Received 21 February 2024; revised 5 June 2024; accepted 18 June 2024. Date of publication 24 June 2024; date of current version 16 July 2024.
The review of this article was arranged by Editor S. Menzel.

Digital Object Identifier 10.1109/JEDS.2024.3418036

A 3-D Bank Memory System for Low-Power Neural Network Processing Achieved by Instant Context Switching and Extended Power Gating Time

KOUHEI TOYOTAKA[1], YUTO YAKUBO[1] (Member, IEEE), KAZUMA FURUTANI[1] (Member, IEEE), HARUKI KATAGIRI[2], MASASHI FUJITA[1], YOSHINORI ANDO[3], TORU NAKURA[4] (Member, IEEE), AND SHUNPEI YAMAZAKI[5] (Life Fellow, IEEE)

1 CAD Division, Semiconductor Energy Laboratory Company Ltd., Atsugi 243-0036, Japan
2 Equipment Division, Semiconductor Energy Laboratory Company Ltd., Atsugi 243-0036, Japan
3 NOS Development Division, Semiconductor Energy Laboratory Company Ltd., Atsugi 243-0036, Japan
4 Graduate School of Engineering, Fukuoka University, Fukuoka 814-0180, Japan
5 Semiconductor Energy Laboratory Company Ltd., Atsugi 243-0036, Japan

CORRESPONDING AUTHOR: K. TOYOTAKA (e-mail: kt1024@sel.co.jp)

ABSTRACT Using a 3-D monolithic stacking memory technology of crystalline oxide semiconductor (OS) transistors, we fabricated a test chip having AI accelerator (ACC) memory for weight data of a neural network (NN), backup memory of flip-flops (FF), and CPU memory storing instructions and data. These memories are composed of two-layer OS transistors on Si CMOS, where memories in each layer correspond to a bank. In this structure, bank switching of the ACC memory and the FF backup memory work together, and thus inference of different NNs is switched with low latency and low power so that the power gating standby time can be extended. Consequently, a 92% reduction in power consumption is achieved in inference at a frame rate of 60 fps as compared with a chip using static random access memory (SRAM) as the ACC memory.

INDEX TERMS Oxide semiconductor, IGZO, monolithic stacking, endpoint AI, power gating, context switching.

I. INTRODUCTION

Recently, in view of information security in various AI applications, not a traditional AI system that calculates uploaded data on a cloud but edge AI that completes processing within a closed network has been highly demanded [1], [2], [3], [4]. The edge AI that can complete inference in a device is called end point AI. For the end point AI, which is often used for data analysis or the like of Internet of things (IoT) devices, a small chip area and low power consumption are highly demanded. In order to perform various types of data analysis, highly flexible AI processing with frequent switching of neural networks (NNs) is further required (Fig. 1).

For highly efficient AI processing, AI accelerators (ACCs) that perform fast NN inference have been studied. The power efficiency of ACC decreases due to an increase in power and processing time for transferring NN weight data and temporary data. As a countermeasure, it is known to be effective to place local memory for those data near the ACC [5], [6], [7], [8]. Among the data stored in the local memory, weight data need not be rewritten as long as no change is made in the NN. Thus, memory being capable of long-term data retention and having no increase in standby power with increasing storage capacity is suitable as the local memory. However, static random access memory (SRAM), which is capable of high-speed data reading and writing but has high leakage current, is used as the ACC's local memory in many cases [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]. Various configurations of ACC chips including SRAM local memory, which have

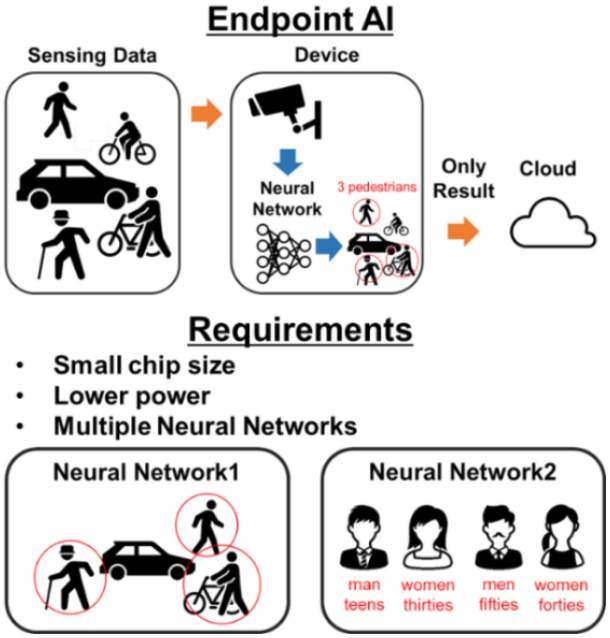


FIGURE 1. Requirements of endpoint AI.

low power consumption and high AI processing capability, are reported [32]. However, increasing the scale of AI processing requires increasing the capacity of SRAM in order to process it efficiently, causing problems such as increased chip size and standby power. For example, when an IoT device continuously processes multiple NNs with different weight data, it is difficult to meet the needs for a small chip size and low power consumption.

Transistors using crystalline oxide semiconductor (OS) such as indium–gallium–zinc oxide (IGZO) have extremely low off-state current [32]. Thus, a charge-holding memory composed of OS transistors has much longer retention time than dynamic random access memory (DRAM) [32]. Furthermore, since multiple layers of memory composed of OS transistors (OS memory) can be 3-D stacked, memory capacity can be increased without the chip area increase [32]. Moreover, OS transistors have high compatibility with Si CMOS processes and thus can be monolithically stacked over Si FETs. A normally-off (Noff) CPU using OS memory as flip-flop (FF) backup memory to enable instant power gating (PG) has been reported [32].

We fabricated a test chip, which achieves all of a small area, low power consumption, and AI processing with instant context switching between two NNs, by adding a highly power-efficient ACC to an Noff CPU and stacking two layers of OS memory over Si CMOS, where the OS memory in each layer corresponds to a bank, with use of the monolithic stacking technology [32].

