Session 28 Overview:
High-Density Memories and High-Speed Interface
MEMORY SUBCOMMITTEE

Session Chair: Seung-Jae Lee
Samsung, Hwaseong, Korea

Session Co-Chair: Dongkyun Kim
SK Hynix, Icheon, Korea

Innovations in 3D NAND Flash, very-high density 5b/cell, will be introduced. New challenges and solutions to improving reliability for DRAM will be presented: including probabilistic aggressor tracking against row-hammer attacks and core bias modulation to overcome process limitations. Meanwhile, the evolution of the memory high-speed interface continues: new technologies are introduced such as single-ended PAM4 signaling to achieve speeds exceeding 16Gb/s/pin, offset-calibration HBM3-interface technology to achieve a 1.15TB/s bandwidth, an input jitter filtering digital PLL technology, and an edge boosting equalizer using a t-coil.

8:30 AM

28.1 A 1.67Tb, 5b/Cell Flash Memory Fabricated in 192-Layer Floating Gate 3D-NAND Technology and Featuring a 23.3Gb/mm² Bit Density

Ali Khakifirooz, Intel, Santa Clara, CA

In Paper 28.1, Intel presents a 1.67-Tb 5b/cell Flash memory fabricated in a 192-layer floating-gate 3D-NAND technology, featuring a 23.3Gb/mm² bit density with a die capacity of 1.67Tb within an 73.3mm² area, and a t_R and t_{PROG} of 354µs and 5500µs.

9:00 AM

28.2 A High-Performance 1Tb 3b/Cell 3D-NAND Flash with a 194MB/s Write Throughput on over 300 Layers

Byungryul Kim, SK hynix Semiconductor, Icheon, Korea

In Paper 28.2, SK Hynix presents a high-performance 1-Tb 3b/cell 3D-NAND Flash with 194MB/s write throughput for over 300 layers. Five new schemes are introduced and these design technologies enable a high-performance (a t_R of 34µs and a program throughput of 194MB/s) 1-Tb 3b/cell 3D-NAND Flash memory with a greater than 20Gb/mm² bit density, which uses the peripheral-circuit-under-cell-array architecture.

978-1-6654-9016-0/23/$31.00

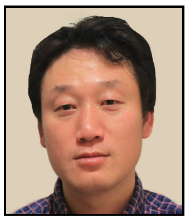
9:30 AM

28.3 A 4nm 16Gb/s/pin Single-Ended PAM4 Parallel Transceiver with Switching-Jitter Compensation and Transmitter Optimization

Jahoon Jin, Samsung Electronics, Hwaseong, Korea

In Paper 28.3, Samsung presents a 4-nm 16-Gb/s/pin single-ended PAM4 parallel transceiver with switching-jitter compensation and transmitter optimization. This paper achieves 0.764pJ/b within 0.0073mm^2. A relaxed TX termination of 20Ω and RX termination of 50Ω is adopted to maximize the eye opening.

10:15 AM



28.4 A 4nm 1.15TB/s HBM3 Interface with Resistor-Tuned Offset-Calibration and In-Situ Margin-Detection

Kwanyeob Chae, Samsung Electronics, Hwaseung, Korea

In Paper 28.4, Samsung presents a 4nm 1.15TB/s HBM3 interface with resistor-tuned offset-calibration and in-situ margin-detection for reliable high-speed memory access. In this work, a compact slim bit-slice architecture in conjunction with a stacked I/O structure achieves a reliable high bandwidth, up to 1.15TB/s.

10:45 AM

28.5 A 900μW, 1-4GHz Input-Jitter-Filtering Digital-PLL-Based 25%-Duty-Cycle Quadrature-Clock Generator for Ultra-Low-Power Clock Distribution in High-Speed DRAM Interfaces

Yuhwan Shin, Korea Advanced Institute of Science and Technology, Daejeon, Korea

In Paper 28.5, KAIST shows a 900-μW 1 – 4-GHz input-jitter-filtering digital-PLL-based 25%-duty-cycle quadrature-clock generator for ultra-low-power clock distribution for high-speed DRAM interfaces. This work presents a low-power clock-distribution scheme for DRAM, using a quadrature clock generator, which can generate accurate 25%-DC quadrature signals over a 1 - 4GHz range.

11:00 AM

28.6 A 32Gb/s/pin 0.51pJ/b Single-Ended Resistor-less Impedance-Matched Transmitter with a T-Coil-Based Edge-Boosting Equalizer in 40nm CMOS

Jung-Hun Park, Seoul National University, Seoul, Korea

In Paper 28.6, Seoul National University presents a 32-Gb/s/pin 0.51-pJ/b single-ended resistorless impedance-matched transmitter with a t-coil-based edge-boosting equalizer in 40nm CMOS. The 2-tap t-coil-based edge-boosting equalizer compensates for the high-frequency impedance drop and does not consume static current for non-transition sequences achieving a power efficiency of 0.51pJ/b.

11:15 AM

28.7 A 1.1V 6.4Gb/s/pin 24Gb DDR5 SDRAM with a Highly-Accurate Duty Corrector and NBTI-Tolerant DLL

Daehyun Kwon, Samsung Electronics, Hwaseong, Korea

In Paper 28.7, Samsung presents a 1.1-V 6.4-Gb/s/pin 24-Gb DDR5 SDRAM with a highly-accurate duty-cycle corrector and an NBTI tolerant DLL. The 24-Gb density DDR5 occupies 71.8mm^2/channel, and is implemented in a 4th-generation 10-nm DRAM technology.

11:45 AM

28.8 A 1.1V 16Gb DDR5 DRAM with Probabilistic-Aggressor Tracking, Refresh-Management Functionality, Per-Row Hammer Tracking, a Multi-Step Precharge, and Core-Bias Modulation for Security and Reliability Enhancement

Woongrae Kim, SK hynix Semiconductor, Icheon, Korea

In Paper 28.8, SK Hynix presents a 1.1-V 16-Gb DDR5 DRAM with probabilistic-aggressor tracking, a refresh-management function, per-row hammer tracking, a multi-step precharge, and core-bias-voltage modulation for security and reliability enhancement. This comprehensive scheme leads to a failure-probability reduction due to row hammer attacks by 93.1%, and an improvement to cell-retention time of 17%.

28.1 A 1.67Tb, 5b/Cell Flash Memory Fabricated in 192-Layer Floating Gate 3D-NAND Technology and Featuring a 23.3Gb/mm² Bit Density

Ali Khakifirooz[1], Eduardo Anaya[2], Sriram Balasubrahmanyam[2], Geoff Bennett[1], Daniel Castro[2], John Egler[2], Kuangchan Fan[2], Rifat Ferdous[1], Kartik Ganapathi[1], Omar Guzman[2], Chang Wan Ha[1], Rezaul Haque[2], Vinaya Harish[2], Majid Jalalifar[2], Owen W. Jungroth[2], Sung-taeg Kang[1], Golnaz Karbasian[1], Jee-Yeon Kim[1], Siyue Li[2], Aliasgar S. Madraswala[2], Srivijay Maddukuri[2], Amr Mohammed[1], Shanmathi Mookiah[2], Shashi Nagabhushan[2], Binh Ngo[2], Deep Patel[2], Sai Kumar Poosarla[2], Naveen V. Prabhu[2], Carlos Quiroga[2], Shantanu Rajwade[1], Ahsanur Rahman[2], Jalpa Shah[2], Rohit S. Shenoy[1], Ebenezer Tachie Menson[2], Archana Tankasala[1], Sandeep Krishna Thirumala[1], Sagar Upadhyay[2], Krishnasree Upadhyayula[2], Ashley Velasco[2], Nanda Kishore Babu Vemula[2], Bhaskar Venkataramaiah[2], Jiantao Zhou[1], Bharat M. Pathak[2], Pranav Kalavade[1]

[1]Intel, Santa Clara, CA, [2]Intel, Folsom, CA

Successful deployment of multiple generations of the 4b/cell (QLC) floating-gate 3D-NAND technology has paved the way for the industry-wide adoption of QLC [1-4]. The transition to 5b/cell (PLC) will be another steppingstone to accelerating bit density growth and expanding Flash storage to wider markets, where a lower cost at a reasonable performance is the paramount requirement.

In this paper, we present the first PLC NAND chip that is fabricated in a 192-layer floating-gate (FG) technology. With a die capacity of 1.67Tb and area of 73.3mm², it delivers a bit density of 23.3Gb/mm². The chip can also be configured as a 1.33Tb QLC or a 1Tb 3b/cell (TLC), achieving bit densities of 18.6Gb/mm² and 14.0Gb/mm², which are 24% and 21% better than the best previously reported QLC [4] and TLC [5] bit densities. Figure 28.1.1 shows the bit density scaling trend with the number of layers, demonstrating superior scaling efficiency of this work compared to other QLC implementations. We describe key innovations to enable reliable PLC operation and the features implemented to support system-level usage, including a fast soft-bit read algorithm capable of handling the presence of defective BLs; a fast read-calibration algorithm, and a reverse-read waveform to improve the read margin, SLC-write-through and program suspend, as well as a resume algorithm compatible with the above read operations.

Programming 32 states to encode 5b of data per cell, within a limited threshold voltage window, poses a significant challenge. To minimize the interference from neighboring WLs, we use a two-pass coarse/fine programming algorithm. The resilience of floating gate technology to charge loss, compared to charge-trap Flash technology that suffers from lateral charge diffusion in the nitride layer, is a key enabler to increasing the number of bits per cell. However, both technologies are affected by random telegraph noise (RTN) due to traps in the polysilicon channel and the interfaces, which imposes a lower bound on how tight the states can be placed. As shown in Fig. 28.1.2, reducing the program gate step is an efficient way to tighten the threshold voltage distributions for TLC and QLC at the cost of increased program time; however, it offers diminishing benefits beyond what is typically used for QLC. Therefore, increasing the error correction code (ECC) capability is required to reliably read the data. Most QLC implementations have already increased the number of ECC bytes, compared to their TLC counterparts. However, to avoid the area penalty we kept the number of ECC bytes unchanged and augmented ECC correction capabilities with a fast soft-bit read (FSBR) algorithm. To maximize the information that can be encoded in 2b (for a total of 3b including the hard-bit data), we implemented a 7-strobe read algorithm, which groups the bits into four buckets from the strongest to weakest confidence. This is achieved by sensing the cells at different sense currents instead of different WL voltages, by modulating the voltage applied to the back of the sensing capacitor after it is discharged in proportion to the BL current, as shown in Fig. 28.1.3. The average read time (t_R) for the proposed FSBR is 354µs and a balanced 6-6-7-6-6 Gray code was used to limit the maximum t_R to 386µs.

NAND-Flash memories typically include additional redundant columns to repair defective BLs. In this work, to further decrease the die area, we reduced the number of redundant columns by more than 70%, and allow for unrepaired defective BLs, which may be present in a small percentage of the dies, so long as the unrepaired BLs contribution to the raw bit error rate (RBER) is significantly smaller than the error correction capability. However, the presence of unrepaired defective BLs adversely impacts the quality of soft-read operation since these bits are sensed as the strongest 0s and 1s. To circumvent this, special open/short sensing operations were added to the read algorithm to identify defective BLs and place them in the weakest confidence bucket, as shown in Fig. 28.1.3.

With the tight spacing between the threshold voltage states, it is extremely important to place the read levels at the optimal location between neighboring states. While optimum read levels are set during NAND manufacturing, die-to-die variations during the lifetime of the NAND operation and under cross-temperature (x-temp) conditions cannot be fully compensated. To address this, we implemented a 5-strobe fast read calibration algorithm by modulating the voltage applied to the back of the sensing capacitor and counting the number of bits that flip between strobes. Experimental data shows that this scheme is more accurate than the 3-strobe algorithm proposed earlier [1]. Moreover, compared to algorithms that are based on counting the total number of bits that belong to different states [3], the proposed algorithm does not require a perfectly uniform threshold voltage distribution, which is difficult to achieve with an increased number of states. Figure 28.1.4 reports the RBER distribution; thereby, demonstrating the robustness of the proposed algorithm to bring the RBER well below the ECC correction capability even under x-temp conditions.

To further improve the read margin and reduce the RBER, a reverse read waveform is implemented, as shown schematically in Fig. 28.1.5. Traditionally, the motivation to implement a reverse read has been to reduce t_R by avoiding the slow ramp down of the pass voltage to the lowest read level, which is present in a forward read waveform. However, in this work the main motivation is to improve the read margin for the higher read levels with a negligible effect on the lower levels: shown through experimental data in Fig. 28.1.5. With a forward read waveform, the cells with a higher threshold voltage are kept in the depletion regime during earlier read levels, with a significantly different trap occupancy compared to the inversion regime where they are being sensed. The reverse waveform improves the read margin by maintaining these cells in the inversion regime prior to their corresponding sense operation.

In order to enable a balanced Gray data encoding, all five pages of data are needed in both the first and second pass of the program algorithm. While this is the norm for most QLC implementations except [1], it requires the storage of a few megabytes of data per die in a DRAM or similar media. Instead, we use a 1b/cell (SLC) cache on the NAND die to store the data needed for the two-pass PLC programming algorithm. To keep the area overhead of the SLC cache to less than 2%, we improved the SLC reliability to 250k program/erase (P/E) cycles, commensurate with 1k of P/E cycle capability in the present PLC work.

The capability to suspend the program algorithm, to service read requests, is extremely important for enterprise-level mixed workloads. To minimize the static page buffer (SPB) area, we did not add extra data latches beyond what is needed for the QLC program operation and encoded inhibit information during the program algorithm as erase data (L0). To support FSBR during a program suspend, a minimum of 3 data latches are needed per BL. To enable this, we rely on the fact that a copy of the data being programmed is available in the SLC cache. When a program suspend command is received, the die constructs the inhibit information (INH) by performing a logical AND operation between the data latches, keeps INH in one of the latches, and releases the rest of the latches for the read operation. To resume the program operation, the user data is first read from the SLC cache, combined with the INH information through a logical OR operation, and then restored to the corresponding data latches, as illustrated in Fig. 28.1.6.

A die photograph of the fabricated NAND chip is shown in Fig. 28.1.7 along with key metrics of the present work.

References:

[1] A. Khakifirooz et al., "A 1Tb 4b/Cell 144-Tier Floating-Gate 3D-NAND Flash Memory with 40MB/s Program Throughput and 13.8Gb/mm² Bit Density," *ISSCC*, pp. 424-425, 2021.

[2] T. Pekny et al., "A 1-Tb Density 4b/Cell 3D-NAND Flash on 176-Tier Technology with 4-Independent Planes for Read Using CMOS-Under-the-Array," *ISSCC*, pp. 132-133, 2022.

[3] W. Cho et al., "A 1-Tb, 4b/cell, 176-stacked-WL 3D-NAND Flash Memory with Improved Read Latency and a 14.8Gb/mm² Density," *ISSCC*, pp. 134-135, 2022.

[4] J. Yuh et al., "A 1-Tb 4b/Cell 4-Plane 162-Layer 3D Flash Memory with a 2.4-Gb/s I/O Speed Interface," *ISSCC*, pp. 130-131, 2022.

[5] M. Kim et al., "A 1Tb 3b/Cell 8th-Generation 3D-NAND Flash Memory with 164MB/s Write Throughput and a 2.4Gb/s Interface," *ISSCC*, pp. 136-137, 2022.

 978-1-6654-9016-0/23/$31.00

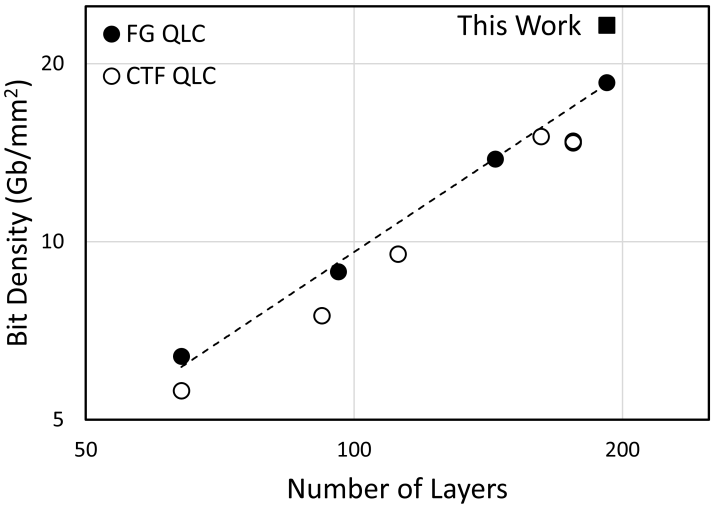


Figure 28.1.1: Bit density comparison of the proposed PLC die vs prior QLC implementations in floating-gate (FG) and charge-trap Flash technologies.

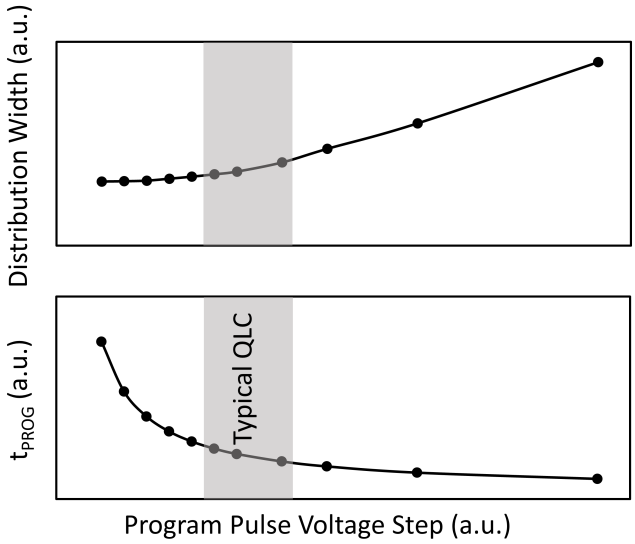


Figure 28.1.2: Average threshold voltage distribution width and the 2nd pass program time as a function of the program gate step. Plots show the diminishing benefit of reducing the gate steps in tightening the distributions.

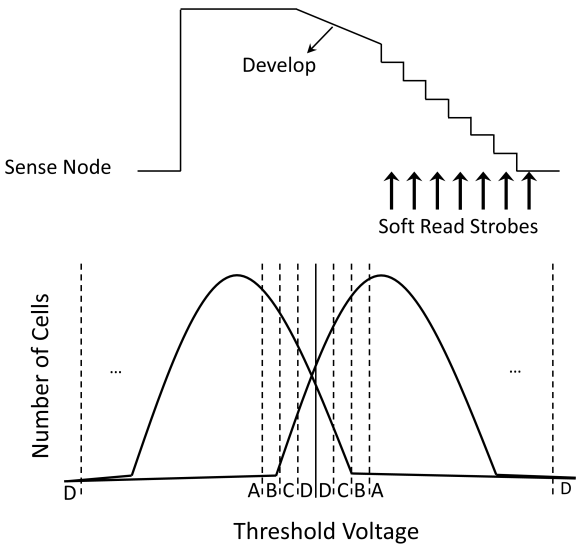


Figure 28.1.3: The FSBR algorithm uses boost modulation to group the bits into four buckets from highest (A) to lowest (D) confidence. Defective (open/short) BLs are placed in the lowest confidence bucket.

