

Implementation of Sram Architecture to Perform in Memory Computation

Adrish Ray, Devadathan P K

Department of Electronics and Communication Engineering
National Institute of Technology, Tiruchirappalli

Abstract—The Von Neuman architecture for computing systems has been around for decades. In present times, with the advent of machine learning and artificial intelligence, the need for continuous evolution of the computing systems is required. This has led to the development of newer and more efficient modes of computation. This is where In Memory computation comes into picture. In Memory computation refers to the notion of performing computation in the memory unit of the computer system to reduce delay and conserve power. We have presented the implementation of NAND Boolean logic using existing SRAM bit cell architectures. The NAND is considered as the universal logic; thus it can be the building block of various other operations inside the memory unit. The designs were made using 65nm technology. We have presented the computations using 8T and dual Vt 7T SRAM bit cell.

Index Terms—In memory computing, dual Vt 7TSRAM, 8T SRAM, NAND operation.

I. INTRODUCTION

A. Motivation

Artificial Intelligence (AI) and Machine learning (ML) are changing the world and making it more accessible to us. One of the main sub-branches of Artificial Intelligence and Machine Learning is Convolutional neural networks abbreviated as CNN that are used in a wide variety of applications, like image classification and speech recognition. For this reason, there is growing pressure among engineers to improve the hardware resources to meet the many needs.

The traditional Von Neumann architecture uses traditional techniques for transferring data between memory and the CPU. The main drawback of the system is the significant energy loss and latency costs along with the huge processing time. Most of the power consumption and latency is expended for accessing the memory and moving of data. For this reason, that is, to minimize the power and latency, in-memory computation (IMC) has recently been proposed for processing data directly inside the on-chip memory macro.

B. Introduction to 6T, 7T and 8T SRAM

SRAM stands for static random access memory. These are a type of Random Access Memory. They have found applications in a variety of microelectronic applications because of their very fast operation. They are primarily used as memory cache for embedded applications. In this of types of memory the data is stored if the power is supplied. The main challenge of SRAM is their bigger size. A basic 6T architecture requires

for six transistors to be interconnected. This way the density of transistors over a given area reduces. Hence the SRAM bit cells usually are designed with transistor devices of minimum size in order to achieve high packing density. The advantage of the SRAM is that it is less power hungry. It is thus employed when bandwidth, low power, or both are important requirements. There are two design aspects considered while designing SRAM bit cell, which are the power it dissipates during the read and write operations and the propagation delay that occurs while reading and writing. The power dissipation is dynamic. The power analysis It aids in the estimation of portable devices' battery life. The delay in reading and writing determines the speed of SRAM. Because of its high storage density and short access time, SRAM has become a key component in many VLSI chips. Due to the rapid growth of low power, low voltage memory design in recent years due to increased demand for notebooks, laptops, IC memory cards, and hand-held communication devices, SRAM has become a major research area. Because of their ease of use and low standby leakage, SRAMs are commonly employed in mobile applications as on-chip and off-chip memory.

6T static random-access memory is a type of semiconductor memory that stores each bit using bistable latching circuitry. The phrase "static" means "not moving." It is distinguished from dynamic RAM, which must be periodically updated. SRAM displays information. However, traditional remembering is still volatile. When memory is not available, data is finally lost. Figure 1 depicts the structure of a 6T SRAM cell. Four transistors constitute a pair of inverters in this construction, which are used to store a bit of information, while the remaining two transistors are called access transistors, which are used to read and write data to the inverter pair. The word line (WL) controls these access transistors, allowing the two-bit lines (BL and BL Bar) to access the memory elements. In this research, dual bit lines are used because they have a better noise margin than single bit lines. 7T SRAM cell is designed to improve read cycle and reduce static power. In the feedback structure, an extra transistor is utilized to amplify high values on one inverter from low values on another inverter, and vice versa. The WRITE SELECT and READ SELECT signals are used to write and read from the memory cell utilizing two different transistors. The read operation turns on the transistor, whereas the write operation turns it off. By sizing the transistors so that the read bit line

