3-D Flash as NAND Replacement

Andrew J. Walker

Schiltron Corporation

1638 Cornell Drive
Mountain View, CA 94040 USA
(email: andy@schiltron.com ; tel: +1 408 425 4150)

Abstract

With the slowing down in NAND Flash scaling, three dimensional stacking of Flash cells is being touted as a replacement for the classic NAND approach. This article looks at the various 3-D approaches reported in the literature and discusses the challenges that each face in their path to product. The growing popularity of 3-D approaches is due to their apparent cost advantage over 2-D NAND. A detailed look at the cost leverage is given. Finally, a prediction is made about the time to the first 3-D Flash product.

I. Introduction

Classic NAND Flash in no way can be considered to be a "universal" memory. Indeed, its very structure of a linear string of memory MOSFETs limits its use to areas where relatively slow access times are good enough. The first academic paper on NAND Flash was given by Toshiba in 1988 at the International Electron Device Meeting[1] and consisted of 8 floating gate MOSFETs in a string between two access devices using 1um design rules. We've come along way since then. The most advanced products at the time of writing this article use devices with a half-pitch of around 30nm. In the 20 years since its inception, many innovations have kept the linear scaling going at an aggressive pace.

[1] M. Momodomi et al., "New Device Technologies for 5V-Only 4Mb EEPROM with NAND Structure Cell", International Electron Devices Meeting Tech. Dig. pp. 412-415, Dec. 1988.

The reason why NAND Flash has become so popular is the low cost per bit combined with the spectacular rise in personal storage. The cell size of NAND is close to the theoretical limit of $4F^2$ where F is the minimum feature size in the process. The linear string architecture allows the pitch to be defined by the contactless gate pitch in one direction and the contactless field oxide pitch in the other direction. The bitline contact is then shared between two strings and the source can be shared between many strings.

As I write this article, the Flash Memory Summit has just ended in Santa Clara with some interesting comments from Eli Harari, founder and CEO of SanDisk Corporation. Besides the admission that the industry will see "a slowing down in NAND scaling", 3-D Flash was mentioned by him as a possible viable alternative to NAND. Specifically, the 3-D technology being worked on together by SanDisk and Toshiba, could replace NAND "assuming there are material breakthroughs". In addition, "That transition (to 3-D) is years ahead of us".

So what are the reasons why 3-D Flash is being talked about as a possible NAND replacement? And what are the various candidates that are vying for the NAND Flash crown? Section II discusses the reasons why NAND Flash scaling could be slowing down and looks at the various 3-D contenders for its crown. Section III looks at the cost of 3-D Flash compared to 2-D. Section IV makes some predictions about the advance to a 3-D product and the gradual demise of the supremacy of standard 2-D NAND Flash.

II. NAND Scaling and 3-D Contenders

The best article I know that looks at the problems of NAND scaling below 30nm is by Kirk Prall at Micron[2]. In it, he shows the floating gate structure as being the main reason for problems. In particular, the capacitive coupling interference from surrounding floating gates causes a pattern sensitive voltage on any given floating gate that degrades the ability to store multiple bits on each cell.

[2] K. Prall, "Scaling Non-Volatile Memory Below 30nm", Technical Digest, pp. 5-10, NVSMW, 2007.

The economics of scaling are also to blame in that continued shrinkage may call for the use of extreme ultraviolet (EUV) lithography and all the manufacturing infrastructural changes that this involves.

Here we shall deal with the various approaches to 3-D Flash that have been proposed.

A. Resistance Change Approaches

There are several approaches that can be classified here. For the smallest cell in a 3-D architecture, most if not all would have to follow the approach shown in Fig.1[3] which has been reproduced from page 28 of this reference. Notice the steering element which is most likely to be a diode since this is a two-terminal device and can be made in a small area.

