2.5 A 20nm 1.8V 8Gb PRAM with 40MB/s Program Bandwidth

Youngdon Choi, Ickhyun Song, Mu-Hui Park, Hoeju Chung, Sanghoan Chang, Beakhyoung Cho, Jinyoung Kim, Younghoon Oh, Duckmin Kwon, Jung Sunwoo, Junho Shin, Yoohwan Rho, Changsoo Lee, Min Gu Kang, Jaeyun Lee, Yongjin Kwon, Soehee Kim, Jaehwan Kim, Yong-Jun Lee, Qi Wang, Sooho Cha, Sujin Ahn, Hideki Horii, Jaewook Lee, Kisung Kim, Hansung Joo, Kwangjin Lee, Yeong-Taek Lee, Jeihwan Yoo, Gitae Jeong

Samsung Electronics, Hwasung, Korea

Phase-change random access memory (PRAM) is considered as one of the most promising candidates for future memories because of its good scalability and cost-effectiveness [1]. Besides implementations with standard interfaces like NOR flash or LPDDR2-NVM, application-oriented approaches using PRAM as main-memory or storage-class memory have been researched [2-3]. These studies suggest that noticeable merits can be achieved by using PRAM in improving power consumption, system cost, etc. However, relatively low chip density and insufficient write bandwidth of PRAMs are obstacles to better system performance. In this paper, we present an 8Gb PRAM with 40MB/s write bandwidth featuring 8Mb sub-array core architecture with 20nm diode-switched PRAM cells [4]. When an external high voltage is applied, the write bandwidth can be extended as high as 133MB/s.

Cost-effectiveness is one of the key prerequisites to be competitive in the market for main memory systems [2], and PRAM's possibility can be found in both fabrication process and chip size. Compared to the DRAM fabrication process, PRAM can be integrated in fewer process steps due to its capacitor-less cell formation. In addition, the cell size of diode-switched PRAM is $4F^2$, approximately 60% of that of DRAM, which is around $7F^2$. However, the demands for high voltage level and current in PRAM write operation result in significant circuit area for the wordline (WL) and bitline (BL) selection switches in cell sub-arrays. Hence, the number of cells per WL and BL in the cell sub-array should be maximized to diminish the area overhead of these WL and BL selection switches. Figure 2.5.1 shows the designed core architecture, where a stacked partition structure is adopted for better area efficiency. The 8Gb PRAM is configured as 8 partitions of 1Gb with each partition consisting of 128 tiles (sub-arrays). Each tile is built with 4096 WLs and 2048 BLs, where WL strapping contacts are located between every 64 cells. In this manner, the effective unit cell area is scaled to 10% of that in the previous work [1], achieving 19% area gain, excluding the technology scaling from 58nm to 20nm. In terms of chip size, it occupies only 70% of DRAM chip size at the same design rule. However, this architecture requires improved read and write schemes due to the increased parasitic resistances and capacitances.

Figure 2.5.2 shows the sensing scheme employed in this work. For better understanding of the sensing mechanism, a simplified version of the sensing scheme is illustrated in the right part of the figure. The clamp voltage (V_{CMP}) applied at the NMOS transistor (MN1) confines the voltage of SDL (V_{SDL}) to (V_{CMP}-V_{THN}), where V_{THN} is the threshold voltage of the MN1. Hence, the current through the cell (I_{CELL}) is calculated as: $I_{CELL}=(V_{CMP}-V_{THN}-V_{THD})/(R_{CELL}+R_{PAR})$, where V_{THD}, R_{CELL} and R_{PAR} are the threshold voltage of the cell diode, the cell resistance and the total parasitic resistance of the read path, respectively. As I_{CELL} is a monotonic function of the R_{CELL}, the cell state can be detected by whether I_{CELL} is higher than the reference current (I_{REF}) or not. The sensing margin (SM) can be defined as the difference between the minimum I_{CELL} at SET state and the maximum I_{CELL} at RESET state. Thus the SM is calculated and found to increase as R_{PAR} decreases. Therefore, R_{PAR} minimization for the given condition is focused as one of the main design targets.