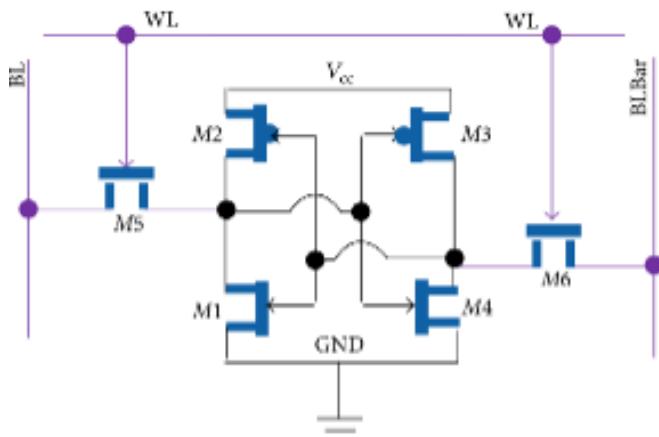


Fig. 1. Basic 6T SRAM

may be charged faster, the read cycle can be improved. Figure 2 depicts the structure of a 7T SRAM cell. Figure 3 shows

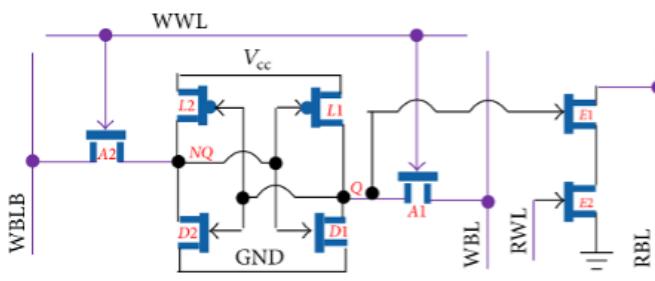


Fig. 2. Conventional 7T SRAM Design

the 8T SRAM cell structure, which is like the 6T but includes additional transistors to protect the internal inverter against inadvertent write during the read cycle. The read bit line (RBL) is precharged to the supply voltage before the read cycle. The read operation then begins by asserting RWL, with RBL remaining in either a logic "1" or a logic "0" state depending on whether the internal node is 0 or 1. The 6T cell has a similar write cycle.

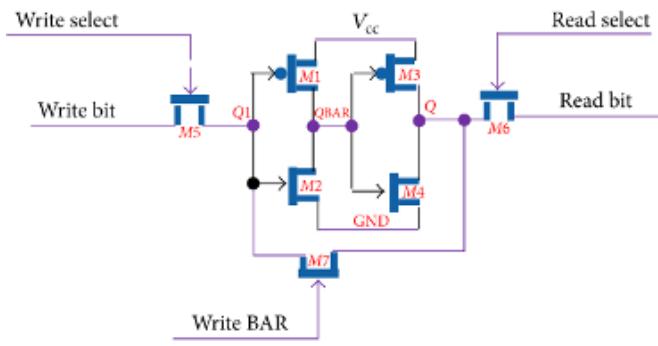


Fig. 3. 8T SRAM Bitcell

C. 1.3 Comparison of Existing SRAM Architectures

8T SRAM structures have read SNM (RSNM) that is more than 50 percent higher than 6T SRAM structures, however 7T

SRAM structures have no significant improvement in RSNM. 7T and 8T SRAM structures exhibit roughly a 10 percent improvement over 6T SRAM structures. The dynamic power consumption of the 7T SRAM structure is 36 percent less than the 6T structure. The dynamic power of the 8T structures is boosted by 36 percent over the 6T structure. Similarly, the read latency for the 7T structure is the shortest compared to the 6T and 8T structures. When compared to alternative designs, the 7T SRAM structure provides the lowest dynamic power usage as well as the shortest read delay. The reason for this necessitates a more detailed examination of its structure. Two cross-couple inverters, INV1 and INV2, are formed by the four transistors in the middle. Low input values in the first inverter INV1 generate high input values in the second inverter INV2, and low input values in the second inverter INV2 generate high input values in the first inverter INV1. This structure is shared by all the SRAM structures in this study. The benefit comes from the additional n-type transistor in the feedback link between the output of the INV2 and the input of the INV1, which performs feedback connection and disconnection during read and write operations, respectively. This enables a quick transfer of logic values from the write bit line "WRITE BIT" into the memory cell, where the WRITE SELECT turns on the transistor. During this cycle, the transistor is turned off. Because the switching activity during memory access is reduced, the dynamic power is reduced.

