Razor Flop Based On Chip Delay Measurement

Student Name: Aditya Vasisth
M.Tech-Electronics and Communication Engineering-2013-2015
June 2015

Indraprastha Institute of Information Technology, New Delhi

Advisor
Dr. Mohammad S. Hashmi
Mr. Abhishek Jain

Submitted in Partial fulfilment of the requirements
For the degree of M.Tech in Electronics and Communication Engineering

Student's Declaration

I declare that the dissertation titled "Razor Based On Chip Delay Measurement" submitted by Aditya Vasisth for the partial fulfilment of the requirements for the degree of Master of Technology in Electronics and Communication Engineering is carried out by me under the guidance and supervision of Dr. M. S. Hashmi at Indraprastha Institute of Information Technology, Delhi and Mr. Abhishek Jain at STMicroelectronics, Greater Noida. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... Place and Date: …….....
Aditya Vasisth

CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

……..
Dr. Mohammad. S. Hashmi

ABSTRACT

Moving towards deep submicron technologies, consideration for design margin will increase. In these technologies, process variation will increase the requirements of design margins. Lower supply voltage and higher clock frequency will further alleviate the issue of design margin. As technology progresses, STA (static timing analysis) is becoming a very important issue. With the increase in frequencies, there is a clear trend towards increasing setup and hold time violations. Thus, there is an increasing requirement of accurate and quick measurement techniques, to take care of all such requirements. Over reliance on tools for STA analysis can lead to inflexible designs. Off chip measurement techniques, such as spectrum analyzer, sampling oscilloscope and time interval analyzer measure delay but as the complexity of the system and chip integration increases, these off chip instruments become unviable.

This dissertation discusses available techniques for on chip delay measurement. Several metrics, such as resolution, area and linearity have been compared. Furthermore, discussions on setup time violations are provided. Violation in setup time occurrence has been ingeniously used to predict the access time for data in NVM. (Non Volatile Memory).

ACKNOWLEDGEMENTS

The work for this dissertation was carried out at STMicroelectronics, Greater Noida, India, during the year 2014-2015.

Firstly, I would like to thank my advisers Dr. Mohammad. S. Hashmi and Mr. Abhishek Jain for providing excellent guidance and encouragement throughout the journey of this work. Without their guidance, this work would never have been a successful one. I also take this opportunity to thank my friends, Mr. Jitendra Kumar Yadav and Mr. Arvind Srinivasan from IIIT-D for the technical discussion whenever I was in need of any. Last but not the least, I would like to thank my parents for supporting me spiritually and emotionally.

PUBLICATIONS:

LIST OF FIGURES

CONTENTS

1 POWER

1.1 INTRODUCTION

The current trend in the VLSI technology is towards high performance and energy efficient devices. Such requirements are often conflicting and warrant tradeoffs. Higher performance requires faster transistors but faster transistors would consume more power. Thus techniques to reduce power consumption become essential.

Power in electrical circuit is the time rate of change of energy

The instantaneous power absorbed or dissipated by a given element is equal to the multiplication of the current through the element and voltage across the element

$$Power(t) = Voltage(t) * Current(t) \quad (1)$$

Since power is the time rate of change of energy, now in order to calculate energy over a time interval T, take the integral of (1).

$$Energy = \int_0^T Power(t)\, dt \quad (2)$$

Now the Average Power over this interval will be given as

$$Pavg = \frac{Energy}{T} = \int_0^T Power(t)dt/T \quad (3)$$

Power dissipated in the resistor will be given as

$$Power = V^2/R \quad (4)$$

Similarly the energy stored by the capacitor when it charges from 0 to V is given as

$$Energy = C * V^2/2 \quad (5)$$

When the input changes (V_{IN}) from 1 to 0 PMOS turns on and it charges the load capacitance to

$$Ec = 1 * Cl * Vdd^2/2 \quad (6)$$

On the other hand total energy supplied by the supply is

$$Etotal = Cl * Vdd\text{^}2 \quad (7)$$

Equations (6) and (7) show that only half of the energy provided by the supply is being dissipated as heat and the other half of the energy is being stored in the capacitor. Now, when the when the input changes from logic low to logic high, the NMOS turns on and it discharges the output node to logic low. The energy stored in the load capacitor during the previous cycle is dissipated in the NMOS transistor and no energy is drawn from the supply during this transition as there is no path from VDD to Ground during this cycle. The transfer characteristics of CMOS inverter shows that there is a short period of time during which both NMOS and PMOS are ON and the current which flows during that time is called short circuit current.

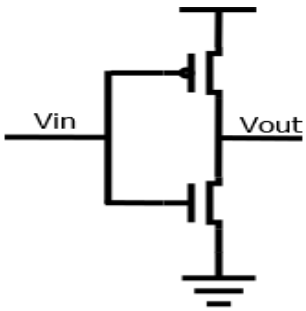

Figure 1: CMOS INVERTER

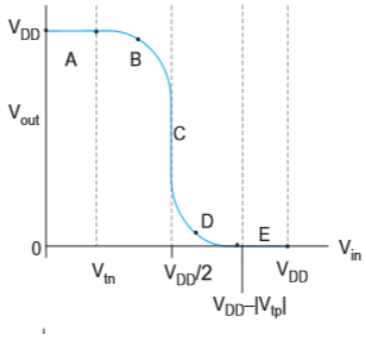

Figure 2: Transfer Characteristics [1]

Region	Condition	P Device	N Device	Output
A	$0 \leq V_{in} < V_{TN}$	Linear	Cutoff	V_{DD}
B	$V_{TN} \leq V_{in} < V$	Linear	Saturation	$V_{OUT} \geq V_{DD}/2$
C	$V_{in} = V_{DD}/2$	Saturation	Saturation	V_{OUT} drops sharply
D	$V_{DD}/2 \leq V_{in} \leq V_{DD} - \lvert V_{TP} \rvert$	Saturation	Linear	$V_{OUT} \leq V_{DD}/2$
E	$V_{in} \geq V_{DD} - \lvert V_{TP} \rvert$	Cutoff	Linear	$V_{OUT} = 0$

TABLE 1: CMOS INVERTER OPERATION REGIONS [1]

Let F_{SW} be the switching frequency of the gate, then over the time interval T the gate will switch $T*F_{SW}$ times. Then average power dissipation will be given as

$$Pavg = \frac{E}{T} = C * Vdd^2 * Fsw \quad (8)$$

This dissipated power is called the Dynamic Power since it happens due to the switching of load. The power dissipated due to the sub- threshold leakage through the OFF transistors, leakage through gate dielectric and junction leakage through source and drain diffusion is called Static Power. The total power, which is the sum of static and dynamic power is calculated using (9)

$$Total\ Power = Static\ Power + Dynamic\ Power \tag{9}$$

1.2 DYNAMIC POWER

Dynamic power dissipation occurs mainly due to the switching. The expression for the dynamic power is defined as (10)

$$Dynamic\ Power = \alpha * Fsw * Vdd\text{^}2 \tag{10}$$

Here α is the activity factor.

