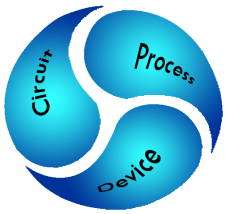

Ron@Maltiel-consulting.com - Semiconductor & Patent Expert Consulting

Litigation expert consultant and patent expert witness
for process, device, and circuit of Dynamic

Ram (DRAM), Flash (NAND, NOR, EEPROM), and Static Ram (SRAM) Memories,

and Microprocessor, Logic, and Analog Devices.

SEMICONDUCTOR FABTECH

SPECIAL FEATURES

Manufacturing technology for sub-50nm DRAM and NAND Flash memory

Kinam Kim, Samsung Electronics, Korea

ABSTRACT

This article discusses whether memory technologies can continue to evolve beyond the sub-50nm node, in particular, for dynamic random access memory (DRAM) and NAND flash memories. First, the technological barriers that need to be overcome will be addressed, based on the inherent features of the two memory types. Second, technological solutions will be introduced in detail. Last, manufacturability of each solution will be examined critically. Beyond the 30nm technology node, it is expected that novel transistor schemes with 3-dimensional structures will be incumbent upon both logic and memory array designs, along with the development of new materials.

Over the past three decades, the semiconductor memory business has flourished due to the great success of DRAM technology. With the advent of the mobile era, another growing semiconductor memory engine is flash memory, especially of the NAND variety. It is certain that the future memory business will be much brighter than in the past and the present because we have two strong sources of business momentum in DRAM and NAND flash memory to push us forward.

At the current speed of fast technology growth, 4Gbit DRAM and 16–32Gbit NAND flash memory will be produced en masse with 50nm or below technology node near 2010. Recently, much effort has been dedicated to clarifying the issues that might be raised at 50nm or below, together with suitable solutions [1, 2]. In spite of the successful manufacturing of these products while maintaining cost-effectiveness, most of the concerns come from the technical complexity, which is inevitable due to the fact that product requirements continue to grow to meet customers' needs. For instance, such requirements include: an operating speed of DDR3 DRAM of more than 1Gbps; a retention time for mobile DRAM of greater than 1sec, and programming speeds of 32Gbit NAND flash memory of more than 20MBytes/s.

Furthermore, a marginal process window and the wide spread-out of process variations will impose further challenges on successful manufacturing of 50nm DRAM and NAND flash memory. Thus, conventional 'shrink technology', which is primarily based on dimension scaling, cannot solely provide complete answers for sub-50nm DRAM and NAND flash memory. In order for successful manufacturing of memories, 'shrink technology' must be supplemented with novel approaches such as new device structures, new process technologies and new materials. In this article, not only will some challenges of 50nm technology

node be reviewed, but the solutions needed to ensure manufacturing the 50nm DRAM and NAND flash memory will also be discussed in detail.

DRAM

Since 1kbit DRAM was demonstrated by Intel in the early 1970s, DRAM has been doubled up in density every 18 months. DRAM at the 80nm technology node and 2Gbit densities is about to reach the market. Also, DRAM beyond the 50nm node is now under consideration at R&D centers around the world. In addition, DRAM technology has focused on ever-increasing retention times (refresh cycle). Such a need for retention-time increases confronts DRAM developers with great challenges in terms of sufficient cell capacitance and less leakage current from the storage junction [2–4].

To meet the memory cell capacitance requirement, the cell capacitor has moved from planar structures to 3D structures such as the stacked and trench capacitors as shown in Figure 1 [5]. On the other hand, to reduce the leakage current of the storage junction, the cell transistor has been designed to minimize sub-threshold leakage current. As a result, the cell transistor tends to have a high threshold voltage, leading to two adverse effects: worsening the leakage current of the storage junction and weakening current drivability. Therefore, the memory cell has been optimized in a trade-off between the least off-current of the cell transistor and the allowable leakage current of the storage junction. Thus, one of the most critical challenges for DRAM scaling is the constraint of prolonging the retention time, posing big questions for new cell transistor designs.

