Materials Science-Poland, Vol. 25, No. 1, 2007

Scaling of nonvolatile memories to nanoscale feature sizes*

T. MIKOLAJICK**, N. NAGEL, S. RIEDEL, T. MUELLER, K.-H. KÜSTERS

Qimonda Dresden GmbH & Co. OHG, Technology Center flash QD TC FL, Dresden, Germany

The market for nonvolatile memory devices is growing rapidly. Today, the vast majority of nonvolatile memory devices are based on the floating gate device which is facing serious scaling limitations. Material innovations currently under investigation to extend the scalability of floating gate devices are discussed. An alternative path is to replace the floating gate by a charge trapping material. The combination of charge trapping and localized channel hot electron injection allows storing two physically separated bits in one memory cell. The current status and prospects of charge trapping devices are reviewed, demonstrating their superior scalability. Floating gate as well as charge trapping memory cells suffer from severe performance limitations with respect to write and erase speed and endurance driving system overhead. A memory that works like random access memory and is nonvolatile would simplify system design. This, however, calls for new switching effects that are based on integrating new materials into the memory cell. An outlook to memory concepts that use ferroelectric switching, magnetic switching, phase change, or other resistive switching effects is given, illustrating how the integration of new materials may solve the limitations of today's semiconductor memory concepts.

Key words: *nonvolatile memories; flash memories; NAND; NOR; organic memories; molecular memories*

1. Introduction

Driven by the demand in mobile devices, the market for nonvolatile memories is growing rapidly [1]. Figure 1 shows the market development expected until 2010. In recent years, floating gate flash memories have evolved as the mainstream nonvolatile memory solution. Traditionally, the flash market is divided into two parts. In code flash it is important to execute software directly from the flash memory, therefore fast random access is required. Typical applications for such memories are cellular phones, where the software of the phone as well as user data can be stored on the

*Presented at the joint events 1st Workshop "Synthesis and Analysis of Nanomaterials and Nanostructures" and 3rd Czech-Silesian-Saxony Mechanics Colloquium, Wrocław, Poland, 21–22 November, 2005.

**Corresponding author, e-mail: Thomas.Mikolajick@esm.tu-freiberg.de

same flash device. In the data flash arena large amounts of data are transferred between the memory and system. Performance is therefore achieved by handling large amounts of data in parallel.

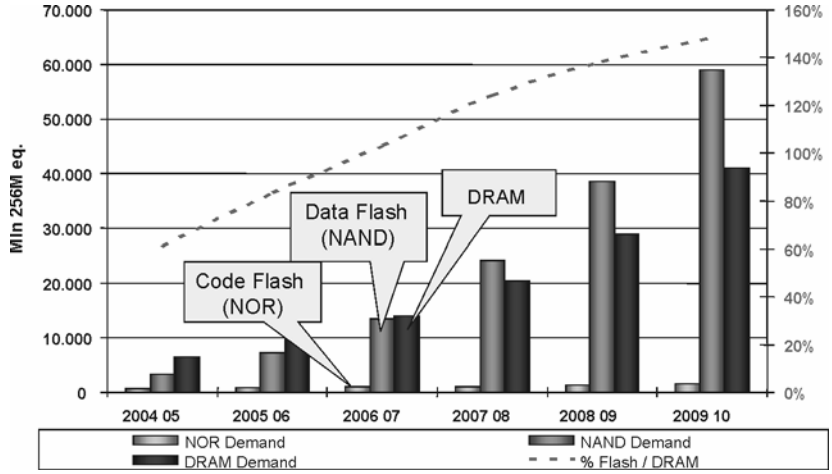

Fig. 1. Market development in terms of bit consumption for code and data flash memories in comparison to DRAM memories for the years 2004–2010

Typical applications here are memory cards for digital still cameras or USB sticks. Dynamic random access memories were traditionally used as the technology driver for the semiconductor industry. Since the 1Gb generation, data flash memories have caught up with DRAM and recently data flash is scaling ahead of DRAM in terms of density as well as minimum feature size [2]. To achieve the required 10 years of retention, however, the tunnel oxide cannot be scaled below 6 nm. Moreover, the coupling between floating gates will narrow down the available window between different states of a memory cell. These effects will limit further scaling of floating gate devices, which is today's workhorse of nonvolatile memories. This paper discusses the scaling down of nonvolatile memory cells, with focus on material innovations essential for extending nonvolatile memory scalability down to a feature size of tens of nanometers.

