A 96-MB 3D-Stacked SRAM Using Inductive Coupling With 0.4-V Transmitter, Termination Scheme and 12:1 SerDes in 40-nm CMOS

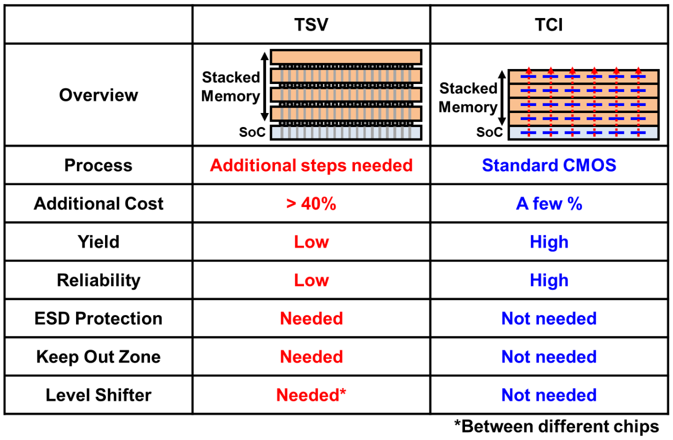
Kota Shiba, *Student Member, IEEE*, Tatsuo Omori, Kodai Ueyoshi, *Member, IEEE*, Shinya Takamaeda-Yamazaki, *Member, IEEE*, Masato Motomura, *Senior Member, IEEE*, Mototsugu Hamada, *Member, IEEE*, and Tadahiro Kuroda, *Fellow, IEEE*

***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 35% with a 0.4-V power supply. The SRAM utilizes an inverted bit insertion scheme that compensates for the degradation of the first transmitted bit, a coil termination scheme that aims to eliminate the ringing of 3D inductive coupling bus, and a 12:1 SerDes that minimizes power consumption and area overhead in inductive coupling channels. Low-power, large-capacity, 3-cycle latency 3D-stacked SRAM for a DNN accelerator is achieved with the combination of these techniques to serve as a replacement of 3D-stacked DRAM. The performance of the proposed 3D-SRAM is compared with HBM DRAM and achieves more than 50% lower energy consumption. The scaling scenario of the SRAM module is discussed in light of the scaling of the inductive coupling technology and logic process.**

***Index Terms*—3D integration, 3D memory architecture, deep neural networks (DNNs), inductive coupling, static random access memory (SRAM), through silicon via (TSV), ThruChip interface (TCI).**

I. Introduction

DEEP neural networks (DNNs) have become widespread in machine-learning applications. DNNs need large and high-bandwidth memories to process large quantities of data [1]. In conventional work, dynamic random access memory (DRAM) dies are stacked on an artificial intelligence (AI) accelerator and connected by through-silicon vias (TSVs) to provide the required large and high-bandwidth memories [2]. 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 [3], [4]. Moreover, the long latency of DRAM access causes computation stalls which limits its performance [2].



	TSV	TCI
Overview	Stacked Memory SoC	Stacked Memory SoC
Process	Additional steps needed	Standard CMOS
Additional Cost	> 40%	A few %
Yield	Low	High
Reliability	Low	High
ESD Protection	Needed	Not needed
Keep Out Zone	Needed	Not needed
Level Shifter	Needed*	Not needed

*Between different chips

Fig. 1. Comparison of TSV and TCI.

To solve these problems, a solution with 3D static random access memory (SRAM) stacked on a DNN inference engine [5], [6] using an inductive coupling wireless link, namely the ThruChip Interface (TCI), has been proposed [7]. TCI is an inter-chip wireless communication interface between vertically stacked chips using on-chip coils which circumvents the issues of TSVs [8]–[11]. Hundreds of on-chip coils enable high-bandwidth memories with high-speed and low-power links. In addition, as seen in Fig. 1, TCI, unlike TSV, needs only a standard complementary metal-oxide semiconductor (CMOS) process to create the coils and transceiver circuits, which leads to low costs and good yield [12]. TCI has high reliability thanks to its wireless nature. Furthermore, while TSV needs electrostatic discharge (ESD) protection circuits, keep-out zone (KOZ), and level shifter circuits, TCI requires none of them, and its coils can be placed even directly above circuits.

Manuscript received May 11, 2020; revised October 14, 2020; accepted November 10, 2020. Date of publication November 26, 2020; date of current version January 12, 2021. This work was supported by JST ACCEL, Japan, under Grant JPMJAC1502. This article was recommended by Associate Editor J.-O. Klein. *(Corresponding author: Kota Shiba.)*

Kota Shiba, Tatsuo Omori, Mototsugu Hamada, and Tadahiro Kuroda are with the Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan (e-mail: shiba@kuroda.t.u-tokyo.ac.jp).

Kodai Ueyoshi is with the Department of Electrical Engineering, ESAT-MICAS, KU Leuven, 3001 Leuven, Belgium (e-mail: kodai.ueyoshi@kuleuven.be).

Shinya Takamaeda-Yamazaki is with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan (e-mail: shinya@is.s.u-tokyo.ac.jp).

Masato Motomura is with the Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 152-8550, Japan (e-mail: motomura@artic.iir.titech.ac.jp).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2020.3037892.

Digital Object Identifier 10.1109/TCSI.2020.3037892

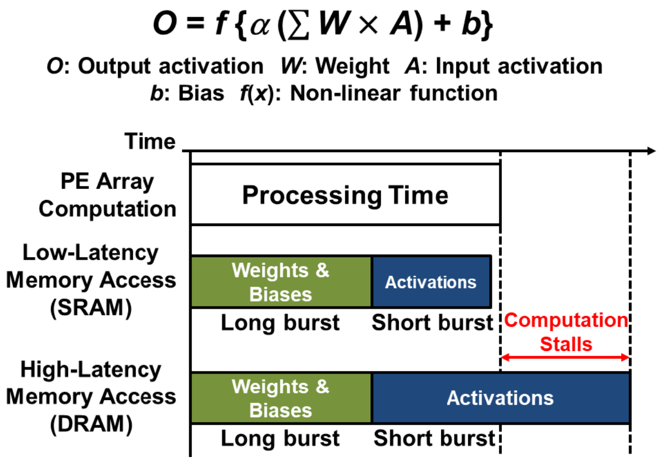


Fig. 2. Comparison between DRAM and SRAM in DNNs [5], [6].

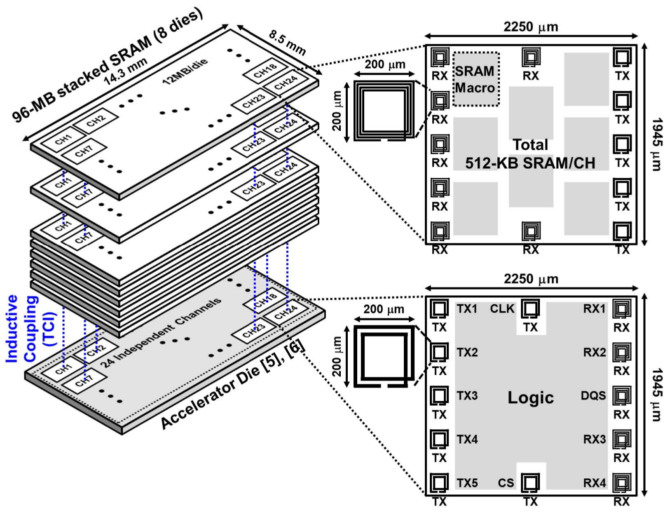


Fig. 3. Overview of 3D-stacked SRAM on DNN accelerator [7].

Whereas the long-latency DRAM is problematic in many DNN applications, the low-latency SRAM plays a key role in DNNs in improving its performance [5], [6]. The computation in DNNs is represented by the equation in Fig. 2 using their weight W, input activation A, and bias b. While the weights and biases need long burst access, for which the memory latency is almost the same in DRAM and SRAM, the activations tend to need short burst access especially for low-bit quantization, for which DRAM and SRAM latencies are drastically different. As illustrated in Fig. 2, while the long-latency DRAM access for retrieval of activations tends to take longer than the processing time resulting in computation stalls, the short-latency SRAM access is within the processing time. As a result, it was reported in [6] that the long DRAM latency adversely affects performance [TOPS] in different types of DNNs, and that utilizing an SRAM instead of a DRAM solution improves DNN performance. In particular, the LeNet benchmark performance with 3D-SRAM is more than twice as high as that with 3D-DRAM. Therefore, the 3D-stacked SRAM module using TCI is a promising technology for achieving a low-power, low-cost and high-performance AI accelerator.

Edge and mobile devices have strict requirements for power efficiency, which are directly linked to their overall performance. TCI has previously been demonstrated as a low power interface technology due to its wireless nature. Besides, this work proposes a new type of TCI transmitter operating at a lower voltage than that of conventional ones. The transmission power of the new transmitter is reduced by 35% using a 0.40V supply. Its operation between stacked dies was experimentally confirmed in this work.

This work also proposes a new termination scheme to solve a problem caused by multi-stacked coils. In addition, this paper presents an inverted bit insertion scheme to compensate for the reduced transmission strength of the first transmitted bit after the transmitter's turn-on sequence.

This paper is an extended version of our previous conference paper [7] and adds extra description of the 3D-SRAM and the newly proposed transmitter and termination scheme, along with experimental results and performance discussion. This work is organized as follows. Section II describes the proposal of a 3D-stacked SRAM module using an inductive coupling technology with an overview, power distribution, block diagram, packet format, and timing diagram. Section III describes the inductive coupling link for the 3D-SRAM with a termination scheme that solves the problem where an open coil of a transmitter in sleep state degrades received signals, a new transmitter operating at a lower voltage than conventional ones, and an inverted bit insertion scheme to compensate for the transmission weakness in the first transmitted bit. Section IV discusses the experimental results, performance of the 3D-SRAM, and a scaling scenario of the module. Section V concludes this paper.