Figure 2.5.3 shows the dual-LY and multi-WL schemes. Compared to the previous work using single local-Y (LY) switch to select the local BL (LBL), LY switches are placed at both sides of sub-array in the dual-LY scheme, which reduces the worst-case value of the LBL resistance (R_{LBL}) to 25%. The multi-WL scheme reduces the effective WL resistance (R_{WL}) of the read path by selecting fewer cells per WL, while keeping the number of simultaneously read bits the same. Moreover, data-pattern dependency of the SM is improved because a WL collects multiple I_{CELL}'s from the selected read paths, which are dependent on the R_{CELL}. The lower part of the figure demonstrates the dual-WL scheme, where the effective R_{WL} is reduced by half. In this manner, the negative R_{PAR} impact on the SM can be mitigated.

Figure 2.5.4 shows the measured write performance of integrated 20nm PRAM cell [4], where the pulse of 150ns duration is sufficient to crystallize the PRAM cell to "SET" state. In addition to the improved program time (t_{PGM}), the reset current (I_{RESET}), and accordingly, the program current (I_{PGM}) is also reduced around 100μA. To take advantage of the reduced I_{PGM}, 128b parallel write operation is performed, compared to the 32b parallel write operation of the previous work [1]. Consequently, the write performance of this work is improved to 40 MB/s, which enables about 6× write-throughput improvement. The dead time noted as t_A and t_B in the figure mainly comes from delay to supply the high voltage for LY selection. Thus the total write period (t_{WR}) is reduced when an external high voltage supply can be applied. In addition, when optionally implemented 256b parallel write operation is used, the write throughput can be improved to as high as 133MB/s.

Figure 2.5.5 shows the designed write driver (WDRV). It is an important feature of the WDRV that it supplies every PRAM cell with the same I_{PGM} irrespective of its location, because the amount of I_{PGM} determines the level of R_{CELL}. For this reason, the main concept of the WDRV is to make the output resistance (R_{OUT}) of the current source high enough to neglect the impact of the load resistances of the path to each cell. To comply with this requirement, we implement a cascode type current source, which increases the R_{OUT} as high as several MΩ and satisfies the condition mentioned above. It is our experience in the previous technology that, with the proper choice of the transistor width and length, I_{PGM} variations between WDRVs are less than 1% under process and voltage variations, which alleviates R_{CELL} variations and enhances the SM. As the parasitic RC value in the write path is increased due to the high chip density, the rise-time (t_{RISE}) of I_{CELL} is also increased, resulting in longer write period. To compensate the increased t_{RISE}, a pre-emphasis method is implemented, where the current magnitude (I_{PRE}) and the time duration (t_{PRE}) are controllable, for example, by MRS.

Figure 2.5.6 summarizes the chip performances and process technology. Figure 2.5.7 shows a micrograph of the fabricated chip in which 8Gb PRAM cells are integrated with 20nm CMOS technology. The device is functional at 1.8V with a low-power double-data-rate nonvolatile memory (LPDDR2-NVM) interface of 800Mb/s/pin [5]. The programming throughput is 40MB/s.

References:

[1] H. Chung, et al., "A 58nm 1.8V 1Gb PRAM with 6.4MB/s Program BW," *ISSCC Dig. Tech. Papers*, pp. 500-501, Feb. 2011.

[2] M. Qureshi, et al., "Scalable High Performance Main Memory System Using Phase-Change Memory Technology," *ISCA*, 2009.

[3] A. Akel, et al., "Onyx: A prototype Phase-Change Memory Storage Array," *Proc. of HotStorage '11*, June 2011.

[4] M.J. Kang, et al.,"PRAM cell technology and characterization in 20nm node size," *IEDM Dig. Tech. Papers,* 2011.

[5] "Low Power Double Data Rate 2 (LPDDR2)", *JEDEC STANDARD, JSED 209-2D,* Apr. 2010.

