28.2 A 16Gb 27Gb/s/pin T-coil based GDDR6 DRAM with Merged-MUX TX, Optimized WCK Operation, and Alternative-Data-Bus

Daewoong Lee, Hye-Jung Kwon, Daehyun Kwon, Jaehyeok Baek, Chulhee Cho, Sanghoon Kim, Donggun An, Chulsoon Chang, Unhak Lim, Jiyeon Im, Wonju Sung, Hye-Ran Kim, Sun-Young Park, HyoungJoo Kim, Hoseok Seol, Juhwan Kim, Jungbum Shin, Kil-Young Kang, Yong-Hun Kim, Sooyoung Kim, Wansoo Park, Seok-Jung Kim, Chanyong Lee, Seungseob Lee, TaeHoon Park, ChiSung Oh, Hyodong Ban, Hyungjong Ko, Hoyoung Song, Tae-Young Oh, SangJoon Hwang, Kyung Suk Oh, JungHwan Choi, Jooyoung Lee

Samsung Electronics, Hwaseong, Korea

Graphic DRAMs have been developed to increase maximum I/O interface speeds to satisfy the demand of high-performance graphic applications [1-5]. Recently, PAM4 signaling was utilized to increase the I/O bandwidth up to 22Gb/s/pin [5]. However, the reduced voltage margin of PAM4, compared to NRZ, complicates circuit design; margins also become worse with a reduced power supply. This paper achieves 27Gb/s in NRZ, a 1.5× speed enhancement, by improving on previous GDDR6 [3]. A T-coil is designed, for the first time in a DRAM process, so that the maximum operation frequency is increased. The proposed merged-MUX TX increases the maximum speed and reduces power and area consumption. A quad-skew training technique enables a wider clock sampling margin for WCK: up to 3ps, which is 8.1% of 1UI at 27Gbp/s/pin. Furthermore, a dual-mode frequency divider allows a wide-range operation from sub-1Gb/s/pin to 27Gb/s/pin. An alternative-data-bus (ADB) is proposed to solve the frequency limit of the data bus.

Among various equalization methods the T-coil has been widely used because it is a passive equalizer that consumes no power. T-coils have been designed with more than two top-metal layers with low resistance to obtain bandwidth improvements in the I/O interface [6, 7]. However, in a DRAM process technology the number of available layers is limited and only the recently supported redistribution layer (RDL) has a low sheet resistance (R_S). The R_S of the RDL is about ten times smaller than the nearest lower layer. Therefore, a conventional T-coil design that uses more than two metal layers, cannot be used in a DRAM process technology. In the proposed GDDR6, a T-coil design is enabled using RDL-based T-coil layer, where only a single metal layer is utilized. Figure 28.2.1 shows a schematic of the proposed T-coil. Since each DQ has its own T-coil and the RDL is also used for the power network, an area-efficient design is considered; the T-coil is drawn in rectangular aspect ratio, rather than as a square, to allow space for the power lines. With an RDL-based T-coil layer that supports an optimum thickness and width for the T-coil design, each length of the inductance L1 and L2 is determined to achieve the bandwidth enhancement for both RX and TX (L1 < L2). As a result, an asymmetric T-coil is designed considering C_{TX} as well as $C_{ESD} + C_{RX}$ of the bi-directional I/O interface of DQ.

Figure 28.2.2(a) shows a previous ZQ-coded transmitter [3], which improved ISI and power-supply-induced jitter (PSIJ) by reducing the number of critical path stages from 3 to 2 stages. However, the scattered multiple 4:1 multiplexers increase the distributed WCK2 metal loading and mismatch among multiplexers cause the skew of multiplexer output, which results in DOUT SI degradation. To solve these problems, the multiplexer-merged transmitter is, shown in Fig. 28.2.2(b), is proposed. All 4:1 multiplexers of [3] are merged into a single 4:1 multiplexer to reduce WCK2 metal loading and to reduce skew. Moreover, the logic gates in [3] are not required in the proposed transmitter, which allows for power and area reduction as the ZQ codes are applied to the pull-up and pull-down drivers. As a result, 27Gb/s is achieved in the proposed merged-MUX transmitter by improving ISI and PSIJ as well as power consumption.