II. 3D-Stacked SRAM Using Inductive Coupling

A. Overview

Fig. 3 illustrates the proposed 28.8-GB/s 96-MB 3D-stacked SRAM module using inductive coupling. An accelerator die [5], [6] and stacked SRAM dies are wirelessly connected by TCI. Each SRAM die is thinned down to 8 μm and vertically stacked [13], [14]. Each die is composed of 24 channels, and each channel in an SRAM die has a 512-KB capacity SRAM and measures 2250 x 1945 μm^2. The maximum operating frequency is 300 MHz, which is limited by the SRAM access path. Each channel has 12 TCI transceiver circuits and 200 x 200 μm^2 coils to cover the maximum communication distance of 64 μm across 8 stacked SRAM dies. Each TCI link has the capability of communicating at a data rate of 3.6 Gbps. The channel pitch is 400 μm to minimize channel-to-channel crosstalk [15].

B. Inductive Coupling Multi-Drop Bus

The accelerator die and SRAM dies in the module are wirelessly connected through an inductive coupling, multi-drop bus. The inductive coupling, multi-drop bus has the advantage of high energy-efficiency. Fig. 4 shows comparisons of the multi-drop TSV and multi-drop TCI busses in power, delay,

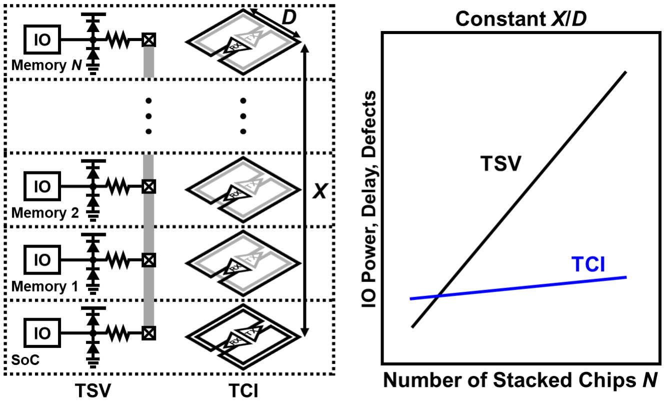


Fig. 4. Comparison of TSV and TCI in multi-drop bus.

and defects. According to the figure, all of these parameters of the multi-drop TSV bus increase rapidly in proportion to the number of stacked chips, as the capacitance of the transceivers, ESD protection circuits and the paths of the TSV and μ-bump increase. On the other hand, the capacitance in multi-drop TCI increases slowly against the number of stacked dies because TCI utilizes small coupling between coils (coupling coefficient <0.2) which reduces the effective loading of each receiver on the bus. In addition, by adjusting the coil dimensions as N increases, the transmitter power can be kept constant such that the only power increase of the multi-drop TCI bus is from receiver circuits, each of which consumes 1/4 the power of the transmitter. Therefore, power consumption of the inductive coupling, multi-drop bus increases quite slowly against the number of stacked dies. The delay is also fairly constant over different numbers of dies, as small coupling between coils reduces the effective loading and the magnetic field penetrates through chips at the speed of light. Therefore, this work achieves reduced IO power with a multi-stacked SRAM module using an inductive coupling, multi-drop bus.

C. Power Distribution

While data communication between stacked chips is conducted by using inductive coupling technology, power to stacked SRAM dies is supplied through bump-less TSVs [14]. While this solution does not fully eliminate the use of TSVs, there are some important differences between TSVs for data communication and TSVs for power distribution.

Table I summarizes the comparison between signal TSVs and power TSVs. Firstly, while signal TSVs are high-aspect ratio and small in diameter to reduce parasitic capacitance load and area overhead, power TSVs are low-aspect ratio and large in diameter to reduce parasitic resistance and the power supply voltage (IR) drop. High-aspect-ratio TSVs are generally difficult to manufacture and have worse manufacturing yield [16], [17]. Therefore, power TSVs have better manufacturability than signal TSVs.

Secondly, the number of signal TSVs per net is one or two in order to keep total number of TSVs small. On the other hand, many power TSVs are placed across the whole chip in parallel to suppress IR drop. One of the most common malfunctions

TABLE I
COMPARISON OF SIGNAL TSVS AND POWER TSVS

	Signal TSV	Power TSV
Size	Small	Large
Yield	Low	High
# of pins per net	A few	Many across chip
Open failure	Fatal problem	No problem

which occurs in TSVs is open failure, which makes the TSV's connection open. Whereas open failure causes fatal errors on signal TSVs, it rarely causes problems on power TSVs thanks to redundantly placed TSVs.

Thus, it is vital to replace signal TSVs with the high-yield and high-reliability TCI since signal TSVs have low yield and their open failures cause fatal problems. While highly doped silicon via (HDSV) will provide a low-cost TSV-less power distribution solution, the technology is still in an early investigation stage [18], [19]. On the other hand, existing wireless power transfer technologies have low area and power transfer efficiencies [20]. As a result, TSVs are utilized for power distribution in this work.

D. Block Diagram

Fig. 5 depicts the 3D-stacked SRAM module block diagram. Each channel of the accelerator die can access any stacked SRAM die and can enter into sleep mode independently from the other channels. This architecture enables extendibility and low-power operation. The data communication is carried out in a source-synchronous manner, and SRAM macros operate at a frequency of 300 MHz obtained by dividing down the 3.6-GHz system clock distributed using the TCI bus. A 12:1 serial-to-parallel (S/P) and parallel-to-serial (P/S) conversion system is utilized to exchange data between a TCI channel and its associated SRAM macro or accelerator die logic in the proposed SRAM module. Because the TCI transceiver power consumption is independent of data rate, adopting a 12:1 serializer/deserializer (SerDes) to increase the channel data rate reduces per-bit TCI power consumption and layout area by a factor of 12 as seen in Fig. 6 [21]. While the area overhead is as large as 216%, which reduces the area available for SRAM macros, in the case of many parallel data links without SerDes, the area overhead is as low as 18% in the case of 12:1 serial data links with SerDes. In addition, the area and energy overhead for the SerDes circuits are low compared to the IO area. The read latency including the 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.

E. Packet Format

Fig. 7 shows the packet format. Each packet is a 12-bit unit, because 12:1 P/S or S/P conversion is used in the TCI-SRAM interface. There are seven TCI downlinks to transmit a system clock signal (CLK), a chip select signal (CS),

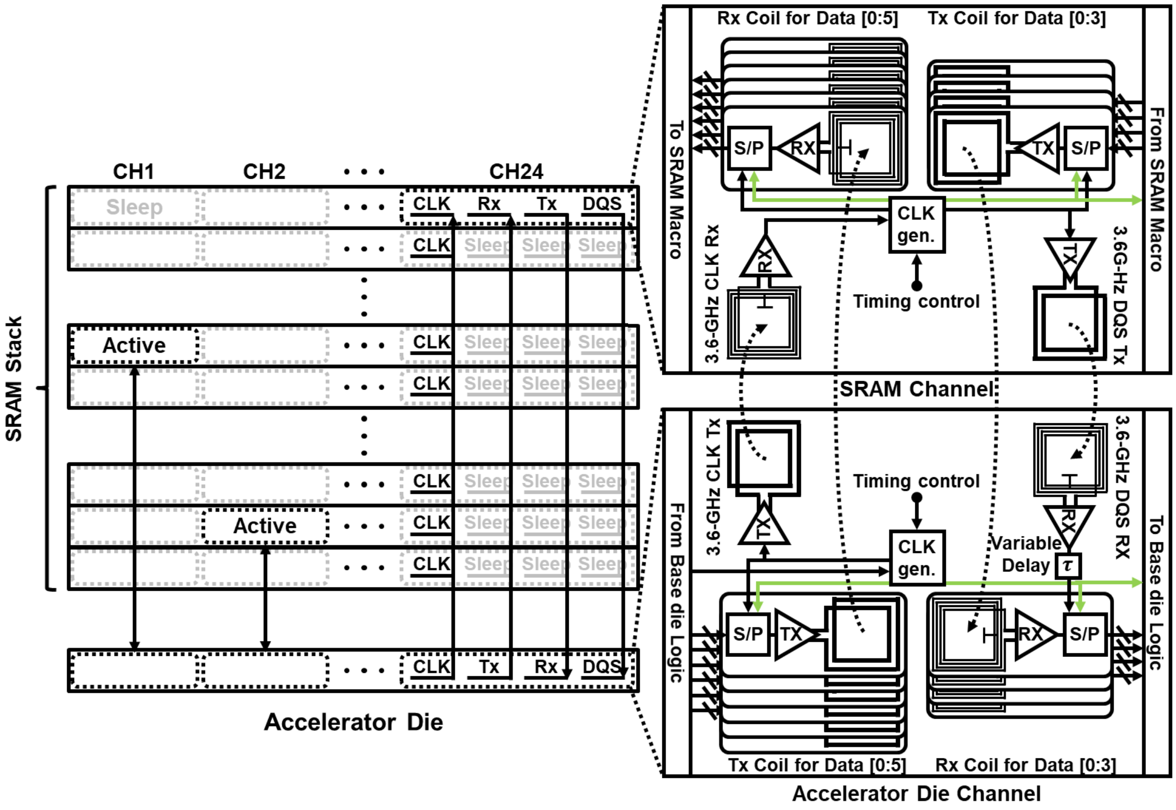


Fig. 5. 3D-stacked SRAM block diagram [7].