1.3 TECHNIQUES TO REDUCE DYNAMIC POWER

1. Clock Gating : Clock Gating is one of the techniques to reduce dynamic power dissipation. It is based upon the principle that sequential blocks do not require clock all the time. So during their idle mode if somehow clock gets stopped then in that case activity factor would be zero and thus according to equation (10) dynamic power would automatically get reduced to zero. Clock gating can be implemented by ANDING the clock with the ENABLE signal.
2. Switching Capacitance: Switching capacitance occurs due to wires and transistors in a circuit. Device capacitance can reduced by using smaller sized transistors and by employing less number of logic-stages in a circuit. On the other hand wiring capacitance can be reduced by doing good floor-planning, placement and routing.
3. Dynamic Voltage Scaling: Dynamic power shows a quadratic dependence on voltage. Lowering the voltage would significantly reduce the dynamic power dissipation. The circuit can be divided into multiple domains according to the amount of voltage it requires for proper functioning for example memory circuitry might require high voltage as it has to keep the memory cell stable on the other hand processor might require medium voltage where as low voltage is required for input output peripherals as they run at low speeds [1].

4. Dynamic Frequency Scaling: Dynamic power dissipation has direct dependence on frequency. So avoid running circuits at higher frequencies other than necessary. Reduction in frequency enables use of smaller transistors as well as lower supply voltages.
5. Dynamic Voltage and Frequency Scaling: Dynamic Voltage and Frequency scaling is nothing but the combination of previous two techniques. In this technique both voltage and frequency are getting scaled simultaneously. Reduce frequency to a point so that task gets completed on schedule and simultaneously reduce voltage to a minimum point sufficient to operate at that frequency. To implement the DVS system we need three components.
    1. DVS Controller: DVS Controller will determine the supply voltage and frequency required to complete the task on schedule. It will take information about the workload from the system [1]
    2. Core Logic: Core Logic is required to generate the specified clock frequency as per the information conveyed by the DVS Controller. For generating the specified clock frequency, the Core Logic might contain Phase Locked Loop or any other clock synchronizer [1].
    3. Switching Voltage Regulator: Switching Voltage Regulator will be required to step up and step down the voltage as required by the DVS Controller.
6. Short Circuit Current: Short circuit current, as discussed previously, will flow in a circuit and will lead to power dissipation when both NMOS and PMOS are on while the input switches. This problem will become worse when the input edge rates become slower as both NMOS and PMOS will be on for the large amount of time. However, the problem can be resolved by using a larger load capacitance because it will not allow output to switch to a higher value during the input transition. This leads to a small V_{DS} across the transistor that is generating the short circuit current [1].
7. Resonant Circuits**:** Resonant Circuits reduces the switching power consumption by letting energy flow back and forth between the storage elements like capacitors and inductors rather than dumping energy to ground. This technique is best suited for applications that require the clock to operate at constant frequency.

1.4 STATIC POWER

Static power is consumed even when there is no switching. It arises from sub-threshold, gate and junction leakage and due to contention current.

1. Sub-Threshold Current

This current flows in a transistor when it is supposed to be off and it is given by the equation (11).

$$Ids = Idso\ e^{(Vgs-Vt0+\eta vds-kVsb)}(1- e^{-Vds/VT}) \qquad (11)$$

Here, η is the DIBL effect coefficient, V_T is thermal voltage, k represents the effect of body on threshold voltage. The above equation shows that higher the value of V_{DS} then higher will be the leakage. From the above equation it can also be conclude that more the body effect modulates the threshold voltage then lesser will be the leakage. Sub-threshold current can be reduced by connecting the transistors in stack mode [1].

In figure 3 the drain of transistor N_2 is at V_{DD} so the transistors will leak. To establish the same current in both the transistors, node V_X will go high as a result of which drain to source voltage of N2 will go down and in this way leakage current will be reduced.

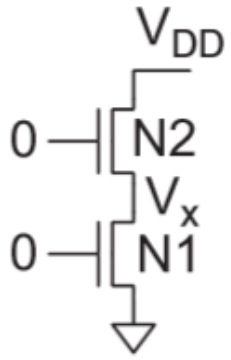

Figure 3: Reduction in sub-Threshold leakage due to stack effect[1]

2. Gate Leakage

Gate leakage occurs when carries tunnel through the gate oxide dielectric on application of high voltage to the gate. Gate leakage is a strong function of the oxide thickness and voltage applied to the gate.

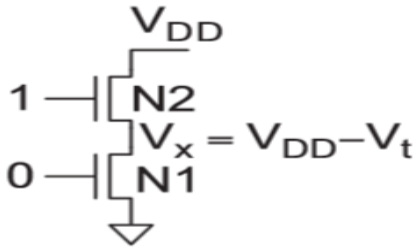

Figure 4: Reduction in gate leakage due to stack effect [1]

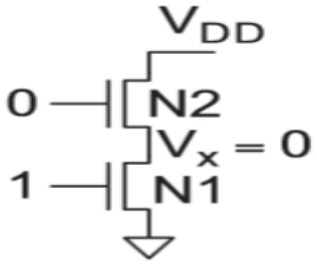

Figure 5: Gate leakage [1]

In figure 5, N1 is ON and the gate to source voltage seen by the transistor N1 will be V_{DD}. On the other hand, in figure 4, N2 is ON and the gate to source voltage seen by the transistor will be $V_{T.}$ As gate leakage is a function of $V_{GS,}$ the effect of leakage in the figure 4 will be less. In both the cases, it is assumed that OFF transistor experiences no leakage. Thus, the gate leakage can also be reduced by stacking transistors such that OFF transistor is nearer to the rail [1].

