IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 22 August 2022; revised 12 October 2022, 7 November 2022, and 20 November 2022; accepted 22 November 2022.
Date of publication 30 November 2022; date of current version 22 December 2022.

Digital Object Identifier 10.1109/JXCDC.2022.3225744

Binarized Neural Network Accelerator Macro Using Ultralow-Voltage Retention SRAM for Energy Minimum-Point Operation

YUSAKU SHIOTSU (Graduate Student Member, IEEE), and SATOSHI SUGAHARA (Member, IEEE)

Laboratory for Future Interdisciplinary Research of Science and Technology, Tokyo Institute of Technology, Yokohama 226-8502, Japan

CORRESPONDING AUTHOR: Y. SHIOTSU (y.shiotsu@isl.titech.ac.jp)

This work was supported in part by Japan Society for the Promotion of Science under KAKENHI Grant JP22H03556 and in part by VLSI Design and Education Center (VDEC), The University of Tokyo, in collaboration with Nihon Synopsys G.K., Cadence Design Systems, Mentor Graphics, and Renesas Electronics Corporation.

This article has supplementary downloadable material available at https://doi.org/10.1109/JXCDC.2022.3225744, provided by the authors.

ABSTRACT A binarized neural network (BNN) accelerator based on a processing-in-memory (PIM)/computing-in-memory (CIM) architecture using ultralow-voltage retention static random access memory (ULVR-SRAM) is proposed for the energy minimum-point (EMP) operation. The BNN accelerator (BNA) macro is designed to perform stable inference operations at EMP and substantive power-gating (PG) using ULVR at an ultralow voltage (<EMP), which can be applied to fully connected layers (FCLs) with arbitrary shapes and sizes. The EMP operation of the BNA macro, which is enabled by applying the ULVR-SRAM to the macro, can dramatically improve the energy efficiency (TOPS/W) and significantly enlarge the number of parallelized multiply–accumulate (MAC) operations. In addition, the ULVR mode of the BNA macro, which also benefits from the usage of ULVR-SRAM, is effective at reducing the standby power. The proposed BNA macro can show a high energy efficiency of 65 TOPS/W for FCLs. This BNA macro concept using the ULVR-SRAM can be expanded to convolution layers, where the EMP operation is also expected to enhance the energy efficiency of convolution layers.

INDEX TERMS Binarized neural network (BNN) accelerator, computing-in-memory (CIM), energy minimum-point (EMP) operation, power-gating (PG), processing-in-memory (PIM), static random access memory (SRAM).

I. INTRODUCTION

IN FUTURE smart society, artificial intelligence (AI) technology becomes more important not just for cloud computing systems but for mobile-edge computing devices. In particular, the demands for the mobile-edge AI technology would be further expanded. This is because the usage of a cloud-based AI system for mobile-edge applications causes unwanted not sufficiently short latency for data-processing owing to communication between them, and also the mobile-edge devices would be frequently used in situations unconnected to the cloud-based AI system. The mobile-edge AI technology can broaden a range of AI applications owing to the portability.

For these mobile-edge applications, energy-efficient low-power design of AI systems is indispensable. Neural network (NN) accelerators are promising for implementing mobile-edge AI systems [1]. Performances of NN accelerators, such as their processing performance (TOPS) and energy efficiency (TOPS/W), can be improved from the point of view of their hardware and/or architecture.

On the hardware of NN accelerators, a system organization based on the processing-in-memory (PIM) methodology has attracted considerable attention [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. This new type of computing paradigm is also called computing-in-memory (CIM) and can trace history back to a memory-based architecture named as ''functional memory'' [12]. Hardware implementation of NNs has been an important target for functional memory since that time. In this article, the term PIM is mainly used.

PIM/CIM can be broadly defined as a computing hardware/architecture that performs data processing inside the memory subsystem. In this type of hardware for NN accelerators, the multiply–accumulate (MAC) unit is placed in/near memory arrays, and the data retrieved from the memory array are directly processed without transferring them through a bus. Therefore, the PIM-type hardware is effective

at improving the energy efficiency. In addition, the PIM structure can effectively parallelize the MAC operations in NNs without the constraint originating in bus usages.

The architecture implementing NNs is also important for processing and energy performances of NN accelerators. Binarized NNs (BNNs) are a promising architecture [5], [6], [7], [8], [9], [10], [11], which uses single-bit activations and also restricts each weight in the network to a single bit. The BNN architecture results in significant reduction in the memory capacity for the weight data. In addition, the binary weight data can be read out fast with lower power consumption. The low memory capacity is also beneficial to suppress the standby power.

The BNN architecture can simplify the MAC operations at each neuron node, where the MAC operations can be replaced by the XNOR and population count (PPC) operations, respectively. This contributes to significant reduction in the power consumption and circuit area of the MAC unit. The simple MAC unit is also preferable to parallelization of the MAC operations. The above-described features on the binary activations and weight data are highly promising for improving the energy efficiency of NN accelerators. Note that the BNN architecture has many types of variations, some of which can achieve relatively high accuracy despite the binary activations and weight data, such as XNOR-net [13], [14].

The PIM-type NN accelerators would be a suitable hardware for implementing the BNN architectures. Hereafter, BNN accelerators are referred to as BNAs. In general, the energy efficiency of NN accelerators can be enhanced by reducing the driving voltage [7], [8], [9], [10]. In particular, the energy minimum-point (EMP) operation having the maximum energy efficiency is promising [15]. In BNAs, owing to the simplified MAC circuits, the energy performance of the on-chip memory strongly affects their energy efficiency. Thus, for PIM-type BNAs, the EMP operation of the memory becomes an important challenge. Moreover, for mobile-edge applications, the implementation of power-gating (PG) is also demanded. Most of the area of PIM-type BNAs is occupied by their memory array, and thus the standby power generated by data retention needs to be diminished. Therefore, both the EMP and PG operations of on-chip memory have a great impact on implementation of PIM-type BNAs, although to achieve both the requirements is not easy for any on-chip memory.

It is worthy to note that the EMP operation is also effective at achieving a high degree of parallelization of the MAC operation. The processing performance (TOPS) can be enhanced by parallelization of the MAC operation. The allowable number of parallelized MAC operations is restricted by the total power consumption of the simultaneously executed MAC operations. The EMP operation can dramatically lower the active power required for the MAC operation, enlarging (maximizing) the number of parallelized MAC operations.

NN accelerators can use a fixed or reconfigurable network structure, which also becomes an important factor for their power and energy performances. Since the main target of mobile-edge AI applications is considered to be for image/biological cognition and the related field, fixed-structure networks that are beneficial for lower power/energy consumption can be applied to NN accelerators. From workload (that may be extended into multiple ones) for the applications, the required network structure can be predetermined using simulation, which can be implemented as a fixed network structure. This methodology without the usage of redundant reconfigurable network structures is effective for low-power energy-efficient NN accelerators. When weight data can be updated and the network structure can have some margin, there is some degree of freedom of the substantive network structure even for the fixed-structure network. The fixed-structure network approach is promising for mobile-edge NN accelerators, which also allows PIM-type NN accelerators to be configured with multiple small-scale macros.

In this article, a PIM-type BNA macro using ultralow-voltage retention static random access memory (ULVR-SRAM) is proposed for the EMP and PG operations, and its design and performance are demonstrated. The ULVR-SRAM is a new type of SRAM having both the EMP and PG operation abilities. The BNA macro is simply configured by adding a MAC unit to an ULVR-SRAM array designed with careful consideration of statistical variation in the constituent devices. Fully connected layers (FCLs) with arbitrary shapes and sizes can be configured using the multiple BNA macros with tiny additional peripherals. The impact of the EMP operation on energy efficiency and the substantive PG execution using the ULVR mode on the standby power reduction are analyzed from postlayout large-scale simulations. Performances of FCLs configured with the multiple BNA macros are discussed. The concept of the ULVR-SRAM-based BNA macro can also be applied to convolution layers.

The important and fascinating results of our work are as follows: the EMP operation of the proposed PIM-type BNA macro, which is enabled by applying the ULVR-SRAM to the macro, can dramatically improve the energy efficiency (TOPS/W) and significantly enlarge the number of parallelized MAC operations. In addition, the ULVR mode of the BNA macro, which also benefits from the usage of ULVR-SRAM, is effective at reducing the standby power. From these features, the proposed macro is promising for mobile-edge BNAs.