978-1-4673-0377-4/12/$31.00 ©2012 IEEE

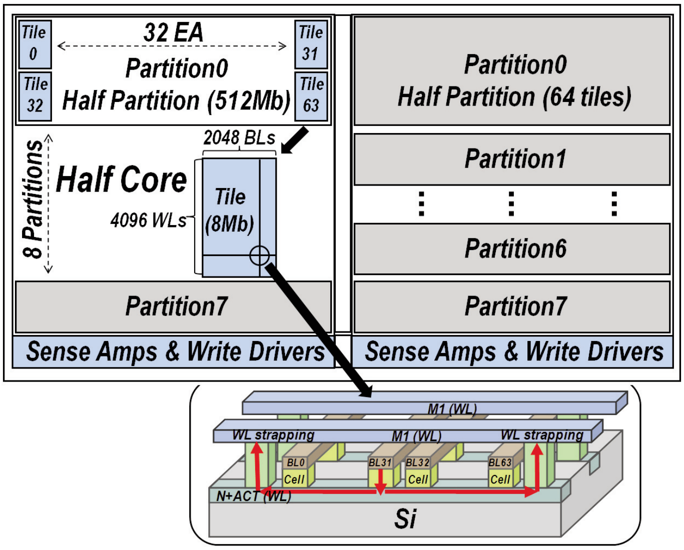

Figure 2.5.1: Core architecture of 20nm 8Gb PRAM.

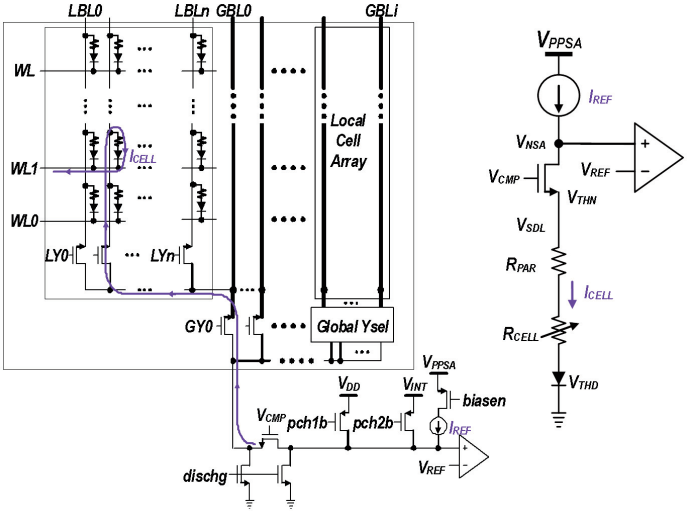

Figure 2.5.2: Read path and sensing scheme.

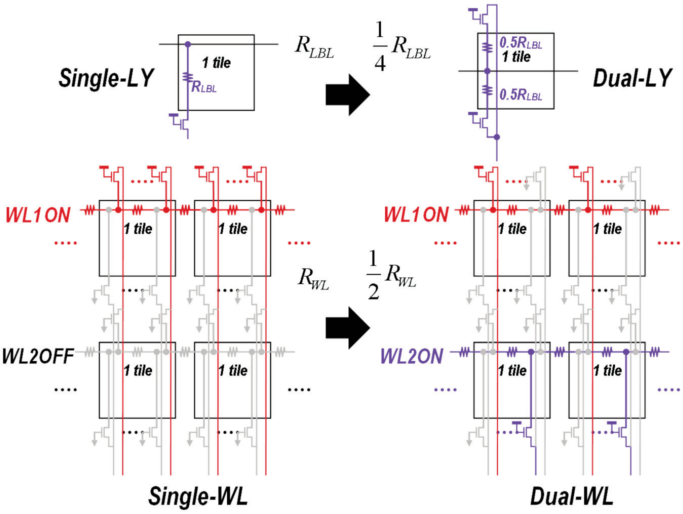

Figure 2.5.3: Dual-LY and multi-WL schemes for reduced effective parasitic resistance.

