IEICE TRANSACTIONS

on Electronics

DOI:10.1587/transele.2023ECP5051

Publicized:2024/05/08

A PUBLICATION OF THE ELECTRONICS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Area-efficient Binarized Neural Network Inference Accelerator Based on Time-multiplexed XNOR Multiplier Using Loadless 4T SRAM

Yihan ZHU[†a)], *Nonmember and* **Takashi OHSAWA**[†b)], *Member*

SUMMARY A binarized neural network (BNN) inference accelerator is designed in which weights are stores in loadless four-transistor static random access memory (4T SRAM) cells. A time-multiplexed exclusive NOR (XNOR) multiplier with switched capacitors is proposed which prevents the loadless 4T SRAM cell from being destroyed in the operation. An accumulator with current sensing scheme is also proposed to make the multiply-accumulate operation (MAC) completely linear and read-disturb free. The BNN inference accelerator is applied to the MNIST dataset recognition problem with accuracy of 96.2% for 500 data and the throughput, the energy efficiency and the area efficiency are confirmed to be 15.50TOPS, 72.17TOPS/W and 50.13TOPS/mm^2, respectively, by HSPICE simulation in 32nm technology. Compared with the conventional SRAM cell based BNN inference accelerators which are scaled to 32nm technology, the synapse cell size is reduced to less than 16% (0.235μm^2) and the cell efficiency (synapse array area/synapse array plus peripheral circuits) is 73.27% which is equivalent to the state-of-the-art of the SRAM cell based BNN accelerators.

key words: *In-memory computing, binarized neural network (BNN), loadless 4T SRAM, time-multiplexed XNOR multiplier, current sensing scheme, MNIST dataset*

1. Introduction

In-memory computing has been extensively studied for the purpose to alleviate the von Neumann bottleneck in some specific fields [1]. Inference accelerators of deep neural networks (DNNs) are suitable applications of the in-memory computing architecture [2]. Compared with the all-digital accelerators, they largely improve throughput and energy efficiency. Since SRAM is the most user-friendly memory with unlimited endurance and is fabricated by a process which is almost compatible with the CMOS one, many works on inference accelerators based on SRAM cells have been published [3-7].

However, when the conventional 6TSRAM cell is used in DNN inference accelerators, it suffers from the nonlinearity and the read disturb under the condition that multiple word-lines (WLs) are selected at the same time [8]. As the BL is discharged to a lower voltage, the access transistors of the 6TSRAM cell enter the linear region, which makes the drain current depend on the BL voltage. This leads to the nonlinear dependence of the BL voltage on the number of cells discharging the BL. In addition to this, the lowest voltage of a bit-line (BL) in multiply-accumulate (MAC) operations must be larger than a write trigger voltage to avoid the read disturb [8]. These reduce the signal-to-noise ratio (SNR) in the MAC operations and impose the maximum number of cells which are to be connected to a BL [8]. The charge-domain scheme was proposed to solve the nonlinearity issue [4]. However, it requires control circuits for WL pulse width, which makes the array size larger and the system vulnerable to process variations. To avoid the read disturb, non-6T SRAM (8T, 10T, etc.) cells have been adopted in which their read paths are separated from the write paths [3, 9-14]. However, they obviously make the crossbar array size larger and the nonlinearity issue remains unsolved.

In this paper, we propose a current sensing scheme for an XNOR multiplier based on the loadless 4T SRAM cell [15-21] which is applied to the MAC operation in binarized neural network (BNN) inference accelerators. The current sense amplifiers for SRAM were proposed about 30 years ago [22-24]. They receive a current from an SRAM cell and detect the signal without discharging a BL. The advantage in this scheme is that the access time is fast and almost independent of the BL capacitance. The proposed current sensing scheme is customized for the XNOR multiply operation so that the current drawn from the multiple loadless 4T SRAM cells are accumulated without affecting the BL voltage and disturbing the cells' data states. Thus, the BL read disturb issue in loadless 4T SRAM BNN inference accelerators can be avoided, and the BL current is totally linear to the MAC values, making it possible to design a high-density crossbar array for a BNN accelerator based on the loadless 4T SRAM cell. To avoid the data disturb of the loadless 4T SRAM cells during the XNOR multiply operation, a time-multiplexed XNOR operational architecture for the loadless 4T SRAM multiplier is proposed. It is shown that the accuracy of the MNIST dataset recognition, the energy efficiency and the area efficiency for the proposed BNN accelerator are 96.2%, 72.17 TOPS/W and 50.13 TOPS/mm^2, respectively.

† The authors are with Graduate School of Information, Production and Systems, Waseda University, Kitakyushu-shi, 808-0135 Japan.
a) E-mail: zyh@akane.waseda.jp
b) E-mail: ohsawa@kxa.biglobe.ne.jp

2. Current Sense Accumulator for SRAM Cells

Current sense amplifiers were proposed for the purpose of achieving SRAM's fast access time [22-24]. Contrary to the conventional voltage sense amplifier, they amplify the signal read from a 6TSRAM cell based on the current supplied to the amplifiers. Since they don't require BLs to be discharged, the access time is substantially independent of the BL capacitance. The basic concept of the current sense amplifier can be applied to the MAC operations in DNN inference accelerators. This scheme makes it possible to use 6TSRAM cells for the weight memory cells without the read disturb and the nonlinearity issues which must be considered in the conventional DNN inference accelerators using 6TSRAM cells [8, 9]. Fig. 1 shows the principle of the circuits which achieve MAC operations in a BNN inference accelerator. The 6TSRAM cell works as an XNOR operator with the input signals of complementary WLs (WL_i, WLB_i; i=1, ···, m) and a pair of BLs combined together. The write operation to the cells is conducted by making BL low and selecting only one of the pair of WLs (selecting either WL or WLB). The gate voltage of a PFET load for a BL is controlled by an operational amplifier (OP-AMP) with a negative feedback so that the BL voltage is clamped at a constant voltage V_{ref} which is lower than the power supply voltage V_{dd}. The current that flows through the PFET is the summation of all the current drawn by the cells with XNOR=+1 connected to the BL. This accumulated current is mirrored to another smaller PFET whose drain is set to V_{ref} by another negative feedback by a smaller OP-AMP and an NFET. The current is converted to an analog voltage V_{out} by a load resistor R_L.

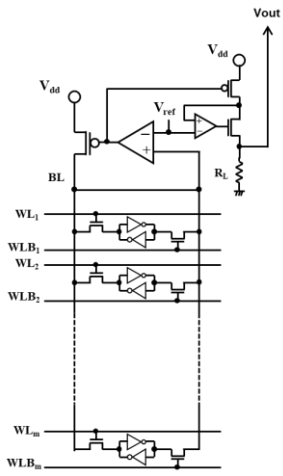


Fig. 1 Circuits for multiply-accumulate (MAC) operation based on XNOR 6TSRAM with current sensing scheme in deep neural networks (DNNs).

This current sense accumulator clamps BLs at V_{ref} even if many WLs are raised at the same time. The 6TSRAM cell's power supply voltage and the high voltage of WLs are also set to be V_{ref}. The read disturb problem, thus, is completely avoided. Furthermore, since the BL voltage is constant regardless of the MAC values, the current and the converted voltage V_{out} are totally linear to the number of cells which draw the current or whose XNOR value is +1. This makes an offline training's forward path equivalent to the inference accelerator's forward operation. Since this ensures the consistency of the weight optimization by the training software with the weight values stored in the 6TSRAM cells, higher inference accuracies are expected in the high-density accelerator

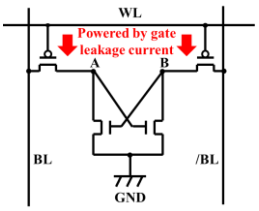


Fig. 2 The loadless 4T SRAM cell powered by the gate tunneling leakage current with BL disturb free [18-21].

Fig. 2 shows the loadless 4T SRAM cell whose data are sustained by the access PFETs' subthreshold leakage currents [15-17] or their gate leakage currents [18-21]. The cell size is reduced by 17% compared with the 6TSRAM cell (see Fig. 21). Fig. 3 shows the comparison of the write trigger voltages between the 6TSRAM cell and the loadless 4T SRAM cell (It is shown that the write margin of the loadless 4T SRAM is larger than the 6TSRAM [19].). It is shown that the write trigger voltage of the loadless 4T SRAM cell is much higher than that of the 6TSRAM cell. The read-disturb issue is thus more sever for the loadless 4T SRAM than the 6TSRAM. Fig. 4 shows the concept of the current sensing scheme which is straightforwardly applied to the XNOR multiplier based on the loadless 4T SRAM cell. The operation of this scheme will be analyzed in the next section and will be shown that this configuration does not work as expected.