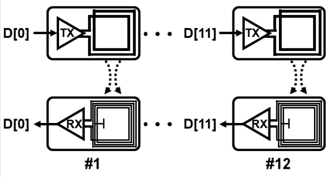
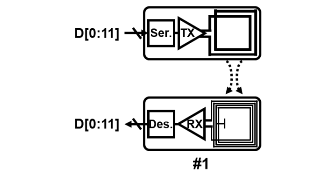
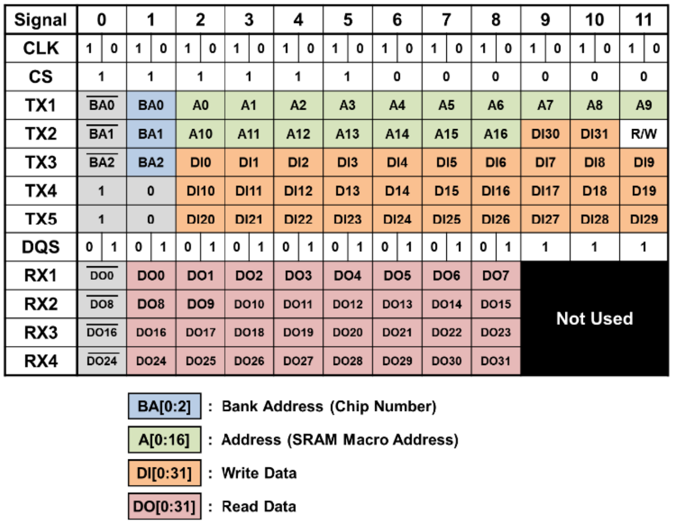
12 Parallel data links w/o SerDes	Serial data link w/ SerDes
 D[0] TX • • • D[11] TX D[0] RX D[11] RX #1 #12	
216% I/O area overhead	18% I/O area overhead
-	1/12 less power

Fig. 6. Comparison of 12 parallel data links without SerDes and serial data link with SerDes.



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. 7. Packet format for 3D SRAM access [7].

bank address (BA[0:2]), SRAM address (A[0:16]), write data (DI[0:31]) and a read/write flag (R/W), and five TCI uplinks to transmit a strobe signal (DQS) and read data (DO[0:31]), using 12-bit packets. Therefore, each channel has a total of 12 TCI links. The first bit of the packet is the inversion of the second bit to implement an inverted bit insertion scheme, which is discussed in detail in Section III. This packet format enables a three-cycle SRAM latency.

F. Timing Diagram

Fig. 8 is a timing diagram of an accelerator die and SRAM die for a single SRAM read operation. First, five transmitters in an accelerator die send a CLK, CS, read address, bank address, and read flag to all SRAM dies after serializing the signals from the accelerator die logic. Then, only one SRAM die is activated based on the bank address. The received address is deserialized and sent to the target SRAM macro, from which data are read within one cycle. The read data are transmitted from the SRAM die to the accelerator die through five TCI links together with a DQS after serialization. Finally, the received read data are deserialized and sent back to the accelerator logic. Therefore, 3D-SRAM read latency is three cycles. Moreover, the TCI operation delay is almost constant,

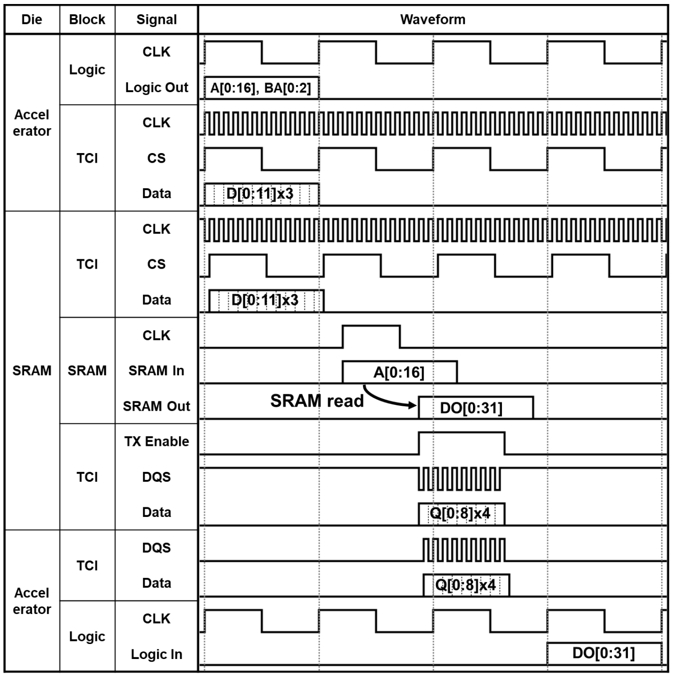


Fig. 8. Timing diagram of a single read access [7].

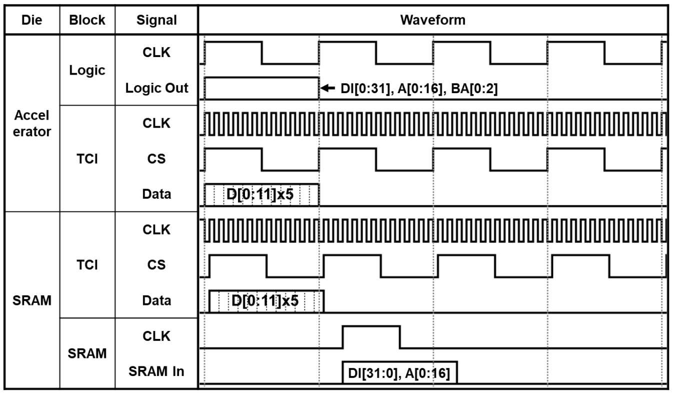


Fig. 9. Timing diagram of a single write access.

irrespective of the distance between the accelerator die and the selected SRAM die, and the three-cycle latency is uniform over the eight SRAM dies.

Fig. 9 shows a timing diagram of a single SRAM write operation. The write operation is conducted in similar fashion to the read operation. The difference is the number of activated transmitters in the accelerator die where seven transmitters send a CLK, CS, write data, write address, bank address, and write flag to the SRAM dies. The data received in an SRAM die are written within one cycle. Thus, 3D-SRAM write latency is two cycles.

Fig. 10 illustrates a timing diagram of back-to-back burst read and write operations. The example of burst operations in the figure include two consecutive reads, two consecutive writes, and a read. As can be seen in the figure, read and write operations are well pipelined and have no memory stall caused by write-after-read (WAR) and read-after-write (RAW) memory hazard. Therefore, the accelerator die can access the 3D-SRAM every cycle just like on-chip SRAM.

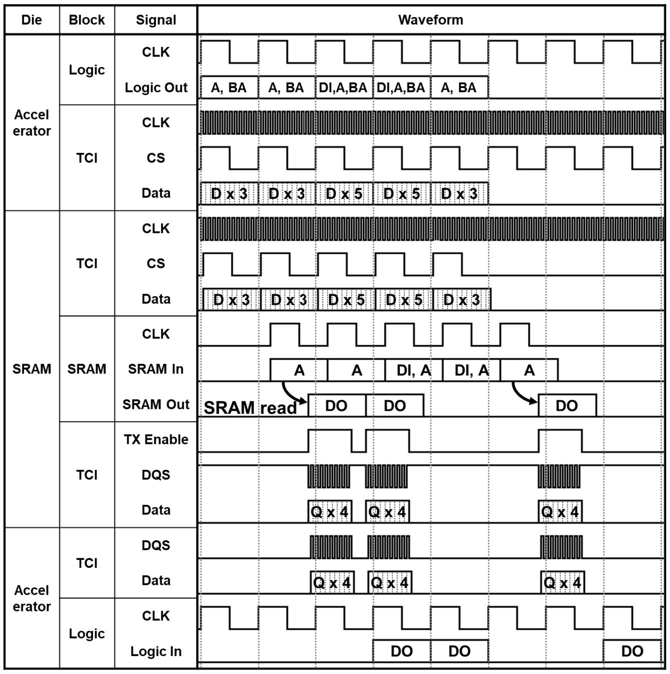


Fig. 10. Timing diagram of burst read and write accesses.

III. Inductive Coupling Link

In this section, we propose three key techniques to implement 3D stacking with TCI. First, we introduce a scheme to terminate the open coil of a transmitter in sleep mode to suppress its degradation of the received signal. Second, we propose a new NMOS push-pull transmitter that operates at 0.4 V to reduce power consumption by 35%. Third, we propose a scheme in which an inverted bit is inserted at the beginning of the packet to compensate for the reduced signal strength of the first transmitted bit after the turn-on sequence. Finally, our 3D-stacked SRAM performance is compared with that of conventional work.