2. Floating gate devices

In Figure 2, basic structures of a floating gate memory cell are shown together with a brief explanation of the cell operation as well as main programming and erase mechanisms. The amount of charge present on a floating gate determines the threshold voltage of the transistor. By sensing the current at an appropriate gate voltage, two states of the cell can be discriminated according to the current that will flow through the cell. Electrons can be transferred to the floating gate using either channel hot electron injection or Fowler–Nordheim tunnelling. In channel hot electron programming, the current is passed trough the channel by applying both a high drain as well as

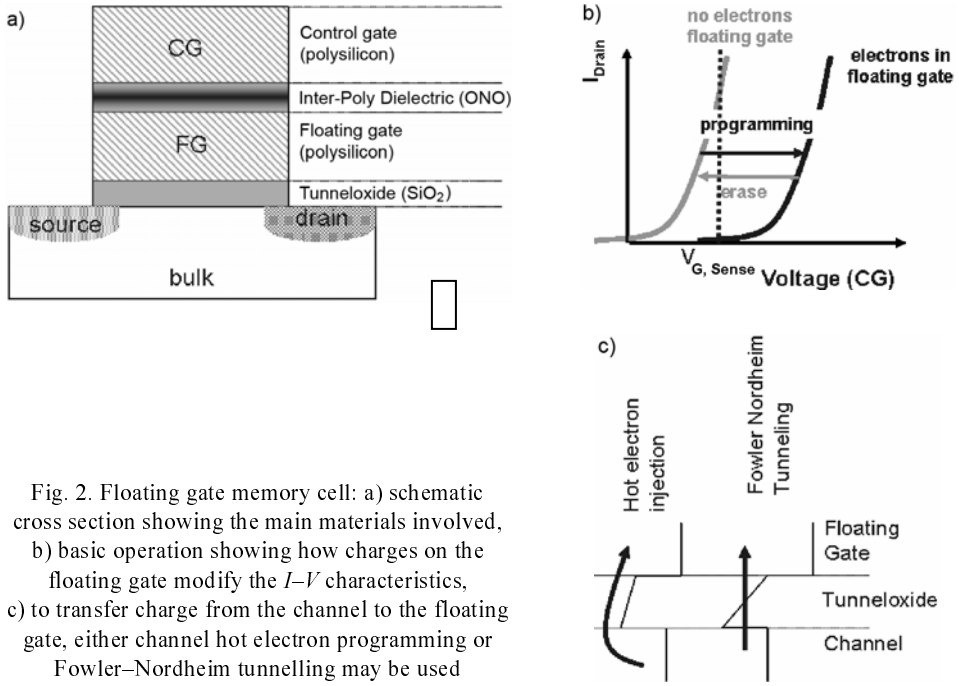

Fig. 2. Floating gate memory cell: a) schematic cross section showing the main materials involved, b) basic operation showing how charges on the floating gate modify the *I–V* characteristics, c) to transfer charge from the channel to the floating gate, either channel hot electron programming or Fowler–Nordheim tunnelling may be used

a high gate voltage with respect to the source. At the drain side of the device, some of the electrons may have enough energy to surmount the potential barrier between the silicon and tunnel oxide and can be injected into the floating gate. In Fowler–Nordheim tunnelling, a high field is applied between the channel and floating gate, leading to the reduction of the effective barrier for electrons.