The cell capacitance is one of the most important parameters for DRAM implementation because it determines the sensing signal margin. In spite of shrinking the cell-capacitor area drastically, the capacitance of DRAM cells has maintained a value of at least 25fF/cell. This has been achieved by either increasing the cell area through a high stack or changing to a material that has a higher dielectric permittivity (high k). This leads to dramatic changes in the capacitor structures. Probably the first option represents the easier route because it avoids material challenges, which will be discussed later. Unfortunately, increased stack heights provoke mechanical instability if pushed beyond a certain limit. Figure 2 is a typical example showing a bridge between storage nodes. A new capacitor structure has successfully been developed for the 80nm technology to get rid of such bridges. The storage node of the newly developed capacitor is mechanically supported by a nitride net as shown in Figure 3 [6]. This novel structure can provide a solution to tackle persistent issues arising from a high aspect ratio stack. The details of the 'mesh' capacitor can be found elsewhere [6]. By using the mesh capacitor, or its modifications, we can extend the current capacitor materials to the next generation.

The materials approach to solving capacitor issues is to develop a cell MIM (Metal-Insulator-Metal) structure having both high-k dielectrics and metal electrodes. The development of deposition technology is making it easier to realize a high-k dielectric capacitor. For example, the ALD (atomic layer deposition) process has many advantages not only because it provides a high-quality film but also through an excellent step coverage on top of the metal electrode, even at a very high aspect ratio. From the technological point of view, the AHO being used now at 90nm can be replaced with the higher-k HAH at 80nm, ZrO at 60nm, and STO or BST with a TiN electrode at 50nm. As an aside, we emphasize that the TiN electrodes might be changed at some point for the noble metal Ru in order to ensure a higher-k capacitor. However, in practice, the integration complexity of a Ru-based capacitor might cause many problems for a mass production implementation. Alongside high-k, enlarging the capacitor area could extend current integration technology down to the 30nm node as well.

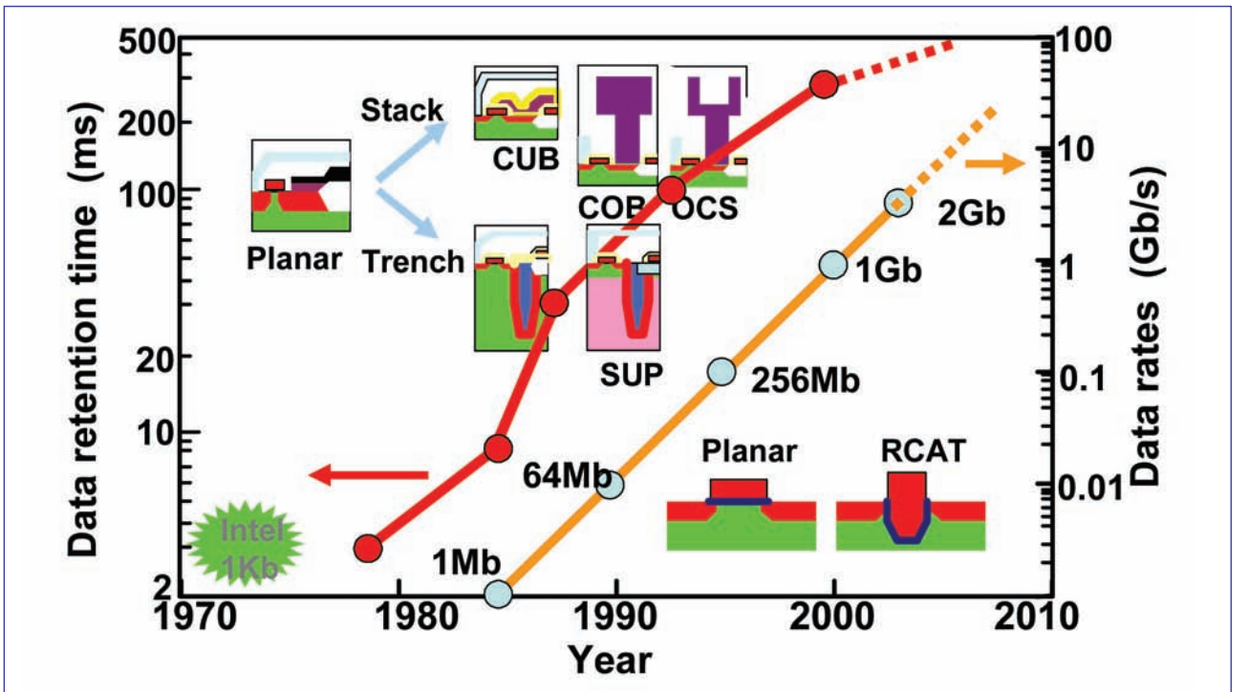

Figure 1. Evolution trend for DRAM technology in terms of data retention time and data rate (speed).

