Energy-efficient Spin-orbit Torque MRAM Operations for Neural Network Processor

Liang Chang[*], Zixuan Zhu[*], Zhen Zhu[*], Siqi Yang[*], Weihang Li[*], and Jun Zhou[*]
[*]University of Electronic Science and Technology of China, Chengdu, China, 611731.
liangchang@uestc.edu.cn, 2018270102018@std.uestc.edu.cn, 201921010224@std.uestc.edu.cn,
siqi_yang@std.uestc.edu.cn, liwh15@std.uestc.edu.cn, zhouj@uestc.edu.cn

2021 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-7281-9201-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCAS51556.2021.9401449

***Abstract*—Emerging energy-efficient neural network processor is a promising hardware design to accelerate neural network algorithms with high performance and low power consumption. Typically, static random-access memory (SRAM) is employed to develop large buffers using in the processor. The bit cell of SRAM contains six transistors, leading to low density and large leakage current. In particular, several AI processors need multiple port and transfer-based SRAMs, which decrease the density and increase the power consumption. Recently, emerging spin-orbit torque magnetic random-access memory (SOT-MRAM) becomes a possible solution to replace the SRAM as working memory. However, more operations should be supported by the SOT-MRAM to provide sufficient functions, such as multiple-port memory, transpose memory, data-streaming operations. In this paper, we develop the working memory of neural network processor with SOT-MRAM to build the design library including the transpose operations, multiple-port memory, and data-streaming based buffer arrays. Equiped with those operations provided by SOT-MRAM, we can build high performance and energy-efficient neural network processors.**

***Index Terms*—Artificial Intelligent, Transpose operation, Neural Network, Spin Orbit Torque, Magnetic Random Access Memory**

I. Introduction

Energy efficiency is a critical factor for wearable devices and internet of things (IoTs), such as wearable biomedical signal monitoring, industrial detecting, and voice classification systems [1] [2]. In these systems, the low-power artificial intelligent (AI) processor is required to identify emergency/critical signals. Recent developments indicate that the deep neural networks (DNNs) can improve the accuracy of the detection and classification. Typically, deep convolutional neural network (DCNN) applications are both compute- and memory- intensive algorithm, which is difficult to be deployed on the energy-efficient neural network processor [3]. Several optimization techniques should be implemented to improve the efficiency of the neural network processor from both algorithm and hardware sides.

In the algorithm side, light-weighted neural networks are used to optimize DNNs algorithm, which uses minimal neural layers and channels to complete the signal detection and classification [4]. In addition, the weight quantization and pruning optimization techniques have demonstrated superior performance [5] [6]. Both of them can improve the computation performance by reducing the scale of parameters hence to decrease computation operations and memory requirements.

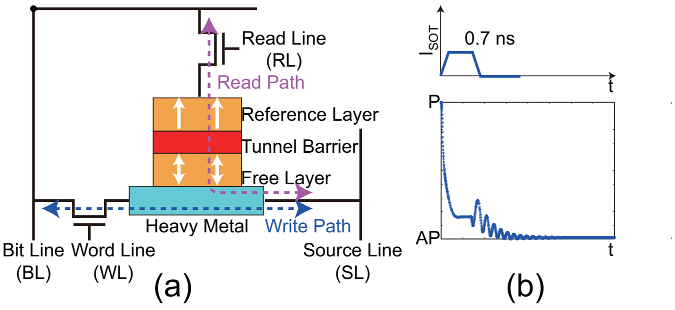

Fig. 1. The memory cell of SOT-MRAM [8]. (a) In this memory, there are two control transistors to connect to BL, and a spin-orbit torque magnetic tunnel junction (SOT-MTJ). (b) The write pulse of the write operation of the memory cell.

Furthermore, the irregular network structure becomes another issue increasing the training complexity and decreasing the inference accuracy. This limitation can be solved by reshaping the matrices with block-circulant neural network which has been validated by previous literature [7].

