F-RAM: Simultaneous computing 512 bit SRAM Architecture with a flash data rate of 4.8 GHz

Debasmita Banerjee
Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90007
dbanerje@usc.edu

Ravi teja Lakkireddy
Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90007
lakkired@usc.edu

Abstract

In the current era of high speed devices, it is crucial to develop hardware solutions that supports the requirement for high speed, low power and dense manufacturing of on-chip integrated circuits. A great number of advances have been made that reduces the power consumption of the overall circuit while maintaining the necessary data rate, especially in memory design. In this article we have applied a low power technique of connecting a virtual ground potential to the access transistors of SRAM and also proposed a dynamic decoder circuit in conjunction with body biasing for the sense amplifier circuit that allows a modest leakage and dynamic power consumption with high performance output. The novelty of the circuit lies in devising an in-memory computing method that harbors both the conventional computing and in memory computing operation at the same time resulting in further increment in the data rate to multi-fold. The circuit also uses a dual port 10T SRAM architecture for added boost in the functionality. This article explores different SRAM architecture that employs pragmatic low power techniques to achieve a modest power consumption but very high data frequency.

Keywords 10T SRAM · In memory computing · Dual port SRAM · Dynamic decoder · High speed SRAM

1 Introduction

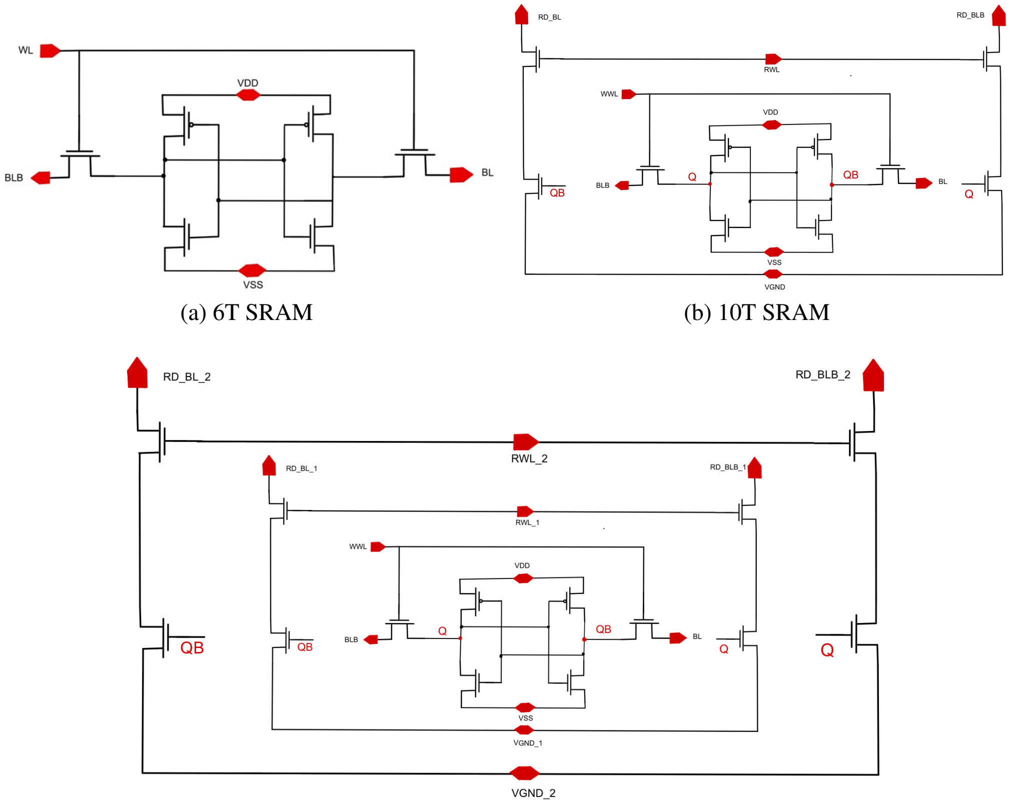
Device scaling, a need based approach by academia and industry, has produced significant number of ultra small channel devices with increased functionality per cost. This further helps to meet the common requirement of high device density and performance but often results into higher leakage, making the devices power hungry. On the surface level the effect of high power dissipation is well identified in consumer electronics products through heated devices and short battery span. This puts forward the need for energy efficient devices which is now following the trend for optimization on the algorithmic level but is largely hardware limited. The situation intensifies if the area limitation of this high speed devices are considered. Plenitude of current applications such as high speed computing, communication and sensing devices, biomedical detection devices, space technologies require the presence of innovative memory solutions. Therefore a significant amount of on-chip area is dedicated to memory design. Owing to the system performance, SRAM in modern design consumes crucial amount of area that otherwise helps in its parallelism [1]. However shrinking technology size, demands optimized SRAM which should also counteract the growing leakage power and delivers high performance output. The ideality of the goal although convoluted has been severely researched and reduction in supply voltage is touted as a power optimization mechanism [2] at the cost of performance. Furthermore different SRAM architectures such as 7T, 8T, 9T, 10T are used consecutively or as stand alone memory architecture with conventional 6T ones in search of better performance and stability. Other techniques to find a superior power-performance-area solution involve the mindful use of logical effort to reduce transistor counts, usage of multi-threshold cmos, body-biasing technique, sub-threshold operation that work on static and standby power to provide an overall improvement of power dissipation. In this article we have used a virtual ground to control the leakage power and an optimized 3 i/p dynamic NAND and 2 i/p static NOR [3] to receive a refined dynamic power performance of a 10T dual port SRAM architecture. The novelty of the design is in its use of in-memory computing technology which refers to the computation of logic inside the memory itself instead of an external unit. Additionally, the proposed design also allows simultaneous extraction of data from both the in-memory computation and traditional computation which further enhances the performance. This project also shows slightly unconventional design of the sense amplifier and optimized layout for both the SRAM and sense amplifier. The schematic simulation and layout is performed on Cadence Virtuoso Analog Design Environment for the 45nm node. The layout designs are further post processed using LayoutEditor.