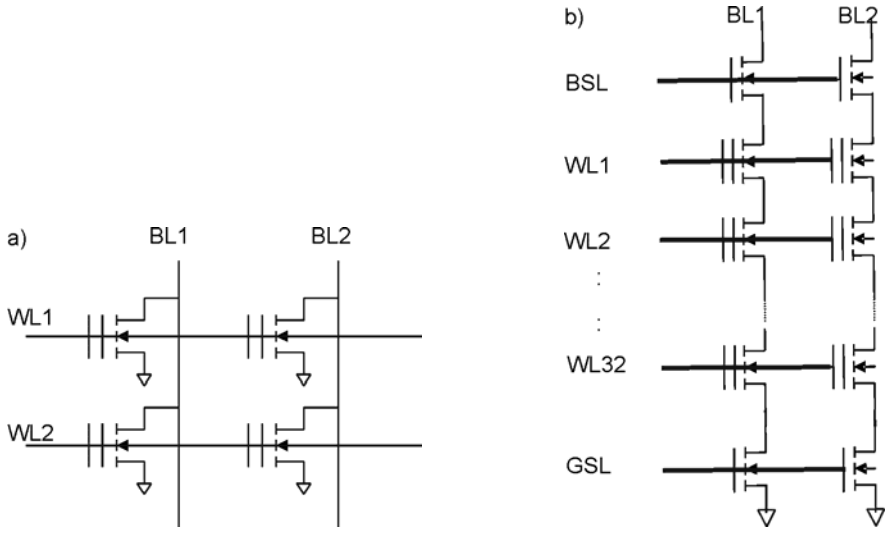

Fig. 3. Basic architectures for flash memory arrays: a) NOR, b) NAND

Figure 3 shows the main array architectures which can be implemented in this basic cell [3]. In the NOR architecture, each cell is connected to a separate bitline by a bitline contact. This allows fast random access. In the NAND architecture, however, an individual cell is connected to the bitline through a string of 16 or 32 cells. This leads to a very small physical cell size, since contacts to the source and drain regions are shared between all 16 or 32 cells of one NAND string. The high series resistance created by the series connection of the cells leads to slow random access, which has to be compensated by massive parallelisation.

Looking into the future, floating gate memories are facing serious scaling limitations. A general issue is the non-scalability of the tunnel dielectric. To maintain the required nonvolatile retention, the tunnel dielectric has to be thicker than 6 nm [4]. Further scaling can only be obtained by radically reengineering the tunnel barrier. Materials with a higher electric permittivity constant, such as HfO_2, ZrO_2, etc., which are currently investigated in logic transistors, could help. Crested barriers [5] could further improve the basic memory cell by increasing the ratio between on and off currents, leading to much faster write times as well as lower programming voltages.

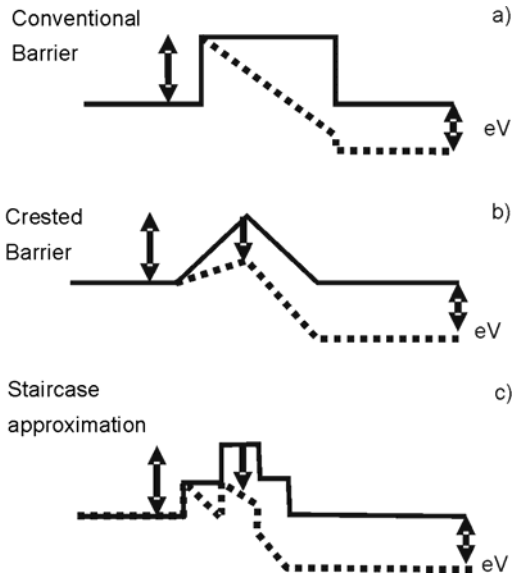

Fig. 4. Schematic presentation of crested barriers: a) conventional barrier, b) crested barrier, c) an approximation of a crested barrier by a staircase function

Figure 4 shows the principle of such an approach. Since a crested barrier is not achievable with materials having the required barrier heights, a staircase approximation using three layers with different band offsets as well as different electric permittivities is a reasonable approach. In the optimum structure, the centre layer would have a high band offset and a high electric permittivity and the surrounding layer a lower band offset as well as a lower electric permittivity. In most materials, how-

ever, a high band offset is correlated with a low electric permittivity and vice versa, making the optimum choice very difficult. A stack consisting of $Si_3N_4/Al_2O_3/Si_3N_4$ could be a reasonable and producible compromise [6]. Another serious constraint is that in the current cell architecture, the inter-poly dielectric together with the wordline has to fit into the space between two floating gates (see Fig. 5a).

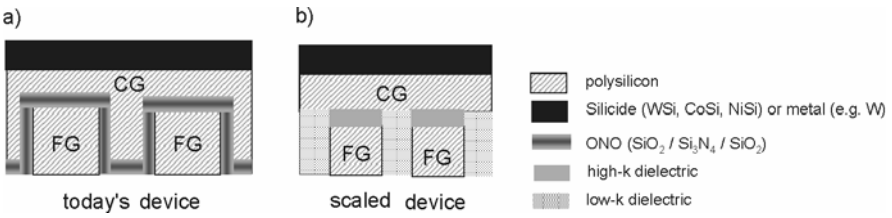

Fig. 5. Cross section of a floating gate cell along the wordline: a) today's solution with an ONO interpoly-dielectric, b) scaled down version with a high-*k* interpolydielectric and low-*k* decoupling dielectric

With the currently used triple dielectric consisting of $SiO_2/Si_3N_4/SiO_2$, with a total thickness of about 15–20 nm, this will limit cell scaling. Without using the floating gate sidewalls a high-*k* dielectric will be required to achieve the necessary coupling between the control gate and floating gate. Furthermore, a low-*k* dielectric will be necessary to decouple two neighbouring floating gates (see Fig. 5b). Typical materials are very similar to the ones discussed for gate dielectrics in conventional MOS transistors, including HfO_2 and Hf/Al microlaminates [7]. While scaling down the floating gate device, the spacing between floating gates will continuously decrease. This leads to a higher capacitive coupling between floating gates, resulting in cross talk between cells. This calls for a material with a lower electric permittivity between the floating gates, like that already shown in Fig. 5b, which also has to be implemented in the area between the word lines. Replacing the silicon nitride spacer of the cell transistor by a silicon dioxide spacer as shown in [8] may already help to significantly reduce the effect. In the long term, real low-*k* materials will be necessary.