II. CONCEPT

Fig. 2 is a concept view of the structure we propose. The ACC memory for NN weight data, the FF backup memory, and the CPU memory storing instructions and data are each 3-D stacked using the monolithic stacking technology of OS

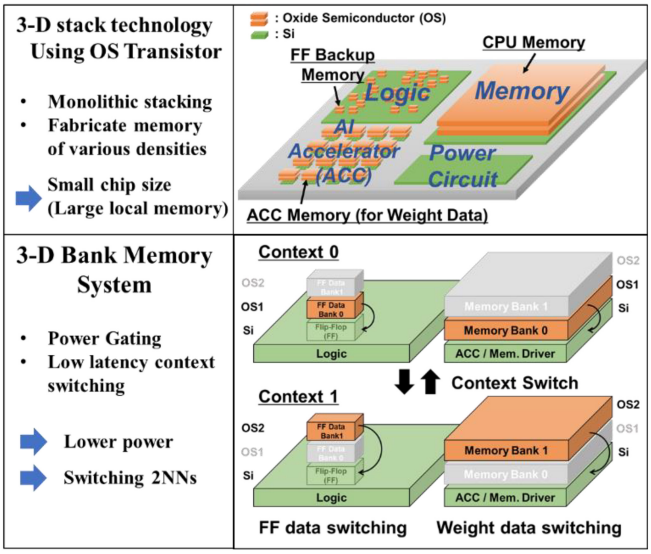


FIGURE 2. 3-D bank memory system.

transistors, thereby suppressing the increase in chip size due to the area of memories. Each memory is composed of OS transistors, and thus can retain data for a long period, so that standby power can be reduced by PG. Since the OS memory in each layer corresponds to a bank, not only weight data but also data stored in the FF with the OS memory (OSFF) can be quickly switched to data for another NN by context switching, thereby enabling fast switching between two NNs. We consider that these functions enable a small and power-saving chip that can perform AI processing with switching multiple NNs.

III. CHIP CONFIGURATION

Fig. 3a shows the configuration of the test chip fabricated as a proof of this concept. The chip is composed of an Noff CPU including Cortex-M0 (CORE), two layers of 4 KB CPU memory, a power management unit (PMU), and peripheral circuits connected to a BUS, and an ACC including 128 multiply accumulate (MAC) processing elements (PEs) and two layers of 32 KB OS memory. These circuits are placed as shown in a die photo of Fig. 3b. The ACC memory is stacked in regions where the PEs are lined up, with less area overhead for the PEs. All logic circuits except the PMU placed in the upper left of Fig. 3b use OSFF as registers, and thus are capable of PG. Fig. 3c is a cross-sectional view of the test chip, and a portion surrounded by a red frame corresponds to the OS memory cell. This chip was fabricated through the OS/OS/Si process [32] where two layers of 200-nm-node OS are stacked over 130-nm-node Si CMOS. The cross-sectional view shows that the two OS transistor layers are stacked directly over the Si back end of line (BEOL) interconnect layer and this allows incorporating more memory banks without increasing the area.

IV. CIRCUIT

The ACC shown in the block diagram in Fig. 4a supports a binary NN (BNN) for low-power AI operation. In addition

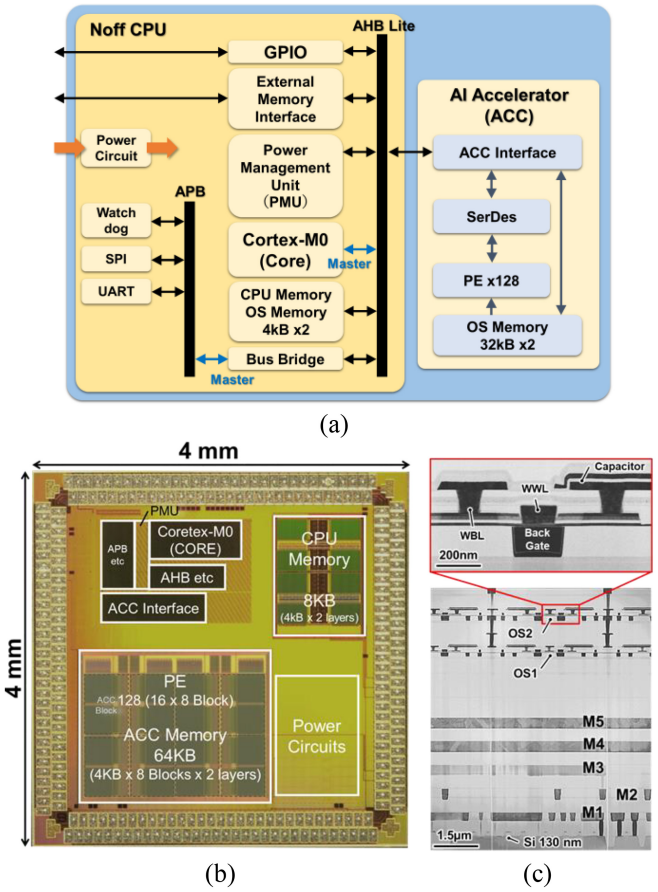


FIGURE 3. (a) Chip configuration, (b) Die photo, and (c) Cross-section.

to a mechanism for changing the number of parallel-driven PEs in accordance with the NN, the ACC includes a control logic circuit having a mode change function for memory operation and AI processing and including a Serializer-Deserializer (SerDes). In consideration of the trade-off between a reduction in driver area and an improvement in latency due to the memory block division number, a block-by-block arrangement has been adopted in which eight blocks, each comprising 16 PEs sharing two layers of 4 KB memory, are arranged. This allows parallel driving of 128 PEs (16 PEs × 8 blocks) at the maximum. The PE shown in Fig. 4b executes eight MAC operations of a binary input and a binary weight by XNOR-Popcount in parallel [32]. The MAC operations are executed in one PE clock (PECLK), and the results are temporarily stored in an accumulator. The stored data is added to MAC operation results of input data in the next clock period, and then stored in the accumulator again. In the case of a fully connected network with 784 inputs (neurons) as shown in Fig. 4c, the eight parallel operations are repeated 98 times, the threshold value processing (biasing) is performed in 11 PECLKs, and then 1 PECLK is consumed to output result, so that operation in a first-layer network is completed. In the case of a fully connected network with three hidden layers, inference is possible in 194 PECLKs as shown in Fig. 4d.