There is another persistent DRAM limit – the performance trade-off of the cell transistor. This is because the cell array's data retention is affected critically by the properties of this transistor. As a device is scaled in dimension, the doping concentration of the transistor substrate is likely to increase, e.g. to ~10^{18}/cm^3 in order to improve short-channel effect (SCEs), leading to an increase of the electric field across the junction involved. The increased junction electric field induces a high leakage current that results in the degradation of the retention time, which becomes very significant below 100nm due to a rapid increase in the junction field.

This issue can be tackled by introducing a 3-dimensional (3D) structure for the cell transistor. Recently, we have successfully developed a 3D transistor, a so-called RCAT (Recessed Channel Array Transistor), where the junction electric field is greatly suppressed due to a relatively low channel doping concentration [7, 8]. In RCATs, the effective gate length can be lengthened because the drain-to-source current flows along the recessed gate, i.e. the current path detours around the gate, thus solving the SCE without area penalty. As a consequence, we can improve data retention [8]. Investigations of the scalability of RCAT tell us that it is very promising down to 50nm node with minor modifications. One of the modified versions of RCAT is S-RCAT that has a lower threshold voltage.

Beyond 50nm, we need another breakthrough because the sub-threshold swing factor does not allow us to obtain proper cell transistor performance. Use of a body-tied FinFET is strongly indicated due to its excellent immunity against SCEs, a high trans-conductance and, most importantly, a superb sub-threshold swing [9, 10]. Apart from the FinFET, a transistor with an elevated source/drain structure formed by selective epitaxy can give another solution for engineering an extremely low-doped channel. Transistors with these structures can provide not only greatly suppressed SCEs in device performance but wider windows for cell contacts in the following process integration. With this approach, scaling down to 30nm node has not been much trouble. Beyond that, we believe that a vertical transistor with a surrounding gate structure would be strongly recommended because it does not have constraints caused by lithography, and also gives better performance as a cell array transistor. *It should be noted that we keep an open mind on the eventual use of a cross-point transistor architecture for the cell array, on the grounds that these will be entangled with an extreme design rule.*

NAND flash

A big market for NAND flash memory has been anticipated since its first appearance at the 1984 International Electron Devices Meeting (IEDM). This memory device has the smallest cell size ($4F^2$ or $2F^2$: F is a feature size: bit-line width or space in nm) among commercially available memories due to its simple one-transistor structure and the serial connection of each cell in a string. However, a true take-off for NAND flash memory has yet to arrive because mobile applications have only just started.

Ubiquitous NAND flash memory has likely been pushed toward a higher node and much higher density together with higher performance features such as improved programming throughput (Figure 4). The key technology to achieving higher density is simply to build smaller memory cells. This can also be achieved by either increasing the string size or developing technology with 2 or more bits per cell.

NAND flash technology is at the 55nm technology node and 16Gbit densities, with mass production due around 2007. Not far into the future, several candidates for improvement are likely such as upgraded versions of 2-bit cells, multi-bit cells, and other sophisticated cell technologies like stacking multiple cells in the vertical dimension. However, as NAND flash