In the hardware side, the systolic-array and shift-register array techniques were proposed to improve the accelerating efficiency [9] [10]. The systolic-based neural network accelerator is organized with processing-element (PE) array obtaining data from on-chip memory. The benefit of the systolic-based accelerator is simple design with high parallelism and balanced input/output. For the shift-register based array, the execution is performed instructions operating on the image pixel from the two-dimensional shift register array. The data can be transferred between different shift registers to provide flexible computation. In these hardware accelerator, large data movements between on and off-chip become significant issue for the energy-efficient neural network processor. Normally, the static random-access memory (SRAM) is distributed inside the accelerator to reduce data movements between on- and off-chip. However, using large SRAMs occupy too much area and consume high power including both dynamic and static power consumption, due to the six-transistor memory cell.

Recently, the non-volatile memory technologies have been employed to alternative the SRAM as working memory [11] [12]. Advantages of non-volatile technologies contain high density, ultra-low leakage power, simple one/two-transistor one-resistance structure, and nonvolatility [13] [14] [8] [15]. Several specific characteristics can be used to design novel memory-array structure to improve the performance of neural

978-1-7281-9201-7/21/$31.00 ©2021 IEEE

network processor. In this paper, we propose a transpose-based spin-orbit torque magnetic random-access memory (SOT-MRAM) array to implement the in-situ transpose operation. The transpose memory array can be used to build the online training neural network processor. In addition, we develop an odd-even memory array structure to support the multi-port SOT-MRAM with parallel write and read operations. Based on the simulation and evaluation, our proposed transpose- and multi-port based SOT-MRAM outperform the SRAM design in terms of area and power consumption.

The rest of this paper is organized as follows: Section II introduces the background and organization of the transpose memory. Section III presents the memory array structure and modifications of the design components for the proposed transpose SOT-MRAM. We discuss analytic and evaluation results in Section IV. Finally, Section V concludes the paper.

II. Background and Motivation

This section provides the preliminary and development of the SOT-MRAM. Also, we present the transpose-operation based neural network accelerator. The SOT-MRAM is not only suitable for the memory structure, but also can be used to design the advanced computing architecture.

A. The Development of the SOT-MRAM

In the past two decades, the interests in magnetic materials with strong spin-orbit coupling has attracted many attentions thanks to theirs potential for applications. Many semiconductor companies have equipped product line for the MRAM, such as Everspin, Globalfoundary, Samsung technology, Toshiba, and Taiwan Semiconductor Manufacturing Company (TSMC) [16]. Traditionally, the MRAM has been employed on satellite, airplane, and wearable watch. With the embedded MRAM, the instrument can be used for a long period.

The basic feature of the SOT-MRAM includes slow write latency and high energy-efficiency. The core element of SOT-MRAM is the SOT- magnetic tunnel junction (MTJ) that consisting of reference layer, oxide barrier, and free layer attached to a heavy metal, as shown in Fig. 1(a). There are two stable states of the SOT-MTJ, parallel state represents the low resistance, and anti-parallel state represents the high resistance. It is possible to switch the magnetization of the free layer for the SOT-MTJ when sufficient large current is added on heavy metal with large spin-Hall angle [17]. Normally, a 0.4 $0.7ns$ pulse current is applied in the x-direction of the SOT-MTJ, and the Hall voltage can be measured in the y-direction. This phenomenon can be concluded that the spin current generated by spin-Hall can induce a torque to switch the magnetization. Typically, the memory cell of the SOT-MRAM consists of two transistors to separately control the read and write operations. Even though two control transistors are required, the area overhead is still smaller than that of SRAM. Moreover, the read and write paths are separated in the SOT-MTJ, which is suitable for developing the multi-port MRAM [8].

We analyze the behavior of memory cell by the transient spice simulation, as shown in Fig. 1 (b).With the three-terminal SOT-MTJ, we can develop a symmetric memory cell and high-speed memory array structure. The paper is based on the transpose SOT-MRAM to analyze the possible architecture for the future neural network processors. Therefore, we mainly focus on the analyses of advantages provided by the SOT-MRAM. For the end-to-end experiment of a system designed with SOT-MRAM is beyond the scope of this work, which is our future work direction.

