International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Jan-2021)

http://dx.doi.org/10.12785/ijcds/100110

Design A Synchronous Single-Port Sram 1024x32xMUX4 Using 28NM Technology

Tran Nguyen Phu Phu[2], Dang Phuong Gia Han[2], Nguyen Cong Luong[2] and Nguyen Van Cuong[1]

[1] Faculty of Electronics and Telecommunications, Danang University of Technology (DUT), Danang, Vietnam
[2] Savarti Company Limited

Received 16 Mar. 2020, Revised 23 May 2020, Accepted 31 Jul 2020, Published 1 Jan. 2021

Abstract: Static Random-Access Memory (SRAM) has become phenomenally crucial to a board spectrum of VLSI designs and applications, ranging from high-performance CPUs to low-power mobile hand-held devices. As technology scaling helps to drive density and performance, it also poses some technical challenges in designing. This paper presents the pre-layout design for a 32kbit 6T synchronous single-port SRAM using 4-bit column multiplexer method and 28nm technology. Several sessions of this paper list out some methods such as self-control the timing constraints of the design, optimize shape and size to reduce difference in delay paths, thus improve performance.

Keywords: SRAM, VLSI, 32kbit, 6T, Single-Port, Synchronous, Column Multiplexer.

1. Introduction

In integrated circuit design flow, pre-layout verification takes place as a step to generate a gate-level netlist, using standard cell library to consider design constraints such as timing, area, function and power. However, as this is pre-layout level, we assume that all connection wires are ideal, which means that we can neglected the unwanted impacts of parasitic factors that harm the chip performance and power in reality. [1]

The single-port SRAM that we are working on has the size of 1024 wordlines, each word contains 32 bits of data. The type of SRAM is 6T, which means that each bitcell contains 6 transistors, including two access NMOSs and two cross-coupled inverters. Besides, 28nm technology is applied since it is conventional enough to feature high performance, low power and supports a wide range of applications that slightly catches up with market trend. The support tools for designing and simulation are Synopsys, HSPICE, Cadence and Custom Sim.

This paper proposed 3 methods that we applied to achieve the results: folding (column multiplexer – Mux4), pre-decoding and tracking technique. The remaining part is organized as follow. Section 2 describes 3 proposed methods, Section 3 explains methods to measure some important parameters, while section 4 illustrates how our SRAM functions based on waveform analysis and comparative table of parameters.

2. Proposed Method

The design of SYCHRONOUS SINGLE-PORT SRAM 1024x32mux4 is a combination of 4 main blocks including Memory, I/O, RWROW and RWCTL. Memory is supposed to be a shortage of data. Before input data comes to memory array, they must be latched at I/O block to keep them stable during operation. RWCTL

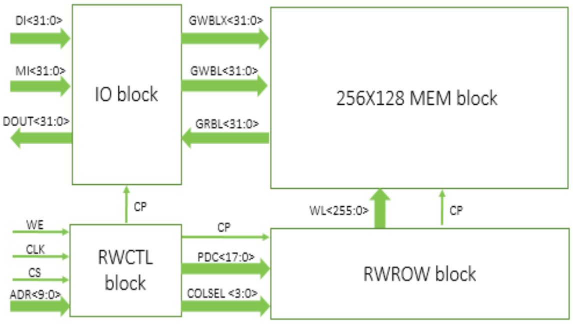

Figure 1. The top structure of the SRAM design.

Email: phuphunk@gmail.com, giahandn1602@gmail.com, nvcuong@dut.udn.vn, luongnguyen@savarti.com

block is responsible for generating internal signals CP (clock pulse) as well as pre-decoding task, while RWROW is a set of 256 wordlines. [2]

A. Folding – Divide into halves:

As the configuration defines, the circuit should be designed to have 1024 wordlines, each word contains 32 bits of data. The fact that this unbalanced rectangular strucutre has a large gap between horizontal and vertical views leads to the proposal of "Folding" method. Specifically, memory array is folded by a factor of 4 to optimize the size ratio to 256 wordlines – 128 bitcells per word. However, a little gap still remains, that is why the array is kept on dividing into halves. [3]

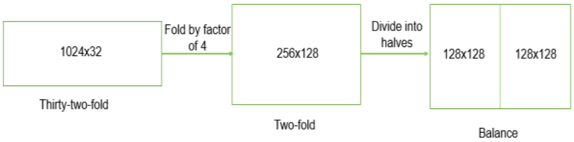

Figure 2. Folding and divide into halve.