2 10T SRAM with dual port technology architecture and stability analysis

This section includes the details for 10T SRAM with dual port architecture and its functionality. Also, Static Noise Margin (SNM) is plotted to determine the stability of the SRAM cell.The least noise voltage needed to change the cell state is SNM and one of the methods of calculating the Static Noise Margin (SNM) is by plotting the butterfly curve which is given below.

2.1 10T Dual port SRAM: choice of architecture

Static Random Access Memories (SRAMs) are volatile memory technology that are also “static” which means they need not to be refreshed but stores the data unless the power supply is removed. The inherent architecture is based on a cross coupled inverter pair where each side is connected to an access transistor which are enabled using wordlines through the respective gate terminals. Other nodes of the access transistors are connected to bitline and $\overline{\text{bitline}}$. Conventional 6T SRAMs have limited read and write noise margins (RSNM and WSNM). Therefore many different architecture such as 7T,8T,9T,10T were proposed with potentially improved read SNM and write SNM. Unlike 6T SRAM cell which uses same access path for the read and write operation , 10T SRAM [4] uses two different path for read and write operation keeping the data undisturbed. Our proposed architecture has 10 transistors in its unit cell that also exhibits dual port access to the data. This improves the data access speed by two folds by allowing access of two data from two different rows of the SRAM or the same SRAM, depending on the usage need of the memory bank.