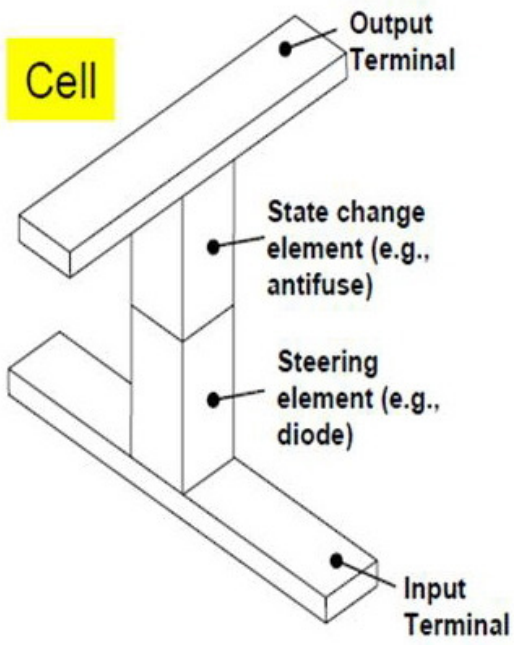

FIGURE 1

A first example of the resistance change approach is Phase Change Memory (PCM) which involves the use of chalcogenide material usually in the form of Ge_2-Sb_2-Te_5 (GST). Programmable resistive states depend on the reversible change between amorphous and (poly)crystalline phases. The

[3] M.A. Vyvoda, "3-Dimensional Monolithic Nonvolatile Memories and the Future of Solid-State Data Storage", http://microlab.berkeley.edu/text/seminars/MonolithicMem.pdf

reset current to change the state from low resistance to high resistance is usually on the order of several hundred microamps.

Another approach within this family involves the use of some form of perovskite material. As for the PCM approach, it seems that at least 100μA may be needed to switch from the low to high resistance states.

The third approach involves the use of simple metal oxides as the switching material. As above, it appears that at least 100μA is needed for reset although promising data of sub-100μA reset current have appeared recently.

Other interesting approaches within the realm of switchable resistance include solid-state electrolytes where programming currents as low as 1μA have been reported.

It appears from the literature on switchable materials that several challenges remain before a high density 3-D approach could be envisaged, namely the need to integrate high current drive steering devices at low enough process thermal budget and the need to lower the reset currents to a level that allows massive programming parallelism that already exists in NAND Flash. A further consideration perhaps is the need to lower reset currents to a level that allows the integration of selection transistors into the 3-D memory stack. Current thin-film transistor technology based on poly- or nanocrystalline silicon would be unable to cope with such demands that the above switchable materials would at present require.

B. Floating gate and Charge Trap Flash NAND in Horizontal Plane

The second main class of approaches for 3-D Flash involves some form of series string of transistors. The most obvious is to stack the existing NAND floating gate structure. This then has the same lateral scalability constraints as normal NAND floating gate and would be expected to run up against the same difficulties at around 30nm half pitch. All other 3-D series string approaches published so far involve the use of a Charge Trap Flash (CTF) approach that uses some form of silicon nitride to replace the floating gate.

To appreciate the challenges of CTF NAND approaches, it is useful to consider Fig.2.

