A 3D-Stacked SRAM Using Inductive Coupling with Low-Voltage Transmitter and 12:1 SerDes

Kota Shiba[1,2], Tatsuo Omori[1,2], Kodai Ueyoshi[3], Kota Ando[4], Kazutoshi Hirose[4],
Shinya Takamaeda-Yamazaki[5], Masato Motomura[4], Mototsugu Hamada[1,2], and Tadahiro Kuroda[1,2]
Email: {shiba, omori, hamada, kuroda}@kuroda.elec.keio.ac.jp, ueyoshi.kodai.6a@ist.hokudai.ac.jp,
{ando.kota, hirose.kazutoshi, motomura}@artic.iir.titech.ac.jp, shinya@is.s.u-tokyo.ac.jp
[1]Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
[2]Graduate School of Science and Technology, Keio University, Yokohama, Japan
[3]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
[4]Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan
[5]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

***Abstract*— A 28.8-GB/s 96-MB 3D-stacked SRAM is presented. A total of eight SRAM dies, designed in a 40-nm CMOS process, are vertically stacked and connected using an inductive coupling wireless link with a low-voltage NMOS push-pull transmitter that reduces the power of the link by 45% with a 0.4-V power supply. The SRAM utilizes an inverted bit insertion scheme that compensates the degradation of the first signal, a coil termination scheme that aims to eliminate the noise of 3D inductive coupling bus, and a 12:1 SerDes. The data density of the SRAM should reach 12.3-MB/mm^3, which extends beyond that of state-of-the-art stacked DRAMs.**

Keywords—inductive coupling, ThruChip Interface (TCI), 3D integration, through-silicon via (TSV), static random access memory (SRAM)

I. Introduction

Deep neural networks (DNNs) have become wide spread in machine-learning applications. DNNs need large and high-bandwidth external memories to process large quantities of data. In conventional work, an artificial intelligence (AI) accelerator stacked on dynamic random access memory (DRAM) dies and connected by through-silicon vias (TSVs) was presented, achieving large and high-bandwidth memories [1]. However, TSVs, which are also found in high-bandwidth memory (HBM) and hybrid memory cubes (HMCs), require an additional mechanical process that results in higher manufacturing costs and worse yields [2]. Moreover, the long latency of DRAM access causes memory stalls and limits its performance [1].

To solve these problems, in this work, a 3D-stacked static random access memory (SRAM) using an inductive coupling wireless link, namely the ThruChip Interface (TCI), is proposed TCI is an inter-chip wireless communication interface between vertically-stacked chips using on-chip coils and suppresses the issues of TSVs [3]. A hundred or more on-chip coils enable high-bandwidth memories with high-speed and low-power links. In addition, TCI, unlike TSV, needs only a standard back-end-of-line (BEOL) CMOS process to create the coils and transceiver circuits, which leads to low costs and a good yield. Whereas the long-latency DRAM is problematic, the low-latency SRAM plays a key role in DNNs and improves their performance [4]. Therefore, the 3D-stacked SRAM module using TCI is a promising effective technology for achieving a low-power and high-performance AI accelerator.

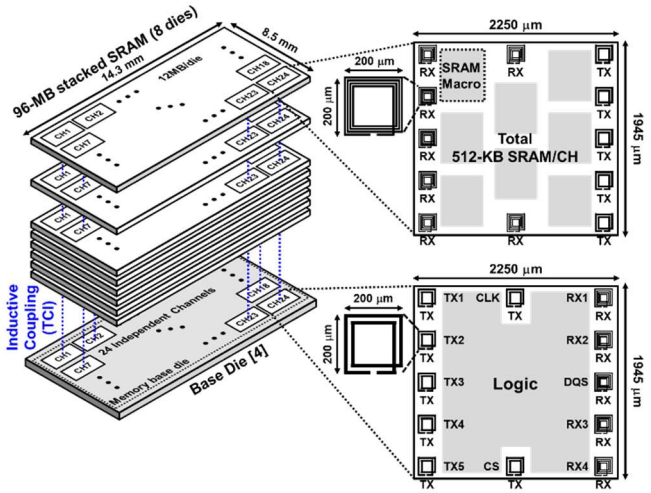

Fig. 1. 3D-Stacked SRAM overview.