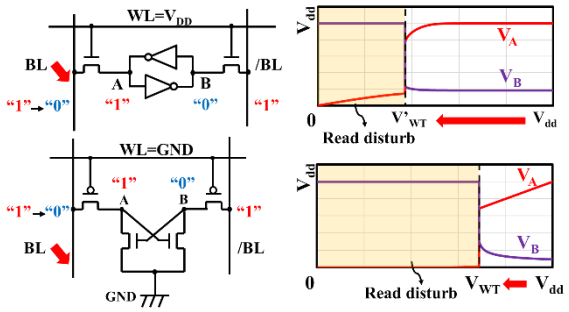


Fig. 3 Comparison of write trigger voltages between the 6TSRAM cell and the loadless 4T SRAM cell. The higher write trigger voltage makes the loadless 4T SRAM more vulnerable to the read-disturb under multi-WL selection

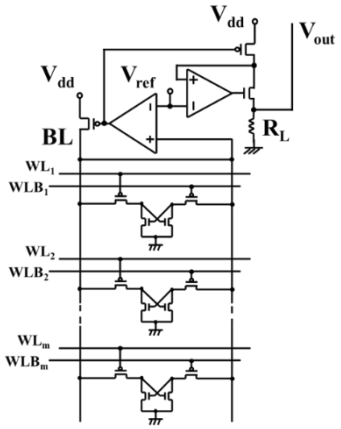


Fig. 4 A straightforward application of the current sensing scheme used in the XNOR 6TSRAM cells (Fig. 1) to the XNOR loadless 4T SRAM cells in the DNN. This scheme is not shown to be feasible in section 3.

3. Time-multiplexed XNOR

The left schematic in Fig. 5 (a) is the loadless 4T SRAM synapse cell based on the conventional XNOR reading mode shown in Fig. 4 whose parameters are shown in Table 1 (a). The weight is stored in the cells' nodes A and B. The two word-lines WL_m and WLB_m provide two input data +1 (WL_m=low and WLB_m=high) and -1 (WL_m=high and WLB_m=low). The read result is based on the current value which flows from the bit-line (BL) whose voltage is fixed at V_{ref} through the synapse cell to GND. The right waveform in Fig. 5 (a) shows the simulated loadless 4T SRAM nodes' voltages during reading operation for the input +1 which is based on the conventional XNOR logic. Though the gate leakage of Q1 provides enough current to sustain the data in the loadless 4T SRAM cell during the holding period [18-21], this current is much smaller than the current flowing through Q2 to GND due to a high voltage in the gate of Q2 by the current flow from BL to the node B. So, the synapse suffers from the read disturb under the conventional XNOR reading mode, showing that the XNOR multiplier and the accumulator scheme in Fig. 4 does not work.

The left schematic in Fig. 5 (b) is a synapse cell which is based on the proposed time-multiplexed XNOR read mode whose parameters are shown in Table 1 (a). This is nothing other than the conventional read mode for the loadless 4T SRAM cell to make the currents which flow from BL and BLB to the nodes A and B balanced for avoiding the read disturb. The synapse cell is designed in the same manner as our previous cell [18-21]. The BL and BLB's voltage will be clamped at a reference voltage V_{ref} during this proposed XNOR read mode. The input data on WL will no longer be a constant voltage value during the

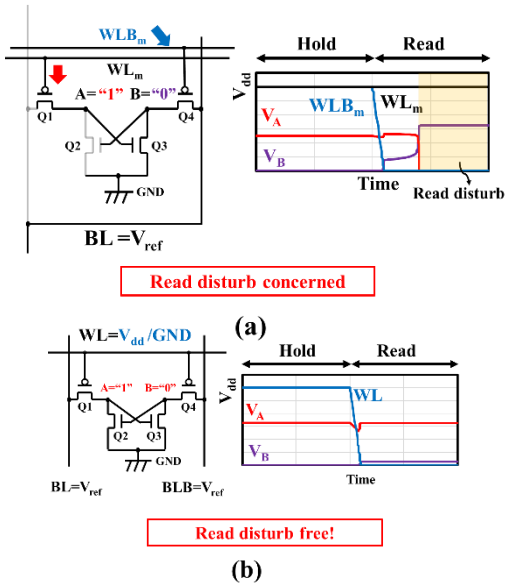


Fig. 5 (a) The loadless 4T SRAM synapse cell suffers from the read disturb under the XNOR logic operation shown in Fig. 4. (b) The proposed loadless 4T SRAM synapse cell which is accessed symmetrically from the pair of PFETs is free from the read-disturb.

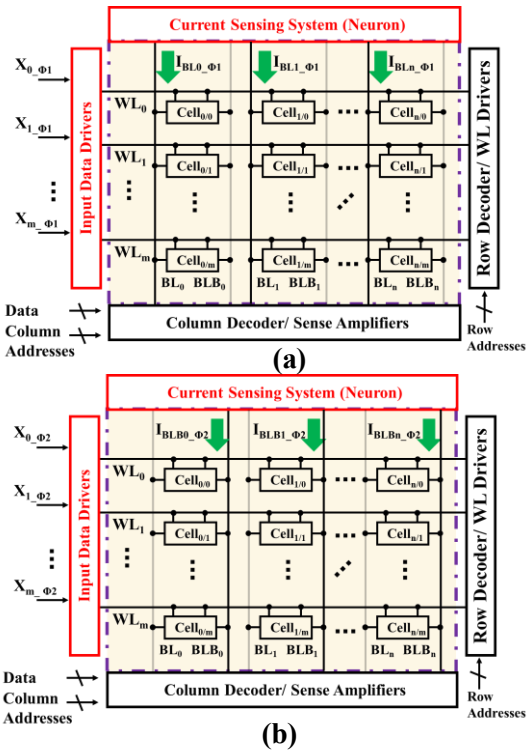


Fig. 6 (a) The operation of the proposed BNN inference accelerator with the synapse cell array in the time phase Φ1. In this phase (Φ1), the information on each current which flows through each BL is stored in each neuron. (b) The operation of the proposed BNN inference accelerator with the synapse cell array in the time phase Φ2. In this phase (Φ2), the information on each current which flows through each BLB is also stored in each neuron.

proposed XNOR read mode but a step voltage signal applied on WL during the two different timing phases $V_{WL\Phi1}$ or $V_{WL\Phi2}$. Due to the simultaneous opening and closing of Q1 and Q4 in the Fig. 5 (b), the input currents which flows into the cell nodes A and B are balanced without the read disturb. The write operation to the synapse cell is the same as the write operation to the loadless 4T SRAM cell, i.e., only a selected WL is 0V and unselected WLs are kept at V_{dd}. Therefore, there is no concern about disturbing other cells by this write operation. The write operation is to be conducted by the sense amp. circuit after disabling the neuron circuit by setting V_{C1}=0V and V_{CB1}=V_{dd} (see Fig. 10 and Fig. 11).

Fig. 6 (a) and (b) show the proposed BNN inference accelerator with the synapse cell array consisting of the loadless 4T SRAM cells and the neuron circuits consisting of the current sensing circuits. The input data drivers are also attached to the synapse cell array. In addition to them, a row decoder with WL drivers and a column decoder with sense amplifiers are attached to the cell arrays for setting the weight data to the synapses before conducting the inference and for testing the cell functionality in advance by the conventional memory access mode. In the proposed XNOR read mode, BL and BLB voltages are clamped at V_{ref} and all synapses in the array are accessed simultaneously. In the memory access mode for setting the weights in the array or

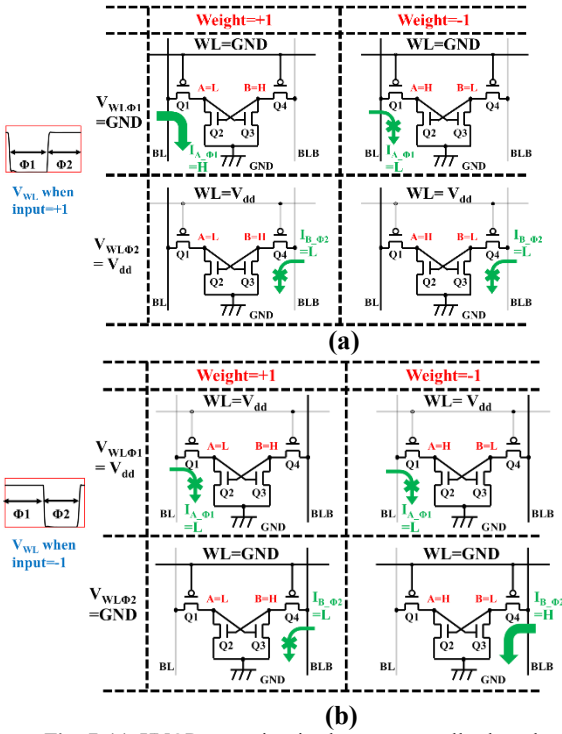


Fig. 7 (a) XNOR operation in the synapse cell when the input data is +1. (b) When the input data is -1.