B. The memory-centric and transpose-based NN accelerator

Rapid development of neural network applications promotes the memory-centric neural network processor to distribute more memory inside the processor. The data movements between memory and process unit can be reduced hence to increase energy-efficiency of the neural network processor. Recently, SRAM-embedded convolution computation architecture was presented for running binary-weight neural networks. The classification accuracy is close to the digital implementations and may much better than prior in-memory approaches [18]. After that, many high precision SRAM-based in-memory approaches have been proposed to accelerate neural networks beyond the MNIST and LeNet-5 convolutional neural network (CNN) [19]–[21]. In addition, the circulant matrices-based CNN was proposed to overcome limitations induced by the irregular network structure. The weight matrices can be reorganized to decrease the training complexity, which utilizes the fast Fourier transform-based fast multiplication [7]. In this work, the transpose SRAM can be developed to accelerate the matrix transpose operation. Very recently, a computing-in memory architecture is proposed to support on-chip training using the transpose SRAM arrays [22] [23]. Both seven-transistor and eight-transistor SRAM are employed to implement bi-directional vector-to-matrix multiplication and data flow of back-propagation process. However, the potential of existing solutions is limited by the buffer with SRAM. The buffer can be replaced by the SOT-MRAM to accelerate the computation. In reference [15] [24], the SOT-MRAM was used to develop the data-streaming based computing-in memory architecture, which has demonstrated the significant energy-efficiency of the SOT-MRAM based solution [15]. Consequently, we develop the transpose, multi-port, and data-streaming SOT-MRAM, as a design kit, to the energy-efficient neural network processor [8] [15].

III. Architecture of Energy-efficient SOT-MRAM Operations

This section introduces the basic idea of the energy-efficient SOT-MRAM operations including transpose operation, multi-port memory array, data-streaming based operation. Those operations can be used to design the neural network processor to replace the SRAM. Also, detailed memory cell and peripheral circuits are provided with the control flow.

A. In-situ Transpose Operation of SOT-MRAM

As shown in Fig. 2 (a) and (b), the SOT-MTJ is symmetrical between Bit Line (BL) and Source Line (SL). The sense