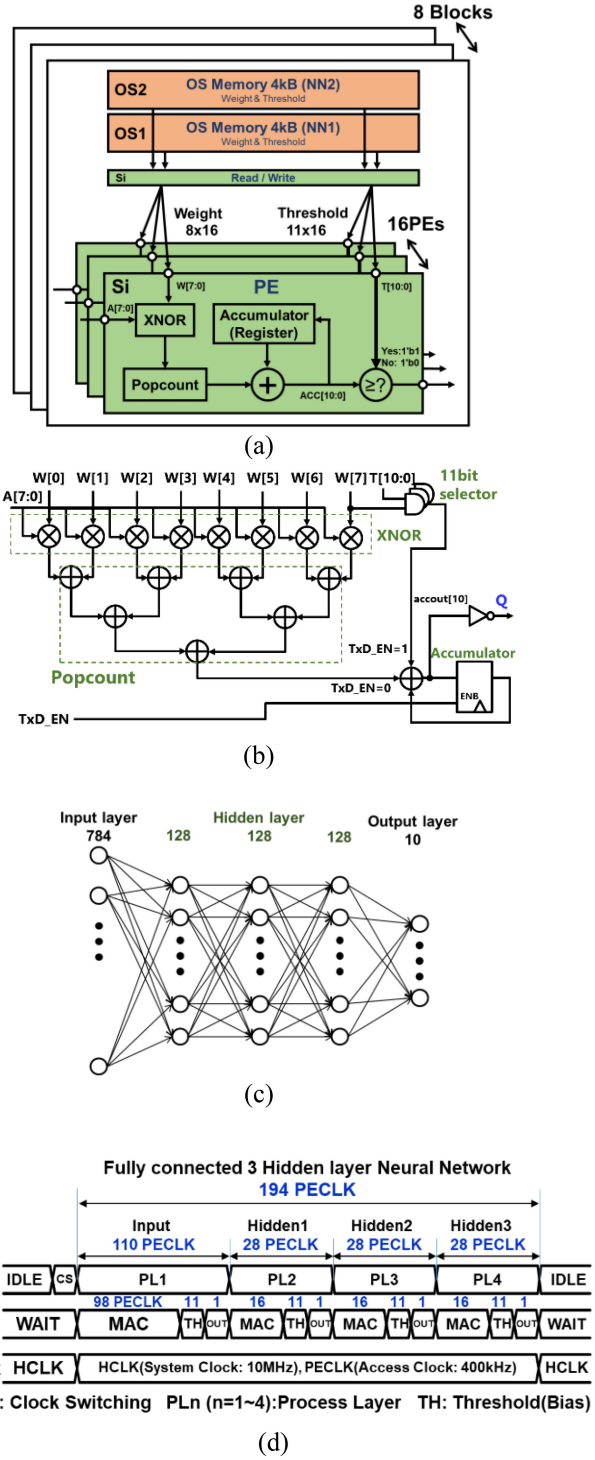


FIGURE 4. (a) Block diagram of ACC, (b) Circuit diagram of PE, (c) Fully connected NN, and (d) Timing chart of inference.

Fig. 5a shows the OSFF schematic. The OSFF is a circuit where a register data backup memory is stacked directly over a Si CMOS scan FF without area overhead. The register data backup memory is composed of the following four elements: a backup control OS transistor connected to an FF output (Q); a restore control OS transistor connected to a selector input; a scan control OS transistor connected to

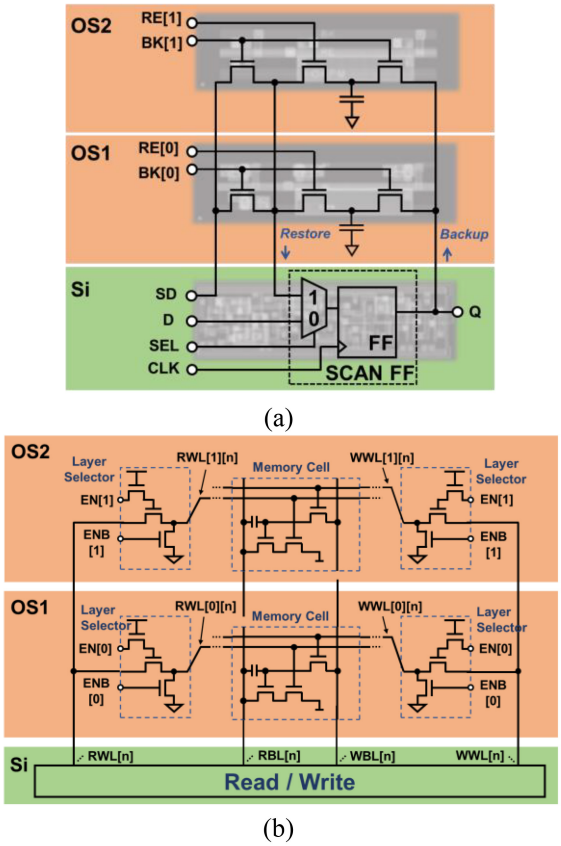


FIGURE 5. Circuit diagram of (a) OSFF and (b) OS memory.