Output XNOR value (y)		$I_{A\_\Phi1}$ and $I_{B\_\Phi2}$ current sum		Data node A/B voltage (W)		
				+1	-1	Data node
				L	H	A
				H	L	B
Input Value (X)	+1	GND	V_{dd}	+1/High	-1/Low	
	-1	V_{dd}	GND	-1/Low	+1/High	
	Data	$V_{WL\Phi1}$	$V_{WL\Phi2}$			

Fig. 8 Logical table for the synapse operation.

in the test mode for checking the functionality of all the synapse cells in the array, WL and BL/BLB's voltages are controlled by the row and the column decoders, respectively. And the data are written and read by the conventional sense amplifiers in the bit-by-bit manner.

Fig. 7 (a) and (b) show the proposed XNOR read mode's operation. The WL voltages under the situations

$$\begin{cases} V_{\mathrm{WL}\Phi1} = GND \\ V_{\mathrm{WL}\Phi2} = V_{dd} \end{cases} \quad (1)$$

or

$$\begin{cases} V_{\mathrm{WL}\Phi1} = V_{dd} \\ V_{\mathrm{WL}\Phi2} = GND \end{cases} \quad (2)$$

are defined as the input data equal to +1 or -1, respectively. There are two time-multiplexed phases: phase1 (Φ1) and phase2 (Φ2). The situation (1) corresponds to the condition in which WL voltage transitions from low to high from Φ1 to Φ2, while the situation (2) corresponds to the condition in which WL voltage transitions from high to low from Φ1 to Φ2. The voltages in the nodes A and B under the situations

$$\begin{cases} V_{\mathrm{A}} = Low \\ V_{\mathrm{B}} = High \end{cases} \quad (3)$$

or

$$\begin{cases} V_{\mathrm{A}} = High \\ V_{\mathrm{B}} = Low \end{cases} \quad (4)$$

are defined as the weight equal to +1 or -1, respectively. The current flowing from BL to node A in Φ1 and the current flowing from BLB to node B in Φ2 are defined as $I_{A\_\Phi1}$ and $I_{B\_\Phi2}$, respectively. The summed value of $I_{A\_\Phi1}$ and $I_{B\_\Phi2}$ is high (+1) or low (-1) depending on the weight and the input data as shown in Fig. 8. This table shows that the summed value of $I_{A\_\Phi1}$ and $I_{B\_\Phi2}$ represents XNOR operation between the weight and the input data. This reflects the mathematical principle:

$$\overline{\mathrm{A} \oplus \mathrm{B}} = \mathrm{A} \cdot B + \bar{A} \cdot \bar{B}, \quad (5)$$

where A and B stand for input and weight, respectively. The first term on the right-hand side is the result in the phase 1, while the second term is the result of phase 2. The currents flow through BL_n and BLB_n from the current sensing system to the synapse array are defined as $I_{BLn\_\Phi1}$ and $I_{BLBn\_\Phi2}$ as shown in Fig. 6 (a) and (b), respectively. Based on the summed value of $I_{BLn\_\Phi1}$ and $I_{BLBn\_\Phi2}$, the current sensing system can generate a multiply (XNOR)-accumulate (MAC) voltage signal and output data to the next layer.

4. Crossbar Array and Current Sensing System Design

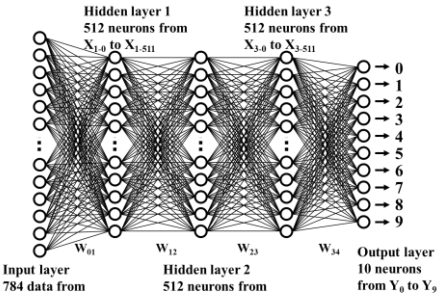


Fig. 9 DNN design for the MNIST dataset classification used in this work.

We design a crossbar array (CBA) which corresponds to the DNN for the MNIST dataset classification as shown in Fig. 9 [25]. There are one input layer, three hidden layers and one output layer. The input layer contains 784 (28×28) different voltage signals. There are 512 neurons and 10 output voltage signals in each hidden layer and output layer, respectively.

Fig. 10 (a) shows one of the first hidden layer neuron's current sensing systems for the proposed CBA by using the proposed XNOR read mode whose parameters are shown in Table 1 (b). The BL and the BLB's voltages V_{BL} and V_{BLB} are clamped at a constant voltage V_{ref} by two PFETs (Q_{BlinP} for the first hidden layer and Q'_{BlinP} for the other layers) with negative feedback loops using two operational amplifiers (OPA1). During the phase Φ1 and the phase Φ2, two OPA1s transmit the signals $V_{out1\_0\_j}$ and $V_{out2\_0\_j}$ to the different two storage capacitors whose capacitance value equal to C_0 by the signal CLK1_0 and CLK2_0 as shown in Fig. 10 (a), respectively. The voltages on the two storage capacitors are named as $V_{out1C\_0\_j}$ and $V_{out2C\_0\_j}$ as shown in Fig. 10 (a), respectively. After the $V_{out1\_0\_j}$ and the $V_{out2\_0\_j}$ transmissions complete, the currents that flows through the two PFETs

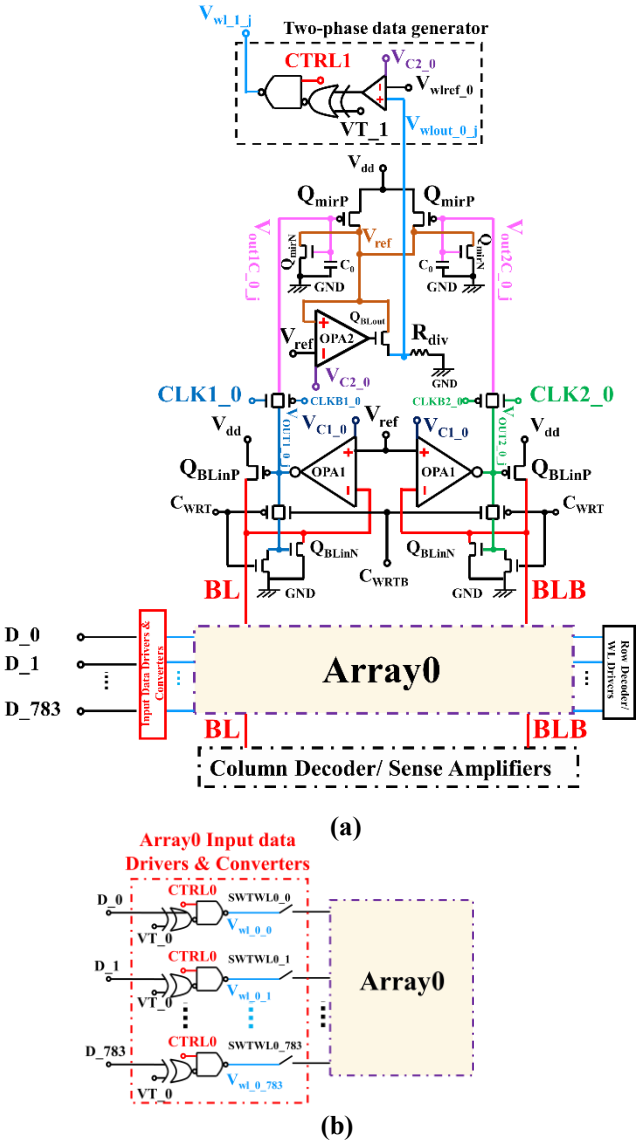
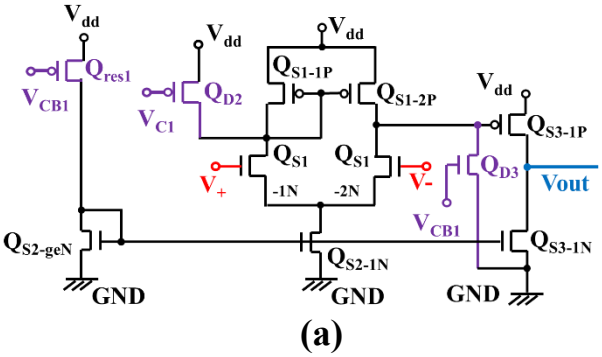


Fig. 10 (a) Schematic of one of the neuron circuits in array0 based on the proposed time-multiplexed XNOR loadless 4T SRAM multipliers. (b) Schematic of the input data converters.