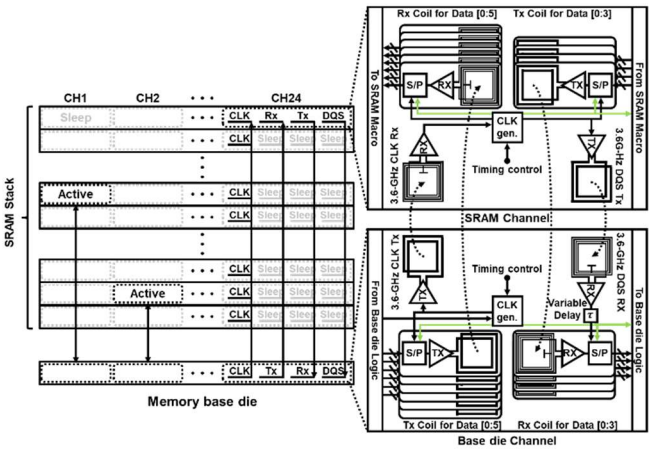

Fig. 2. 3D-Stacked SRAM block diagram.

Edge and mobile devices have strict requirements for power efficiency that are directly linked to its overall performance. Besides, this work proposes a new type of TCI transmitter operating at a lower voltage than conventional ones. The transmission power of the new transmitter is reduced by 45%. This work also proposes a new termination scheme and inverted bit insertion scheme to solve a problem caused by multi-stacked coils and to compensate for the attenuation of the received signal during the transmitters' turn-on sequence, respectively.

978-1-7281-3320-1/20/$31.00 ©2020 IEEE

Signal	0	1	2	3	4	5	6	7	8	9	10	11
CLK	1 0	1 0	1 0	1 0	1 0	1 0	1 0	1 0	1 0	1 0	1 0	1 0
CS	1	1	1	1	1	1	0	0	0	0	0	0
TX1	$\overline{BA0}$	BA0	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9
TX2	$\overline{BA1}$	BA1	A10	A11	A12	A13	A14	A15	A16	DI30	DI31	R/W
TX3	$\overline{BA2}$	BA2	DI0	DI1	DI2	DI3	DI4	DI5	DI6	DI7	DI8	DI9
TX4	1	0	DI10	DI11	DI12	D13	D14	D15	DI16	DI17	D18	D19
TX5	1	0	DI20	DI21	DI22	DI23	DI24	DI25	DI26	DI27	DI28	DI29
DQS	0 1	0 1	0 1	0 1	0 1	0 1	0 1	0 1	0 1	1	1	1
RX1	$\overline{DO0}$	DO0	DO1	DO2	DO3	DO4	DO5	DO6	DO7	Not Used		
RX2	$\overline{DO8}$	DO8	DO9	DO10	DO11	DO12	DO13	DO14	DO15			
RX3	$\overline{DO16}$	DO16	DO17	DO18	DO19	DO20	DO21	DO22	DO23			
RX4	$\overline{DO24}$	DO24	DO25	DO26	DO27	DO28	DO29	DO30	DO31			

BA[0:2] : Bank Address (Chip Number)
A[0:16] : Address (SRAM Macro Address)
DI[0:31] : Write Data
DO[0:31] : Read Data

Fig. 3. Packet format for 3D SRAM access.