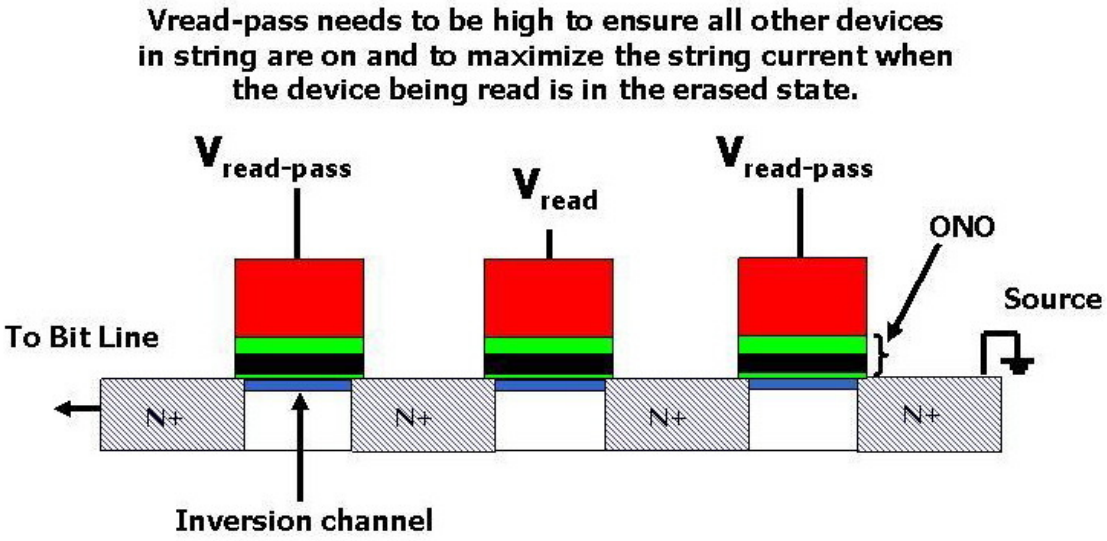

FIGURE 2

The first challenge is to make sure all pass cells are conducting when one cell needs to be read. Therefore the read-pass voltage needs to be higher than the highest possible programmed threshold voltage plus margin. If this voltage margin is small, then any programmed pass cells have high impedance resulting in a small string current when the cell being read is in the erased (low threshold voltage) state. The second challenge is to ensure a large enough "worst case" string current when all pass cells are in the highest programmed state and the cell being read is erased. This again requires high read-pass voltages. A third challenge is to make sure that programming never results in a threshold voltage greater than the read-pass voltage. Otherwise, a pass cell will be off when it should connect the cell being read or programmed to the end of the string. Similarly, any upturn in threshold voltage as seen in endurance cycling either causes a serious reduction in string current if the read-pass voltage is kept constant or results in higher read-pass voltages and a resultant exponentially larger read-pass disturb. A common method used in NAND Flash to increase area efficiency is to increase

the number of cells in a string. In this way, the area overhead of the bitline contacts and the source is amortized over more cells. If the read-pass voltage cannot be increased by much, then this approach reduces the worst case string current.

The tendency for thin tunnel oxide Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) devices to soft-program with such pass voltages has been partially tackled by either thickening up the tunnel oxide from its initial ~2.5nm to around 5nm. This however results in high program and erase voltages, unavoidable exponential threshold voltage effects from pass voltages and limited endurance for the resulting TANOS architecture. The situation with thin-film transistor (TFT) series strings is especially painful due to the low TFT carrier mobility.

Samsung has been the main proponent of 3-D versions of CTF NAND in the form of TANOS. The complete eradication of pass disturbs at the expense of higher program and erase voltages has however not been achieved even with the relatively thick TANOS gate dielectric stack.

A method to remove pass disturbs completely and still use thin ONO gate dielectric stacks has been pioneered by my own company, Schiltron Corporation. Figure 3 shows the approach.

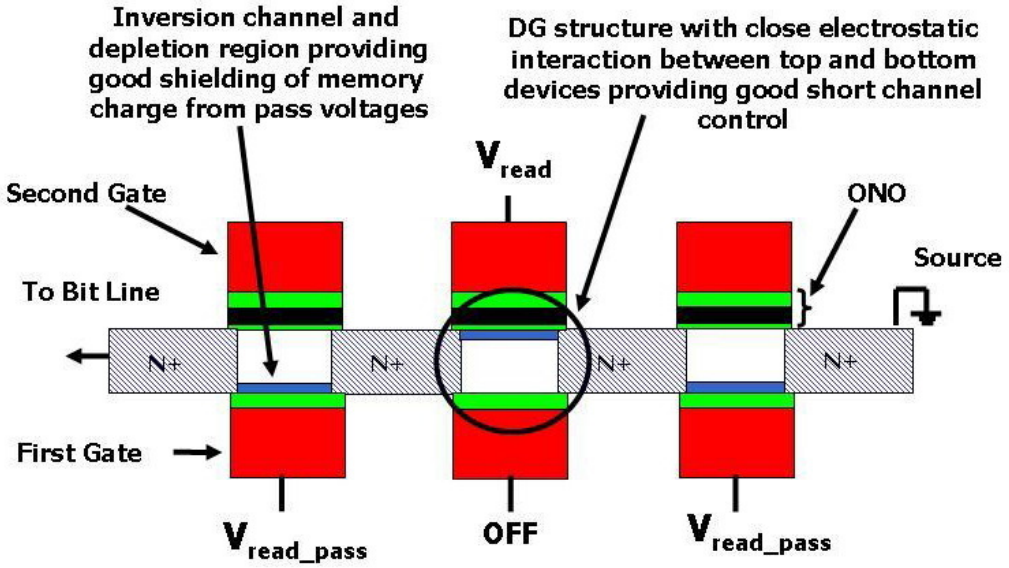

FIGURE 3

Here, access to a memory device in the series string is achieved by forming an inversion channel in the lower, non-memory devices. The inversion channel and its associated depletion region provide the charge trapped in the memory dielectric a high level of electrical screening from the pass voltages applied to these bottom access devices. In addition, the dual-gate architecture is a known good laterally scalable approach by having close electrostatic interaction between top and bottom devices that cuts out short channel effects. In a paper[4] at last year's International Electron Device Meeting, the smallest silicon-based TFTs (sub-50nm for both length and width) were reported in series strings of up to 64 cells with worst case string currents of 100's of nA's. Additional benefits include the use of standard materials, standard program and erase mechanisms and a process temperature budget low enough to stack several layers on top of each other. It is interesting to contrast this result with the approach from others that need material breakthroughs to work.

