The Inevitability of Monolithic 3-D Flash

Andrew J. Walker

At the time of writing (May 2009), the most aggressive NAND half-pitch is at 30nm, the total NAND market is between 10 and 20 B$ and there are no multi-programmable Flash products made from monolithic 3-D technology. Despite the current economic downturn, all's well with the world then? Well, no actually: most experts agree that NAND Flash scaling (i.e. shrinking) is running out of steam. Some even say that we are probably within a generation (NAND, not human) of no more shrinks. If we are now at 30nm, that would put the final generation half-pitch at around 21nm. What then?

As most who have been in the semiconductor industry can attest, engineers will try everything in their power to keep shrinking the existing technology. We can imagine a 20.5nm version followed by perhaps a 19.5nm version. The question is, at what cost? If there were no alternative to scaling to achieve ever-higher chip capacities, we can imagine that manufacturing would be left to perhaps one big player (perhaps a Flash Foundry) and that product companies would differentiate themselves in other ways.

Fortunately (and you know where I'm going), there is an alternative to scaling to achieve higher single chip memory capacities, namely monolithic 3-D stacking. I can hear a collective "well he would say that wouldn't he" but now we can see using simple mathematics the **inevitability of monolithic 3-D Flash**.

First of all, what is "monolithic"? Monolithic defines integrated circuits where circuits are made on a substrate and at least one layer above this substrate in a single linear process flow with no material bonding used.

Removing any mystique, we can see that monolithic 3-D stacking is just an alternative to scaling in that we complicate a semiconductor process to increase the number of cells in a chip while reducing the cost per cell. It has been viewed as an alternative for quite some time but since scaling has been the easier path it has remained in the shadows. Now, with the specter of NAND Flash scaling coming to an end, it can finally show its face.

The **inevitability of monolithic 3-D Flash** can be put in a mathematical form and is derived from my (peer-reviewed) article entitled "A Manufacturing Cost Model for 3-D Monolithic Memory Integrated Circuits"[1]. If you slog through that paper, you can come up with the following equation for the cost of a good 3D die divided by the cost of a good 2D die both with the same total capacity:

$$\frac{C_{die}^{3D}}{C_{die}^{2D}} = \left(\frac{Y_{2D}}{Y_{3D}}\right)\left(\frac{1}{N_L}\right)\left(\frac{F_{3D}}{F_{2D}}\right)^2\left(\frac{MLC_{2D}}{MLC_{3D}}\right)\left(\frac{C_o + C_{crit\_mask}^{3D}.N_{crit\_mask}^{3D}}{C_o + Z^n.C_{crit\_mask}^{3D}.N_{crit\_mask}^{2D}}\right)$$

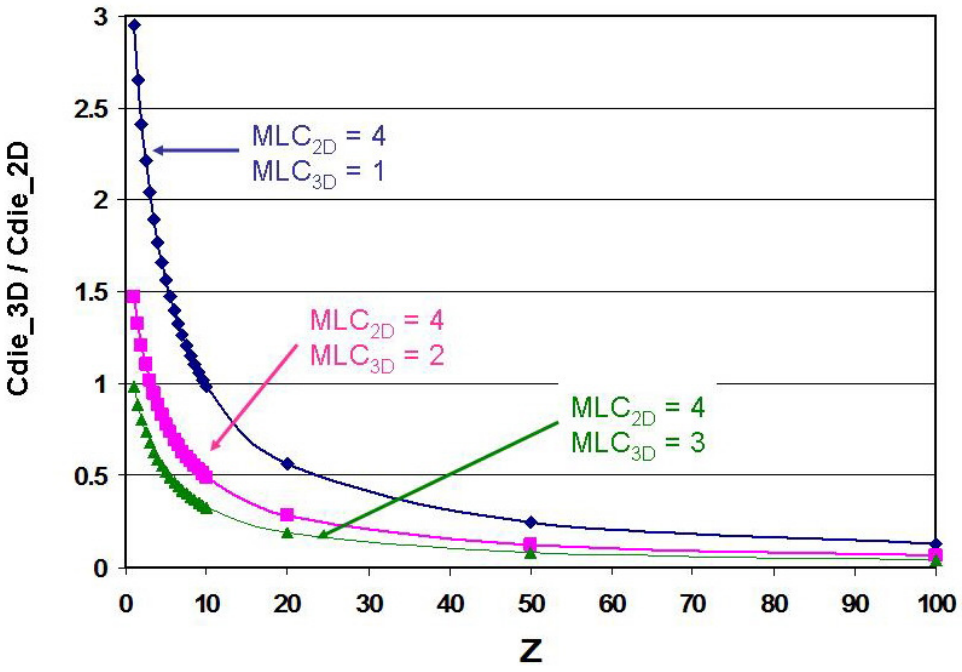
[1] IEEE Transactions on Semiconductor Manufacturing, vol.22, no.2, pp.268-275, May 2009.

where the Y's are the total yields, N_L is the number of memory device layers in 3-D, the F's are the minimum half-pitches for each approach, the MLC's are the number of bits stored in each physical cell (1,2,3 and so on), C_o is the base wafer cost excluding the difficult memory processing, C_{crit_mask} is the cost of each critical memory masking layer along with its associated processing, N_{crit_mask} is the total number of such critical memory masking layers, Z is the rate of increase in wafer cost between generations and n is the number of generations between the 3-D version and the 2-D version. For more details, see the above-mentioned paper.

Z is the key. It can be seen to be the repository of all costs going to the next generation. It can be used to measure the difficulty of making the 2-D generation at a more advanced node compared to the 3-D version at an older node. As the difficulties mount, Z multiplies.

Let's take an example where a Flash chip of a certain capacity is manufactured in 3-D using 21nm minimum half-pitch while the same capacity memory is manufactured at the next full node of 14.7nm in 2-D. As was shown in the above cost paper, we can assume that the total final yields are very similar. Let's take the base wafer cost to be $2800, each 3-D memory layer costs $200, the number of device layers in 3-D is 4, the number of critical memory layers per device layer in 3-D is 3, the number of critical memory layers in 2-D is 4.

Figure 1 below shows the cost of a good 3-D die normalized to the cost of a good 2-D die as a function of the cost multiplier Z. Notice I have assumed that the 2-D die can be done using 4 bits per cell which is highly arguable at that node.



The simple fact is that as the 2-D costs multiply, monolithic 3-D will take over no matter how many bits per cell can be practically stored in a 2-D cell. The question remains: when? Clearly the incumbent manufacturers will do the sensible thing and try to prolong the life of their 20 year old 2-D NAND technology as far as it can be scaled. But eventually the frog will die.

Now is the time to invest in real monolithic 3-D Flash technology. The next question is: which one? In Schiltron's opinion, any technology that uses existing materials and tools, and existing program and erase mechanisms that preserve hard-won NAND program bandwidths and powers will be the technology that allows further reductions in cost per bit.

And of course this technology is Schiltron's !