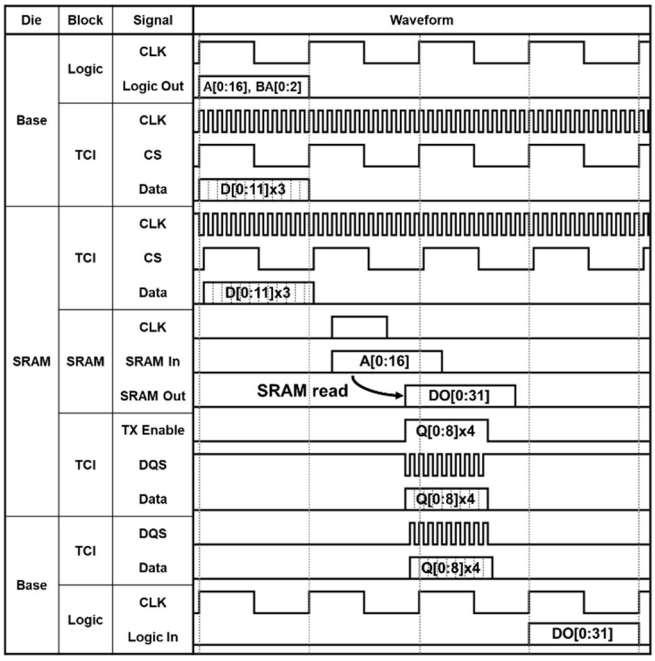

Fig. 4 Time chart of a single read access.

This work is organized as follow. Section II describes the proposal of a 3D-stacked SRAM module using inductive coupling technology with an overview, block diagram, packet format, and time chart. Section III describes the inductive coupling link for a 3D SRAM with a termination scheme that solves the problem in which sleep-transmitters' open coils deteriorate received signals, a new transmitter operating at a lower voltage than conventional ones, and an inverted bit insertion scheme to compensate for the received signal attenuation, where the performance of the 3D SRAM is discussed. Section IV concludes this paper.

II. 3D-STACKED SRAM

A. Overview

Fig. 1 illustrates the proposed 28.8-GB/s 96-MB 3D-stacked SRAM module using inductive coupling. A base die [4] and stacked SRAM dies are wirelessly connected by TCI. Each SRAM die is thinned down to 8 μm and vertically stacked [5]. Each die is composed of 24 channels, and each

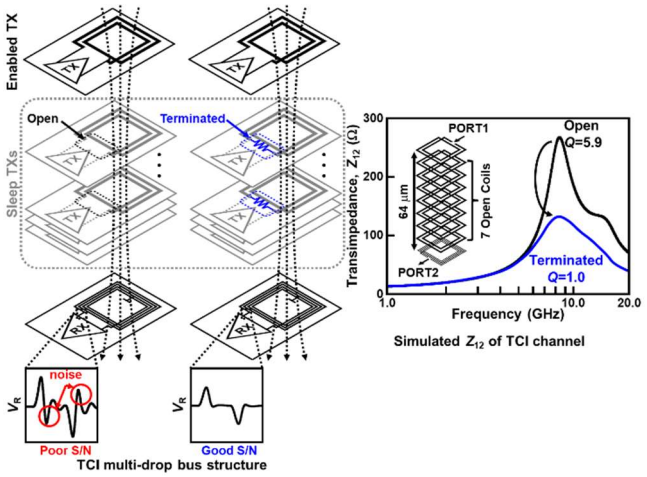

Fig. 5. Termination scheme.

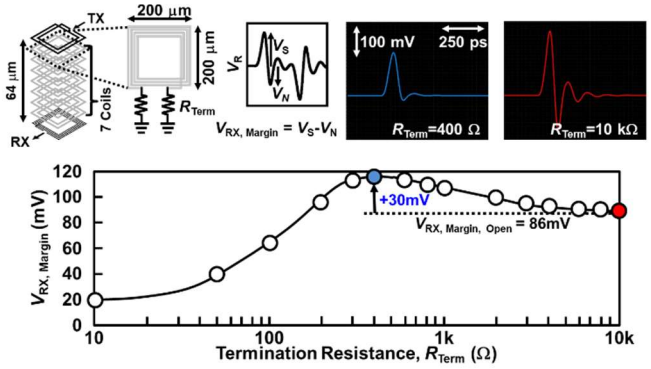

Fig. 6. Process variation tolerance of termination resistance.

channel in an SRAM die has a 512-KB capacity SRAM and is 2250 x 1945 μm^2. Each channel has 12 TCI transceiver circuits and 200 x 200 μm^2 coils to cover the maximum communication distance of 64 μm. Each TCI link has the capability of communicating at a data-rate of 3.6 Gbps.

