18.2 A 1.2V 20nm 307GB/s HBM DRAM with At-Speed Wafer-Level I/O Test Scheme and Adaptive Refresh Considering Temperature Distribution

Kyomin Sohn, Won-Joo Yun, Reum Oh, Chi-Sung Oh, Seong-Young Seo, Min-Sang Park, Dong-Hak Shin, Won-Chang Jung, Sang-Hoon Shin, Je-Min Ryu, Hye-Seung Yu, Jae-Hun Jung, Kyung-Woo Nam, Seouk-Kyu Choi, Jae-Wook Lee, Uksong Kang, Young-Soo Sohn, Jung-Hwan Choi, Chi-Wook Kim, Seong-Jin Jang, Gyo-Young Jin

Samsung Electronics, Hwaseong, Korea

Demand for higher bandwidth DRAM continues to increase, especially in high-performance computing and graphics applications. However, conventional DRAM devices such as DDR4 DIMM and GDDR5 cannot satisfy these needs since they are bandwidth limited to less than 30GB/s. Also, if multiple GDDR DRAMs are used simultaneously for higher bandwidth, then high power consumption and routing congestion on PCBs become a big concern. In order to overcome these limitations, the high-bandwidth memory (HBM) DRAM was recently introduced[1]. HBM-DRAM uses TSV and interposer technologies enabling multiple chip stacks and wide I/Os between the processor and memory: providing high capacity, low power and high bandwidth. This paper proposes the 2nd generation HBM to double the bandwidth from 128GB/s to more than 256GB/s and support pseudo-channel mode and 8H stacks [2]. In the pseudo-channel mode, a legacy channel is divided into two pseudo channels and the two pseudo channels share the command-address pins. Thus, one HBM has 16 pseudo channels instead of 8 legacy channels. To support various stack configurations including 8H stacks, a new architecture is adopted for flexible density ranging from 16Gb to 64Gb maintaining the same bandwidth. Finally, the bandwidth increase requires an active thermal solution to manage hotspots that develop from highly concentrated power consumption; we propose an adaptive refresh considering temperature distribution (ART) scheme as a solution.

HBM is composed of stacked DRAM dies over a buffer die as shown in Fig. 18.2.1. All 3 configurations of HBM are shown on the right upper side. The thickness of the top core die is different depending on the number of stacks used, but the total height of the HBM is the same regardless of the number of stacks for compatibility. The purpose of the buffer die is to 1) provide routes from the TSVs related to the DRAM dies to the micro bump IOs in the PHY area. 2) provide test functionality to system makers and DRAM vendors; DRAM tests are all executed through DA (direct access) PADs. Figure 18.2.1 shows that the test block covers all of the normal paths from the PHY to the TSV for each test item. Functionality and timing margins can be guaranteed by this architecture since a DFT and a SerDes module with a PLL block is implemented for low frequency test equipment. BIST and IEEE1500 blocks support HBM tests on silicon in package (SiP) since it is difficult and inconvenient to test HBM after the chip on wafer (CoW) process is completed and the HBM is connected to a processor. System makers can test HBMs using these functions and isolate failure points when they occur.

Figure 18.2.2 shows the architecture of the core die (DRAM die). Each die has a 9Gb cell array including 1Gb cells for optional ECC. The upper diagram shows two configurations for 4H/8H and 2H cases: where 4H means that four core dies are stacked over the buffer die. In 4H/8H case the HBM is composed of two channels and each channel has two pseudo channels (PC0/PC1) which consists of 16 banks (4 bank-groups and 4banks per group). In case of 2H, one pseudo channel is divided into two different channels and each channel has eight banks which is necessary to keep the same bandwidth as in 4H/8H case. In addition, each core die has an additional 1Gb cell array for optional ECC.