3. Charge trapping devices

A natural way to extend the scalability of a floating gate device is to replace the charge storing floating gate by a dielectric material, in which the charges are stored in deep traps. The main drawback of this approach is that electrons that are erased via the bottom oxide by either electron or hole tunnelling may be replaced by electrons tunnelling from the control gate trough the top oxide to the nitride. This will lead to erase saturation, which limits the erase speed at a given thickness of the bottom oxide. A very thin bottom oxide of the order of 1 nm is not practical, since the retention requirement cannot be achieved. This was the biggest obstacle for the commercial success of charge trapping devices. New materials can greatly help improve this issue. A high-*k* top oxide can reduce the voltage drop across the top oxide and a high work

function gate can increase the potential barrier for electrons that travel across the top oxide. For the top oxide, Al_2O_3 or a combination of Al_2O_3 and HfO_2 are the best candidates [9], and As gate electrodes p^+ have the potential of drastically improving the situation [10]. Poly-depletion, however, may limit the actual gain in this approach. Since materials like Pt or Ir, which would be very well suited from a work function point of view, are hard to integrate into a CMOS flow, TaN seems to be a very good choice [11]. The combination of both approaches allows for an erase speed like in charge trapping memory cells similar to tht in NAND flash [12]. For the charge trapping material itself, silicon nitride has been well established for many years. Silicon oxinitride [13] as well as hafnium oxide and aluminium oxide [14], however, are possible alternatives with potential benefits.

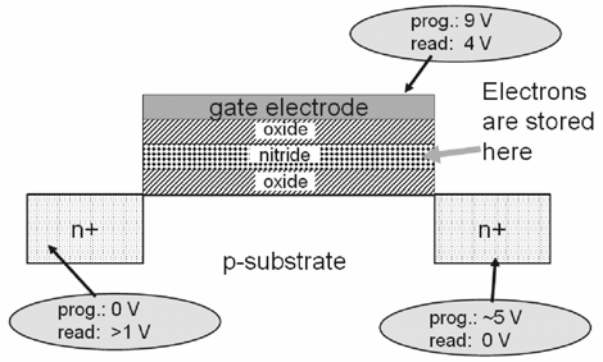

Fig. 6. Schematic of a multi-bit charge trapping memory cell illustrating the programming, erase, and read operations

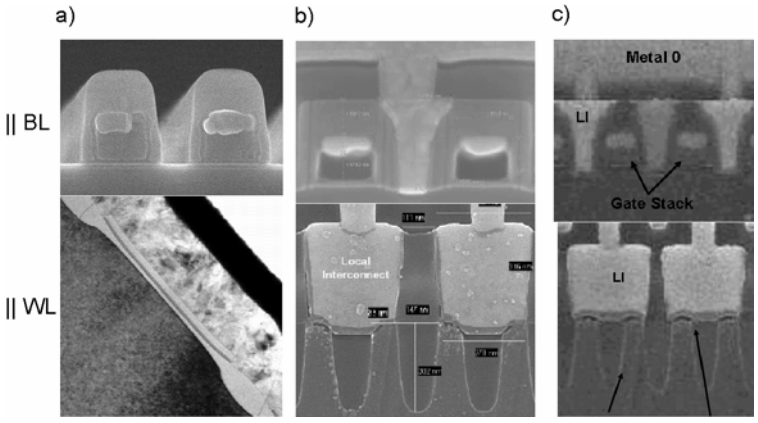

Fig. 7. Cross section parallel to bitline (top row) and parallel to WL (bottom row) of TwinFlash memory cells of the: a) 170 nm, b) 110 nm, and c) 90 nm generations

Another way of solving the erase saturation issue is to change the erase mechanism to hot hole injection [15]. If hot electrons are used for the programming, then

two bits can be stored and physically separated in a single cell [16]. Figure 6 illustrates such a multi-bit charge trapping memory cell as well as its basic programming, erase and read functions. Figure 7 shows the real cross sections of three generations of TwinFlash, which is an advanced version of the multi bit charge trapping concept [17, 18].

