18.1 A 20nm 9Gb/s/pin 8Gb GDDR5 DRAM with an NBTI Monitor, Jitter Reduction Techniques and Improved Power Distribution

Hye-Yoon Joo, Seung-Jun Bae, Young-Soo Sohn,
Young-Sik Kim, Kyung-Soo Ha, Min-Su Ahn, Young-Ju Kim,
Yong-Jun Kim, Young-Ju Kim, Ju-Hwan Kim, Won-Jun Choi,
Chang-Ho Shin, Soo Hwan Kim, Byeong-Cheol Kim,
Seung-Bum Ko, Kwang-Il Park, Seong-Jin Jang, Gyo-Young Jin

Samsung Electronics, Hwaseong, Korea

A 9Gb/s/pin 8Gb GDDR5 DRAM is implemented using a 20nm CMOS process. To cover operation up to 9Gb/s, which is the highest data-rate among implemented GDDR5 DRAMs [1], this work includes an NBTI monitor, a WCK clock receiver with equalizing and duty-cycle correction modes, CML-to-CMOS converters with wide range operation, active resonant loads at the end of WCK lane, and an on-chip de-emphasis circuit at a 4-to-1 multiplexer output as shown in Fig. 18.1.1. In addition, extra power pads improve the power distribution and release the frequency limitation at the memory core.

As CMOS technology scales down, negative-bias temperature instability (NBTI) is becoming a major barrier to further speed improvements, especially in DRAM because of the absence of a high-k gate dielectric in low-cost DRAM processes. After NBTI stress, PMOS devices are degraded and the propagation delay is increased by 8% as shown in Fig. 18.1.2. This leads to decreased internal margins, duty-cycle distortion and, consequently, frequency degradation. Ring oscillators or delay lines can monitor the amount of degradation [2]. To remove the speed degradation caused by the NBTI effect, this work includes an NBTI monitor with an inverter delay chain as shown in Fig. 18.1.2. Since the reset state of the input is set to 0, only the odd-numbered inverters are exposed to NBTI stress and the width of the delayed pulse becomes wider than the reference pulse. A time-to-digital converter (TDC) converts the two pulse widths to digital codes, respectively, and the up/down counter updates the delay chain by controlling the strength of PMOS transistors. To eliminate the effect of TDC offsets, one TDC is used for both the reference pulse and the delayed pulse. The monitor makes 8b thermometer codes and feeds them back into long delays at data paths and I/O interface circuits to compensate for the decreased margin and to correct the duty-cycle distortion. The monitor operates at power-up, and updates the codes at every auto-refresh. In addition, transistors within the critical paths use PMOS header transistors, which are enabled in active mode to reduce the NBTI stress.

The memory controller transfers two half-data-rate differential clocks (WCK, WCKB) to the DRAM. The WCK clocks are divided by 2 and used to sample the write data and read data at the interface. The duty-cycle distortion or inter-symbol interference (ISI) on WCK clocks can cause jitter on the data that the DRAM transmits or receives and limits the maximum data-rate at which the DRAM can operate. Figure 18.1.3 shows a WCK clock receiver that operates in two different modes: duty-cycle correction mode and equalizing mode. The DC suppression capacitor corrects the duty-cycle distortion [3] by disconnecting the mode select transistor. The equalizing mode is for WCK stopping. DRAM uses WCK stop mode to reduce the power consumption and minimize the time required on WCK2CK training. While in power down or self-refresh, the memory controller transfers a constant 1 or 0 for WCK/WCKB. After power down exit or self-refresh exit, WCK clocks start to toggle and the first edge after the long 1 or 0 can cause ISI on the clocks [4]. Connecting the mode select transistor operates as a channel equalizer to receive the first toggle edge of WCK without ISI. If the clock receiver maintains the duty-cycle correction mode when WCK stops, the output of the receiver may oscillate because of the lack of gain in the DC signal. By changing the modes of the WCK receiver, both precise duty-cycle correction and power savings are achieved.