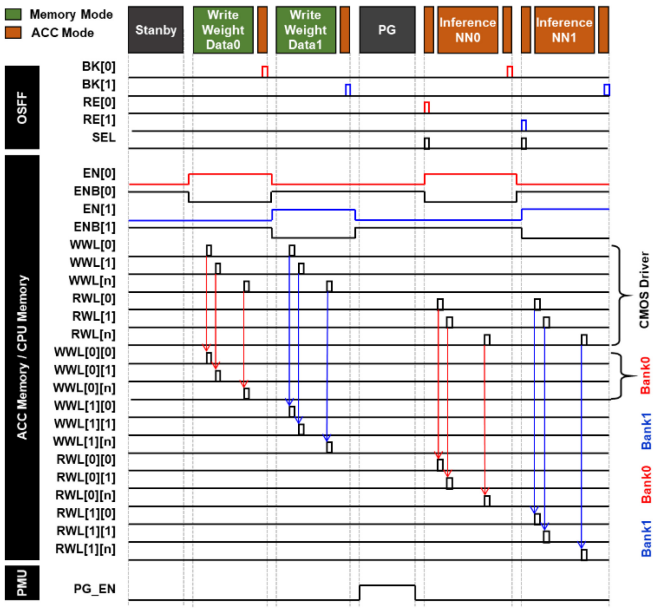


FIGURE 6. Timing chart of context switching.

an FF scan data input (SD); and a data retention capacitor. Taking advantage of monolithic stacking, fine-grained and random arrangement is possible for the OSFFs. In the register data backup memory in each layer, register data of a scan FF is backed up in response to a backup signal BK [0] or BK [1] of the layer corresponding to context switching, and then the backup data can be restored through the selector input in response to a restore signal RE [1] or RE [0].

Fig. 5b is a circuit diagram of the OS memory used as the ACC memory and the CPU memory. Since a memory layer to be accessed by the ACC or CPU is determined by a layer selection driver formed using only OS transistors, there is no need to increase the address size of the Si CMOS read and write circuits. Furthermore, the layer selection drivers can be directly stacked using the same mask pattern, thereby increasing neither area nor power consumption of the Si CMOS circuit even with the increased number of layers. The layer selection driver is composed of a layer selection control OS transistor connected to an EN signal, a buffer OS transistor connected to a word line of the OS memory, and an OS transistor that pulls down the word line in response to an ENB signal, and employs a bootstrap circuit to inhibit a Vth drop of the word lines (RWL and WWL) caused by the use of an NMOS transistor as the buffer OS transistor. The OS memory cell is composed of 3Tr1C [32] and has a retention capacity of 2.4 fF and a cell size of 4.304 μm^2. Like the layer selection drivers, the OS memory cells can be stacked using the same mask pattern.

Fig. 6 shows a timing chart of NN switching with context switching and PG, which can be achieved by the above circuits. First, with layer selection signals EN [0] and ENB [0] corresponding to Bank 0 of the ACC memory, the weight data of the first NN is written to the OS memory in the first layer. This makes the system to be ready for starting AI processing using the weight data stored in Bank 0. The register data in this state is stored as context 0 in the OS memory in the first layer of the OSFF in response to the signal BK [0]. Next, similar processing is performed on the second NN, and the obtained OSFF register data is stored as context 1 in the OS memory in the second layer. Then, it is possible to fall in a sleep mode with PG. In restoration from PG, either of the context data needs to be written back to the FF input. When the context 0 is written back in response to the signal RE [0], AI processing can be performed on the first NN immediately after the restoration. After this processing, the context 1 is written back in response to the signal RE [1], so that AI processing can be performed on the second NN with quick switching. Owing to long-term retention of the NN weight data by the OS memory, weight data rewriting is unnecessary and subsequent processing can be performed with flexible and low-latency switching between PG, the first NN processing, and the second NN processing.

V. MEASUREMENT RESULT

The ACC performance, power consumption in each operation mode, the static characteristics of OS transistors, and the data retention characteristics of OS memories were evaluated using the test chip.

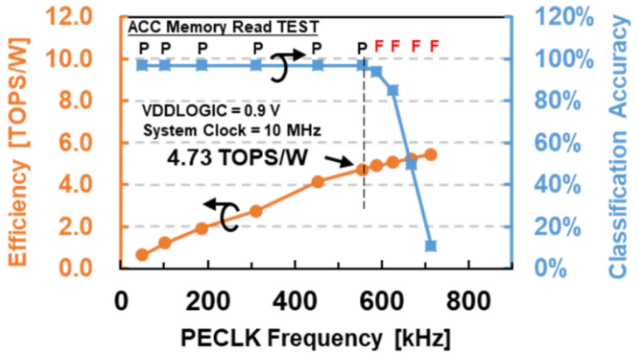


FIGURE 7. ACC performance.

A. ACC PERFORMANCE

The chip evaluation reveals that the energy and processing time for MNIST inference using the ACC were 0.21 μJ and 350 μs, respectively, indicating that ACC incorporation enables inference in accordance with the frame rate of imaging data (e.g., 60 fps and 16 ms), even in a low-performance and low-power microcontroller. The ACC's top energy efficiency was 4.73 TOPS/W (PECLK: 555 kHz, system clock: 10 MHz, including ACC control logic power). The critical pass of the system is memory reading for inference, and the classification accuracy degrades over the maximum frequency. Increasing the operation speed of the OS memory would further improve the performance (Fig. 7).

B. POWER CONSUMPTION

Fig. 8 shows the measurement results of power consumption in each of data writing from an external memory to the ACC memory, inference using ACC, clock gating (CG), and PG. In an active mode, the power consumption for writing data to ACC memory was the largest, which was 595.8 μW, due to large power consumption by the CORE and ACC memory driver circuits. The power consumption of ACC inference was 464.5 μW, which is smaller than that of writing data to the ACC memory because the power consumption of the ACC is increased by the PE operation, but the power consumption of the CORE is greatly reduced. In a standby mode, the power consumption was 5.0 μW in CG and 0.35 μW in PG. Since circuits other than the PMU are powered off in PG, almost only the PMU consumes power.

We also measured power for backup and restoration in PG. In backup and restoration, energy is mainly used to drive the gate of the OS transistor; thus, the energy can be calculated from charge and discharge power and the execution time of 200 ns when backup and restoration are performed concurrently on 4,045 FFs. The results are 510 fJ/bit in backup and 111 fJ/bit in restoration. Although instantaneous power at the time of the backup and restoration in 4,045 FFs becomes a large value on the mW order, the actual energy consumption integrated with respect to time is negligibly small because the execution time of 200ns is short (see Section VI).

