PCM progress report no. 7: A look at Samsung's 8-Gb array

Here's a discussion on the features of Samsung's 8-Gb array.

By Ron Neale

After Samsung's presentation [1] of their 8-Gb PRAM at ISSCC2012 and while phase-change memory (PCM) watchers wait for the other shoe to drop in the form of a possibility of an associated PCM product announcement, there are a number of features of the array that are worth discussing.

The architecture of Samsung's 8-Gb array, from the top down to the individual PCM cells runs as follows: eight partitions of 1 Gb, each of which is divided into 128 sub-arrays (tiles), organized as a matrix of 4096 word lines (WLs) by 2048 bit lines (BLs). WL strapping contacts are situated at 64-cell intervals, which the authors assert provides a 19% area gain in over earlier 1 Gb work [2] in addition to the advantages introduced by moving from 58-nm to 20-nm technology. The result is a 9.43 x 6.30 mm$^2$ chip. The device operates at 1.8 V and uses a low power, double-data-rate non-volatile memory (LPDDR2-NVM) interface. No quantitative data on chip power dissipation was provided.

Write bandwidth

Although the headline write bandwidth claimed for the 8-Gb array was 40 MB/s, under certain conditions this could be increased. The array architecture employed parallelism to increase the bandwidth over the Samsung's earlier 1-Gb demonstration array [2]. Parallel write operation, upped to 128bit from the 32bit of the earlier 1-Gb array, provides the increased bandwidth. There was an indication that when the device was optimally implemented, this could be increased to 133 MB/s. To understand in part how that is possible, it is necessary to explore the form of the write/erase (w/e) pulse. Pulse shaping of the leading and trailing edges and multipulse trains has long been a feature of PCM development, usually in order to optimize a particular device characteristic such as w/e lifetime, on/off resistance ratios or values, and elevated temperature data retention characteristics.

Figure 1: Features of the programming pulses illustrate how pre-emphasis reduces rise time and contributes to reducing total pulse width.
For their 8-Gb array, the Samsung team brings something new to the PCM table in the form of the write pulse presented to the PCM cell. By design, the pulse that is delivered to the cell is formed in part by the parasitic resistance and capacitance of the array. The write pulse uses what is described as “pre-emphasis” technique (figure 1). In this approach, the write generator produces a pulse output in which the initial period of the pulse is of greater amplitude than is actually required to reset or set the device. This pre-emphasis pulse is then integrated by the parasitic capacitance and resistance of the word and bit lines to produce, at the PCM cell, a leading edge with a rise time to the required current that is significantly less than would be the case if pre-emphasis was not used. The net effect is to reduce the overall write time.

Other effects that increase the write time, shown as $T_A$ and $T_B$ in figure 1, are there to allow the high-voltage programming charge pump to recover. The reason they are not lumped together is because while charge pump recovery is the main purpose, the authors state that it is not the only role. I would suggest that other roles might be to allow for the thermal recovery of the PCM cell and to allow discharge of all of the parasitic capacitance. The latter is especially important when the passive parasitic parts of the array are used for pulse shaping.

The earlier mentioned ability to increase the bandwidth from 40 MB/s to 133 MB/s is facilitated by providing the required high voltage from an external source, thereby removing any increase in program time related to charge pump recovery. The high voltage also extracts a chip area penalty.

To provide a constant programming current to all PCM cells, irrespective of the position and resistance along the array interconnect, a cascode write-pulse generator design, with an output impedance of several mega-ohms, is used. This means the current at any cell is less influenced by variable parasitic series resistance. As a result, the programmed cell resistance is more constant and adds to the width of the sense amplifier read window.

The reductions in write pulse-widths and the reduced write current allow the opportunity for the parallelism that has provided the bandwidth. The authors indicate that the write current is on the order 80 to 100 μA and show the form of on/off (set/reset) resistance as function of current for reset/set pulses of duration 100 ns and 150 ns, respectively. The device uses pulses of the same maximum current amplitude for both set and reset, with, for any pair, the longer set pulse having a sloping trailing edge.