3. Junction Leakage

It takes place when source or drain region are at different potential from substrate because of formation of a reverse biased diode between n+(source or drain) and p+(substrate). Although the leakage current of reverse biased diode is usually negligible, but the leakage currents due to BTBT (Band to Band Tunneling) and GIDL (Gate Induced Drain Lowering) effects can be comparable to the sub-threshold current in High V_T transistors. BTBT is maximum when the voltage between the drain and body is high and the effect of GIDL is maximum when transistor is OFF and high voltage is applied to the drain.

4. Power Gating

Power Gating is the most effective technique to reduce static current by turning off the power supply in the sleep mode.

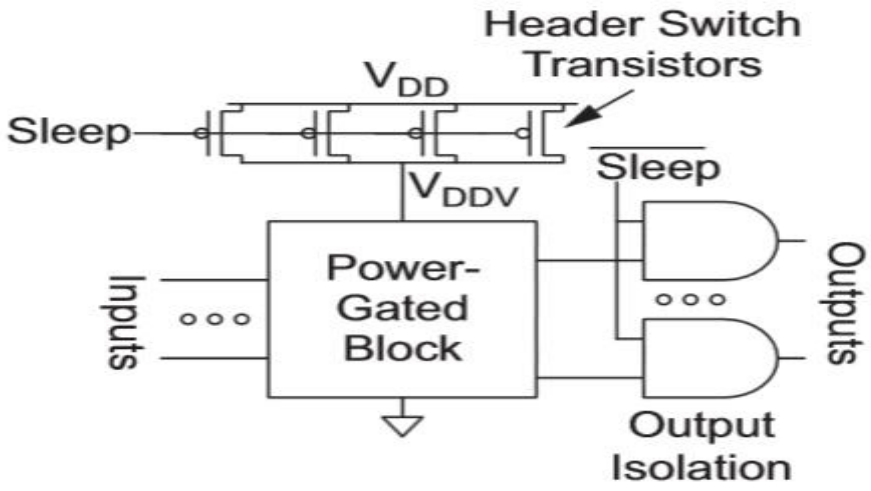

Figure 6: Power Gating [1]

For Sleep =0 , V_{DD}= V_{DDV} and when Sleep = 1 , V_{DDV} eventually settles to logic low . During this time node, V_{DDV} might acquire some unwanted voltage level. This requires output isolation gates to force the voltage to an acceptable level, so that they do not cause problems.

5. Multiple Threshold Voltages and Oxide Thickness

Low V_T transistors are used on critical paths to maintain the high speed whereas, high V_T transistors can be used on remaining paths use to minimize leakage. Similarly gate leakage is minimum in high oxide devices however, such devices they are not adequate for high speed applications [1].

6. Variable Threshold Voltages

To achieve high current in active mode and low leakage current in sleep mode, threshold voltage can be varied by applying proper body bias. This technique is called Variable Threshold CMOS.

2. RAZOR AND CANARY LOGIC

The previous chapter mainly focused on power, its types and how to minimize it. In this chapter focus will be on the implementation of power and speed efficient design. In deep submicron technologies, consideration of design margin will increase. In these technologies process variations will increase the requirements of design margins. Lower supply voltage and higher clock frequency will further alleviate the issue of design margin. Thus in order to realize various designs in these technologies only typical case design methodology is considered assuming rare occurrence of worst case design.

2.1 TYPICAL CASE DESIGN METHODOLOGY

In this methodology designer divide its design into two parts

1) Performance Oriented Design
2) Function Guaranteed Design

In Performance oriented design, designer has to think only about the typical cases and need not worry about worst possible cases. Since worst possible cases are not considered in this part that's why lesser number of design constraints will be there resulting in easy designs [2].

In function guaranteed design designer has to consider only about the function and not about the performance therefore all the worst cases possible will be included here [2].

Every critical function in LSI chip is designed by combing two methods. One is called the main part and other is called the checker part. While designing the main part, main emphasis is on the functionality. The main part might have some errors in its design and the error part provides reliability to the main part. The error part detects the errors that occur in the main part and thus has to satisfy all the design margins.

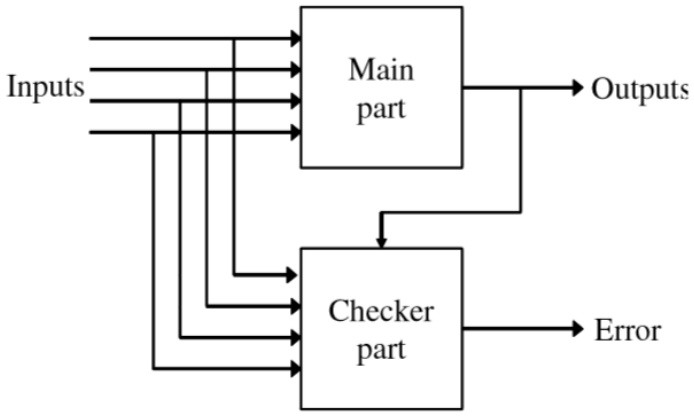

Figure 7: Typical case design methodologies [2]

Various examples of typical case design methodology are Razor and Canary Logic, CTV [3,4] and approximation circuits [5], ANT [6] and Teatime Circuits [7].

CTV [13] circuits generally makes use of input value variations. It is based on the fact that critical path are not always selected. In case of a failure CTV circuits has an approximate circuit to determine the timing violation and also has a desired back up to remove that violation.

ANT [6] circuit uses theoretical approach to determine lower bounds on energy and performance. Various circuit techniques and algorithms are generally developed to determine these bounds.

Teatime Circuits [7] uses a tracking circuit to identify the worst case margins. As long as tracking circuit is working fine we can lower the supply voltage and frequency.

In the approximation circuits [5], approximate circuit is used instead of actual circuit. The approximate circuit runs at higher frequency when compared to actual circuit and usually produces the correct results. In case it fails then the system utilising the approximate circuit must recover to a safe point.

This thesis will mainly focus on Razor and Canary Logic.

2.2 RAZOR LOGIC

The main concept behind the Razor logic is Dynamic Voltage Scaling. As mentioned previously in dynamic voltage scaling voltage is getting scaled to save dynamic power, in razor logic we are reducing the voltage margins to save power. Based upon the error difference voltage, controller will model the voltage, as shown in figure 8. A low voltage difference indicates that the supply voltage can be decreased and a high difference indicates that supply voltage must be increased to maintain a predefined error rate.