II. METHODOLOGY

In our work we have presented the implementation of In Memory computation in the SRAM bit cell. We have first studied the various existing SRAM bit cell architectures. The standard 6T and 8T architectures and a single ended low Vt 7T SRAM architecture. The read and write operations of each design were simulated. The design of the SRAM architectures was all done in 65nm technology node. The sizing was then done to get the least amount of disturbance in the operations of the Bit cells and the computation. The supply voltage for all the architectures is taken as 1.2 V. The following figure is of the 6T SRAM cell. The six transistors with the above arrangement form a 6T bit cell. It consists of two CMOS inverters connected back-to-back. This setup is used to latch up to the input during when the write is enabled. The inverter pairs use up 4 of the six transistors in the design. The other two transistors act as pass transistors. When the input in the WL is made high the pass transistors get switched on and hence [7] the input is directly connected to the Q. The moment the WL input is disabled the pass transistors disconnect the input from the inverter and the value gets stored inside the bit cell. The 8T SRAM cell is an extension of the 6T SRAM cell with the addition of two extra transistors. Unlike in the 6T SRAM in the 8T design the read and write operations are decoupled from each other. The extra transistors are solely for the purpose for a stable read operation. The following figure is of an 8T SRAM cell. This architecture has a RBL through

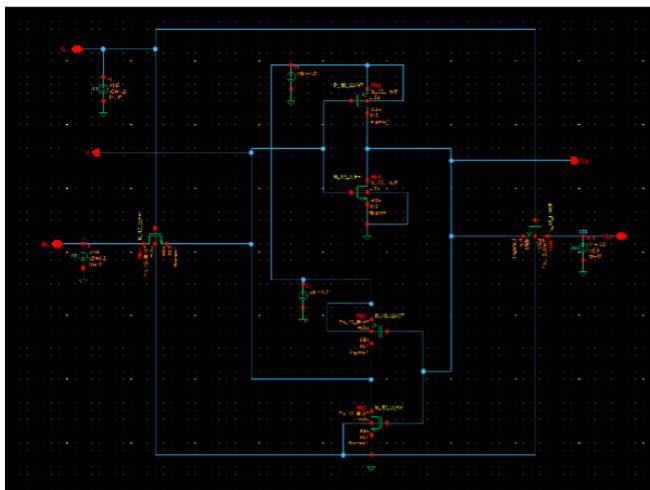


Fig. 4. 6T SRAM cell

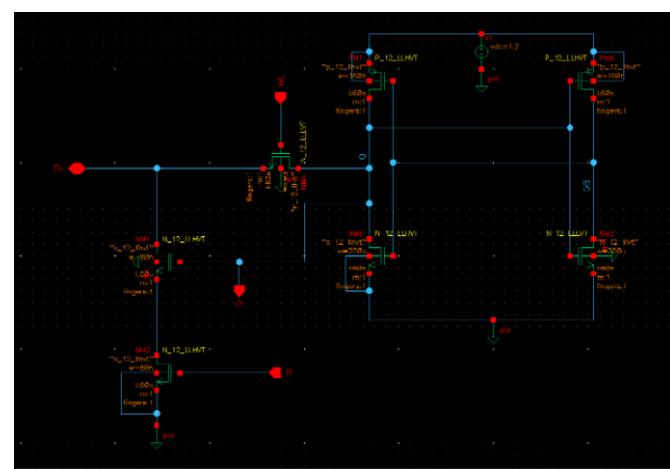


Fig. 6. Dual Vt 7T SRAM Cell

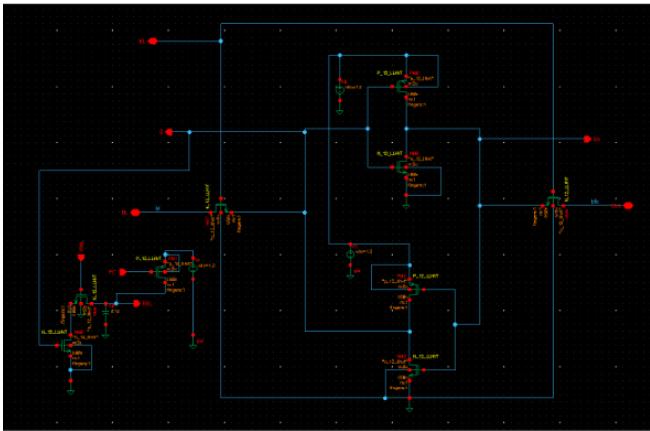


Fig. 5. 8T SRAM cell

which the output is taken. The write operation of this design is same as that of the standard 6T bit cell. For the read operation the part with six transistors gets completely decoupled. While reading operation is performed the RBL is pre-charged to the Vdd value. To start the reading operation the RWL is pulled up high. In the case, when the cell stores '0' the