II. RELATED WORK

In this section, related work on several types of memories used for PIM-type NN accelerators is briefly reviewed. For PIM-type NN accelerators, there are several variations that can be roughly classified by the type of on-chip memory used in them. SRAMs (that often have modified structures suitable to NN accelerator architectures) and emerging nonvolatile memories (NVMs), such as resistive-switching random access memory (ReRAM), phase-change random access memory (PRAM), and magnetoresistive random access memory (MRAM), have been investigated for PIM-type NN accelerators [3], [4], [5], [6], [7], [8], [9], [10].

NVMs are highly effective at reducing the standby power of PIM-type NN accelerators. Their smaller cell sizes are also beneficial. Nevertheless, higher energies required for the write operation would restrict their applications. Namely, these NVMs are suitable for NN accelerators for exclusive use of inference. Also, the EMP operation might be difficult (which could be caused by the peripheral circuits) [16]. Furthermore, the embedded technology of NVMs costs a lot.

SRAMs are useful to implement PIM-type NN accelerators owing to the sophisticated design methodology, accumulated design technologies, and excellent compatibility to CMOS logic circuits/processes, although the area overhead is not so small. Recently, SRAM-based PIM-type NN accelerators are commercially available. In general, the energy efficiency of SRAM-based PIM-type NN accelerators can be enhanced by reducing the driving voltage. Nevertheless, the conventional 6T cells are difficult to reduce it to the EMP voltage that gives the maximum energy efficiency. This is because the 6T cells cannot ensure sufficient noise margins at such lower voltages. Therefore, specially designed cells such as isolated read port cells and Schmitt trigger (ST) cells [17], [18] can be applied to EMP operation NN accelerators. However, their volatile nature prohibits implementation of PG to NN accelerators for standby power reduction.

Recently, fully CMOS-based ULVR-SRAM has been proposed [19], [20], which can have three operating modes, i.e., the retention mode at an ultralow voltage V_{UL} such as 0.2 V (hereafter, referred to as the ULVR mode), the SRAM-operating mode at the EMP voltage V_{EMP} (referred to as the SRAM^{EMP} mode), and the normal SRAM-operating mode at the ordinary supply voltage V_{DD} (referred to as the SRAM^{Norm} mode). The ULVR-SRAM cell is configured with ST-based dual-mode inverters, which is designed so as to have strong noise immunity for the ULVR mode. During the ULVR mode, the dual-mode inverters in the cell act as an ST inverter having rectangular-shaped transfer characteristics with wide hysteresis, and thus the ULVR-SRAM cell can stably retain data even at V_{UL} (=~0.2 V). Since the ULVR mode can effectively reduce the standby power [20], substantive PG using the ULVR mode can be achieved. This ST mode of the dual-mode inverters can also be applied to the stable energy-efficient SRAM^{EMP} operation. Namely, the ST mode enables the cell to ensure sufficient noise margins for the SRAM operations even at V_{EMP}. The SRAM^{Norm} operation can be performed using the normal inverter mode of the dual-mode inverters. High-performance SRAM operations comparable to the conventional 6T-SRAM operations can be achieved at the ordinary supply voltage V_{DD} ($V_{UL} < V_{EMP} < V_{DD}$) [20]. Therefore, ULVR-SRAM is promising for PIM-type NN accelerators. The SRAM^{EMP} mode is highly beneficial not just to enhance energy efficiency but also to enlarge parallelized MAC operations. The ULVR mode can effectively reduce the standby power through substantive PG operation.

In our previous paper [20], the ULVR-SRAM cell was designed to achieve the two-mode operations of the SRAM^{Norm} and ULVR modes, particularly to achieve

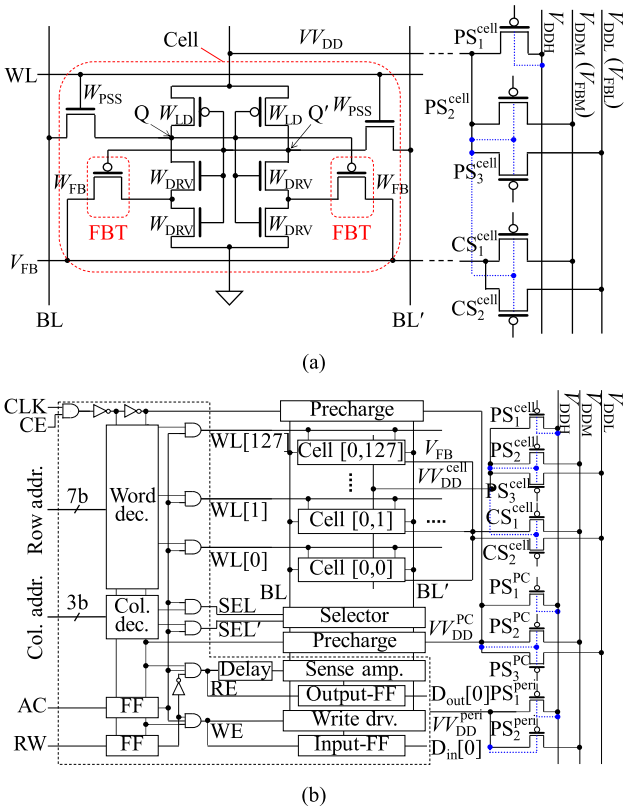

FIGURE 1. (a) Circuit configuration of the ULVR-SRAM cell with power and control switches. (b) Block diagram of the ULVR-SRAM subarray.

efficient standby power reduction during the ULVR mode. Thus, this cell used the header and footer power switches (HFPSs). The HFPS configuration can easily introduce the automatic body bias control during the ULVR mode, which can enhance leakage power reduction. However, using this cell architecture, to achieve the three-mode (SRAM^{Norm}, SRAM^{EMP}, and ULVR) operations causes difficulty in hardware implementation that required many power rails with complex power switch control (the details are shown in Section B of the supplementary material). In this article, the ULVR-SRAM cell is redesigned so as easily to achieve the EMP operation. The power switch configuration is changed from HFPSs to header power switches (HPSs), and automatic body bias control is not used. Design of this new type of ULVR-SRAM with three-mode operations and performance of the proposed EMP-BNA macro using it are demonstrated.

III. PROPOSED BNA MACRO

A. ULTRALOW-VOLTAGE RETENTION SRAM

Fig. 1(a) shows the circuit configuration of the ULVR-SRAM cell that consists of the dual-mode inverters with the pMOS feedback transistors (FBTs) [20]. The storage nodes are denoted by Q and Q′ in the figure. The power switch organization is also shown in the figure. The virtual supply voltage (VV_{DD}) is supplied using the HPSs (PS_1^{cell}, PS_2^{cell}, and PS_3^{cell}) with three power rails of V_{DDH} (=1.2 V), V_{DDM} (=0.4 V), and V_{DDL} (=0.2 V). During the SRAM^{Norm},

SRAM^{EMP}, and ULVR modes, VV_{DD} is set to V_{DDH}, V_{DDM}, and V_{DDL}, respectively. The bias V_{FB} of the FBTs is controlled by the control switches (CSs) CS_1^{cell} and CS_2^{cell}. V_{FB} is set to V_{FBM} (=0.4 V) for the SRAM^{EMP} mode and to V_{FBL} (=0.2 V) for both the SRAM^{Norm} and ULVR modes [20]. As (VV_{DD}, V_{FB}) = (V_{DDH}, V_{FBL}), the dual-mode inverters in the cell operate with the normal inverter mode, and thus the SRAM^{Norm} operation can be achieved. For the conditions of (VV_{DD}, V_{FB}) = (V_{DDM}, V_{FBM}) and (V_{DDL}, V_{FBL}), the dual-mode inverters act as ST inverters. Thus, even at lower voltages of VV_{DD} = V_{DDM} (=V_{EMP}) and V_{DDL} (=V_{UL}), the cell can perform stable SRAM^{EMP} and ULVR operations owing to the strong noise immunity for the ST mode. The SRAM^{EMP} mode can exhibit the highest energy efficiency. Also, the ULVR mode can be used as substantive PG and effectively diminish the standby power. Note that hardware implementation only for the two-mode operations with the SRAM^{EMP} and ULVR modes (without the SRAM^{Norm} mode) can be achieved by directly connecting the V_{FB} terminal of the FBTs to the VV_{DD} rail. Also note that although general ST cells have nMOS FBTs, the pMOS FBTs of the ULVR-SRAM cell are more beneficial for reducing the cell area and enhancing the noise margins at lower voltages [20]. Thus, even for exclusive use of two-mode operations with the SRAM^{EMP} and ULVR modes, the ULVR-SRAM cell is advantageous compared with general ST cells.