**Maximizing the sense window**

The way in which BL and WL resistance are effectively reduced to provide the widest possible window for the sense amplifier also relies on parallelism. The technique is novel with respect to Samsung’s previously reported PCM array work. First, the bit lines are driven from both ends, reducing the bit line resistance to 25% of the more normal single-end method. Although for the read process, two word lines are used rather than one—in other words, the same number of cells are read but with half on one word line and half on the other—this can only be achieved in an orthogonal matrix by accessing to two tiles. The benefit is that only half the combined cell read currents flows in each word line.

This focus on the BL and WL resistance is essential because the read window for the minimum reset current is only $2 \times 10^{(n)}$ to $4 \times 10^{(n+1)}$ A.U. (figure 2).

*Figure 2: The volume of amorphous material reset results in the resistance increasing.*
The positive slope of the resistance of the reset state curve would suggest to this writer that the either the volume of material being returned to the amorphous state is increasing, or for a fixed volume, the volume fraction of crystallized material in the reset cell is decreasing; figure 2 illustrates the manner in which one of those two options might be explained. While it is any author's privilege to use arbitrary units for, in this case, resistance and current values, it does prevent much by way of analysis of what might be happening or the consequences. The reduced size of the reset volume, or the low volume fraction reset, might offer an explanation for the poor elevated temperature data retention characteristics and why this 8-Gb PCM is now more DRAM than NVRAM. The consequence of employing a "soft" reset mode is most likely the reason why this PCM array was in the DRAM section of the ISSCC2012 conference and not the non-volatile memory session. In reference 3, to which this work on the 8-Gb PRAM is closely linked by the authors of reference 1, the elevated-temperature data-retention time at 85°C was reduced from better than 10 years—the required timespan for any accepted NV memory—to 15 months, making the array more appropriate for DRAM applications. In a historical context, it is worth noting that for a similar PCM structure, the previous claim was a retention time of four years at 85°C [4]. Based on other published work by Samsung on the dash PCM structure, however, I have different a observation regarding the suitability of this new 8-Gb array for even RAM, or at least some RAM applications in relation to w/e lifetime. I touched on those misgivings in an earlier PCM Progress report [5] and they relate to the safety of an extrapolation of a lifetime of $10^{15}$ write/ erase cycles made in reference 4 for the Samsung PCM dash structure. My rule, and that of most mathematicians, is you can invoke a logarithmic relationship and safely extrapolate it if you have data points covering at least three decades. To support their claims for the possibilities of a PCM DRAM based on w/e lifetime, the authors of the "dash" paper [4] took data points over two decades and extrapolated them over seven decades. Even the real data points they showed are more accurately extrapolated by a curve. As a blind experiment, I reproduced as an exact copy the "dash" paper data and data points and without the curve or any other text data and asked three experienced engineering professionals to draw the best straight line and the best curve through the data points. None gave a curve that was anything like the one presented in the Samsung paper or even a similar straight line. In figure 3, I have added to the original an overlay, in orange, of what in my view is a more reasonable version of this curve. To make claims and draw any conclusions from the results as presented with respect to future application possibilities was unsafe. Although the PCM lifetime, from a more realistic curve, taken as the maximum of the orange curve, that would be good enough for most NOR applications, such a curve requires at least a hypothesis to account for it. I would offer the following: There is some minimum energy at which this particular structure will no longer operate as a PCM. Referring back to figure 2 and the red curve, this would be a current value lower than that needed to make the step to the higher resistance state, so the device will fail to reset. At that point, the w/e lifetime will have dropped to zero. I would therefore postulate that for this particular PCM device structure, there is a maximum in the w/e lifetime curve at about $10^{12}$ w/e cycles for which the operating conditions are optimized so as to minimize reset failure.