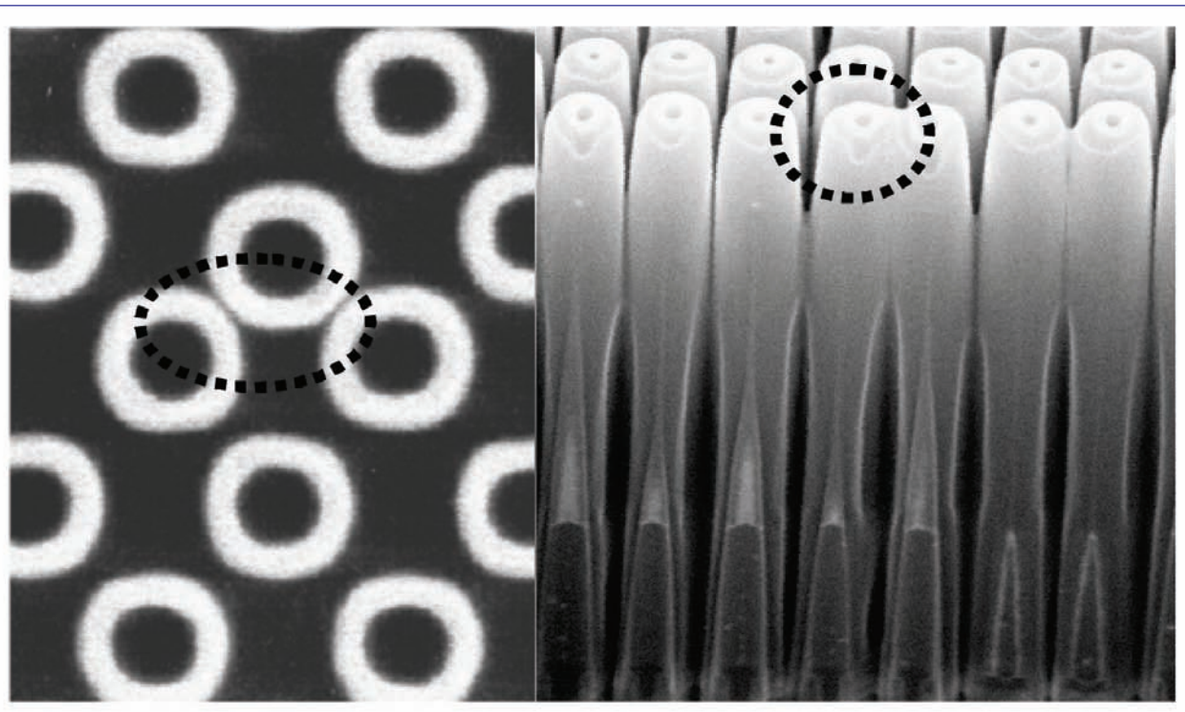

Figure 2. SEM micrographs illustrating cell-to-cell bridge as marked by the dotted line due to mechanical instability of a DRAM capacitor with a very high aspect ratio.

memory enters the technology node of 50nm or less, memory based on the floating gate (FG) structure will face serious scaling problems [1, 12]. These problems are often posed not only by physical aspects of the cell structure but also by the cell transistor performance. The physical constraints include the fact that both tunnel oxide and inter-poly ONO dielectrics have already reached their lower thickness limit.

As the cell size scales down, it becomes too narrow to lay out the space between the floating gates, making the control gate difficult to lay out in the confined area. This will probably lead to large interference between floating gates. Therefore, as scaling proceeds, the width and spacing of the floating gates converge to an equal size. This then makes it difficult to thin the floating gate because it increases its aspect ratio. In other words, there is a very critical point one should consider in terms of scaling limits: the bit-line pitch $2F$ should contain one floating gate, one control gate, and two ONO films between them. Since the ONO thickness reached its limit of 15–18nm, there is not much room to reduce and the floating gate width is extremely difficult to take below 10–15nm because of the depletion effect of the poly-silicon layer, which becomes pronounced at thinner dimensions. In practice, all these considerations lead to a 40nm node scaling limit for FG NAND flash memory.

Another thing we have to consider in dimension scaling is the floating gate's mechanical stability. The aspect ratio of floating gate tends to increase in order to minimize cell-to-cell interference while maintaining an adequate coupling ratio. Beyond the 50nm node, floating gate mechanical stability could be seriously compromised. Much worse, if the control gate is included, the mechanical stability of the total gate stack becomes very critical. To constitute a reliable product, the floating gate aspect ratio had better be less than 1.5–2. Taking into account both limits, scaling of FG NAND flash memory below 40nm is not practical at this moment.

Another problem when scaling down FG NAND flash is the cell-to-cell interference as pointed out above. As the space between word lines becomes narrower and narrower, the cell-to-cell coupling ratio increases. We cannot simply dismiss this because it gives an adverse effect on the V_{th} window, especially in the case of multi-level cells (MLCs). Potential techniques to reduce the cell-to-cell coupling interference from the integration point of view include using inter-dielectric materials with lower permittivity, reducing the stack height, modifying the floating gate structure and changing design schemes. Nevertheless, the cell-to-cell interference will eventually slow FG NAND flash memory scaling.