Since the quad skew of WCK reduces the sampling margin, quad-skew adjustment is required for a high-speed operation. Therefore, this paper proposes a quad-skew training technique. The quad-skew between the divided WCKs can be removed by controlling the delay of the differential WCK signals, WCK0 and 180 or WCK90 and 270, as shown in Fig. 28.2.3(a). Each delay is current (I1 and I2) programmable and the delay adjustment is performed after checking the quad skew of the transmitted clock-pattern data from DRAM. In the proposed design, a quad skew compensation of 3ps (8.1% of 1UI at 27Gbps) is possible. Figure 28.2.3(b) shows the ½-frequency divider consisting of two dual-mode CML latches: it operates in dual-mode and supports a wide range of input frequencies. As shown in Fig. 28.2.3(b), when LF_ON is high (LF_ONB is low) P1, P2 and N1 are turned off, hence the CML resistance increases and its current decreases. Thus, the divider operates in a low-frequency mode with a 8GHz (16Gb/s) center frequency. When LF_ON is low (LF_ONB is high), the CML resistance and its current changes in the opposite direction and the center frequency is close to 12GHz (24Gb/s). As a result, the proposed frequency divider covers the sub-1Gb/s/pin to 27Gb/s/pin operating range.

The previous ZQ calibration method shown in Fig. 28.2.4(a) obtains pull-up (PU) and pull-down (PD) code from loop 1 and 2 without using a T-coil. However, when an I/O T-coil is used the ZQ calibration needs to include T-coil impact as shown in Fig. 28.2.4(b); the total resistance of the T-coil is about 10% of the 40Ω pull-down resistance. In Fig. 28.2.4(b) loop 1 performs the same operation, as in Fig. 28.2.4(a), for the PU code. Then this PU code is used to calibrate the PD code, considering the T-coil, in loop 2. Finally, the calibrated PD code is used to update the PU code, considering the T-coil, in loop 3. To minimize the chip size two T-coils are removed; Fig. 28.2.4(c) shows the final ZQ calibration method that uses a single T-coil. Although two T-coils are removed, loop 3 still calculates the same PU code as CODE2PU of Fig. 28.2.4(b), because both T-coils for PU and PD are removed and CODE1PD is obtained with T-coil. Therefore, only a single T-coil is used in the proposed calibration method. Note that the calculated PU code without T-coil (CODE1PU) is used for ODT and the calculated PU and PD codes with T-coil (CODE1PD, CODE2PU) are used for OCD codes.

The operating frequency, considering a fixed t_{CCDS} of $2t_{CK}$ for the timing window of the data-line fetch (data window), is limited. Fig. 28.2.5 shows the chip and bus architecture of the proposed GDDR6: the data bus of GDDR6 is divided into banked group bus (BG-bus) and global bus (G-bus). The data window of the BG-bus follows t_{CCDL}, which is $2t_{CK}$ without bank-group and $3t_{CK}$ or $4t_{CK}$ with bank-group operation. In high-frequency operation the bank group is enabled; thus, the data window of BG-bus is enlarged to $3t_{CK}$ or $4t_{CK}$. On the other hand, the data window of G-bus is t_{CCDS} ($2t_{CK}$) regardless of bank-group (i.e., frequency), making the G-bus the frequency limiter. To overcome this problem, we propose a bussing scheme that enlarges the data window for G-bus to $2t_{CK}$. The key idea is to transmit data in an alternating fashion using two identical data buses. Figure 28.2.5 shows the timing diagram of the proposed bussing scheme: there are two G-BUSes, even and odd. The data generated from even column accesses is transmitted on the even G-BUS while data generated from odd column accesses is transmitted on the odd G-BUS. The proposed scheme increases the data window of G-BUS by 2× to achieve a high-bandwidth operation up to 27 Gb/s.

The measured t_{CK} shmoo shown in Fig. 28.2.6(a) indicates that 27Gb/s operation is achieved at 1.35V and above. Fig. 28.2.6(b) shows the measured 0101 pattern on the error detect code (EDC) pin for a 27Gb/s data rate. Fig. 28.2.7 shows the fabricated die chip photo.

References:

[1] H.-Y. Joo et al., "A 20nm 9Gb/s/pin 8Gb GDDR5 DRAM with an NBTI monitor, jitter reduction techniques and improved power distribution," *ISSCC*, pp. 314-315, 2016.

[2] M. Brox et al., "An 8Gb 12Gb/s/pin GDDR5X DRAM for cost-effective high-performance applications," *ISSCC*, pp. 388-389, 2017.

[3] Y.-J. Kim et al., "A 16Gb 18Gb/S/pin GDDR6 DRAM with per-bit trainable single-ended DFE and PLL-less clocking," *ISSCC*, pp. 204-206, 2018.

[4] K. Kim et al., "A 24Gb/s/pin 8Gb GDDR6 with a Half-Rate Daisy-Chain-Based Clocking Architecture and IO Circuitry for Low-Noise Operation," *ISSCC,* pp. 344-346, 2021.

[5] T. M. Hollis et al., "An 8Gb GDDR6X DRAM Achieving 22Gb/s/pin with Single-Ended PAM4 Signaling," *ISSCC*, pp. 348-350, 2021.