A. Termination Scheme

In the proposed module, any unused transmitter is put in sleep mode in order to save power. When a selected SRAM die sends data back to the accelerator die in the uplink, the magnetic field reaches the receiver (RX) coil from the transmitter (TX) coil after passing through intervening coils in sleep mode, which causes a specific problem. As shown in Fig. 11, despite a high-Q (Q >5) TX coil, TCI has low-Q insertion gain characteristics thanks to damping by the transmitter when there are no intervening coils on the transmission path. In a TCI link for baseband communication, the ringing of the received pulse is suppressed by designing the gain to be low. However, when there are unused TX coils sandwiched in between which are open with high-Q due to their transmitters being in sleep mode, it makes the insertion gain characteristics high-Q. To make matters worse, the peak frequency of the gain is shifted lower and closer to the cutoff frequency of the received pulse, which makes ringing more likely to occur. Whereas wireless power transfer techniques utilize this mechanism [22] to increase power

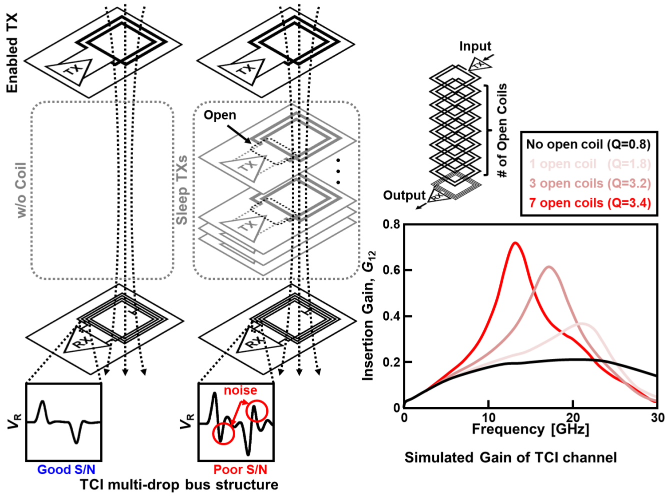


Fig. 11. Ringing issue caused by sleep TX coils in TCI bus.

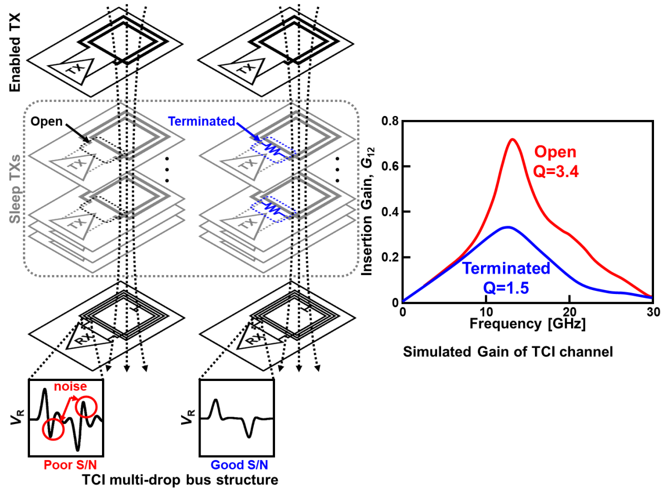


Fig. 12. Termination scheme suppressing ringing [7].

delivery efficiency, TCI is a baseband communication interface where such characteristics cause ringing in the RX coil, which results in communication error and low-speed operation due to lower signal-to-noise-ratio (S/N) and inter-symbol interference (ISI) [23].

To suppress the ringing, we propose a scheme to terminate the TX coils in sleep mode. As seen in Fig. 12, terminating sleep-mode TX coils with resistance can damp the insertion gain and suppress the ringing in RX, which enables reliable high-speed, low-power communication.

Fig. 13 shows the simulation results of received pulse amplitude margin $V_{\mathrm{RX,Margin}}$ against termination resistance. The received pulse amplitude margin $V_{\mathrm{RX,Margin}}$ is defined as the absolute value of the first received pulse amplitude (V_{S}) minus that of the second one (V_{N}) which results from ringing, where the threshold voltage of the hysteresis comparator receiver is set between V_{S} and V_{N}. When the termination resistance is 400 Ω, $V_{\mathrm{RX,Margin}}$ is 116 mV, which represents a 30-mV improvement over when there is no termination. Furthermore, as $V_{\mathrm{RX,Margin}}$ changes slowly against termination resistance around 400 Ω, the termination scheme is tolerant of process variations.

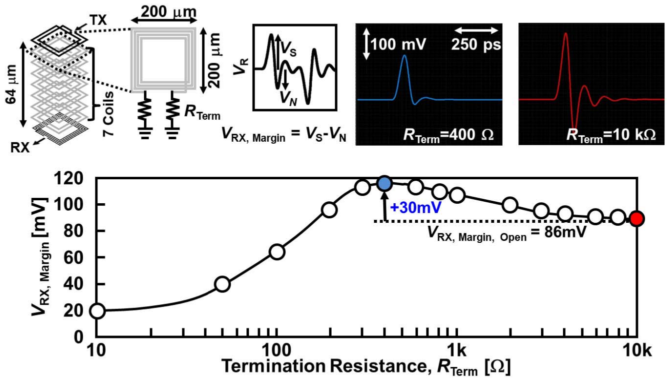


Fig. 13. Process variation tolerance of termination resistance [7].

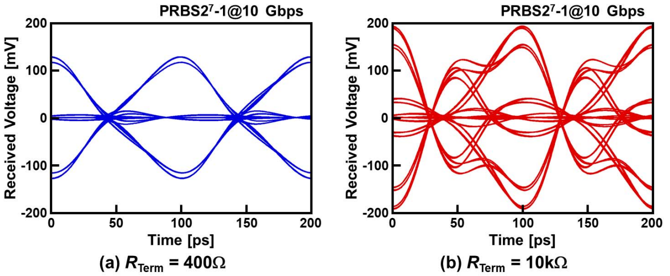


Fig. 14. Simulated eye diagram of (a) $R_{\mathrm{Term}} = 400\Omega$ and (b) $R_{\mathrm{Term}} = 10\mathrm{k}\Omega$ at 10 Gb/s.

Fig. 14 shows simulated eye diagrams when the termination resistance R_{Term} is 400Ω and 10kΩ respectively. When R_{Term} is 400Ω, clear eye opening is achieved at a 10 Gbps data rate. On the other hand, when R_{Term} is 10kΩ, the eye opening is greatly reduced. The ISI caused by the ringing in the receiver coil limits the maximum data rate of the inductive coupling link. Therefore, the termination scheme improves not only S/N but also maximum data rate.

B. NMOS Push-Pull Transmitter

The conventional CMOS H-bridge transmitter is a full-swing transmitter that efficiently generates a large pulse signal in the receiver to enable high-speed and low-noise operation. However, the CMOS transmitter cannot operate at a low voltage, which results in large energy dissipation. Though low-voltage, open-drain NMOS transmitters have been reported, these types of transmitters are pulse-modulated and have issues related to small received signals, low data rates, and large switching noise [24], [25]. Therefore, this work proposes a low-voltage NMOS push-pull transmitter that can maintain the same transmission strength, data rate, and switching noise as the conventional full-swing CMOS H-bridge transmitter with lower power. Fig. 15 illustrates the TCI transceiver block diagram, schematic, and waveforms of the new NMOS push-pull transmitter. The TX coils are designed with low resistance to compensate for the reduction of the flowing current due to reduction of the transmitter power supply voltage. In addition, pre-drivers operate at a standard voltage of 1.1 V, and NMOS

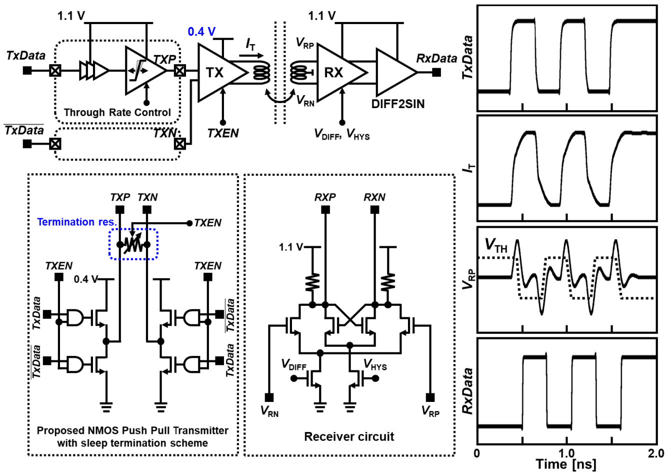


Fig. 15. Inductive coupling transceiver and diagram with NMOS push-pull transmitter [7].

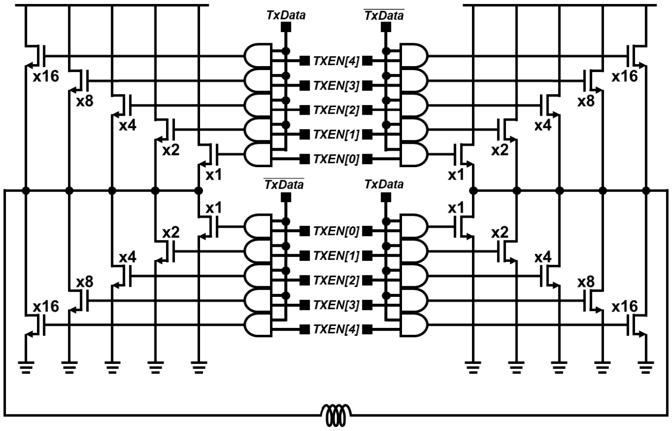


Fig. 16. Detailed structure of NMOS push-pull transmitter with current adjustment scheme.

transistors in the 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 by the transmitter without any degradation caused by the termination scheme.

Fig. 16 illustrates the detailed structure of the NMOS transmitter. The transmitter is equipped with a binary-weighted current adjustment scheme controlled by a 5-bit control signal. After fabrication and stack assembly, the amount of transmitter current can be optimized based on the location of the stacked chip and the power supply voltage of the transmitter to save power. The settings of the current strength are saved in the register of each chip.

