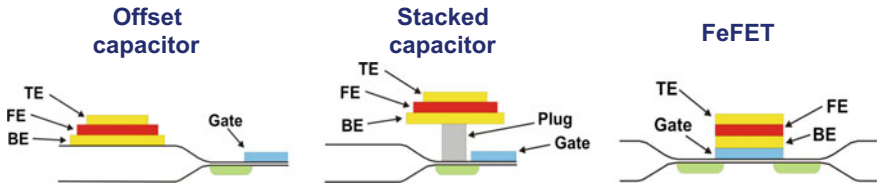

Fig. 5 FeRAM cell architectures with the ferroelectric material (FE) integrated either into a separate storage element, i.e., a ferroelectric capacitor (offset capacitor or stacked capacitor option) or into the selection element, i.e., a ferroelectric FET transistor (FeFET)

opposite state, after being kept programmed in one state for long periods of time. This effect is related to the polarization shift in the hysteresis loop, and it is proportional to the increasing number of switching cycles. Nevertheless, the write cycle (but also the read cycle, since several proposed cell structures have a destructive reading) is expected to have an endurance level of about 10^{12}, which will be enough for a wide majority of high-demanding storage applications. Up to now several FeRAM cell structures have been proposed, with the ferroelectric material integrated either into a separate storage element, i.e., a ferroelectric capacitor [24], or into the selection element, i.e., a ferroelectric FET transistor [25] (Fig. 5). In the latter case, the storage and the selection elements are merged. The first cell type can be used in both the two-transistor/two-capacitor (2T/2C) cell and the one-transistor/one-capacitor (1T/1C) cell, while the latter has been proposed with a one-transistor (1T) approach. Moreover, a NAND-type FeRAM array configuration has also been proposed with the name of chain-type FeRAM memory. All FeRAM architectures have high-speed access cycles and provide genuine random access to all memory locations. Among the proposed architectures, the 1T/1C FeRAM approach is characterized by quite large cell size that cannot compete with today's high-density solutions (flash for NVM and DRAM for volatile storage). However, the very low-voltage operation and the superior electrical performance in terms of programming speed and endurance make FeRAM technology a valuable solution for specific application in the embedded market. The 1T FeRAM architecture is a very promising alternative for high-density application with very small cell size and practically an infinite endurance level, but its processing has proven to be very complicated, and no products have been so far announced or released. Apart from the FeRAM potentialities discussed above, there is a quite important issue for cell scaling that could impact the further development of this technology. In fact, the cell sensing in capacitor-based architectures relies on the capability to detect the displacement current associated with this capacitance, similarly to what is usually done in DRAM technology. With a planar capacitor approach, the continuous shrinking of the cell size corresponds to a reduction of the capacitor surface, degrading the available signal for reading the cell status. As for DRAM, this issue can be solved just by moving from the simple planar capacitor to the more complicated three-dimensional (3D) capacitor architecture, analogously with what had already happened in the DRAM-scaling roadmap. Moreover,

ferroelectric properties tend to disappear in very thin layers, thus making the scaling of the active material below 50-nm thickness a huge issue, at least for the proposed ferroelectric alloys. It follows that, to scale FeRAM technology below the 90-nm technological node, a more complicated 3D approach is mandatory, really challenging the already difficult fabrication process associated with the integration of the ferroelectric material into a standard CMOS process.

3.2 Magneto-Resistive Memories

Up to the early 2000s, all the development efforts for MRAM technology were MTJ cell based [26], with an architecture composed of one transistor and one resistor (1T/1R). This technology relies on the adoption of a tunnel junction coupled to magneto-resistive materials that exhibit changes in their electric resistance when a magnetic field is applied. The MTJ is composed of a pinned magnetic layer, a tunnel barrier, and a free magnetic layer. Electrons spin polarized by the magnetic layers traverse the tunnel barrier. A parallel alignment of the free layer with respect to the pinned layer results in a low resistance state, while an anti-parallel alignment results in a high resistance state [27, 28]. Therefore the storing mechanism consists of the permanent magnetization of the ferromagnetic material in the MTJ. The datum can be sensed as the resistance in the MTJ which can be high (low current) or low (high current). The writing can be performed through the magnetic field produced by the current flowing in the bit- and digit-lines.