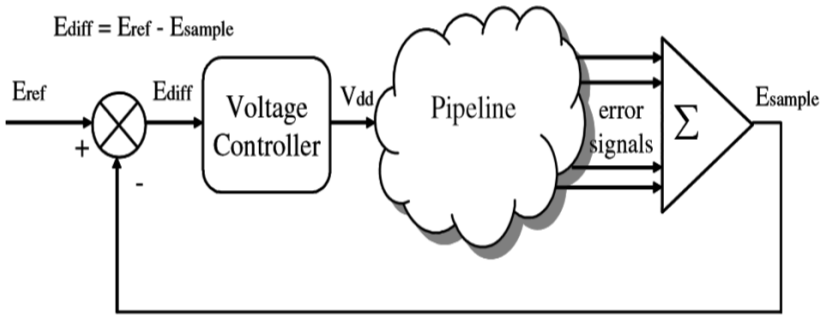

Figure 8: Razor DVS System [2]

To implement the logic discussed above, a Razor flop was proposed.

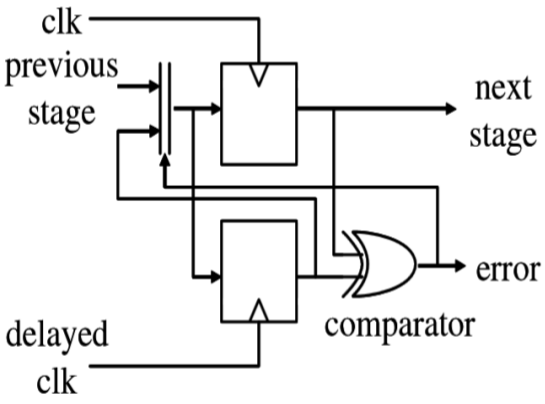

Figure 9: Razor system [2]

In a Razor flop, each main flop has its own shadow flop. A clock is provided to the main flop but a delayed version of the clock is provided to the shadow flop. Thus, the shadow flop is always expected to hold correct values even though a timing violation may have occurred in the main flop. In the design shown in figure 9, apart from error detection, error correction has also been done. An EXOR gate is used to compare the values of output of the main flop to the shadow

flop. In case of a mismatch, timing error is reported and output of the EXOR gate will go high. The output of the EXOR gate is also the select line of the multiplexer which will go high in case of mismatch. During timing violation, instead of taking input from previous stage, the Razor flop will reflect the output of the shadow flop and during the next cycle, correct data will be forwarded to the next stage.

One of the major disadvantages of the flop design is the vulnerability of the shadow flop to timing violations, thereby, making the design too dependent on the working of the shadow flop.

2.2 CANARY FLOP

Razor flop has several shortcomings. Canary flop is an improvement over the Razor flop. Analogous to the canary bird used by the miners to test a mine, the canary flop can be used to test for possible timing violations. The timing error is predicted by comparing the value of the main flop with that of the canary flop. It is obvious from the design mentioned in figure10, that data input to the canary flop is delayed instead of clock. So the canary flop will run into timing violation before the main flop .An alert signal triggers a warning for possible timing violations.

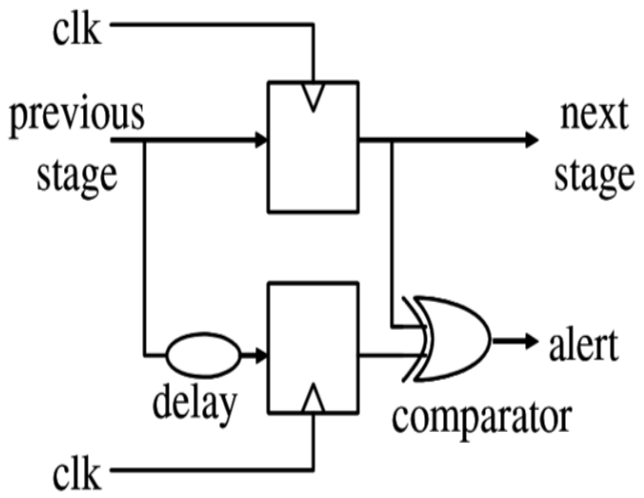

Figure 10: Canary System [2]

This design has following advantages over the Razor design:

1. Elimination of delayed clock [2]**:** The usage of a single phase clock greatly simplifies the design. It also eliminates the short path problem [8] that we counter in the razor flop

2. Robustness for variations [2]**:** the delay buffer used in the canary flop has a positive delay. Canary flop will always run into timing error much before the main flop no matter how the PVT variations affect the delay buffer.
3. Protection against timing errors [2]: Canary flop protects the main flop from timing errors i.e. it gives a precautionary warning for timing violations thus preventing timing violations on the main flop, making the correction mechanism redundant. Alert signal here triggers voltage or frequency control.

3 PROPOSED DESIGN

In the previous chapter, block level implementation of Razor and Canary flop was discussed. In this chapter, focus will be on transistor level design of Razor and Canary flops .Previously designed Razor flops shall be compared with the proposed design.