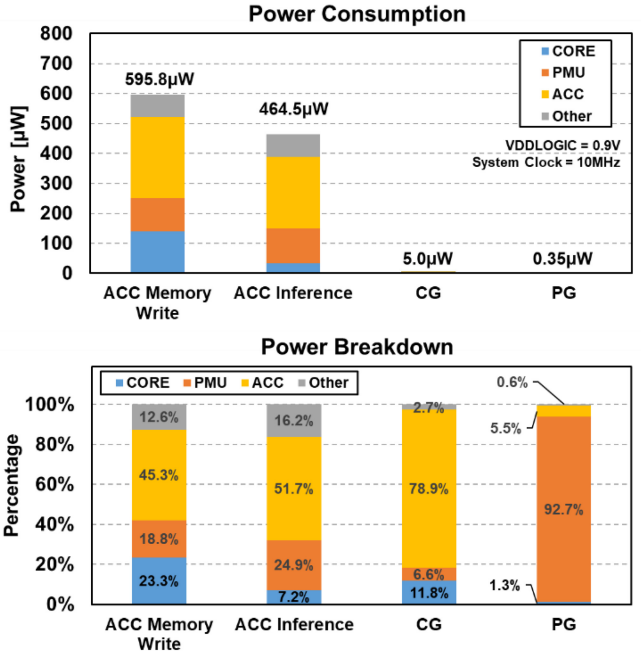
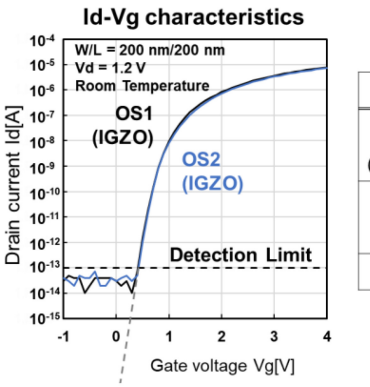


FIGURE 8. Power consumption and breakdown.



Param.	OS1	OS2
Ion [µA] @Vg=3.3V	4.81	4.42
S.S. [mV/dec]	91.03	88.4
Vth [V]	1.12	1.15

FIGURE 9. I_d–V_g characteristics and data retention.

C. ID-VG CHARACTERISTICS & MEMORY RETENTION

Fig. 9 shows the I_d–V_g characteristics of the OS transistors in the first and second layers, which are measured from a test element group of the OS transistors in the OS memory cells placed at the periphery of the chip. These OS transistors are fabricated through the same process as "3D-Stacked CAAC-In-Ga-Zn Oxide FETs with Gate Length of 72 nm" [32]: the OS transistors in the first and second layers each have a channel length of 200 nm and c-axis aligned crystalline In-Ga-Zn oxide (CAAC-IGZO) as a channel material. The threshold voltage is approximately 1.1 V, showing that the OS transistors have normally-off characteristics. The off-state current of the transistor shown in the graph is the lower measurement limit, and is actually very small [32]. The graph also shows that stacking does not cause variation in transistor characteristics. Such transistor characteristics contribute to data retention by the ACC memory. Fig. 10

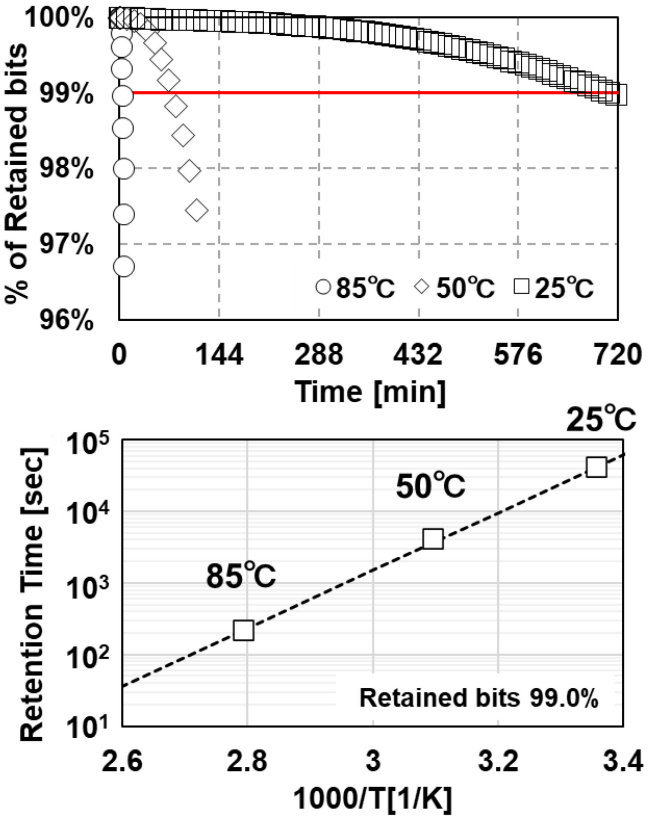


FIGURE 10. Data retention of 32 KB ACC memory.

shows the retention characteristics of the 32 KB memory cells in the ACC memory, which were measured excluding 0.03% of the memory cells with initial defects. The results show that the capability of data retention in 99% of the memory cells is 11 hours at 25 °C, 1 hour at 50 °C, and 5 minutes at 85 °C. This allows a low refresh rate that makes the data refresh energy negligibly small compared with the inference energy.

VI. DISCUSSION

In order to verify the effectiveness of our OS/OS/Si test chip for area reduction and power saving, the OS/OS/Si, OS/Si, and Si (SRAM) structures are compared.