Fig. 1(b) shows the block diagram of the ULVR-SRAM subarray used for the proposed BNA macro. The macro is configured with four memory subarrays and peripheral circuits. The memory capacity of the macro is 8.25 kB, which is organized for 256 × 256 b weight data and 256 × 8 b bias data. The bit-width for the readout is 264 b, i.e., 256 b weight and 8 b bias data can be simultaneously read out.

The rails of VV_{DD}^{cell}, VV_{DD}^{PC}, and VV_{DD}^{peri} are separately arranged for the cell array, the precharge circuits, and the other peripherals, respectively, to individually analyze the power characteristics of these circuit blocks. The VV_{DD}^{PC} rails are also used for voltage control of the bit lines (BLs) during the ULVR mode. Note that in practical implementation of the macro, these rails can be appropriately merged.

These virtual supply voltages are controlled using the power switches PS_α^β (β = cell, PC, peri; α = 1–3 for β = cell, PC; α = 1, 2 for β = peri), as shown in the figure. Using PS_1^{cell}–PS_3^{cell}, one of the supply voltages V_{DDH}, V_{DDM}, and V_{DDL} is selected for VV_{DD}^{cell}. The same means is used for VV_{DD}^{PC}. VV_{DD}^{peri} is supplied from V_{DDH} or V_{DDM} using PS_1^{peri} and PS_2^{peri} (i.e., V_{DDL} is not used for the peripherals except the precharge circuits). V_{FB} is controlled using CS_1^{cell} and CS_2^{cell} with the power rails V_{DDM} and V_{DDL}. The body bias connections of these PSs and CSs are shown by the dotted lines in Fig. 1(b). These connections can effectively suppress back-flow currents from the V_{DDH} to the V_{DDM}/V_{DDL} rail during the shutoff phases of these PSs. Note that although the body bias effect lowers the threshold voltages of PS_2^{cell}, PS_2^{PC}, and CS_1^{cell} during the ULVR mode, the unwanted leakage current flows are negligible owing to the small difference between VV_{DD} (V_{DDL}) and V_{DDM}.

TABLE 1. Operating conditions.

mode	Control voltage				Gate voltage				
	BL (V)	AC (V)	CE (V)	RW (V)	PS_1^β (V)	PS_2^β (V)	PS_3^β (V)	CS_1^{cell} (V)	CS_2^{cell} (V)
ULVR	0.20	0.0	0.0	0.0	1.4	1.4	0.0	0.0	1.4
SB_1^{EMP}	0.40	0.0	0.40	0.0	1.4	0.0	1.4	1.4	0.0
SB_2^{EMP}	FL	0.0	0.0	0.0	1.4	0.0	1.4	1.4	0.0
$READ^{EMP}$/ $INFER^{EMP}$	0.40	0.40	0.40	0.0	1.4	0.0	1.4	1.4	0.0
$WRITE^{EMP}$	0.40/0.0	0.40	0.40	0.40	1.4	0.0	1.4	1.4	0.0
SB_1^{Norm}	1.2	0.0	1.2	0.0	0.0	1.4	1.4	0.0	1.4
SB_2^{Norm}	FL	0.0	0.0	0.0	0.0	1.4	1.4	0.0	1.4
$READ^{Norm}$/ $INFER^{Norm}$	1.2	1.2	1.2	0.0	0.0	1.4	1.4	0.0	1.4
$WRITE^{Norm}$	1.2/0.0	1.2	1.2	1.2	0.0	1.4	1.4	0.0	1.4

PS_1^β, PS_2^β: β= cell, PC, peri. PS_3^β: β= cell, PC. FL: Floating

Table 1 shows the bias conditions of the PSs and CSs for all the operation modes. The modes SB_1 and SB_2 represent the standby states without and with clock-gating, respectively, where the BL rails are precharged (clock-driven) to VV_{DD}^{PC} and floating, respectively. The BLs during the ULVR mode is charged to 0.2 V, which can effectively suppress leakage currents through the pass transistors during the ULVR mode [20]. Other notes for VV_{DD} control are described in Section E of the supplementary material.

B. BNA MACRO USING ULVR-SRAM

Fig. 2(a) shows a network structure having FCLs that can be configured with the proposed BNA macros. Let n and m to be the numbers of neuron nodes in each layer and of layers, respectively. $x_i^{(j-1)}$ and $x_i^{(j)}$ ($i = 1, \ldots, n$, $j = 0, \ldots, m$) represent the elements of input and output vectors for the jth layer, respectively, and $w_{ii'}^{(j)}$ ($i, i' = 1, \ldots, n$) and $b_i^{(j)}$ ($i = 1, \ldots, n$) are the weight and bias data for the jth layer, respectively. The elements of the input and output vectors for each layer and the weights have single-bit binary value, and the bias data are integer. The MAC operations can be carried out through the n-to-1 connections in the network, such as the red, blue, and green lines shown in Fig. 2(a).

In this article, an ULVR-SRAM-based PIM-type BNA macro that can achieve the three-mode operations based on the SRAM^{Norm}, SRAM^{EMP}, and ULVR modes of the ULVR-SRAM is demonstrated. Fig. 2(b) shows the block diagram of the BNA macro that consists of the ULVR-SRAM array and a MAC unit with the output generation circuits (activation (ACTV) circuit and output latch). The ULVR-SRAM array has the memory capacity of 8.25 kB for weight and bias data, as noted above. The MAC unit is simply configured with XNOR gates and an adder tree PPC circuit [7], [8], [9], [10], [11]. The ACTV circuit is a simple adder (with the carry-out port) for MAC results and bias data. The macro can simultaneously read 256 b weight data stored in it and operate MAC calculations for these weight data and a 256 b input

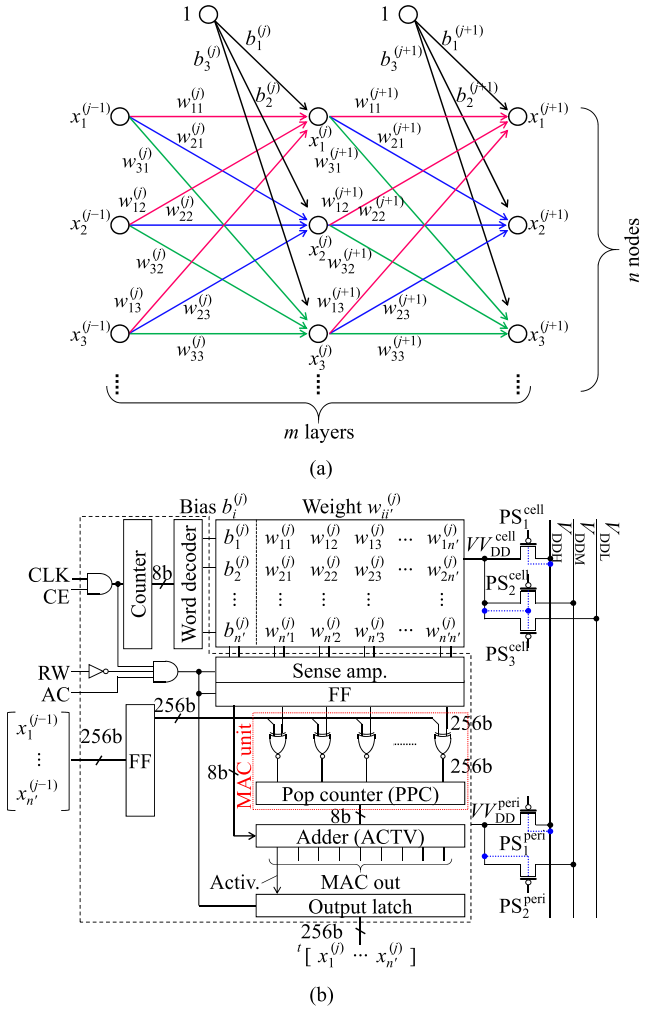

FIGURE 2. **(a) Structure of an FCL network. (b) Block diagram of the BNA macro. CS_α^{cell} and PS_α^{PC} are not shown in this figure. The region enclosed by the dashed lines represents the circuit block where VV_{DD}^{peri} is supplied.**

vector. Namely, a single FCL with $n' = 256$ (n' represents the number of neuron nodes in a single macro) can be configured using a single macro. The elements ($x_1^{(j-1)}$, ..., and $x_{n'}^{(j-1)}$) of an input vector are XNORed with the weights ($w_{u1}^{(j)}$, ..., and $w_{un'}^{(j)}$) for a single neuron node (for $x_u^{(j)}$) and then the population of "1" in the results is counted using the PPC circuit. At this stage, the MAC result is integer. After adding 8 b bias data (that are also stored in the memory array of the macro) to the integer MAC result in the ACTV circuit, the activation (firing) is judged by the most significant bit (MSB) of the final result, i.e., the neuron is activated when there exists the carry flag. The activation result returns to a single-bit binary value.