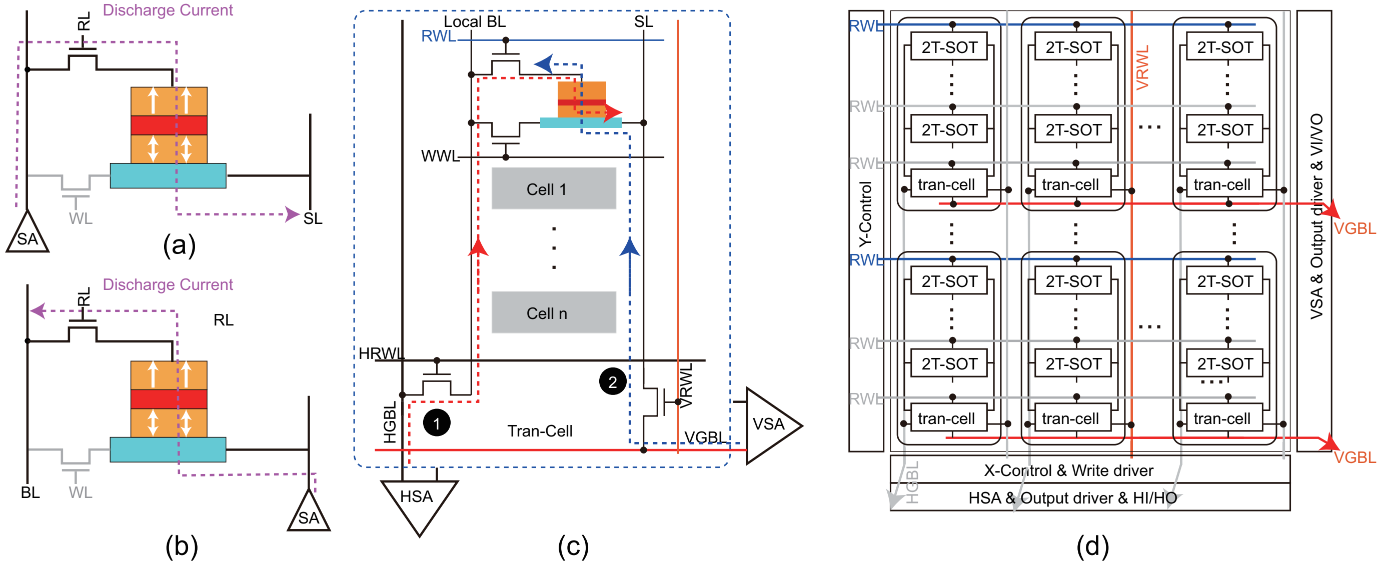

Fig. 2. The transpose SOT-MRAM array structure [8]. Read circuit connection with(a) the normal read circuit connected to the BL and (b) The transpose read circuit connected to the BL. (c) The transpose sub-array structure with two transpose transistors. (d) The proposed memory array structure with the transpose sub-array (We only indicate read wires for this structure).

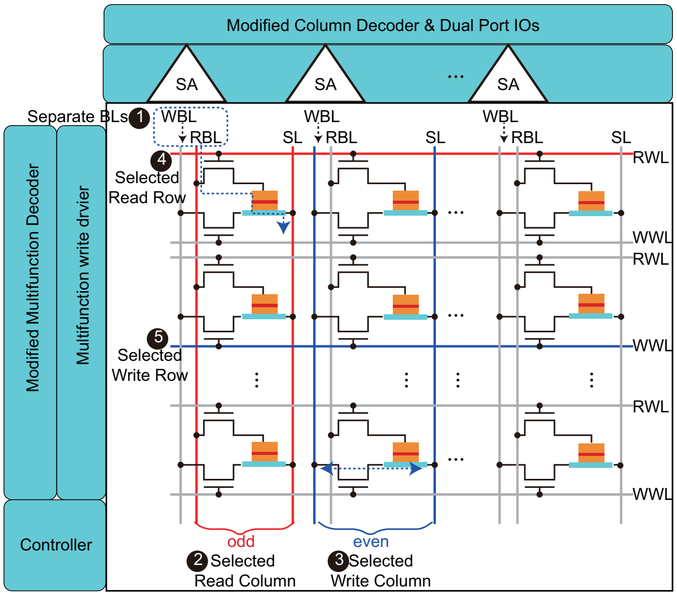

Fig. 3. The proposed dual-port memory array with separated WBLs and RBLs, respectively. Several components should be modified including multi-functional row and column decoders, write driver and IOs [8].

amplifier (SA) can be located to connect BL and the discharge current flows through BL to SL. Equally, the SA can be located to connect SL and the discharge current passes through SL to BL. Therefore, we can use this characteristic of the SOT-MRAM to design the transpose memory array in the block level, as shown in Fig. 2. In this memory sub-array, there are several memory cells and a transpose cell. The transpose cell contains two control transistors, each of them is connected to corresponding SA. Based on the analyses of Fig. 2 (a) and (b), the discharge current path is the same between the horizontal and vertical read paths, as indicated by the red and blue lines, respectively. Based on this memory sub-array, the transpose SOT-MRAM array can be organized as shown in Fig. 2 (d). The architecture includes horizontal I/O (HI/HO), additional vertical I/O (VI/VO), sub-arrays, and vertical and horizontal RLs (VRL and HRL). The VI/VO supports the transpose operation by activating the VRL and corresponding RLs. The HRL and HI/HO are used for the normal access operation. With this T-SOT-MRAM, both normal and transpose access operation can be supported to accelerate the computation for the transpose and circulant matrices. In the following, we introduce the T-SOT-MRAM in more details. It is worth mentioning that the T-STT-MRAM could be designed with a similar structure.

B. Dual-Port Operation of SOT-MRAM

Fig. 3 gives the memory array for the dual-port SOT-MRAM. The BL of the memory array is separated into WBL and RBL as write and read operations (①), respectively. The column of the memory array is defined as odd and even as shown in Fig. 3 (② and ③). At meanwhile, the row of the memory array is labeled as selected read and write rows (④ and ⑤). For supporting the dual-port read/write operations, the control logic of the memory array is modified to support multi-functional operation. Firstly, both row and column decoders should be modified to select difference control lines according to the required operations. Secondly, the SA and write driver are connected to RBL and WBL, respectively, so that read and write operation could be independently controlled. In the dual-port SOT-MRAM, the read and write operation are performed in the odd and even columns in a clock cycle. For the read operation, there are pre-charge and discharge operations controlled by SAs. During the pre-charge operation and negative clock edge, a read-enable pulse is triggered on