B. Block Diagram

Fig. 2 depicts the 3D-stacked SRAM module block diagram. Each channel of the base die can access any stacked SRAM die and can sleep independently from channel to channel. This architecture enables extendibility and low-power operation. The data communication is carried out by using a source-synchronous manner, and SRAM macros operate at a frequency of 300 MHz, obtained by dividing the 3.6-GHz system clock distributed by TCI bus. A 12:1 serial-to-parallel (S/P) and parallel-to-serial (P/S) conversion system is utilized to exchange data in the TCI channel between an SRAM macro or a base die logic in the proposed SRAM, because adopting 12:1 Serializer/Deserializer (SerDes) reduces TCI power consumption and required layout area by a factor of 12 [6]. The read and write latency including TCI, S/P, and P/S process time is three cycles at 300 MHz. The phase difference between the clock and data is controlled by the clock generator allocated in each channel.

C. Packet Format

Fig. 3 shows the packet format. Each packet is a 12-bit unit, because 12:1 P/S or S/P conversion is used as the TCI-SRAM interface. To read from or write to an SRAM macro, a system clock signal (CLK), chip selection signal (CS), strobe signal (DQS), bank address (BA[0:2]) to designate an SRAM die,

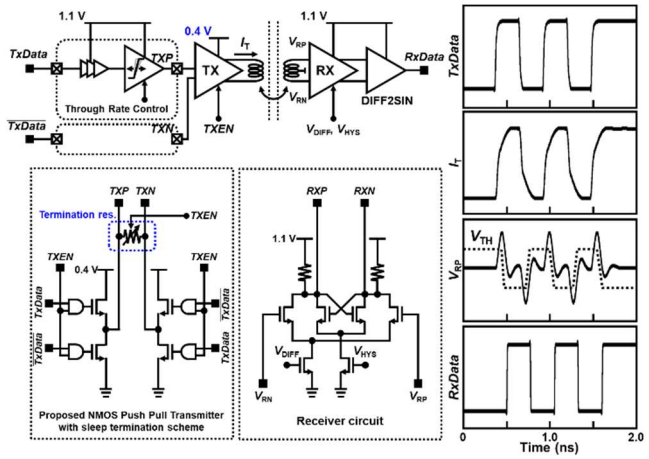

Fig. 7. Inductive coupling transceiver and diagram with NMOS push-pull transmitter.

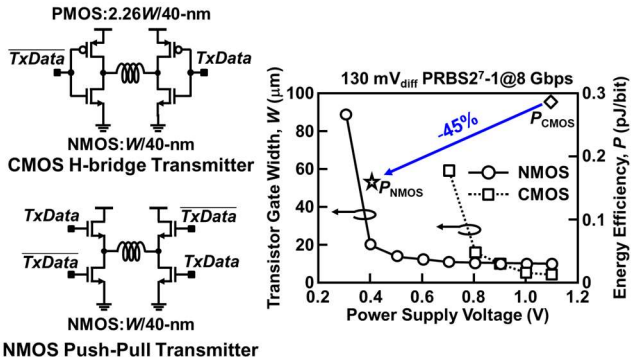

Fig. 8. Comparison of CMOS H-bridge and NMOS push-pull transmitter.

SRAM address (A[0:16]), read/write flag (R/W), and read/write data (DO[0:31]/DI[0:31]) are required. These data are assigned to the 12-bit-unit packet. The first bit of the packet is the inversion of the second bit following an inverted bit insertion scheme, which is discussed in detail in Section III. This packet format enables three-cycle SRAM latency.

D. Time Chart

Fig. 4 is a time chart of a base die and SRAM die for a single SRAM read operation. First, five transmitters in a base die send a CLK, CS, read address, bank address, and read flag to the SRAM dies after serializing those from a base die logic. Then, only one SRAM die is activated following the bank address. The received address is deserialized to be sent to an SRAM macro and the data are read from the SRAM macro within one cycle. The read data are transmitted from the SRAM die to the base die through five TCI channels including a DQS after serialization. Finally, the received read data are deserialized and sent back to the logic. Therefore, 3D SRAM latency is three cycles. Moreover, a TCI operation delay is almost constant, irrespective of the distance of the base die to a selected SRAM die, and the three-cycle latency is uniform over the eight SRAM dies.