(Q_{mirP}) are respectively mirrored whose drain voltages are clamped at V_{ref} by another negative feedback loop using an operational amplifier (OPA2). The total current flowing through the two Q_{mirP} is converted to a voltage $V_{wlout\_0\_j}$ by a resistor (R_{div}). Fig. 11 (a) and (b) show the schematics of the



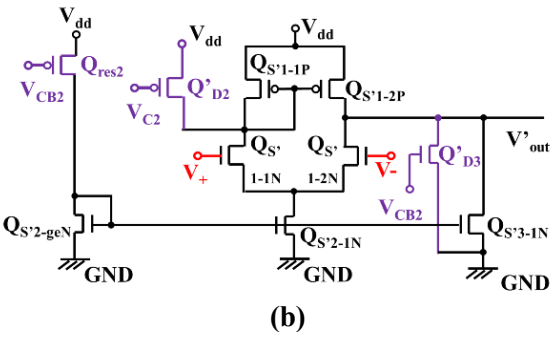


Fig. 11 (a) Schematic of op-amp OPA1. (b) Schematic of op-amp OPA2.

two operational amplifiers OPA1 and OPA2, respectively. By the signal CTRL1, VT_1, V_{C2_0} and the reference voltage for binarization V_{wlref_0}, $V_{wlout\_0\_j}$ is converted to a two-time phase step voltage signal WL_{1_j} as shown in Fig. 10 (a). The input layer's 784 data from D_0 to D_783 are converted to from $V_{wl\_0\_0}$ to $V_{wl\_0\_783}$ in the same way as shown in Fig. 10 (b), respectively.

5. Simulation Results and Layout

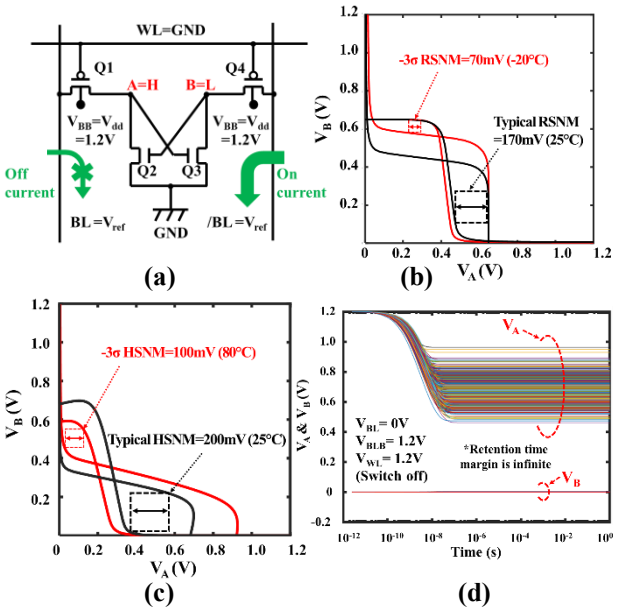


Fig. 12 (a) The proposed synapse cell's bias and current conditions with transistor definition. (b) Butterfly curves in read for the proposed synapse cell. Typical and -3σ worst RSNMs equal to 170mV (25℃) and 70mV (-20℃), respectively. (c) Butterfly curves in hold for the proposed synapse cell. Typical and -3σ worst HSNMs equal to 200mV (25℃) and 100mV (80℃), respectively. (d) Retention curve 1000 Monte Carlo simulation where the BL pairs' voltages are opposite to the data node voltages.

Fig. 12 (a) shows the equivalent circuit of the proposed loadless 4T SRAM synapse cell during the read period. The black curves in Fig. 12 (b) and (c) show the butterfly curves of the proposed synapse cell during the read (proposed XNOR read mode) and the hold periods under 0V and 1.2V word-line high voltage (V_{WL}) and 0.65V V_{BL}/V_{BLB} at 25℃, respectively. The backgate (N well) voltages of the access PFETs Q1 and Q4 are V_{dd}=1.2V. Table 1 (a) shows the cell's

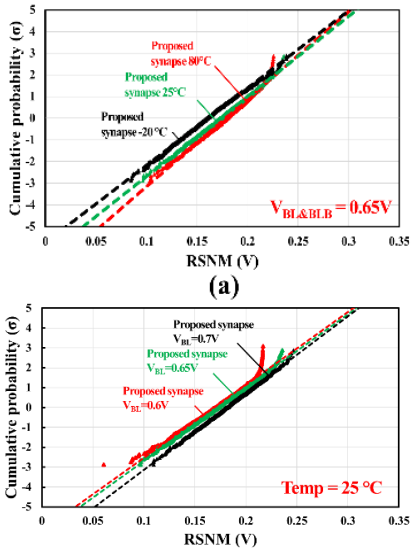


Fig. 13 (a) Cumulative probability distributions for RSNM of the proposed synapse cell obtained by Monte Carlo simulation under the different temperatures with V_{BL}=V_{BLB}=0.655V. (b) Cumulative probability distributions for RSNM of the proposed synapse cell obtained by Monte Carlo simulation under the different V_{BL}=V_{BLB} with Temp=25°C.

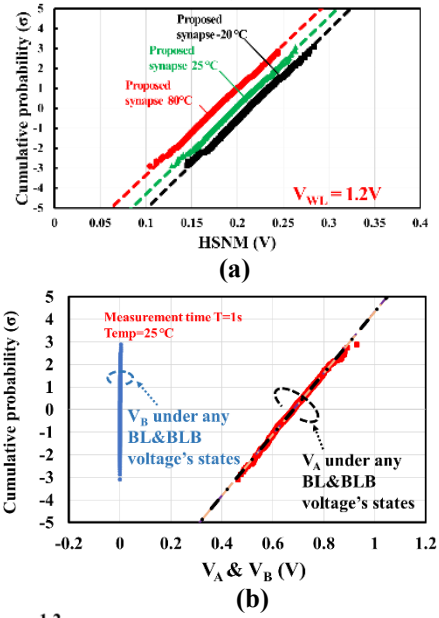


Fig. 14 (a) Cumulative probability distributions for HSNM of the proposed synapse cell obtained by Monte Carlo simulation under the different temperatures with V_{WL}=1.2V. (b) Cumulative probability distributions for data node voltages of the proposed synapse cell obtained by 1000 Monte Carlo simulation at 25 ℃ with V_{WL}=1.2V and all possible BL pairs voltage states at Time=1s for the transient simulations such as in Fig 12 (d).

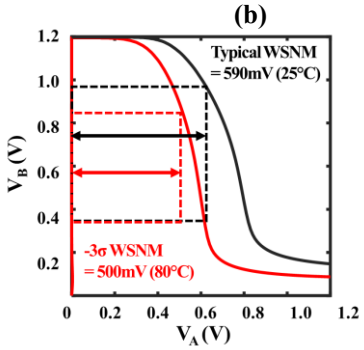


Fig. 15 Butterfly curves in write for the proposed synapse cell and the definition of the write static noise margin (WSNM). The typical WSNM and the -3σ worst WSNM are equal to 590mV at 25 ℃ and 500mV at 80 ℃, respectively.

Table 1 (a) Proposed synapse cell's parameters.

Parameter	Loadless 4TSRAM	
Technology node	32nm LP	
Transistor	Q2,Q3 (N)	Q1,Q4 (P)
Average EOT (nm)	1.14	0.95
V_{th} (V) /σV_{th} (6% V_{th}) (mV)	0.48 /27	-0.50 /30
W/L (nm)	40/40	40/50
σEOT (for ON /OFF state) (Å)	0.16 /0.60	0.13 /0.60

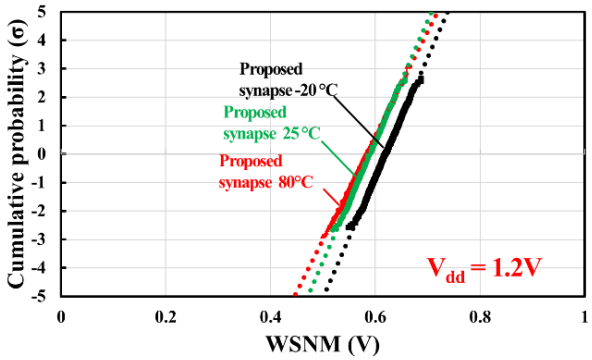


Fig. 16 Cumulative probability distributions for WSNM of the proposed synapse cell obtained by Monte Carlo simulation under the different temperatures with V_{dd}=1.2V.

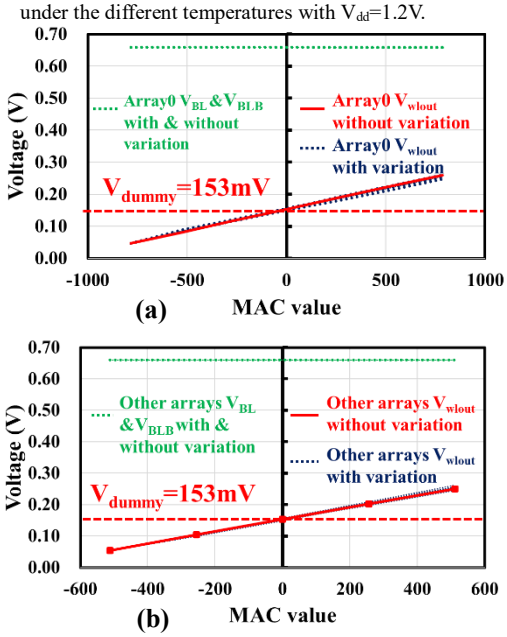


Fig. 17 (a) the simulation results for V_{wlout} vs. MAC value for Array 0. The MAC value's range is from -784 to +784. (b) The similar results for other arrays for the MAC value ranging from -512 to +512. In the both graphs, we added 10 different chips' relations by dotted lines with their V_{th} and EOT fluctuated according to the local variations based on Table I shown below. We also added the voltage fluctuation for BL and BLB for the 10 different chips.