Fig. 17 shows comparisons between the new NMOS push-pull transmitter and the conventional CMOS H-bridge transmitter. The right-hand graph plots the transistor gate width W of the output stage required to maintain the received pulse voltage of 130 mV against the power supply voltage of the transmitter. If the power supply voltage is reduced, the required transistor width W becomes drastically larger below a certain critical voltage. Whereas the critical voltage of the CMOS transmitter is around 0.8 V, that of the NMOS transmitter is around 0.4 V. Furthermore, the required transistor width W of the NMOS transmitter increases more slowly than that of the CMOS transmitter. Therefore, the NMOS transmitter enables a much lower supply voltage with smaller area penalty. The proposed NMOS transmitter achieves a 45% power reduction at 0.4 V compared with that of the CMOS transmitter.

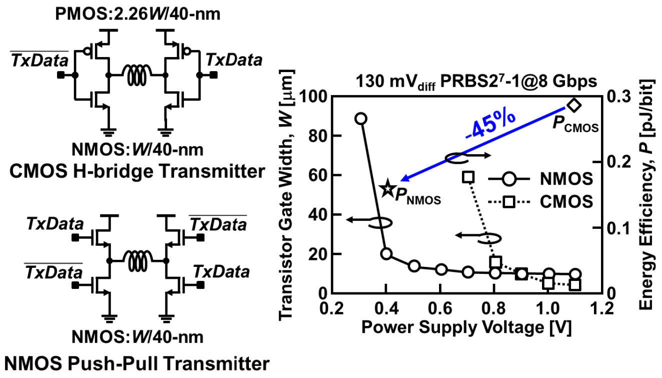


Fig. 17. Comparison of CMOS H-bridge and NMOS push-pull transmitter [7].

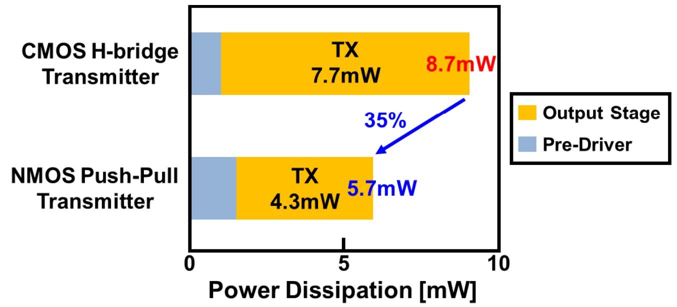


Fig. 18. Power dissipation in buffer for CMOS H-bridge and NMOS push-pull transmitter.

However, for the low voltage operation, pre-drivers have to drive a higher gate capacitance load. The comparison in power consumption between CMOS and NMOS transmitters along with pre-drivers are seen in Fig. 18. Under the low-voltage operation of the NMOS transmitter, the pre-drivers are required to drive higher capacitance load than that of the CMOS transmitter operating at a standard voltage. Therefore, the pre-drivers of the NMOS transmitter consume more power than that of the CMOS transmitter. Despite this higher drive requirement, because the pre-drivers consume much less power than the output stage of the transmitter, the NMOS transmitter achieves a 35% overall power reduction compared to the CMOS transmitter even with the increased power of the pre-drivers taken into account.

C. Inverted Bit Insertion Scheme

TCI's full-swing transmitters consume standby current even when idle, so the unused transmitters should be turned off so

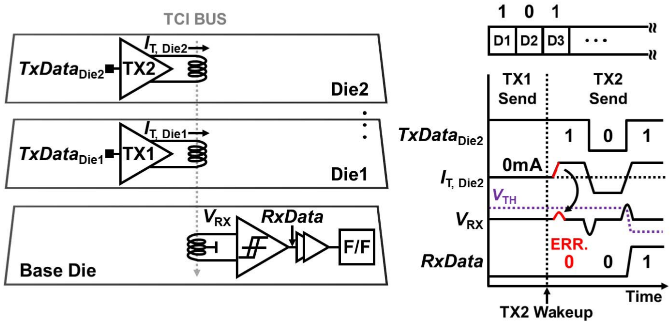


Fig. 19. Turn-on pulse causing a bit error at the first bit.

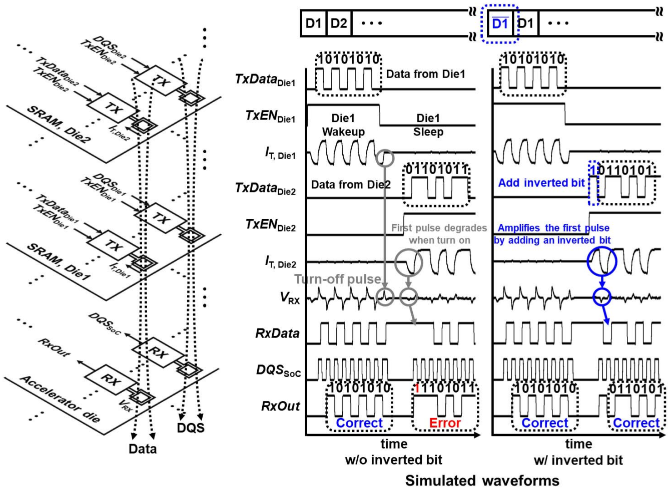


Fig. 20. Inverted bit insertion scheme [7].

that they do not consume unnecessary power. However, as seen in Fig. 19, the initial received pulse when the transmitter is first turned on (turn-on pulse) has only half the normal amplitude. If the receiver is unable to detect the turn-on pulse, it will lead to bit error(s).

In this work, we propose an inverted bit insertion scheme that inserts an inverted bit right before the first bit to prevent the error caused by the turn-on pulse. Fig. 20 illustrates the simulated waveforms without and with an inverted bit inserted when a transmitter in Die1 sends data, followed by Die2 transmitter. Although it causes no fatal error, a halved received pulse (turn-off pulse) is generated during the turn-off of a transmitter. If the receiver detects the turn-off pulse generated by the turn-off of Die1 transmitter, the data is flipped, and it holds a 1 and awaits data from Die2. After that, though Die 2 transmitter starts sending data with a 0 in the first bit, without an inverted bit inserted, the first bit is halved and cannot be detected by the receiver, so received data remains to be 1. This leads to a bit error. However, this problem can be solved by flowing current in the opposite direction just before sending the 0. The insertion of an inverted bit 1 enables full pulse amplitude when sending the 0, which eliminates the bit error. Furthermore, it is not problematic whether the first bit (inverted bit) is detected by the receiver or not because it is masked in the receiver. Therefore, thanks to the inverted bit insertion scheme, data are correctly received whether the receiver detects the turn-on pulse or turn-off pulse or neither.

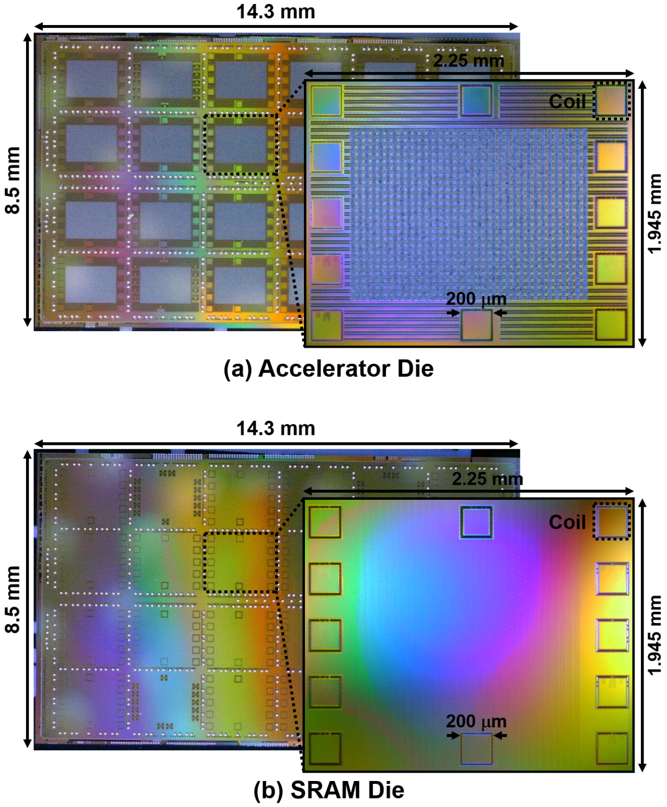


Fig. 21. Chip micro-photographs of test chips.

As described in Section II and seen in Fig. 7, 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-off and turn-on pulses. The inverted bit occupies a portion of a packet. However, it occupies only one bit in a 12-bit packet, and it enables unused transmitters to be turned off without causing errors upon turn-on. In other words, sacrificing only 8% of packets makes the standby current consumption from nearly 1,000 transmitters on stacked dies almost 0mA.

IV. Results and Scaling Scenario

A. Measurement Results

The photographs of the test chips are illustrated in Fig. 21. Each of the accelerator die in Fig. 21 (a) and the SRAM die in Fig. 21 (b) was fabricated in a 40-nm CMOS low-power (LP) process.

The shmoo test result of the proposed NMOS push-pull transmitter is shown in Fig. 22. As explained in Section III, the transistor gate width of the output stage in the transmitter is adjustable after fabrication of the test chip. By adjusting the gate width, wireless data communication using the NMOS push-pull transmitter operating at 0.40 V was experimentally confirmed with successful operation of the 3D-SRAM system.