A. ACC BLOCK SIZE VS INCREASING MEMORY BANK

First, the chip area is considered. Fig. 11 shows the ACC block size when SRAM or OS memory is used for NN weight data with a varying number of memory banks. The SRAM size is calculated by an SRAM generator for 130 nm technology. The OS memory cell, which can be stacked over the Si CMOS circuit, can suppress an increase in Si circuit area. Furthermore, Si read and write circuits can be shared by all layers of OS memory in the structure utilizing the monolithic stacking technology of OS transistors, so that the memory capacity can be increased without increasing the area and power consumption of the Si circuit. Accordingly,

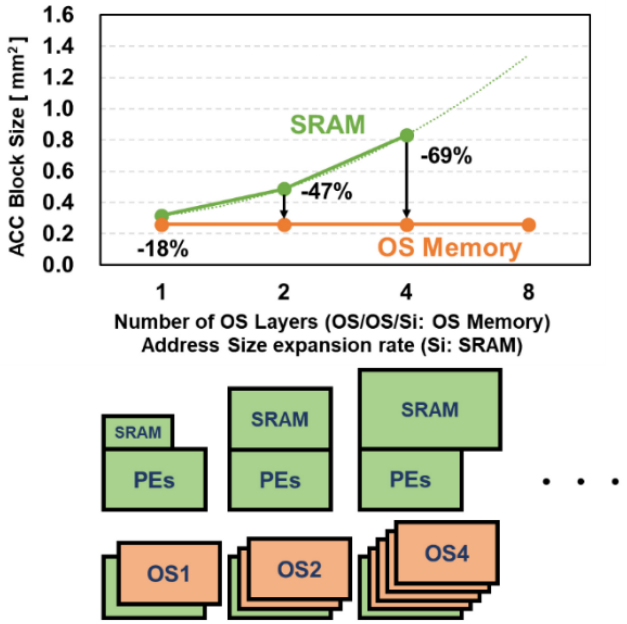


FIGURE 11. Comparison of implementation size.

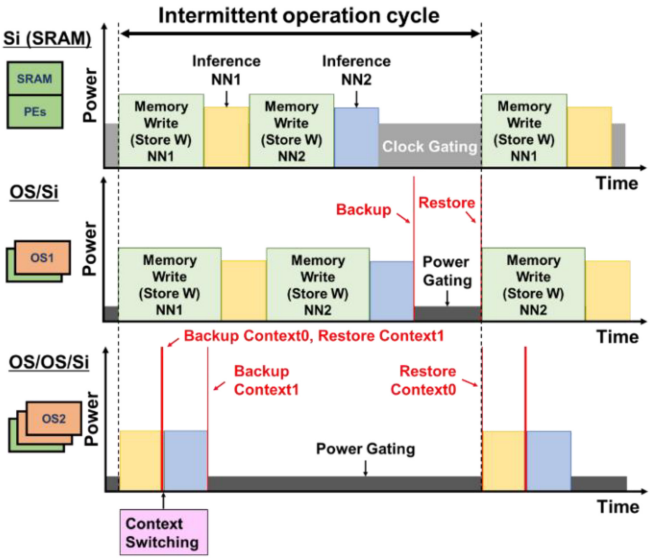


FIGURE 12. Intermittent operation.

it is estimated that the ACC block can be significantly downsized as compared with the Si (SRAM) structure.

B. POWER CONSUMPTION OF INTERMITTENT OPERATION

Next, power consumption is considered. Fig. 12 shows intermittent operation of the Si (SRAM), OS/Si, and OS/OS/Si structures assuming processing where data is obtained at regular intervals. Processing of one operation cycle is performed in the following manner: data analysis is performed using first NN, data analysis is performed using second NN, and then the system is kept in a standby mode until the end of the cycle time. The SRAM, which is a

TABLE 1. Power consumption of each configuration and each operation.

unit μW

Configuration	Si (SRAM)			OS/OS/Si and OS/Si				
Operations	W Write	Inference	CG	W Write	Inference	PG	Backup	Restore
Memory&ACC	1800[a)]	192.2[a)]	15.63[a)]	269.7	240.1	0.00	-	-
PMU	112.1	115.6	0.331	112.1	115.6	0.32	-	-
CORE	139.1	33.45	0.588	139.1	33.45	0.00	-	-
Other	74.95	75.29	0.134	74.95	75.29	0.00	-	-
Summary	2126	416.52	16.68	595.8	464.5	0.34	10315	2225

a) Including simulation values with SRAM generators

TABLE 2. Operation time of each configuration.

unit μs

Configuration	Si (SRAM)			OS/OS/Si and OS/Si				
Operations	W Write	Inference	CG	W Write	Inference	PG	Backup	Restore
	901.1	349.55	T_{CG}	4505.7	349.55	T_{PG}	0.20	0.20

volatile memory, cannot perform PG and only uses CG to save standby power. The OS/Si and Si (SRAM) structures require weight data rewriting before inference, whereas the OS/OS/Si structure, which can quickly switch NN weight data by context switching, does not require weight data loading from an external memory. When PG or context switching is performed, it is necessary to take into account the power consumption due to the backup and restoration processing of all registers' data.

Time average power consumption of the intermittent operation in Fig. 12 is estimated by averaging power P_x with a weight of execution time T_x in each operation, as shown in (1).

$$P_{\mathrm{AVE}} = \frac{\sum_x P_x \times T_x}{\sum_x T_x} \quad x = Wwrite,\ Inference, CG, PG, Backup, Restore \tag{1}$$

Table 1 and Table 2 show P_X and T_X in each structure, which are used in (1). Note that the ACC power in the Si (SRAM) structure is obtained by replacing the OS memory power of the test chip with SRAM power calculated by an SRAM simulator. Standby time T_{CG} and T_{PG} are each a variable that increases with cycle time T_{cycle} of the intermittent operation, and is obtained by subtracting active time T_{active} from T_{cycle}, as shown in (2).

$$T_{CG,PG} = T_{\mathrm{cycle}} - T_{\mathrm{active}} \tag{2}$$

T_{active} varies between the Si (SRAM), OS/Si, and OS/OS/Si structures as shown in (3). In the OS/OS/Si structure, T_{Backup} and $T_{Restore}$ are much smaller than T_{Wwrite}, and accordingly T_{active} is shortened, that is, the standby time T_{PG} is extended.