The XNOR, PPC, and ACTV circuits share the PSs and VV_{DD}^{peri} rail for the peripheral circuits of the ULVR-SRAM array, as shown in Fig. 2(b). Based on the operating modes of the ULVR-SRAM, the BNA macro can perform the inference operations at V_{DD} and V_{EMP} (hereafter, referred to as the $INFER^{Norm}$ and $INFER^{EMP}$ modes, respectively) and the

TABLE 2. **Design of the ULVR-SRAM cell.**

L	60 nm	W_{PS1}	1000 nm	V_{DDH}	1.2 V
W_{LD}	100 nm	W_{PS2}	1000 nm	V_{DDM}	0.40 V
W_{DRV}	220 nm	W_{PS3}	100 nm	V_{DDL}	0.20 V
W_{FB}	140 nm	W_{CS1}	2000 nm	V_{FBM}	0.40 V
W_{PSS}	100 nm	W_{CS2}	100 nm	V_{FBL}	0.20 V

PS_1^{cell}, PS_2^{cell}, and CS_1^{cell} are shared by 128 cells in a single column.

ULVR operation at V_{UL}. These operating states are controlled by the PSs and CSs, as shown in Table 1. Note that the BNA macro is used for the inference operation with updatable weight data.

An FCL with arbitrary neuron numbers can be implemented by the multiple BNA macros with additional peripherals (adders and output latch), as described in Section V (also see Section G in the supplementary material). These additional peripherals circuits can be synthesized computationally so as to adapt its network structure. Note that in general, the logic blocks synthesized using standard cells would perform the EMP operation (except SRAM circuits). However, design refinement could be required for the various timing conditions and resulting operating frequency depending on the scale of the network (particularly, adjustment of the drivability of buffers becomes important). Also note that it can be considered to use reconfigurable hardware for the additional peripherals. Generally, reconfigurable peripherals require the large area and power overhead due to their redundant configuration. Therefore, in this article, the multiple BNA macros with custom-synthesized additional peripherals are supposed for implementation of FCLs with arbitrary neuron number.

The BNA macro concept can be expanded to convolution layers. Namely, convolution layers can also be configured with ULVR-SRAM-based macros having memory capacity adopted to the size and number of kernels and appropriately modified MAC units for kernels, which can introduce the EMP operation to convolution layers.

IV. DESIGN AND PERFORMANCES OF BNA MACRO

A. ULVR-SRAM DESIGN

In Sections IV and V, the operations in the $SRAM^M$ mode (M = EMP, Norm) are denoted by $READ^M$, $WRITE^M$, and SB^M, in which SB^M represents the standby retention (HOLD) mode. Note that the cell operations of the $SRAM^{Norm}$ and $SRAM^{EMP}$ modes are based on the normal inverter mode at $VV_{DD} = V_{DDH}$ (=1.2 V) and the ST mode at $VV_{DD} = V_{DDM}$ (=0.4 V), respectively.

In this study, the low-power devices of the 65-nm silicon on thin buried oxide (SOTB) technology were used [21]. A methodology described in [20] was used for the ULVR-SRAM cell design. Namely, the cell was designed so as to ensure sufficient noise margins for the EMP-operating mode and also the ULVR mode (see Section C in the supplementary material).

Table 2 shows the design result for the ULVR-SRAM cell with the HPS configuration. Note that this design result

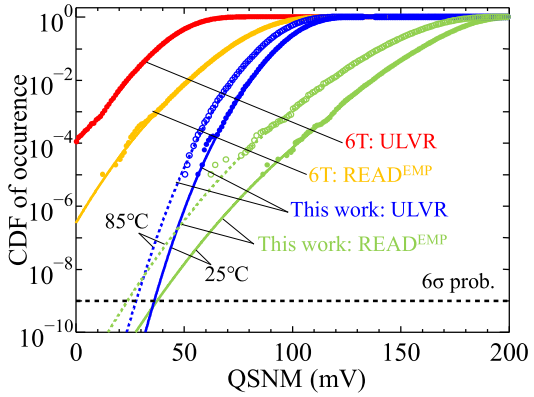

FIGURE 3. CDF of the QSNM distributions for the ULVR-SRAM and 6T-SRAM cells in the ULVR and READ^{EMP} modes. The solid and dotted curves with the filled and open circles represent the operating temperatures of 25 °C and 85 °C, respectively. The dashed line represents the 6σ failure probability. The number of trials for each Monte Carlo simulation is 100 000.

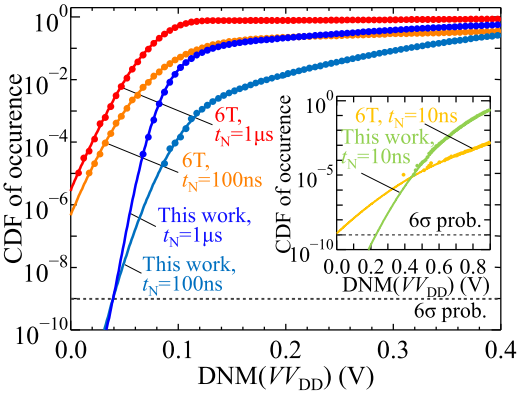

FIGURE 4. CDF of the VV_{DD}-noise-induced DNM distributions for the ULVR-SRAM and 6T-SRAM cells. The dashed line represents the 6σ failure probability. The number of trials for each Monte Carlo simulation is 100 000.

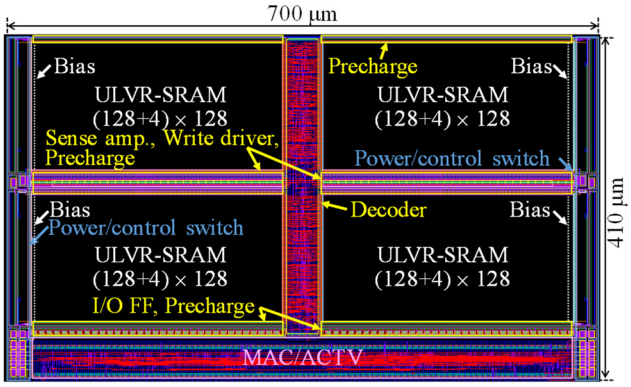

FIGURE 5. Layout of the ULVR-SRAM-based PIM-type BNA macro.

differs from the design for the ULVR-SRAM cell with the HFPS configuration [20]. The HPS configuration cell has a larger driver size (W_{DRV}) than the HFPS-configuration cell, resulting in sufficiently high noise margins comparable to the HFPS-configuration cell.

The ULVR-SRAM cell was laid out based on the logic design rule [since the SRAM design rule was unavailable for our used process design kit (PDK)]. From the size of the PDK-provided 6T cell (designed with the SRAM design rule), the area overhead for the ULVR-SRAM cell can be estimated. For the 6T cell, the logic design rule causes 2.4 times larger layout than the SRAM design rule. For the logic design rule, the ULVR-SRAM cell has 1.8 times larger layout than the 6T cell. Thus, the ULVR-SRAM cell layout using the logic design rule is 4.2 times larger than the 6T cell layout using the SRAM design rule. Nevertheless, assuming that the SRAM design rule is used for the ULVR-SRAM cell, its layout can be estimated to be ∼1.8 times larger than that of the 6T cell using the SRAM design rule.

Using the Monte Carlo simulations, the cell design was verified from noise immunity under the random local variation in the constituent devices. The details of the Monte Carlo simulations were described in [20]. Fig. 3 shows the cumulative distribution function (CDF) for the quasi-static noise margin (QSNM) distribution of the cell during the ULVR and READ^{EMP} modes at 25 °C. The results for a conventional 6T cell are also shown in this figure as a reference (the 6T cell was designed by reference to [22]). The data (filled circles) were fit by the superposition (solid curves) of several Gaussians. The CDF tails for the ULVR-SRAM cell can satisfy the 6σ failure probability for both the READ^{EMP} and ULVR modes. On the other hand, the 6T cell cannot satisfy this criterion for both the READ^{EMP} and ULVR modes. Note that the ULVR-SRAM cell can also satisfy the 6σ failure probability criterion for both the READ^{EMP} and ULVR modes even at 85 °C, as shown by the open circles with the dotted fitting curves in the figure.