WCK clocks are transmitted to DQ modules in current mode logic (CML) levels rather than CMOS levels to improve power supply voltage sensitivity and reduce common-mode noise. In the DQ modules, WCK clocks are converted to CMOS level signals by a converter shown in Fig. 18.1.4. In order to enable operation over a wide range of data-rates, there are two modes. In high frequency mode, an AC-coupled inverter with resistive feedback and cross-coupled inverters convert the CML input to CMOS output with duty-cycle correction [3]. If the input WCK clocks have a frequency lower than a certain cutoff frequency, the attenuated

signal between rising and falling edges may oscillate due to the inverter with resistive feedback. To solve this problem, another technique for low-frequency mode is used. The inverter A with feedback directly injects the logic threshold voltage at node B in Fig. 18.1.4. The hysteresis latch detects the rising and falling edges and recovers the edge information to a CMOS-level clock signal. The strength of the hysteresis latch is relevant to the cutoff frequency of the high-pass filter. With high cutoff frequency, the strength of the hysteresis latch should be weak enough to detect the low-frequency clock signals, but should be strong enough not to recover the glitches caused by power supply noise. Using a low cutoff frequency may ease the design of the hysteresis latch, but the time required for WCK clocks to settle to a constant state is increased. In this work, the boundary of high and low frequency modes is set near 2.5Gb/s and the size of the AC coupling capacitor is designed considering the WCK clock settling time specification when operating at 9Gb/s.

The LC resonance improves the skew and jitter on the clock distribution [5]. In this work, the active resonant load that includes the CML oscillator with tunable RC time constant is used as shown in Fig. 18.1.5. The 3b frequency codes control the oscillation frequency to maximize the ability of jitter reduction. The simulated results show that the jitter caused by power noise reduced by 30% at the optimum oscillation frequency. At the read data path, there is a 4-to-1 multiplexer where the data toggles at the full data-rate at the output. To improve signal integrity, an on-chip de-emphasis circuit [1] is used. It receives the 4-to-1 multiplexer output and injects the delayed and inverted data at the same point, so it operates as a de-emphasis circuit unless the delay of three inverters exceeds 1UI. To prevent oscillation or direct current while the DRAM is not in read operation, the inverter has enable/disable control transistors.

To operate up to 9Gb/s with 8b pre-fetch data, the core operation cycle is 1.125Gb/s. There is a fundamental limitation to increasing the frequency of the memory core because of the poor DRAM process for low cost. Moreover, if there is large amount of voltage drop due to complex core operations, the core frequency degradation gets worse. In this work, extra power pads in the memory core region improve the power distribution. The simulated results show that the dynamic and static voltage drop in burst read mode is reduced by 46% and 60%, respectively.

In Fig. 18.1.6, the device pass/fail shmoo is shown with various supply voltages and clock periods. The maximum operating frequency at a 1.45V supply voltage is improved from 6.9 to 9Gb/s by the NBTI monitor and compensation in accelerated degradation tests. The maximum WCK period of 220ps is the tester limit. The measured pass/fail eye shmoos during read operation at 9Gb/s with and without jitter reduction techniques – active resonant loads at the WCK lane and on-chip de-emphasis at the 4-to-1 multiplexer output – are shown. The pass window is increased by 25ps. Figure 18.1.7 shows the chip micrograph and summary.