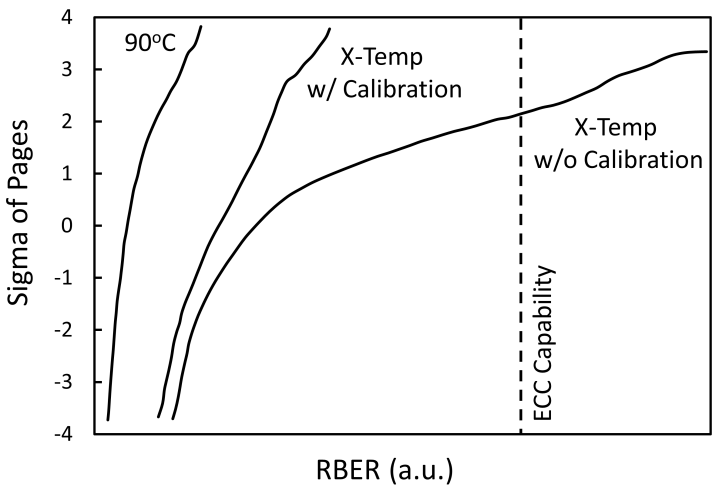


Figure 28.1.4: RBER distribution under cross-temperature condition, demonstrating the strength of the proposed fast-read calibration algorithm to lower the RBER well below the ECC correction capability.

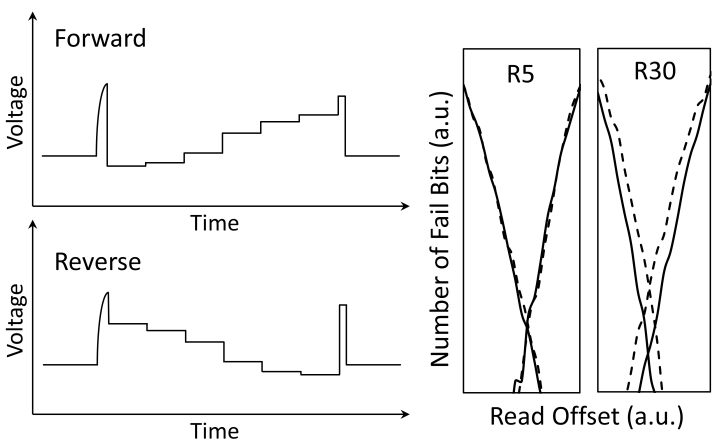


Figure 28.1.5: Schematic illustration of the forward and reverse read waveforms and representative experimental data, which demonstrates an improved read margin for higher read levels with reverse read (solid line: RR, dashed line: FR).

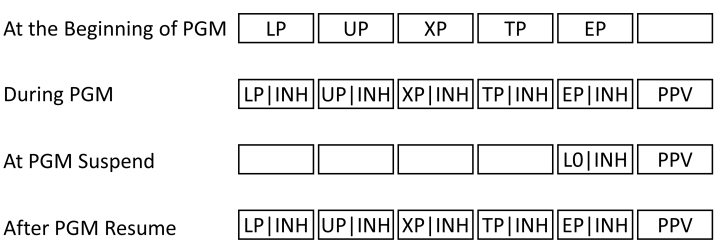
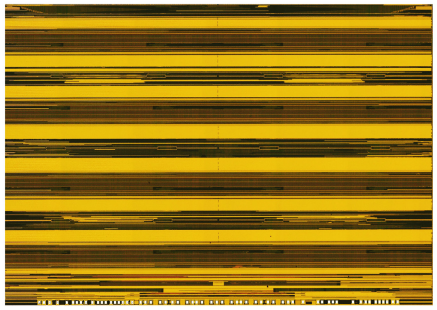


Figure 28.1.6: SPB latch allocation for user data, inhibit, and pre-program verify (PPV) information, to support FSBR during program suspend.



Number of Layers	192
Capacity	1.67 Tb
Number of Planes	4
Program Time (µs)	5500
Read Time (µs)	354
Endurance (P/E Cycle)	1K
Die Size (mm^2)	73.3
Bit Density (Gb/mm^2)	23.3
I/O Rate (MT/s)	1600

Figure 28.1.7: Die photograph and key metrics of the proposed work.

978-1-6654-9016-0/23/$31.00

28.2 A High-Performance 1Tb 3b/Cell 3D-NAND Flash with a 194MB/s Write Throughput on over 300 Layers

Byungryul Kim, Seungpil Lee, Beomseok Hah, Kangwoo Park, Yongsoon Park, Kangwook Jo, Yujong Noh, Hyeoncheon Seol, Hyunsoo Lee, Jaehyeon Shin, Seongjin Choi, Youngdon Jung, Sungho Ahn, Yonghun Park, Sujeong Oh, Myungsu Kim, Seonguk Kim, Hyunwook Park, Taeho Lee, Haeun Won, Minsung Kim, Cheulhee Koo, Yeonjoo Choi, Suyoung Choi, Sechun Park, Dongkyu Youn, Junyoun Lim, Wonsun Park, Hwang Hur, Kichang Kwean, Hongsok Choi, Woopyo Jeong, Sungyong Chung, Jungdal Choi, Seonyong Cha

SK hynix Semiconductor, Icheon, Korea

As data produced by multimedia explodes and demand for data storage increases, the most important topics for the NAND-Flash memory field are continuous performance improvements and cost/bit reduction. To improve performance, features to improve the quality of service (QoS) as well as the read/write performance [1] are required. To reduce the cost/bit, the number of stacked layers needs to increase, while the pitch between stacked layers decreases. It is necessary to manage the increasing WL resistance produced by a decreased stack pitch. To overcome these challenges, this paper presents techniques applied to a >300-layer 1Tb 3b/cell (TLC) 3D-NAND Flash memory: 1) A triple-verify program (TPGM) technique is used to improve program performance. 2) An adaptive unselected string pre-charge (AUSP) technique is used to reduce disturb and program time (t_{PROG}). 3) A programmed dummy string (PDS) technique is used to reduce WL settling time. 4) An all-pass rising (APR) technique is used to reduce the read time (t_R), 5) A plane-level read retry (PLRR) technique is used during erase to improve the QoS.

The TPGM scheme reduces t_{PROG} by narrowing the cell threshold voltage (V_{TH}) distribution. Increasing the step voltage (V_{STEP}) is one way to reduce program time, whereby an incremental step pulse programming method increases the step voltage (V_{STEP}) but makes the V_{TH} distribution wider. However, improving the V_{TH} distribution is essential to increasing the step voltage and reducing the program time. In a program operation, the threshold voltage difference (ΔV_{TH}) is determined by difference between the step voltage applied to WL and the channel voltage (V_{CH}). Figure 28.2.1 (a) and Fig. 28.2.1 (b) present the difference between the double-verify program (DPGM) and the TPGM scheme. The DPGM scheme [2] divides cells into three groups, according to the program verify (PV) levels and then controls the channel voltage of each group by applying three different BL voltages (V_{BL}). Appling V_{DD} to the group 1 (GR1) BLs to isolate the channels; the cells of GR1 are not programmed. V_A is applied to group 2 (GR2) BLs, and $\Delta V_{TH} = V_{STEP} - V_A$. 0V is applied to group 3 (GR3) BL and $\Delta V_{TH} = V_{STEP}$. In DPGM, the V_{TH} distribution can be improved by two kinds of ΔV_{TH}. Adding one more group ($\Delta V_{TH} = V_{STEP} - V_B$, $V_A > V_B$) to existing three groups in DPGM. TPGM categorizes cells into four groups according to their PV levels and drives the channel voltage of each group by applying four different BL voltages. Figure 28.2.1(c) illustrates the counter driving scheme that prevents BL coupling effect. BL1 is driven by the series connection of NMOS and is set to $V_{REF1} - V_{THN}$, while BL2 is initially set to V_{DD} and is discharged to $V_{REF2} + V_{THP}$ by the series connection of PMOS and NMOS. V_{THN} and V_{THP} represents the threshold voltages of the NMOS and the PMOS. BL1 rising is affected by BL2 falling, however the BL1 level does not exceed the target level due to inverse coupling. The counter driving scheme enhances BL settling and TPGM efficiency. By converting the V_{TH} distribution improvements into program time reduction results in approximately a 10% of program time reduction.

The AUSP scheme reduces t_{PROG} by tightening the cell's V_{TH} distribution. A program pulse is preceded by an unselected-string precharge (USP) [3] period to initialize all channels. USP prevents lack of channel boosting in a program pulse by precharging channels with V_{DD}, but a hot-carrier injection (HCI) disturbance occurs, as shown in Fig. 28.2.2(a). A voltage below V_{PASS} (V_{LOW}) is applied to all WLs, and the selected cell with a V_{TH} higher than V_{LOW} is turned off. The source-selection line (SSL) side channel is pre-charged to V_{DD} and the Drain Selection Lines (DSL) side channel is undriven. Due to the voltage difference between the SSL- and DSL-side channel, the HCI disturbance is produced by the high electric field. In the AUSP scheme, the SSL-side dummy WL is controlled by V_{DWL}, and $V_{DWL} - V_{TH(DummyCell)}$ is applied to the channel. HCI disturbances are reduced due to a lower electric field. Figure 28.2.2(b) illustrates the incremental channel initialization voltage that is proportional to the number of program loops. The channel initialization voltage corresponds to the SSL-side channel voltage; a higher channel initialization voltage is required for higher program loops. The channel initialization voltage can be lowered for lower program loops, thereby reducing HCI disturb further. As shown in Fig. 28.2.2(c), the cell's V_{TH} distribution becomes widen after programming, while programming with AUSP results in a narrower V_{TH} distribution, compared to a conventional USP. This reduced V_{TH} distribution contributes to around 2% t_{PROG} reduction.

The PDS scheme reduces t_R and t_{PROG} by programming dummy cells of the dummy strings. DSLs are divided by the DSL cut, as shown in Fig. 28.2.3(a), which separates each DSL; meanwhile, the dummy WLs, main WLs, and SSLs are connected to several strings in the 3D-NAND cell array. A dummy string produced by the DSL cut acts as capacitive load for the case of a rising/falling WL; hence, delaying WL settling time. Figure 28.2.3(b) and 28.2.3(c) present different channel conditions between an unprogrammed dummy string and a programmed dummy string. In an unprogrammed dummy string, all the cells are turned on, and the channel voltage becomes 0V via the source-line voltage (V_{SL}) when V_{PASS} is applied to all WLs. The non-floating channel acts as a capacitive load and affects the WL settling time. The PDS scheme programs the V_{TH} of dummy string's SSL-side dummy cell above V_{PASS} to turn off the dummy cell. As the SSL-side dummy cell is turned off, the floating channel no longer acts as capacitive load and the WL settling time is reduced.

The APR scheme reduces t_R by reducing the WL rise time. The different resistance and capacitance characteristics of each WL require different V_{PASS} sources to be connected to each WL group, and one source is selected by the switch circuits. As depicted in Fig. 28.2.4(a), in a conventional scheme one target V_{PASS} source is selected and applied to the dedicated WL during V_{PASS} rise time. As in shown in Fig 28.2.4(b), the APR scheme divides the V_{PASS} rise time into two parts, A and B. In part A, all V_{PASS} sources are connected to all WL to reduce the WL rise time. In part B, one target V_{PASS} source is applied to the dedicated WL so that it is same as the conventional V_{PASS} rising scheme. The APR scheme reduces t_R by around 2%.

As program/erase (P/E) cycles increase, the number of erroneous bits also increase; adjusting the read voltage bias can reduce the number erroneous bits. The read retry (RR) scheme with read level change is one effective method to overcome these situations. However, in a conventional RR the read level can only be changed when the read operation for all planes in the NAND device are completed. As a result, the read performance is determined by the last plane terminated. In this work, a PLRR scheme is used to alleviate read performance deterioration in the NAND controller. Figure 28.2.5 shows an example PLRR sequence: the read level is changed regardless of the operations occurring in other planes. Therefore, the read performance can be improved compared to the previous one since subsequent read commands can be issued immediately. In addition, the PLRR effect becomes greater when the number of planes increases.

In this work, five new techniques are introduced to achieve a high-performance 1-Tb 3bit/cell 3D-NAND Flash memory using a peripheral circuit under cell array architecture. The key comparison table, shown in Fig 28.2.6, reports a 20Gb/mm^2 bit density, which is achieved by using over 300-stacked WLs with an improved program throughput, t_R and bit density compared to prior work [4]. A die microphotograph of the fabricated TLC NAND chip is shown in Figure 28.2.7.

References:

[1] A. Grossi et al., "Quality-of-service implications of enhanced program algorithms for charge-trapping NAND in future solid-state drives," *IEEE Trans. Device Mater. Rel.*, vol. 15, no. 3, pp. 363-369, Sept. 2015.

[2] C. Miccoli et al., "Investigation of the programming accuracy of a double-verify ISPP algorithm for nanoscale NAND Flash memories," IEEE IRPS, pp. 5.1-5.6, 2011.

[3] R. Yamashita et al., "A 512Gb 3b/cell flash memory on 64-word-line-layer BiCS technology", *ISSCC*, pp. 196-197, 2017.

[4] M. Kim et al., "A 1Tb 3b/Cell 8th-Generation 3D-NAND Flash Memory with 164MB/s Write Throughput and a 2.4Gb/s Interface," *ISSCC*, pp. 136-137, 2022.

 978-1-6654-9016-0/23/$31.00

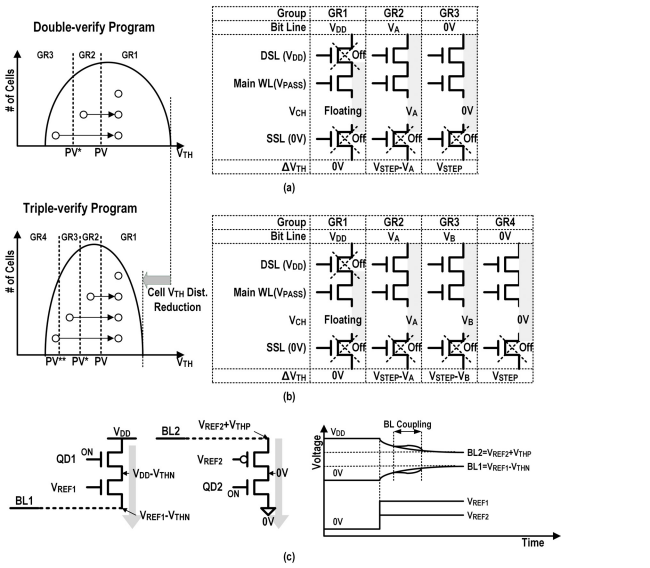


Figure 28.2.1: Comparison of (a) the DPGM scheme, (b) the TPGM scheme, and (c) a schematic and a timing diagram for the counter driving scheme.

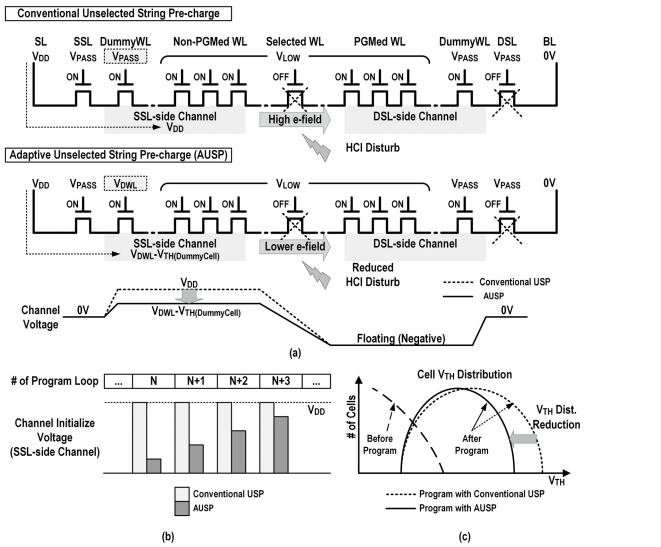


Figure 28.2.2: (a) HCI disturb comparison between a conventional USP scheme and the AUSP scheme. (b) AUSP incremental channel voltage and (c) the cell resulting V_{TH} distribution.

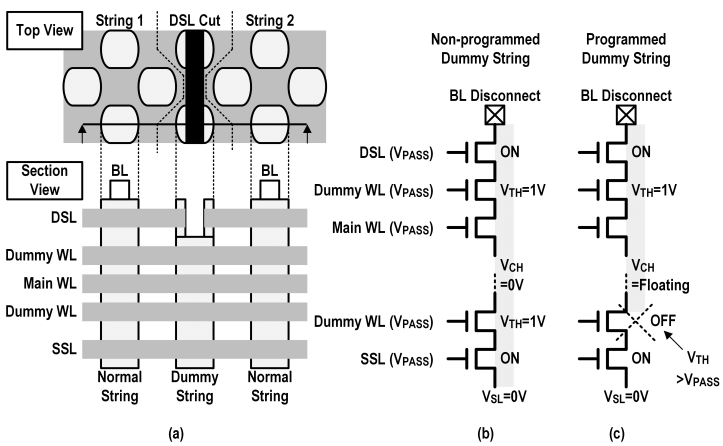


Figure 28.2.3: (a) DSL cut diagram. (b) A dummy string with an unprogrammed dummy WL cell and (c) a dummy string with a programmed dummy WL cell.

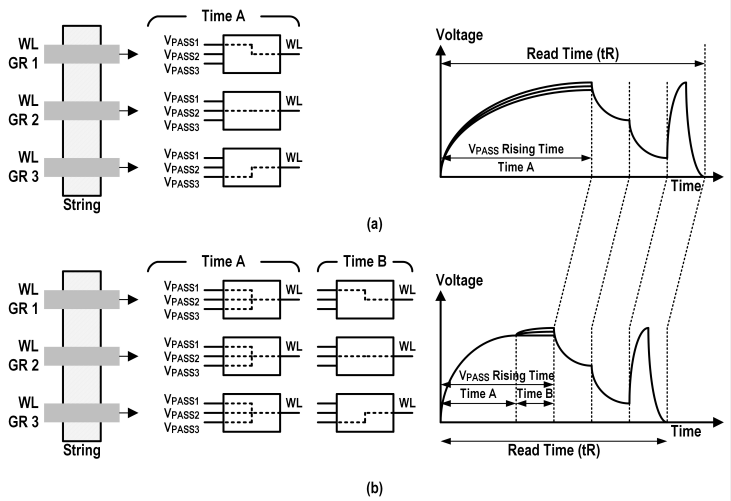


Figure 28.2.4: (a) A timing diagram for the conventional WL rise and (b) a timing diagram for APR scheme.

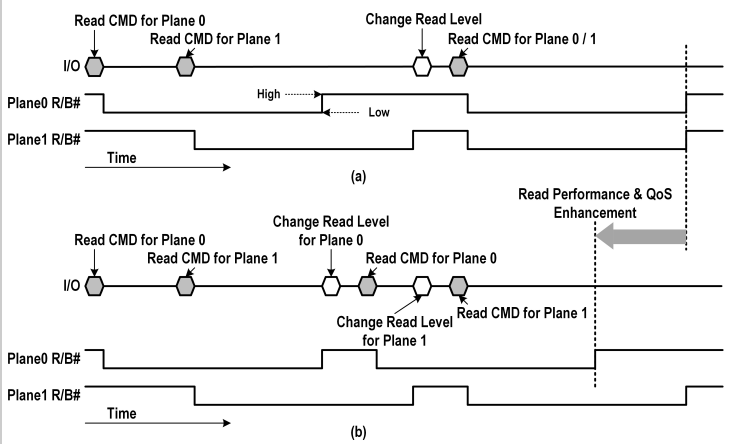


Figure 28.2.5: Timing diagrams for operations (a) without and (b) with PLRR scheme.

	ISSCC 2022 [4]	This Work
# Bit/Cell	3	3
Capacity (Gb)	1024	1024
# of Planes	4	4
Page Size (KB/Page)	16	16
Program Throughput (MB/s)	164	194
16KB tR (us)	45	34
IO Speed (Gbps)	2.4	2.4
Vccq (V)	1.2	1.2
Bit Density (Gb/mm2)	11.55	>20

Figure 28.2.6: Key comparison table.

28

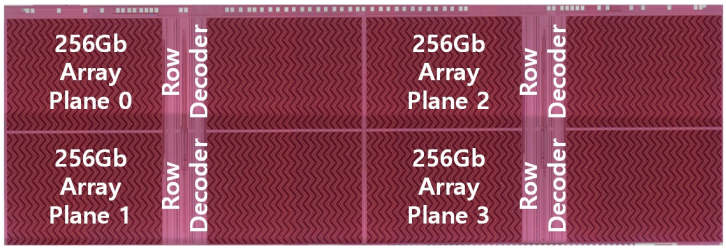


Figure 28.2.7: A die microphotograph.