B. Performance Comparison

Table II summarizes the performance comparison of the proposed 3D-stacked SRAM module and the 3D-stacked

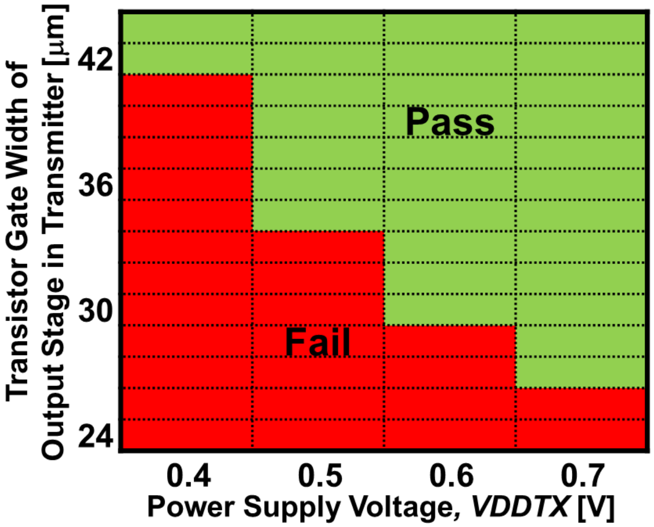


Fig. 22. Experimental result of NMOS push-pull transmitter.

TABLE II
PERFORMANCE COMPARISON TABLE

	JSSC'15 [4]	JSSC'17 [3]	This Work	Scaled to 10 nm
Memory Type	HBM DRAM	HBM2 DRAM	SRAM	SRAM
Capacity/Module	1 GB	8 GB	96 MB	~ 500 MB
Bandwidth	128 GB/s	307 GB/s	28.8 GB/s	> 512 GB/s
Data-rate	1.0 Gb/s/pin	2.4 Gb/s/pin	3.6 Gb/s/pin	> 30 Gb/s/pin
I/O Energy Consumption	3.8 pJ/bit/pin	~ 2 pJ/bit/pin	1.5 pJ/bit/pin (1 mW/RX standby)	< 0.05pJ/bit (~0.2 mW/RX standby)
Chip Size	5.1 mm x 6.9 mm	12 mm x 8 mm	14.3 mm x 8.5 mm	14.3 mm x 8.5 mm
Technology Node	29-nm DRAM	20-nm DRAM	40-nm CMOS	10-nm CMOS

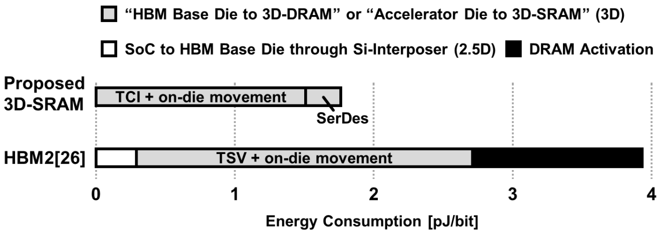


Fig. 23. Proposed 3D-SRAM and HBM2 DRAM energy consumption breakdown.

DRAM modules from [3] [4]. Each SRAM die is composed of 24 512-KB capacity channels, and the module has 8 12-MB SRAM dies adding up to 96 MB in total capacity. Twenty-four channels can access the 32-bit SRAM at a system clock frequency of 300 MHz, resulting in a total bandwidth of 28.8 GB/s. The total capacity and bandwidth are relatively small and low in a 40-nm process compared to HBM2 [3], respectively. However, if this module is designed in a 10-nm CMOS that is a comparable process node to the DRAM module in [3], its capacity and bandwidth are about 500 MB and more than 512 GB/s, respectively, which is discussed in detail later.

Fig. 23 illustrates the energy consumption breakdown of the proposed 3D-SRAM and the HBM2 DRAM from [26] for transferring data from a memory cell output to an SoC, such as an accelerator die or GPU. In HBM2, the DRAM activation needs 1.21 pJ/bit. The energy consumption required to transfer it from there to the HBM base die output accounts for the largest percentage and is 2.24 pJ/bit. In addition, transferring the data from the HBM base die to a GPU through a silicon interposer requires 0.3 pJ/bit, considering the toggle rates. In total, the HBM2 DRAM access requires 3.92 pJ/bit including ECC overhead. On the other hand, the energy consumption required to transfer the data from an SRAM to the accelerator die through TCI is 1.5 pJ/bit and the total energy consumption is 1.76 pJ/bit with a 0.25-pJ/bit 12:1-SerDes and 0.01-pJ/bit on-die movement. In total, the 3D-SRAM consumes only 1.76 pJ/bit for the SRAM access, which is less than half of the energy in the HBM2. Therefore, the 3D-SRAM makes the memory access energy half thanks to the 0.4-V transmitter and 12:1-SerDes.

The 3D-stacking structure reduces the board area by more than half compared to the 2.5D system integration with a horizontally arranged SoC and HBM [27], [28], which results in high area-efficiency in a PCB. There may be concern that the stacking of the SoC and memory dies causes difficulty in heat removal. However, it is not problematic when this is applied to accelerators for edge devices which consume less power. Though heat dissipation is a limiting factor for high-performance devices, 3D-stacking of thinned dies without μ-bumps reduces thermal resistance compared to conventional TSV-based stacking and eases the heat removal problem [29].

Moreover, there is a lot of room for further improvement in bandwidth and memory capacity. In this work, although SRAM macros and TCI coils are placed without overlapping in a conservative design, and the bandwidth is relatively low compared to HBM with approximately same I/O area overhead of 18%, the coils can potentially be placed above the SRAM macros and the bandwidth can be improved. The only concern is that eddy current flowing on the power distribution mesh causes attenuation of the coils' magnetic field. By using previously reported design techniques to suppress the adverse effect of power mesh, coils can be placed above SRAM macros to improve 3D-stacked SRAM bandwidth and capacity by a few hundreds and a few tens of percent, respectively [30].

C. Scaling Scenario

Furthermore, these numbers will be improved through technology scaling. A scaling scenario of the SRAM module is presented here. Firstly, a technology node trend of SRAM and DRAM respectively are illustrated in Fig. 24. The trend of SRAM is based on previous papers [31]–[39] and International Roadmap for Devices and Systems (IRDS) 2017 edition [40], while that of DRAM is based on TechInsights report [41] and IRDS [40]. Both SRAM and DRAM are expected to continue to shrink and memory density continue to rise. In addition, scaling will continue at least for the upcoming decade according to IRDS. However, as shown in the figure, DRAM has more difficulties in shrinking than SRAM, and the gap in capacity between SRAM and DRAM will probably be smaller. Note that a 10-nm CMOS is comparable to a 20-

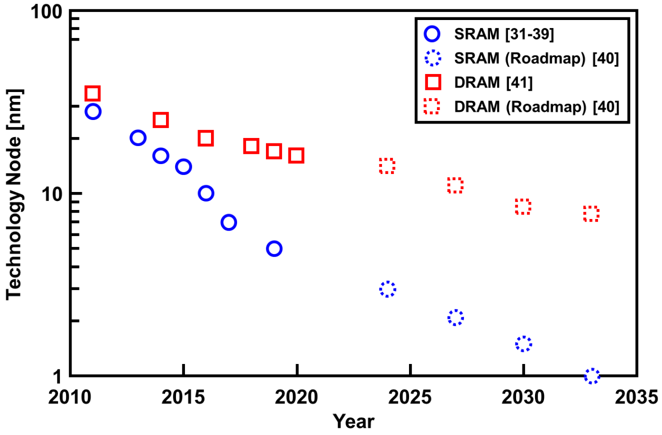
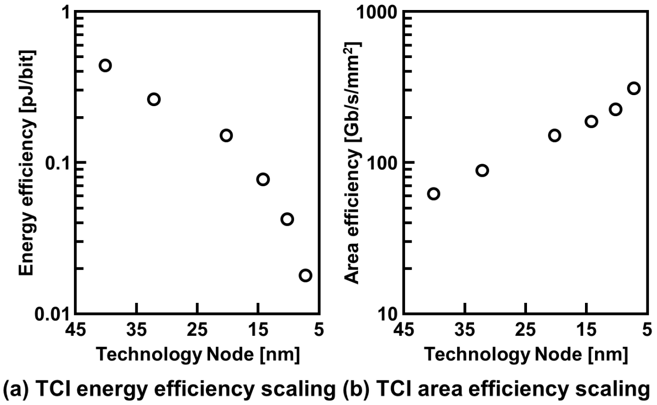


Fig. 24. Technology node trend of SRAM and DRAM.

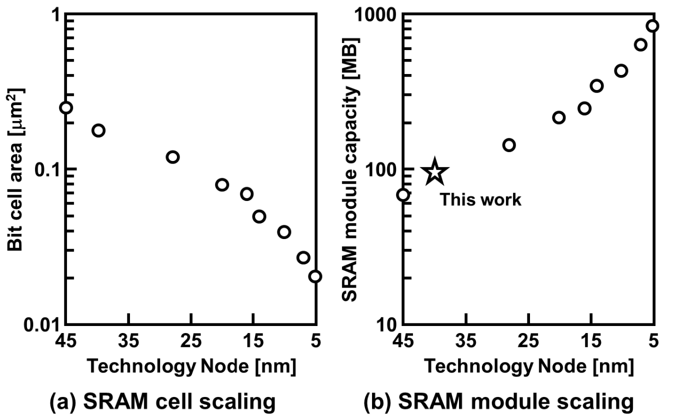


(a) TCI energy efficiency scaling (b) TCI area efficiency scaling

Fig. 25. TCI scaling scenario.

nm DRAM in the figure, which is used for the following discussion.