Obviously, third structure has identified the travelling distance of data and wordline, as well as reduces the recovery time of wordline. In the previous design, data must go horizontally through 256 rows in the worst case, for reaching sense amplifier. Meanwhile, update design allows circuit to reduce the former distance by half. This change helps to reduce a duration of one cycle time, thus greatly improves access time, propagation delay and higher frequency. Besides, modifying from unbalanced rectangle to nearly square, which is supposed as a common shape of SRAM, will be more appropriate for the task of integration on real chip.[7]

B. Pre – decoding

Because of re-arrangement, the extraction of 10 bits address is not only for wordlines as standard designs but also for column selection (COLSEL). In detail, it would take the first two bits (ADR<1:0>) for COLSEL and the rest 8 bits (ADR<9:2>) to decode wordlines. Below are our proposed algorithm to select 1 among 256 wordlines. This method is called "Pre-decoding", in which we used 4 simple decoders to process 8 bits ADR. [2]

Since we separated the memory and decoder into 2 parts, a logical approach is first using 1 ADR bit to justify which half to be accessed. Afterward, a chosen part of 128 wordlines is extracted into 4 parts that have size of 32, at this point, 2 bits are decoded for selection. This rule of division is also applied for the following level. At the lowest level, 3 bits are assigned to pick up one among 8 parts. Initially, without pre-decoding, there would be 256 connection wires coming from RWCTL to RWROW block, but after adding, the figure is reduced phenomenally to 22 lines. Hence, the fundamental purpose of pre-decoding method is to minimize the total number of metal layers, thus brings benefits to physical design stage. [6].

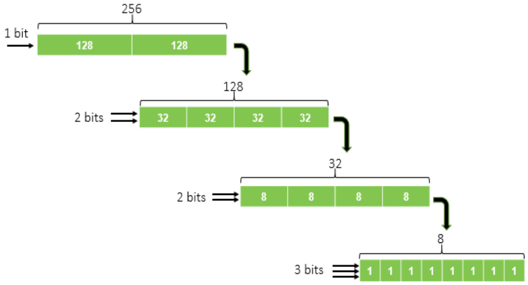

Figure 3. The algorithm of pre-decoding.

C. Tracking:

Nanoscale SRAM design must ensure robust operations against process variation, temperature degradation and difference in supply voltage over the usage lifetime [9], [10]. This fact rises the need for a well-track timing control circuit in order to ensure a compatible timing to activate/deactivate the synchronous clock.

By optimizing corresponding operation duration over difference conditions, it will improve the access time, reduce leakage power due to redundant interval and avoid unwanted incorrect function.

The idea starts from generating an Internal Clock Pulse (CP) from External Synchronous Clock (CLK) and use "Tracking" method to measure suitable pulse-width of CP. To solve this method, we have to answer one by one all its surrounding concerns: how to access SRAM? which signals are decisive in Read/Write operation? when can we turn off CP?
SRAM design composes of logic control circuits and bitcells. Logic control circuits include Decoder

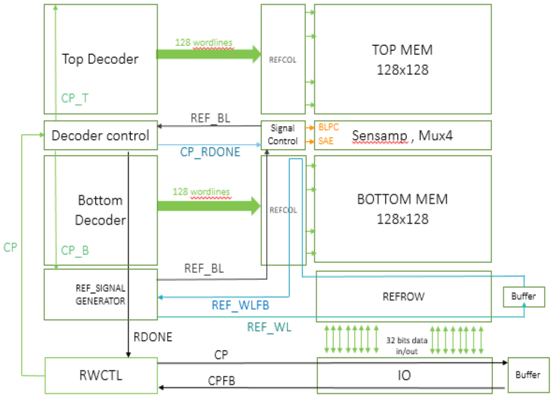

Figure 4. The signal flow of tracking

(RWROW), Input/Output control (IO) and Read/Write timing control (RWCTL) blocks. CP is generated from RWCTL and is sent out to drive the function of SRAM by two access ways: IO path and RWROW path. [7]

In Write operation, the turn-on duration of Wordline plays a key role. Wordline's high pulse-width must not only adapt with the worst case of CP's travelling distance to turn on the furthest Wordlines from RWCTL (in this case is Word 0 and Word 255) but also must be wide enough to cover the required duration for a bitline to travel from furthest bitcell to column multiplexer. To trace the motion of real wordline and real bitline, we will need the support from replica circuits: REFROW and REFCOL that emulate real structure of a row and a column in Memory Block (but do not affect SRAM real operation). These blocks create paths for referenced signals (REF_WL, REF_WLFB and REF_BL) to travel simultaneously with real signals.