III. Inductive Coupling Link

In this section, we discuss three key inductive coupling techniques. First, we introduce a scheme terminating sleep transmitter's open coils to suppress the deterioration of the received signal. Second, we propose a new NMOS push-pull transmitter that operates at 0.4 V and reduces power consumption by 45%. Third, we propose a scheme in which

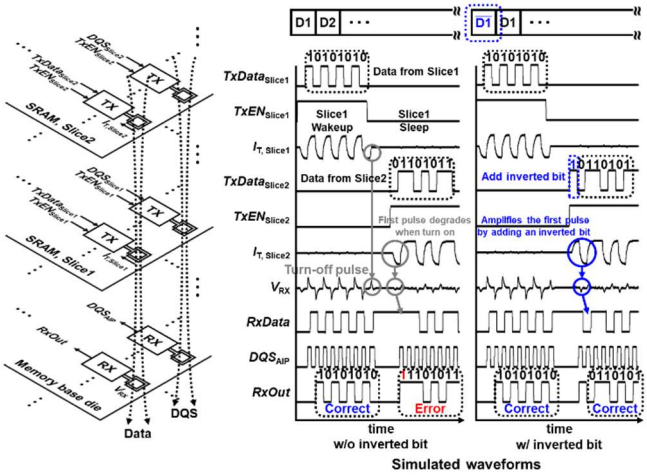

Fig. 9. Inverted bit insertion scheme.

the inverted bit is inserted at the beginning of the packet to compensate the attenuation of the signal during the turn-on sequence. Finally, our 3D-stacked SRAM performance is compared with that of conventional work.

A. Termination Scheme

Fig. 5 shows a termination scheme for the inductive coupling multi-drop bus communication. In the proposed module, when an SRAM die sends data back to the base die in the uplink, the magnetic field reaches the receiver (RX) coil from the transmitter (TX) coil after passing through several sleep-TX coils. When these sleep-TX coils are at an open state, the transimpedance from an 8th SRAM die's TX coil to a base die's RX coil has high-Q characteristics. Whereas wireless power transfer techniques utilize this mechanism [7], TCI is a baseband communication interface; therefore, the characteristics cause ringing in the RX coil, which results in worse power efficiency and low speed due to the bad signal-to-noise-ratio (S/N) and intersymbol interference (ISI), respectively [8]. To suppress the ringing, we propose a scheme terminating sleep-TX coils: termination scheme. As seen in Fig. 5, terminating sleep-TX coils with resistance can dump the transimpedance and suppress the ringing in RX, which enables reliable high-speed low-power communication.

Fig. 6 shows the simulation results of received pulse amplitude margin $V_{RX,Margin}$ against the termination resistance. The received pulse amplitude margin $V_{RX,Margin}$ is defined as the first received pulse amplitude (V_S) minus the second one (V_N), where the threshold voltage of a hysteresis comparator receiver is set between V_S and V_N. When the termination resistance is 400 Ω, $V_{RX,Margin}$ is 116 mV, which is 30-mV improvement from one without a termination scheme. Furthermore, as $V_{RX,Margin}$ is changing slowly against the termination resistance around 400 Ω, the termination scheme is tolerant to the process variation.

B. NMOS Push-Pull Transmitter

Fig. 7 illustrates the TCI transceiver block diagram, schematic, and waveform with the new NMOS push-pull transmitter. Though low-voltage open-drain NMOS transmitters have been reported, these types of transmitters are pulse-modulated and have issues related to low received signals, low data rates, and large switching noise [9], [10]. Therefore, this work proposes a low-voltage NMOS push-pull transmitter that can maintain the same received voltage amplitude, data rate, and switching noise as the conventional

	Nvidia Pascal [2]	This Work	This Work at 10 nm
Memory Type	HBM2	SRAM	SRAM
Capacity/Module	4 GB	96 MB	432 MB
Bandwidth	180 GB/s	28.8 GB/s	-
I/O Energy Consumption	-	1.5 pJ/bit/pin	-
Volume	79.5 mm^3	7.8 mm^3	7.8 mm^3
Capacity Density	50 MB/mm^3	12.3 MB/mm^3	55.4 MB/mm^2
Technology Node	20-nm DRAM	40-nm CMOS	10-nm CMOS