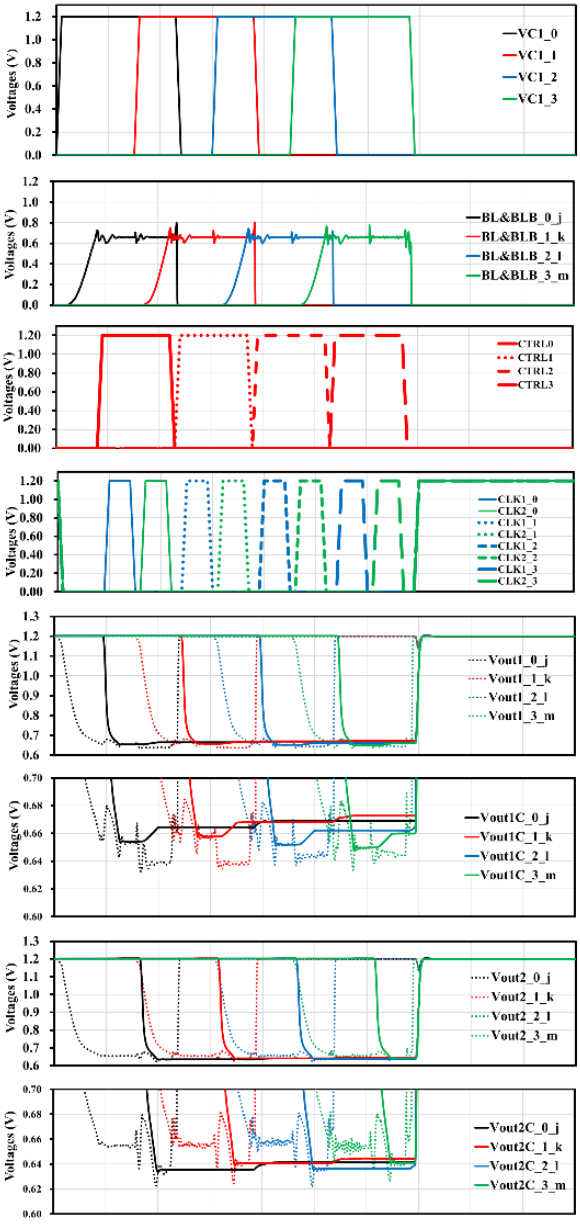
Table 1 (b) Neuron circuits' parameters. Q_{BLinN} and Q_{BLinP} are for the first layer. Q'_{BLinN} and Q'_{BLinP} are for the other layers.

Transistor	Q_{BLinN}(N) (Q'_{BLinN})	Q_{BLinP}(P) (Q'_{BLinP})	Q_{mirN}(N)	Q_{mirP}(P)	Q_{BLout}(N)
EOT (nm)	1.6	1.6	1.6	1.6	1.6
V_{th} (V)	0.48/0.48	-0.50/-0.50	0.48	-0.50	0.48
W/L	0.8μm (0.52μm)/50nm	12μm (7.8μm)/50nm	50nm/50nm	750nm/50nm	150nm/50nm
Other	R_{div}		C_0		Technology
	15.5KΩ		10fF		32nm LP

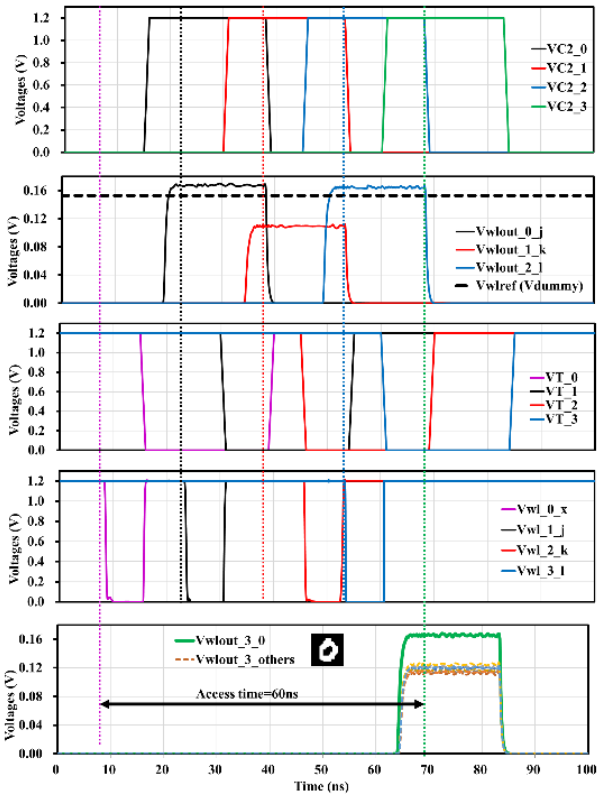
Table 1 (c) -3σ worst case's proposed synapse cell's parameters

	-3σ RSNM Sim.	-3σ HSNM Sim.	-3σ WSNM Sim.
Q2's V_{thn} (V)	0.538	0.453	0.500
Q3's V_{thn} (V)	0.474	0.475	0.424
Q1's V_{thp} (V)	-0.441	-0.467	-0.490
Q4's V_{thp} (V)	-0.547	-0.542	-0.568
Q2's EOT_n (nm)	1.143	1.155	1.145
Q3's EOT_n (nm)	1.137	1.150	1.140
Q1's EOT_p (nm)	0.945	0.770	0.955
Q4's EOT_p (nm)	0.969	1.040	0.932
-3σ worst case (mV)	70 (@-20°C)	100 (@80°C)	500 (@80°C)

parameters which are used in the simulations. The proposed synapse's read static noise margin (RSNM) and hold static noise margin (HSNM) under the proposed XNOR read mode and the hold mode are shown in Fig. 12. Almost the same static noise margins are guaranteed in the read and the hold conditions. It is observed that V_B (V_A) becomes close to 0V (6.56mV) when V_A (V_B) =0.65V in the read butterfly curves. This phenomenon is understood if we consider that the backgate voltage of the access PFET Q4 (Q1) is V_{dd}=1.2V and that the voltage of the bit-lines (BL /BLB) is set close to V_{ref}=0.65V. This voltage condition makes the threshold voltage for Q4 (Q1) high due to the backgate bias effect to reduce the cell on current (for XNOR= '+1') to 299nA that flows from the bit-lines to the storage node B (A) in the read condition (The cell off current is 3.8nA for XNOR= '−1'). Although this on current per cell in read is very small, the total cell current which is to be sensed by the current sensing system to measure the current (see Fig. 10 (a)) is the accumulation of 784 on or off cells which ranges from 2.98μA (=784× 3.8nA) to 234μA (= 784 × 299nA), because the 784 cells are connected to a BL and are activated at the same time for the Array0 (see Fig. 9, Fig. 19 (a)). This backgate bias design contributes to the reduction in the power consumption of the accelerator. The maximum voltage in V_A(V_B) is observed in the hold butterfly curve when V_B (V_A) is around 0.15V. This phenomenon is explained by the situation that the gate leakage current that flows from node A (B) to B (A) through Q2 and Q3 decreases more dominantly than the increase in the subthreshold leakage current when V_B (V_A) increase from 0V until around 0.15V. The red curves in Fig. 12 (b) and (c) show the -3σ worst RSNM and HSNM in 1000 Monte Carlo simulations and the corresponding parameters for the cell are shown in Table 1 (c). The statistical conditions are shown in Table 1 (a). Fig. 12 (d) illustrates the retention



(a)



(b)

Fig. 18 (a) Simulation results of OPA1 for the neuron circuits shown in Fig. 10 for array0, array1, array2 and array3. (b) Simulation results of OPA2 and later for the neuron circuits shown in Fig. 10 for array0, array1, array2 and array3.

Table 2 Accuracies for different input digits.