(a) 6T SRAM

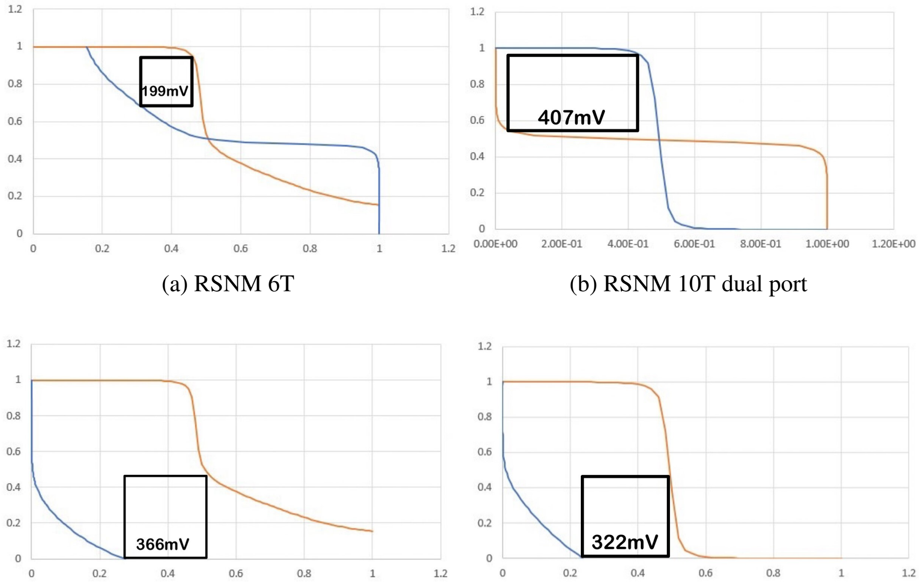
(b) 10T SRAM

(c) 10T dual port SRAM

Figure 1: Schematic of (a) 6T and (b) 10T and (c) 10T dual port SRAM

2.2 Stability analysis of 10T Dual port SRAM

One of the figure of merit and popular choice of evaluating the stability of memory cell is performed through SNM calculation in which the static noise margin is evaluated by encapsulating the largest possible square in the two voltage transfer curves (VTC) of the involved CMOS inverters. A formal definition of SNM of SRAM is said to be the minimum amount of noise voltage present on the storing nodes of SRAM required to flip the state of the unit cell. The write SNM is a specific measure to the writing ability of SRAM in which The minimum voltage required to feed new value into the SRAM cell is known as write margin whereas Write stability is the ability of the SRAM to allow the modification in the stored value. Similarly, the read margin is used to find out read stability of the SRAM which is defined as the ability to prevent the SRAM cell to flip the stored value while the stored value is being read. The simulation result shows the comparison of 6T and 10T dual port which exhibits an improved read SNM for 10T dual port SRAM.



(a) RSNM 6T

(b) RSNM 10T dual port

(c) WSNM 6T

(d) WSNM 10T dual port

Figure 2: SNM graphs for (a) RSNM 6T SRAM and (b) RSNM 10T Dual port SRAM and (c) WSNM 6T SRAM (D) WSNM 10T dual port SRAM

2.3 Testbench description:

Conventional Memory Access: Considering the Memory bandwidth bottleneck which hinders the program performance, the prime research focus has been shifted from reducing the latency in the processors to improving the memory speed inorder to reduce the gap between the processor's speed and memory. To write data into a particular location in the memory, write enable is given as logic high, the respective address is given to the memory block which enables the specific row of the memory using Write Word Line (WWL) and the data is given to the Write block which writes the data into the SRAM cell through bitline(BL) and $\overline{\text{bitline}}$ (BLB). To read the data which is already stored, Read enable (rd_en) is given as logical high. Then, according to the memory address from which data should be read is given to the memory block which enables the specific row by enabling Read Word line (RWL) of that row. This data from the memory is sent through sense amplifier and muxed using the column multiplexers.

Dual-port Memory Access: When compared to the conventional 6T SRAM, dual port 10T SRAM has excellent Read SNR and write SNR values, which enables us to use multiple access ports from the same SRAM cell. In our proposed design we have implemented a dual port SRAM by adding an extra pair of access transistors and BL and BLB rails. Which results in increasing the data rate by two times at the expense of a comparatively higher area. This 2 data accessed from the memory is given to two different sense-amplifiers to detect the data and then multiplexed using Column Multiplexer.