3.1 DESIGN OF RAZOR FLOP

3.1 Basic Design of Razor Flop

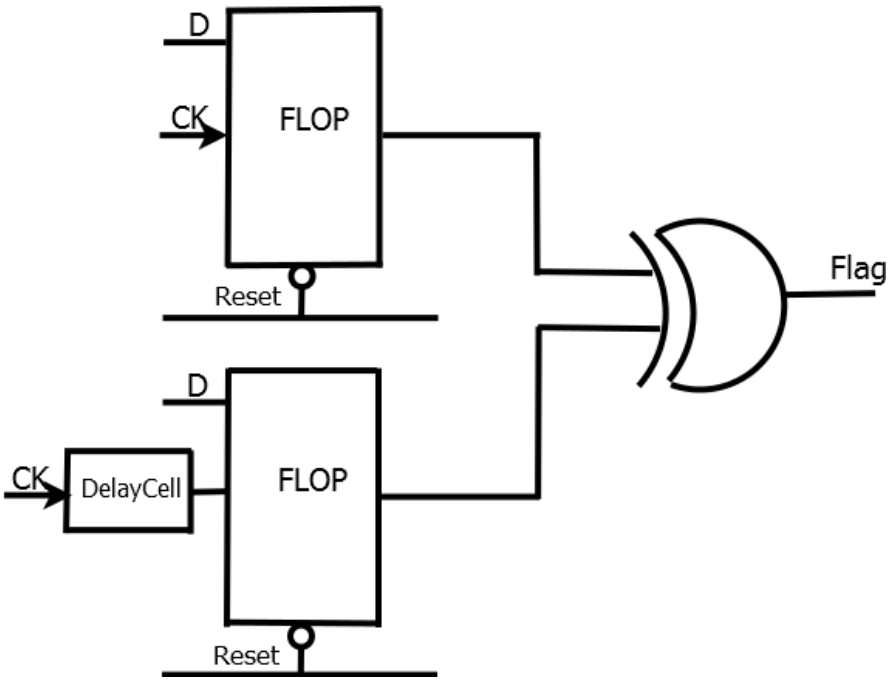

FIGURE 11: RAZOR FLOP

This is the most basic design of Razor Flop. This circuit has three components

1) Flop
2) Delay Cell
3) Exor Gate

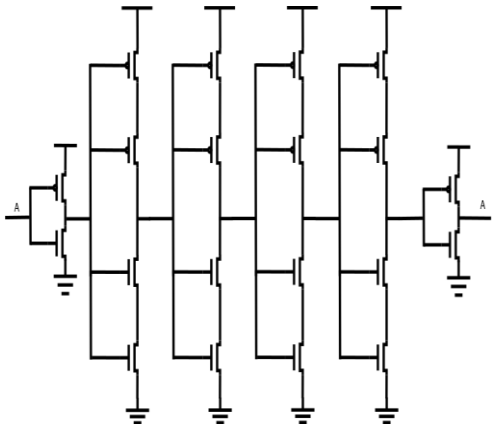

Figure 12: delay cell

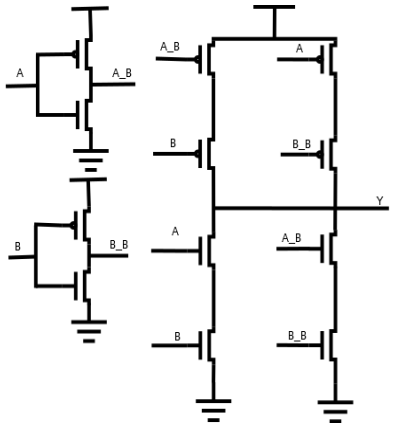

Figure 13: exor gate

Figures 12, 13, 14 are the schematics of the Delay Cell, XOR gate and D flip flop respectively. These are used in the design of Razor Flop. D flip flop in figure 14 is a master slave flip flop and is having an asynchronous reset. Asynchronous reset when activated will force the output Q to go high and is incorporated in the design of D flip flop by using a NAND gate. Delay cell is implemented by connecting a series of buffers in series on the other hand XOR gate is implemented by connecting two chains of NMOS and PMOS transistors in parallel.

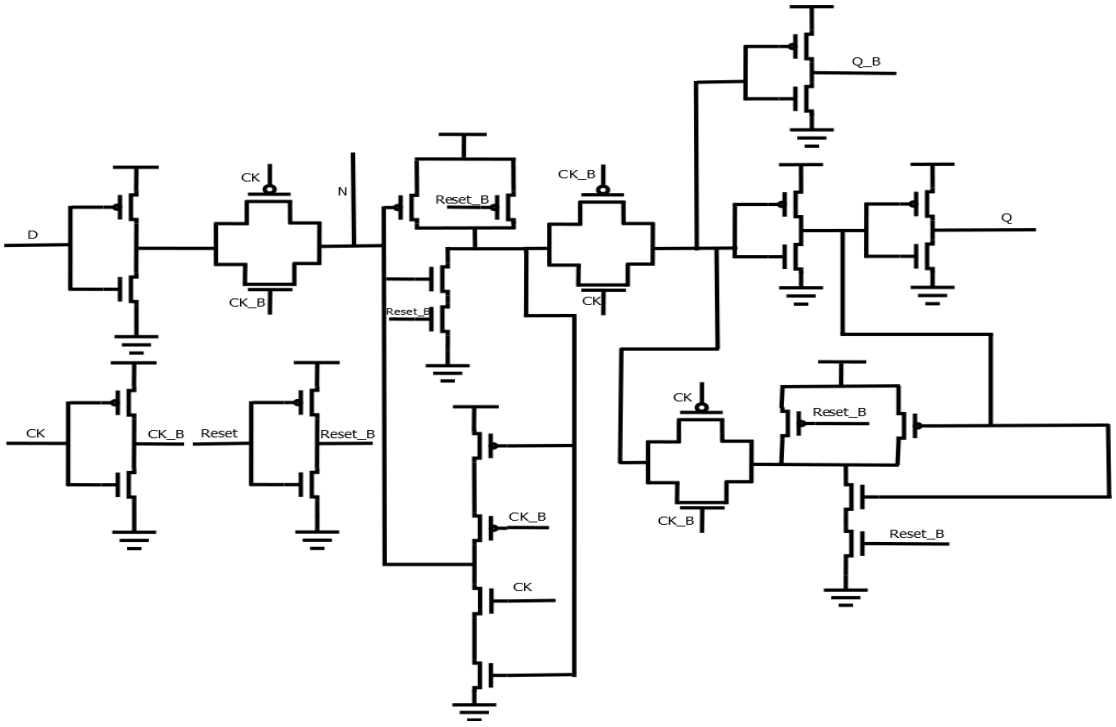

Figure 14: D flip flop

As previously mentioned, this design of razor flop will predict timing error by comparing the output of two flops. Output will remain low when data has setup before the rising edge of the clock and will go high when one of the flops used in the design has undergone timing violation. A plot of optimal working of the Razor flop is shown in figure 15. No setup violations were present, thereby keeping the output of the EXOR gate low. Setup violations force the value of EXOR gate high, as obvious from figure 16.