Secondly, a scaling scenario of the inductive coupling technology is discussed. As discussed in [42], when the device size is scaled down by a factor of $1/\alpha$ following Moore's Law, the energy efficiency and area efficiency of inductive coupling links are improved by a factor of α^3 and α, respectively, based on constant electric field scaling. Fig. 25 shows a scaling scenario of the inductive coupling technology based on simulation results, since voltage scaling is limited in actual devices. Fig. 25 (a) shows the scaling of the energy efficiency of the inductive coupling link with the NMOS push-pull transmitter operating at half of the standard voltage in each technology node. The simulation results are obtained by using predictive technology model (PTM) PDK for 32 to 10 nm [43] and Arizona State Predictive PDK (ASAP) for 7 nm [44]. The simulation results suggest that a 42-fJ/bit TCI link can be achieved in a 10-nm process. Fig. 25 (b) shows the scaling of the area efficiency at the maximum data rate in each node. The advance of the process node makes higher-speed inductive coupling link operation possible, which results in the improvement in area efficiency. Therefore, TCI memory bandwidth is improved through technology scaling and more than 512-GB/s SRAM is achievable when the overlapping placement techniques discussed in the previous subsection are also applied, which exceeds that of HBM2 DRAM.



(a) SRAM cell scaling (b) SRAM module scaling

Fig. 26. SRAM scaling scenario.

Thirdly, the scaling of SRAM is discussed in Fig. 26. Fig. 26 (a) shows the scaling of an SRAM bit cell from 45-nm to 5-nm CMOS process [31]–[39]. Within the same SRAM area as the proposed module, an SRAM module fabricated in a 10-nm CMOS technology, which is a comparable node to a 20-nm DRAM process, should achieve a 500-MB SRAM capacity, as shown in Fig. 26 (b). While this is still small compared to DRAM, we expect the gap to gradually close based on Fig. 24.

V. Conclusion

A 28.8-GB/s 96-MB 3D-stacked SRAM module is proposed. The proposed low-voltage NMOS push-pull transmitter achieves a 35% power reduction compared to that of the conventional CMOS H-bridge transmitter. The proposed termination scheme removes the ringing induced in a 3D inductive coupling bus. The proposed inverted bit insertion scheme compensates for the reduced transmission strength of the first bit after the turn-on sequence of the transmitter. With these techniques, the proposed 3-cycle latency 3D-SRAM can potentially replace 3D-DRAM to improve the accelerator performance in DNN. A scaling scenario of the proposed SRAM module is discussed through scaling of the inductive coupling link and logic process.

Acknowledgment

The authors would like to thank T. Miyata, J. Kadomoto, W.-Y. Yip, and M. Okada for their invaluable support.

References

[1] D. Kim, J. Kung, S. Chai, S. Yalamanchili, and S. Mukhopadhyay, "Neurocube: A programmable digital neuromorphic architecture with high-density 3D memory," in *Proc. IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 380–392.

[2] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "TETRIS: Scalable and efficient neural network acceleration with 3D memory," *ACM SIGOPS Operating Syst. Rev.*, vol. 51, no. 2, pp. 751–764, Apr. 2017.

[3] K. Sohn *et al.*, "A 1.2 V 20 nm 307 GB/s HBM DRAM with at-speed wafer-level IO test scheme and adaptive refresh considering temperature distribution," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 250–260, Jan. 2017.

[4] D. U. Lee *et al.*, "A 1.2 V 8 GB 8-channel 128 GB/s high-bandwidth memory (HBM) stacked DRAM with effective I/O test circuits," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 191–203, Jan. 2015.

[5] K. Ueyoshi *et al.*, "QUEST: A 7.49TOPS multi-purpose log-quantized DNN inference engine stacked on 96 MB 3D SRAM using inductive-coupling technology in 40 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 216–217.

[6] 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, Jan. 2019.

[7] K. Shiba *et al.*, "A 3D-stacked SRAM using inductive coupling with low-voltage transmitter and 12:1 SerDes," in *Proc. IEEE Int. Symp. Circuits Syst.*, Oct. 2020, pp. 1–5.

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

[9] T. Kuroda, "Circuit and device interactions for 3D integration using inductive coupling," *IEDM Tech. Dig.*, Dec. 2014, pp. 18.6.1–18.6.4.

[10] M. Saito, Y. Yoshida, N. Miura, H. Ishikuro, and T. Kuroda, "47% power reduction and 91% area reduction in inductive-coupling programmable bus for NAND flash memory stacking," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 9, pp. 2269–2278, Sep. 2010.

[11] N. Miura, M. Saito, and T. Kuroda, "A 1 TB/s 1 pJ/b 6.4 mm^2/TB/s QDR inductive-coupling interface between 65-nm CMOS logic and emulated 100-nm DRAM," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 2, pp. 249–256, Jun. 2012.

[12] I. A. Papistas, V. F. Pavlidis, and D. Velenis, "Fabrication cost analysis for contactless 3-D ICs," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 5, pp. 758–762, May 2019.

[13] Y. S. Kim *et al.*, "Ultra thinning down to 4-μm using 300-mm wafer proven by 40-nm node 2Gb DRAM for 3D multi-stack WOW applications," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2014, pp. 1–2.

[14] Y. S. Kim *et al.*, "A robust wafer thinning down to 2.6-μm for bumpless interconnects and DRAM WOW applications," in *IEDM Tech. Dig.*, Dec. 2015, pp. 8.3.1–8.3.4.

[15] N. Miura, T. Sakurai, and T. Kuroda, "Crosstalk countermeasures for high-density inductive-coupling channel array," *IEEE J. Solid-State Circuits*, vol. 42, no. 2, pp. 410–421, Feb. 2007.

[16] M. Murugesan *et al.*, "Fully-filled, highly-reliable fine-pitch interposers with TSV aspect ratio >10 for future 3D-LSI/IC packaging," in *Proc. IEEE 69th Electron. Compon. Technol. Conf. (ECTC)*, May 2019, pp. 1047–1051.

[17] J. Su, F. Wang, and W. Zhang, "Capacitance expressions and electrical characterization of tapered through-silicon vias for 3-D ICs," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 5, no. 10, pp. 1488–1496, Oct. 2015.

[18] K. Shiba, M. Hamada, and T. Kuroda, "3D SoC design with TSV-less power supply employing highly doped silicon via," in *Proc. JSAP Int. Conf. Solid State Devices Mater.*, Sep. 2019, pp. 515–516.

[19] K. Shiba, M. Hamada, and T. Kuroda, "3D system-on-a-chip design with through-silicon-via-less power supply using highly doped silicon via," *Jpn. J. Appl. Phys.*, vol. 59, Apr. 2020, Art. no. SGGL04.

[20] B. J. Fletcher, S. Das, and T. Mak, "CoDAPT: A concurrent data and power transceiver for fully wireless 3D-ICs," in *Proc. Des., Automat. Test Eur. Conf. Exhib.*, Mar. 2019, pp. 1343–1348.

[21] N. Miura, Y. Kohama, Y. Sugimori, H. Ishikuro, T. Sakurai, and T. Kuroda, "A high-speed inductive-coupling link with burst transmission," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 947–955, Mar. 2009.

[22] A. G. Pelekanidis, A. X. Lalas, N. V. Kantartzis, and T. D. Tsiboukis, "Optimized wireless power transfer schemes with metamaterial-based resonators," in *Proc. Int. Workshop Antenna Technol.*, Mar. 2017, pp. 289–292.

[23] L.-C. Hsu, J. Kadomoto, S. Hasegawa, A. Kosuge, Y. Take, and T. Kuroda, "Analytical thruchip inductive coupling channel design optimization," in *Proc. 21st Asia South Pacific Des. Automat. Conf. (ASP-DAC)*, Jan. 2016, pp. 731–736.

[24] 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*, vol. 46, no. 4, pp. 965–973, Apr. 2011.

[25] S. Hasegawa, J. Kadomoto, A. Kosuge, and T. Kuroda, "A 1 Tb/s/mm^2 inductive-coupling side-by-side chip link," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2016, pp. 469–472.

[26] M. O'Connor *et al.*, "Fine-grained DRAM: Energy-efficient DRAM for extreme bandwidth systems," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchit. (MICRO)*, Oct. 2017, pp. 41–54.

[27] Y. Jeon, H. Kim, J. Kim, and M. Je, "Design of an on-silicon-interposer passive equalizer for next generation high bandwidth memory with data rate up to 8 Gb/s," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 7, pp. 2293–2303, Jul. 2018.

[28] S. Ma, H. Yu, Q. J. Gu, and J. Ren, "A 5–10-Gb/s 12.5-mW source synchronous I/O interface with 3-D flip chip package," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 2, pp. 555–568, Feb. 2019.

[29] H. Ryoson, K. Fujimoto, and T. Ohba, "A design guide of thermal resistance down to 30% for 3D multi-stack devices," in *Proc. Int. Conf. Electron. Packag. (ICEP)*, Apr. 2017, pp. 522–525.

[30] L.-C. Hsu, J. Kadomoto, S. Hasegawa, A. Kosuge, Y. Take, and T. Kuroda, "A study of physical design guidelines in thruchip inductive coupling channel," *Trans. Fundam. Electron., Commun. Comput. Sci.*, vols. E98–A, no. 12, pp. 2584–2591, Dec. 2015.

[31] M.-S. Kim *et al.*, "12-EUV layer surrounding gate transistor (SGT) for vertical 6-T SRAM: 5-nm-class technology for ultra-density logic devices," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2019, pp. T198–T199.

[32] J. Chang *et al.*, "A 7 nm 256 Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 206–207.

[33] 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, Jan. 2017.

[34] E. Karl *et al.*, "A 0.6 V, 1.5 GHz 84 Mb SRAM in 14 nm FinFET CMOS technology with capacitive charge-sharing write assist circuitry," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 222–229, Jan. 2016.