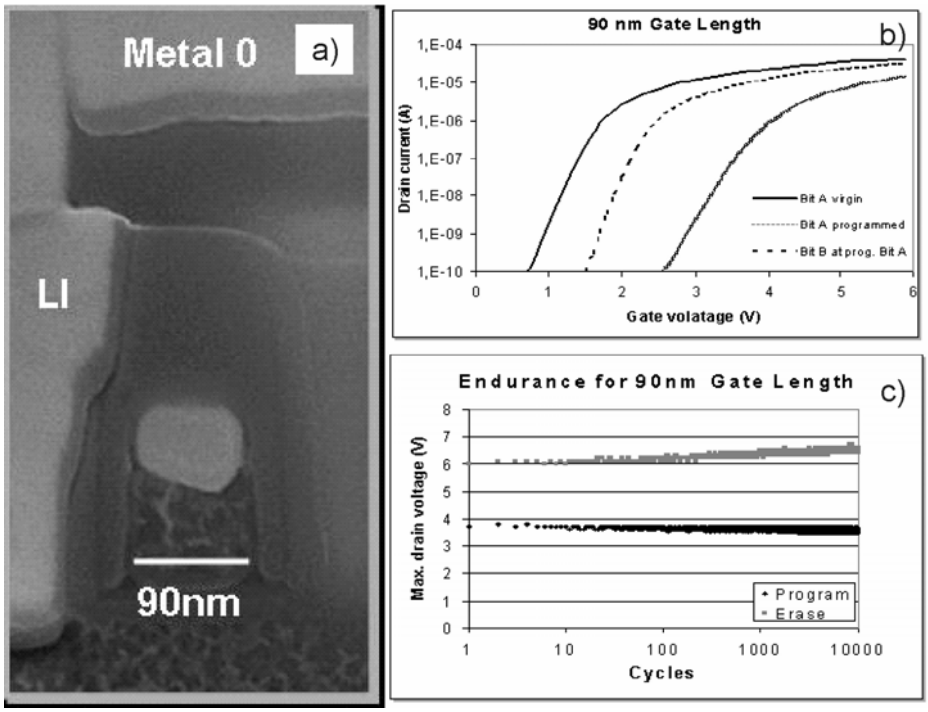

Fig. 8. TwinFlash memory cell from the 60 nm generation: a) SEM cross section, b) *I–V* characteristics of native and programmed cell, c) cycling behaviour

Figure 8 demonstrates the scalability of this type of cell down to the 60 nm node. Further scaling down to about 40 nm groundrules is possible using standard approaches [18]. For even smaller groundrules, 3D devices can help overcome the scaling issues [19]. Again, high-*k* materials that replace existing charge trapping and barrier materials may further extend the scalability of also this type of device [20].

4. Alternative memories

All charge-based nonvolatile memories described in the previous chapters suffer from the fact that a high potential barrier is needed to achieve nonvolatile retention. The barrier, however, has to be overcome by charges during programming and erasing operations. This contradiction leads to severe performance drawbacks of all charge-

based memory concepts that include the necessity for high programming and erase voltages (in the range 10–20 V), slow write and erase times (from µs up to seconds in contrast to ns, which are common in random access memories), and very limited endurance (typically up to 10^6 cycles; 10^{16} cycles are required for a random access memory). From a system point of view, a random access type of memory that is nonvolatile would be of great benefit. To achieve such a memory, new switching effects realized in new materials are required [21].