C. Charge Trap Flash NAND in Vertical Plane

An important approach within the series string of CTF devices is the vertical NAND. This method attempts to reduce the cost per bit even further than lateral 3-D NAND approaches. Toshiba's Bit-Cost Scalable (BiCS) approach[5] consists of etching holes through dielectrically separated conducting plates. These holes are then filled with the memory dielectric stack and the silicon channel. The result is a vertical CTF NAND structure where each series device has a gate all around the channel. Several important challenges remain however. For example, the same pass-disturb analysis presented above applies. In addition, the lateral scalability is linked to the need to fill the hole with both the CTF dielectric stack and the silicon channel which limits the lateral half pitch to probably greater than ~55 nm. Finally, the method of increasing memory density by lengthening each vertical NAND string not

[4] A.J. Walker, "Sub-50nm DG-TFT-SONOS – The Ideal Flash Memory for Monolithic 3-D Integration", International Electron Devices Meeting Tech. Dig. pp. 847-850, Dec. 2008.Technical presentation to be found at: [Online]: http://www.schiltron.com/PDF_files/Schiltron_IEDM_2008_full.pdf

[5] H. Tanaka at al., "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory", Symp. VLSI Technology Tech. Dig. pp. 14-15, 2007.

only increases pass disturbs but reduces the worst case string current. Indeed, for every doubling in density, the worst case string current halves. Since the channel of these devices is polysilicon, the worst case string current may quickly tend towards an unreadably low value as density increases. Also, the fact that a rather thin stack of CTF dielectric needs to be used to make a small cell not only could seriously affects the pass disturbs but also the retention. Despite these points, the BiCS approach is a valiant attempt to lower cost.

III. The Cost Advantage of 3-D Flash

First of all, let me define what I mean by 3-D Flash. This is Flash cells stacked one on top of the other over a substrate that may contain the driving circuitry. The method of stacking is through material deposition with the Flash cells fabricated in the deposited material. This approach is call "monolithic" and would be the cheapest way of making a 3-D Flash chip when compared to chip stacking for example. There are excellent reviews that have appeared recently[6].

The cost advantage of stacking Flash cells is not intuitively obvious since the process expense is increasing to get to a smaller total chip size for a certain memory capacity. The added process complexity may also impact yield which would lower the cost advantage too. The key is to work out the cost advantage since we cannot simply say that it is always advantageous to go 3-D due to the smaller chip.

It was for this reason that I developed a comprehensive model on monolithic 3-D cost that was published recently[7]. Here I shall summarize the approach and conclusions.

---

[6] C. Petti, S.B. Herner and A.J. Walker, "Monolithic 3D Integrated Circuits", in *Wafer Level 3-D ICs Process Technology*, C.S. Tan, R.J. Gutmann and L.R. Reif (Eds), Springer 2008.

[7] A.J. Walker, "A Manufacturing Cost Model for 3-D Monolithic Memory Integrated Circuits", IEEE Trans. Semicond. Manufacturing, vol. 22, pp. 268-275, May 2009.

In monolithic 3-D, there is one factor that reduces cost, namely more chips per wafer for a fixed memory capacity. Many advocates of 3-D concentrate on this one point. However, there are two factors that increase cost, namely increasing process complexity in the form of extra process steps) and yield. How do these factors affect the total cost of a working 3-D Flash chip?

My cost model paper combines all these factors into one cost equation and looks at the cost of a 3-D Flash chip compared to that of a 2-D Flash chip of the same capacity. Figure 4 shows one example from the paper where the 2-D chip is made using NAND at a half pitch of 25nm and the 3-D version uses 32nm half pitch. The factor Z is the multiplier in cost of equipment going from one node to the next more advanced node. Here it has been varied and can be seen as the number that defines the increasing cost of making 2-D NAND.

Several interesting features appear from this analysis.