Data	**0**	**1**	**2**	**3**	**4**	
Test num.	**50/50**	47/50	47/50	48/50	49/50	
Accuracy	**100.0%**	94.0%	94.0%	96.0%	98.0%	
Data	**5**	**6**	**7**	**8**	**9**	**Total**
Test num.	49/50	47/50	**50/50**	48/50	46/50	481/500
Accuracy	98.0%	94.0%	**100.0%**	96.0%	92.0%	96.2%

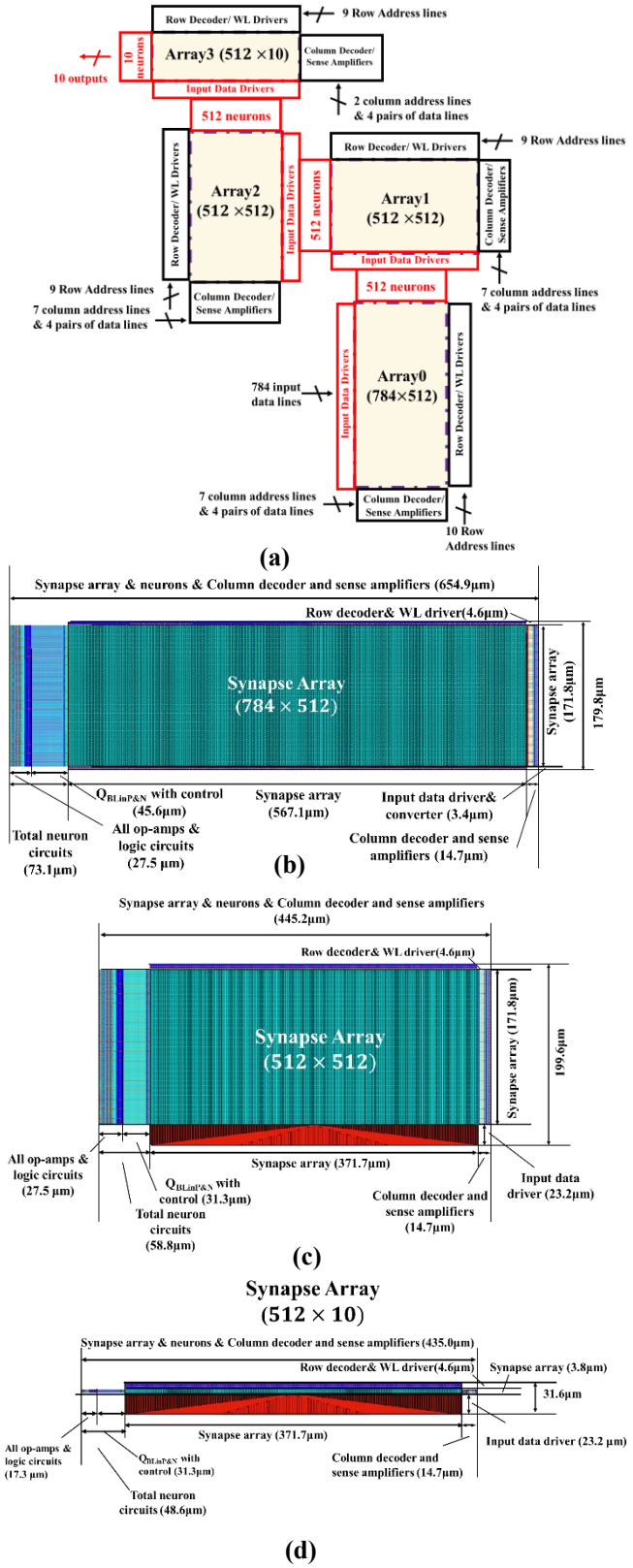


Fig. 19 (a) Overall diagram of the DNN for the MNIST dataset classification including 4 synapse arrays (Array 0-3). (b) The layout of Array 0, (c) Array 1 and 2 and (d) Array 3, all arrays including neuron, input data buffers, row decoder, column decoder and sense amplifier

curves in the 1000 Monte Carlo simulations where V_A and V_B are set at 1.2V and 0V at Time=0 as initial conditions, respectively, while V_{BL} and V_{BLB} are set at 0V and 1.2V after Time > 0, respectively. It is observed that although V_A's decreases from 1.2V to lower voltages, they stop at the positive voltages stably. This is because the proposed cell's data is maintained by the gate leakage current of the PFET access transistor from WL which is kept 1.2V. This is an advantage of the proposed cell over the conventional loadless 4T SRAM cell [15-17] which maintains the data by the subthreshold leakage current of the PFET access transistor from BL. It is to be noted that the read condition shown in Fig. 12 (b) is the same as the half select condition, because both V_{BL} and V_{BLB} are set at V_{ref}=0.65V during the read.

Fig. 13 (a) and (b) show the proposed synapse's RSNM 1000 Monte Carlo simulation results (cumulative probability distribution) under different temperatures and different V_{BL}, respectively. Each cell's effective oxide thickness (EOT) and threshold voltage (V_{th}) are fluctuated independently in the Monte Carlo simulation. The statistical conditions are shown in Table 1. The standard deviations are reflecting the local variation [19]. From Fig. 13, the probability distributions for RSNM are found to be Gaussian ones in this Monte Carlo simulations (1000 times), because the data points in the cumulative distributions in the unit of the standard deviation are found to be lined up in straight lines. So, it can be extrapolated easily to -4.74σ which corresponds to the worst cell among 9.31×10^5 synapse cells in our DNN crossbar arrays (see Fig. 9, Fig. 19 (a)). Fig. 13 (a) and (b) show the proposed synapse's RSNM under different temperatures from -20 ℃ to 80 ℃ and V_{BL}. The smallest RSNM for the worst cell is estimated as 30mV for V_{BL}=0.65V and -20℃ at -4.74σ.

Fig. 14 (a) shows the proposed synapse's HSNM 1000 Monte Carlo simulation under the same V_{WL} 1.2V and different temperatures from −20 ℃ to 80 ℃ (also Gaussian) which are acceptable even at -5σ. The smallest HSNM for the worst cell is estimated as 66mV for 80℃ at -4.74σ. Fig. 14 (b) shows the cumulative distributions for V_A and V_B in the 1000 Monte Carlo simulations of Fig. 12 (d) at Time=1s. Fig. 14 (b) includes the data for all the possible conditions (V_{BL}=0V & V_{BLB}=1.2V, V_{BL}=1.2V & V_{BLB}=0V, VBL=1.2V & V_{BLB}=1.2V, V_{BL}=0V & V_{BLB}=0V) with the same initial condition V_A=1.2V and V_B=0V. It is shown that the retention margin for all the cases are almost the same and there is no concern about the retention (BL disturb).

Fig. 15 shows the butterfly curves in write for the proposed loadless 4T SRAM cell with the definition of the write static noise margin (WSNM). The black and red curves correspond to the typical case at 25℃ and the -3σ worst case at 80℃, respectively. The parameters for the cell corresponding to the worst case are shown in Table 1 (c). The Fig. 16 shows the results of 1000 Monte Carlo simulation for WSNM at V_{dd}=1.2V (also Gaussian). The cumulative probabilities at different temperatures are

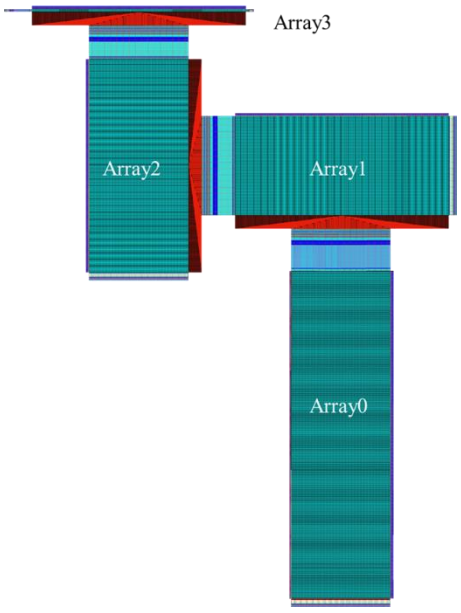


Fig. 20 The total layout for the DNN for the MNIST dataset classification using the synapse arrays shown in Fig. 19 (b), (c) and (d).

simulated under the local variations of V_{th} and EOT shown in Table 1. The worst WSNM is estimated as 0.44V for 80℃ at -4.74σ.

Fig. 17 (a) and (b) show the simulation results of V_{wlout} as a function of the MAC value ranging from -784 to +784 for the first array between the input layer and the first hidden layer and the MAC value ranging from -512 to +512 for the

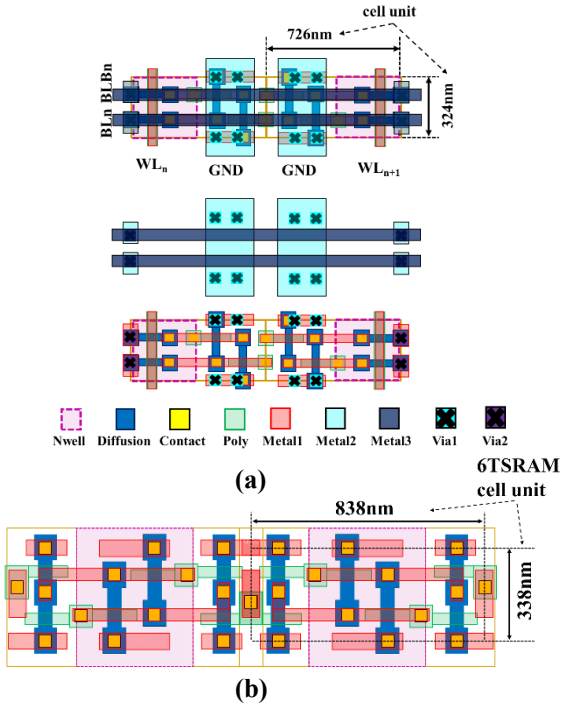


Fig. 21 (a) Synapse layout for the proposed loadless 4T SRAM cell (2cells). (b) Cell layout (2cells) for the conventional 6TSRAM. Both layouts are drawn by using a 32nm technology design rule in which a shared contact is not allowed. The size of the loadless 4T SRAM cell is 83% of the 6TSRAM cell.