The non-destructive read with a very fast access cycle is the premise for high performance, equal-long read and write cycles, and for low-power operation. Moreover, the structure is radiation-hard with a potentially unlimited read/write endurance, which makes MRAMs suitable for write intensive storage applications. The major MRAM disadvantage appears to be the high write current. While this technology has enough read current to guarantee a fast access time, it requires a very large write current (mA range), which increases power consumption. The current requirements become even more challenging when MRAM devices are scaled. In fact, since the data-retention capabilities of MRAM memory cells are related to the total volume of magnetic material used in the free layer of the MTJ, it is expected that the capability to retain the stored information will be degraded in scaled devices, where the MTJ geometrical features are shrunk. It follows that suitable materials with higher magnetic coercivity must be adopted to retain the data in scaled devices, thus demanding more current to be programmed and erased.

One of the key milestones for the development of MRAM technology was the introduction of the toggle-MRAM writing scheme in order to achieve better program-disturb immunity. With respect to the conventional MRAM, a toggle-MRAM employs a programming technique based on the current amplitude (as in conventional MRAM) and on the timing of the applied programming pulses. Only the correct sequence of pulses delivered to the selected cell are able to switch its magnetization, leaving unchanged all the other cells along the programmed bitline

and wordline. Beginning in 2006, low-density (4- to 16-Mb) chips based on the toggle MRAM concept have been commercialized and are today available on the market for very specific applications.

In order to mitigate the scaling issues of the MRAM concept, a novel programming technique has been recently investigated and is fuelling a renewed interest in the MRAM technology. This approach is based on the spin-polarizing effect [29], in which magnetization orientations in magnetic multilayer nanostructures can be manipulated via spin-polarized current. The so-called Spin-Transfer Torque-MRAM (STT-MRAM) technology is based on an MTJ structure where a current-induced switching caused by spin-transfer torque is exploited. Despite that this approach enables mitigation of some of the conventional MRAM issues, particularly for scaling, there are still several challenges that must be faced (e.g., self-read disturbance, writing times, cell integration). Today the STT-MRAM developments are very active with prototypes of 64 Mb and 1 Gb at 90 nm [30] and 54 nm [31], respectively. However no volume production has yet been started.

3.3 Phase-Change Memories

Among the different NVM based on mechanisms alternative to the floating-gate concept, Phase-Change Memories (PCM) are one of the most promising candidates to become mainstream NVM, having the potentiality to improve the performance compared to flash — random access time, read throughput, direct write, bit granularity, endurance — as well as to be scalable beyond flash technology.

PCM exploits thermally reversible phase transitions of some chalcogenide materials or alloys (e.g., $Ge_2Sb_2Te_5$). In fact some alloys based on the VI group elements (usually referred to as chalcogenides) have the interesting characteristic of being stable at room temperature, both in their amorphous and crystalline phases. In particular, the most promising are the GeSbTe alloys that follow a pseudo-binary composition (between GeTe and Sb_2Te_3), hereafter referred to as GST. The most interesting feature of these alloys is their capability to reversibly switch between a high-resistance amorphous phase and a low-resistance crystalline one in a few hundreds of nanoseconds.