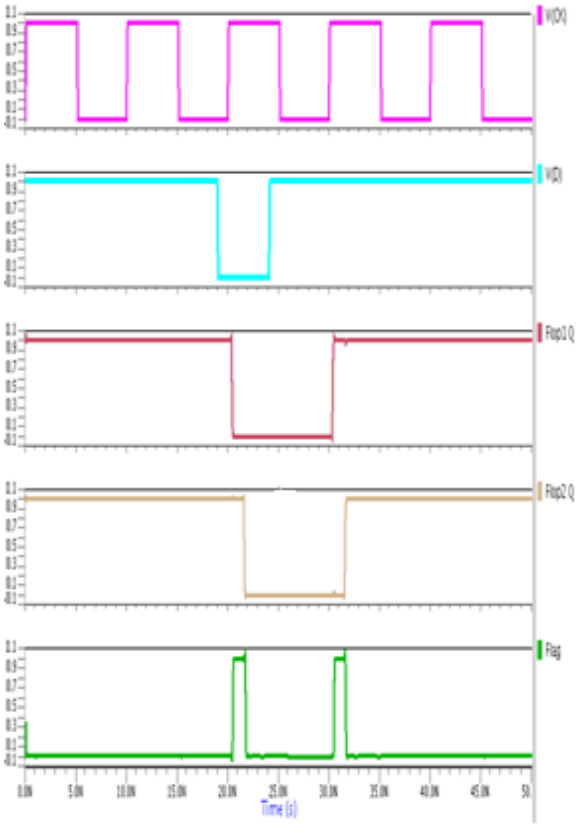

Figure 15: Output keeping low due to no set up violations

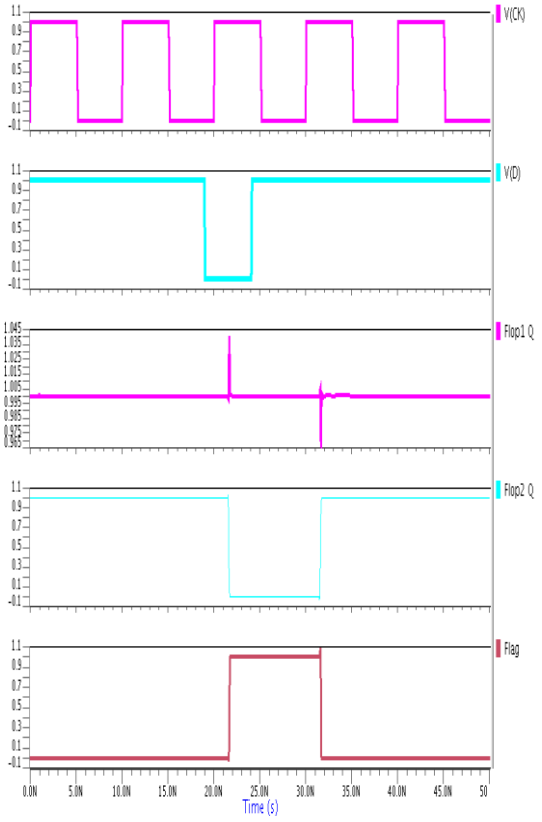

Figure 16: Set up violations forcing output high

The entire mechanism of detection of timing violation is based on delay cell, shadow flop and the EXOR Gate. But the problem with this design is that if somehow data gets delayed with the delay greater than the delay by which clock to the second flop is being delayed, then both the flops will miss the data and the flag output instead of going high will remain low. Figure 17 shown below represents above mentioned drawback of Razor flop.

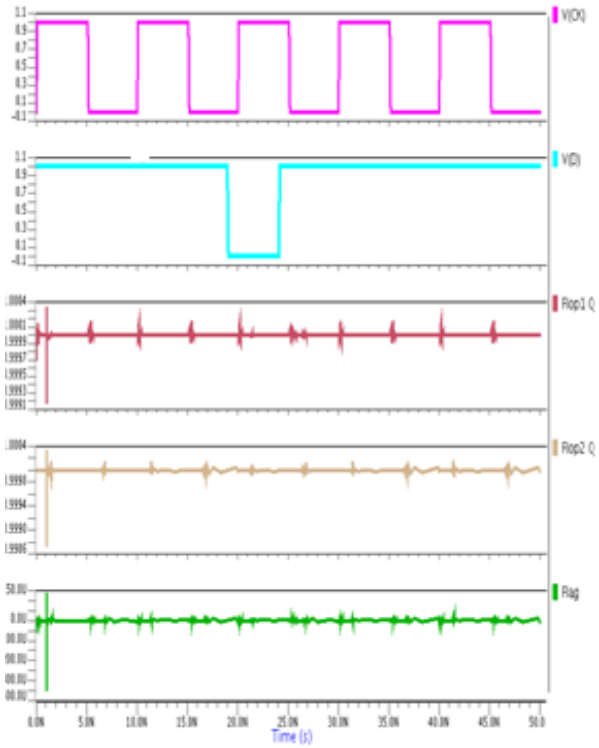

Figure 17: Drawback of razor flop

3.2 CANARY FLOP

The main difference between the canary flop and razor flop is that in the canary flop input to one of the flop was getting delayed whereas in the razor flop clock was being delayed. So the razor flop actually detect timing violation in a given flop whereas canary flop gives the warning that timing violation is about to occur in a flop. In the schematic design of the canary flop same components were used that were used in the construction of the razor flop but in this design some improvements were made in comparison to the razor circuit like usage of single phase clock, design was more robust and complexity was reduced.

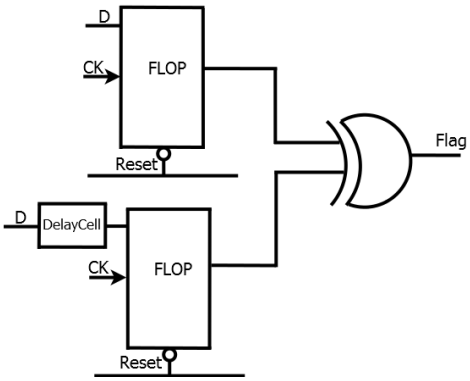

Figure 18: Canary Flop

Figures 19 and 20 explain the working of canary flop. Figure 19 represents output staying low when both the both the flops were able to catch the edge of data before the rising edge of the clock. Figure 20 represents the canary flop giving precautionary warning that timing violation is about to occur in main flop.