References:
[1] Seung-Jun Bae et al., "A 40nm 2Gb 7Gb/s/pin GDDR5 SDRAM with a Programmable DQ Ordering Crosstalk Equalizer and Adjustable Clock-Tracking BW," *ISSCC Dig. Tech. Papers*, pp. 498–500, Feb. 2011.
[2] E. Saneyoshi et al., "A Precise-Tracking NBTI-Degradation Monitor Independent of NBTI Recovery Effect," *ISSCC Dig. Tech. Papers*, pp. 192–193, Feb. 2010.
[3] C. Menolfi et al., "A 16Gb/s Source-Series Terminated Transmitter in 65nm CMOS SOI," *ISSCC Dig. Tech. Papers*, pp. 446–447, Feb. 2007.
[4] J. Zerbe et al., "A 5.6Gb/s 2.4mW/Gb/s Bidirectional Link With 8ns Power-On," *IEEE Symp. VLSI Circuits*, pp. 82-83, June 2011.
[5] S. Chan et al., "A Resonant Global Clock Distribution for the Cell Broadband-Engine Processor," *ISSCC Dig. Tech. Papers*, pp. 512–513, Feb. 2008.
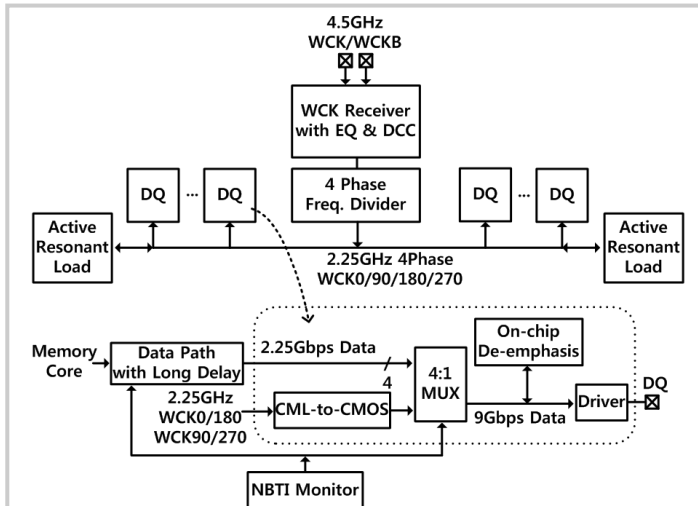
Figure 18.1.1: Block diagram of the DRAM interface that includes NBTI monitor, WCK receiver with active resonant loads, DQ modules with CML-to-CMOS converter and on-chip de-emphasis circuit.
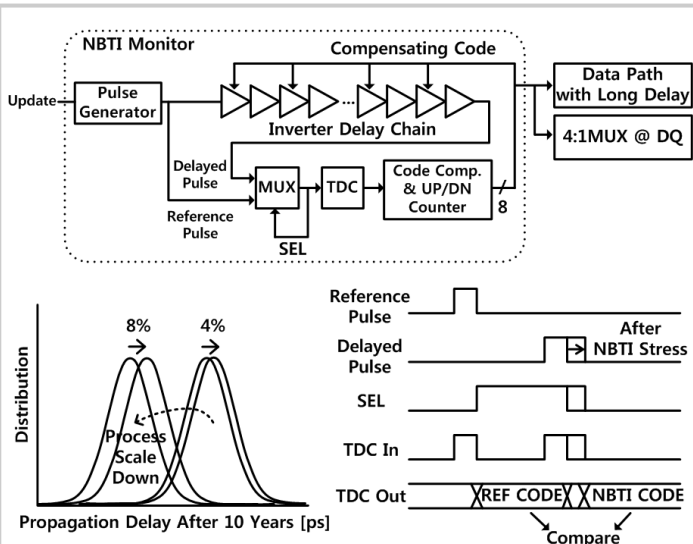
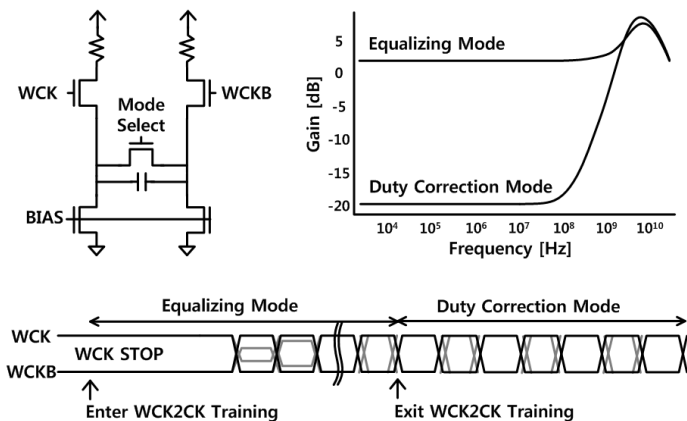**Figure 18.1.2: Effect of NBTI stress and the block diagram of the NBTI monitor.**