978-1-6654-9016-0/23/$31.00

28.3 A 4nm 16Gb/s/pin Single-Ended PAM4 Parallel Transceiver with Switching-Jitter Compensation and Transmitter Optimization

Jahoon Jin, Soo-Min Lee, Kyunghwan Min, Sodam Ju, Jihoon Lim, Hyunsu Chae, Kwonwoo Kang, Yunji Hong, Yeongcheol Jeong, Sang-Ho Kim, Jongwoo Lee, Joonsuk Kim

Samsung Electronics, Hwaseong, Korea

Ever-growing applications, such as 5G communication, deep learning, advanced driver-assistance systems (ADAS), and extended reality (XR), have fueled demand for increased computing power and per-pin interface bandwidth. Recently, four-level pulse-amplitude modulation (PAM4) has been adopted as a solution [1-3]: the throughput is doubled without increasing the baud (Nyquist) rate. Compared to a conventional non-return-to-zero (NRZ) signaling, PAM4 requires more design effort: varying from the precise design of I/O circuits to the off-chip characterization. This is in part due to SNR degradation and an increased switching jitter (SWJ). For a 1st-order low-pass filter with a Nyquist-frequency cutoff, SWJ is 35% for the middle eye and 51.2% for the top and bottom eyes [4]. Maximum-transition-avoidance (MTA) encoding [3] can be used to reduce SWJ, but at the cost of additional encoder/decoder hardware and an auxiliary channel to compensate for data loss.

This paper presents a 16-Gb/s/pin 0.764-pJ/b single-ended PAM4, NRZ compatible, parallel transceiver for short-reach and low-power applications. An SWJ compensation (SWJC) technique is proposed for the receiver (RX) to improve timing margins by adjusting transitions of the thermometer-coded signals. The transmitter (TX) performs 1-tap fractionally spaced feedforward equalization (FS-FFE) [4] with customized tap spacing and capacitive-peaking equalization (C-peaking) to further extend bandwidth. To improve SNR by more than 3dB, we use relaxed impedance matching [5] and optimize the termination values (R_{TX} = 20Ω, R_{RX} = 50Ω). Asynchronous I/O and a synthesized serializer/deserializer design [6] are used to reduce power consumption and area.

Figure 28.3.1 shows a diagram of the proposed SWJC for a PAM4 RX. In a conventional RX, three comparators convert a PAM4 signal into thermometer-coded signals (DH, DM, and DL), and the thermometer-to-binary (T2B) decoder converts them into a 2b output (MSB and LSB). In the proposed RX, the SWJ of the thermometer signals are compensated by the proposed SWJC circuit before being converted into 2b output. Since PAM4 eyes are asymmetrical in shape, the thermometer signals have different SWJ. On the other hand, the MTA [3] reduces SWJ by removing the maximum transitions between the lowest and highest levels, which contribute the most to SWJ. However, pin efficiency is reduced as two of 16 transitions are not used. The proposed SWJC technique reduces SWJ while maintaining the same pin efficiency. Considering the DH and its SWJ (SWJ_H), the falling edges come earlier than the rising edges, and vice-versa DL and it's SWJ (SWJ_L). SWJC adjusts the transitions of the thermometer signals so that the falling edges and rising edges are aligned. In the case of DH, the falling edges would be delayed until the leftmost falling edge crosses the leftmost rising edge. Hence, the SWJC circuit generates jitter reduced thermometer signals: DH1, DM1, and DL1.

The overall architecture of the proposed parallel-interface chip consists of a synthesized digital PHY, a phase-locked loop (PLL), and dual-mode PAM4/NRZ I/O circuits, as shown in Fig. 28.3.2. A source-synchronous clock signal (DQS) is used to sample the data. In PAM4 mode the Gray-coded 2b data (*MSB and *LSB) are obtained from the serializer, which is implemented in the synthesized PHY, and provided to a TX I/O circuit (TX DQ). TX DQ generates and transmits PAM4 signals. The RX DQ on the receiver side converts the PAM4 signal into the transmitted binary data. In NRZ mode only the *MSB is transmitted. Figure 28.3.2(top-right) shows the TX DQ block diagram. Three identical thermometer output stages are used to obtain a high ratio of level mismatch (RLM). The binary-to-thermometer converter (B2T) converts the Gray-coded binary signals to three thermometer-coded signals, DH, DM, and DL, and passes them to each output stage. Each output stage includes a segmented source-series terminated driver terminated to 60Ω. The TX performs a 1-tap FS-FFE using a tap spacing (Δt) of 0.8UI, and a 3b controllable C-peaking driver is used. Figure 28.3.2(center-right) shows how the thermometer driver generates the four levels while maintaining the same impedance. In the case of V_{11}, the PAD-to-supply (Z_{UP}) and PAD-to-ground (Z_{DN}) impedances are 20 and 50Ω; the resulting voltage is $0.714 \cdot V_{DDQ}$. Note that a 50-Ω RX termination is Z_{DN}, in this case. Figure 28.3.2(bottom) shows the B2T circuit details. The small tails are added in series to match the pull-down and pull-up strength, allowing for jitter improvement as shown by the simulation results.

The RX DQ block consists of three data paths and a Gray decoder, as shown in Fig. 28.3.3(top left). The data path is comprised of a single-to-differential converter (S2D), a CML-CMOS, and an SWJC circuit. The RX DQ first generates three differential thermometer signals based on the three reference voltages: VH, VM, and VL. Full-swing conversion and SWJ compensation are achieved via the CML-CMOS and SWJC circuit, respectively. Then, the Gray decoder converts the thermometer signals into 2b data (MSB and LSB). Here, an inverted DH (DHb) is used for two reasons: 1) the Gray decoder is constructed from two AND gates, minimizing the required area and MSB-to-LSB skew; 2) it is possible to minimize the mismatch between the DH1b and DL1 that are the outputs of the SWJC circuits. It is because DHb and DL have the same characteristics in that the rising edges lead the falling edges; hence, the corresponding SWJC circuits are adjusted in the same manner. DHb can be generated by crossing the differential signals between the top S2D and the CML-CMOS blocks, resulting in no additional delay. In the SWJC circuit, the strength of each pull-up/down path is controlled by separate thermometer controls to control the transition time. Essentially, when the code goes up, the propagation delay of the rising edge (t_{RISE}) decreases while that of falling edge (t_{FALL}) increase, as shown by the simulation results. The SWJC circuit has a resolution of 4.5ps and a dynamic range of 67.6ps. The schematic diagram shows an example case where t_{RISE} is greater than t_{FALL}. Once the input offset for the three S2Ds are cancelled, the reference voltages are optimally set. Then, the SWJ compensation begins. Figure 28.3.3(bottom-right) shows an SWJC example. Starting from the default code of 7, the control code is decreased while monitoring the eye width. Once SWJ reaches its minimum value, compensation is done.

To improve the eye opening further, a relaxed impedance matching scheme [5] is adopted. Figure 28.3.4(top) shows the behavioral simulation results for the various termination sets of R_{TX} and R_{RX} at 16Gb/s. R_{TX} and R_{RX} are swept from 10 to 90Ω in increment of 5Ω. The extracted channel includes all interconnect: ranging from the traces on the redistribution layer (RDL), the package, and the board. The insertion loss is -4.65dB at 4GHz. Note that a tap spacing of 0.8UI is used, and the FS-FFE's coefficient (A_{EQ}) is optimized for each case. The simulation results are summarized as a normalized contour map of eye area. Considering power and area, the chosen optimal set for this work is R_{TX} = 20Ω and R_{RX} = 50Ω. Figure 28.3.4(bottom) shows FS-FFE behavioral simulation results across various tap spacing, from 0.4 – 1.0UI, for three different termination sets; A_{EQ} is optimized for each case. Although, as reported in [4], a termination of R_{TX} = 50Ω and R_{RX} = 50Ω shows that a tap spacing of 0.6UI is the optimum, however the optimal value is different for other terminations sets. When applying a tap spacing of 0.8UI for a termination of R_{TX} = 20Ω and R_{RX} = 50Ω, the eye opening is improved by 2.25× compared to the maximum eye opening for R_{TX} = 50Ω and R_{RX} = 50Ω. The prototype IC is fabricated in a 4nm FinFET process with the following configurations: four data lanes (TX DQ + RX DQ), one strobe lane (TX DQS + RX DQS), a VREF DAC, and a synthesized PHY for test. Note that the length of every PCB trace is matched to 60mm, and each DQ of the prototype IC is verified at a data rate of 16Gb/s/pin. An overall data rate of the prototype is 64Gb/s. The measured eye diagrams shown in Fig. 28.3.5 (top) highlight the effect of FS-FFE and C-peaking on bandwidth extension: the eye width is improved by 2×, from 31.46 to 62.93ps, equivalent to 0.25 to 0.50UI. Figure 28.3.5 (bottom) shows the measured RX eye Shmoo plots and bathtub curves. Using optimized TX settings SWJC improves the eye opening of the LSB from 0.31 to 0.37UI at 10^{-12} BER. The eye opening of the MSB is 0.34UI at 10^{-12} BER.

Figure 28.3.6 summarizes the performance metrics of the proposed transceiver and compares it to recent single-ended PAM4 transceivers. The power efficiency of the unit TX DQ and RX DQ is 0.436 and 0.328pJ/b. A power breakdown of the 64-Gb/s prototype IC is also presented. Figure 28.3.7 shows the die photograph. The total area of the four-DQ configurations is 0.222mm^2, and the area of a unit TX and RX DQ is 0.00363mm^2 each.

References:

[1] T. M. Hollis et al., "25.3 An 8Gb GDDR6X DRAM Achieving 22Gb/s/pin with Single-Ended PAM4 Signaling," *ISSCC*, pp. 348-349, 2021.

[2] H. Jin et al., "A 24Gb/s/pin PAM-4 Built Out Tester chip enabling PAM-4 chips test with NRZ interface ATE," *IEEE A-SSCC*, pp. 1-3, 2021.

[3] H. N. Rie et al., "A 40-Gb/s/pin Low-Voltage POD Single-Ended PAM-4 Transceiver with Timing Calibrated Reset-less Slicer and Bidirectional T-Coil for GDDR7 Application," *IEEE Symp. VLSI Circuits*, pp. 148-149, 2022.

[4] X. Zheng et al., "A 50–112-Gb/s PAM-4 Transmitter with a Fractional-Spaced FFE in 65-nm CMOS," *IEEE JSSC*, vol. 55, no. 7, pp. 1864-1876, July, 2020.

[5] M. Choi et al., "An FFE Transmitter Which Automatically and Adaptively Relaxes Impedance Matching," *IEEE JSSC*, vol. 53, no. 6, pp. 1780-1792, June, 2018.

[6] S.-M. Lee et al., "A 0.6V 4.266Gb/s/pin LPDDR4X Interface with Auto-DQS Cleaning and Write-VWM Training for Memory Controller," *ISSCC*, pp. 398-399, 2017.

 978-1-6654-9016-0/23/$31.00

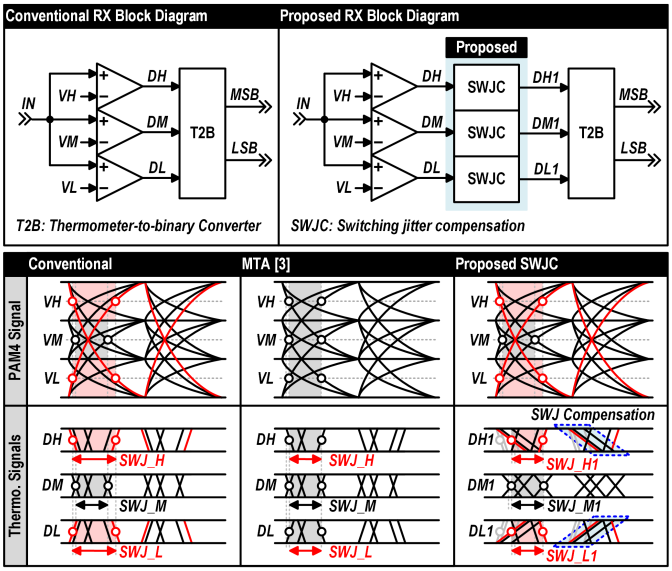


Figure 28.3.1: Simplified RX block diagrams (top) and a conceptual diagram for the switching-jitter compensation technique (bottom).

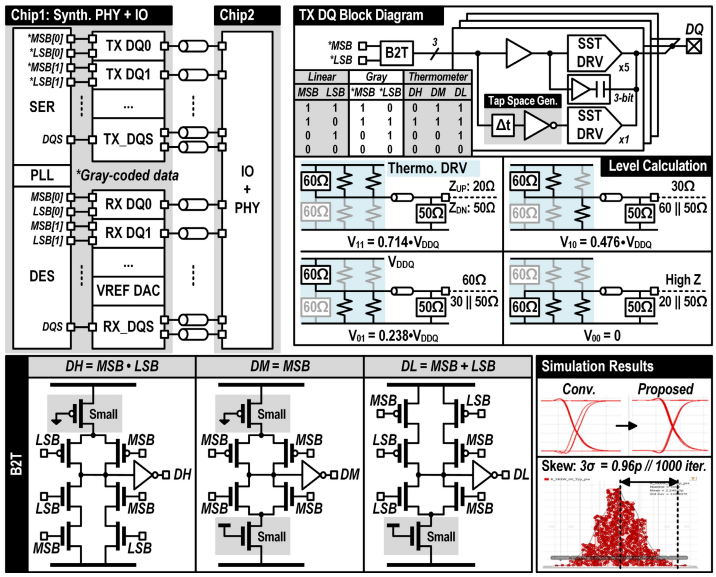


Figure 28.3.2: An overall architecture (top left), a TX block diagram (top right), and B2T circuit details (bottom).

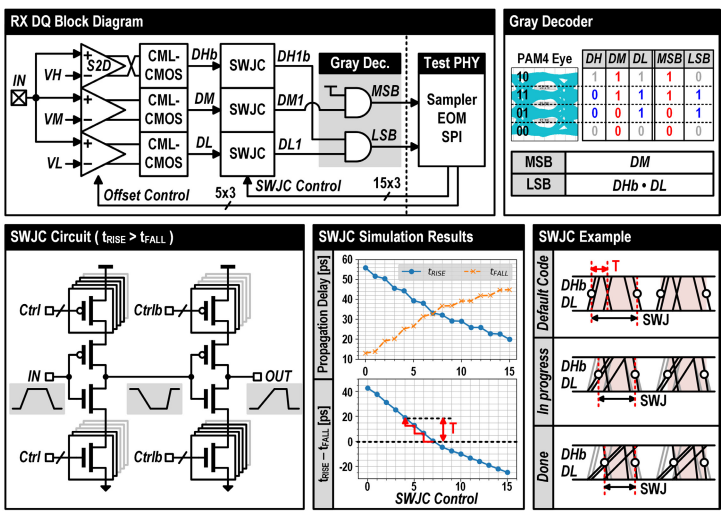


Figure 28.3.3: An RX block diagram, an SWJC circuit, SWJC simulation results, and an SWJC example.

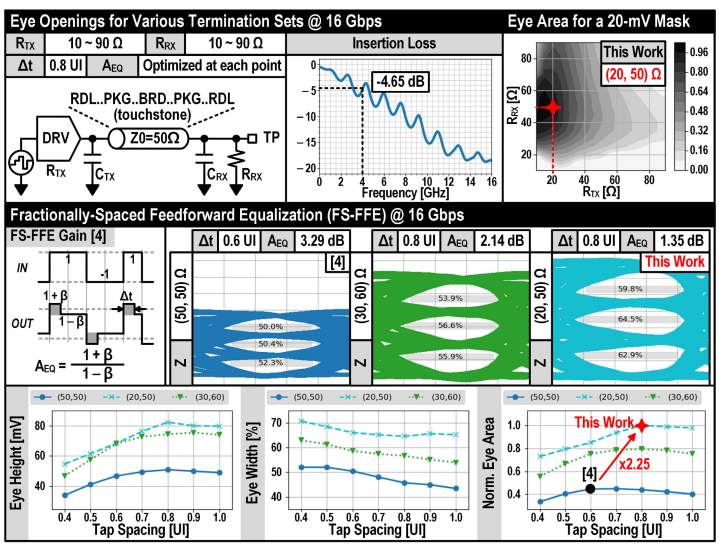
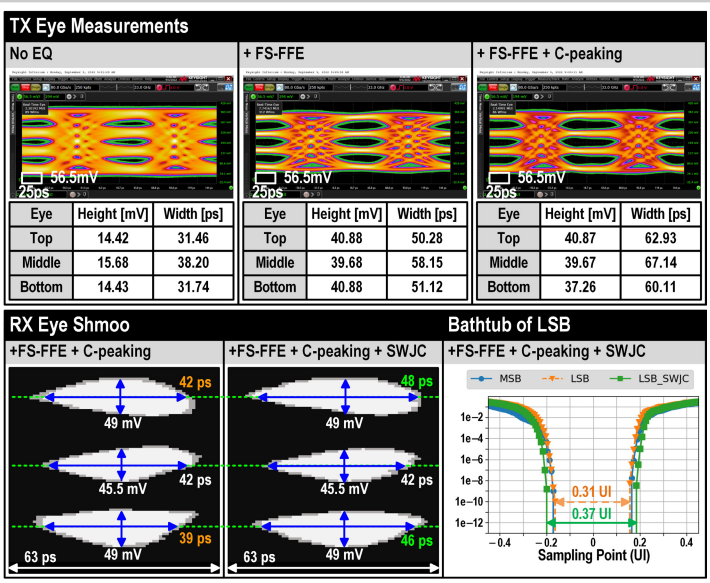


Figure 28.3.4: Behavioral verification: the optimum termination set for a tap spacing of 0.8UI and the optimum tap-spacing for various termination sets.



Eye	Height [mV]	Width [ps]
Top	14.42	31.46
Middle	15.68	38.20
Bottom	14.43	31.74

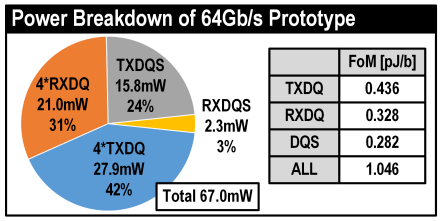
Eye	Height [mV]	Width [ps]
Top	40.88	50.28
Middle	39.68	58.15
Bottom	40.88	51.12

Eye	Height [mV]	Width [ps]
Top	40.87	62.93
Middle	39.67	67.14
Bottom	37.26	60.11

Figure 28.3.5: TX eye measurement for various equalization settings (top), and RX eye Shmoo plots and the resulting bathtub curves at 16Gb/s/pin (bottom).

	This Work		ISSCC'21 25.3 [1]	ASSCC'21 17.2 [2]	VLSI'22 [3]	CICC'22 J. Kim
Process	4nm FinFET		1Ynm CMOS	28nm CMOS	28nm CMOS	28nm CMOS
Supply	0.75V / 0.6V		1.35V	1V	1.2V / 0.95V	1.2V / 1.2V
Data Rate	8G/16G		11G/22G	12G/24G	40G	60G
Modulation	NRZ/PAM4		NRZ/PAM4	NRZ/PAM4	PAM4	PAM4
Termination	TX20 - RX50		TX40 - RX40	N/A	N/A	1:1
Equalization	1T-FS-FFE C-peaking		1T-FFE CTLE	1T-DFE	T-Coil, 2T-FFE CTLE, 4T-DFE	2T-FFE
SWJ Reduction	SWJC	X	X	X	MTA	LPMTA
Eye Opening (BER)	0.37 UI (1e-12)	0.31 UI (1e-12)	0.23 UI(C) (N/A)	0.3 UI (1e-11)	0.2 UI (1e-6)	0.36 UI(D) (N/A)
FoM [pJ/b]	0.764(A), 1.046(B)		N/A	N/A	2.02	1.67(D)

(A) DQ I/O only (B) 64Gb/s prototype IC including DQS I/Os (D) TX only
(C) Estimated from the write shmoo plot @ 24 Gb/s/pin



	FoM [pJ/b]
TXDQ	0.436
RXDQ	0.328
DQS	0.282
ALL	1.046

Figure 28.3.6: Performance summary and power breakdown.