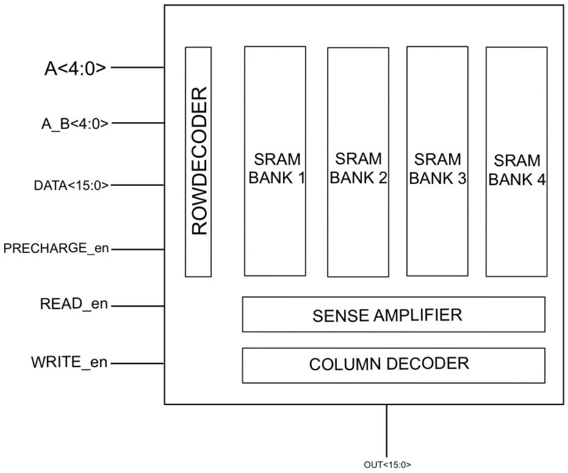


Figure 3: Block diagram of 512 bit memory architecture

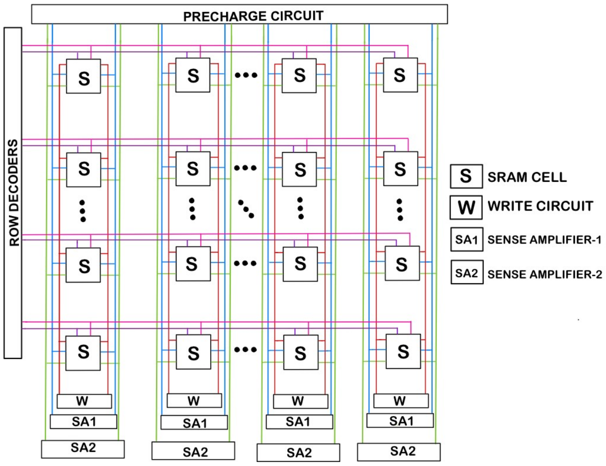


Figure 4: Memory bank architecture

In-Memory computing: In a conventional architecture, to perform an operation, data is accessed from the memory and sent to a processor (ALU) for computations. Then after the computation is done, the data result is again written back into the memory address. Whereas, In-Memory computing [5] is a method of computing the first set of computations inside the memory while accessing the data, which will save the data transfer time from memory to processor and vise-versa. In the proposed design, we have implemented this technology using skewed sense-amplifier instead of symmetric one which can give NAND and OR logical output by skewing the transistor sizes of left side and right side of the differential pair in sense amplifier.

3 Layout design of 10T dual port SRAM and Sense Amplifier

One of the crucial design strategy for on-chip memory architecture is the optimal layout creation with highest possible density of design. In the proposed 10T dual port SRAM design the cross couple inverter pair has been allocated the central region whereas Read Bitline(RBL), Read $\overline{\text{bitline}}$ (RBLB), BL, BLB, Read Bitline due to the dual port (RD_BL1), and Read $\overline{\text{bitline}}$ due to the dual port (RD_BLB1) are shifted to the edge of the layout for easy integration of multiple unit cell to create a larger memory bank specific to the goal of 512 bit memory architecture. In contrast to that, another equally important architecture, the sense amplifier possesses an unconventional design methodology to accommodate higher voltage connection to the body. The unit cell area of 10T dual port SRAM is 1.61um × 1.49um = 2.39 pm and the same for the sense amplifier is given by 1.195um × 2.035um = 2.431pm.

