Proposal and performance prediction of a BNN accelerator using ULVR-SRAM

ULVR-SRAM を用いた BNN アクセラレータの提案と性能予測

°Y. Shiotsu, S. Yamamoto, and S. Sugahara, *FIRST, Tokyo Inst. of Tech.*
東工大未来研 °塩津勇作, 山本修一郎, 菅原聡
E-mail: y.shiotsu@isl.titech.ac.jp

Introduction: In future smart society, neural network accelerators (NNAs) play an essential role for deep learning applications in mobile edge computing, in which these devices need to have much lower computation energy for the realization. Processing-in-memory (PIM) architectures [1] are expected to be effective at improving the performance of these NNAs. In the PIM architecture, data retrieved from the cells are directly processed without transferring them through a bus. Therefore, the PIM architecture can effectively parallelize the multiply-accumulate (MAC) processing in neural networks without the constraint originating in bus usages, which is promising for enhancing the energy performance of NNAs. For PIM-based NNAs, reducing active/standby power of the constituent SRAM is indispensable for the energy-efficient operations. The SRAM should equip ever-more-sophisticated operating modes, in particular, the minimum energy point operation and power gating (PG) modes. Note that the former is promising for large-scale parallelization of the MAC processing in neural networks.

Recently, we have proposed a fully CMOS-based ultralow-voltage-retention SRAM (ULVR-SRAM), which can retain data even at an ultralow voltage (V_{UL}) and the substantive PG can be achieved using the ULVR [2,3]. In addition, the ULVR-SRAM can perform the energy minimum SRAM operation at V_{min} and the high-performance SRAM operation at an ordinary supply-voltage V_{DD} ($V_{UL} < V_{min} < V_{DD}$) [3]. In this study, an ULVR-SRAM-based PIM-type NNA is proposed and the predicted performance is discussed. The V_{min} operation enables it large-scale parallelization of the MAC processing and the ULVR operation can achieve substantive PG. This type of accelerator based on binarized neural network (BNN) can show a high processing performance of 100 TOPS/W.

Circuit/PIM configurations: Fig. 1 shows the ULVR-SRAM cell that uses pMOS feedback Trs (FBTs) in the Schmitt-trigger-type dual-mode inverters. A virtual supply-voltage (VV_{DD}) is generated through the header power switches (PS1, PS2, and PS3) from three supply-voltages (V_{DDH}=1.2V, V_{DDM}=0.4V, and V_{DDL}=0.2V). During the normal SRAM operation, energy-minimum V_{min} operation, and ULVR modes, VV_{DD} is set to V_{DDH}, V_{DDM}, and V_{DDL}, respectively. The bias (V_{FB}) of the FBTs are set as follows: V_{FBM} (=0.4V) for the V_{min} operation mode and V_{FBL} (=0.2V) for both the normal SRAM operation and ULVR modes [2,3]. Fig. 2 shows a layout of designed PIM macro configured with an 8kB ULVR-SRAM macro [3] and an MAC processing unit. In this study, a BNN architecture is used and thus the MAC unit can be simply configured with XNOR gates and a POP counter. This PIM macro can simultaneously read 256-bit weight data and execute MAC calculations for these weight data and a 256-bit input vector. The cell array is custom-designed and the peripheral circuits including the MAC unit are synthesized based on RTL. The LP model of the 65nm SOTB devices is used for the CMOS transistors. Post-layout large-scale simulations are carried out for the designed macro using FineSim. A 6T-SRAM-based PIM macro is also designed as a reference.

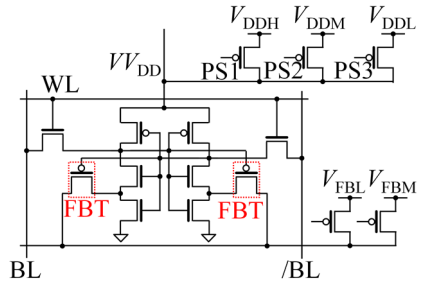

Fig. 1. ULVR-SRAM cell with power- and control-switches.

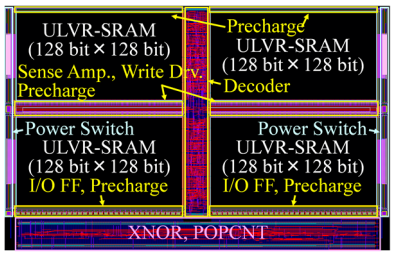

Fig. 2. Layout of the PIM macro.

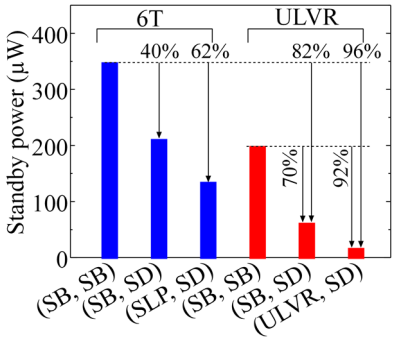

Fig. 3. Standby power of the PIM macros using ULVR-SRAM and 6T-SRAM.

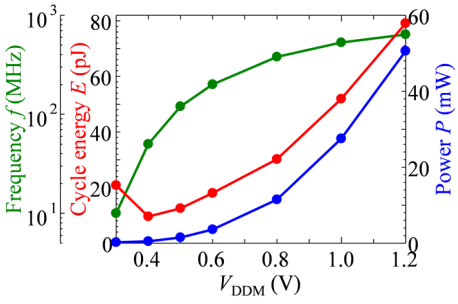

Fig. 4. P, f, and E as a function of V_{DDM}.

Analysis results: Fig. 3 shows standby power of the PIM macro, where the standby (SB) and ULVR modes are performed at 1.2V and 0.2V, respectively, and the SB and shutdown (SD) modes are used for the peripheral circuits. The results of the 6T-SRAM-based PIM macro with the sleep (SLP) mode at 0.8V are also shown in the figure. The ULVR mode of ULVR-SRAM can reduce the standby power by 92% from that in the SB mode. Considering the replacement of the 6T-SRAM by the ULVR-SRAM, the standby power reduction can reach to 96%, as shown in the figure. Fig. 4 shows the operation frequency f, average power P, and cycle energy E of the ULVR-SRAM-based PIM macro as a function of V_{DDM}. E is minimized at V_{DDM}=0.4V, i.e., V_{min}=0.4V. Note that at this V_{min} point, P is reduced by 1/100, while f is degraded only by ~1/10. This feature is highly effective at parallelizing the MAC processing. For instance, when 100 parallel MAC operations are employed for the V_{min} operation, the computation amount is 10 times higher than that for the V_{DD} operation. Nevertheless, it causes almost the same power consumption as the V_{DD} operation. The PIM macro can easily parallelize the MAC processing, and also the large-scale parallelization can be achieved using multiple these PIM macros. Using the parallelization of the MAC processing with the V_{min} operation, the BNN PIM can show a processing performance of ~100 TOPS/W.

Acknowledgement: This work was supported by the VDEC, the University of Tokyo, in collaboration with Synopsys, Inc.

References: [1] K. Ando *et al.*, IEEE J. Solid-State Circuits **53**, 983–994, 2018. [2] Y. Shiotsu *et al.*, the 81st JSAP Autumn Meeting, 2020, 11a-Z09-10. [3] Y. Shiotsu *et al.*, the 68th JSAP Spring Meeting, 2020, 17a-Z26-5.