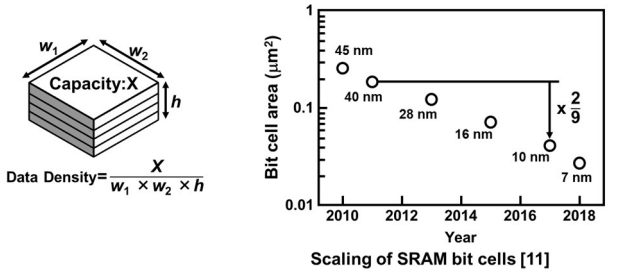

Fig. 10. Performance comparison.

full-swing CMOS transmitter with lower power. The TX coils are designed with low resistance to compensate for the reduction of the flowing current by the reduction of the transmitter power supply voltage. In addition, pre-drivers are operating at a standard voltage of 1.1 V, and NMOS transistors in a driver are highly over-driven to reduce on-resistance. In the event that the low-resistance system causes a self-resonance issue, the pre-drivers are designed to be able to control the slew-rate. When the transmitter is in a sleep state, all four NMOS transistors of the output stage are disabled and a termination scheme is enabled, as mentioned in the previous subsection. On the other hand, when the transmitter is enabled, the termination scheme is disabled and the coil is driven only by the transmitter without any of the drawbacks caused by the termination scheme.

Fig. 8 shows comparisons between the new NMOS push-pull transmitter and the conventional CMOS H-bridge transmitter. The right-hand graph plots the power supply voltage of the transmitters against the transistor gate width W of the output stage required to maintain the received pulse voltage of 130 mV. If the power supply voltage is down, the required transistor width W becomes drastically larger at one voltage point. Whereas the voltage point of the CMOS transmitter is about 0.8 V, that of the NMOS transmitter is 0.4 V. Furthermore, the required transistor width W of the NMOS transmitter increases more slowly than that of the CMOS transmitter. Thus, the NMOS transmitter is capable of operating at low voltage without any area overhead and additional buffer. The proposed NMOS transmitter achieves a 45% power reduction at 0.4 V compared with that of the CMOS transmitter.

C. Inverted Bit Insertion Scheme

The above-proposed NMOS transmitter generates halved pulse noise (turn-off pulse) in the receiver, as seen in Fig. 9, when the transmitter becomes sleep and the current flowing into a TX coil transits from $-I_{T,max}$ to 0 mA. The turn-off pulse generated by the turn-off of Slice1 potentially flips the received data and causes critical errors in the system. For example, if Slice1 transmits data $TxData_{Slice1}$ and becomes sleep, where the last data is "0," the hysteresis comparator receiver holds *RxData* of "1" flipped by the turn-off pulse of Slice1 and awaits the data from Slice2. When the transmitter of Slice2 becomes activated and transmits the first data of "0," the current flow $I_{T,Slice2}$ transits only from 0 mA to $-I_{T,max}$, where it normally transits either from $-I_{T,max}$ to $+I_{T,max}$ or from $+I_{T,max}$ to $-I_{T,max}$. Therefore, the current flow transition is halved and the received pulse V_{RX} is also halved (turn-on pulse). As a result, this turn-on pulse can not flip *RxData* of "1" held by the receiver, which leads to bit error.

In this work, we propose an inverted bit insertion scheme that inserts the inverted bit in front of the first bit. By utilizing this scheme, the current $I_{T,Slice2}$ flows against the first data before transmitting it. This is how the scheme compensates the received pulse V_{RX} attenuation during turn-on. For the above-mentioned example, as the scheme lifts up the current $I_{T,Slice2}$ to $+I_{T,max}$ beforehand, the current transmits from $+I_{T,max}$ to $-I_{T,max}$ and the received pulse V_{RX} is amplified. Thus, the transmitter can correctly flip *RxData* to "0".

As described in Section II and seen in Fig. 3, the first bit of the packet is the inversion of the second bit, which helps the inductive coupling links avoid a bit error caused by the turn-on pulse.