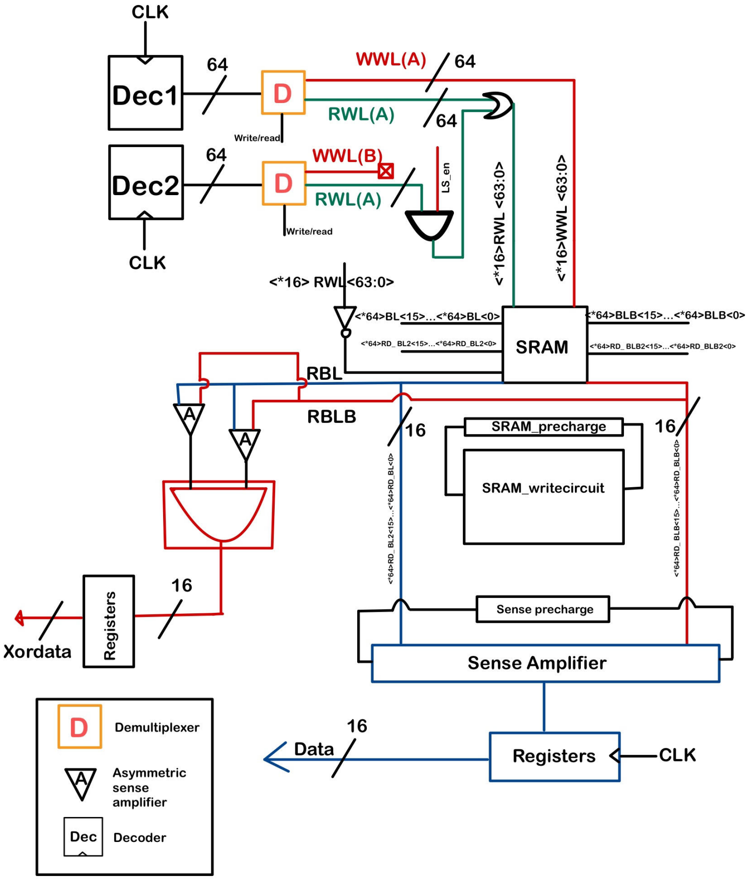


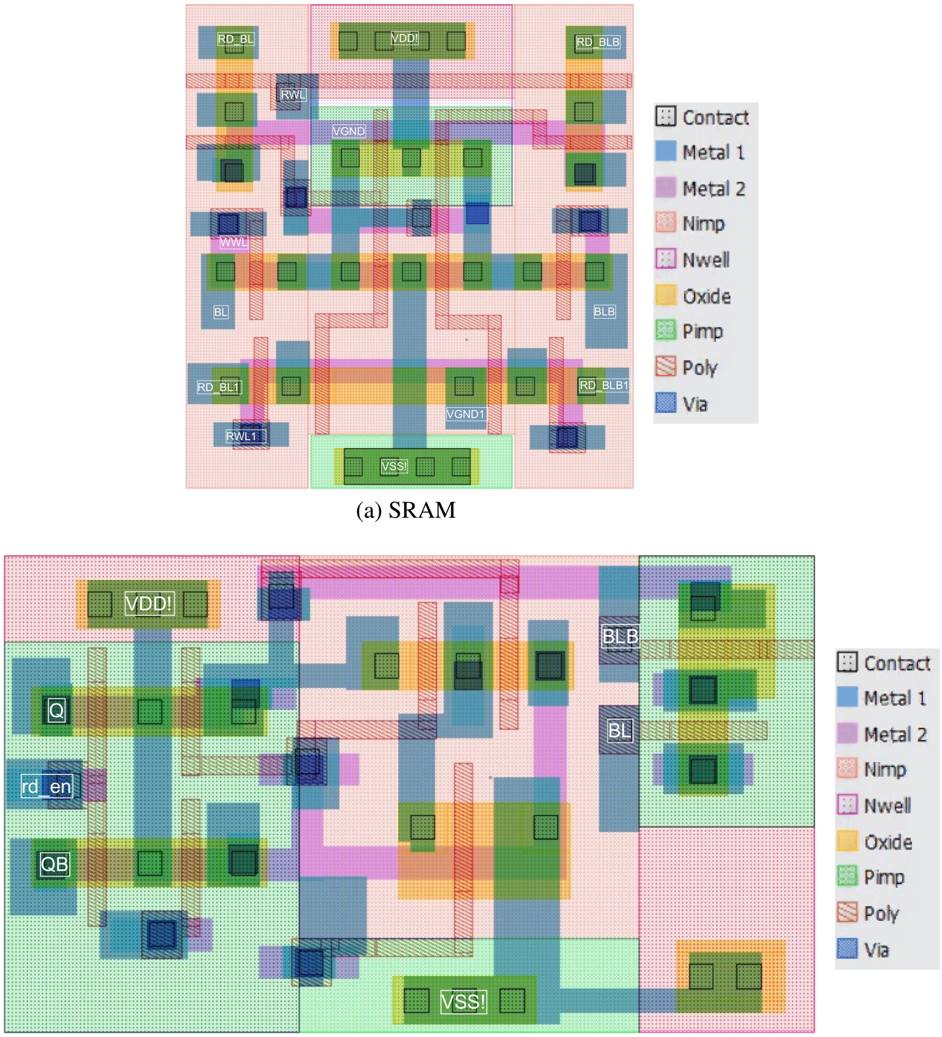
Figure 5: Full scale test bench description including in memory computing and low power technique

4 Results: Timing diagram, Write energy, Read energy, In memory computing energy, leakage power

The key highlight of the proposed design is that we can access two different word bits from the memory along with its logical computation result at the same clock edge. From the transient wave forms shown, it is noticeable that the data is written into the memory block up-to 6.25nS. Now, the two data addresses of interest is given to two row decoders to enable the respective rows in the memory block. Using the first decoder, First data (in Figure 6. from the dataout_1) word is accessed and sent through first set of sense amplifiers and after a certain delay, second row is also activated depending on the second decoder output for in-memory computing, then this combined data is sent through the asymmetric sense amplifiers to access the logical computed outputs (in the figure NAND<15:0>, OR<15:0>, xor_of _data1 _data2) (In this case its XOR logic). Simultaneously utilizing the second port of SRAM second data (dataout_2) is also accessed depending on the 2nd decoder output and sent through second set of sense amplifiers. Therefore the proposed design leverages the mentioned method to access two different data with the computed result of them in one clock edge. This amounts to a data rate speed of 4.8 GHz, three folds of the clock speed of 1.6 GHz used for the design. The processing is power heavy but with the help of low power techniques used we limited the write, read and in memory computation energy to the order of pico watts while the leakage power is found to be in the nanowatts range.

5 Conclusion

In this article, the main aim was to achieve a low power SRAM that offers high performance and displays considerably compact area. We have chosen 10T SRAM for its high read noise margin and tweaked it to achieve a dual port SRAM



(a) SRAM

(c) Sense amplifier

Figure 6: Layout of (a) SRAM and (b) sense amplifier with the legend

Table 1: Energy and power numbers for reading and writing 16 bits

Name	Value	Units
Write Energy	1.175×10^{-12}	J
Read Energy	1.156×10^-12	J
In memory computation energy	1.508×10^-12	J
Leakage power	249.44×10^-9	W

contributing to the performance of the device. The performance was further improved owing to the intuitive application of in memory computing where the data is processed within the memory itself. The outcome of such design was a data rate of 4.8 GHz in expense of a relatively low read and write energy as mentioned in Table 1. Although there are many room for improvements such as modifying the design for low leakage power, the simultaneous processing of in memory computation and traditional computation could be a potential contribution to the increasing need of data speed with constant clock speed which could be limited due to the parasitic factors of the design.