Although the HBM DRAM has nearly twice the power efficiency compared to GDDR5, the power density is 3 times worse than GDDR5 because of its small area. Since high power density causes thermal issues which have critical impact on DRAM cell retention, both DRAM and controller should know the exact temperature condition of the DRAM cells so as to send proper refresh commands. This is achieved by using the proposed adaptive refresh considering temperature distribution (ART) scheme, shown in Fig. 18.2.3. Temperature sensors are placed at corner and middle areas of each die. The refresh controller for each die detects the spatial temperature difference (ΔT_X, ΔT_Y) and sets the proper refresh rate accordingly for the eight sections of the core die. The DRAM also sends the temperature code to the DRAM controller for setting the external refresh rate. By combining the external and internal refresh rate, the final refresh rate of the DRAM macro is determined as shown in Fig. 18.2.3. Thus, macros in thermal hotspots would have a high refresh rate for data retention, while refresh frequency in the cold macros are reduced, thus reducing the overall power consumption of DRAM refresh.

The bandwidth of HBM has been doubled over the previous generation HBM, yet the internal signal speed of TSV is limited. To overcome the speed limit HBM uses multiple TSVs in parallel, currently requiring more than 5000 TSVs, including some for power power. TSV defect detect and repair schemes are essential for massive TSVs, hence the number of robust TSVs for test and repair must also increase [3]. A conventional robust TSV consists of three TSVs that majority vote on the actual output values. To reduce the number of TSVs required for the robust TSVs, a built-in self test and selector scheme is proposed. Figure 18.2.4 shows the diagram of the proposed robust TSV module including the self test and selector logic. During the power-up sequence, test pulses are generated and transferred through TSVs at the buffer die first. Then the output value of a flip-flop, which enables the receiver of TSV, is changed to the high level. If the output of TSV does not toggle, due to TSV failures, then the output value of flip-flop does not change and the receiver of the TSV under test maintains its off-state.

HBM uses a micro-pillar grid array (MPGA) package and it is being tested as CoW before being individually separated. For either MPGA or CoW, an on-chip test method is required, since directly probing over 1000 IOs simultaneously is difficult. To resolve this issue, a PLL with a PI (phase interpolator) is implemented and a design for everything (DFX) IO is proposed: it includes a dedicated clock tree and DFX-integrated control with multiple-input shift register (MISR) loopback function. This IO DFX function is enabled by IEEE1500 special instructions. By utilizing JEDEC functions such as MISR, LFSR compare, and parity, the internal timing margins can be checked with simple toggle and random noisy patterns. The DFX IO consists of the PLL, PI, programmable clock pattern generator, data pattern generator, and DFX control as shown in Fig. 18.2.5. The PLL derives eight 45°-phased clocks for the PI, three phase-movable clocks are selected, and used for data pattern generation, WDQS and CLK, respectively. The clock pattern generation block guarantees that each clock has the proper number of clock cycles and phase for timing margin check. When the DFX IO feature is enabled, the internal VREF tuning can be extended from 38~66% to 20~80% of VDD with finer resolution. A 2D shmoo plot can be obtained by varying the programable clock phases and VREF according to test sequences described in Fig. 18.2.5.

Figure 18.2.6 shows the shmoo plots for t_{CK} and VREF, along with a performance summary. The DFX IO achieves 2.4Gb/s/pin operation at a 1.0V supply voltage, which results in 307GB/s of total bandwidth. The measured VREF shmoo from DFX IO shows similar result with the one from ATE-based measurements via test package as shown in Fig. 18.2.6(b). The output terminal of HBM is a micro bump which cannot be tested by a conventional test socket. Therefore, these results are obtained by the test package using the test interposer die and HBM cube. The HBM chip is fabricated using a 20nm DRAM process and the chip size is 12x8mm^2. Supply voltages are 1.2/1.2/2.5V for VDD/VDDQ/VPPE. VPPE is the word-line driver supply voltage. The refresh rate is 8k/32ms at room temperature. Fig. 18.2.7 shows the chip photographs of the core die, buffer die and the cross-sectional photos of 4H and 8H stack dies. This paper presents the 2nd generation HBM having twice the IO bandwidth, enriched test functionality, and thermal solution. While other existing candidate approaches, such as GDDR5, have limitation in bandwidth, it is expected that this device with its wide-IO and 2.5D architecture will continue to satisfy the needs of future high-performance computing systems in a power-efficient manner.