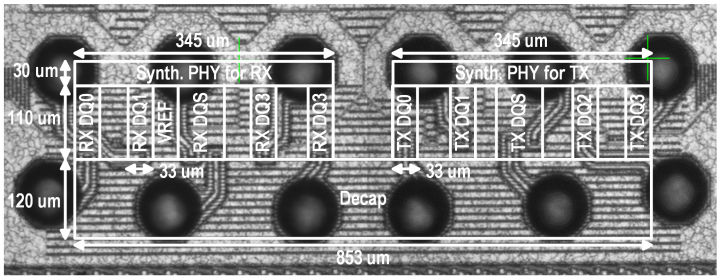


Figure 28.3.7: Die Photo.

978-1-6654-9016-0/23/$31.00

28.4 A 4nm 1.15TB/s HBM3 Interface with Resistor-Tuned Offset-Calibration and In-Situ Margin-Detection

Kwanyeob Chae, Jiyeon Park, Jaegeun Song, Billy Koo, Jihun Oh, Shinyoung Yi, Won Lee, Dongha Kim, Taekyung Yeo, Kyeongkeun Kang, Sangsoo Park, Eunsu Kim, Sukhyun Jung, Sanghune Park, Sungcheol Park, Mijung Noh, Hyogyuem Rhew, Jongshin Shin

Samsung Electronics, Hwaseung, Korea

A critical performance bottleneck for memory-bound applications such as high-performance computing (HPC), artificial intelligence (AI), and machine learning (ML) applications is the limited memory bandwidth [1]. An HBM3 DRAM interface, based on a WDQS-clocking scheme with high-speed low-voltage-swing terminated logic (LVSTL) I/O, is a promising energy-efficient high-bandwidth solution. The WDQS-clocking scheme is expected to improve the read-valid-window margin (VWM), which is a major limiting factor in achieving DRAM access reliability at high speed. However, the long turn-around read path limits the read VWM improvement. In addition, it is essential to minimize interface area for clustered channel implementations by considering controllers, processing units and bus interconnects. Memory-access latency can be minimized by the close placement of functional units to the HBM3 interface.

To address these technical challenges, we introduce a digital HBM3 interface that is assisted by read VWM improvement techniques: an offset-calibration and an in-situ read margin-detection scheme. Furthermore, a low-power delay sensor, which enables accurate tracking and compensation of delay variation, is proposed to maintain DQS-to-DQ centering for DRAM access reliability under environmental variations. A slim bit-slice architecture, with stacked I/Os, enables a scalable and compact structure that improves the signal integrity by minimizing the interposer channel length.

Figure 28.4.1 shows a block diagram of the proposed digital bit-slice-based HBM3 interface. The bit-slice is the minimum macro unit for transmitting and receiving a data bit; it is placed near the corresponding I/O thereby minimizing the duty and jitter performance degradation between digital bit-slices and the I/O block. The bit-slice includes digital delay lines, which are used for reliable memory accesses. The delay lines are controlled to maintain DQS-to-DQ centering under environmental variations in conjunction with a digital delay sensor that tracks on-chip delay changes. The standard-cell-based digital bit-slice achieves high-speed and low-voltage operation while occupying a small area. This digital interface architecture achieves a 9.0Gb/s/pin operation from a 660mV supply.

Figure 28.4.2 shows the physical structure of the HBM3 interface. I/O arrays are stacked in two rows to accommodate a wide bit-width in a limited area. To interface the stacked I/Os, a slim bit-slice is designed to have ports on one side and routing channels on the other side. This stacking pattern, with non-flipped first row I/Os and flipped second row I/Os, provides a seamless interface between the digital slim bit-slices and the stacked I/Os. The routing channel in the non-flipped I/Os includes empty metal routing tracks to allow for connections to ports in the flipped I/O. As a result, the odd numbered slim bit-slices interface with the ports of the second row, and the even numbered slim bit-slices interface with the ports of the first row I/O. DQ I/Os are symmetrically placed on top and bottom, like a sandwich, for a 32b data PHY configuration. Using this structure, pins for the controller interface can be placed in the shared controller logic area, which improves the logic speed. Moreover, the divided I/O stacks relaxes signal routing congestion between I/Os and μ-bumps.

For the HBM3 interface, receiver input mismatches severely degrade the read VWM, since per-bit V_{REF} generation is not feasible, due to area and power limitations. Figure 28.4.3 shows the proposed resistor-tuned offset-calibration method in combination with a V_{REF} generator. A single V_{REF} covering a wider bit-width is more area efficient, but results in larger input-offsets. In this work, a coarse V_{REF} generator with an offset-calibration scheme is proposed to simultaneously minimize the size of V_{REF} generator and offset mismatch. The coarse V_{REF} step resolution reduces the number of multiplexers required by half, which contributes to a 20% V_{REF} circuit area reduction. However, this coarse V_{REF} step resolution degrades the read VWM due to an increased quantization error. To compensate for this degradation, an offset-calibration circuit is used with the coarse V_{REF} generator. The per-bit offset tuning calibrates out the remaining offset after the coarse V_{REF} control; compensating for both the mismatch and the V_{REF} quantization error, while at the same time minimizing the area consumption. In this work, stacked-transistor-based resistor tuning is used to tune the output resistance, while consuming minimal area; effectively calibrating the input mismatch and providing a per-bit suitable V_{REF}. The tunable transistor-array is added to the secondary resistor to control the output impedance with fine resolution. Since the primary resistor is connected to the output node, the high-frequency degradation, due to an increased parasitic capacitance, is insignificant compared to the conventional input-transistor-based offset tuning [6]. As a result, the proposed solution achieves optimized per-bit offset within compact area. In addition, a feed-forward equalization scheme, based on an AC-coupling capacitor, is used in this work to boost the write performance; thereby, leading to a 9.0Gb/s/pin operation from a 300mV supply (VDDQ). The measured results of the AC boost are shown in Fig. 28.4.3. The internal interposer signal channel is connected to an external ball for board-level testing.

The HBM3 interface includes an all-digital delay sensor to detect on-chip delay changes to maintain access stability under dynamic delay fluctuation. The proposed delay sensor includes a pre-processing logic block, as shown in Fig. 28.4.4. The pre-processing logic generates Di and Ds, which are generated from multiple clock edges. These are used for delay measurement; thus, the phase-detection and delay error caused by process variation is minimized. In addition, the control logic can be simplified since pre-processed control signals are used; the output of the phase detector, 1 or 0, leads to a delay increase or decrease. The DQ delay for each slim bit-slice is calculated based on the primary selection value, from the all-digital delay sensor, to compensate for delay changes. As a result, DQS-to-DQ centering is accurately maintained under environmental variations.

External environmental variations significantly impact memory access reliability. To maintain access reliability, periodic training is essential to compensate for any changes in the external delays. A periodic read training results in a 0.6μs access black-out, per our estimate, once initiated. A periodic memory-access pause causes more excessive performance degradation than required, considering the infrequency of delay changes. This work proposes in-situ margin-detection to measure the real-time read VWM for the data traffic. This scheme is used to initiate the read training process only when required. Figure 28.4.5 shows the in-situ margin-detection block; including a digital DQ-de-skewing delay line with multiple output ports, each with different DQ delay. The center DQ is the data sample used for data read-out, the other DQs are used to measure the left and the right distances from the center, which indicate the read VWM. If DQS-to-DQ centering shifts, then the left or the right edge detection signal indicating a centering shift is set. Only when an edge is detected is the read training initiated to compensate for a delay change. Using the proposed margin-detection, on live data traffic, allows unnecessary periodic read training to be eliminated.

Experimental results, shown in Fig. 28.4.6, illustrate the wide operating range of the proposed HBM3 interface. The maximum operating frequency of DRAM cell accesses is 8.0Gb/s/pin with an eye with of 0.39UI, which is limited by the DRAM bit rate. The HBM3 interface achieves multiple-input shift-register (MISR) access at 9.0Gb/s/pin from a 660mV supply. The proposed offset-calibration improves the read VWM by 16.7%. In addition, the proposed in-situ read margin-detection minimizes the performance penalty, by eliminating unnecessary periodic training. The proposed HBM3 interface also shows an impressive supply noise tolerance: a ±220mV supply voltage change during memory accesses.

The chip micrograph and the assembled 2.5D package are shown in Fig. 28.4.7. The proposed slim bit-slice architecture in conjunction with a stacked I/O structure achieves a reliable high bandwidth, up to 1.15TB/s, while consuming only 0.0246mm^2/bit. The area- and energy-efficiency comparison to prior work is summarized in Fig. 28.4.7.

References:

[1] M.-J. Park et al., "A 192-Gb 12-High 896-GB/s HBM3 DRAM with a TSV Auto-Calibration Scheme and Machine-Learning-Based Layout Optimization," *ISSCC*, pp. 444-445, 2022.

[2] K.-Y. Chae et al., "An 8nm All-Digital 7.3Gb/s/pin LPDDR5 PHY with an Approximate Delay Compensation Scheme," *IEEE Symp. VLSI Circuits,* pp. 96-97, 2019.

[3] S.-M. Lee et al., "An 8nm 18Gb/s/pin GDDR6 PHY with TX Bandwidth Extension and RX Training Technique," *ISSCC*, pp. 338-339, 2020.

[4] S.-Y. Hwang et al., "A 3.2 Gbps/pin HBM2E PHY with Low Power I/O and Enhanced Training Scheme for 2.5D System-in-Package Solution," *IEEE Hot Chips Symp.*, pp.1-13, 2020.

[5] H.-G. Ko et al., "A 370-fJ/b, 0.0056 mm^2/DQ, 4.8-Gb/s DQ Receiver for HBM3 with a Baud-Rate Self-Tracking Loop," *IEEE Symp. VLSI Circuits*, pp. 94-95, 2019.

[6] K.-H. 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. 344-345, 2021.

 978-1-6654-9016-0/23/$31.00

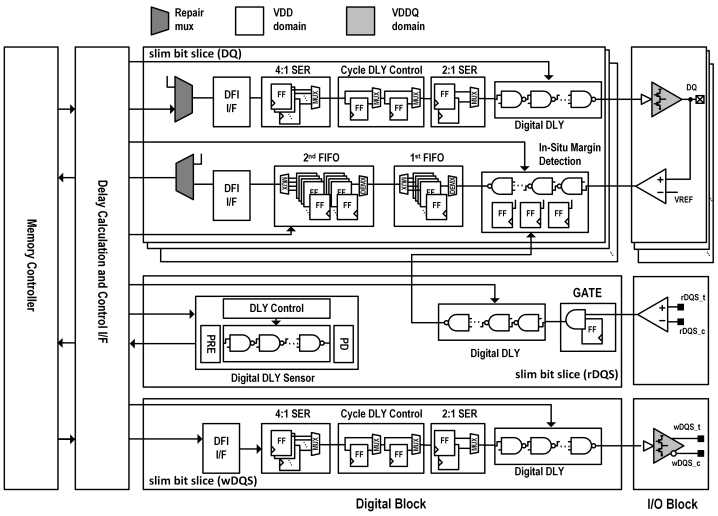


Figure 28.4.1: Block diagram of the proposed HBM3 interface.

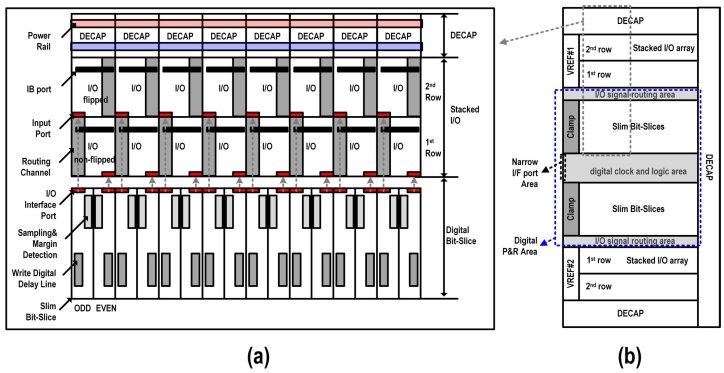


Figure 28.4.2: (a) The physical structure of the stacked I/O with slim bit-slice for compact configuration and (b) the sandwich structure for 32b data PHY unit.

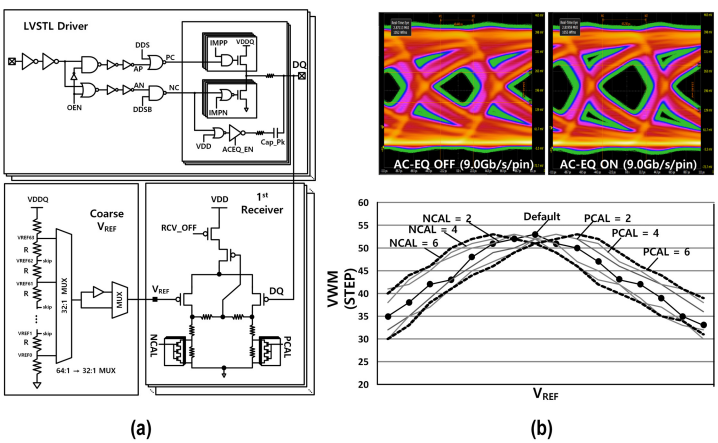


Figure 28.4.3: (a) A 0.3V LVSTL driver and a receiver with resistor-tuned offset-calibration combined with a coarse V_{REF} generator. (b) Measured AC-EQ effect and offset-tuning results.

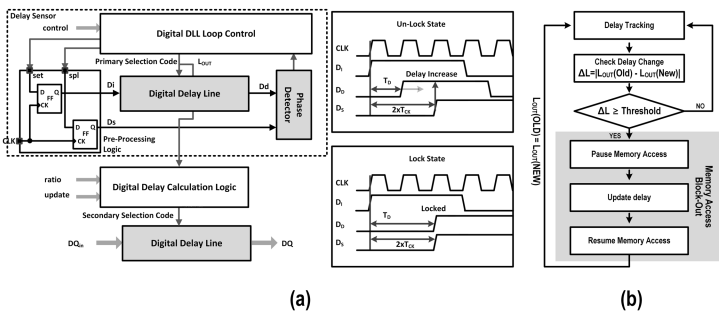


Figure 28.4.4: (a) The proposed digital-delay sensor. (b) Proposed delay update procedure.

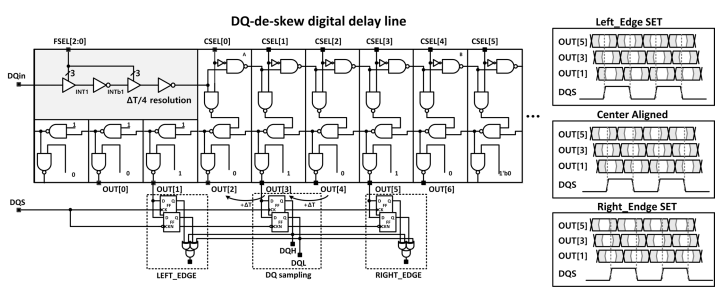


Figure 28.4.5: In-situ read margin-detection block.

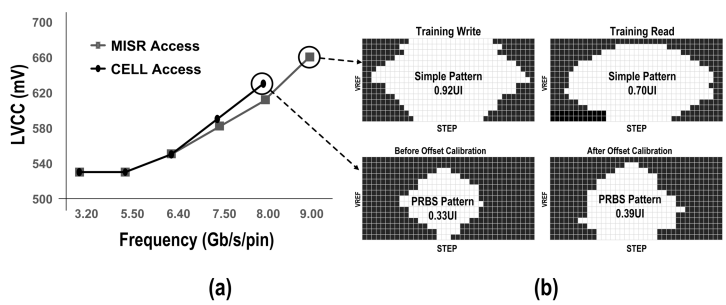


Figure 28.4.6: (a) Measured operating frequency and voltage. (b) Cell read VWM with offset-calibration at 8.0Gb/s/pin and MISR write/read VWM at 9.0Gb/s/pin.

Reference	[2]	[3]	[4]	[5]	This Work
Interface	LPDDR5	GDDR6	HBM2E	HBM3 RCV.	HBM3
Technology	8nm	8nm	7nm	65nm	4nm
Speed	7.3Gb/s/pin	18Gb/s/pin	3.2Gb/s/pin	4.8Gb/s/pin	9.0Gb/s/pin
B/W	14.6GB/s	36GB/s	409.6GB/s	N/A	1.15TB/s
VDD/VDDQ	0.79V/0.5V	0.85V/1.35V	0.75V/1.2V	1.1V	0.66V/0.3V
Area Per-Bit (mm²/bit)	0.0246	0.1038	0.0056	0.0056	0.0046
Energy Efficiency (pJ/bit)	1.17	N/A	1.07	0.37	0.29

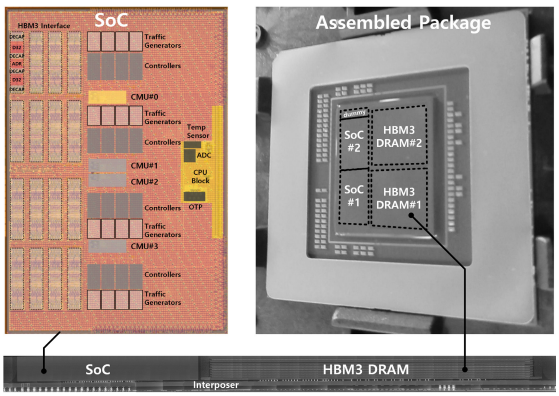


Figure 28.4.7: Chip micrograph and comparison table to prior work.

978-1-6654-9016-0/23/$31.00

28.5 A 900μW, 1-4GHz Input-Jitter-Filtering Digital-PLL-Based 25%-Duty-Cycle Quadrature-Clock Generator for Ultra-Low-Power Clock Distribution in High-Speed DRAM Interfaces

Yuhwan Shin*, Yongwoo Jo*, Juyeop Kim, Junseok Lee, Jongwha Kim, Jaehyouk Choi

Korea Advanced Institute of Science and Technology, Daejeon, Korea
*Equally Credited Authors (ECAs)

To secure sufficient timing margins, despite ever-increasing data rates, advanced DRAM interfaces use quadrature clocks, which run internally at a quarter-rate frequency, f_{QCLK}. The top of Fig. 28.5.1 shows a conventional clock-distribution scheme: a DLL in the middle of the peripheral distributes the quadrature clocks, $S_{IN,x}$, where $x = \{I, Q, IB, QB\}$, to all TXs that may be more than a few millimeters away. The fundamental problem with this conventional scheme is that it requires excessive power to distribute four (or at least two) high-frequency clocks over long distances across the chip. With f_{QCLK} = 2GHz, the clock distribution power consumption reaches tens of milliwatts and will continue to increase proportionally to f_{QCLK} for future standards. Moreover, the quadrature relationship between $S_{IN,x}$ is degraded as well, thus a quadrature-error corrector (QEC) can be used before each TX. Polyphase-filter- or analog-DLL-based QECs have issues with accuracy and area. Recently, digital DLL-based QECs [1-4] overcome these problems, but still have many disadvantages. (1) DLL-based ones cannot filter input jitter: the RMS jitter of the quadrature clocks at each TX should be less than 6mUI, hence the jitter-filtering capability becomes more important as f_{QCLK} increases. (2) They consume large power when operating at f_{QCLK}: the QEC in [1] consumes 8mW at f_{QCLK} of 2.3GHz. (3) They use DTCs to detect quadrature errors, the range of f_{QCLK} is limited to that of the DTC. (4) Their outputs, $S_{OUT,x}$, have a 50% duty-cycle (*DC*), an additional 25%-*DC* converter is required before the serializer at each TX.