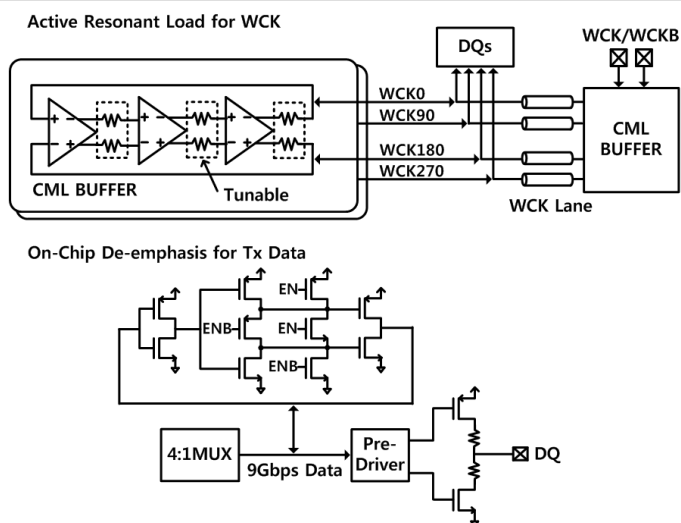**Figure 18.1.3: WCK receiver with equalizing and duty-cycle correction modes.**

Figure 18.1.4: CML-to-CMOS converter with high-frequency and low-frequency modes.

Figure 18.1.5: Active resonant load for WCK jitter reduction and on-chip de-emphasis circuit for Tx data jitter reduction.

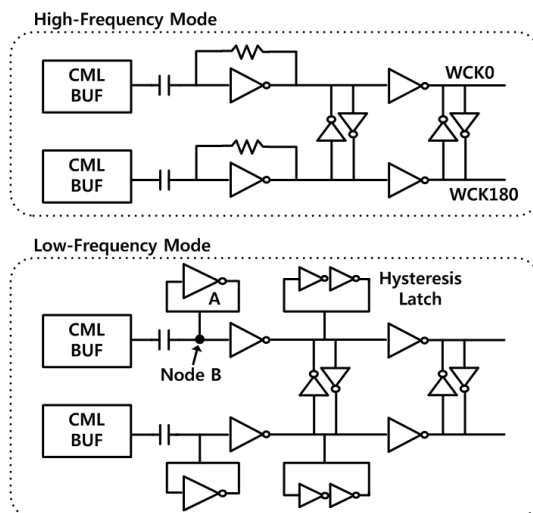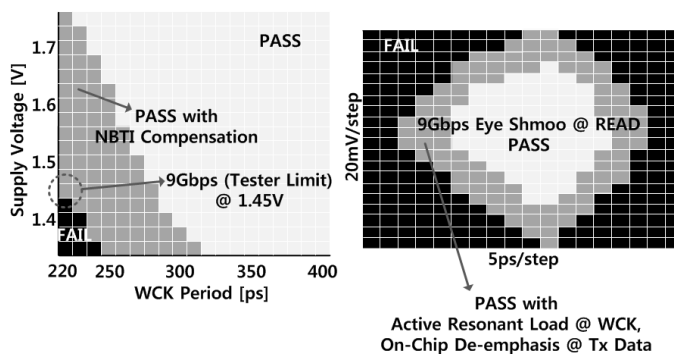**Figure 18.1.6: Pass/fail voltage-clock period shmoo and pass/fail eye shmoo at 9Gb/s.**

18

Chip Micrograph

128M

DQs
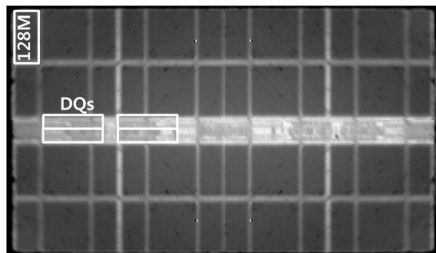
20nm CMOS 3 metal
170 Ball Flip-Chip
9Gbps/pin @ CL23
16 Banks
4 Bank Group
128M x 64
8 Burst Length
X32 I/O
62.04mm^2
1.5V

Figure 18.1.7: Chip micrograph and summary.