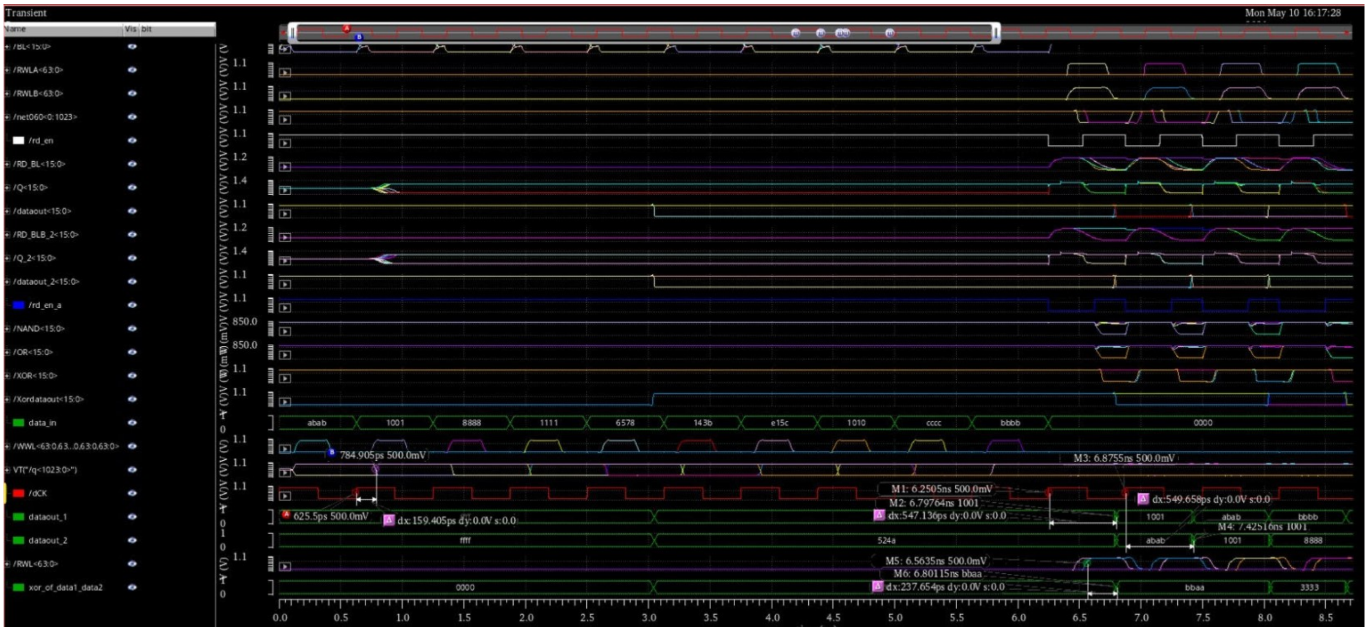


Figure 7: Timing diagram

Acknowledgments

This work is an extension of the project done as a requirement to the partial fulfilment of the course credits for EE577A: VLSI System Design at the University of Southern California, instructed by Prof. Akhilesh Jaiswal in Spring 2021. The authors would like to acknowledge the invaluable feedback received during the project timeline from Prof. Jaiswal and thank him for his utmost support.

References

[1] M. E. Sinangil and A. P. Chandrakasan, IEEE Journal of Solid-State Circuits **49**, 1, 107-117 (2014).

[2] H. Kumar, and V. K. Tomar, Wireless Pers Commun **117**, 1959-1984 (2021).

[3] M. Bennaser, Y. Guo and C. A. Moritz, IEEE Transactions on Very Large Scale Integration (VLSI) Systems **16**, 12, 1631-1638 (2008).

[4] Z. Lin et al., IEEE Journal of Solid-State Circuits **56**, 9, 631-1638 (2008).

[5] A. Agrawal, A. Jaiswal, C. Lee and K. Roy, IEEE Transactions on Circuits and Systems I: Regular Papers **65**, 12, 4219-4232 (2018).