References:

[1] D. U. Lee, et al., “A 1.2V 8Gb 8-Channel 128GB/s High-Bandwidth Memory (HBM) Stacked DRAM with Effective Microbump I/O Test Methods Using 29nm Process and TSV,” ISSCC Dig. Tech. Papers, pp. 432-433, Feb. 2014.
[2] “High Bandwidth Memory (HBM) DRAM Specification,” *JEDEC Standard*, Oct 2013.
[3] D. U. Lee, et al., “An exact measurement and repair circuit of TSV connections for 128GB/s high-bandwidth memory (HBM) stacked DRAM,” *IEEE Symp. VLSI Circuits*, pp. 1-2, June 2015.

 978-1-4673-9467-3/16/$31.00 ©2016 IEEE

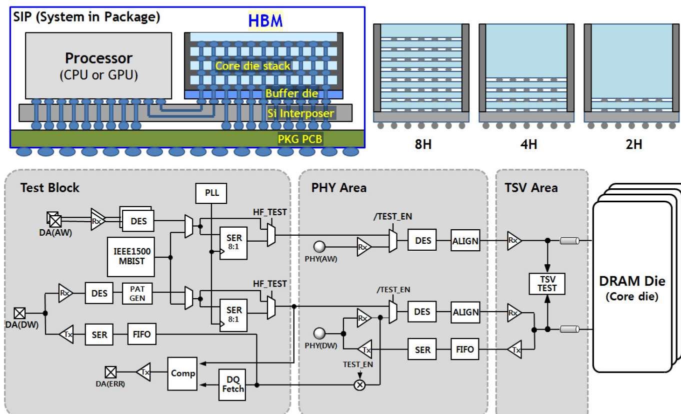


Figure 18.2.1: HBM system, various stack configurations and buffer die architecture.

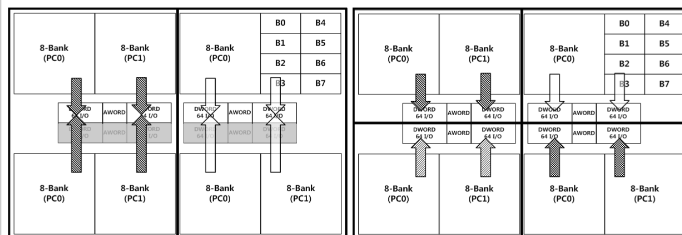


Figure 18.2.2: Core die architecture considering various configurations.

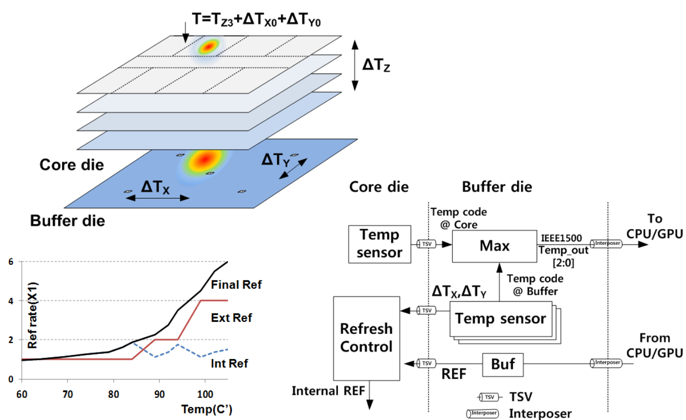


Figure 18.2.3: Adaptive Refresh considering Temperature distribution (ART) scheme.

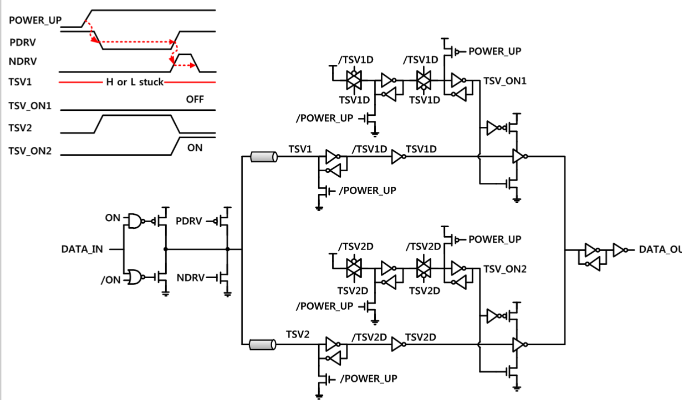


Figure 18.2.4: TSV defect auto-detect and selection scheme.