The basic cell structure is composed of one transistor and one resistor (1T/1R) that can be programmed through the current induced Joule heating, and can be read by sensing the resistance change between the amorphous and the polycrystalline phase. The PCM cell is essentially a resistor of a thin-film chalcogenide material with a low-field resistance that changes by orders of magnitudes, depending on the phase state of the GST in the active region. The switch between the two states occurs by means of local temperature increases. Above the critical temperature, crystal nucleation and growth occur, and the material becomes crystalline (Set operation). To bring the chalcogenide alloy back to the amorphous state (Reset operation), the temperature must be increased above the melting point of hundreds of °C and then very quickly quenched down to preserve the disorder and not let the

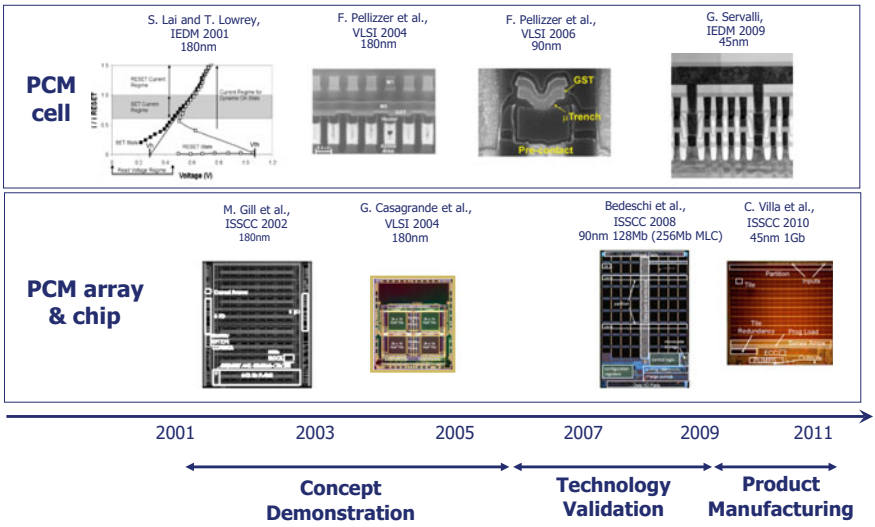

Fig. 6 Evolution of PCM development

material crystallize. From an electrical point of view, it is possible to use the Joule effect to reach locally both critical temperatures using the current flow through the material by setting proper voltage pulses. The cell read out is performed at low bias. Programming thus requires a relatively large current in order to heat-up the GST and results in to a thermally induced, local phase change. Phase-transitions can be thus easily achieved by applying voltage pulses with various amplitudes and with durations in the range of tens or hundreds of nanoseconds.

Although the phase-change concept is well known for years and the first studies date back to the 1970s [32, 33], its application for NVM experienced renewed interest in 1999, when the idea for integrating a phase-change material into a NVM cell was presented again [34]. As result, the effort to bring the PCM basic concept to a mature technology level has constantly increased since then, and many groups have started to study, to develop, and to integrate a Multi-Megabit array in the memory cell.

The development roadmap of PCM technology is summarized in Fig. 6. A 180-nm technology node has been used to develop the first demonstrator vehicles and to prove the technology viability [35]. The BJT-selected cell has been chosen for the high-performance and high-density application, since the cell size can be $\sim$5 F^2. The MOS-selected cell is suitable for system-on-a–chip or embedded application [36], because, in spite of the larger cell size ($\sim$20 F^2), the memory integration adds only very few masks to the logic process with a clear cost advantage.

A 90-nm technology node has been developed and commercialized using a 128-Mb product [37]. A 1-Gb-PCM product fabricated at the 45 nm technology node with a cell size of 5.5 F^2 has been developed, and it is in volume production

for wireless applications [38]. This 45-nm-PCM architecture [39] demonstrates the maturity of the technology. The energy delivered to program a bit is on the order of 10 pJ, with a state-of-the-art, random access time of 85 ns, read throughput 266 MB/s and write throughput 9 MB/s [40]. These peculiar features combined with data retention, single bit alterability, execution in place, and good cycling performance enable traditional NVM utilizations but also novel applications in the LPDDR field. Moreover, PCM is considered the essential ingredient to push to the market the so called Storage-Class Memory (SCM) [41], a non-volatile, solid-state-memory technology that is capable of filling the gap between CPUs and disks.