$$T_{\mathrm{active}} = \begin{cases} 2(T_{\mathrm{Wwrite}} + T_{\mathrm{Inferece}}) & (\mathrm{Si\ only}) \\ 2(T_{\mathrm{Wwrite}} + T_{\mathrm{Inferece}}) + T_{\mathrm{Backup}} + T_{\mathrm{Restore}} & (OS/Si) \\ 2\left(T_{\mathrm{Inference}} + T_{\mathrm{Backup}} + T_{\mathrm{Restore}}\right) & (OS/OS/Si) \end{cases} \tag{3}$$

Fig. 13 shows a graph based on (1) to (3) and Table 1 and Table 2, where a horizontal axis represents T_{cycle} and a vertical axis represents power consumption. The standby time of CG and PG increases with increasing cycle time; thus, when the cycle time is as long as 1,000 ms, the structure including the OS memory capable of PG consumes less power. On the other hand, when the cycle time is as short as 16 ms (60 fps), the power consumption of the OS/Si structure and the Si (SRAM) structure exceeds 100 μW because the standby time is shortened and the power for writing weight data to memory becomes dominant. By contrast, the average power consumption of the OS/OS/Si structure, where operation can be performed with instant switching between two NNs only by weight data switching by context switching, is as low as 21.71 μW, achieving a 92% reduction from 276.44 μW of the Si (SRAM) structure. The results show that our OS/OS/Si structure can reduce power consumption in any cycle time, and verify the effectiveness of the test chip for processing two NNs while achieving a small area and power saving.

VII. CONCLUSION

We fabricated an Noff CPU incorporating a power-saving ACC with use of our OS/OS/Si process technology, where two OS memory layers can be stacked over Si CMOS. ACC memory cells for storing NN weight data are formed over the ACC formed of Si CMOS utilizing the monolithic stacking technology of OS transistors, thereby minimizing an increase in Si circuit area due to local memory implementation. Forming two OS memory layers as bank memories corresponding to context switching eliminate the need for rewriting ACC memory in switching two NNs, contributing to longer PG duration time. Consequently, power consumption can be reduced compared with the structure employing SRAM as the ACC local memory.

These results reveal the potential of our OS/OS/Si process technology for end point AI achieving all of a small chip area, low power consumption, and AI processing with multiple NNs switching.

a) Including simulation values with SRAM generators

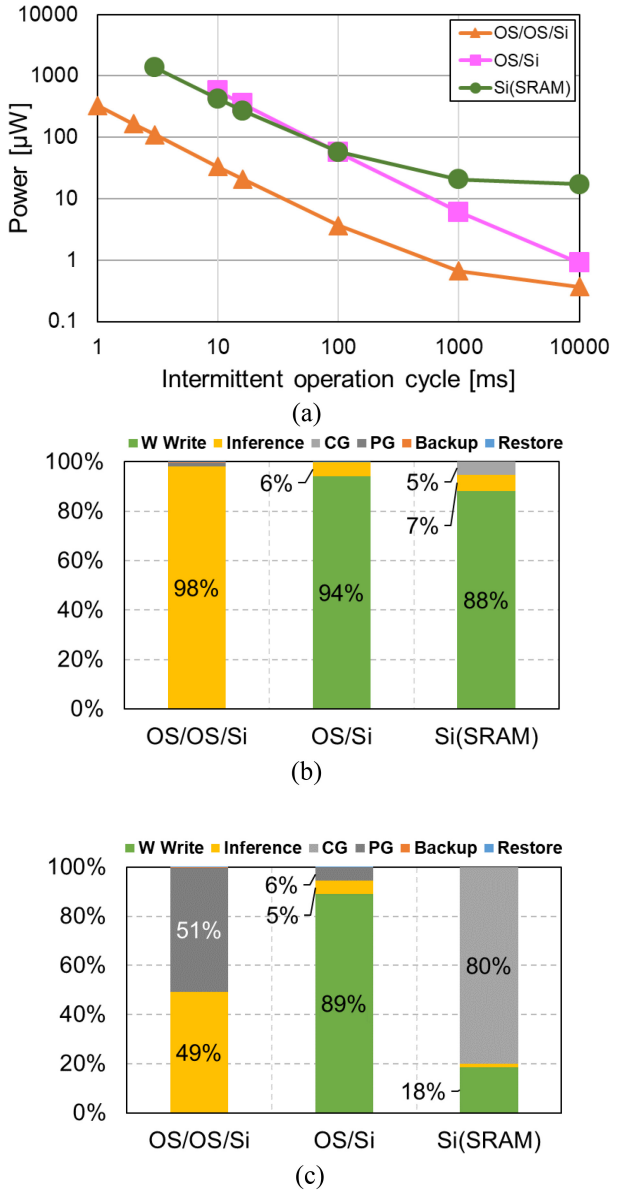


FIGURE 13. (a) Power consumption (b) Power breakdown at 16ms cycle, and (c) Power breakdown at 1000ms cycle.

REFERENCES

[1] S. Deng et al., "Edge intelligence: The confluence of edge computing and artificial intelligence," Feb. 2020, *arXiv:1909.00560v2*.

[2] R. Sachdev, "Towards security and privacy for edge AI in IoT/IoE based digital marketing environments" in *Proc. 5th Int. Conf. Fog Mobile Edge Comput. (FMEC)*, Paris, France, 2020, pp. 341–346.

[3] H. Hu and C. Jiang, "Edge intelligence: Challenges and opportunities," in *Proc. Int. Conf. Comput. Inf. Telecommun. Syst. (CITS)*, Hangzhou, China, 2020, pp. 1–5.

[4] Y. Zhao et al., "Exploiting near-memory processing architectures for Bayesian neural networks acceleration," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, 2019, pp. 203–206.

[5] G. Brown, V. Tenace, and P. E. Gaillardon, "NEMO-CNN: An efficient near-memory accelerator for convolutional neural networks," in *Proc. IEEE 32nd Int. Conf. Appl. Spec. Syst. Architect. Process. (ASAP)*, 2021, pp. 57–60.

[6] M. Hassanpour, M. Riera, and A. González, "A survey of near-data processing architectures for neural networks," in *Proc. Mach. Learn. Knowl. Extraction*, 2022, pp. 66–102.