Probably the first difficulty for FG NAND flash memory scaling will be a drop of the coupling ratio. Below 50nm, the coupling ratio decreases rapidly because there is no room to enlarge the area of the floating gate unless the stack height is increased. The choice is between severe interference between floating gates or less mechanical stability. Another effect of decreased coupling ratios could be a higher voltage for programming and erasing operations. This might cause device scaling to be even more difficult because higher voltage means less reliability.

As design rules shrink, one of the most fundamental issues in FG NAND flash memory is the reduction of stored charge, and hence the number of electrons. Below 40nm, it is likely that less than a few hundred electrons will be stored in each memory cell. We believe that the FG NAND flash

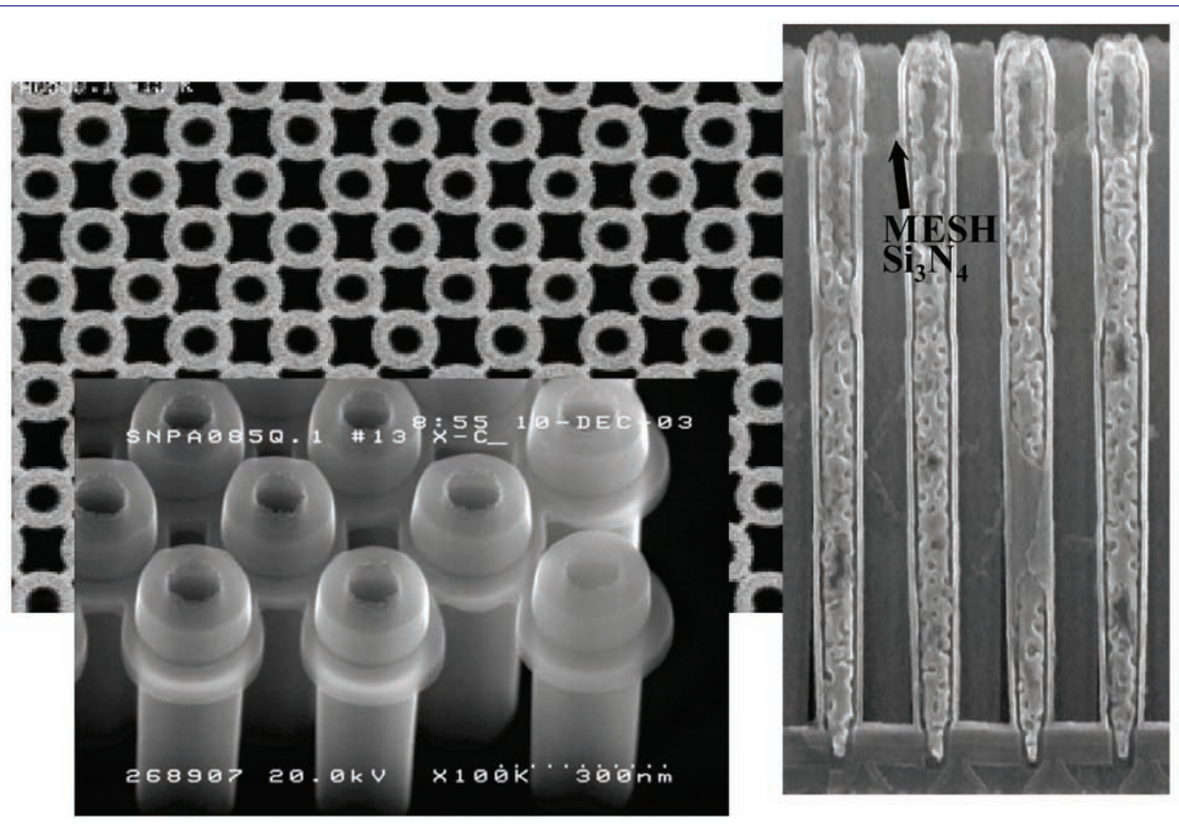

Figure 3. SEM images of the MESH (Mechanically Enhanced Storage node Height) capacitor using a nitride net, which substantially improve mechanical stability of the DRAM capacitor with a high aspect ratio.

memory will face serious electron-storage problems at 40nm node because the total charge stored in a single floating gate will not be enough for an adequate retention period.