In Read operation, however, the time to activate/deactivate Sense Amplifier Enable signal (SAE) is greatly important. This is where we need support from a signal called CP_RDONE, generated when data from bitcell has just reached Sense Amplifer. CP_RDONE's high level will help to stop pre-charging bitlines and turn on SAE by its falling edge. When data is sensed to full rail-to-rail, self-tracking circuit will generate a signal to turn off Sense Amplifier.

CP is turned off when self-tracking circuit ensures that all latches inside IO block have been activated and a duration turning on Wordline has just finished. Two feedback signals: Clock Pulse Feedback (CPFB) and Read Done (RDONE) from two access paths will then be sent back to RWCTL to pull down CP.

3. Evaluation Methods

A. Power Consumption:

Power dissipation has dominant effect on reflecting the quality of SRAM compared to the other parameters. The total power consumption is distributed by two primary components: dynamic power and static power. This paper mainly points to the former which happens during the circuit switching.

$$\mathbf{P_{Dynamic} = P_{Short\ circuit} + P_{Switching}}$$

Dynamic current, which is defined as the flow of electric charges through a conductor or the connecting wires forming the electric current, is a key factor to approach the value of switching power. On simulation test, the value of I_D is calculated by taking an integration of current at the voltage source V_{DD} node during an interval at which the status of circuit is flipped. Then, the figure for $P_{Dynamic}$ will be released following the formula below:

Formula:

$$\boldsymbol{P_{Dynamic} = f.V_{DD}.\int_{t1}^{t2} I_D(t)dt}$$

Where, V_{DD}: applied voltage,

I_D: current measured at V_{DD} node,

f: operating frequency,

$t1$, $t2$: the duration time during the SRAM switching. [5]

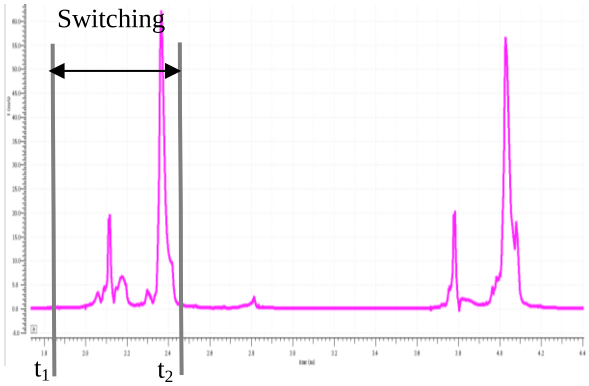

Figure 5. The waveform of I_D during circuit switching

B. Leakage Current:

When it leans toward the adaptation of scaling down speed, a noticeable problem of leakage current is interested by designers. Normally, the circuit operates on three modes including active, sleep and standby. The leakage current - an electric current in an unwanted conductive path under normal operating conditions, is dominated in standby mode at when SRAM is jumping into idle status. To get the value of I_{LEAK}, the immolation of standby mode must be set up in testbench by forcing the circuit to deactivate after finishing read and write cycles. Leakage current is measured at the point when its value becomes stable.

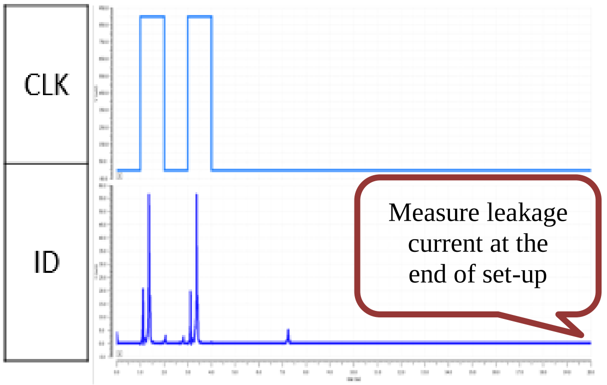

Figure 6. The waveform of measuring $I_{Leakage}$

C. Cycle Time:

Cycle time refers to the minimum period between two successive requests. This paper proposed three methods to measure cycle time, final result is the maximum value among three methods.

a) Method 1: Based on Access time

Access time refers to how quickly the memory can respond to a read or write request. Read access time is measured in a Read cycle, from when CLK turns on half way to half-way rising of data at output. In the other hand, Write access time is measured from Clk is turned on to when new data is overwritten inside bitcells.