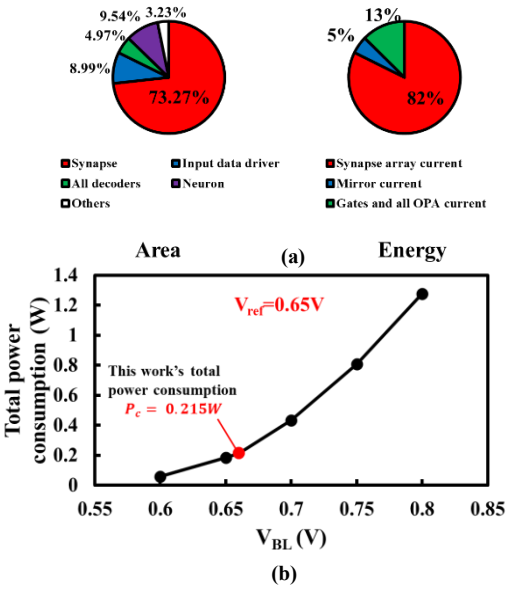


Fig. 22 (a) Breakdowns of area and energy consumption. (b) Dependency of energy consumption on V_{BL}. The design point is V_{BL}=0.655V which is realized by setting V_{ref}=0.65V.

other arrays. The current sensing system schematic's and the synapse's parameters are shown in Table 1 (a) and (b), respectively. Due to the proposed current sensing system, V_{wlout} is shown almost linear to the MAC value. V_{dummy} is used for V_{wlref} for binarizing V_{wlout} and provided by the dummy neuron which is connected to a pair of bit-lines whose MAC value is equal to 0 in each layer as shown in Fig. 17 and Fig. 10 (a). The values of V_{dummy}'s in all the arrays are tuned to be the same. Fig. 17 also shows 10 different chips' V_{wlout} and V_{BL} /V_{BLB} as a function of the MAC for different set of V_{th} and EOT local variations which are distributed according to the Gaussian distributions

Table 3 Performance comparison.

	CICC'20 [29]	JSSC'19 [11]	JSSC'20 [28]	This work
Technology	65nm (32nm LP)	65nm (32nm LP)	65nm (32nm LP)	**32nm LP**
SRAM bitcell	8T	10T	12T	**4T**
Operating V_{dd} (V)	0.80V	1.00V	0.60V	**1.2V**
Bit cell area (μm^2)	3.34 (1.46[30])	3.85 (1.69[30])	3.92 (1.72[30])	**0.235**
On-chip Mem.	49/8KB	2KB	2KB	**49/32 KB**
Throughput (TOPS)	6.72 (9.58[30])	0.455 (0.447[30])	34.90 (78.10[30])	**15.50**
Area Efficiency(TOPS/mm^2)	7.72 (25.12[30])	0.057 (0.13[30])	5.46 (27.90[30])	**50.13**
Cell Efficiency (synapse/total)	51.6%	42.3%	70.75%	**73.27%**
Energy Efficiency(TOPS/W)	15.80 (14.95[30])	40.30 (59.07[30])	403.00 (305.11[30])	**72.2**
MNIST accuracy	96.2%	98%	98.65%	**96.2%**

whose standard deviations are shown in Table 1 (a). Though V_{wlout} fluctuates as MAC value increases, the linearity is almost guaranteed. And V_{BL}/V_{BLB} are closed to V_{ref}=0.65V, showing that there is no concern about read disturb under the local variations.

Fig. 18 shows the simulated waveforms in the inference accelerator described in Fig. 10 and Fig. 11 which is applied to the MNIST dataset recognition problem shown in Fig. 9. The four arrays (array0, 1, 2 and 3) are controlled sequentially by the precharge signals (VCi_0, VCi_1, VCi_2 and VCi_3 with i=1, 2), the output control signals (CTRL0, CTRL1, CTRL2 and CTRL3) and the phase transition

signals (VT_0, VT_1, VT_2 and VT_3). The signals VCBi_0, VCBi_1, VCBi_2 and VCBi_3 are the inverted signals of VCi_0, VCi_1, VCi_2 and VCi_3, respectively. Fig. 18 (a) explains the operation of OPA1 which controls the bit-line pair BL and BLB and transmits the OPA1 output voltage to the gate of the mirror PFET (Q_{mirP}). The operation in the array0 is explained. The operations in other arrays are exactly the same as the array0. When the precharge signal VC1_0 rises high, the bit-line pair BL and BLB are regulated to V_{ref}=0.65V by the negative feedback loop with OPA1. There are two phases for realizing the time-multiplexed XNOR operation. In the first phase, CLK1_0 rises high with CLK2_0 remains low. In this phase, the capacitor C_0 on the gate of the mirror PFET Q_{mirP} corresponding to BL is charged by the voltage in $V_{out1C\_0\_j}$. In the second phase, CLK2_0 rises high with CLK1_0 returned low. In this phase, the capacitor C_0 on the gate of the mirror PFET Q_{mirP} corresponding to BLB is charged by the voltage in $V_{out2C\_0\_j}$. The gate voltages of the two PFETs (Q_{mirP}) are charged to voltages which represent the addition of the logical sates in the two phases. Fig. 18 (b) explains the later operation, i.e., the operation to flow the mirror current in the two PFETs (Q_{mirP}) by using OPA2 and convert the current to a voltage at $V_{wlout\_0\_j}$. The analog voltage is binarized by the comparator and the binarized signal is output to $V_{wl\_1\_j}$ in two phases distinguished by the phase transition signal VT_0. Finally, the output signal $V_{wlout\_3\_0}$ is output as the highest voltage signal than other signals in around 60ns after the data input to show that the digit '0' is successfully classified.

By using an offline training software for the corresponding BNN, 96.8% accuracy was obtained [26]. The optimized binarized weights binary were deployed into the accelerator's synapse data. The proposed inference accelerator's hardware accuracy is tested as 96.2% under totally 500 data (50 data per each digit) as shown in Table 2. Fig. 19 (a) shows the overall diagram of the DNN for the MNIST dataset classification. Fig. 19 (b), (c) and (d) shows the layout for the Array 0, Array 1-2 and Array 3, respectively in 32nm technology [27]. All the layouts include the synapse array, neuron, input data buffer, row decoder, column decoder and sense amplifier circuits. Fig. 20 is the total layout of the DNN. As shown in Fig. 19 (a), Fig 20 and Fig. 22 (a), the area of proposed neural network inference accelerator is mainly consumed by the synapse array where the array efficiency (synapse array area/synapse array + peripheral circuits) is 73.27% where the peripheral circuits include neuron, input data driver, column decoder/sense amplifiers and row decoder/WL drivers. By using the standard CMOS design rules, the parameters and the bit size of the proposed synapse cell unit are shown in Table 1 and Fig. 21, respectively. The proposed synapse bit cell consumes 0.235μm^2, which is significantly smaller than the conventional 8T, 10T and 12T SRAM synapse area consumption as shown in Fig. 21 (a) and Table 3 [11, 28, 29], respectively. It is also 17% smaller than the 6TSRAM cell as shown in Fig. 21. In Table 3, bit cell area, throughput, area efficiency and energy efficiency in the papers [11, 28, 29] were scaled from 65nm with individual V_{dd} to 32nm with V_{dd}=1.2V according to the scaling equations proposed in paper [30]. The total power of the inference accelerator at V_{dd}=1.2V is simulated to be 0.215W for V_{BL}=0.655V. The energy per classification in the cycle time T_c=60ns is calculated to be 0.215W×60ns=12.9nJ. The number of synaptic operations required for a single data classification is 784×512+2×512×512+512×10=9.31×10^5. Therefore, the energy efficiency is calculated to be 9.31×105/12.9nJ=72.2×10^{12} operations/J=72.2TOPS/W. The throughput is 9.31×105/60ns=15.5TOPS. Area efficiency for the total arrays is 15.50 TOPS/0.309mm^2=50.13 TOPS/mm^2. It is worth noting that all the data compared in Table 3 are excluding the decoders and the sense amplifiers to be used in the memory access mode.