Partial solutions of the aforementioned scaling limits of the FG NAND flash memory will come either from introducing high-k inter-poly dielectrics to the cell structure or from a new floating gate structure. But we believe that these kinds of modifications will only serve as temporary solutions between 60 and 40nm.

However, we need not only solutions to the 40nm node, but even to beyond the 25–20nm node, and this going beyond FG NAND flash memory. Any attempts to do this must consider the following: first, the most critical issue is cell-to-cell interference, which should be either absent or at least as little as possible; second, the coupling ratio should be sufficient for fast program and erase performance while maintaining reliability; last, it should contain as many electrons as possible, so that we can have multi-bit cell capability and also better on-current behaviour. In addition, it would be advantageous to have some clues on how to scale both the physical dimensions and the mechanical stability of the stack.

With all these requirements, are there any possible candidates for replacing the FG NAND? In particular, what about FG MLC NAND? There are indeed many proposals: charge-trap; nano-dot; and nano-crystal. However, it has been found that nano-crystals or nano dots cannot contain enough charges so that they will probably not be the first choice. According to the research efforts aimed at finding the solution, the most likely next step is a charge trap NAND flash memory based on a nitride trap. There are several reasons for this: first, the structure is able to store more electrons than other candidates; second, it is free from cell-to-cell interference; and, finally, nitride film is not only a relatively well-known material but has also been successfully used in CMOS technology. However, there are two big quandaries for the charge-trap-based device to tackle: one is retention period and the other is speed concerns for the erase operation.

Recently, we have successfully developed a 4Gbit TANOS NAND flash memory with 63nm technology [12]. Taking into account the MLC option, the chip can readily double its density. In the TANOS structure (Figure 5) TaN serves as the top electrode, AlO as the top oxide, nitride as the charge trap, and SiO_2 as the bottom oxide. In order not just to meet the requirements for erase speed but to obtain a wide V_{th} window, we pioneered a novel structure with a high-k dielectric as top oxide and TaN as top electrode. From this approach, we can achieve several essential properties for a flash memory: higher programming and erasing performance characteristics and a wide V_{th} window that is adequate for MLC operation. It is the first demonstration that a non-FG type of NAND flash memory is able to operate with an MLC scheme. It should be noticed that the TANOS NAND flash memory has much better mechanical stability than a FG type memory. This is due to the fact that the stack height of the TANOS NAND is much smaller than that of the FG type memory thanks to the absence of a floating gate.

In the TANOS NAND flash memory, interference among cells barely occurs due to the nature of the charge trap mechanism in the cells – the nitride traps act as point charges. This is the single biggest advantage of the charge trap NAND flash memory. We have also investigated retention characteristics of the memory. It has been found that the retention time in TANOS NAND flash memory depends heavily on both the top and bottom oxides. What is important in the former is categorized into three factors: quality, thickness, and nitride traps in the top oxide. The latter is divided into both quality and thickness of the bottom oxide. By optimizing the formation processes of both oxides, we obtained reasonable retention characteristics. Data retention of the TANOS NAND flash memory need to be pursued with improvements expected from further optimization of a TANOS stack, with a W control gate, TaN top electrode, top Al oxide, nitride film, and