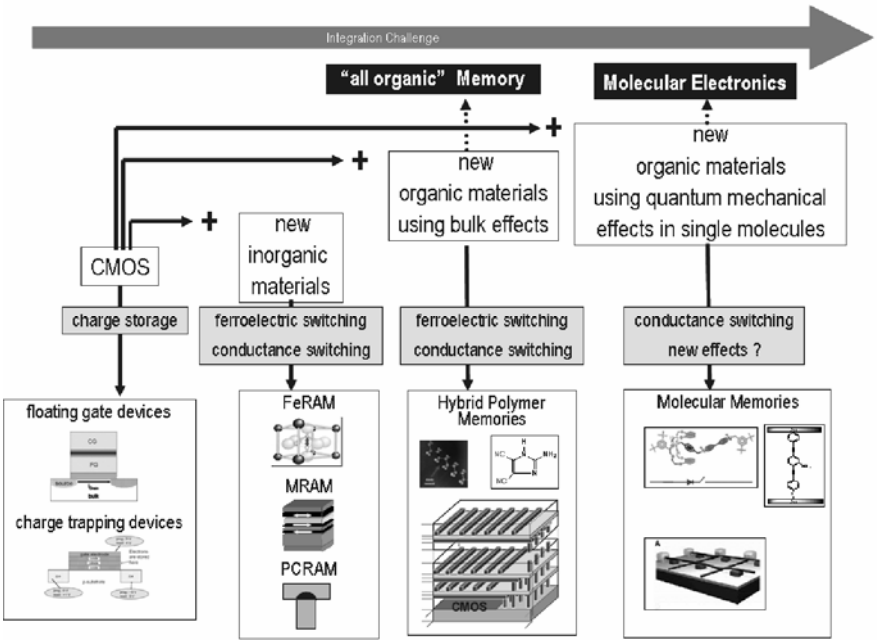

Fig. 9. Hierarchy of alternative nonvolatile memories from a materials perspective

Figure 9 gives an overview over a number of concepts discussed in the literature from a material point of view. In general, the concepts can be classified [22] into concepts that use switching in inorganic materials, concepts that use bulk effects in organic materials (referred to as organic memories in this overview), and concepts that use quantum mechanical effects in single molecules (referred to as molecular memories). Due to a scaling potential down to the molecular level, memories based on carbon nanotubes are included in the later class. A detailed overview of the material aspects of many possible options can be found in Chapter 3 of [22]. Due to their similarity to CMOS processing, the concepts that use inorganic switching materials are the most advanced. Among them ferroelectric memories (FeRAM), magnetoresistive memories (MRAM), and phase change memories (PCRAM) are close to production or already in production for niche applications. FeRAM [23, 24] uses the switchable electrical polarization of ferroelectric materials such as lead-zirconium titanate (PZT) and strontium-bismut tantalate (SBT) to store information. The main

challenge is the material integration of the ferroelectric material and electrodes. Since the ferroelectric has to maintain the right phase, high temperature annealing is necessary and exposure to hydrogen needs to be avoided. Moreover, since the charge transferred during polarization switching is detected, a three dimensional structure is necessary in order to maintain a minimum sensed charge when the device is scaled to nanometer dimensions [25]. In MRAM [26], tunnel magnetoresistance is used to distinguish between different states. This approach consists of a thin tunnelling dielectric like Al_2O_3 or MgO placed between two ferromagnetic electrodes. One of the two ferromagnetic electrodes is pinned to an antiferromagnetic layer to define a reference. The resistance of the stack then depends on the orientation of the magnetization in both electrodes with respect to one another. A resistance ratio of 50-200% can be obtained when the magnetizations of the two layers are parallel (lower resistance case) or when they are antiparallel (higher resistance case). Writing is traditionally done by passing current through the word and digit lines, leading to a situation where the superposition of the two magnetic fields at the intersection of both lines is high enough to switch one of the two ferromagnetic layers, while the field generated by each of the lines separately is not high enough to change the magnetization state. This approach, however, requires high currents (in the mA range) to be passed trough the lines, and is therefore a major scaling limitation. Spin transfer switching, where a current is passed trough the tunnelling barrier, has recently been proposed to overcome this issue [27]. Even in this approach the required switching current has to be further reduced. PCRAM is based on the reversible phase change of chalcogenide materials such as $Ge_2Se_2Te_5$ between high resistive amorphous and low resistive crystalline phases [28]. Integration as well as scalability is much simpler than in FeRAM and MRAM. Some issues remain, however, the most prominent one being the reset current that is required to melt the material in the process of transforming the crystalline phase to an amorphous one. Other issues are the asymmetric write/erase as well as the still limited endurance. Recently, a 64Mb memory was demonstrated using 0.12 μm technology [29].

More challenging than the integration of inorganic materials is the integration of organic materials. Here, a wide variety of concepts have been shown in literature [30–32]. Polymer ferroelectrics seem to be the most advanced. Nonetheless, resistive switching devices promise better scalability.

Switching in single molecules directly paves the way to the nanoscale world. Rotaxane [33], porphyrines [34], and phenyl-based molecules with attached nitro-redox centres [35] are among the most prominent molecules that show promising switching effects. In all these, however, the defined contact with the outside world and integration with CMOS logic are challenges that remain to be solved. An alternative path to molecular memories is the utilization of carbon nanotubes. A mechanical memory based on a cross bar arrangement of nanotubes, separated by support pillars and switched by electrostatic forces, is among the most frequently discussed concepts in this direction [36]. An alternative is the implementation of a charge trapping device using a carbon nanotube transistor. Carbon nanotubes, however, have basic uncertainties with respect to reproducible mass production.