This work presents an ultra-low-power clock-distribution scheme with a digital PLL (DPLL)-based quadrature clock generator (QCG) for DRAM interfaces. As shown at the bottom of Fig. 28.5.1, the proposed clock-distribution scheme sends only a single-phase clock, S_{IN}, at a much lower frequency (f_{QCLK}/8) to the TX across the chip; reducing power consumption by 32×. The proposed scheme also provides an idle mode to minimize the quiescent power; by further decreasing the S_{IN} frequency to f_{QCLK}/64. The synchronous mode-switching divider (SMS-DIV) enables a fast return to active mode. The proposed QCG generates precise quadrature clocks at f_{QCLK} with a 25% *DC* right before the TX, minimizing DQ skew. The QCG is designed based on a type-II PLL operating at f_{QCLK}/8, hence it can filter S_{IN}'s jitter while using only 900μW to generate quadrature S_{OUTx} at 2GHz. The proposed *DC*-comparing quadrature-error calibrator (DCQC) uses the *DC* information of S_{OUTx}, which are frequency independent, allowing the QCG to cover a very wide-range of f_{QCLK}: 1 – 4GHz. This *DC* comparator consumes only 50μW when operating at $f_{QCLK}/(8{\cdot}256)$. Finally, the quadrature error is corrected by independently controlling the delay cell of the individual-delay-controlling ring DCO (IDC-RDCO); removing the need for extra delay cells outside of the RDCO and allowing a further reduction of the output jitter.

The top of Fig. 28.5.2 shows the overall architecture of the proposed DPLL-based QCG. The SMS-DIV on the DLL side divides the DLL output's (S_{DLL}) f_{QCLK} by 8 to transmit S_{IN} to the TX with low power despite the long routing. Using S_{IN} as the reference clock, the TX QCG generates quadrature signals, $S_{RO,x}$, at the recovered frequency (f_{QCLK}). These quadrature $S_{RO,x}$ are generated by a 4-stage RDCO, but intrinsic mismatches between the delay cells of the RDCO cause quadrature errors; which becomes more severe as the target f_{QCLK} increases, making the RDCO more vulnerable to PVT variation. To address this issue, we present a IDC-RDCO with four delay cells, each of which can be controlled individually by the corresponding $D_{QC,x}$ control code. Since the RDCO quadrature errors are internally corrected the extra delay cells that cause additional jitter are not required. To directly provide the 25%-*DC* quadrature S_{OUTx} outputs to TX serializers, the QCG is designed to embed an AND-based 25%-*DC* converter. Since all rising and falling edges of S_{OUTx} come from the rising edges of $S_{RO,x}$ (e.g., the rising and falling edges of $S_{OUT,I}$ are from the rising edges of $S_{RO,I}$ and $S_{RO,Q}$), the S_{OUTx} outputs do not overlap one another. Hence, the four 25%-*DC* S_{OUTx} outputs can guarantee the precise quadrature relationship between them; the proposed DCQC is designed based on this property. First, the DCQC selects two consecutive S_{OUTx} based on D_{SEL}[1:0], which rotates in a 00→10→11→00→ order. Second, the *DC* comparator compares the DCs of the two selected signals. Third, the output code of the *DC* comparator, D_{COMP}, updates the corresponding $D_{QC,x}$, which individually adjusts the delay of the corresponding RDCO delay cell until all S_{OUTx} have a 25% *DC*. Since $S_{OUT,I}$ is used as the reference, $D_{QC,I}$ is fixed at the mid-code. The inverter's and pass-gate's phase mismatch at the differential inputs of the AND gate can also be corrected by DCQC. If D_{SEL}[1:0] is 0 (bottom-left of Fig. 28.5.2) then the *DC* of $S_{OUT,I}$ is larger than that of $S_{OUT,Q}$, and $D_{QC,Q}$ is decreased to move the rising edge of $S_{RO,Q}$ forward, simultaneously correcting the *DC* of $S_{OUT,I}$ and $S_{OUT,Q}$. A 50%-*DC* also is available by using S_{ROx}. Via SMS-DIV internal resampling, $S_{OUT,I}$ is synchronized with the output of SMS-DIV, S_{DIV}, thus is also synchronized with S_{IN} in steady state. The common controls to the RDCO, conveying the phase error information between S_{DIV} and S_{IN} (D_I and V_P) are applied to all four delay cells: thereby, correcting the frequency and the jitter of the RDCO. A frequency-acquisition path, with a dead-zone PD, is initially used to expedite frequency lock. To ensure stability, the bandwidth of the PLL is set much larger than that of the DCQC (bottom-right of Fig. 28.5.2), so that the PLL can settle quickly. Once the PLL settles, mode transitioning between idle and active states effectively takes zero transition time, due to SMS-DIV.

The top of Fig. 28.5.3 shows the *DC* comparator of DCQC. The comparison unit cycle is $256T_{IN}$, where T_{IN} is the period of S_{IN}, and has three non-overlapped phases: Φ_1, Φ_2, and Φ_3. When D_{SEL} is 0, during Φ_1, $S_{OUT,I}$ and $S_{OUT,Q}$ pass through 10-MHz low-pass filters, and their *DC* information is extracted as V_{DC1} and V_{DC2}. In this phase, auto-zeroing is enabled eliminates the differential pre-amplifier's input offset. Next, during Φ_2, the difference between V_{DC1} and V_{DC2} (V_{ERR}) is amplified by A_V via the pre-amplifier. Finally, a voltage comparator is triggered on the rising edge of Φ_3, generating a D_{COMP} code. The *DC* comparator has a resolution of 100μV, which translates to a 0.04° quadrature error, while consuming only 50μW. Since the proposed quadrature-error *DC*-comparing method is frequency independent, it can cover a much larger range of f_{QCLK} than prior methods using a DTC as an intermediate for comparing the timing of consecutive edges.

An idle mode is provided to save clock-distribution and QCG power while TXs are not being used. For an immediate transition between the idle (/64) and active (/8) modes, we use synchronous mode-switching with two SMS-DIVs: one on the DLL side (*SIDE_SEL* = 0) and the other on the TX side (*SIDE_SEL* = 1). In idle mode, as shown at the bottom of Fig. 28.5.3, the SMS-DIV on the DLL side transmits S_{IN} at f_{QCLK}/64 to the QCG's BBPD and the mode-switching code, D_{MD} (=0), to the TX-side SMS-DIV via a long routing. While the QCG is locked in steady-state, S_{IN} and S_{DIV} are synchronized with each other at the input of the BBPD. For a seamless transition between idle and active states, S_{IN} and S_{DIV} should stay aligned despite the sudden mode switch. To do this, regardless of when the external request comes (via EX_{MD}), the internal code selecting either S_{DIV64} or S_{DIV8}, (SEL_{MD}) is designed to change synchronously on the falling edge of S_{DIV8} right after the D_{MD} edge on both sides. Since the first rising edge of S_{IN} in active mode experiences almost the same delay, S_{IN} and S_{DIV} can stay aligned despite the mode switch. Hence, the DPLL can maintain lock without any loop disturbance, and the QCG can resume providing a low-jitter S_{OUTx} immediately. In the worst case, the mode switch is guaranteed to occur in less than 40ns after EX_{MD} toggles.

This work's QCG uses 900μW at 2GHz and 0.011mm² of area in 40nm CMOS. Figure 28.5.4 shows the measured 25%-*DC* quadrature S_{OUTx} waveforms at f_{QCLK} = 2GHz (top) and 4GHz (bottom). Without DCQC the intrinsic IDC-RDCO mismatches cause a ~2° quadrature error; however, with DCQC the quadrature error is reduced to <0.5°. The top of Fig. 28.5.5 shows that the jitter is significantly filtered by the QCG from the input (S_{DLL}) to the output ($S_{OUT,I}$), reducing it from 2.94 to 1.22ps$_{RMS}$ at 2GHz. This intrinsic jitter-filtering capability is more evident in phase-noise (PN) measurements at the bottom, which show that the S_{DLL}'s out-of-band PN is suppressed greatly at $S_{OUT,I}$. The top of Fig. 28.5.6 shows the effect of SMS-DIV, allowing $S_{OUT,I}$ to maintain exactly the same frequency without experiencing any disturbance despite abrupt transitions between the idle and active states. The bottom of Fig. 28.5.6 shows that the proposed QCG can provide jitter-filtering capabilities and dual *DC* options: 25% and 50%. This work also achieves tiny quadrature errors consistently over the largest range of f_{QCLK} while consuming the lowest power among the state-of-the-art QECs and QCGs [5].

Acknowledgement:
This work was supported by Samsung Electronics Co., Ltd (IO220321-09459-01). Chip fabrication was supported, in part, by the IC Design Education Center.

References:
[1] S. Shin et al., "22.6 A 0.8-to-2.3GHz Quadrature Error Corrector with Correctable Error Range of 101.6ps Using Minimum Total Delay Tracking and Asynchronous Calibration On-Off Scheme for DRAM Interface," *ISSCC*, pp. 340-341, 2020.
[2] Y. Kim et al., "A 2.3-mW 0.01-mm² 1.25-GHz Quadrature Signal Corrector With 1.1-ps Error for Mobile DRAM Interface in 65-nm CMOS," *IEEE TCAS-II*, vol. 64, no. 4, pp. 397-401, April 2017.
[3] J. Chae et al., "A Quadrature Clock Corrector for DRAM Interfaces, with a Duty-Cycle and Quadrature Phase Detector Based on a Relaxation Oscillator," *IEEE Tran. VLSI*, vol. 27, no. 4, pp. 978-982, April 2019.
[4] H. Yoon et al., "A 3.2-12.8Gb/s Duty-Cycle Compensating Quadrature Error Corrector for DRAM Interfaces, With Fast Locking and Low Power Characteristics," *ESSCIRC*, pp. 463-466, Sept. 2021.
[5] H. Park et al., "A 1.3-4-GHz Quadrature-Phase Digital DLL Using Sequential Delay Control and Reconfigurable Delay Line," *IEEE JSSC*, vol. 56, no. 6, pp. 1886-1896, June 2021.

 978-1-6654-9016-0/23/$31.00

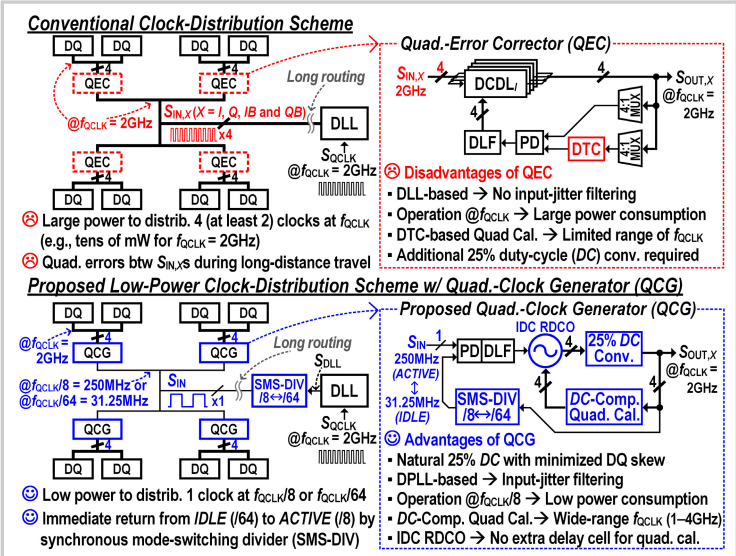


Figure 28.5.1: Conventional clock-distribution scheme with a quadrature-error corrector (QEC) (top) and the proposed low-power clock-distribution scheme with a quadrature-clock generator (QCG) (bottom).

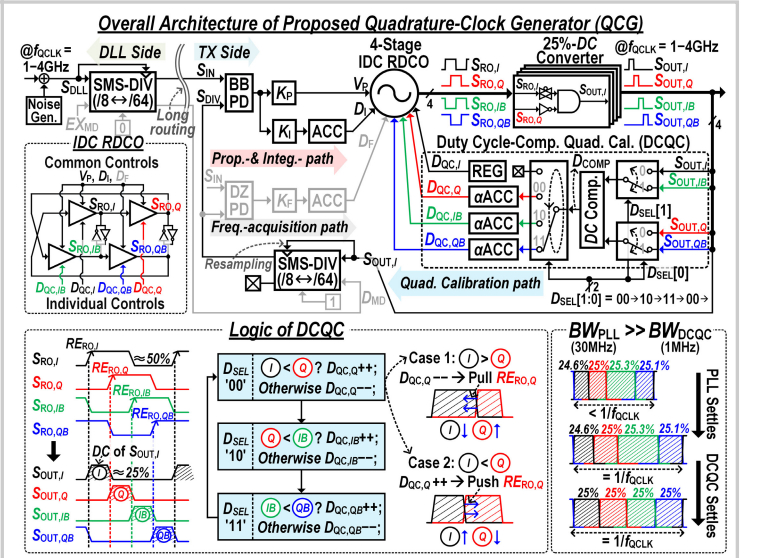


Figure 28.5.2: Overall architecture of the proposed QCG (top), the *DC*-comparing logic for the quadrature calibrator (DCQC) (bottom left), and the PLL and the DCQC bandwidth settings (bottom right).

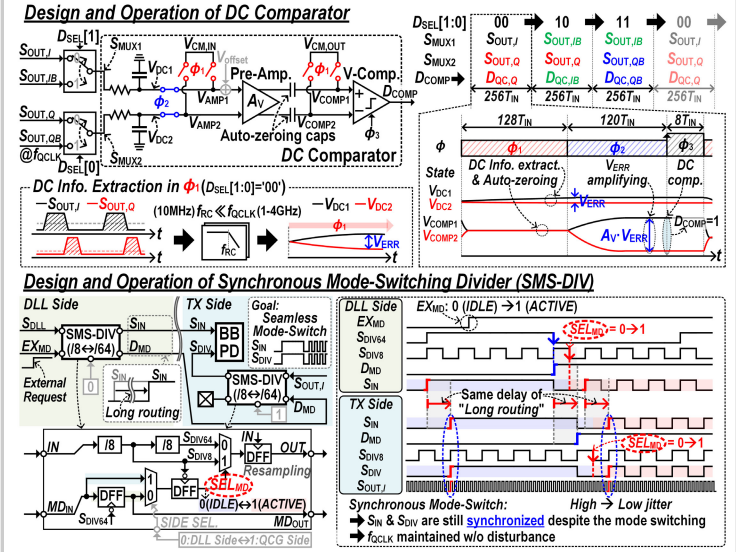


Figure 28.5.3: Design and operation of the *DC* comparator (top) and the design and operation of the synchronous mode-switching divider (SMS-DIV) (bottom).

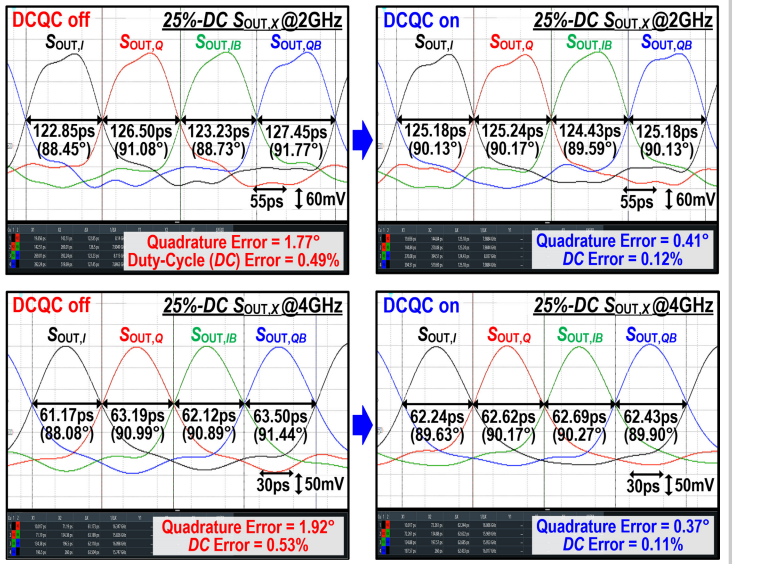


Figure 28.5.4: Measured waveforms of 25%-*DC* quadrature $S_{OUT,X}$ with DCQC disabled and enabled at f_{QCLK} = 2GHz (top) and 4GHz (bottom).

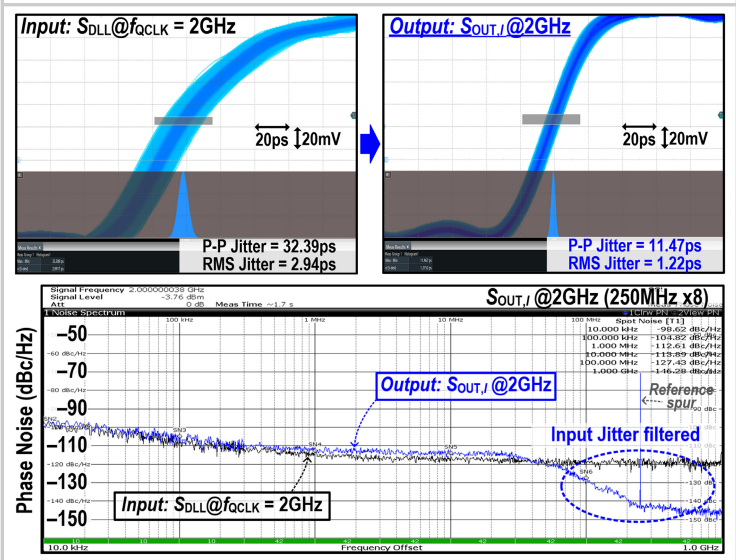
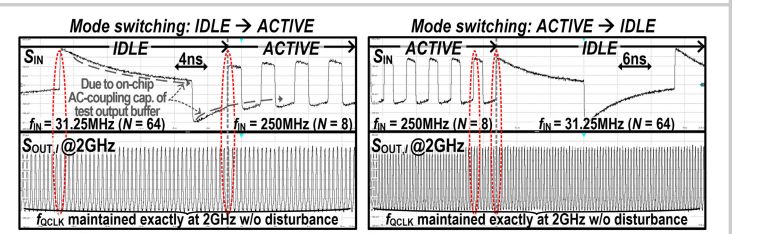


Figure 28.5.5: Measured RMS and peak-to-peak jitter for S_{DLL} (input) and $S_{OUT,I}$ (output) at 2GHz (top). Measured S_{DLL} (input) and $S_{OUT,I}$ (output) PN at 2GHz, showing the input-jitter-filtering capability of QCG (bottom).