transistor connected to the Q output is turned off and thus there is no closed path formed till ground. Hence the capacitor at the RBL retains the voltage. Hence a 0 value is read by the 8T cell. In the case when a '1' is stored, the transistor turns on. It then gives the capacitor a closed path till ground to discharge its voltage. Hence the voltage at the RBL drops and a read action of data value '1' is performed. The Dual Vt 7T SRAM has low Vt transistors NM3 and NM4 from the figure so as to facilitate writing 1. It uses high Vt transistors to reduce leakage currents. The reading and writing operations are carried out on a single bit line.[8] Hence, it is also referred to as a single ended architecture. This reduces the power consumption of the overall bit cell. For writing in the cell, the bit line is pre charged to logic high of around 1.2 V or bit line is discharged

so as to get a logic low of around 0V. Then the WL is enabled so as to latch the input to the cell. For reading from the cell, the R line is enabled. If the bit cell stores a 1, the bit line gets discharged indicating the presence of '1' stored. In case of a logic '0' stored the bit line retains the charge indicating a value of logic low is stored.

A. Read and Write waveforms

In this section we discuss about the read and write operations of the SRAM architectures. We also observe the output waveforms for the same. The final computation of NAND output is also presented. The figure represents the write '1' operation of the 6T SRAM cell. From the figure it is evident

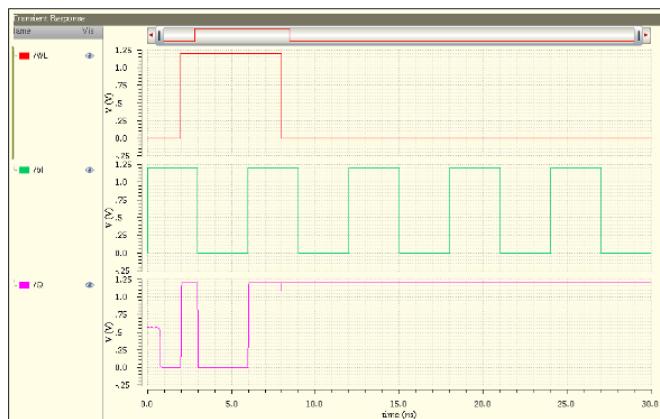


Fig. 7. Write 1 operation in the 6T SRAM Bit cell

that the WL when enabled passes on the input to the inverter pair connected back-to-back. The input is then latched into the cell. It then gets stored into the cell once the WL is disabled by passing a low voltage to the pass transistors. For the 8T SRAM cell the following waveform demonstrates the case of read '0' and write '0' operation. The write operation is very similar to that of the 6T cell. The WL is made high to latch up the input which eventually stores the data in the cell. For the read