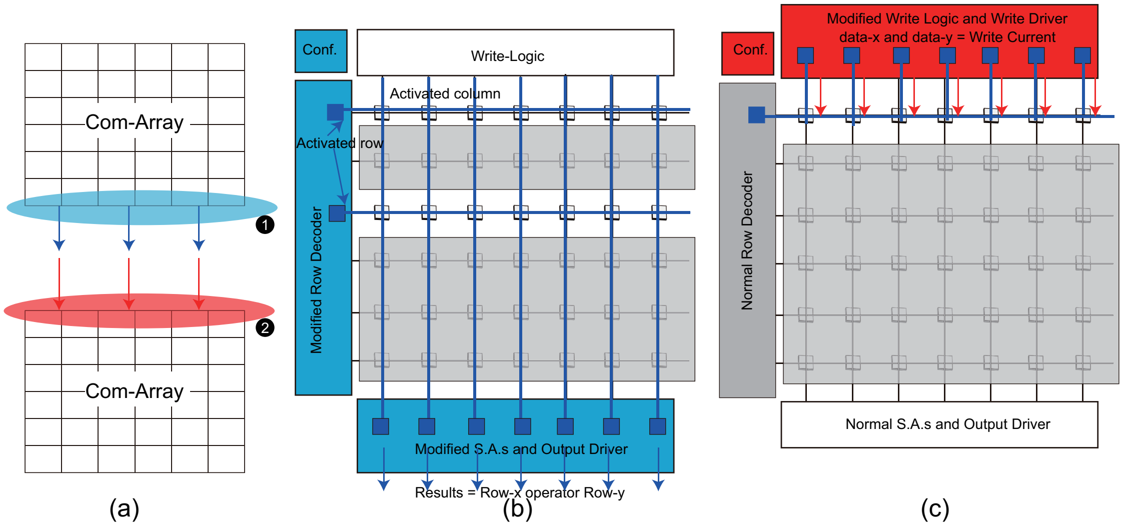
Com-Array
1
2
Com-Array
(a)
Conf.
Write-Logic
Activated column
Activated row
Modified Row Decoder
Modified S.A.s and Output Driver
Results = Row-x operator Row-y
(b)
Conf.
Modified Write Logic and Write Driver
data-x and data-y = Write Current
Normal Row Decoder
Normal S.A.s and Output Driver
(c)

Fig. 4. The proposed data-streaming computing blocks [15]. (a) The proposed data streaming model with both read- and write-computing operations. (b) The read-computing based memory array with modified control unit, sense amplifier, and reconfigurable logic unit. (c) The write-computing based memory array with modified input control unit, reconfigurable logic unit, and write driver.

the selected column. Meanwhile, the write enable is triggered at the WWL and activates one memory row. As a result, the write current passes through the heavy metal of SOT-MTJ to program the selected memory cell.

C. Data-streaming Operation of SOT-MRAM

The data streaming model contains the data movements with computation, as shown in Fig. 4 (a). For example, ① is a read-computing operation, while ② is a write-computing operation. The SOT-MRAM can be used to implement this model by modifying several peripheral components. For supporting the data streaming operations, we provide the computation memory array including read and write computing operations, as shown in Fig. 4 (b) and (c). For the read-based computation block, as shown in Fig. 4 (b), the control logic, row decoder and sense amplifier should be modified to support multiple rows (at least two rows) activation. The benefit of this structure is provide high parallel operations, such as AND, OR, XOR and XNOR operations. Normally, several special logic units are required including accumulator and shift logic.

For the write-based computation patterns, as shown in Fig. 4 (c), we modify the input control unit and write driver to support the bitwise operation such as AND, OR, XNOR, Majority/Minority, and Max/Min. Particularly, the SOT-MRAM support the threshold write operation. The SOT-MTJ always holds the original value unless sufficient pulse width/amplitude of current and voltage are applied. With the read- and write-based computation patterns, we can develop the streaming computation model overlapping memory access (write and read) and computing latency by combining the Com-Array. In addition, all data movements can also be recognized as computation of the intermediate data by the buffer or connection.