First, there is an optimum number of 3-D memory cell layers that gives the lowest relative cost. This can be understood as follows. Consider that in a 2-D NAND Flash chip, there is control circuitry that would remain in the substrate even if the memory cells were placed in a 3-D stack. The amount of the 2-D chip area that is taken up by memory cells is usually called the array efficiency. The stuff left over is usually around one quarter of the total 2-D chip area. Therefore, this is the smallest 3-D chip area provided that this circuitry can be placed underneath the memory stack. To get to this minimum area, about four memory layers would be needed since doubling the number of layers would halve the array area each time. Four layers would reach the area almost equal to the area left over by the control circuitry.

Second, 3-D stacking does not reach the 10X reduction in cost as some have claimed. On the contrary, it can be seen as the logical successor to lateral scaling once lateral scaling becomes too difficult to do. The gains in cost reduction through 3-D stacking then are about the same as for lateral scaling.

Third, it is always better to lift all memory cells out of the substrate and put them in the 3-D stack. Otherwise, a minimum in cost will actually never be reached by adding layers and the cost will always be higher than the case where there are no memory cells left in the substrate.

Fourth, techniques to integrate control circuitry into the 3-D stack would be advantageous for cost in that even smaller chip sizes could be realized.

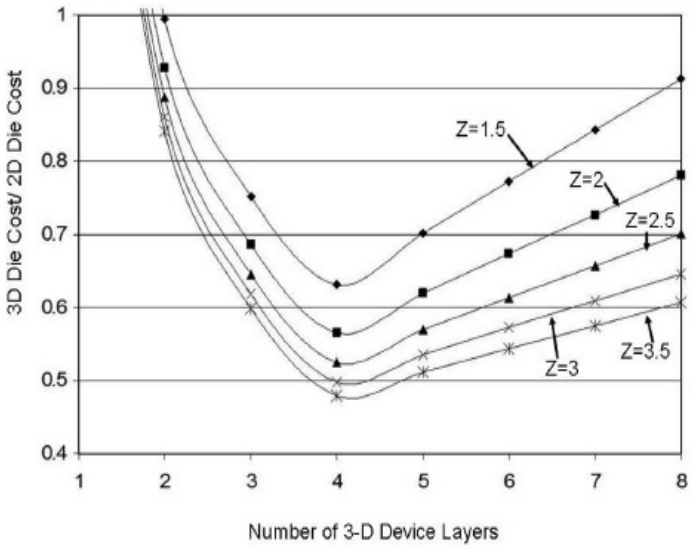

Fig. 5. Ratio of 3-D to 2-D die costs as a function of the number of 3-D device layers added for a fixed capacity memory (64 Gbit in this case) with Z as a parameter. 2-D uses advanced 25-nm half-pitch while 3-D uses more mature 32-nm half-pitch. $D_0 = 0.1$ cm^2, $\mathrm{MLC} = 2$ and $AE_0 = 75\%$. The same base wafer and adder costs were used as in Fig. 4.

FIGURE 4 (taken from cost model reference)

IV. Predictions

It has been said that one should never make predictions, especially about the future. Nevertheless, here goes (and of course I give no guarantees about the future).

First, I predict that there will be a strategic withdrawal of 2-D NAND from certain market segments and an inability by 2-D NAND to address certain market segments. The reason for this is that certain tradeoffs are being made and will be made to allow continued cost per bit reductions. We can already sense a flavor here in the reliability tradeoffs (endurance and retention after cycling)

being made to reach multiple bits per cell at the advanced nodes. Obviously, many markets will remain immune to this and will continue to be served by 2-D NAND, at least for a time.

Second, I predict that 3-D Flash will gradually take over some high density Flash market segments. The progress may be relatively slow but 3-D Flash will have to prove itself in the marketplace. An analogy that springs to mind is the "Japanese motorcycle approach". I remember quite vividly the small 50cc motorbikes from the Japanese manufacturers that arrived in the UK at the time of the UK dominance of the motorbike industry. The experts said that these "sewing machines" would never displace the Triumphs of this world. Well, we saw what happened. Start small and gradually take over.

Third, the time to the first product that contains 3-D Flash is three years from now. The questions remain: which 3-D technology and which market segment? Here I shall dispense perhaps with my objectivity and say the following. The 3-D Flash technology that will take over will be the one that remains within the confines of the semiconductor industry maxim, being "change as little as possible". In this regard, the 3-D technology that uses existing materials and tools and keeps as close as possible to existing mechanisms for program and erase, will eventually dominate.