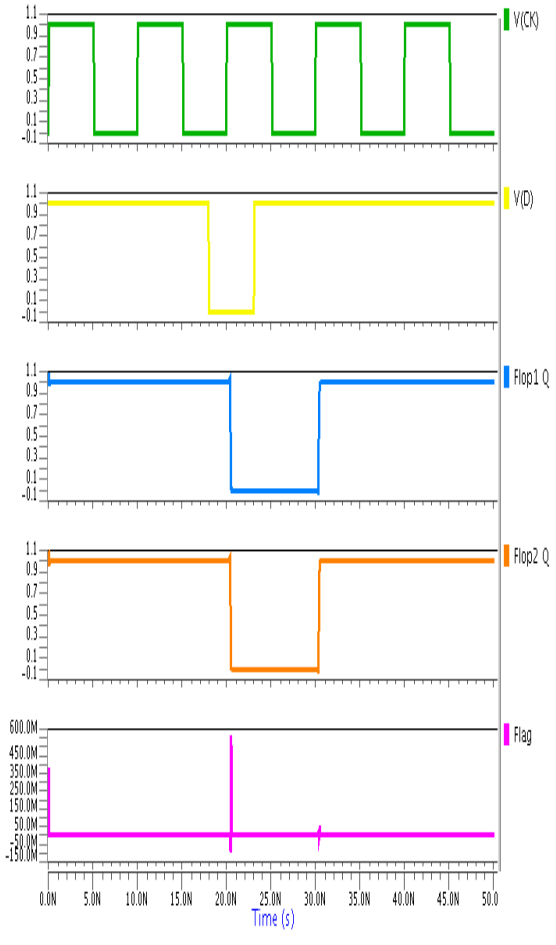

Figure 19: Output staying low when there is no timing violation

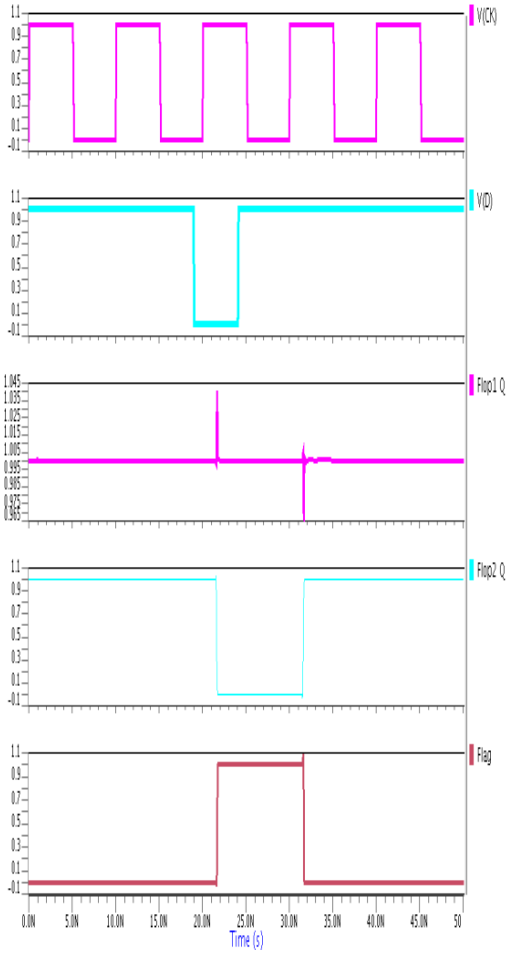

Figure 20: Set up Violation forces the output to go high

Here also the same problems persists which was coming in the razor circuit that is dependency on the other sequential circuit to detect timing violation and the output staying low when both the flops have undergone timing violation as shown in figure 21

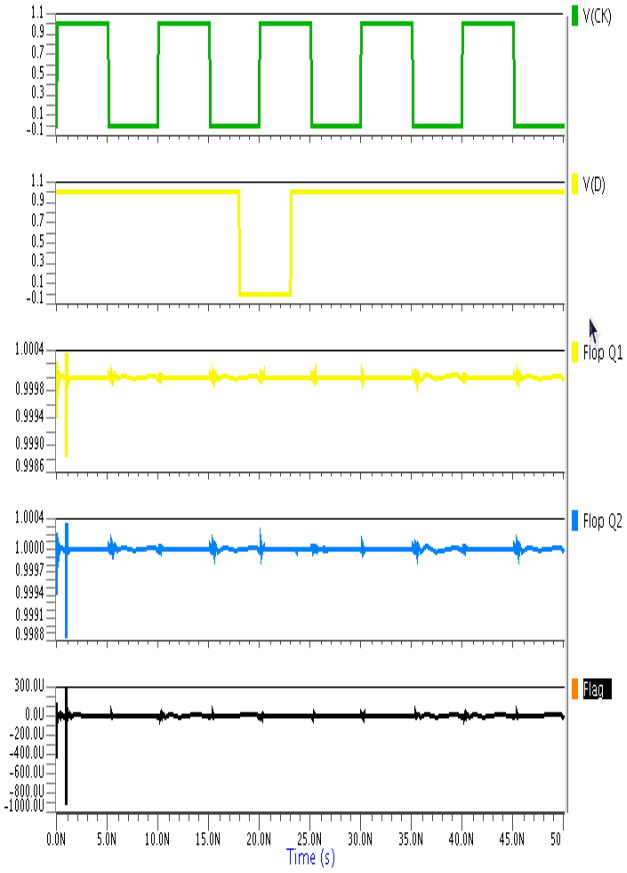

Figure 21: Drawback of canary flop

From the previous designs of razor and canary flops conclusion can be drawn that to design a proper razor flop, design should pass following figure of merits

1. Avoid using sequential circuits like latches and flops in design of razor flops: Since latch and flops are triggered by clocks. So there will always be a possibility of occurrence of timing violations in these circuits.
2. Design should be robust to PVT variations: Design should be constructed in such a way that it should be able to detect timing violation at each and every design corners, temperatures, supply voltages and threshold voltages.
3. Minimum Area: Since in VLSI industry area is money, design should be constructed with minimum possible transistors.

Keeping the following figure of merits in mind a design has been proposed to detect timing errors. The design consists of two major blocks delay cells and transition detector. Transition detector block is controlled by clock and apart from that the entire design is combinational.

3.3 IDEA BEHIND THE DESIGN

The main logic behind the design was to identify the node where the effects of timing violations will appear first and then display that violation at the output using some combinational blocks.