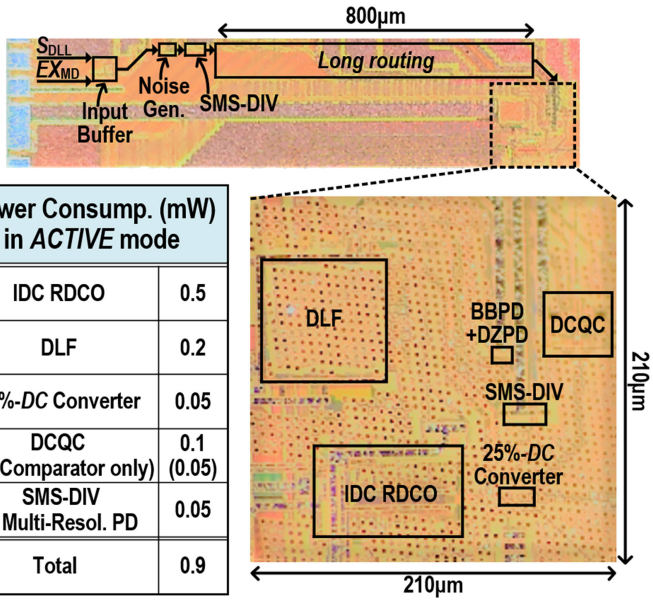


State-of-the-art Quadrature Error Correctors (QECs) and Quadrature Clock Generators (QCGs)

	This work	ISSCC'20 [1]	TCASII'17 [2]	TVLSI'19 [3]	ESSCIRC'21 [4]	JSSC'21 [5]
Process	40nm	40nm	65nm	55nm	28nm	28nm
Architecture	Digital-PLL QCG	Digital-PLL QEC	Digital-DLL QEC	Digital-DLL QEC	Digital-DLL QEC	Digital-DLL QCG
Quadrature Clock	Gen./ Correc.	Correc.	Correc.	Correc.	Correc.	Gen./ Correc.
Duty-Cycle (*DC*)	25% & 50%	50%	50%	50%	50%	50%
Freq. (f_{QCLK}) Range	1.0 – 4.0GHz	0.8 – 2.3GHz	1.25GHz	1.0 – 3.0GHz	0.8 – 3.2GHz	1.3 – 4.0GHz
Jitter Filtering	Yes	No	No	No	No	No
Input → Output $Jitter_{rms}$ @f_{QCLK}	2.94ps→1.22ps @2.0GHz	2.28ps→2.34ps @2.3GHz	1.84ps→2.53ps @1.25GHz	1.85ps→2.14ps @3.0GHz	NA*→1.31ps @3.2GHz	0.96ps→1.82ps @4.0GHz
Quadrature Error	< 0.5°	< 2.18°	< 0.48°	< 1.11°	< 1.84°	< 2.82°
Power Cons. @f_{QCLK}	0.9mW@2.0GHz	8.9mW@2.3GHz	2.3mW@1.25GHz	2.1mW@3.0GHz	9.80mW@3.2GHz	6.5mW@4.0GHz
Power Efficiency	0.45mW/GHz	3.87mW/GHz	1.82mW/GHz	0.69mW/GHz	3.06mW/GHz	1.63mW/GHz
Active Area	0.011mm²	0.012mm²**	0.004mm²**	0.003mm²	0.010mm²	0.004mm²

* Input Jitter was not reported, ** Estimated from die micrograph

Figure 28.5.6: Measured S_{IN} and $S_{OUT,I}$ waveforms, showing seamless transition, due to the SMS dividers, when $S_{OUT,I}$ is 2GHz (top). Performance comparison of this work with the state-of-the-art QECs and QCGs (bottom).



Power Consump. (mW) in *ACTIVE* mode	
IDC RDCO	0.5
DLF	0.2
25%-*DC* Converter	0.05
DCQC (*DC* Comparator only)	0.1 (0.05)
SMS-DIV & Multi-Resol. PD	0.05
Total	0.9

Figure 28.5.7: Die micrograph and power breakdown table.

978-1-6654-9016-0/23/$31.00

28.6 A 32Gb/s/pin 0.51pJ/b Single-Ended Resistor-less Impedance-Matched Transmitter with a T-Coil-Based Edge-Boosting Equalizer in 40nm CMOS

Jung-Hun Park[1], Hyeonseok Lee[1], Hoyeon Cho[1], Sanghee Lee[1], Kwang-Hoon Lee[1], Han-Gon Ko[2], Deog-Kyoon Jeong[1]

[1]Seoul National University, Seoul, Korea
[2]ONEsemiconductor, Gyeonggi, Korea

To cope with the rapidly growing data demand, the DRAM interface bandwidth also increases steeply each year; for graphics applications, the bandwidth per pin has increased to 27Gb/s/pin, thanks to T-coils implemented using RDL layers [1]. As I/O hardware expands, its area and power consumption are also increasing. To alleviate the DRAM interface burden two key ideas are proposed in this paper: (1) a PN-over-NP driver capable of impedance matching, without the use of a resistor, significantly reduces the chip area, and; (2) a T-coil-based edge-boosting equalizer, which does not consume static current when there no sequence transition, that enables impedance matching at high frequencies. In addition, a CMOS clock-edge corrector is introduced to reduce clock skew. As a result, the proposed transmitter achieves 32Gb/s, while maintaining high signal integrity, small area, and low-power consumption.

Figure 28.6.1 shows the overall architecture of the proposed transmitter. Quarter-rate clocks are generated in the IQ divider from a 16GHz external differential clock. The edge corrector adjusts the phase and duty cycle for each output clock. A 4:1 serializer generates full-rate data (D_{in}) and the inverted data for the driver using a corrected clock. A capacitor-coupled edge-boosting equalizer uses D_{in} to assist transitions via pulse generation. The output of the equalizer is connected to the center tab of the asymmetric T-coil along with an ESD protection circuit to maintain constant output impedance and to extend the bandwidth.

The proposed PN-over-NP driver overcomes the limitations of existing drivers by adding an N-over-P driver that compensates for non-linearity. A source-series-termination (SST) driver occupies a large area, due to the series resistor. Moreover, it also increases the pre-driver power consumption, since the size of the transistor of the SST driver must be increased to ensure high linearity. An inverter-based driver without a series resistor reduces the area significantly, but is only capable of far-end matching [2]. An N-over-N driver operates at a low voltage and consumes only a small amount of power. However, its output impedance varies greatly depending on the output voltage, due to transistor non-linearities. Linearity can be improved, by adding an encoder and a transistor [3], but the area, power consumption, and latency increase. On the other hand, a PN-over-NP driver does not use a passive resistor, so it is more scalable than a SST driver and occupies less area. In addition, since it operates at a low voltage it consumes less power and has a high linearity; thereby, improving signal integrity through better impedance matching.

The operating principle and characteristics of the PN-NP driver are illustrated in Fig. 28.6.2. When the CMOS inverter operates in the linear region, a non-linearity, due to V_{DS}^2. remains and the current drive of the pull-up or pull-down are thus weakened. To compensate for the insufficient current, an N-over-P pair is added to the CMOS inverter using the inverted data as input and operates in the saturation region. Since the I-V characteristic is symmetrical the sum of the currents shows excellent linearity. Post-layout simulation results show that output impedance is maintained within ±10Ω across the output swing. Digitally-controlled transistors located in the pull-up and pull-down path of the driver, allow for the fine adjustment of target impedance from 40 – 60Ω. Moreover, the PN-over-NP driver allows the output swing to be greater than $½V_{DDQ}$. The DC impedance (V/I) of the SST driver is approximately equal to $\partial V/\partial I$ due to the characteristic of a passive resistor. However, the DC impedance of the PN-NP driver is reduced when the output voltage is low since the pull-down PMOS is turned off. Hence, the DC voltage corresponding to logic 0 is lowered, thereby improving the swing, which makes it possible to secure a more considerable voltage margin in the face of a decreasing supply voltage.

Due to an increase in the Nyquist frequency, the transmitter also needs to equalize for channel losses via a feed-forward equalizer (FFE). For the case of conventional FFE, the advantage of low-power consumption of a pseudo-open drain structure is diminished since it consumes static current even in the absence of transitions, thereby wasting power. Moreover, it also requires a lot of power to secure the timing margins needed for the 1-UI delayed signal. There have been attempts to overcome these shortcomings in the design of memory interfaces. For example, an addition-only FFE (AFFE) is more power-efficient by removing the current path required for FFE tap subtraction, but it still requires re-timers and consumes power for calculating the tap coefficients [2]. An edge-boosting equalizer, as an alternative, does not have a static current path and does not require retiming blocks [4]. However, since it is coupled to the output node, the signal integrity deteriorates as the output impedance decreases at high frequencies. Using edge detectors to maintain a high-Z state during idle data periods requires a short-UI pulse generator, which consumes additional power. Furthermore, an impedance drop during data transitions is still present [5]. These problems can be solved by capacitively coupling the equalizer to the load of the asymmetric T-coil. Figure 28.6.3 shows the combination of the proposed edge-boosting equalizer and T-coil, and their RLC equivalent small-signal model. Input data is capacitively coupled to node C to boost transition edges. The output impedance remains relatively constant at high frequencies due to the inductance and parasitic capacitance (C_b), which also conceal the load impedance. Post-layout simulation results show that the impedance drop is improved by 47% compared to the non-T-coil configuration. The T-coil is designed using mostly thick metal, emulating the RDL layer of the DRAM process; less than 20% of the layout uses thinner-metal layers.

As the Nyquist frequency increases a sophisticated clock control is essential to ensure sufficient eye margin. A typical 4-phase clock requires a duty-cycle corrector and a quarter-rate phase error corrector, which consumes large area and power. In the proposed transmitter, a passive-less CMOS-based clock edge corrector (CEC) is used to minimize area. Eight edges of the quarter-rate clock can be independently controlled to simultaneously adjust their phase and duty cycle. Figure 28.6.4 shows the structure and behavior of the proposed CEC. The two tri-state inverter stages adjust the rise and fall time of the input clock; the last starved-inverter stage provides fine adjust. The each stage slice is weighted differently to improve CEC linearity and to maintain its monotonicity. If the rise and fall times are adjusted in the same direction, the phase leads or lags. In contrast, the duty-cycle changes when the rise and fall times are adjusted in the opposite direction. It is also possible to change the phase and duty cycle simultaneously using a linear combination. The table in Fig. 28.6.4 shows the operation method of the CEC for nine cases. The 5b control code is converted into a 12b thermometer code by a look-up table. The edge control resolution is 250fs, based on post-layout simulations, with an INL and DNL less than ±1 LSB.

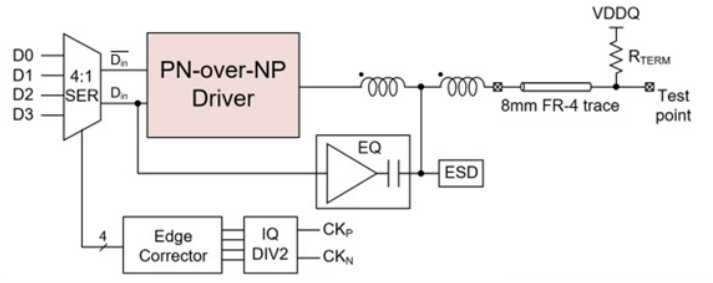
The prototype chip is fabricated with a 40nm CMOS process and tested with a 1V V_{DD} and a 0.6 V_{DDQ}. As briefly described in Fig. 28.6.1, the output of the transmitter is 50Ω terminated to V_{DDQ} through an 8mm long FR-4 PCB trace. Figure 28.6.5 shows eye diagrams measured at 12.8, 20, and 32Gb/s. At 12.8 and 20Gb/s a sufficient voltage margin is obtained without the equalizer. At 32Gb/s the vertical eye opening is 87 and 114mV for the case when the equalizer is turned off and on; in both cases, the horizontal timing margin is more than 0.5UI.

The performance of the proposed transmitter is summarized and compared with other single-ended transmitters for memory interfaces in Fig. 28.6.6. The proposed PN-NP driver enables TX impedance matching while reducing area and saving power. The proposed 2-tap edge boosting equalizer offers better power efficiency compared to a conventional FFE, especially during non-transition in a data sequence. In addition, the output impedance is well-matched during transitions. The total power consumption of the transmitter is 16.3mW, and its energy efficiency is 0.51pJ/b. Figure 28.6.7 shows the chip micrograph and a power breakdown graph, calculated based on post-layout simulation results. The chip area, including the T-coil, is 5008um^2. The driver and V_{DDQ} termination consume 1.35mW, the serializer consumes 9.81mW, and the CEC and clock buffers consume 5.14mW.

References:

[1] D. Lee et al., "A 16Gb 27Gb/s/pin T-coil based GDDR6 DRAM with Merged-MUX TX, Optimized WCK Operation, and Alternative-Data-Bus," *ISSCC*, pp. 446-447, 2022.
[2] C. Moon et al., "A 20 Gb/s/pin 1.18pJ/b 1149μm^2 Single-Ended Inverter-based 4-tap Addition-Only Feed-Forward Equalization Transmitter with Improved Robustness to Coefficient Errors in 28nm CMOS," *ISSCC*, pp. 450-451, 2022.
[3] Y.-U. Jeong et al., "A 0.64-pJ/Bit 28-Gb/s/Pin High-Linearity Single-Ended PAM-4 Transmitter with an Impedance-Matched Driver and Three-Point ZQ Calibration for Memory Interface," *JSSC*, vol. 56, no. 4, pp. 1278-1287, April. 2021.
[4] S.-M. Lee et al., "An 8nm 18Gb/s/pin GDDR6 PHY with TX Bandwidth Extension and RX Training Technique," *ISSCC*, pp. 338-339, 2020.
[5] J. M. Wilson et al., "A 1.17pJ/b 25Gb/s/pin ground-referenced single-ended serial link for off- and on-package communication in 16nm CMOS using a process- and temperature-adaptive voltage regulator," *ISSCC*, pp. 276-277, 2018.

 978-1-6654-9016-0/23/$31.00



	VDD/VDDQ	VDDQ	VDDQ
Structure	SST driver	N-over-N driver	PN-over-NP driver
Linearity	Good	Bad	Good
Scalability	Bad	Good	Good
Swing	Large	Small	Small
Power Consumption	Large	Small	Small
Area	Large	Small	Small

Figure 28.6.1: Overall architecture of the proposed transmitter, including PN-over-NP driver(top). Comparison of driver structures (bottom).

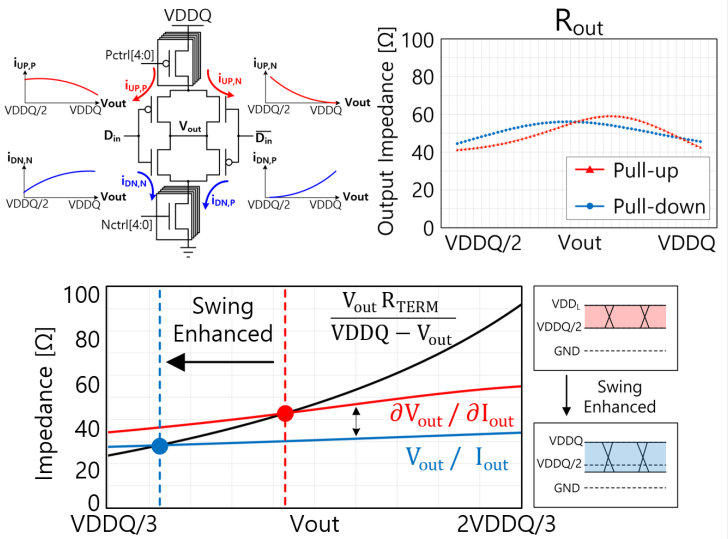


Figure 28.6.2: Proposed PN-over-NP driver operation (top-left), output impedance (top-left), and swing characteristics (bottom).

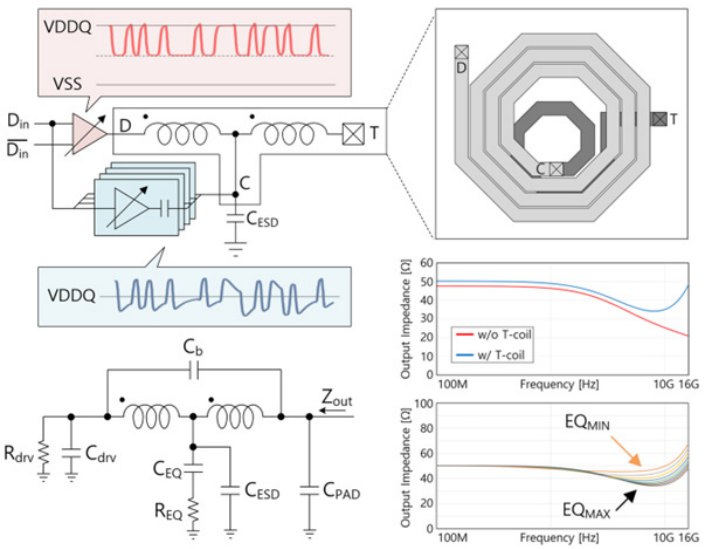
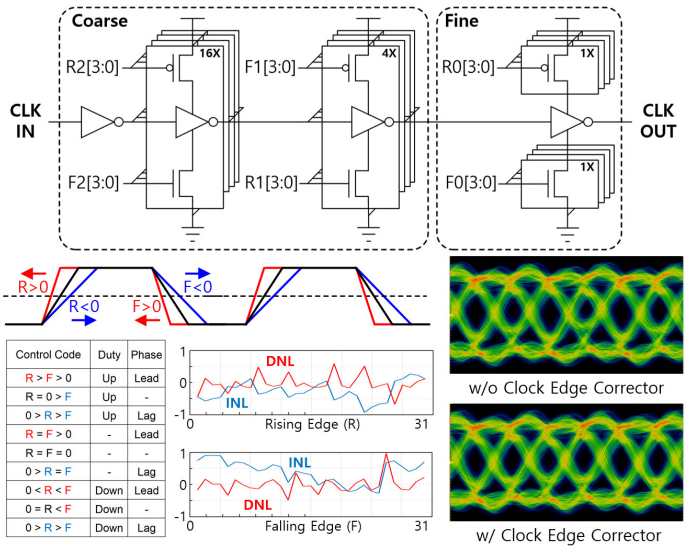


Figure 28.6.3: Asymmetric T-coil-based edge-boosting equalizer schematic and T-coil implementation (top). RLC small-signal model of the proposed equalizer and output impedance characteristics (bottom).



Control Code	Duty	Phase
R > F > 0	Up	Lead
R = 0 > F	Up	-
0 > R > F	Up	Lag
R = F > 0	-	Lead
R = F = 0	-	-
0 > R = F	-	Lag
0 < R < F	Down	Lead
0 = R < F	Down	-
0 > R > F	Down	Lag

Figure 28.6.4: Clock-edge corrector schematic (top). Its operation and post-layout simuilated INL & DNL (bottom-left), Measured eye diagrams with and without the clock edge corrector (bottom-right).

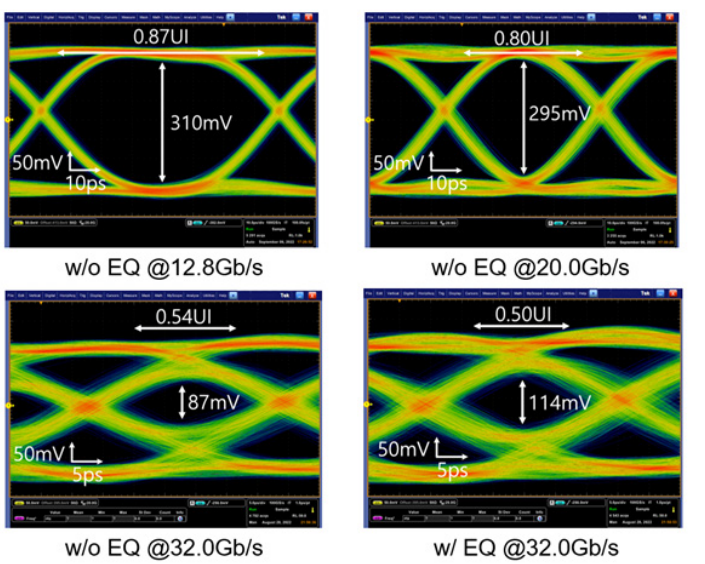


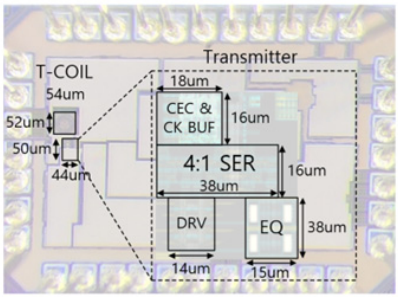
Figure 28.6.5: Measured eye diagrams at 12.8 and 20.0Gb/s without equalizer (top). Measured eye diagrams for 32Gb/s with and without equalizer (bottom).