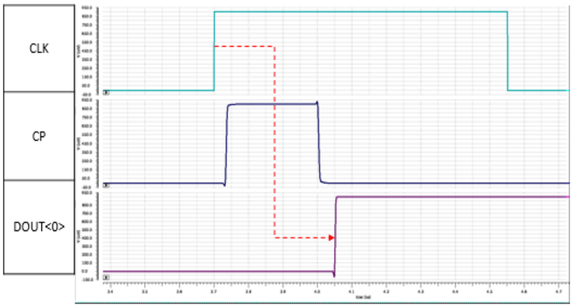

Figure 7. The waveform of measuring read access time.

b) Method 2: Based on Setup time & Hold time

Notice that a cycle time consists of hold time of current request and setup time of next request. The interval between these two must satisfy both the minimum range for an operation to happen and the minimum range that input data can change without affecting the function of current operation. This interval is proved to be equal to high pulse width of CP.

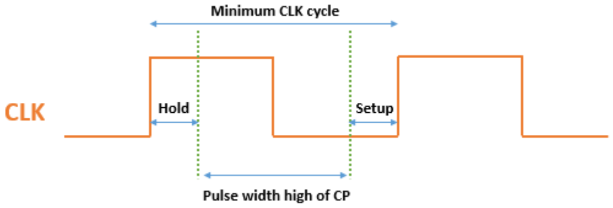

Figure 8. The waveform of calculating cycle time basing on setup and hold time

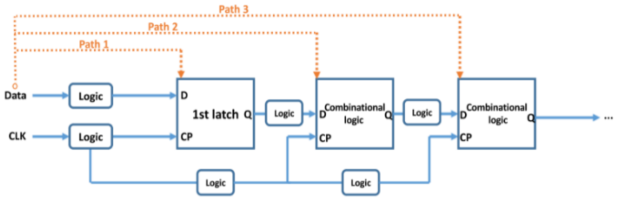

Figure 9. Delay paths for data-CLK.

We obtain the equation to calculate cycle time as follows:

Cycle time = Setup time + Hold time + Pulse width high of

Setup/Hold time is a timing constraint between input signals and external clock. Each input when coming inside circuit will be driven by external clock, through first latch to multiple kinds of combinational logic until it becomes independent of clock itself or clock's inheritors. This divides data-clock journey into different paths. Each path will present different setup time and hold time.

Our proposed method to measure setup/hold time is:

- Measure all timing constraints at all path that output is a combination of data and clock.
- Take the maximum value of Setup time & Hold time of each data.
- Take the maximum value of the sum Setup time + Hold time among all data.

❖ Setup time:

  - Definition: Setup time is an interval before clock rises that data must be stable.
  - Equation:

Setup time = Max data delay – Min CLK delay

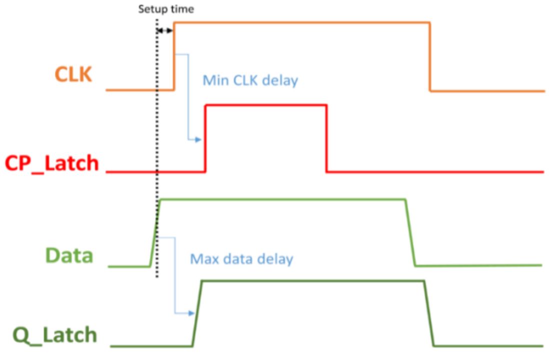

Figure 10. The waveform of Setup time

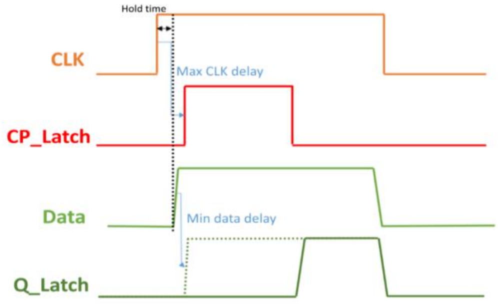

Figure 11. The waveform of Hold time

Hold time

- Definition: Hold time is an interval after clock rises that data has to remain stable.
- Equation:

Hold time = Max CLK delay – Min data delay

4. ANALYSIS

A. *Write operation:*

To write data into a memory cell, the cell must be selected using its row and column coordinates, the data to be stored must be applied at the data input pins, and the information must be stored in the selected memory cell. In terms of timing, the following steps must occur:

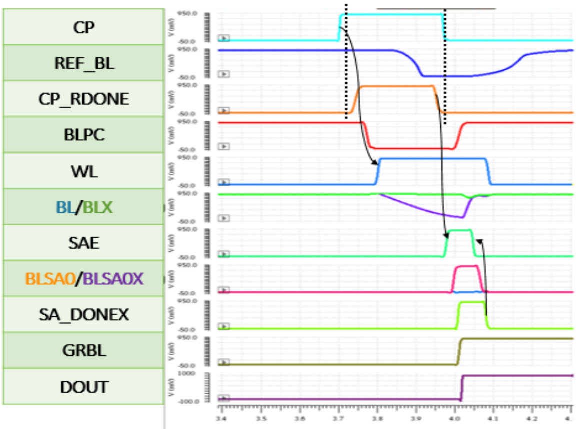

Figure 12. The waveform of write operation.

- Before the clock transition (low to high) that initiates the write operation, the row and column addresses must be applied to the address input pins, the chip must be selected, the Write Enable must be high, data to be written must be applied to the data input pins, and Byte Write Enable (BWE) must be low as well.
- On the rising edge of CLK, CP is pulled up and help to activate one of 256 wordlines thanks to the decoding part. At the meantime, data is driven into Global Write Bitline (GWBL) and Global Write Bitline Bar (GWBLX) and travel to Column Mux to select 32 columns based on one of 4 bits column selection.
- Tracking circuit generates referenced signals, creating CP_RDONE to pull down BLPC. When Pre-charge is off, data is overwritten into bitcell. After that, when CP turns off, a Write Cycle finishes.

B. *Read operation:*

To read data from a memory cell, all external input values must also been applied to the Logic Control Circuit, similar to Write Operation. In terms of timing, the following steps must occur:

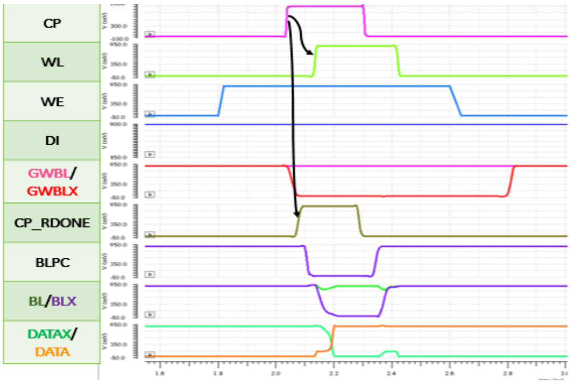

Figure 13. The waveform of read operation.

- On the rising edge of CLK, CP is turned on, one Wordline is activated.
- When Wordline turns on, data from bitcells is overwritten to its corresponding bitlines and travel to Column Multiplexer to select only 32 among 128 bits of data to come to Sense Amplifier.
- Tracking circuit helps to turn off pre-charging and activate Sense Amplifier. When data is all sent to full rail-to-rail, self-tracking inside Sense Amplifier will deactivate itself. After that, CP turns off, a Read Cycle finishes.