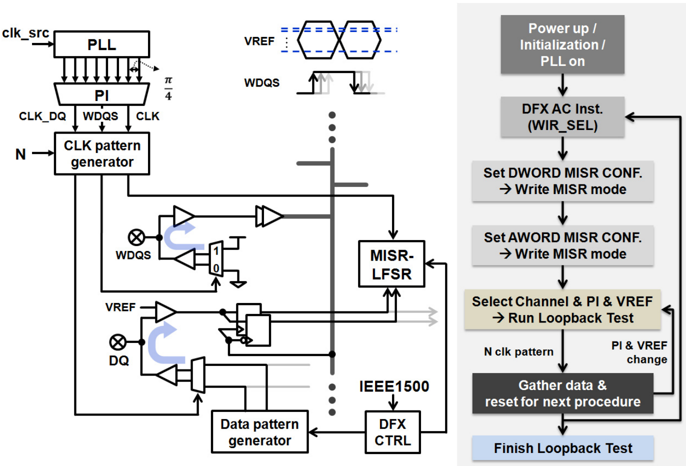
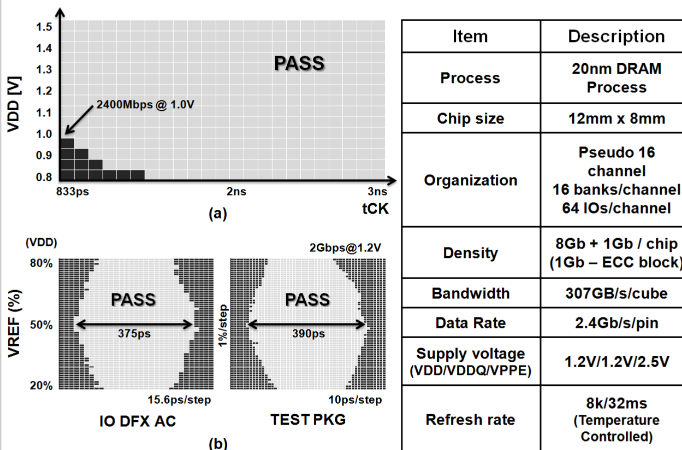


Figure 18.2.5: The block diagram of the IO DFX and its test sequence.



Item	Description
Process	20nm DRAM Process
Chip size	12mm x 8mm
Organization	Pseudo 16 channel 16 banks/channel 64 IOs/channel
Density	8Gb + 1Gb / chip (1Gb – ECC block)
Bandwidth	307GB/s/cube
Data Rate	2.4Gb/s/pin
Supply voltage (VDD/VDDQ/VPPE)	1.2V/1.2V/2.5V
Refresh rate	8k/32ms (Temperature Controlled)

Figure 18.2.6: Measurement results of (a) tCK shmoo (b) VREF shmoo plot and performance summary.

18

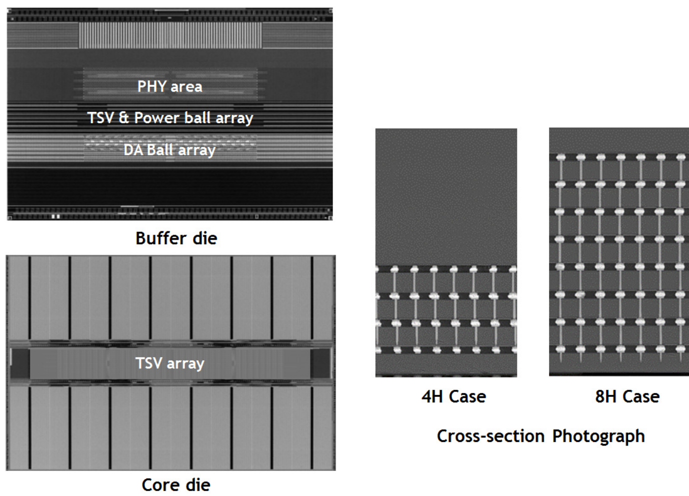


Figure 18.2.7: The chip photographs of the core die, buffer die and the cross-sectional photos of 4H and 8H stack dies.

978-1-4673-9467-3/16/$31.00 ©2016 IEEE