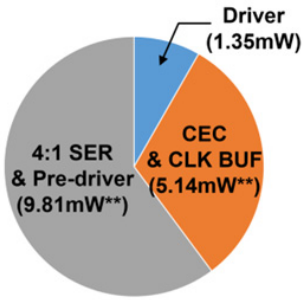
		ISSCC'22 [2]	JSSC'21 [3]	ISSCC'20 [4]	ISSCC'18 [5]	Kang JSSC'22	Ko ISSCC'20	Chiu ISSCC'20	This Work
Technology		28nm LPP	65nm CMOS	8nm FinFET	16nm FinFET	28nm CMOS	65nm CMOS	65nm CMOS	40nm CMOS
Data rate [Gb/s]		20	28	18	25	21	4	32	32
Signaling		NRZ	PAM-4	NRZ	GRS	Duobinary	NRZ	PAM-4	NRZ
Supply voltage	VDD [V]	1.1	1.0	0.85	0.75	1.0	1.2	1.2	1.0
	VDDQ [V]		0.6	1.35		0.8			0.6
Driver & Equalizer	Driver type	Inverter	N-over-N	High-voltage SST	Charge Pump	SST	Inverter	SST	PN-over-NP
	TX equalization	4-tap AFFE (pre & post 2)	2-tap pre-emphasis	2-tap Edge boosting	2-tap Edge boosting	3-tap FFE (pre & post)	2-tap FFE (post + XT)	3-tap FFE (half-rate)	2-tap Edge boosting
	No static current during IDLE state	O	O	O	X	X	X	X	O
	Impedance matching during transition	X	X	X	X	O	X	O	O
Energy efficiency [pJ/b]		1.18	0.58*	N/A	1.17**	0.67	0.9	0.97**	0.51
Area [mm²]		0.00115	0.033	4.15	0.0102**	0.0072	0.0027***	0.009**	0.00501

* Excludes PRBS generator and 32:8 serializer (according to the power breakdown)
** TRX
*** Area / (# of I/O)

Figure 28.6.6: Performance comparison table for state-of-the-art single-ended TX-equalized transmitters.

28





T-coil		54um X 52um
Transmitter	CEC & CK BUF	18um X 16um
	4:1 SER	38um X 16um
	DRV	14um X 38um
	EQ	15um X 38um

* Includes the VDDQ termination
** Measurement result is separated based on post-layout simulation results

Figure 28.6.7: Chip micrograph and power-breakdown summary chart.

978-1-6654-9016-0/23/$31.00

28.7 A 1.1V 6.4Gb/s/pin 24-Gb DDR5 SDRAM with a Highly-Accurate Duty Corrector and NBTI-Tolerant DLL

Daehyun Kwon, Heon Su Jeong, Jaemin Choi, Wijong Kim, Jae Woong Kim, Junsub Yoon, Jungmin Choi, Sanguk Lee, Hyunsub Norbert Rie, Jin-il Lee, Jongbum Lee, Taeseong Jang, JunHyung Kim, Sanghee Kang, Jungbum Shin, Yanggyoon Loh, Chang Yong Lee, Junmyung Woo, Hyeseung Yu, Changhyun Bae, Reum Oh, Young-soo Sohn, Changsik Yoo, Jooyoung Lee

Samsung Electronics, Hwaseong, Korea

The need for high-quality multi-media data increases the amount of data to be stored and processed, necessitating DDR5 to achieve high-density and high-speed with low-power consumption [1]. However, high-speed with low-power operation makes DRAM more vulnerable to process-voltage-temperature (PVT) variations, negative-bias thermal instability (NBTI), etc. In this work, a mono-die based 24-Gb high-density DDR5 achieving 6.4Gbps/pin is implemented. To lower power consumption, GIO switching is reduced by using a GIO separation switch and a read-only GIO pre-charge scheme. The proposed DRAM has a higher tolerance to NBTI, since the delay-locked loop (DLL) experiences slow toggling during self-refresh operations where the DLL is not necessary. Also, adaptive body bias (ABB) is used to combat process variation [2], thereby achieving high-performance I/O circuits. In addition, a low-pass filter is added for higher operations and sensitivities in front of charge pump, which is used by a duty cycle error detector (DCD) and a quadrature error detector (QED). Additionally, a balanced MUX and a bandwidth booster are also used in the transmitter for high-speed operations.

Figure 28.7.1 shows the 24-Gb/ch DRAM architecture. To increase the per channel density, the physical size of each bank needs to become larger, and in turn increase the GIO length resulting in increased power consumption. In our design, two solutions are proposed to reduce power consumption: (1) a switch is added to the GIO lines to separately access near and far blocks. This switch is only turned on to access the far blocks, and is turned off when the DRAM accesses near blocks; thereby reducing the loading capacitance of the GIO lines, therefore decreasing the dynamic power consumption. (2) GIO pre-charge scheme is applied only for read operations, as shown in Fig. 28.7.1(c), to solve the increased current consumption due to longer GIO lines. The scheme is not used for write operations, as the write drivers are strong enough to drive the GIOs during write. Consequently, our DRAM achieves a lower power consumption than the DRAM having a 1.5× smaller density as shown in Fig. 28.7.1(d).

Header-only power gating was used in [3] to suppress device degradation, but the high-speed performance of I/O circuits is sacrificed as the virtual power supply level drops. For the case of DLL operation, the phase and delay are controlled by adapting the current digitally, which causes the virtual power to drop differently according to the amount of current used. As a result, the DLL can experience unwanted duty and delay variations, which can also induce noise and jitter. In our design, instead of using header-only power gating, a low frequency oscillator-based toggling scheme [3] is applied to the DLL clock path, as shown Fig. 28.7.2, during the standby state, where the DLL output clock is unused. The number of delay cells can be changed to compensate for clock delays against voltage and temperature variations when the DLL is operating; thus, the proposed schemes make all delay cells in the DLL toggle forcefully regardless of whether the delay cells are being locked or not. As a result, the DLL can experience NBTI half of the time so that it can decrease V_t variations from NBTI, which might vary the the duty and/or delay in Fig. 28.7.2(a). As shown in the simulation results, Fig. 28.7.2(b) and (c), the DLL with the self-toggling schemes performs much better with NBTI degradation compared no protection.

Figure 28.7.3 conceptually shows the charge pump used in the DLL for DCD and QED. The input to the DCD is $f_{CK}/2$, and the f_{CK} output from CK0 and CK90 is used as the input to the QED. Both the DCD and the QED measure the duty cycle of each input to adjust the duty cycle and quad-skew of $f_{CK}/2$ respectively. The charge pump makes proportional charges to the duty cycle and integrates it on a capacitor, as shown in Fig. 28.7.3(a). Since the charge is only proportional to the time difference of the logically high and low pulse width, the amount of integration is so small that a duty cycle close to 50% takes longer for the sampler to determine as a logical high or low. Furthermore, the time is highly related with the loop delay, which might make the loop unstable, especially in high-speed operation because the DLL updates the QED/DCD codes in proportion to the input clock frequency. A low pass filter added in front of the charge pump to make the input common mode level different according to the duty cycle, as can be seen in Fig. 28.7.3(c); thus, integrating the charge faster. Figure 28.7.3(d) shows the Monte-Carlo simulation results of the DCD. Accuracy is calculated as the number of right decisions according to an input duty cycle among 1000 simulations. This shows indirectly that the charge pump with an LPF has a much higher loop-gain and sensitivity compared to one without. Although it has advantages in the aspects of DLL performance, it should be carefully designed considering the charge pump bandwidth including the LPF to not sacrifice the stability of the DLL loop. In our design, the bandwidth of the charge pump including LPF is set to 1.6GHz which is ten times larger than the loop bandwidth of the DLL.

Figure 28.7.4 shows a block diagram and the operation of the ABB [2]. In our design, only the reverse body bias (RBB) is applied to decrease the effect of process variation. The body bias level is controlled by the ABB monitoring circuits based on the process corner. For example, -0.8V is applied to the NMOS body in the fast corner, -0.4V for the typical corner, and 0V for the slow corner; thereby, aligning all corners to the slow side. As shown in Fig. 28.7.4(b), the distribution of the propagation delay of inverter chains, t_{PD}, moves to a lower speed; however, the variations of t_{PD} and I_{DD2P} are decreased. Additionally, ABB is also applied to the I/O to not only decrease the process variation but to also improve the I/O performance: such as CIO reduction and R_{on} variations of the output drivers across all process corners. Figure 28.7.4(c) shows that CIO decreases with increasing ABB voltage, as the junction capacitance of the output drivers is decreased with the ABB voltage.

Since ABB makes all process corners align to the slow side, the I/O circuits need high-speed techniques, especially in the transmitter where most of the circuits are designed as CMOS logic. A balanced multiplexer and a bandwidth booster [4], shown in Fig. 28.7.5, are designed for high-speed operation. The balanced multiplexer consists of a transmission gate for data transitions and reset gates for data ignoring, which can decrease the size of a serializer compared with conventional one [5]. A bandwidth booster is designed with 3 inverters using ABB to regulate the delay in the face of process variation, so that it can compensate inter-symbol interference simply without using 1UI delayed pre-emphasis techniques that are generally used [6].

The measured automated test equipment (ATE) results are shown in Fig. 28.7.6. The proposed DDR5 achieves 6.4Gbps/channel at 1.1V. Our DRAM can achieve 7.7Gbps at 1.1V, as shown by the frequency voltage Shmoo. To satisfy low-voltage specifications, the valid DQ window is measured at 6.4Gbps, and the window size is 116ps for read and 88ps for write operations. The write window size is relatively small compared to the read as the equalizers, such as the decision feedback equalizer and the continuous time linear equalizer, are not used. Additionally, as shown in Fig. 28.7.6(c), our DRAM can tolerate NBTI degradation by using toggling schemes. The chip microphotograph is shown in Fig. 28.7.7. The 24-Gb density DDR5 occupying 71.8mm^2/channel is implemented in a 4th generation 10-nm DRAM technology.

References:

[1] C. Lee et al., "An 8.5-Gb/s/Pin 12-Gb LPDDR5 SDRAM with a Hybrid-Bank Architecture, Low Power, and Speed-Boosting Techniques," *IEEE JSSC, vol. 56, no. 1, pp.* 212-224, Jan. 2021.

[2] Y. Kim et al., "A 16Gb Sub-1V 7.14Gb/s/pin LPDDR5 SDRAM Applying a Mosaic Architecture with a Short-Feedback 1-Tap DFE, an FSS Bus with Low-Level Swing and an Adaptively Controlled Body Biasing in a 3rd-Generation 10nm DRAM," *ISSCC*, pp. 346-347, 2021.

[3] K. Chun et al., "A 16Gb LPDDR4X SDRAM with an NBTI-Tolerant Circuit Solution, an SWD PMOS GIDL Reduction Technique, an Adaptive Gear-Down Scheme and a Metastable-Free DQS Aligner in a 10nm Class DRAM Process," *ISSCC*, pp. 206-207, 2018.

[4] H. 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.

[5] W. Choi et al., "A 0.45-to-0.7V 1-to-6Gb/s 0.29-to-0.58pJ/b Source-Synchronous Transceiver Using Automatic Phase Calibration in 65nm CMOS," *ISSCC*, pp. 66-67, 2016.

[6] P. Peng et al., "A 56Gb/s PAM-4/NRZ Transceiver in 40nm CMOS," *ISSCC*, pp. 110-111, 2017.

 978-1-6654-9016-0/23/$31.00

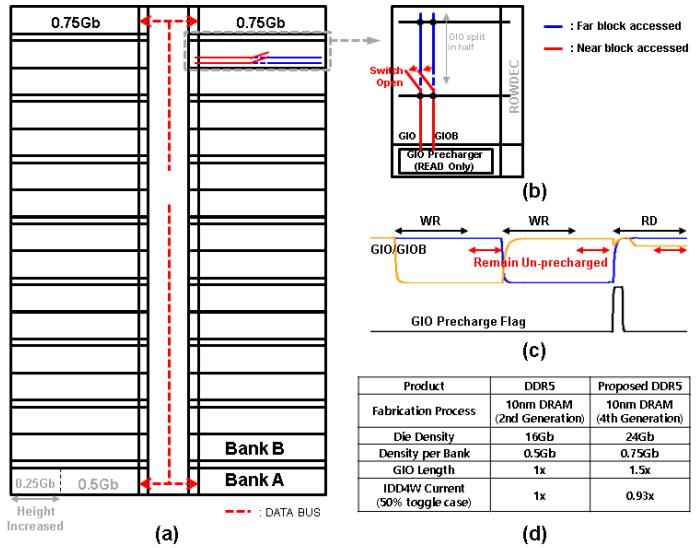


Figure 28.7.1: (a) Proposed 24Gb DDR5 architecture (b) diagram of switched GIO and RD only GIO precharger (c) GIO behavior under RD only GIO precharge system, and (d) architecture comparison.

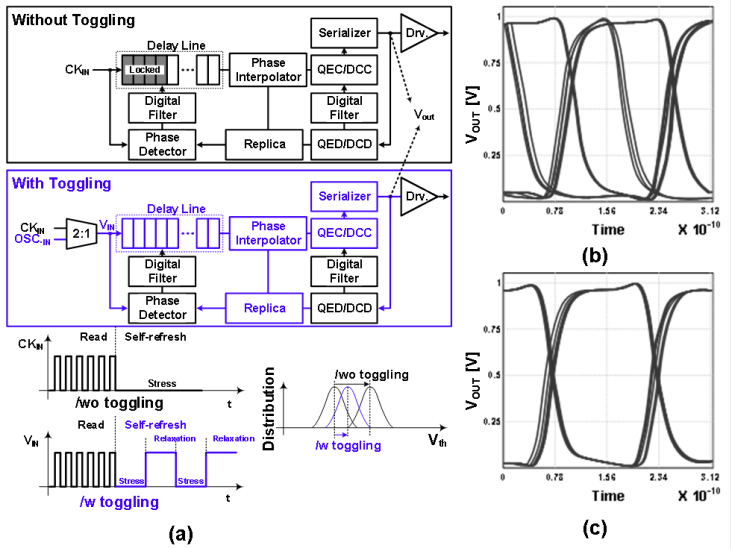


Figure 28.7.2: (a) Conceptual DLL block diagram with/without toggling schemes (b) DLL experienced NBTI simulation results without toggling schemes, and (c) with toggling schemes.

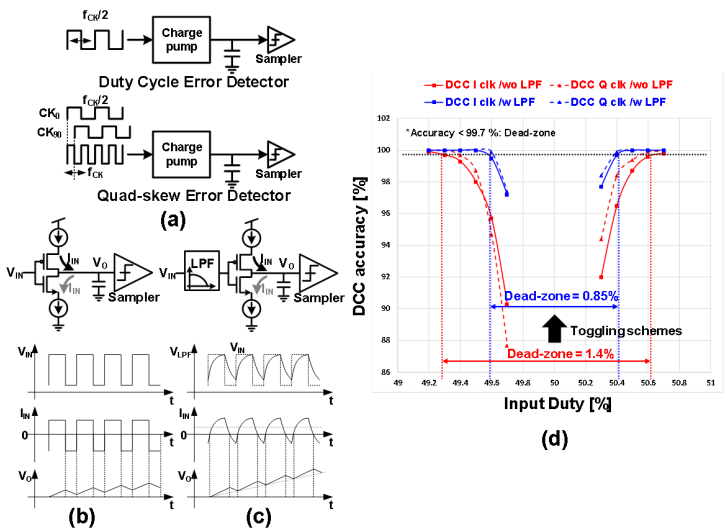


Figure 28.7.3: (a) DCD/QED schemes (b) a conventional charge pump (c) a proposed charge pump with a low pass filter, and (d) monte-carlo simulation results.

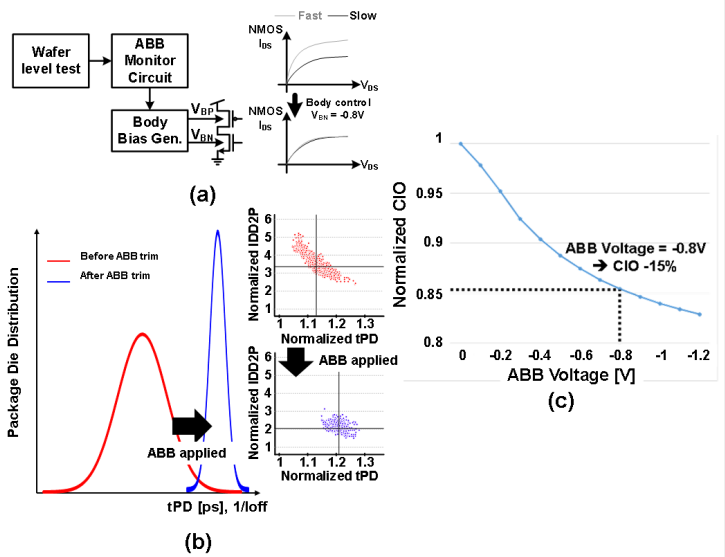


Figure 28.7.4: (a) ABB block diagrams (b) package die distributions, and (c) CIO simulation results with ABB voltage.

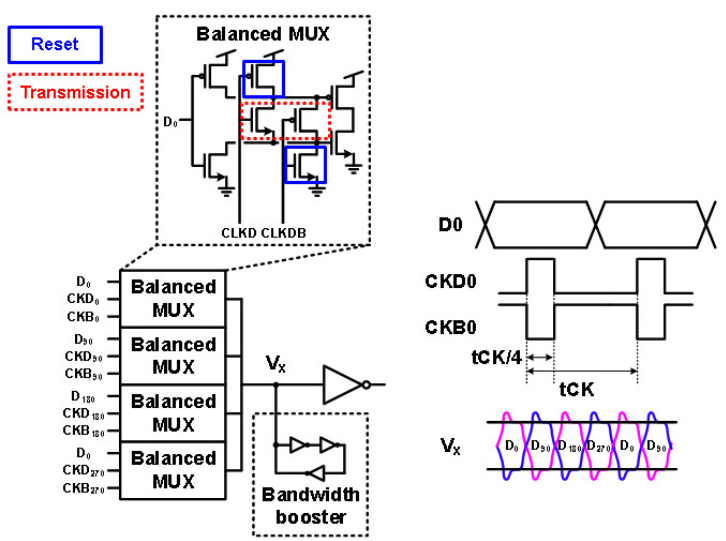


Figure 28.7.5: Balanced MUXs and a bandwidth booster.

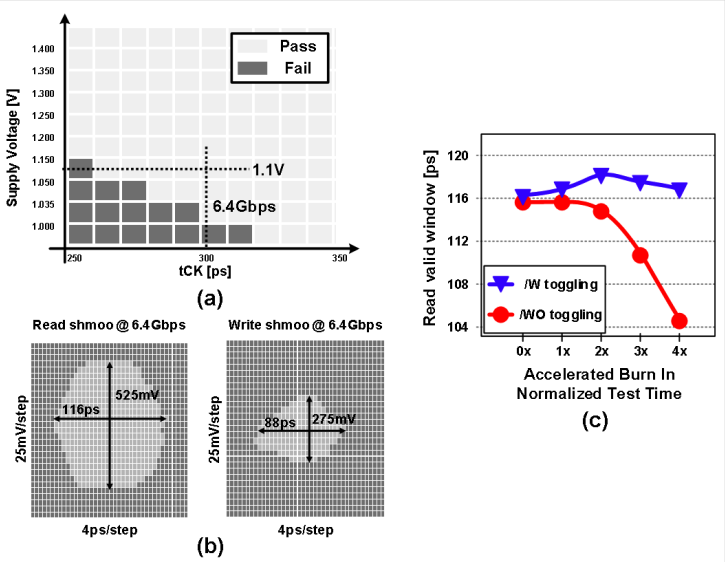


Figure 28.7.6: Measured (a) frequency voltage shmoo (b) RD/WR shmoos, and (c) read valid window in DDR burn in test.