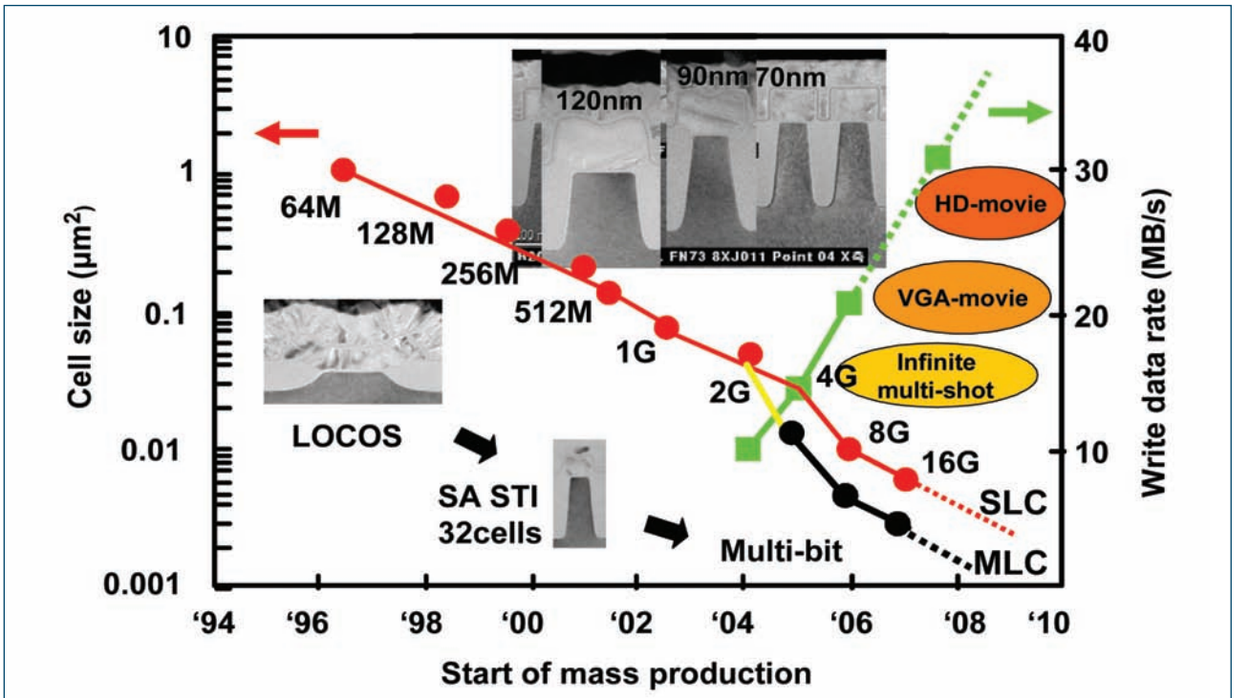

Figure 4. Evolution trend of NAND flash technology in terms of cell size and write data rate.

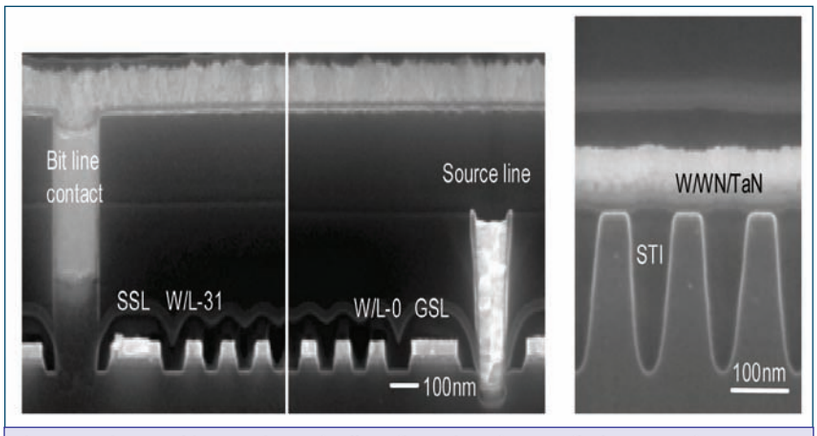

Figure 5. Micrographic SEM views of cell configuration of TANOS flash memory for MLC NAND: Left- hand cross-sectional SEM image shows the cell structure in bit-line direction and right-hand cross-sectional SEM image depicts the same structure in word line direction.

thin bottom oxide. It is strongly believed at this stage that the TANOS NAND flash memory could meet the needs of future NAND flash memory. This is due to the freedom from interference, an adequate coupling ratio, reasonable numbers of storage electrons, dimension scalability and mechanical stability, all of which place the TANOS NAND flash memory in a strong position. Although all the data here are obtained from 63nm 8Gbit NAND flash memory, we believe that its cell technology could simply be extended to the 40–35nm node.

In order to further scale down a TANOS NAND flash memory, we may need another cell technology. FinFET [13] could be a very promising candidate mainly because it can effectively increase storage electrons. In pursuit of which, we have also successfully developed FinFET TANOS where we can obtain a uniform distribution of V_{th} because of outstanding transistor performance of FinFETs. However, at this moment, we do not have enough information to decide whether we can scale down FinFET TANOS NAND much further. But what we know is that FinFET TANOS technology can be extended down to the 30nm node. We can even go down to 25nm as long as high-k top oxide can be improved. Beyond the 25nm node, we believe that the only possible way to increase the density of NAND flash memory is to stack the cells vertically because of the many limitations encountered: scaling limitations of FinFET TANOS, difficulties in manufacturing, and non-availability of lithography. The lithography concern is not discussed here, but it might be the first stumbling block on the road of NAND flash memory's evolution unless we prepare properly.