The noise immunity for VV_{DD} rail was also analyzed from dynamic noise margins (DNMs) using the Monte Carlo simulations with careful consideration of local variation in the constituent devices. Fig. 4 shows the CDF for the DNMs of the ULVR-SRAM cell during the ULVR mode, where the noise pulse widths t_N of 10 ns, 100 ns, and 1 µs are examined. The cell sufficiently satisfies the 6σ failure probability even for longer pulse widths. In the figure, the CDF of the 6T cell is also shown. The 6T cell cannot ensure sufficient DNMs for the 0.2 V retention mode (i.e., the 6T cell cannot satisfy the 6σ failure probability). The ULVR-SRAM cell also has high immunity for power rail noises.

B. MACRO DESIGN

The PIM-type BNA macro was designed using the above-described ULVR-SRAM cell (see Table 2). Fig. 5 shows the layout of the designed BNA macro. The memory decoder, XNOR, PPC, and ACTV circuits were computationally synthesized and laid out. These circuits were configured with high threshold voltage (HVT) devices, while the clock drivers and the FFs for the address latch were configured with low threshold voltage (LVT) devices. The MAC unit was configured with XNOR gates and an adder-tree-type PPC circuit, and the ACTV unit is organized using an adder circuit. The simple latch-style sense amplifiers and the other peripherals

(write driver, precharge driver, and selector) for the memory array were custom-designed using HVT devices. Note that the delay circuit for the sense amplifiers was designed to ensure secure sensing operations at 1.2 and 0.4 V.

The threshold voltages of the HVT devices are slightly higher than those of the LVT devices (∼0.05 and ∼0.1 V higher for the nMOS and pMOS devices, respectively), and thus the HVT devices allow to perform the near-threshold voltage (NTV) operation for logic circuits synthesized using the standard cells without any modification. The HVT devices were effective at reducing the power consumption of the peripherals, resulting in enhancing the energy efficiency of the BNA macro. When the LVT devices were used for the peripherals, the highest energy efficiency of the macro was 58 TOPS/W. Using the HVT devices for the peripherals, the energy efficiency can be improved to 65 TOPS/W, as discussed later. The designs of the power switches are described in the supplementary material (see Section D).

The computationally synthesized XNOR, PPC, and ACTV circuits occupy 12% of the total area of the macro, i.e., the area overhead from the ULVR-SRAM array is only 12%. The maximum operating frequency of 620 MHz can be achieved at 1.2 V. The main limiting factor is the memory decoder synthesized computationally. Note that the designed cell has the ability to operate with higher than 1 GHz at 1.2 V. The macro can successfully operate with 32 MHz even at 0.4 V, as shown later.

A PIM-type BNA macro using a 6T-cell array (hereafter, referred to as a $\text{BNA}_{6\text{T}}$ macro) was also designed for comparison. The $\text{BNA}_{6\text{T}}$ macro was simply organized by replacing the ULVR-SRAM cells with the 6T cells without changing the array area for simplicity. The peripherals of the cell array and the XNOR, PPC, and ACTV circuits for the BNA macro were diverted for the $\text{BNA}_{6\text{T}}$ macro (where the circuits for V_{DDM} and V_{FB} were omitted). The low-voltage retention (LVR) mode at $V_{\text{DDL}} = 0.65$ V [17] was used for the $\text{BNA}_{6\text{T}}$ macro instead of the ULVR mode, since the 6T cell cannot achieve sufficient noise immunity at 0.2 V, as described above.

C. POWER AND ENERGY PERFORMANCES

The standby power of the BNA macro was analyzed. The SB_1^{M}, SB_2^{M} (M = Norm, EMP), and ULVR modes are applied to the cell array and the SB_1^{M}, SB_2^{M}, and shutdown (SD) modes to the peripherals. The conditions of (cell array, peripheral) = (SB_1^{M}, SB_1^{M}), (SB_2^{M}, SB_2^{M}), (SB_2^{M}, SD), and (ULVR, SD) are examined (the details for these modes are shown in Table 1). The average power for the clock cycle at the maximum frequency (discussed later) and the leakage power for the steady-state are analyzed for the SB_1^{M} and SB_2^{M} modes, respectively. For the following power/energy analyses, the cell array stores data so that 50% of the Q nodes are in the H level and the others in the L level.

Fig. 6 shows the standby power of the BNA macro for the various states described above. The results of the $\text{BNA}_{6\text{T}}$ macro are also shown in the figure. By introducing clock-gating from the ($\text{SB}_1^{\text{Norm}}$, $\text{SB}_1^{\text{Norm}}$) state, both the macros at the ($\text{SB}_2^{\text{Norm}}$, $\text{SB}_2^{\text{Norm}}$) state can effectively reduce the standby power by ∼90%. The standby power during the ($\text{SB}_2^{\text{Norm}}$, $\text{SB}_2^{\text{Norm}}$) mode can be attributed to the leakage currents of the constituent transistors. In the ($\text{SB}_2^{\text{Norm}}$, $\text{SB}_2^{\text{Norm}}$) state, the standby power of the BNA macro can be reduced than that of the $\text{BNA}_{6\text{T}}$ macro, owing to the effect of the stacking driver transistors of the cell. At the (SB_2^{EMP}, SB_2^{EMP}) state, the BNA macro can reduce the standby power by 53% from the ($\text{SB}_2^{\text{Norm}}$, $\text{SB}_2^{\text{Norm}}$) state. The standby power can be further reduced by shutting down the XNOR/PPC/ACTV circuits with the peripheral circuits of the cell array. Using the ULVR mode for the cell array, the BNA macro can reduce the standby power by 84% and 65% from its ($\text{SB}_2^{\text{Norm}}$, $\text{SB}_2^{\text{Norm}}$) and ($\text{SB}_2^{\text{EMP}}$, SB_2^{EMP}) states, respectively. Thus, the substantive PG using the ULVR mode can be achieved. Note that the break-even time (BET) is estimated to several μs or less (0.82–2.9 μs), which can allow fine-grained substantive PG [20] using the ULVR mode. The details are shown in the supplementary material (see Section F). The $\text{BNA}_{6\text{T}}$ macro reduces the standby power by only 65% from the ($\text{SB}_2^{\text{Norm}}$, $\text{SB}_2^{\text{Norm}}$) state using the LVR mode. Considering the replacement of the 6T-SRAM array with the ULVR-SRAM array, the BNA macro enables it to reduce the standby power by 93%, as shown in the figure.

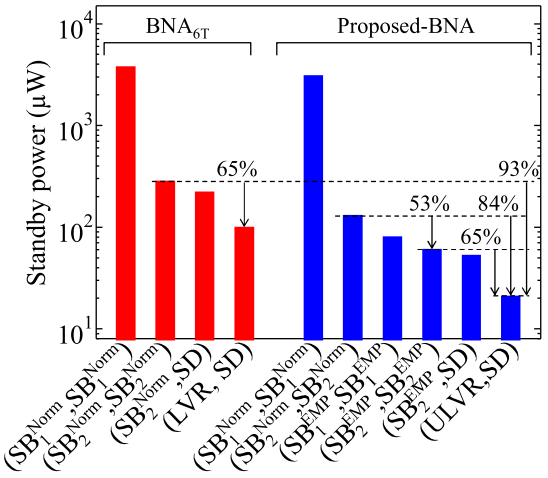

FIGURE 6. Standby power of the proposed BNA and $\text{BNA}_{6\text{T}}$ macros.

Fig. 7 shows the maximum operating frequency f_{m}, average active power P_{avg}, and cycle energy E_{cyc} of the BNA macro for the inference mode as a function of V_{DDM}, in which f_{m} is for the slowest process corner variation condition and E_{cyc} and P_{avg} are for the typical process corner variation condition (the details of the corner variations are described in Section C of the supplementary material). In the BNA macro, the cycle duration of an inference operation is defined for serially executed weight read, XNOR, PPC, and ACTV operations. E_{cyc} and P_{avg} are averaged over several tens of cycles of the inference operation. For the XNOR operations, randomly generated weight data are used for simplicity.