[6] Y. Chen et al., "A 25Gb/s hybrid integrated silicon photonic transceiver in 28nm CMOS and SOI," *ISSCC*, pp. 1-3, 2015.

[7] J. Kim et al., "A 16-to-40Gb/s quarter-rate NRZ/PAM4 dual-mode transmitter in 14nm CMOS," *ISSCC*, pp. 1-3, 2015.

978-1-6654-2800-2/22/$31.00 ©2022 IEEE

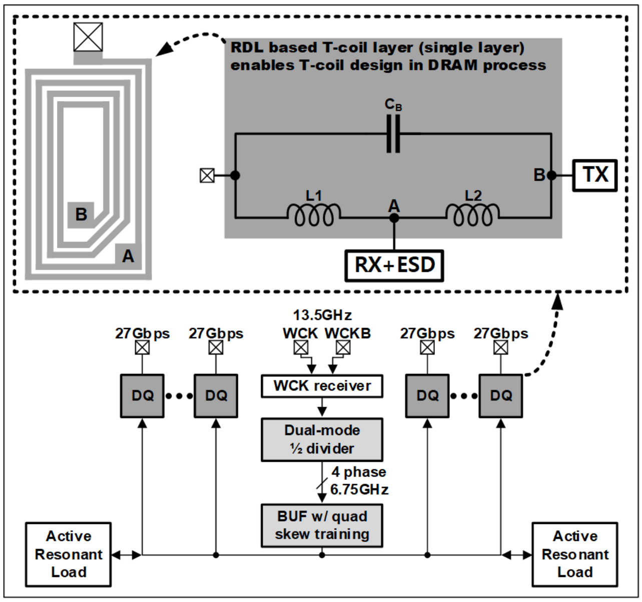

Figure 28.2.1: 27Gbps GDDR6 I/O block diagram with asymmetric T-coil.

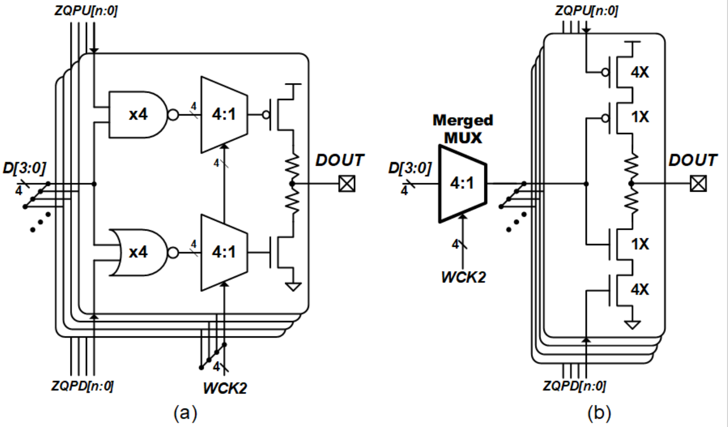

Figure 28.2.2: (a) Previous ZQ-coded transmitter [3] (b) Proposed merged-MUX transmitter.

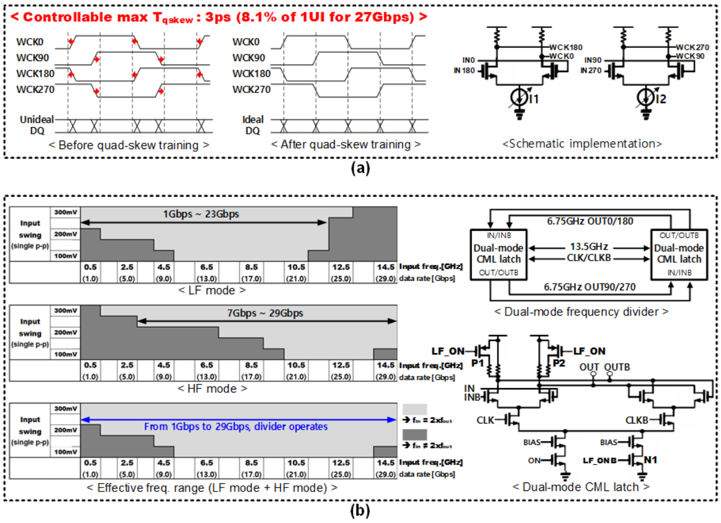

Figure 28.2.3: WCK clocking improvement: (a) Quad-skew training (b) Dual-mode frequency divider.