![Figure 3: Plot of write/erase cycle lifetime versus program energy for PCM devices shows realistic estimated performance for PCM.](image)

EE Times-Asia  |  eetasia.com  Copyright © 2012 eMedia Asia Ltd.  |  Page 3 of 5
When will PCM find applicability?
The simple PCM device insets in figure 3 are included, as in figure 2, to suggest the manner in which for a dash structure volume of material might change as the read/write energy is reduced for a given fixed structure. If correct, each of the changes in set/reset energy produces, in effect, a different device. To have validity, claims for w/e lifetime for any operating conditions must address all other relevant device parameters, e.g. elevated temperature data retention, on-to-off ratio, reliability, and so on for each of these different devices.

It is interesting the latest 8-Gb paper does not mention w/e lifetime. For a similar dash structure [2], w/e lifetime is only mentioned in relation to thermal cross talk and for simulation, citing \(10^7\) cycles as the safe number of w/e cycles. Even reference 4, the source of the suspect extrapolation, only reports experimental test results of \(10^{11}\) w/e cycles, again for a similar dash structure, below the maximum shown in figure 3.

In relation to other effects that might impact reliability, it is possible to extract a current density value for the dash devices, estimated to be in the range \(3 \times 10^7\) A/cm\(^2\) to \(5 \times 10^7\) A/cm\(^2\). This may have current density and reliability consequences with respect to the matrix/isolation components, especially in relation to the cell size claimed.

Clearly the 8-Gb PCM device described by the Samsung team is the result of some admirable work but it raises many questions. Is the 8-Gb device on the verge of product status? We still do not know. Figure 4 illustrates how the PCM over a period of more than 40 years has moved its claimed areas of applicability. From its early potential applications as a random access non-volatile memory, the term electrically alterable read mostly memory or (EARMM) in figure 4 as the original description of a PCM that acknowledged at the time its w/e lifetime limitations. This was followed by mostly promotional, or at best aspirational claims for its role as a universal memory, on to the present, where the Samsung 8-Gb PCM appears to have a role as a possible DRAM replacement, having lost the 10-year data retention at accepted temperature of operation required for a non-volatile memory. Even the claims for use as a DRAM might be troubled by an inability to provide the w/e lifetime required for that role.

Figure 4: Over a period of more than 40 years, the areas of applicability for PCM have changed.

When I started this series of “Myths of Scalability” articles and progress reports on PCM in EE Times [6], I listed a number of known fundamental problems that would work against the claims made for the ability of PCM to scale. Nearly three years later, those same problems still hinder the production of competitive PCM products. At the time, it appeared to me that over promotion would also act limit future progress.

More moderate voices now suggest that 2016, with caveats, might be the time when PCM will find applicability, a date that has more recently been moved to 2024, with the caveat of a “might,” in relation to the IBM square kilometer antenna (SKA) radio telescope project. According to IBM, over the next five years they will be looking at seven technologies as part of a roadmap for the SKA. One of those technologies is PCM. While the SKA will go operational in 2024, IBM sources state that they will need to be fairly certain by 2017 of the technologies needed for the SKA.

Micron recently issued a statement that it has developed a new PCM process that it plans to deploy sometime this year. Yet experience teaches us that plans and implementation are not always the same. Whatever the future holds, if Samsung, or others, do not announce the availability of a fully-qualified competitive 1-Gb or 8-Gb PCM product very soon that will serve to build the base required for any possible future application and predictions, then for PCM, I think the game will surely be lost.
5. http://www.eetasia.com/ART_8800666813_499486_TA_0bd2fd72.HTM

About the author
Ron Neale is the former editor-in-chief of Electronic Engineering. Also, he is the co-author of "Nonvolatile and reprogrammable, the read-mostly memory is here", by R.G.Neale, D.L.Nelson and Gordon E. Moore, Electronics, pp56-60, Sept. 28, 1970.