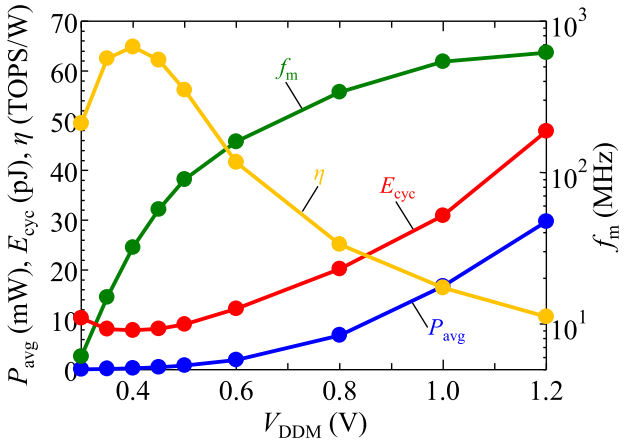

FIGURE 7. **Average active power P_{avg}, operating frequency f_m, cycle energy E_{cyc}, and energy efficiency η as a function of V_{DDM}.**

E_{cyc} is minimized at $V_{DDM} = 0.4$ V, i.e., $V_{EMP} = 0.4$ V. At this minimum point, P_{avg} is largely reduced by 1/99 in comparison to the case of $V_{DDM} = 1.2$ V, while f_m is degraded only by 1/19. This feature is highly effective at parallelizing MAC processing, as shown later. The energy efficiency η (TOPS/W) of the BNA macro is also shown in Fig. 7. η varies depending on V_{DDM} and reaches a peak at 0.4 V ($=V_{EMP}$), resulting in an energy efficiency of 65 TOPS/W. This value is seven times higher than that of the normal voltage (1.2 V) case.

V. PERFORMANCE FOR PARALLELIZED MAC OPERATIONS

In general, the allowable number of parallelized MAC operations for NN accelerators is restricted by their total power consumption. The INFER^{EMP} operation of the proposed BNA macro can drastically reduce the active power, as shown in Section IV, and thus the INFER^{EMP} operation allows the macro to enlarge the degree of MAC parallelization. The parallelization can be achieved by simultaneously processing MAC operations for multiple n-to-1 connections. For instance, the MAC processes shown by the red, blue, and green lines in Fig. 2(a) are simultaneously carried out. Hereafter, this type of parallelization is referred to as in-layer parallelization (ILP).

Fig. 8 shows a macro-based implementation structure of a single FCL using the multiple BNA macros. In the figure, the layer with 1024 neuron nodes consisting of 4 × 4 macros is shown as an example. The weight matrix $[w_{ii'}^{(j)}]$ and the bias vector $[b_i^{(j)}]$ are divided so as to adapt to the memory capacity of the macros and these divided matrices and vectors are divisionally stored in every macro. The input vector $\boldsymbol{x}^{(j-1)} = {}^t[x_1^{(j-1)} \cdots x_n^{(j-1)}]$ is also divided to $\boldsymbol{X}_u^{(j-1)}$. These are divisionally inputted to the corresponding macros. In this configuration, the constituent macros can output their MAC results before activation [also see Fig. 2(b)], and these outputs are summed up using the additional adders, as shown in the figure. Then, each activation can be obtained by the MSB of the corresponding sum (which is given by the carry of the additional adder). Thus, essentially, the macro-based

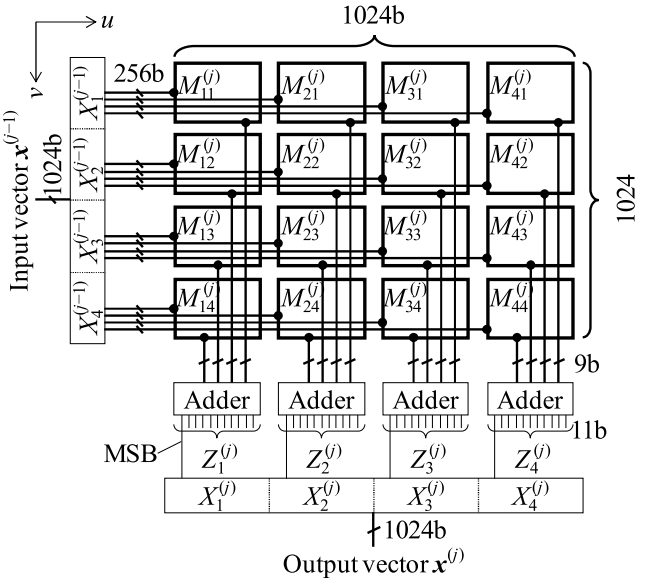

FIGURE 8. **Macro-based implementation structure of a single FCL with 1024 neuron nodes. Also see Section G in the supplementary material for the notations.**

implementation can be achieved by the multiple BNA macros and additional adders. Section G in the supplementary material describes the detailed organization method. Hereafter, an FCL configured with $N \times N$ macros is considered.

As clearly shown in Fig. 8, the macro-based implementation structure can also have parallelized MAC processes for a single neuron node (see the MAC processing along the v-direction in Fig. 8). This type of parallelization is referred to as in-node parallelization (INP). The above-described ILP corresponds to the MAC processing along the u-direction. Let N_{ILP} and N_{INP} to be the numbers of ILP and INP operations, respectively. The total number N_P of parallelized MAC operations is given by $N_P = N_{INP} \times N_{ILP}$. Note that the BNA macro having the single MAC unit can achieve the parallelization of $(N_{ILP}, N_{INP}) = (N, N)$, as shown in Fig. 8. Namely, in this case, the maximum value of N_P is N^2. For $N_{INP} > N$ or $N_{ILP} > N$, the macro needs to have multiple MAC units. Nevertheless, the BNA macro requires no multiport cell for the parallelization. Since weight data stored in the cell array are sequentially read out, adding multiple BLs and modifying the connections enable the single-port cells to multiplex the readout operations. Other notes for the macro-based BNA implementation are described in Section H of the supplementary material.

Fig. 9(a) shows the computation sequence of a single layer in the FCLs shown in Fig. 2(a), which can be configured with the multiple BNA macros, as shown in Fig. 8. Assume that the number of neuron nodes in each layer is n ($>n'$). When the constituent macros are operated one by one without any parallelization of the MAC operations, the MAC operations for an n-to-1 connection are divided into the N-part processes, which are carried out in series (along the v-direction in Fig. 8). Namely, for every macro arranged along the v-direction, divided weight and bias data are read out, and

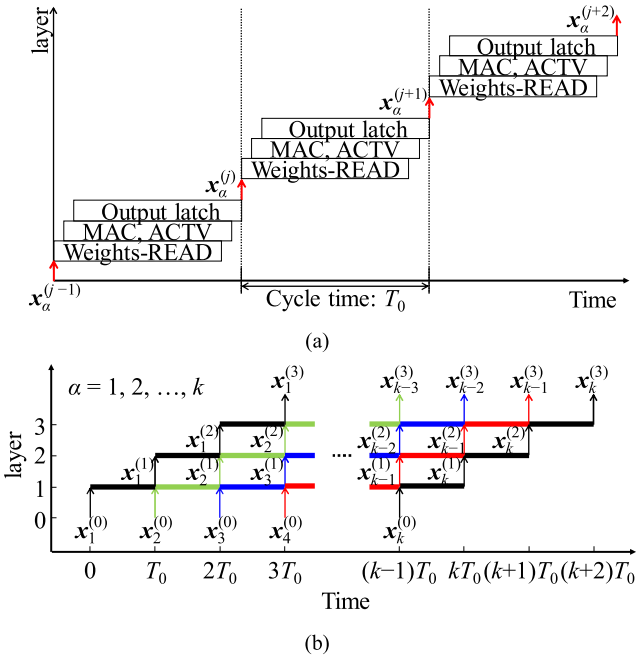

FIGURE 9. **(a) Computation sequence of a single layer in an FCL network. (b) Computation sequence of the FCL network.**

then the MAC operation is carried out with the divided input vectors using the XNOR and PPC operations. The MAC results of all the macros arranged along the v-direction are summed up. The MSB of the summed result gives the output (activation) for the n-to-1 connection. The outputs for 256 nodes can be obtained from a single column of macros arranged in the v-direction. By executing sequentially these operations for the other macro columns, the outputs for all the neuron nodes can be obtained. The output vector is latched as the input vector for the next layer. Let T_0 to be the cycle time for the operations generating an output vector. The INP and ILP MAC processes can shorten T_0, which are executed for the macros arranged along the v- and u-directions, respectively. T_0 can be roughly reduced by a factor of $\sim 1/(N_{\mathrm{INP}} \times N_{\mathrm{ILP}})$.