[7] E. Tam et al., "DRAM-based processor for deep neural networks without SRAM cache," in *Proc. Intell. Comput.*, vol. 284, Jul. 2021, p. 156.

[8] J. Yue et al., "7.5 A 65nm 0.39 to 140.3TOPS/W 1 to 12b unified neural network processor using block-circulant-enabled transpose-domain acceleration with 8.1 × higher TOPS/mm2 and 6T HBST-TRAM-based 2D data-reuse architecture," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2019, pp. 138–140.

[9] Y. Ju and J. Gu, "15.2 a 65nm systolic neural CPU processor for combined deep learning and general-purpose computing with % PE utilization, high data locality and enhanced end to end performance," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2022, pp. 248–249.

[10] Z. Chen, X. Chen, and J. Gu, "15.3 a 65nm 3T dynamic analog RAM-based computing-in-memory macro and CNN accelerator with retention enhancement, adaptive analog sparsity and 44TOPS/W system energy efficiency," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2021, pp. 240–242.

[11] H. Zhu et al., "COMB-MCM: Computing-on-memory-boundary NN processor with bipolar bitwise sparsity optimization for scalable multi-chiplet-module edge machine learning," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1–3.

[12] J. Park, J. Lee, and D. Jeon, "7.6 A 65nm 236.5nJ/classification neuromorphic processor with 7% energy overhead on-chip learning using direct spike-only feedback," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2020, pp. 140–142.

[13] Z. Yuan et al., "14.2 A 65nm 24.7μJ/frame 12.3mW activation-similarity-aware convolutional neural network video processor using hybrid precision, inter-frame data reuse and mixed-bit-width difference-frame data codec," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2020, pp. 232–234.

[14] S. Park, S. Lee, J. Park, H. S. Choi, and D. Jeon, "22.8 A0.81 mm2 740μW real-time speech enhancement processor using multiplier-less PE arrays for hearing aids in 28nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2020, pp. 21–23.

[15] I. Miro-Panades et al., "SamurAI: A 1.7MOPS-36GOPS adaptive versatile IoT node with 15,000× peak-to-idle power reduction, 207ns wake-up time and 1.3TOPS/W ML efficiency," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, 2020, pp. 1–2.

[16] K. Kato et al., "Evaluation of off-state current characteristics of transistor using oxide semiconductor material, indium-gallium-zinc oxide," *Jpn. J. Appl. Phys.*, vol. 51, no. 28, 2012, Art. no. 21201.

[17] N. Saito et al., "High mobility (>30 cm2/Vs) and low S/D parasitic resistance In-Zn-O BEOL transistor with ultralow (<10-20 A/ μm) off leakage current," *Jpn. J. Appl. Phys.*, vol. 58, no. 1, 2019, Art. no. SBBJ07.

[18] J. Guo et al., "A new surface potential and physics based compact model for a-IGZO TFTs at multinanoscale for high retention and low-power DRAM application," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2021, pp. 8.5.1–8.5.4.

[19] S. Subhechha et al., "Ultra-low leakage IGZO-TFTs with raised source/drain for Vt 0 V and Ion 30 μA/μm," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2022, pp. 292–293.

[20] T. Onuki et al., "Fabrication of dynamic oxide semiconductor random access memory with 3.9fF storage capacitance and greater than 1 h retention by using c-axis aligned crystalline oxide semiconductor transistor with L of 60 nm," *Jpn. J. Appl. Phys*, vol. 54, no. 2, 2015, Art. no. 4DD07.

[21] M. Oota et al., "3D-stacked CAAC-in-ga-zn oxide FETs with gate length of 72nm," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 3.2.1–3.2.4.

[22] H. Sawai et al., "Improved C-axis-aligned crystalline oxide semiconductor FET suitable for scaling and monolithic stacking for higher integration of integrated circuit," in *Proc. Int. Conf. Solid-State Devices Mater. (SSDM)*, 2023, pp. 225–226.

[23] T. Ohmaru et al., "Eight-bit CPU with nonvolatile registers capable of holding data for 40 days at 85°C using crystalline In–Ga–Zn oxide thin film transistors," in *Proc. Int. Conf. Solid-State Devices Mater. (SSDM)*, 2020, pp. 1144–1145.

[24] H. Tamura et al., "Embedded SRAM and Cortex-M0 core with backup circuits using a 60-nm crystalline oxide semiconductor for power gating," in *Proc. IEEE COOL Chips XVII*, Yokohama, Japan, 2014, pp. 1–3.

[25] T. Ishizu et al., "A 48 MHz 880-nW standby power normally-off MCU with 1 clock full backup and 4.69-μs wakeup featuring 60-nm crystalline In–Ga–Zn Oxide BEOL-FETs," in *Proc. Symp. VLSI Circuits*, Kyoto, Japan, 2019, pp. C48–C48.

[26] Y. Yakubo et al., "Crystalline oxide semiconductor-based 3D bank memory system for endpoint artificial intelligence with multiple neural networks facilitating context switching and power gating," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 12–14.

[27] N. J. Fraser et al., "Scaling binarized neural networks on reconfigurable logic," Jan. 2017, *arXiv:1701.03400*.

[28] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training deep neural networks with binary weights during propagations," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Dec. 2015, pp. 3123–3131.

[29] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," Aug. 2016, *arXiv:1603.05279v4*.

[30] S. Zhu, L. H. K. Duong, and W. Liu, "XOR-Net: An efficient computation pipeline for binary neural network inference on edge devices," in *Proc. IEEE 26th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Hong Kong, 2020, pp. 124–131.

[31] S. Maeda et al., "A 20ns-write 45ns-read and 1014-cycle endurance memory module composed of 60nm crystalline oxide semiconductor transistors," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, 2018, pp. 484–486.

[32] H. Takahashi et al., "Soft- and hard-error radiation reliability of 228 KB 3T+1C oxide semiconductor memory," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Monterey, CA, USA, 2023, pp. 1–6.