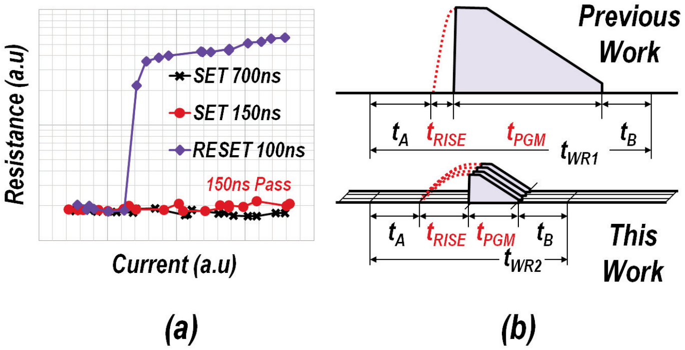

Figure 2.5.4: Write performance improvement: (a) measured cell write time, and (b) elements of write period.

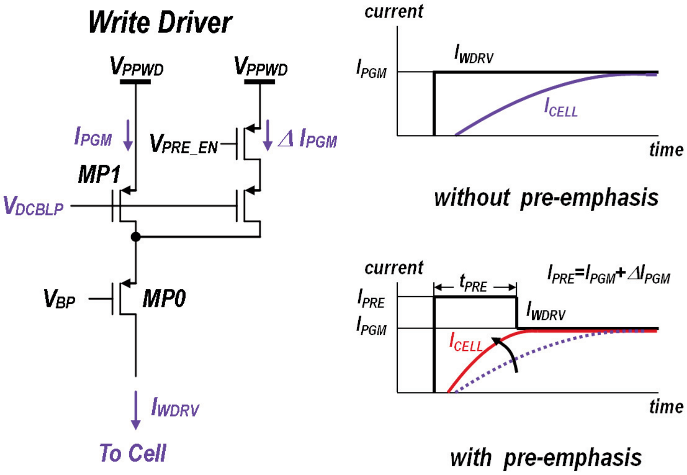

Figure 2.5.5: Write driver circuit.

Process Technology	20nm PRAM Process
Cell Size	$41X41nm^2$
Cell Switch	Diode-switch
Chip Size	$9.43X6.30mm^2$
Power Supply	VDD : 1.8V VDDQ, VDDCA : 1.2V
Temperature Range	-25 ~ 85 °C
Organization	1GbX8 (LPDDR2 interface)
Tile(CPWL/CPBL)	8Mb (2Kb/4Kb)
Tile Array(X/Y)	64/16
tSET	150ns
Parallel write	128b(default), 256b (option)
Write performance	40MB/s (internal power only) 133MB/s (external power+256b parallel write)
tRCD	120ns
I/O Bandwidth	800Mb/s/pin

Figure 2.5.6: Feature summary of the designed chip.

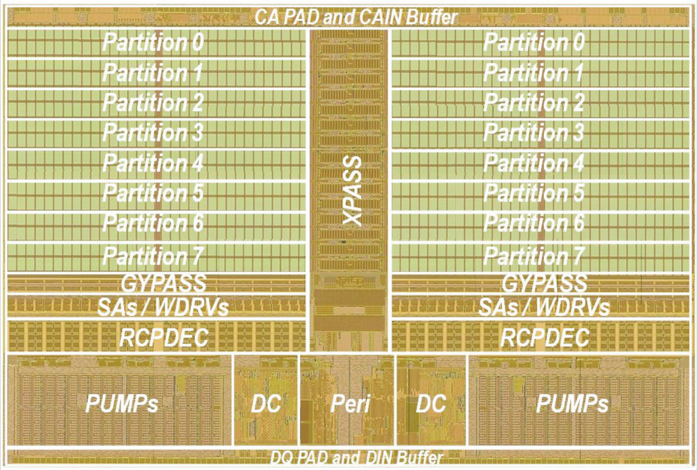

Figure 2.5.7: Chip micrograph.