5. Summary and conclusions

Driven by the demand for more mobile electronic devices, the market for nonvolatile memories is growing rapidly. Today, floating gate flash is the mainstream solution for nonvolatile memories. Floating gate devices face serious scaling limits in the sub 50 nm region, however, and new materials will be required to scale below 50 nm, especially high-*k* dielectrics to increase coupling and low-*k* dielectrics to reduce unwanted coupling between neighbouring cells. Charge trapping devices are an alternative that have developed very rapidly in the past few years. Especially the multibit charge trapping concept has appeared in significant production volumes for both code and data flash products. The scaling of this type of device will continue to the 40 nm generation without major material innovations. In order to overcome the basic limitations of all charge-based nonvolatile memories, new switching materials have to be integrated into the CMOS process flow. Inorganic approaches such as ferroelectrics, magnetoresistive switching, as well as material phase change, are in advanced development stages. They are all, however, far from reaching the small cell size benchmark set for data flash memories. Therefore, they will appear where their performance advantage comes into play or as an alternative to code flash or SRAM memories. For the long-term scalability, organic as well as molecular memories promise to extend nonvolatile memories beyond classical CMOS scaling. Hybrid memories that combine CMOS with organic or molecular memory cells, as well as memories where even the necessary electronic circuits are replaced by organic or molecular circuits, can be envisioned as two possible development steps on the long-term roadmap as indicted in Fig. 9. In the short to mid term, classical charge-based memories fabricated with the CMOS technology will continue to dominate the market, especially in data flash devices.

References

[1] NIEBEL A., Proc. of the 20th Nonvolatile Semiconductor Memory Workshop, Monterey, California (2004), p. 14.
[2] BYEON D.-S., LEE S.-S., LIM Y.-H., PARK, J.-S., HAN W.-K., KWAK P.-S., KIM D.-H., CHAE D.-H., MOON S.-H., LEE S.-J., CHO H.-C., LEE J.-W., KIM M.-S., YANG J.-S., PARK Y.-W., BAE D.-W., CHOI J.-D., HUR S.-H., SUH K.-D., Proc. Int. Solid State Circuits Conference, IEEE, San Francisco (2005), p. 46.
[3] PAVAN P., BEZ R., OLIVO P., ZANONI E., Proc. IEEE, 85 (1997), 1248.
[4] LAI S., Proc. Seventh Biennial International Nonvolatile Memory Technology Conference, IEEE, Albuquerque (1998), p. 6.
[5] LIKHAREV K., Appl. Phys. Lett., 73 (1998), 2137.
[6] CASPERSON J., J. Appl. Phys., 92 (2002), 261.
[7] LEE W.-H., CLEMENS J.T., KELLER R.C., MANCHANDA L., VLSI Technology Digest of Technical Papers (1997), p. 117.
[8] LEE J.-D., SUNG-HOI H., CHOI J.-D., IEEE Electr. Device Lett., 23 (2002), 264.
[9] CHOI S., CHO M., HWANG H., KIM J.W., J. Appl. Phys., 94 (2003), 5408.
[10] BACHHOFER H., REISINGER H., BERTAGNOLLI E., VON PHILIPSBORN H., J. Appl. Phys., 89 (2001), 2791.
[11] LEE C.-H., PARK K.-C., KIM K., Appl. Phys. Lett., 86 (2005), 73510.