C. *Results of parameters simulation over PVT variations:*

The term "PVT variation" refers to the fluctuation of Process, Voltage and Temperature in real fabrication. Simulation of circuit's parameters over PVT is carried out in order to survey how the variation has effects performance, stability and power of SRAM. [4]

a) Process:

When design corners vary, ranging from SS, TT and FF,

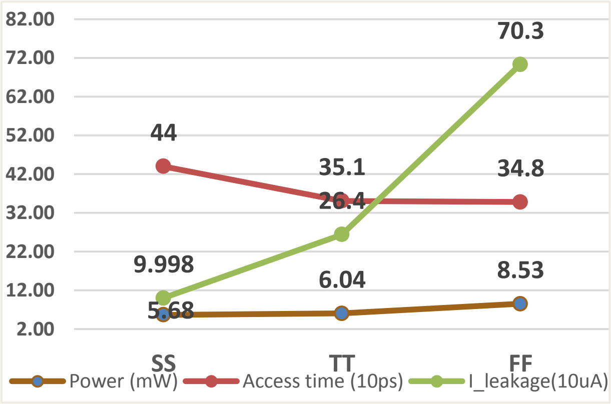

Figure 14. The chart of process variation

it will positively increase the mobility of electrons. As a result, the process of charging and discharging capacitors happens faster, which somehow leads to the decrease in access time. However, figures of dynamic power and leakage current climb up.

b) Voltage:

Let assume that noise voltage source is around 10%. Therefore, with the ideal voltage 0.9V, the simulation will be ran at three level of V_{DD} (0.81V, 0.9V and 0.99V).

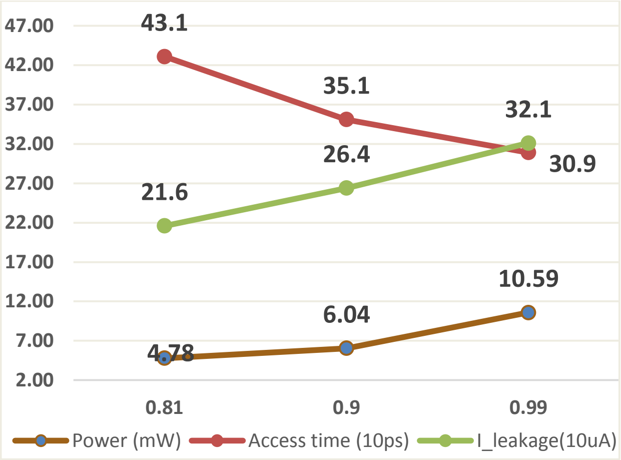

Figure 15. The chart of voltage variation

It is undoubted that higher supply voltage will be followed up by a growth of current. That is why the chart shows both P_D and I_{LEAK} increasing trend. Besides, access time decreases steadily.

c) Temperature:

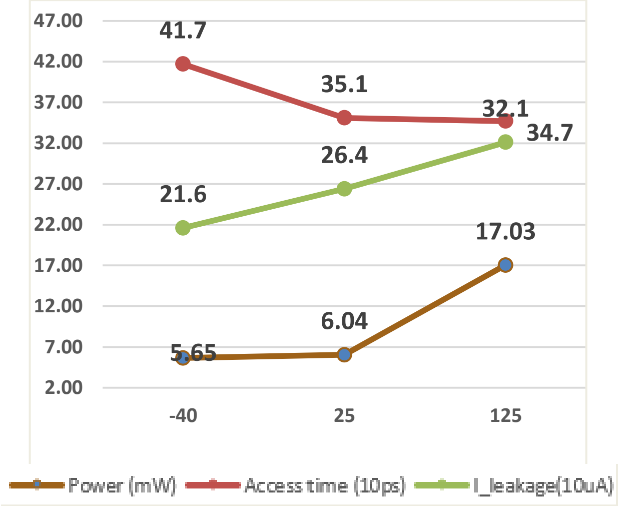

Figure 16. The chart of temperature variation

Recall to a theory that the movement of electrons significantly unpredictably fluctuates in higher temperature condition. We cannot exclude the case when electrons collide and prevent charges from moving toward a certain dimension. In this case, the value of current would drop. However, if the fluctuation makes electrons go faster with less collision, it now leads to the rise of current value. As can be seen from the waveform, the result is suitable with the second assumption in which the figures of dynamic power and leakage current go up.

D. Tables of comparisons

TABLE I.

	Reference [7]	This work
Cell/Technology	6T/28nm	6T/28nm
Capacity	256kbs	32kbs
Total Power	3.43mW	6.6mW
Maximum Frequency	735MHz	1.92 GHz

Above table compares our design with reference work in terms of power consumption and maximum frequency. Note that the reference design has a 8-time larger capacity.