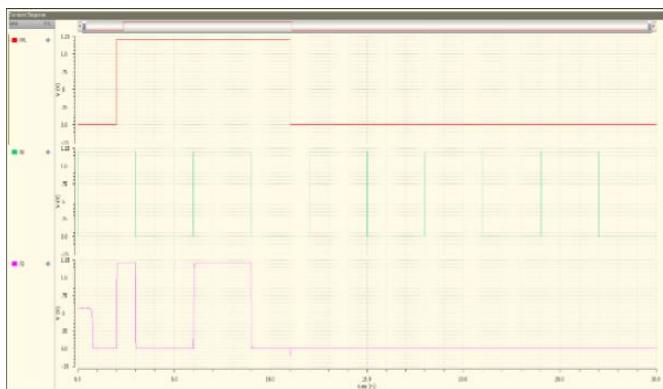


Fig. 8. Write 0 operation in the 6T SRAM cell

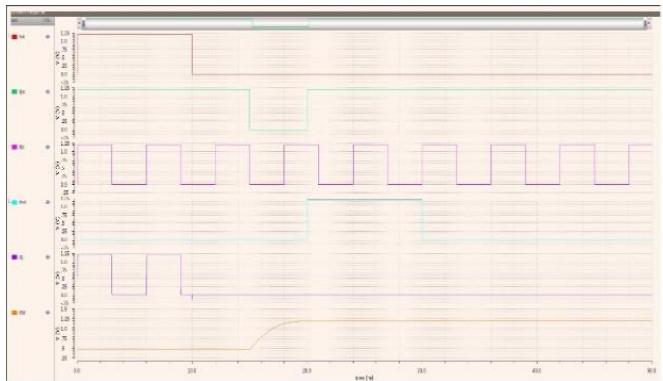


Fig. 9. Write 0 and RBL voltage value during Read 0 operation of the 8T SRAM cell

operation the RBL is pre-charged as shown in the waveform. As the value stored in the memory is '0' the capacitor attached to the RBL does not discharge and hence the value of zero will be read the sense amplifier. It can be observed that during

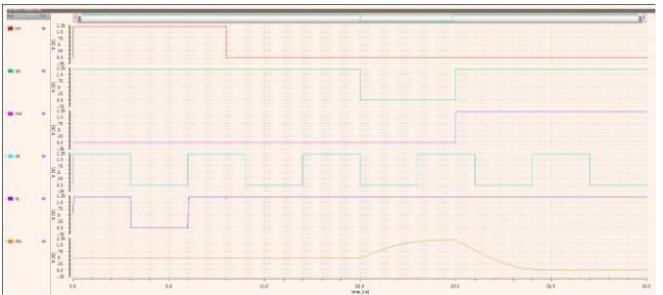


Fig. 10. Write 1 operation and voltage at RBL on Read 1 operation.

a read 1 operation there is a discharge of the capacitor at the RBL and hence the voltage at the RBL line drops as soon as the read operation begins. The 7T SRAM cell also works in a similar fashion. At the start the cell shows some disturbance with an existing voltage at the Q output. This initial value is ignored and the actual functioning starts from the first write pulse. In the 7T design the input and the output both are being taken from the BL. Thus, at the first instance when WL is

pulled up high, at this moment the voltage at the bit line is low (nearly 0), hence a 0 is being stored in the cell. Now the bit line is charged to Vdd and the read line is activated. As the stored value is 0, there is no discharge of the bit line capacitor and hence a 0 is sensed. For the next write cycle, the BL is already pre charged to the Vdd value. Thus, when the WL is enabled, the value of 1 gets stored into the cell. So far, we have seen

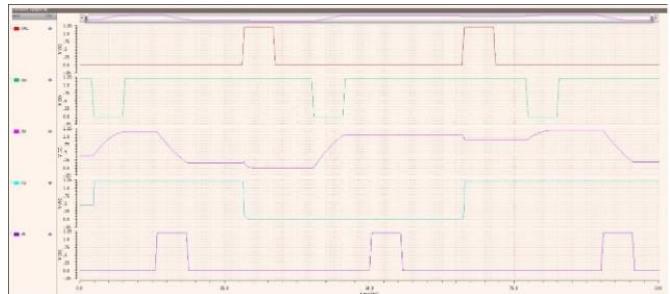


Fig. 11. Write 0 and write 1 operation in the 7T SRAM and the voltage at the bit line during the read operation

how the voltages at the RBL line in case of 8T cell and the BL for the 7T cell varies during the read operations. This fact is then put into action for the in-memory computation. We have focused on performing the NAND operation by using two-bit cells storing the two operands. For the purpose of computation, we have taken bit cells and have connected their bit lines. The figure shows the arrangement of the two cells used for the computation. From the above figure it can be seen that the Voltage from the RBL/BL is connected to the positive terminal of the sense amplifier and the reference voltage is connected to the negative terminal. The choice of the voltage value of the reference terminal is an important step. This determines the success of the computation. As the RBL is common to both the bit cells we get three different voltages for the cases when both the cells store 1 (V11), when both the bit cells store 0 (V00) and when one of them stores a 1 and the other stores a 0 (V10).