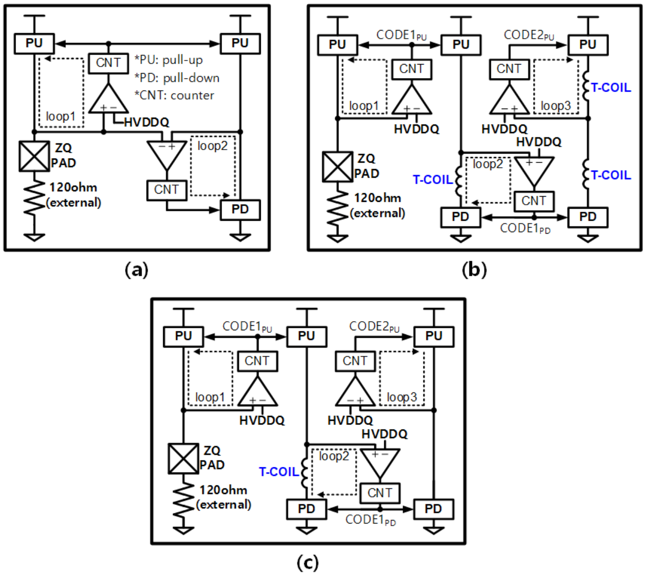

Figure 28.2.4: ZQ calibration: (a) without T-coil (b) with 3 T-coil (area-consuming) (c) with only 1 T-coil (area saving).

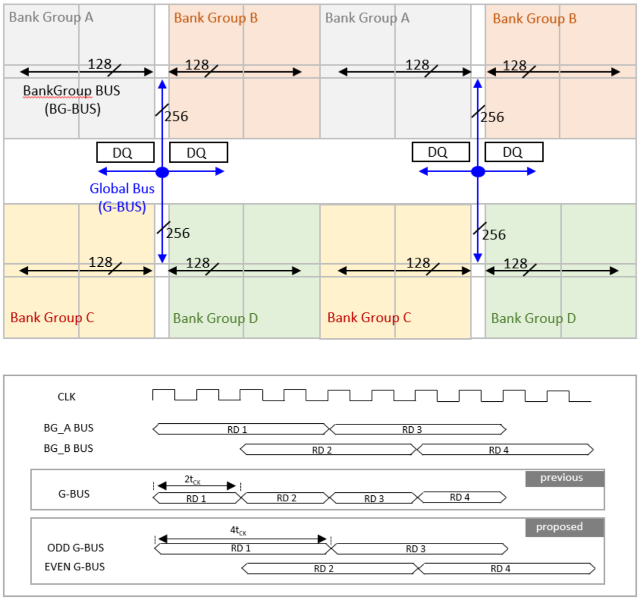

Figure 28.2.5: Data path of GDDR6 1CH data bus and alternative data bus concept.

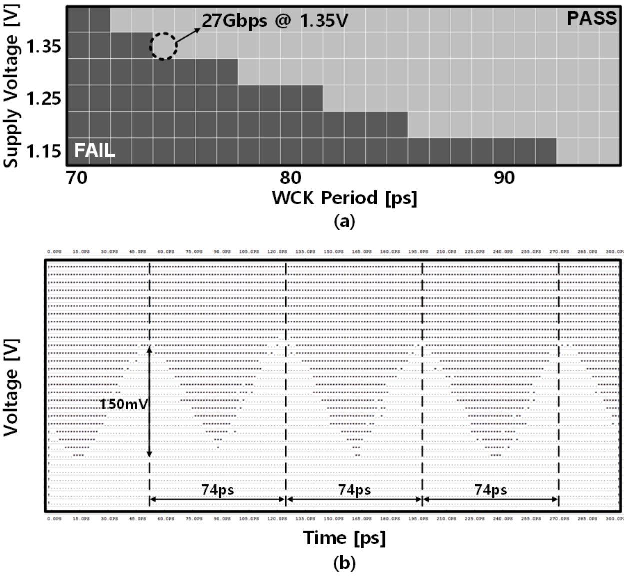

Figure 28.2.6: Measurement result. (a) Frequency-voltage shmoo (b) measured 27Gb/s output waveform for a 0101 pattern at 1.35V.

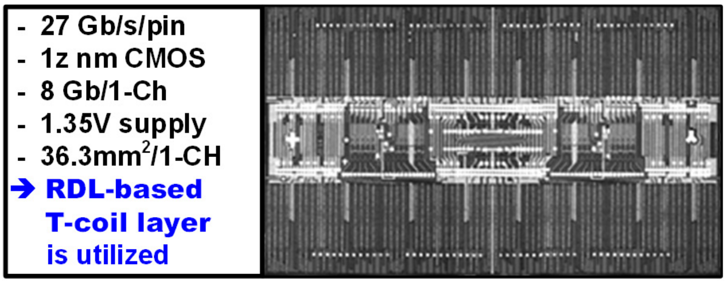

Figure 28.2.7: Chip micrograph.

978-1-6654-2800-2/22/$31.00 ©2022 IEEE