IV. Discussion of the Design Kit of SOT-MRAM

In the design kit of SOT-MRAM, we provide transpose, dual-port and data streaming operations. In the transpose memory array, a transpose cell should be added into the subarray. The function of the transpose SOT-MRAM has been validated by the previous work [8]. Compared to transpose SRAM, our proposed SOT-MRAM can provide similar latency with lower leakage power consumption. For dual-port SOT-MRAM, the memory array structure is adjusted with separating WWL and RWL. The area of the proposed SOT-MRAM is much lower than the dual-port SRAM thanks to the two transistor one resistance memory cell with only one additional WL for the column. In particular, as the memory capacity of dual-port SOT-MRAM scaling up, the benefit of the proposed memory structure is much better. For data streaming based SOT-MRAM, we develop both write- and read-based computing blocks to support data-streaming model [24] [15]. We integrate those three functional memory array into the modified NVSim simulator to evaluate the performance. The parameter of the SOT-MRAM can be used to design the neural network processor.

V. Conclusion

As SOT-MRAM is featured by energy-efficient write operation, ultra-low power consumption, separated write and read path, we investigated different components required by neural network processors. In this paper, we proposed transpose, dual-port, and data streaming operation-based SOT-MRAMs that can be applied in the neural network processor. Based on characteristics of the memory cell for SOT-MRAM, the performance of each solution is better than SRAM counterparts. As future work, we can provide scalable SOT-MRAM array structure with more operations to replace more components in neural network processors.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work was supported by National Key R&D Program of China under grant 2019YFB2204500 and UESTC Research start-up funding under grant Y030202059018052. J. Zhou is the corresponding author.

REFERENCES

[1] J. Liu, Z. Zhu, Y. Zhou, N. Wang, G. Dai, L. Qingsong, J. Xiao, Y. Xie, Z. Zhong, H. Liu, L. Chang, and J. Zhou, "Bioaip: A reconfigurable biomedical ai processor with adaptive learning for versatile intelligent health monitoring," in *2021 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2021.

[2] Y. Wei, J. Zhou, Y. Wang, Y. Liu, Q. Liu, J. Luo, C. Wang, F. Ren, and L. Huang, "A review of algorithm hardware design for ai-based biomedical applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 2, pp. 145–163, 2020.

[3] M. Kim and J. Seo, "Deep convolutional neural network accelerator featuring conditional computing and low external memory access," in *2020 IEEE Custom Integrated Circuits Conference (CICC)*, 2020, pp. 1–4.

[4] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2016.

[5] M. Wess, S. M. P. Dinakarrao, and A. Jantsch, "Weighted quantization-regularization in dnns for weight memory minimization toward hw implementation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2929–2939, 2018.

[6] F. Sun, M. Qin, T. Zhang, L. Liu, Y. K. Chen, and Y. Xie, "Invited: Computation on sparse neural networks and its implications for future hardware," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020, pp. 1–6.

[7] J. Yue, R. Liu, W. Sun, Z. Yuan, Z. Wang, Y. Tu, Y. Chen, A. Ren, Y. Wang, M. Chang, X. Li, H. Yang, and Y. Liu, "7.5 a 65nm 0.39-to-140.3tops/w 1-to-12b unified neural network processor using block-circulant-enabled transpose-domain acceleration with 8.1 × higher tops/mm2and 6t hbst-tram-based 2d data-reuse architecture," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 138–140.

[8] L. Chang, Z. Wang, Y. Zhang, and W. Zhao, "Multi-port 1r1w transpose magnetic random access memory by hierarchical bit-line switching," *IEEE Access*, vol. 7, pp. 110 463–110 471, 2019.

[9] Y. Chen, T. Yang, J. Emer, and V. Sze, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 292–308, 2019.

[10] R. Jason, M. Albert, G.-H. Nathan, V. Altem, and O. Shacham, "Pixel visual core: Google's fully programmable image, vision, ai processor for mobile devices," in *2021 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2021.