One of the key challenges of the PCM-cell scaling is the reduction of the power, mainly due to the programming current. So far, most of the development has been focused on the cell-structure optimization in order to increase the programming efficiency. More recently it has been shown that, also by proper engineering of the active material, it is possible to reduce the programming current of a standard PCM cell of 1/20 [42], enabling even more low-power, innovative applications.

3.4 Resistive RAM

Resistive Memories, usually called ReRAM, are a quite large class of memory concepts that store the information in the resistance value of the cell. It is worth noting that PCM and MRAM fall within this definition too, but the interest and the efforts devoted to their specific development in the last decades set them apart as stand-alone concepts. All the resistive memories are electrically programmable, although they are based on different material classes and different proposed switching mechanisms:

- Formation of metallic bridges by dielectric breakdown and bridge opening by thermal fusion [43];
- "Volume" switching, e.g., by electronic charge transfer (redox) mechanisms [44];
- Switching of filament regions in the resistive material by electronic charge transfer (redox) mechanisms [45];
- Electrochemical growth and dissolution of metallic dendrites by solid electrolyte/electrode processes (programmable metallization cells, PMC) [46].

The resistive memory materials proposed in these concepts range from organic materials (rotaxanes and catananes, polyphenyleneethylenes, Cu- and Ag-TCNQ) to inorganic (chalcogenide alloys, perovskite-type oxides, manganites, binary transition metal oxides), while electrode materials comprise various metals as well as electronically conducting oxides and nitrides. Among the materials that exhibit a resistive switching phenomenon oxide materials have been studied intensively.

On the basis of *I–V* characteristics, the switching behaviors can be classified into two types: unipolar (nonpolar) and bipolar. In unipolar-resistive switching, the switching direction depends on the amplitude of the applied voltage but not on the polarity. An as-prepared memory cell is in a highly resistive state and is put into a low-resistance state (LRS) by applying a high-voltage stress. This is called the 'forming process'. After the forming process, the cell in a LRS is switched to a high-resistance state (HRS) by applying a threshold voltage ('reset process'). Switching from a HRS to a LRS ('set process') is achieved by applying a threshold voltage that is larger than the reset voltage. In the set process, the current is limited by the current compliance of the control system or, more practically, by adding a series resistor. This type of switching behavior has been observed in many highly insulating oxides, such as binary metal oxides.

Bipolar resistive switching shows a resistive change that depends on the polarity of the applied voltage. This type of resistive switching behavior occurs in many semiconducting oxides, such as complex perovskite oxides.

It should be noted that there is a lot of speculation and controversy on the actual, physical switching mechanisms for many of these concepts. Moreover, in many cases, the role of the electrode materials is found to be very important although it is also not exactly understood. Independent of the mechanism, however, the important, basic characteristics for all concepts are represented by the required switching voltage and switching current. Indeed, low switching voltages are required to be compatible with the low supply voltages of scaled technologies, while low switching currents are required to be able to switch with minimal-size selector devices, as well as to limit the switching power. For array fabrication, a transistor-type architecture is preferred while the cross-point architecture and the diode architecture open the path toward stacking memory layers, and therefore are ideally suited for mass-storage devices.

An important advantage of ReRAM technology is the good compatibility with the CMOS processes, in particular for binary-oxide-based memories. The critical issues for the future development of ReRAM devices are reliability, such as data retention and memory endurance (the number of erase and program cycles), and the characteristic variations from cell to cell and from chip to chip. Despite the large potential evident in some of these concepts, the poor control of resistance distribution and the low maturity level reached by the cell integration into sub-micrometric devices represent today the main limitations for several of them.

Despite the huge interest among the semiconductor companies and inside the scientific community, very few attempts to develop ReRAM have materialized as real commercial products. The availability of an evaluation kit including a ReRAM memory fabricated in 180-nn technology has been announced [47]. On the research level, a 32-Gb, ReRAM test chip developed in a 24-nm process, with a diode as the selection device and a 2-layered architecture, has been presented [48]. These interesting results make ReRAM one of the most promising alternative memory technologies for mainstream applications.