All test chip simulation results are measured under typical corner, at 25 Celsius and 0.9V supply voltage. It is clear that our maximum frequency is better than the standard design, thanks to the self-tracking system that helps to reduce redundant amount of time for an actual operation. However, in terms of power, there is a trade-off since we receive help from multiple additional blocks that help to improve the performance but increase the load at the same time.

5. Conclusion

It is obvious that the integrated circuit (IC) field has witnessed some significant issues in terms of process design kits or topologies. However, they are all overcome based on the conventional structure of 6 transistors SRAM. That is why we choose 6T topology to research on the improvement of memory performance.

In this paper, we propose three new methods including folding, pre-decoding and tracking with the aim to optimize design's parameters. The folding method is assigned to enhance parameters related to the timing especially at higher frequency. Apart from dealing with the problem of extraction from 8 bits to 256 bits, the function of pre-decoding also helps to reduce the number of wires that tie RWCTL and RWROW together. The tracking technique, which plays the most important role compared to the others, is used to control actual timing of circuit. Self-tracking technique enables the circuit to work on different frequency instead of being fixed at a certain value. The circuit only operates either read or write in one clock cycle. Therefore, when it comes to performance, the single port works less efficiently compared to its more common realtive - pseudo dual port.

REFERENCES

[1] Neil H. E. Weste and David Money Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, fourth edition, United States of America: Addison – Wesley, 2011.

[2] Behzad Razavi, *Design of Analog CMOS Integrated Circuits*, second edition, New York: McGraw-Hill Education, 2017.

[3] Richard C. Jaeger, Travis N. Blalock, *Microelectronic Circuit Design*, 4th edition, Newyork, McGraw Hill 2010.

[4] Matthew Barlow, Guoyuan Fu, Brent Hollosi, Chris Lee, Jia Di, H. Alan Mantooth, *A PFET-Access Radiation-Hardened SRAM For Extreme Environments*, September 2008.

[5] A. Islam, M. Hasan, *Single-Ended 6T SRAM Cell To Improve Dynamic Power Dissipation By Decreasing Activity Factor*, January 2011.

[6] Cypress Semiconductor, *Introduction To Cypress SRAMs –AN1116*, October 5, 2006.

[7] Shang-Lin Wu, Kuang-Yu Li, Po-Tsang Huang, Member, IEEE, Wei Hwang, Life Fellow, IEEE, Ming-Hsien Tu, ShengChi Lung, Wei-Sheng Peng, Huan-Shun Huang, Kuen-Di Lee, Yung-Shin Kao, and Ching-Te Chuang, Fellow, IEEE, *A 0.5-V 28-NM 256-KB Mini-Aarray Based 6T SRAM With VTrip-Tracking WRITE-ASSIST.*

[8] Nese Chaya Anusha, Sapati Upendar, *High Speed Performence Of SRAM Cache Development*, August 2016.

[9] Vishal Sharma, Santosh Vishvakarma, Shailesh Singh Chouhan, Kari Halonen, *A Write-Improved Low-Power 12T SRAM Cell For Wearable Wireless Sensor Nodes*, August 28, 2018.

[10] Jingcheng Wang, Xiaowei Wang, Charles Eckert, *A Compute SRAM With Bit-Serial Integer/Floating-Point Operations For Programmable IN-MEMORY Vector Acceleration*, Feb 2019.

Assoc. Prof. Dr.-Ing. Nguyen Van Cuong received his Ph.D at BW University in Munich, Germany (2000), Post-Doctoral Researcher at University of Stuttgart, Germany (2004) and Visiting Scholar at UW, Seattle, USA (2007). He is currently at Faculty of Electronics and Telecommunications, DUT, Vietnam.

Eng. Nguyen Cong Luong receives his undergraduate bachelor in Electrical and Electronics Faculty in Ho Chi Minh University of Technology, Vietnam.
He is currently a AMS Engineer at Savarti Company Limited.

Tran Nguyen Phu Phu receives his undergraduate bachelor in Electronic and Telecommunication at DUT, Vietnam.

He is currently a Physical Design Engineer at Savarti Company Limited.

Dang Phuong Gia Han receives her undergraduate bachelor in Electronic and Telecommunication at DUT, Vietnam.
She is currently a Model Design Engineer at Savarti Company Limited.