D. Simulation Results and Performance Comparison

Fig. 10 summarizes the performance comparison of the proposed 3D-stacked SRAM module and the 3D-stacked DRAM module [2] on the basis of the simulation results. The SRAM die is composed of 24 512-KB capacity channels and the module has 8 12-MB SRAM dies, totaling 96 MB. 24 channels can access the 32-bit SRAM at a system clock of 300 MHz, and the total bandwidth is 28.8 GB/s. This work defines the data density as a performance indicator, which is memory capacity per volume. The proposed 96-MB module has a 14.3 x 8.5 mm^2 area and 64-um thickness and its data density is 12.3 MB/mm^3. If this module is designed in a comparable process node to the DRAM module [2], its data density will be 55.4 MB/mm^3, where a 20-nm DRAM process and 10-nm CMOS process appeared at the product level in the same year of 2017 and are considered to be comparable process nodes [11]. Therefore, the proposed SRAM module achieves a data density beyond that of a state-of-the-art DRAM module.

IV. Conclusion

A 28.8-GB/s 96-MB 3D-stacked SRAM module is proposed. The proposed low-voltage NMOS push-pull transmitter achieves a 45% power reduction compared with that of the conventional CMOS transmitter. The proposed termination scheme amplifies the first received pulse. The proposed inverted bit insertion scheme compensates for the attenuation of the received pulse during the turn-on sequence of the transmitter. The 3D-stacked SRAM module with the proposed transmitter, inverted bit insertion scheme, and termination scheme will achieve a data density beyond that of a state-of-the-art DRAM module when a comparable process node is employed.

Acknowledgment

This work was supported by JST ACCEL Grant Number JPMJAC1502, Japan. The authors thank T. Miyata and J. Kadomoto for their invaluable support.

References

[1] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "Tetris: Scalable and efficient neural network acceleration with 3D memory,"

ACM SIGARCH Comput. Archit. News, vol. 45, no. 1, pp. 751–764, April 2017.

[2] D. Foley, and J. Danskin, “Ultra-performance pascal GPU and NVLink interconnect,” IEEE Micro, vol. 37, no. 1, pp. 250-260, 1 2017.

[3] D. Ditzel, T. Kuroda, and S. Lee, “Low-cost 3D chip stacking with ThruChip wireless connections,” in Proc. IEEE Hot Chips Symp. (HCS), 1–37, August 2014.

[4] K. Ueyoshi, et al., “QUEST: Multi-purpose log-quantized DNN inference engine stacked on 96-MB 3-D SRAM using inductive coupling technology in 40-nm CMOS,” IEEE J. Solid-State Circuits, vol. 54, no. 1, pp. 186-196, January 2019.

[5] Y. S. Kim, et al., “Ultra thinning down to 4-um using 300-mm wafer proven by 40-nm node 2Gb DRAM for 3D multi-stack WOW applications,” IEEE Symp. VLSI Technology, 2014.

[6] N. Miura, et al., "A High-Speed Inductive-Coupling Link with Burst Transmission," IEEE J. Solid-State Circuits, vol. 44, no. 3, pp. 947-955, March 2009.

[7] A. G. Pelekanidis, A. X. Lalas, N. V. Kantartzis, and T. D. Tsiboukis, “Optimized wireless power transfer schemes with metamaterial-based resonators,” International Workshop on Antenna Technology, 2017.

[8] L.-C. Hsu, et al., “Analytical ThruChip Inductive Coupling Channel Design Optimization,” ASP-DAC, pp. 731-736, 2016.

[9] N. Miura et al., "A 0.55 V 10 fJ/bit inductive-coupling data link and 0.7 V 135 fJ/cycle clock link with dual-coil transmission scheme," IEEE J. Solid-State Circuits (JSSC), vol. 46, no. 4, pp. 965-973, April 2011.

[10] S. Hasegawa, J. Kadomoto, A. Kosuge, and T. Kuroda, "A 1 Tb/s/mm2 Inductive-Coupling Side-by-Side Chip Link," IEEE European Solid-State Circuits Conference (ESSCIRC), pp. 469-472, September 2016.

[11] T. Song, et al., “A 10 nm FinFET 128 Mb SRAM with assist adjustment system for power, performance, and area optimization,” IEEE J. Solid-State Circuits, vol. 52, no. 1, pp. 240-249, January 2017.