Conclusions

Technical challenges for sub-50nm node of DRAM and NAND flash memory have been reviewed in detail. It has also been suggested how to break scaling barriers with new device structures, new process technologies and new materials. DRAM and NAND flash memory are expected to maintain their dominance in the portable mass-storage market beyond the sub-50nm node. However, below 30nm, we seem to be a long way from solving the ballooning difficulties, particularly those arising from processing wafers caused by both ever-smaller feature sizes and the demand for new materials. Novel solutions to tackle these issues have been suggested and have also started to show promise. These will allow both DRAM and NAND flash memory to extend their scalability further and provide cost-effective solutions for massive data storage.

REFERENCES

[1] Kinam Kim et al, "Future outlook of NAND Flash technology for 40nm node and beyond", (invited) NVSMW Dig. Tech. Papers, pp.9-11, 2006.

[2] Kinam Kim, "Technology for sub-50nm DRAM and NAND Flash manufacturing", (invited), IEDM Dig. Tech. Papers, pp.333-336, 2006.

[3] Kinam Kim et al, "Memory technologies in Nano-era: Challenges and Opportunities", (invited) ISSCC Dig. Tech. Papers, pp. 576-577, 2005.

[4] Kinam Kim et al, "DRAM technology perspective for Gigabit era", (invited) IEEE Trans. Electron Dev. 45, pp.598-608 ,1998.

[5] Y. Park and Kinam Kim., "COB stack cell technology beyond 100nm technology node", (invited) Technical Digest IEDM, pp.391, 2001.

[6] D. H. Kim et al, "A Mechanically Enhanced Storage node for virtually unlimited Height (MESH) capacitor aiming at sub 70nm DRAM", Technical Digest IEDM, pp.69-72, 2004.

[7] J.Y. Kim et al, "The Breakthrough in data retention time of DRAM using Recess-Channel-Array Transistor (RCAT) for 88nm feature size and beyond", VLSI Technical Digest, pp.11-12, 2003.

[8] J.Y. Kim et al, "S-RCAT (Sphere-shaped-Recess-Channel-Array-Transistor) Technology for 70nm DRAM feature size and beyond", VLSI Technical Digest, pp.34-35, 2005.

[9] D. G. Park et al, "3-dimensional-CMOS transistor to overcome scaling limits", Proceedings of ICSICT, pp.35-38, 2004.

[10] Y. S. Kim et al, "Local-Damascene-FinFET DRAM integration with p+ doped poly-silicon gate technology for sub-60nm device generation", will be published in Technical Digest IEDM, 2005.

[11] Kinam. Kim et al, "The future prospect of non-volatile memory" (invited) Technical Digest VLSI-TSA, pp.88-94, 2005.

[12] Y. Shin et al, "A novel NAND-type MONOS memory using 63nm process technology for multi-gigabit flash EEPROMs", Technical Digest IEDM, pp. 327-330, 2005.

[13] P. Xuan et al, "FinFET SONOS flash memory for embedded applications", Technical Digest IEDM, pp 609-612, 2003.

ABOUT THE AUTHOR

Kinam Kim received Ph.D. degree in electrical engineering from University of California, in 1994 and the B.Sc. degree in electronic engineering in 1981 from Seoul National University. In 1983 he received the master degree in electrical engineering from KAIST(Korea Advanced Institute of Science and Technology). In 1983 he joined Samsung Electronics Co., Ltd., where he has been involved in the development of various memory technologies (64Kb to 4Giga-bit densities for DRAM, 2Gb to 16Gb and beyond for NAND Flash, and etc.). Currently he is a senior vice president responsible for the research and development of next generation memory technologies for DRAM, Non-volatile memory, SRAM, and emerging new memories such as FRAM, PRAM, and MRAM.

ENQUIRIES

Kinam Kim – Semiconductor R&D Div.
Samsung Electronics Co. Ltd.
San #24, Nongseo-Ri, Kiheung-Eup
Yongin, Kyunggi-Do 449-900, Korea
Tel: 82-2-760-6170
Fax: 82-31-209-3274
E-mail: kn_kim@samsung.com