The right of Fig. 22 (a) shows power breakdowns. Though the additional OPA and the large BL and BLB load transistors (Q_{BLinP} and Q_{BLinN} in Fig. 10) are required in the proposed neural network inference accelerator, the neuron area consumption only occupies 9.54% of the total area consumption. As for the energy consumption in the proposed accelerator, the synapse array's consumption occupies 82%. The mirror current which flows in Q_{mirP} and Q_{mirN} is converted to V_{wlout} as shown in Fig. 10 (a) accounts for 5% of the total energy consumption. All other power which is consumed in such as logic gates and OPAs in neuron totally accounts 13%. Fig. 22 (b) shows the energy scaling with V_{BL}. Our design chose the case of V_{ref}=0.65V in which the actual V_{BL} was 0.655V.

6. Conclusions

A binarized neural network (BNN) accelerator based on the loadless 4T SRAM cells was proposed which is about 84% efficient in area compared with the state-of-the-art SRAM cell based BNN accelerators. The time-multiplexed XNOR operation applied to the loadless 4T SRAM synapse was adopted to make the operation stable. The current sensing scheme was also verified to be efficient in SRAM based inference accelerators for avoiding the nonlinearity and the read disturb issues. The MNIST dataset recognition problem was solved by using the accelerator and the accuracy was shown by HSPICE to be 96.2% in 32nm technology. The area efficiency in 32nm technology was shown 50.13 TOPS/mm^2 which is one of the highest numbers in the state-of-the-art. The energy efficiency in 32nm technology was 72.17TOPS/W which is higher than the state-of-the-art except the accelerator using the 12T SRAM cell.

Acknowledgments

This work was partly executed under the cooperation of

organization between Kioxia Corporation and Waseda University. This work was supported by JSPS KAKENHI Grant Number JP20K04626. This work was also supported by VLSI Design and Education Center (VDEC), the University of Tokyo with the collaboration with Synopsys Corporation.

References

[1] V. Sze, Y. -H. Chen, T. -J. Yang, J. Emer, Efficient Processing of Deep Neural Networks, Morgan & Claypool Publishers, 2020.
[2] N. Zheng, P. Mazumder, Learning in Energy-Efficient Neuromorphic Computing, John Wiley & Sons Ltd., 2020.
[3] C. Yu, et al, "A 65-nm 8T SRAM compute-in-memory macro with column ADCs for processing neural networks," IEEE J. Solid-State Circuits, vol. 57, no. 11, pp. 3466-3476, Nov. 2022.
[4] Z. Chen et al., "CAP-RAM: A charge-domain in-memory-computing 6T-SRAM for accumulate and precision-programmable CNN inference," IEEE J. Solid-State Circuits, vol. 56, no. 6, pp. 1924-1935, June 2021.
[5] H. Kim, et al, "Colonnade: A reconfigurable SRAM-based digital bit-serial compute-in-memory macro for processing neural networks," IEEE J. Solid-State Circuits, vol. 56, no. 7, pp. 2221-2233, July 2021.
[6] Z. Lin et al., "Cascade current mirror to improve linearity and consistency in SRAM in-memory computing," IEEE J. Solid-State Circuits, vol. 56, no. 8, pp. 2550-2562, Aug. 2021.
[7] H. Valavi, et al, "A 64-tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute," EEE J. Solid-State Circuits, vol. 54, no. 4, pp. 1789-1799, June 2019.
[8] C. -J. Jhang et al, "Challenges and trends of SRAM-based computing-in-memory for AI edge devices," IEEE Trans. Circuits and Systems-I: Regular Papers, vol 68, no. 5, pp. 1773-1786, May 2021.
[9] Y. Zhang, et al, "Recryptor: A reconfigurable cryptographic cortex-M0 processor with in-memory and near-memory computing for IOT security," IEEE J. Solid-State Circuits, vol. 53, no. 4, pp. 995-1005, Apr. 2018.
[10] J. Wang et al., "A compute SRAM with bit-serial integer/floating-point operations for programmable in-memory vector acceleration," International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, pp. 224-225, Feb. 2019.
[11] A. Biswas, A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," IEEE J. Solid-State Circuits, vol. 54, no. 1, pp. 217-230, Jan. 2019.
[12] K. Shin, et al, "Half-select free bit-line sharing 9T SRAM for reliable supply voltage scaling," EEE Trans. Circuits and Systems-I: Regular Papers, vol 64, no. 8, pp. 2036-2048, Aug. 2017.
[13] Z. Jiang, et al, "XNOR-SRAM: In-memory SRAM macro for binary-ternary deep neural networks," Symp. VLSI Tech. Dig. Tech. Papers, pp. 173-174, June 2018.
[14] H. Valavi, et al, "A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement," Symp. VLSI Circuits Dig. Tech. Papers, pp. 141-142, June 2018.
[15] K. Noda, et al, "A loadless CMOS f our-transistor SRAM cell in a 0.18-μm logic technology," IEEE Trans. Electron Devices, vol. 48, no. 12, pp. 2851-2855, Dec. 2001.
[16] K. Imai, et al, "0.13- μ m CMOS technology integrating high-speed and low power/high density devices with two different well/channel structures," IEDM Tech. Dig., pp. 667-690, Dec. 1999.
[17] K. Noda, et al, "An ultrahigh-density high-speed loadless four-transistor SRAM macro with twisted bitline architecture and triple-well shield," IEEE J Solid-State Circuits, vol. 36, no. 3, pp. 510-515, Mar. Mar, 2001.
[18] Y. Zhu, T. Ohsawa, "A Loadless 4T SRAM Powered by Gate Leakage Current with a High Tolerance for Fluctuations in Device Parameters," Jpn. J. Appl. Phys. 61, SC1053 (2022).
[19] Y. Zhu, T. Ohsawa, "A gate leakage current-powered loadless 4T SRAM with immunity against random dopant fluctuation and surface roughness in silicon-silicon dioxide interface," Jpn. J. Appl. Phys. 62, SC1004 (2023).
[20] Y. Zhu, T. Ohsawa, "A Bit-Line Disturb Free Loadless 4T SRAM Using Gate Leakage Current for Sustaining Data with High Immunity against Process Variations," Extended Abstracts of the 2021 International Conference on Solid State Devices and Materials (SSDM), pp. 127-128, Sep. 7, 2021.
[21] Y. Zhu, T. Ohsawa, "A Loadless 4T SRAM Cell Powered by Gate Leakage Current and Tolerant of Random Dopant Fluctuation and Surface Roughness at Si-SiO_2 Interface," Extended Abstracts of the 2021 International Conference on Solid State Devices and Materials (SSDM), pp. 625-626, Sep. 2022.
[22] E. Seevinck, et al, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," IEEE J. Solid-State Circuits, vol. 26, no. 4, pp. 525-536, Apr. 1991.
[23] N. Shibata, "Current sense amplifiers for low-voltage memories," IEICE Trans. Electron., vol. E79-C, no. 8, pp. 1120-1130, Aug. 1996.
[24] B. Wicht, Current Sense Amplifiers, Springer-Verlag, Berlin, 2003.
[25] A. Nguyen, K. Pham, D. Ngo, T. Ngo, L. Pham, "An analysis of state-of-the-art activation functions for supervised deep neural network" International Conference on System Science and Engineering, pp. 215-220, 2021.
[26] Z. Chen, T. Ohsawa, "A Low-Cost Training Method of ReRAM Inference Accelerator Chips for Binarized Neural Networks to Recover Accuracy Degradation due to Statistical Variabilities," IEICE Trans. Electron., vol. E105-C, no. 8, pp. 375-384, Aug. 2022.
[27] R. T. Greenway, et al, "32nm 1-D Regular Pitch SRAM Bitcell Design for Interference-Assisted Lithography," Proc. of SPIE, vol. 7122, pp. 71221L-1-71221L-12, Oct. 2008.
[28] S. Yin, et al, "XNOR-SRAM: In memory computing SRAM macro for binary/ternary deep neural networks," IEEE J. Solid-State Circuits, vol. 55, no. 6, pp. 1733–1743, Jun. 2020.
[29] C. Yu, et al, "A 16K Current-Based 8T SRAM Compute-In-Memory Macro with Decoupled Read/Write and 1-5bit Column ADC," 2020 IEEE Custom Integrated Circuits Conference (CICC), Mar. 2020.
[30] A. Stillmaker, and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm," Integration 58, 74 (2017).

Yihan ZHU received the B.S. degree from Beijing University of Chemical Technology, Beijing, China, in 2017, and the M.S. degree from Graduate School of Information, Production and Systems, Waseda University, Kitakyushu, Japan, in 2019. He is currently focusing on analog circuit design.

Takashi OHSAWA received the B.S. and M.S. degrees in physics from Waseda University, Tokyo, Japan, in 1977 and 1979, respectively, and the Ph.D. degree in electronic engineering from University of Tsukuba, Tsukuba, Japan, in 2009. During 1982-2010, he worked on research and development of semiconductor memories in the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki and Yokohama, Japan. From 2010 to 2014, he was a Professor with Center for Spintronics Integrated Systems, Tohoku University, Sendai, Japan. Since 2017 he has been a Professor with Graduate School of Information, Production and Systems, Waseda University, Kitakyushu, Japan.