Fig. 9(b) shows the computation sequence of the FCL network consisting of m n-node layers, where the layer process shown in Fig. 9(a) is simply represented by the thick horizontal line. The above-described output vector generation operation is carried out from the first to the mth layer in series. The jth layer can serially process input vectors $\boldsymbol{x}_\alpha^{(j-1)} = {}^t[x_{\alpha 1}^{(j-1)} \cdots x_{\alpha n}^{(j-1)}]$ $(\alpha = 1, \ldots, k)$ for every T_0. Hereafter, the number of input vectors is denoted by k, and the symbols of physical quantities for the macro are represented by adding prime. Also, assume that n and n' can be expressed as a power of two for simplicity (e.g., $n = 2^{10}$ and $n' = 2^8$).

The above-described cycle time T_0 and the total time T_{tot} for completing the operations for all the input vectors are given by

$$T_0 = \frac{N^2 n'/N_{\mathrm{P}} + n_{\mathrm{CO}}}{f} = \frac{Nn/\left(N_{\mathrm{INP}} N_{\mathrm{ILP}}\right) + n_{\mathrm{CO}}}{f} \tag{1}$$

$$T_{\mathrm{tot}} = (k+m-1)\, T_0 = \frac{(k+m-1)\left[Nn/\left(N_{\mathrm{INP}} N_{\mathrm{ILP}}\right) + n_{\mathrm{CO}}\right]}{f} \tag{2}$$

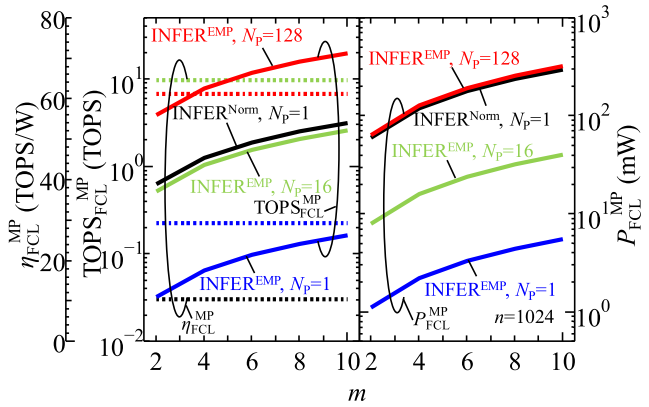

FIGURE 10. **$\mathrm{TOPS}_{\mathrm{FCL}}^{\mathrm{MP}}$, $P_{\mathrm{FCL}}^{\mathrm{MP}}$, and $\eta_{\mathrm{FCL}}^{\mathrm{MP}}$ as a function of m for the FCL networks with $n = 1024$.**

respectively, where f is the operating frequency, and n_{CO} is the cycle overhead factor caused other than the MAC operation such as output latch (in this article, $n_{\mathrm{CO}} = 2$ is used). N can be written as $N = n/n'$. The processing performance $\mathrm{TOPS}_{\mathrm{FCL}}$ (Tera operations per second) of the FCL network can be written as

$$\mathrm{TOPS}_{\mathrm{FCL}} = \frac{2n^2 mk}{(k+m-1)\, T_0}. \tag{3}$$

Assuming that $N_{\mathrm{ILP}} \geq N$ and $N_{\mathrm{INP}} \geq N$, the average power P_{FCL} of the entire network is given by

$$P_{\mathrm{FCL}} = \frac{mkN_{\mathrm{INP}}N_{\mathrm{ILP}}P'_{\mathrm{avg}} + mN^2\,(m-1)\,P'_{\mathrm{L}}}{k+m-1} \tag{4}$$

where P'_{avg} and P'_{L} is the average active power and standby (leakage) power of the BNA macro, respectively. P_{FCL} for the other conditions for N_{ILP} and N_{INP} are described in the supplementary material (see Section I). The maximum processing performance $\mathrm{TOPS}_{\mathrm{FCL}}^{\mathrm{MP}}$ and the corresponding power $P_{\mathrm{FCL}}^{\mathrm{MP}}$ are given by

$$\mathrm{TOPS}_{\mathrm{FCL}}^{\mathrm{MP}} = \frac{2n^2 m}{T_0} \tag{5}$$

$$P_{\mathrm{FCL}}^{\mathrm{MP}} = mN_{\mathrm{INP}}N_{\mathrm{ILP}}P'_{\mathrm{avg}}. \tag{6}$$

In this situation, all the layers operate simultaneously. Therefore, $\mathrm{TOPS}_{\mathrm{FCL}}^{\mathrm{MP}}$ is given by the single-layer performance $\mathrm{TOPS}_{1\mathrm{L}}$ $(=2n^2/T_0)$ multiplied by the number m of layers. $P_{\mathrm{FCL}}^{\mathrm{MP}}$ is also given by the single-layer power $P_{1\mathrm{L}}$ $(=N_{\mathrm{INP}}N_{\mathrm{ILP}}P'_{\mathrm{avg}})$ multiplied by m. Note that (5) and (6) can also be obtained from (3) and (4). In Fig. 9(b), the region where all the layers operate simultaneously is given by eliminating the pipeline prolog and epilog stages. Substantially, this can be achieved by taking the limit of $k \to \infty$ in (3) and (4). In the situation for the maximum processing performance, the energy efficiency $\eta_{\mathrm{FCL}}^{\mathrm{MP}}$ (TOPS/W) is given by

$$\eta_{\mathrm{FCL}}^{\mathrm{MP}} = \frac{2n^2}{N_{\mathrm{INP}}N_{\mathrm{ILP}}P'_{\mathrm{avg}}T_0} = \frac{2n^2 f}{(Nn + N_{\mathrm{INP}}N_{\mathrm{ILP}}n_{\mathrm{CO}})\,P'_{\mathrm{avg}}}. \tag{7}$$

Fig. 10 shows $\mathrm{TOPS}_{\mathrm{FCL}}^{\mathrm{MP}}$, $P_{\mathrm{FCL}}^{\mathrm{MP}}$, and $\eta_{\mathrm{FCL}}^{\mathrm{MP}}$ as a function of m for the FCLs ($n = 1024$), in which $(N_{\mathrm{INP}}, N_{\mathrm{ILP}}) = (1,$

TABLE 3. Performance comparison for state-of-the-art BNAs.

	Intel [7] JSSC' 21	BinarEye [8] CICC' 18	BRein [9] JSSC' 18	UNPU [10] ISSCC' 18	This work (EMP-BNA) Macro	$N_P = 1$	$N_P = 16$	$N_P = 128$
Technology	10nm FinFET	28nm CMOS	65nm CMOS	65nm CMOS	65nm CMOS (SOTB)			
Cell type	8T (Isolated read-port cell)	6T	6T	6T	10T (ULVR-SRAM cell)			
Supported networks	CNN	CNN	FCL	CNN	–	FCL ($n = 1024$, $m = 2$)		
Supply voltage (V)	0.37	0.66	0.55	0.63	0.4	0.4	0.4	0.4
Memory capacity (kB)	161	328	104	256	8.25	264	264	264
Average active power P_{avg} (mW)	5.6	1.6	60	3.2	0.25	1.1	8.0	64
Standby (leakage) power $P_{ULVR/SD}$ (μW)	–	–	–	–	0.020	0.33	0.33	0.33
(Reduction rate from 1.2V/0.4V-states)					(84%/65%)	(84%/65%)	(84%/65%)	(84%/65%)
Processing performance (TOPS)	3.4	0.35	0.36*	0.16*	0.016	0.033	0.52	3.9
Energy efficiency η (TOPS/W)	620	230	6.0	51	65	29	65	61

* Estimated using $\eta \times P_{avg}$

1), (4, 4), and (4, 32) that correspond to $N_P = 1$, 16, and 128, respectively, are examined. The black curves in the figure represent the INFER^{Norm} mode [at 1.2 V with $f = 620$ MHz ($= f_m$ at 1.2 V)] with $N_P = 1$, and the others represent the INFER^{EMP} mode [at 0.4 V with $f = 32$ MHz ($=f_m$ at 0.4 V)] with $N_P = 1$, 16, and 128. The network can be configured with $16 \times m$ BNA macros. $TOPS^{MP}_{FCL}$, P^{MP}_{FCL}, and η^{MP}_{FCL} were calculated using the analysis results of the single macro shown in Fig. 7. When $N_P = 1$, the INFER^{EMP} mode causes ~1/20 degradation of $TOPS^{MP}_{FCL}$ in comparison to the INFER^{Norm} mode. Nevertheless, it can largely reduce P^{MP}_{FCL} by ~1/50. These results reflect the features of the EMP operation for the BNA macro described in Section IV. For the INFER^{EMP} mode with $N_P = 1$, η^{MP}_{FCL} is reduced to 29 TOPS/W from 65 TOPS/W of the single macro. This is because the condition of $N_P = 1$ causes unwanted leakage power of the waiting macros with the longer T_{tot}. However, the η^{MP}_{FCL} value is higher than that of the INFER^{Norm} mode.