The reference voltage is chosen such that it lies in between the V00 and V10. The sense amplifier receives both the reference voltage and the VRBL as the inputs. The output of the sense amplifier is taken from the negative output terminal when the SAE is made high. When the reference voltage is lower than the voltage at the bit line, the output of the sense amplifier is high. This occurs in the case of V00. The value of the reference voltage is set in such a manner that for all the other combinations other than the case of V11, the reference voltage will always be lower than bit line voltage. Thus, only in case of V11 the output given by the sense amplifier is a low value or 0.

III. RESULTS AND DISCUSSION

In the figure the simulation waveforms of the NAND function for the 10/01 computation are demonstrated. One of the cells stores a 1 and the other stores a 0. Initially the bit line is pre charged to a high value (Vdd). When the RWL

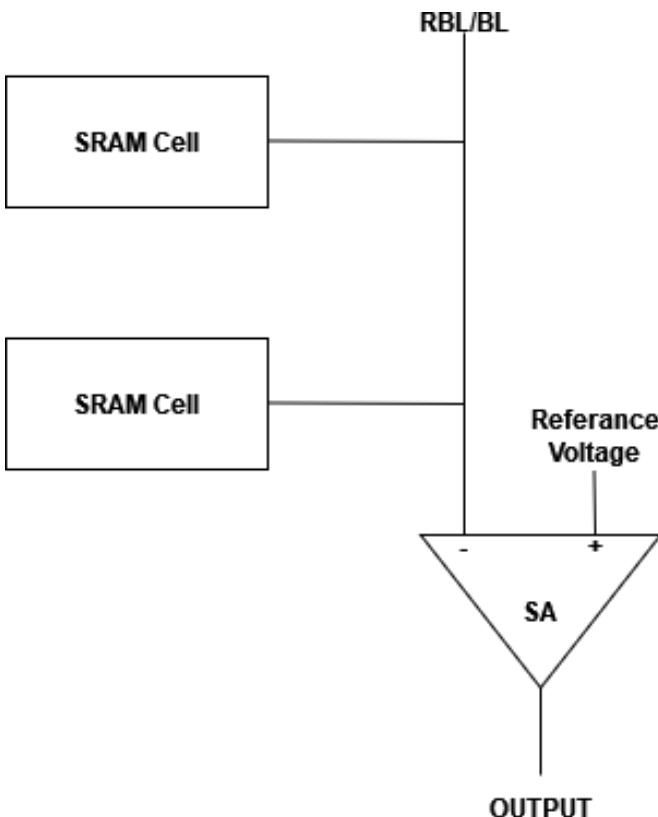


Fig. 12. Arrangement of the Bit cells for the In Memory Computation

TABLE I
 COMPARISON OF READ AND WRITE DELAYS OF 8T AND DUAL VT 7T SRAM CELLS

SRAM Type	Write '1'	Read '1'	Write '0'	Read '0'
8T SRAM	75.86 ps	371.5 ps	~61.4 ps	~0 s
Dual-Vt 7T SRAM	12.9 ps	320 ps	~130.9 ps	~0 s



Fig. 13. Computation of NAND for 10/01 combination with 8T cells

enabled, there is a drop of voltage. At the end of the read, when the SAE is made high the sense amplifier is enabled, and the output is high (logic 1). This is as the expected result as the NAND operation of 1 and 0 combination. This figure



Fig. 14. Computation of NAND for the 11 combinations with 8T cells

shows the computation of NAND operation when both cells store a logic 1. In this case when the sense amplifier is enabled through SAE, the output from the negative output terminal we get is a logic low (around 0V).

A. Logic Verification



Fig. 15. Computation of the NAND for 00 input combination with 8T cells

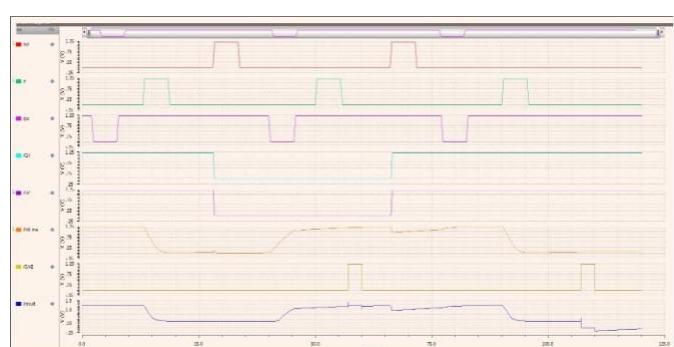


Fig. 16. Computation of NAND with 00 and 11 input combinations using 7T SRAM cells

From the figure it can be observed that when both the inputs that is the Q1 and Q2 are low and when the sense amplifier is enabled the output, we get is a logic high (1.2 V). Similarly, when the input combination is such that both of them are high the output we receive is as expected, logic low.