4 Scaling Path and Issues in Various Emerging Architectures

The current NVM mainstream is based on flash technology, and it is expected that flash will be the high-volume NVM in production for the next years. Flash technology is characterized by a compact structure in which the selecting element and the storage element are merged in a MOS-like architecture. The resulting, full compatibility with the CMOS technology and the compact device size have made flash technology the cheapest solution for stand-alone and embedded-memory applications. However, discordant requirements to shrink the MOS structure, while preserving good selection and storage capabilities, are making flash scaling more and more difficult. Moreover, even if in the long term the cost advantage is important, better performance can speed up a novel technology introduction. In fact, a NVM with low-power and low-voltage capabilities, bit granularity, fast operations, and higher endurance would be a potential game changer for system designers.

To enlarge application segments, offering better performance and scalability, new materials, and alternative memory concepts are mandatory to boost the NVM industry. During recent decades, a total of more than 30 NVM technologies and technology variations have been competing for a piece of the fast growing NVM market, many of them aiming to replace also DRAMs. Although the planar-NAND scaling is becoming harder and it is obtained with several compromises in term of reliability, state-of-the-art NAND technology is 3 bits-per-cell 16 nm. Moreover, 3D NAND is becoming a mature technology. It follows that any NVM development must be able to provide the same capabilities at higher density. If this is not achieved in the next generation, technology scaling will not result in a cost reduction, thus eliminating the interest to continue along this path. Cost structure is therefore a fundamental parameter for benchmarking novel NVM concepts, in particular for those that want to compete for data-storage solutions, where MLC capabilities and/or 3D-stacking are mandatory.

On the other hand, the actual development efforts of alternative NVM concepts are demonstrating that disruptive innovation takes a long time. Figure 6 recounts schematically the development history of PCM technology, the only emerging memory concept that has reached the volume production maturity for large-density arrays. It is worth noting that the continuous need to stay close to the state-of-the-art lithographic node, combined with the necessary learning cycles, necessitated a decade of effort to evolve from concept to mass production. Moreover, we need also to consider the time-to-market, i.e., the time needed to get a significant profit from a novel NVM technology. For the flash NOR, about four years were needed to reach the break-even with the EEPROM in terms of profit for the main semiconductor industries involved in this market. The scaling lesson was also clearly demonstrated by the MRAM developments. MRAM products reached product maturity simultaneously with the last feasible technology node for the conventional toggle-MRAM, thus limiting the commercial success of MRAM technology and

reducing it to a niche market. Only the discovery and exploitation of the STT-MRAM concept enabled a better scalability below the 90 nm node, thus renewing the interest in this technology.

In this perspective, there are two major aspects that must be considered for evaluating the potentials of any emerging memory concept, namely the readiness for moving beyond the leading—edge technology node and the scalability perspective. If we combine the above two statements, it follows that any realistic proposal for a novel NVM technology must prove its feasibility for the sub-1X-nm-technology node.