[35] Y.-H. Chen *et al.*, "A 16 nm 128 Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 238–239.

[36] J. Chang, "A 20 nm 112 Mb SRAM in high-k metal-gate with assist circuitry for low-leakage and low-VMIN applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2013, pp. 316–317.

[37] M. E. Sinangil, H. Mair, and A. P. Chandrakasan, "A 28 nm high-density 6T SRAM with optimized peripheral-assist circuits for operation down to 0.6 V," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 260–261.

[38] O. Hirabayashi *et al.*, "A process-variation-tolerant dual-power-supply SRAM with 0.179μm^2 cell in 40 nm CMOS using level-programmable wordline driver," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 458–459.

[39] N. Verma and A. P. Chandrakasan, "A high-density 45 nm SRAM using small-signal non-strobed regenerative sensing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 380–381.

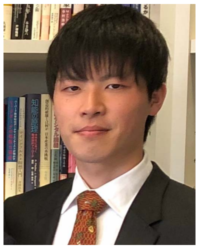
[40] *International Roadmap for Devices and Systems 2017 Edition*. Accessed: Oct. 14, 2020. [Online]. Available: https://irds.ieee.org/images/files/pdf/2017/2017IRDS_MM.pdf

[41] D. James and J. Choe. (Apr. 11, 2019). *TechInsights Memory Technology Update From IEDM18*. TechInsights Inc. [Online]. Available: https://www.techinsights.com/blog/techinsights-memory-technology-update-iedm18

[42] T. Kuroda, "Near-field coupling integration technology," *ESC Trans.*, vol. 72, no. 3, pp. 83–91, May 2016.

[43] *Predictive Technology Model (PTM) PDK*. Accessed: Oct. 14, 2020. [Online]. Available: http://ptm.asu.edu/

[44] *ASAP: Arizona State Predictive PDK*. Accessed: Oct. 14, 2020. [Online]. Available: http://asap.asu.edu/asap/



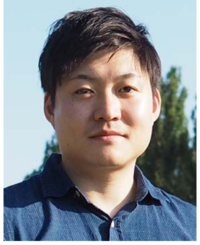
Kota Shiba (Student Member, IEEE) was born in Tokyo, Japan, in 1995. He received the B.S. and M.S. degrees in electronics and electrical engineering from Keio University, Yokohama, Japan, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with The University of Tokyo, Tokyo, Japan. He is a Research Assistant with the Graduate School of Engineering, The University of Tokyo (SEUT-RA) since 2020, and will be a JSPS research fellow (DC2) since 2021.

Since 2017, he has been involved in research on the inter-chip wireless interface for 3D system integration. His current research interests include inductive coupling wireless communication, high-speed I/O interface, 3D-SRAM-based hardware architecture, and 3D system integration.

Mr. Shiba was a recipient of the IEEJ Tokyo Branch Student Encouragement Award in 2018.

Tatsuo Omori was born in Kanazawa, Japan, in 1996. He received the B.S. degree from Keio University, Yokohama, Japan, in 2020. He is currently pursuing the M.S. degree with The University of Tokyo, Tokyo, Japan, both in electrical engineering.

His current research interest includes three-dimensional integrated circuit using near-field coupling. He was a recipient of the IEEJ Tokyo Branch Student Encouragement Award from The Institute of Electrical Engineers of Japan in 2020.



Kodai Ueyoshi (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Hokkaido University, Japan, in 2015, 2017, and 2020, respectively.

He is currently a Post-Doctoral Researcher with KU Leuven. His research interests include energy efficient hardware architecture for machine learning systems and software and hardware co-optimization.

Dr. Ueyoshi received a Research Fellowships for Young Scientists from JSPS in 2017, the Silkroad Award at the 2018 IEEE International Solid-State Circuits Conference, the IEEE SSCS Pre-Doctoral Award, and the Ninth JSPS Ikushi Prize in 2019.

Shinya Takamaeda-Yamazaki (Member, IEEE) received the B.E., M.E., and D.E. degrees from the Tokyo Institute of Technology, Japan, in 2009, 2011, and 2014, respectively.

From 2011 to 2014, he was a JSPS Research Fellow (DC1). From 2014 to 2016, he was an Assistant Professor with the Nara Institute of Science and Technology, Japan. From 2016 to 2019, he was an Associate Professor with Hokkaido University, Japan. Since 2018, he has been a Researcher of JST PRESTO. Since 2019, he has been an Associate Professor with The University of Tokyo, Japan. His research interests include computer architecture, high-level synthesis, and machine learning acceleration.

He is a member of IEICE and IPSJ.

Masato Motomura (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Kyoto University, Kyoto, Japan, in 1985, 1987, and 1996, respectively. In 1987, he joined the NEC Central Research Laboratories, Kawasaki, Japan, where he worked on various hardware architectures, including string search engines, multi-threaded on-chip parallel processors, embedded DRAM-field-programmable gate array (FPGA) hybrid systems, memory-based processors, and reconfigurable systems. From 2001 to 2008, he was with NEC Electronics, Kawasaki, Japan, where he led research and business development of dynamically reconfigurable processor (DRP) that he invented. He was also a Visiting Researcher with the MIT Laboratory for Computer Science, Cambridge, MA, USA, from 1991 to 1992. He has been a Professor with Hokkaido University, Sapporo, Japan, since 2011. He has also been a Professor with the Tokyo Institute of Technology since 2019, and is actively working on reconfigurable and parallel architectures for deep neural networks, machine learning, annealing machines, and intelligent computing in general. He is a member of IEICE, IPSJ, and EAJ. He was a recipient of the IEEE JSSC Annual Best Paper Award in 1992, the IPSJ Annual Best Paper Award in 1999, the IEICE Achievement Award in 2011, and the ISSCC Silkroad Award as the last author in 2018, respectively.

Mototsugu Hamada (Member, IEEE) was born in Nara, Japan, in 1968. He received the B.S., M.S., and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1991, 1993, and 1996, respectively.

In 1996, he joined Toshiba Corporation and has been involved in wireless and low-power electronic circuits design with Toshiba's Center for Semiconductor Research and Development, Kawasaki, Japan. From 2002 to 2004, he was a Visiting Scholar with Stanford University. From 2011 to 2016, he was with the Mixed Signal IC Division as the Group Manager of Power Analog IC Design Group to lead the development of analog mixed signal ICs. In 2016, he joined Keio University and was a Project Professor. In 2020, he joined The University of Tokyo, where he is currently a Project Professor of Systems Design Laboratory. His research interests include low-power, high-speed CMOS design, low-power wireless systems and circuits design, and power management systems design.

Dr. Hamada was a recipient of the 2007 IEEE International Conference on Computer Design (ICCD) Best Paper Award and the recipient of the Design Automation Conference (DAC) 2010 Best User Track Poster Award. He was also recognized in the list of AUTHORS of TEN OR MORE PAPERS IN THE PAST TEN YEARS at the International Solid-State Circuits Conference 2013 (ISSCC2013). He has served as a member of the Technical Program Committee of International Solid-State Circuits Conference from 2003 to 2009 and in 2011) and VLSI Circuits Symposium from 2018 to 2020, and Asian Solid-State Circuits Conference (from 2005 to 2012 and 2017 to 2020, where he served as the RF Subcommittee Chair, the Digital Subcommittee Chair, the Student Design Contest Chair, and the Technical Program Committee Chair.

Tadahiro Kuroda (Fellow, IEEE) received the Ph.D. degree in electrical engineering from The University of Tokyo, Tokyo, Japan, in 1999.

In 1982, he joined Toshiba Corporation, where he designed CMOS SRAMs and ASICs. From 1988 to 1990, he was a Visiting Scholar with the University of California at Berkeley, Berkeley, CA, USA, where he conducted research in the field of VLSI CAD. In 1990, he was back to Toshiba, and involved in the research and development of BiCMOS/ECL ASICs, high-speed CMOS LSIs for telecommunications and low-power CMOS LSIs for mobile applications. He invented a Variable Threshold-voltage CMOS (VTCMOS) technology to control V_{TH} through substrate bias, and applied it to a DCT core processor, in 1995. He also developed a Variable Supply-voltage scheme to control VDD by an embedded DC–DC converter, and employed it to a microprocessor core and an MPEG-4 chip in 1997. In 2000, he moved to Keio University, Yokohama, Japan. He was a MacKay Professor with the University of California at Berkeley, in 2007. In 2019, he moved to The University of Tokyo, where he is currently the Director of the Systems Design Laboratory. His research interests include low-power, high-speed CMOS design, 3D integration using near-field coupling, and artificial intelligence. He has published more than 400 technical publications, including 38 ISSCC articles, 29 VLSI Symposia articles, 19 CICC articles, and 17 A-SSCC articles. He wrote 29 books/chapters and filed more than 200 patents.

Dr. Kuroda is an IEICE Fellow. He was an elected AdCom member of two terms. He was a recipient of the 2005 P&I Patent of the Year Award, the 2007 ASP-DAC Best Design Award, the 2009 IEICE Achievement Award, and the 2011 IEICE Society Award. He serves as an Executive Committee Chair for Symposium on VLSI Technology and Circuits. He served as a Steering Committee Chair for A-SSCC, a Vice Chair for ASP-DAC, sub-committee chairs for A-SSCC, ICCAD, SSDM, and VLSI-DAT, and TPC members for ISSCC, Symposium on VLSI Circuits, CICC, DAC, ASP-DAC, ISLPED, SSDM, ISQED, and other international conferences. He was a Distinguished Lecturer and a representative of Region 10 for the IEEE Solid-State Circuits Society.