IV. CONCLUSION

In our work we have demonstrated performing of NAND operation by using two SRAM cells. The architectures we have used for the design of the cell were 8T and dual-Vt 7T architecture. All the architectures were designed using 65nm Technology node. For both the 8T and 7T cells the outcome of the NAND operation is shown. The delay for the dual Vt 7T and the 8T SRAM is also evaluated. A comparison of both of these design patterns is also presented. The 8T design successfully computes the NAND result for all the three different 2-input combinations of 0,0 1,1 and 1,0. For the 7T structure, we could carry out the NAND of the combinations 0,0 and 1,1 successfully. The computation of 1,0 caused read disturbances in the bit line. Further research is required to improve the In Memory computing dual Vt 7T SRAM design. This technology shows huge potential in the fields of AI and ML. This work can be carried further to perform total arithmetic operations inside the memory using the dual-Vt 7T SRAM bit cell as this architecture is efficient than most of the existing architectures in terms of speed and power.

REFERENCES

- [1] S. Joshi and U. Alabawi, "Comparative Analysis of 6T, 7T, 8T, 9T, and 10T Realistic CNTFET Based SRAM," *Journal of Nanotechnology*, vol. 2017, Article ID 4575013, 9 pages, 2017.
- [2] Z. Jiang, S. Yin, J. Seo and M. Seok, "C3SRAM: An In-Memory Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, July 2020.
- [3] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Power Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.
- [4] Q. Dong et al., "A 4+2T SRAM for Searching and In-Memory Computing With 0.3-V VDDmin," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 1006–1015, April 2018.
- [5] W. Simon, J. Galicia, A. Levisse, M. Zapater and D. Atienza, "A Fast, Reliable and Wide-Voltage-Range In-Memory Computing Architecture," in Proc. 56th ACM/IEEE Design Automation Conference (DAC), 2019, pp. 1–6.
- [6] N. Verma et al., "In-Memory Computing: Advances and Prospects," *IEEE Solid-State Circuits Magazine*, vol. 11, no. 3, pp. 43–55, Summer 2019.
- [7] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.
- [8] K. Absel, L. Manuel and R. K. Kavitha, "Low-Power Dual Dynamic Node Pulsed Hybrid Flip-Flop Featuring Efficient Embedded Logic," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 9, pp. 1693–1704, Sept. 2013.
- [9] D. G. Elliott, M. Stumm, W. M. Snelgrove, C. Cojocaru, and R. McKenzie, "Computational RAM: Implementing processors in memory," *IEEE Design & Test of Computers*, vol. 16, no. 1, pp. 32–41, 1999.
- [10] S. A. Tawfik and V. Kursun, "Low power and robust 7T dual-Vt SRAM circuit," in *IEEE International Symposium on Circuits and Systems*, pp. 1452–1455, Seattle, WA, USA, 2008.
- [11] M. H. Tu, J. Y. Lin, M. C. Tsai, S. J. Jou, and C. T. Chuang, "Single-Ended Sub-threshold SRAM with Asymmetrical Write/Read-Assist," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 12, pp. 3039–3047, Dec. 2010.
- [12] R. E. Aly and M. A. Bayoumi, "Low-Power Cache Design Using 7T SRAM Cell," *IEEE Trans. Circuits Syst. II*, vol. 54, no. 4, pp. 318–322, Apr. 2007.
- [13] A. K. Rajput and M. Pattanaik, "Implementation of Boolean and Arithmetic Functions with 8T SRAM Cell for In-Memory Computation," in *Proc. Int. Conf. for Emerging Technology (INCET)*, 2020, pp. 1–5.
- [14] B. Chen, F. Cai, J. Zhou, W. Ma, P. Sheridan, and W. D. Lu, "Efficient In-Memory Computing Architecture Based on Crossbar Arrays," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2015, pp. 17.5.1–17.5.4.