References

1. R. Bez et al., "Introduction to Flash Memory", Proceedings of the IEEE, vol. 91, n. 4, 2003.
2. "Flash Memories", edited by P. Cappelletti, C. Golla, P. Olivo, E. Zanoni, Kluwer Academic Publishers, 1999.
3. G. Ginami et al., "Survey on Flash Technology with Specific Attention to the Critical Process Parameters Related to Manufacturing", Proceedings of the IEEE, vol. 91, n. 4, p. 503, 2003.
4. A. Fazio, "A High Density High Performance 180 nm Generation ETOX Flash memory Technology", IEEE IEDM Tech. Dig. pp. 267–270, 1999.
5. S. Keeney, "A 130 nm Generation High Density ETOX Flash Memory Technology", IEEE IEDM Tech. Dig. pp. 2.5.1–2.5.4, 2001.
6. G. Servalli et al., "A 65 nm NOR Flash Technology with 0.042 μm^2 Cell Size for High Performance Multilevel Application", IEEE IEDM Tech. Dig., pp. 2.5.1–2.5.4, 2005.
7. H. Hu et al., "K = 0.266 immersion lithography patterning and its challenge for NAND FLASH", Semiconductor Technology International Conference (CSTIC), 2015 China.
8. K. Naruke et al., "Stress Induced Leakage Current Limiting to Scale Down EEPROM Tunnel Oxide Thickness", IEEE IEDM Tec. Dig., pp. 424–427, 1988.
9. J. S. Witters et al., "Degradation of Tunnel Oxide Floating Gate EEPROM Devices and Correlation with High Field Current Induced Degradation of Thin Gate Oxide", IEEE Trans. Electron Devices, vol. 36, p. 1663–1682, 1989.
10. D. Ielmini et al., "A Statistical Model for SILC in Flash Memories", IEEE Trans. Electron Devices, vol. 49, n. 11, p. 1955–1961, 2002.
11. Micron Press Release, "Intel, Micron Extend NAND Flash Technology Leadership With Introduction of World's First 128 Gb NAND Device and Mass Production of 64 Gb 20 nm NAND", December 6, 2011.
12. R. Micheloni et al., "Error Correction Codes for Non-Volatile Memories", Springer-Verlag, 2008.
13. R. Micheloni et al., "A 4 Gb 2b/cell NAND Flash Memory with Embedded 5b BCH ECC for 36 MB/s System Read Throughput", IEEE International Solid-State Circuits Conference Dig. Tech. Papers, pp. 142–143, Feb. 2006.
14. B. DeSalvo et al., "How Far Will Silicon Nanocrystals Push the Scaling Limits of NVMs Technologies?" IEEE IEDM Tech. Dig., p. 597–600, 2003.
15. Y. Shin et al., "A Novel NAND-type MONOS Memory using 63 nm Process Technology for Multi-Gigabit Flash EEPROMs", IEEE IEDM Tech. Dig., p. 327–330, 2005.
16. B. Eitan et al. "NROM: A Novel Localized Trapping, 2-bit Nonvolatile Memory Cell", IEEE EDL, Vol. 21, No. 11, 2000.
17. C. H. Lee et al., "A novel SONOS structure of SiO2/SiN/Al2O3 with TaN metal gate for multi-Giga bit Flash memories", IEDM tech. digest 2003.