Using the INFER^{EMP} mode, $TOPS^{MP}_{FCL}$ can be enhanced by enlarging N_P, while the resulting increase in P^{MP}_{FCL} can be satisfactorily suppressed. For instance, the INFER^{EMP} operation with $N_P = 16$ shows almost the same $TOPS^{MP}_{FCL}$ value as the INFER^{Norm} operation with $N_P = 1$, whereas P^{MP}_{FCL} for the INFER^{EMP} mode can be suppressed to ~1/7 in comparison to the INFER^{Norm} mode. When $N_P = 16$, all the macros operate simultaneously. As a result, η^{MP}_{FCL} reaches the maximum value. This value is equal to that for the single macro (65 TOPS/W). When N_P is enlarged to $N_P = 128$, the INFER^{EMP} operation results in ~six times higher $TOPS^{MP}_{FCL}$ with the almost equivalent P^{MP}_{FCL} value in comparison to the INFER^{Norm} operation with $N_P = 1$. In this case, η^{MP}_{FCL} slightly decreases from the maximum value, since the effect of n_{CO} stands out for larger N_P [see (7)].

Table 3 shows performances of our proposed EMP operation BNAs and other state-of-the-art BNAs [7], [8], [9], [10]. In the table, our proposed BNAs are referred to as EMP-BNAs to distinguish from the others. For EMP-BNAs, an FCL network [configured with 32 macros ($n = 1024$, $m = 2$)] with $N_P = 1$, 16, and 128 and the single macro using the 8.25-kB ULVR-SRAM array are examined.

Except the FinFET-based accelerator, only the EMP-BNAs can lower the operating voltage to the EMP (also see Fig. 7). This operating voltage is comparable to that of the FinFET-based BNA. Although, in general, the average active power P_{avg} of the FCL networks is higher than that of the convolutional NNs (CNNs), the EMP-BNA with $N_P = 1$ has a low P_{avg} value at the same level as the CNN-type BNAs. Nevertheless, the processing performance TOPS of EMP-BNA with $N_P = 1$ remains at a not-so-high value. The TOPS values of the EMP-BNAs can be greatly improved with increasing N_P. A high TOPS value (comparable to the value of the FinFET-based BNA) can be obtained for $N_P = 128$. Although P_{avg} is enlarged owing to the increase in N_P, the P_{avg} value for $N_P = 128$ is suppressed to a relatively low level as an FCL network. Generally, the energy efficiency (TOPS/W) of the FCL networks tends to be lower than that of the convolution layers. The EMP-BNAs can achieve the relatively high energy efficiencies despite the FCL structure.

The FCL networks would be an example in point to analyze the performance of the EMP-BNA macro, since the energy efficiency is always lower than that of CNNs. The application of the ULVR-SRAM-based BNA macros for convolution layers is discussed in the supplementary material (see Section J), where the potential ability for improving the energy efficiency of convolution layers using the EMP operation is shown.

VI. CONCLUSION

A PIM-type BNA macro using ULVR-SRAM is proposed, and the impact of the EMP operation and the ability of the substantive PG using the ULVR mode are demonstrated. The BNA macro is designed so as to achieve stable inference operations at EMP (0.4 V) and low-power ULVR at 0.2 V. The optimally designed ULVR-SRAM cell can exhibit high noise immunity for the EMP- and ULVR-operating modes. The EMP operation of the macro can reduce the active power by 99%, enabling the large-scale parallelization of MAC processing. This EMP operation is strongly effective at achieving a high energy efficiency of 65 TOPS/W for the FCL networks. Using the ULVR mode, the standby power of the macro can be reduced by 84% with a short BET of 2.9 μs, which is applicable to substantive PG for mobile-edge applications.

REFERENCES

[1] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, *Efficient Processing of Deep Neural Networks*. San Rafael, CA, USA: Morgan & Claypool, 2020.

[2] S. Ghose, A. Boroumand, J. S. Kim, J. Gomez-Luna, and O. Mutlu, "Processing-in-memory: A workload-driven perspective," *IBM J. Res. Develop.*, vol. 63, no. 6, pp. 3:1–3:19, Nov. 2019.

[3] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018.

[4] S. Yu, W. Shim, X. Peng, and Y. Luo, "RRAM for compute-in-memory: From inference to training," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 7, pp. 2753–2765, Jul. 2021.

[5] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang, and M.-F. Chang, "Challenges and trends of SRAM-based computing-in-memory for AI edge devices," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 5, pp. 1773–1786, Mar. 2021.

[6] A. Agrawal et al., "Xcel-RAM: Accelerating binary neural networks in high-throughput SRAM compute arrays," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 8, pp. 3064–3076, Aug. 2019.

[7] P. C. Knag et al., "A 617-TOPS/W all-digital binary neural network accelerator in 10-nm FinFET CMOS," *IEEE J. Solid-State Circuits*, vol. 56, no. 4, pp. 1082–1092, Apr. 2021.

[8] B. Moons, D. Bankman, L. Yang, B. Murmann, and M. Verhelst, "BinarEye: An always-on energy-accuracy-scalable binary CNN processor with all memory on chip in 28 nm CMOS," in *Proc. Custom Integr. Circuits Conf. (CICC)*, San Diego, CA, USA, Apr. 2018, pp. 1–4.

[9] K. Ando et al., "BRein memory: A single-chip binary/ternary reconfigurable in-memory deep neural network accelerator achieving 1.4 TOPS at 0.6 W," *IEEE J. Solid-State Circuits*, vol. 53, no. 4, pp. 983–994, Apr. 2018.

[10] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H.-J. Yoo, "UNPU: A 50.6 TOPS/W unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 218–220.

[11] H. Nakahara, H. Yonekawa, T. Sasao, H. Iwamoto, and M. Motomura, "A memory-based realization of a binarized deep convolutional neural network," in *Proc. Int. Conf. Field-Programmable Technol. (FPT)*, Xi'an, China, Dec. 2016, pp. 277–280.

[12] K. Tamaru, "The trend of functional memory development," *IEICE Trans. Electron.*, vol. E76-C, no. 11, pp. 1545–1554, Nov. 1993.

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

[14] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Enabling AI at the edge with XNOR-networks," *Commun. ACM*, vol. 63, no. 12, pp. 83–90, Nov. 2020.

[15] S. Jain et al., "A 280 mV-to-1.2 V wide-operating-range IA-32 processor in 32 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2012, pp. 66–68.

[16] T. Na, S. H. Kang, and S.-O. Jung, "STT-MRAM sensing: A review," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 1, pp. 12–18, Jan. 2021.

[17] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, Mar. 2007.

[18] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mV robust Schmitt trigger based subthreshold SRAM," *IEEE J. Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, Oct. 2007.

[19] D. Kitagata, H. Yoshida, S. Yamamoto, and S. Sugahara, "Virtually nonvolatile retention SRAM cell using dual-mode inverters," in *Proc. IEEE SOI-3D-Subthreshold Microelectron. Technol. Unified Conf.*, San Francisco, CA, USA, Oct. 2018, pp. 1–3.

[20] H. Yoshida, Y. Shiotsu, D. Kitagata, S. Yamamoto, and S. Sugahara, "Ultralow-voltage retention SRAM with a power gating cell architecture using header and footer power-switches," *IEEE Open J. Circuits Syst.*, vol. 2, pp. 520–533, 2021.

[21] *VLSI Design and Education Center: VDEC*, Univ. Tokyo, Tokyo, Japan, 2022. [Online]. Available: http://www.vdec.u-tokyo.ac.jp/English/index.html

[22] S. Ohbayashi et al., "A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits," *IEEE J. Solid-State Circuits*, vol. 42, no. 4, pp. 820–829, Mar. 2007.