[12] SHIN Y., CHOI J., KANG C., LEE C., PARK K.-T., LEE J.-S., SEL J., KIM V., CHOI B., SIM J., KIM D., CHO H.-J., KIM K., IEDM Digest Techn. Papers, IEEE (2005), p. 327.
[13] ISHIMARU T., MATSUZAKI N., OKUYAMA Y., MINE T., WATANABE K., YUGAMI J., KUME H., ITO F., KAWASHIMA Y., SAKAI T., KANAMARU Y, ISHII Y., MIZUNO M., ISHII Y., MIZUNO M., KAMOHARA S., HASHIMOTO T., OKUYAMA K., KURODA K., KUBOTA K., IEDM Digest Techn. Papers. IEEE (2004), p. 885
[14] TAN Y.-N., CHIM W.-K., BYUN J.C., WEE-KIONG C., IEEE Trans. Electr. Devices 51 (2004), 1143.
[15] CHAN T.Y., IEEE Electr. Device Lett., 8 (1987), 93.
[16] EITAN B., PAVAN P., BLOOM I., ALONI E., FROMMER A., FINZ D., IEEE Electr. Device Lett., 21 (2000), 543.
[17] NAGEL N., OLLIGS, D., POLEI, V., PARASCANDOLA S., BOUBEKEUR H., BACH L., MULLER T., STRASSBURG M., RIEDEL S., KRATZERT P., CASPARY D., DEPPE J., WILIER J., SCHULZE J., SCHULZE N., MIKOLAJICK T., KUSTERS K.-H., SHAPPIR A., REDMARD E., BLOOM I., EITAN B., VLSI Techn. Digest Techn. Papers, Kyoto (2005), p. 120.
[18] STEIN V., KAMIENSKI E.G., ISLER M., MIKOLAJICK T., LUDWIG C., SCHULZE N., NAGEL N., RIEDEL S., WILLER J., KÜSTERS K.-H., Proc. Non-Volatile Memory Technology Symposium, Dallas, 2005, p. 5.
[19] WILLER J., LUDWIG C., DEPPE J., KLEINT C., LAU F., PALM H., EITAN B., BLOOM I., Proc. 19th Non-volatile Semiconductor Memory Workshop, Monterey, California (2003), p. 42.
[20] SUGIZAKI T., KOBAYASHI M., ISHIDAO M., MINAKATA H., YAMAGUCHI M., TAMURA Y., SUGIYAMA Y., NAKANISHI T., TANAKA H., VLSI Techn. Digest Techn. Papers, Kyoto (2003), p. 27.
[21] PINNOW C.-U., MIKOLAJICK T., J. Electrochem. Soc., 151 (2004), K1.
[22] *Materials for Information Technology*, E. Zschech, C. Whelan, T. Mikolajick (Eds.), Springer, London, 2005, p. 112.
[23] MIKOLAJICK T., DEHM C., HARTNER W., KASKO I., KASTNER M.J., NAGEL N., MOERT M., MAZURE C., Microelectronics Reliability, 41 (2001), 947.
[24] LEE S.Y., Extended Abstracts of the International Conference on Solid State Devices and Materials, Kobe, Japan (2005), p. 1026.
[25] KOO JU.M., SEO B.-S., KIM S., SHIN S., LEE J.-H., BAIK H., LEE.J.-H., LEE J.H., BAE B.-J., LIM. J.-E., YOO D.-C., PARK S.-O., KIM H.-S., HAN H., BAIK S., CHOI J.-Y., PARK Y.J., PARK Y., IEDM Digest Tech. Papers, IEEE (2005), p. 4.
[26] GALAGHER W.J., IEEE VLSI-TSA Int. Symp. VLSI Technology, Kyoto, Japan (2005), p. 72.
[27] HOSOMI M., YAMAGISHI H., YAMAMOTO T., BESSHO K., HIGO Y., YAMANE K., YAMADA H., SHOJI M., HACHINO H., FUKUMOTO C., NAGAO H., KANO H., IEDM Digest Techn. Papers, IEEE (2005), p. 459.
[28] HUDGENS S., JOHNSON B., MRS Bull. November (2004), p. 829.
[29] OH H.-R., IEEE J. Solid State Circuits, 41 (2006), 122.
[30] SECZI R., WALTER A., ENGL R., MALTENBERGER A., SCHUMANN J., KUND M., DEHM C., IEDM Digest Techn. Papers IEEE (2003), 10.2.1.
[31] YANG Y., Organic Nonvolatile Memories, in [22], p. 197.
[32] KRIEGER J.H., SPITZER S.M., Proc. Non-Volatile Memory Technology Symposium, Orlando-Florida (2004), p. 121.
[33] LUO Y., COLLIER C.P., JEPPESEN, J.O., NIELSEN K.A., DEIONNO E., HO G., PERKINS J., TSENG H.-R., YAMAMOTO T., FRASER STODDART J., HEATH, J. R., ChemPhysChem, 3 (2002), 519.
[34] ROTH K.M., DONTHA N., DABKE R.B., GRYKO D.T., CLAUSEN C., LINDSEY J.S., BOCIAN D.F., KUHR W.G., J. Vac. Sci. Technol., B 18, (2000), 2359.
[35] REED M. A., CHEN J., RAWLETT A.M., PRICE D.W., TOUR J.M., Appl. Phys. Lett., 78 (2001), 3735.
[36] RUECKES T., KIM K., JOSELEVICH E., TSENG G.Y., CHEUNG C.-L., LIEBER C.M., Science, 289 (2000), 94.

Received 3 January 2006
Revised 28 May 2006