18. J. Kim et al., “Novel Vertical-Stacked-Array-Transistor (VSAT) for ultra-high-density and cost-effective NAND Flash memory devices and SSD (Solid State Drive)”, Symposium on VLSI technology 2009.
19. Micron Press Release, “Micron and Intel Unveil New 3D NAND Flash Memory”, March 26, 2015.
20. H. Tanaka et al., “Bit Cost Scalable technology with punch and plug process for ultra-high density Flash memory“, Symposium on VLSI technology 2007.
21. J. Kim et al., “Novel 3-D structure for ultra-high density Flash memory with VRAT (Vertical-Recess-Array-Transistor) and PIPE (Planarized Integration on the same Plane)”, Symposium on VLSI technology 2008.
22. J. Jang et al., “Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra-high density NAND Flash memory”, Symposium on VLSI technology 2009.
23. P. Vettiger et al., “The “Millipede”-More than thousand tips for future AFM storage”, IBM Journal of Research and Development, vol. 44, n. 3, pp. 323–340, 2000.
24. S.-H. Oh et al., “Novel FERAM Technologies with MTP Cell Structure and BLT Ferroelectric Capacitors”, IEEE IEDM Tech. Dig., p. 835–839, 2003.
25. H. Ishiwara, “Recent Progress in FET-Type Ferroelectric Memories”, IEEE IEDM Tech. Dig., p. 263–267, 2003.
26. M. Durlam et al., “A 0.18 um 4 Mb Toggling MRAM”, IEEE IEDM Tech. Dig., p. 995–999, 2003.
27. S. Tehrani et al., “Magnetoresistive Random Access Memory Using Magnetic Tunnel Junctions”, Proceedings of the IEEE, vol. 91, n. 5, p. 703–714, 2003.
28. J. C. Slonczewski, “Current-Driven Excitation of Magnetic Multilayers”, Journal of Magnetism and Magnetic Materials, vol. 159, n. 1–2, p. L1–L7, 1996.
29. C. Demerjian, “Everspin Makes ST-MRAM a Reality”, “LSI AIS 2012: Non-volatile Memory with DDR3 Speeds”, SemiAccurate.com, November 16, 2012.
30. S. Chung et al., “Fully Integrated 54 nm STT-RAM with the Smallest Bit Cell Dimension for High Density Memory Application”, IEEE IEDM Tech. Dig., p. 12.7.1–12.7.4, 2010.
31. S. R. Ovshinsky, “Reversible Electrical Switching Phenomena in Disordered Structures”, Phys. Rev. Lett., vol. 21, p. 1450, 1968.
32. R. G. Neale et al., “Nonvolatile and Reprogrammable, the Read-Mostly Memory is Here”, Electronics, p. 56, Sept., 1970.
33. G. Wicker, “Nonvolatile, High Density, High Performance Phase Change Memory” SPIE Conf. on Elect. and Struc. for MEMS, Australia, 1999.
34. F. Pellizzer et al., “Novel μTrench Phase-Change Memory Cell for Embedded and Stand-Alone Non-Volatile Memory Applications”, Symp. on VLSI Tech., p. 18–19, 2004.
35. F. Ottogalli et al., “Phase-Change Memory Technology for Embedded Applications”, Proc. ESSDERC 04, p. 293–296, 2004.
36. F. Bedeschi et al., “A Multi-Level-Cell Bipolar-Selected Phase-Change Memory”, Solid-State Circuits Conference, ISSCC, p. 428, 2008.
37. G. Servalli, “A 45 nm generation Phase Change Memory technology”, IEEE IEDM Tech. Dig., p. 1–4, 2009.
38. Micron Press Release, “Micron Announces Availability of Phase Change Memory for Mobile Devices”, July 18, 2012.
39. C. Villa, D. Mills, G. Barkley, H. Giduturi, S. Schippers, D. Vimercati, “A 45 nm 1 Gb 1.8 V Phase-Change Memory”, Solid-State Circuits Conference, ISSCC, p. 270–271, 2010.
40. R. F. Freitas and W. W. Wilcke, “Storage-Class Memory: The next Storage System Technology”, IBM Journal of Research and Development, vol. 52(4/5), p. 439–448, 2008.
41. T. Shintami et al., “Properties of Low-Power Phase-Change Device with GeTe/Sb2Te3 Superlattice Material”, EPCOS 2011, p. 110, 2011.
42. K. Szot et al., “Localized Metallic Conductivity and Self-Healing during Thermal Reduction of SrTiO3”, Phys. Rev. Lett., vol. 88, n. 7, p. 075508, 2002.
43. T. Iizuka-Sakano et al., “Stability of the Staging Structure of Charge-Transfer Complexes Showing a Neutral–Ionic Transition”, Phys. Rev. B, vol. 70, n. 8, p. 085111, 2004.

44. B. J. Choi et al., “Resistive Switching Mechanism of TiO2 Thin Films Grown by Atomic-Layer Deposition”, J. Appl. Phys., vol. 98, n. 3, p. 33715, 2005.
45. M. N. Kozicki et al., “Nonvolatile Memory Based on Solid Electrolytes”, NVMTS 2004.
46. Panasonic Press Release, “The New Microcontrollers with On-Chip Non-Volatile Memory ReRAM”, May 15, 2012.
47. T.-Y. Liu et al., “A 130.7mm2 2-Layer 32 Gb ReRAM Memory Device in 24 nm Technology”, Solid-State Circuits Conference, ISSCC, pp. 210, 2012.