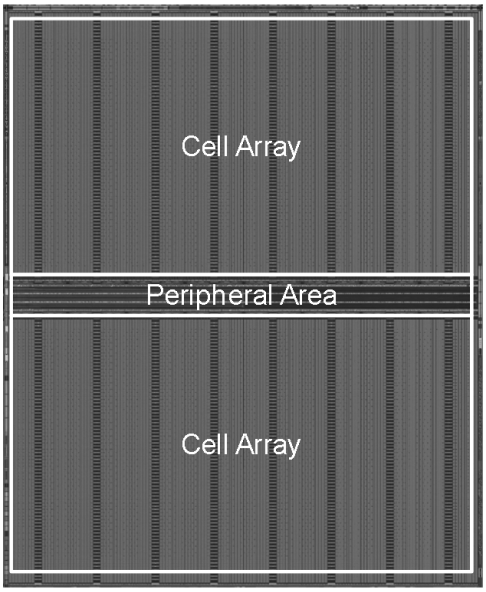


Figure 28.7.7: Chip micrograph of 24Gb DDR5.

978-1-6654-9016-0/23/$31.00

28.8 A 1.1V 16Gb DDR5 DRAM with Probabilistic-Aggressor Tracking, Refresh-Management Functionality, Per-Row Hammer Tracking, a Multi-Step Precharge, and Core-Bias Modulation for Security and Reliability Enhancement

Woongrae Kim, Chulmoon Jung, Seongnyuh Yoo, Duckhwa Hong, Jeongjin Hwang, Jungmin Yoon, Ohyong Jung, Joonwoo Choi, Sanga Hyun, Mankeun Kang, Sangho Lee, Dohong Kim, Sanghyun Ku, Donhyun Choi, Nogeun Joo, Sangwoo Yoon, Junseok Noh, Byeongyong Go, Cheolhoe Kim, Sunil Hwang, Mihyun Hwang, Seol-Min Yi, Hyungmin Kim, Sanghyuk Heo, Yeonsu Jang, Kyoungchul Jang, Shinho Chu, Yoonna Oh, Kwidong Kim, Junghyun Kim, Soohwan Kim, Jeongtae Hwang, Sangil Park, Junphyo Lee, Inchul Jeong, Joohwan Cho, Jonghwan Kim

SK hynix Semiconductor, Icheon, Korea

DRAM products have been recently adopted in a wide range of high-performance computing applications: such as in cloud computing, in big data systems, and IoT devices. This demand creates larger memory capacity requirements, thereby requiring aggressive DRAM technology node scaling to reduce the cost per bit [1,2]. However, DRAM manufacturers are facing technology scaling challenges due to row hammer and refresh retention time beyond 1a-nm [2]. Row hammer is a failure mechanism, where repeatedly activating a DRAM row disturbs data in adjacent rows. Scaling down severely threatens reliability since a reduction of DRAM cell size leads to a reduction in the intrinsic row hammer tolerance [2,3]. To improve row hammer tolerance, there is a need to probabilistically activate adjacent rows with carefully sampled active addresses and to improve intrinsic row hammer tolerance [2]. In this paper, row-hammer-protection and refresh-management schemes are presented to guarantee DRAM security and reliability despite the aggressive scaling from 1a-nm to sub 10-nm nodes. The probabilistic-aggressor-tracking scheme with a refresh-management function (RFM) and per-row hammer tracking (PRHT) improve DRAM resilience. A multi-step precharge reinforces intrinsic row-hammer tolerance and a core-bias modulation improves retention time: even in the face of cell-transistor degradation due to technology scaling. This comprehensive scheme leads to a reduced probability of failure, due to row hammer attacks, by 93.1% and an improvement in retention time by 17%.

The technology scaling platform, shown in Fig. 28.8.1, is implemented in 1a-nm 16-Gb DDR5 DRAM chip. Row control circuits consist of row hammer (R/H) control circuits based on probabilistic approaches to improve row hammer aggressor tracking ability. This protection scheme is cost-effective since it can be implemented within the peripheral circuit area. Aggressor tracking accuracy is improved using the PRHT scheme, which requires additional bank area and counts the number of active commands for each WL [2]. PRHT consists of a control unit, a read-modify-write (RMW) control block in the column control circuit, and additional R/H cells. A multi-step precharge (PCG) control scheme generates row control signals to implement multiple stages of the WL level during precharge operations to improve the intrinsic row-hammer tolerance. A core-bias modulation scheme is adopted to minimize the temperature variation of the intrinsic row hammer tolerance, which leads to increased refresh-retention time for reliability and reduces refresh power consumption.

The DRAM controller counts the number of activation executions (RAACNT) and reads the threshold (RAAIMT) from the mode-register in the DRAM (see RFM algorithm in Fig. 28.8.2) [3]. When the RAACNT value is larger than RAAIMT, the controller executes RFM so that the DRAM conducts an additional refresh operation with the sampled address from the DRAM R/H control circuit. Refresh (REF) command is utilized for both a normal refresh operation to guarantee cell retention time and a target refresh operation for row-hammer mitigation with the sampled activation address [4]. Based on RFM and REF, the refresh-command-generation block in the R/H control circuit generates a row-hammer refresh command (RH_REF) and normal refresh (NREF) command for cell retention time. Probabilistic-aggressor-tracking (PAT) logic is proposed to sample the active addresses for row hammer mitigation with a higher accuracy than based on probabilistic approaches. To sample the active addresses randomly, a random generation block in the PAT generates a random flag (EN) based on a pseudorandom-binary sequence. When randomly selected activation commands latch ACT_ADD<0:15> in the PRE_LATCH, comparators in the latch sets compare the latched addresses in PRE_LATCH with addresses stored in LATCH<0:6> and the hidden latch. The hidden latch is designed to latch additional active aggressor when all of LATCH<0:6> are occupied, and malicious activation is focused on other addresses. Unlike LATCH<0:6>, a hidden latch does not have a corresponding counter. When an address does not exist in LATCH<0:6> and the hidden latch, all values of COMP<0:7> are set to zero and the R/H controller triggers the PIN signal to store the latched address in PRE_LATCH in one of the latches among LATCH<0:6> with ACT, EN, and COMP<0:7> signals. When the active addresses stored in LATCH<0> to LATCH<6> are sampled in the PRE_Latch again, a paired counter value from COUNTER<0> to COUNTER<6> is also increased. The R/H controller finally selects the row hammer address, RH_ADD, using RH_REF and COUNTER<0:6>, which are counter values. RH_ADD<0:15> is chosen among the stored address from LATCH<0:6> and the hidden latch. The refresh address, REF_ADD<0:15>, is generated with RH_ADD for row hammer mitigation and N_ADD, which is generated with address counters for a normal refresh.

Per-row hammer tracking (PRHT) is presented in Fig. 28.8.3. The R/H cells to store the number of activation executions for each WL are added with additional columns. Internal RD (IRD) and internal write (IWR) execute RMW to check and update the number of activation executions between activation and precharge operations. The IRD and IWR control signals are generated from the RMW control block shown in Fig. 28.8.1. When a WL whose address is 0x3 is activated, the IO sense amplifiers (IOSAs) read the cumulative activation counter number (0xA) from the R/H cell for the WL address and the adder updates the Cell_RD register value to 0xB by adding one to 0xA. If the Cell_RD value is larger than the Max_CNT number, then the Max_CNT number is updated to 0xB and the comparator sends the Update signal to latch sets to store the current active address from address latch0 in address latch1 for row hammer mitigation. The Cell_RD value is written in R/H cell with write driver (WTDRV) with IWR internal command. Two additional refresh commands are needed for row hammer mitigation for each WL. One refresh command is used to reset R/H cell with RSTADD_N1 as zero and next refresh conducts row hammer mitigation for adjacent rows with ADD_N2 and ADD_N0 with addresses stored in address latch 2, which are latched with the refresh for reset.

Charge loss due to a row hammer attack occurs during active and precharge operations. The magnitude of the electric field during a precharge operation can determine the amount of electron charges that are dispersed to neighboring cells. The multi-step precharge circuit, shown in Fig. 28.8.4, improves intrinsic row-hammer tolerance by creating an electric field that helps to maximize the amount of returning electron charges into a victim cell and to minimize electron charges to be dispersed into adjacent cells. In the timing diagram shown in Fig. 28.8.4, V_A is the optimized sub-WL level to minimize charge loss for the victim cell under a row-hammer attack. The mutli-step precharge control circuit shown in Fig. 28.8.4 and 28.8.1 generates MWLT, FXB0, and FXB1 signals to generate the multi-stage sub-WL level in Fig. 28.8.4.

The reduction in cell size leads to a degradation of intrinsic row-hammer tolerance and refresh retention time, which is closely related to a cell transistor reliability and power consumption. Moreover, there is a trade-off between the intrinsic row-hammer tolerance and the refresh retention time within a fixed cell size. The body bias of a cell transistor (V_{BB}) can be used to balance the two parameters by controlling the cell transistor's threshold voltage and leakage. A V_{BB} temperature-modulation circuit is proposed in Fig. 28.8.5 to maximize refresh retention time across 25 – 90°C by making the intrinsic row-hammer tolerance similar across temperature. V_{REFB} and V_{REFBH} are reference voltages generated by the reference generation amplifier. V_{BB} is determined by a feedback loop from a charge pump, which is regulated by a variable resistance controlled by the CTRL_CODE from the TEMP_CTRL block, until detector (DET) cannot detect the difference in input voltage levels. The TEMP_CTRL block sets a variable resistance value based on TEMP_CODE from the temperature sensor and the fuse information based on intrinsic row-hammer tolerance.

Figure 28.8.6 presents measurement results of the proposed schemes. Compared to a DRAM device using conventional aggressor tracking logic [5], the DRAM device with the proposed PAT logic functionally passes fifty row-hammer malicious pattern attacks even with a 66% lower intrinsic row-hammer tolerance, which makes DRAM tolerant even with technology scaling. The probability of failure is reduced when the RFM function enabled, by lowering RAAIMT values. The probability of failure with fifty row hammer malicious patterns is reduced by 90.5% using the PRHT scheme. The intrinsic row-hammer tolerance is improved by 37% with the multi-step precharge scheme compared to a conventional single-step precharge scheme [1]. V_{BB} temperature modulation improves refresh retention time by 17% at 90°C. The 16-Gb DDR5 DRAM is fabricated in a 1a-nm high-k metal-gate DRAM process; the micrograph is presented in Fig. 28.8.7.

References:

[1] K. C. Chun et al., "A 16Gb LPDDR4X SDRAM with an NBTI-tolerant circuit solution, an SWD PMOS GIDL reduction technique, an adaptive gear-down scheme and a metastable free DQS aligner in a 10nm class DRAM process." *ISSCC*, pp. 206-207, 2018.
[2] J. S. Kim et al., "Revisiting rowhammer: An experimental analysis of modern dram devices and mitigation techniques." *Intl.. Symp. on Comp. Arch.*, pp 638-651, 2022.
[3] M. Marazzi et al., "ProTRR: Principled yet Optimal In-DRAM Target Row Refresh." *IEEE Symposium on Security and Privacy*, pp 735-753, 2022.
[4] Y.-C. Bae et al., "A 1.2 V 30nm 1.6 Gb/s/pin 4Gb LPDDR3 SDRAM with input skew calibration and enhanced control scheme." *ISSCC*, pp. 44-45, 2012.
[5] P. Jattke et al., "Blacksmith: Scalable Rowhammering in the Frequency Domain." *IEEE Symposium on Security and Privacy*, pp. 716-734. 2022.

 978-1-6654-9016-0/23/$31.00

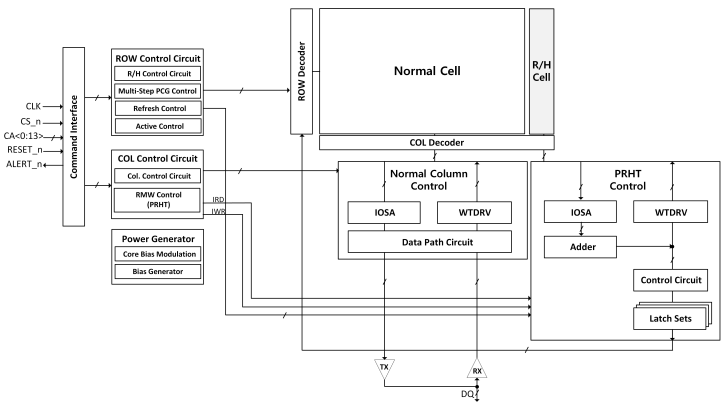


Figure 28.8.1: Block diagram of the technology scaling platform used for evaluating the proposed and conventional schemes.

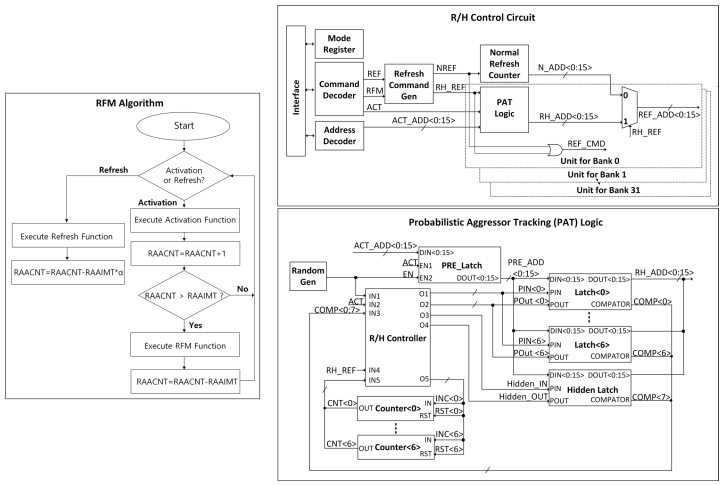


Figure 28.8.2: RFM algorithm with a flow chart to execute RFM function in a memory controller, R/H control circuit for refresh operations, and probabilistic aggressor tracking (PAT) logic for tracking row hammer addresses.

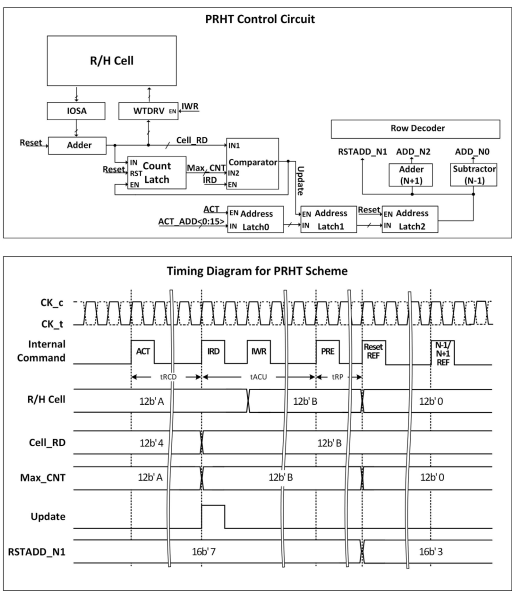


Figure 28.8.3: Control circuit (top) and timing diagram for the PRHT scheme (bottom).

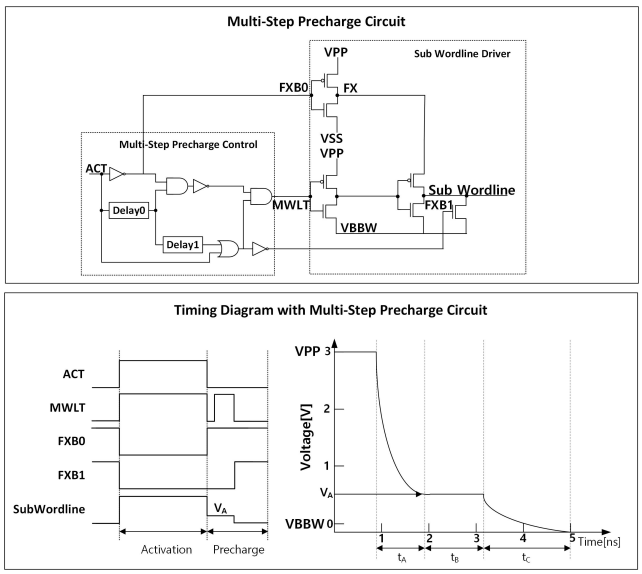


Figure 28.8.4: Circuit (top) and timing diagram (bottom) for multi-step precharge circuit.

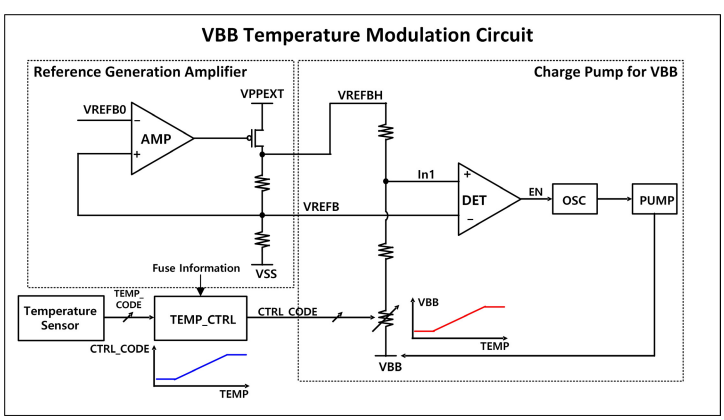


Figure 28.8.5: V_{BB} temperature modulation circuit.

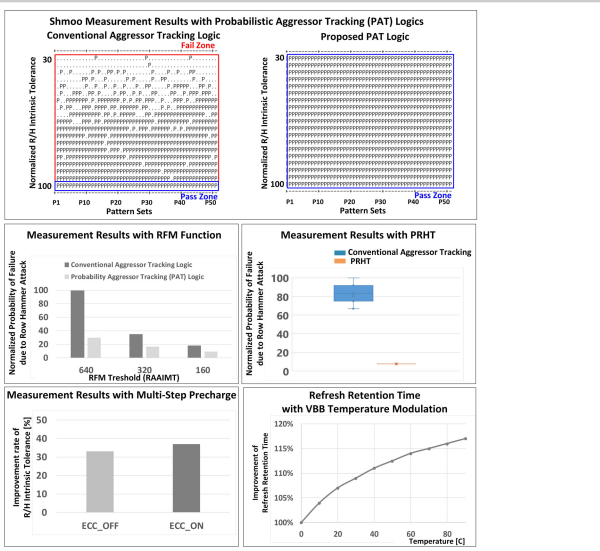


Figure 28.8.6: Measurement results for the proposed schemes: Improvements of row hammer resilience with PAT, RFM, PRHT, and multi-step precharge schemes and refresh retention time with VBB temperature modulation.

Chip summary	
Technology	1anm 5-metal DRAM HKMG process
Data Rate	6.4 Gbps/pin
Burst Length	BC8, BL16 on the fly
Number of IO	X4/X8/X16
Chip Size	7.159mm X 7.444mm

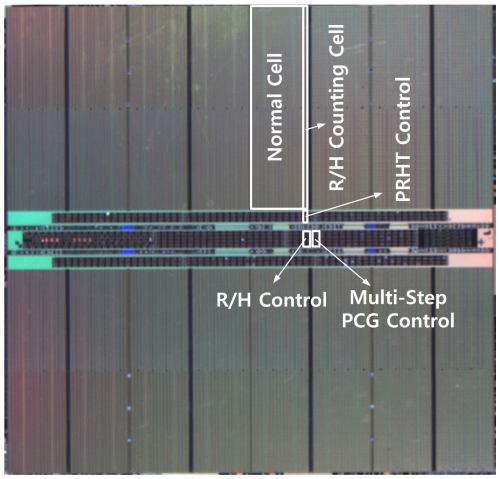


Figure 28.8.7: Chip summary and micrograph.

978-1-6654-9016-0/23/$31.00