[11] T. Chang, Y. Chiu, C. Lee, J. Hung, K. Chang, C. Xue, S. Wu, H. Kao, P. Chen, H. Huang, S. Teng, and M. Chang, "13.4 a 22nm 1mb 1024b-read and near-memory-computing dual-mode stt-mram macro with 42.6gb/s read bandwidth for security-aware mobile devices," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 224–226.

[12] W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, and H. . P. Wong, "33.1 a 74 tmacs/w cmos-rram neurosynaptic core with dynamically reconfigurable dataflow and in-situ transposable weights for probabilistic graphical models," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 498–500.

[13] L. Chang, Z. Wang, A. O. Glova, J. Zhao, Y. Zhang, Y. Xie, and W. Zhao, "Prescott: Preset-based cross-point architecture for spin-orbit-torque magnetic random access memory," in *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2017, pp. 245–252.

[14] L. Chang, X. Ma, Z. Wang, Y. Zhang, Y. Xie, and W. Zhao, "Pxnor-bnn: In/with spin-orbit torque mram preset-xnor operation-based binary neural networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 11, pp. 2668–2679, 2019.

[15] L. Chang, X. Ma, Z. Wang, Y. Zhang, Y. Ding, W. Zhao, and Y. Xie, "Dasm: Data-streaming-based computing in nonvolatile memory architecture for embedded system," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2019.

[16] Z. Wang, Z. Li, Y. Liu, S. Li, L. Chang, W. Kang, Y. Zhang, and W. Zhao, "Progresses and challenges of spin orbit torque driven magnetization switching and application," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

[17] M. Wang, W. Cai, D. Zhu, Z. Wang, J. Kan, Z. Zhao, K. Cao, Z. Wang, Y. Zhang, T. Zhang *et al.*, "Field-free switching of a perpendicular magnetic tunnel junction through the interplay of spin–orbit and spin-transfer torques," *Nature Electronics*, vol. 1, no. 11, p. 582, 2018.

[18] A. Biswas and A. P. Chandrakasan, "Conv-ram: An energy-efficient sram with embedded convolution computation for low-power cnn-based machine learning applications," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 488–490.

[19] W. Khwa, J. Chen, J. Li, X. Si, E. Yang, X. Sun, R. Liu, P. Chen, Q. Li, S. Yu, and M. Chang, "A 65nm 4kb algorithm-dependent computing-in-memory sram unit-macro with 2.3ns and 55.8tops/w fully parallel product-sum operation for binary dnn edge processors," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 496–498.

[20] X. Si, J. Chen, Y. Tu, W. Huang, J. Wang, Y. Chiu, W. Wei, S. Wu, X. Sun, R. Liu, S. Yu, R. Liu, C. Hsieh, K. Tang, Q. Li, and M. Chang, "24.5 a twin-8t sram computation-in-memory macro for multiple-bit cnn-based machine learning," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019, pp. 396–398.

[21] X. Si, Y. Tu, W. Huang, J. Su, P. Lu, J. Wang, T. Liu, S. Wu, R. Liu, Y. Chou, Z. Zhang, S. Sie, W. Wei, Y. Lo, T. Wen, T. Hsu, Y. Chen, W. Shih, C. Lo, R. Liu, C. Hsieh, K. Tang, N. Lien, W. Shih, Y. He, Q. Li, and M. Chang, "15.5 a 28nm 64kb 6t sram computing-in-memory macro with 8b mac operation for ai edge chips," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 246–248.

[22] H. Jiang, X. Peng, S. Huang, and S. Yu, "Cimat: A compute-in-memory architecture for on-chip training based on transpose sram arrays," *IEEE Transactions on Computers*, vol. 69, no. 7, pp. 944–954, 2020.

[23] H. Jiang, S. Huang, X. Peng, J. W. Su, Y. C. Chou, W. H. Huang, T. W. Liu, R. Liu, M. F. Chang, and S. Yu, "A two-way sram array based accelerator for deep neural network on-chip training," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020, pp. 1–6.

[24] L. Chang, X. Ma, Z. Wang, Y. Zhang, W. Zhao, and Y. Xie, "Corn: In-buffer computing for binary neural network," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 384–389.