3.4 IDENTIFICATION OF A NODE

Flop consists of two latches master and slave. Master is active during first half cycle of clock and slave is active during the other half. The data passed on to the master latch during the first half will be passed on to the slave during the other half. Since data is passing through the master latch first so it means if timing error has to occur it would occur in the master latch first and then during the other half of the cycle it would be passed to the slave latch. Thus node where the violation is appearing will be in the master latch. Since setup violation is due to the transition of data after a predefined time before the clock edge. Thus the violated node will definitely be the output of the transistor to which clock is input. Thus in the proposed design violation will be appearing first at the output of the pass transistor connected in the master latch.

3.5 SETUP VIOLATION IN THE CIRCUIT

Let's consider that at the input D of a flip flop is a falling pulse. So the output of inverter will be a rising pulse. If the transition of input D from high to low happens well before the rising edge of a clock then pass transistor will have sufficient time to pass the output of inverter to node N. But if transition of input D happens somewhat closer to the edge of a clock then pass transistor (as pass transistor is active during the negative half of clock) will not be able to pass the output of inverter to node N and instead of taking value at node N as logic 1 it will consider a logic 0. As a result of this output of NAND gate will be 1 instead of 0 (considering R_B as 1). As a result of this output of tri-state inverter becomes 0 and thus instead of logic 1 logic 0 gets feedback to the node N. Figure 22 shown below represents the output of node N during normal and timing violation cases.

Identification of the node is done. So the task now is to represent the violated case at the output such that the Flag (output of entire circuit) should remain low during normal cases and it should go high during violated cases.

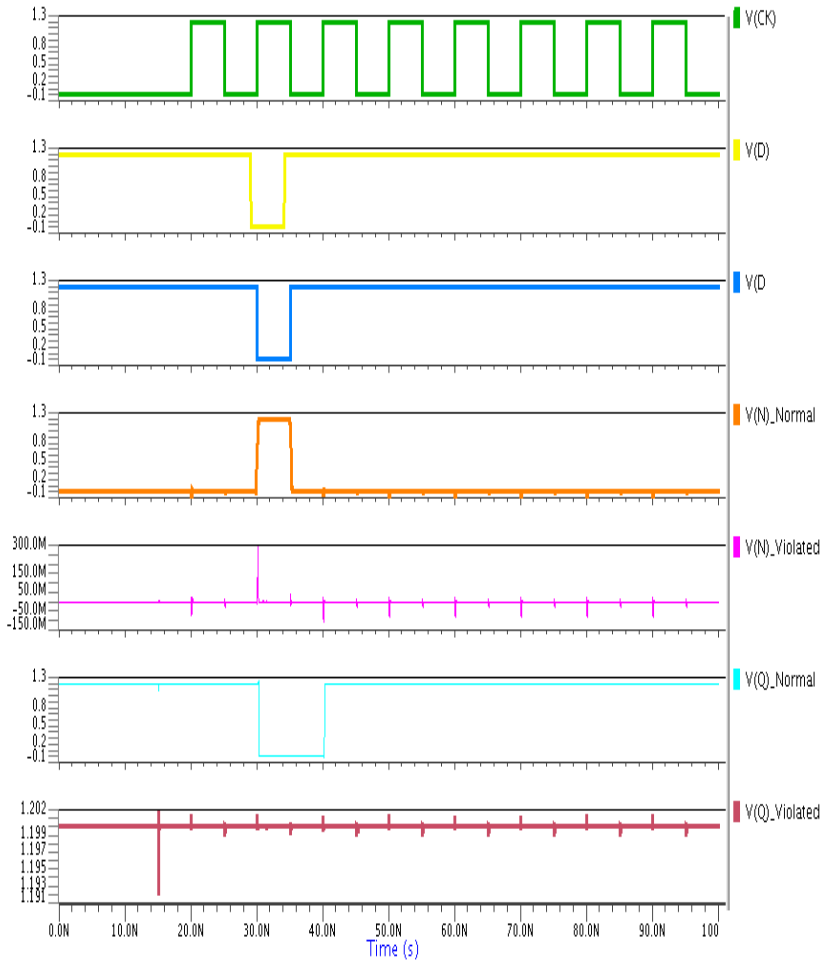

Figure 22: Output of Node N during normal and timing violation cases

3.6 DELAY AND TRANSITION DETECTOR BLOCK

Delay Block as the name suggests adds small amount of delay between the waveforms. Transition detector block is used to capture the transitions happening at various nodes. Transition detector is made up of 2 parallel chains of NMOS and PMOS transistors controlled by clock signal. One chain is for falling transition and other chain is for rising transition. Two chains of NMOS transistors are for passing logic low when setup violation is not occurring in the circuit

whereas chains of PMOS transistors are for passing logic high during set up violation cases. At the output of the transition detector a tri-state inverter is connected in feedback to prevent the output flag node from getting discharged. Figures 23 and 24 represent the schematic of delay Block and transition detector block. In figure 23 DX and DX_BAR are the pulses generated using input pulse D whereas DN and DN_BAR are the pulses generated using pulse at node N.

When input D is a falling pulse, DX_BAR and DN pulse will pass logic low during normal cases on the other hand DX and DN pulse will pass logic high during violation cases. When input D is a rising pulse, DX and DN_BAR pulse will pass logic low during normal cases on the other hand DX_BAR and DN_BAR pulse will pass logic high during violation cases.

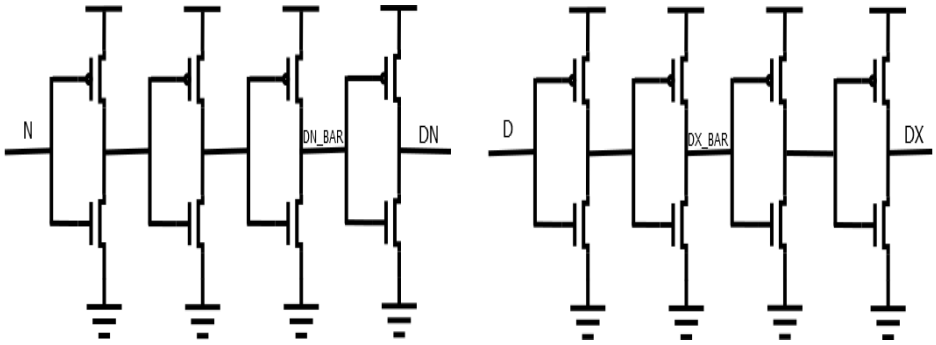

Figure 23: Delay Block

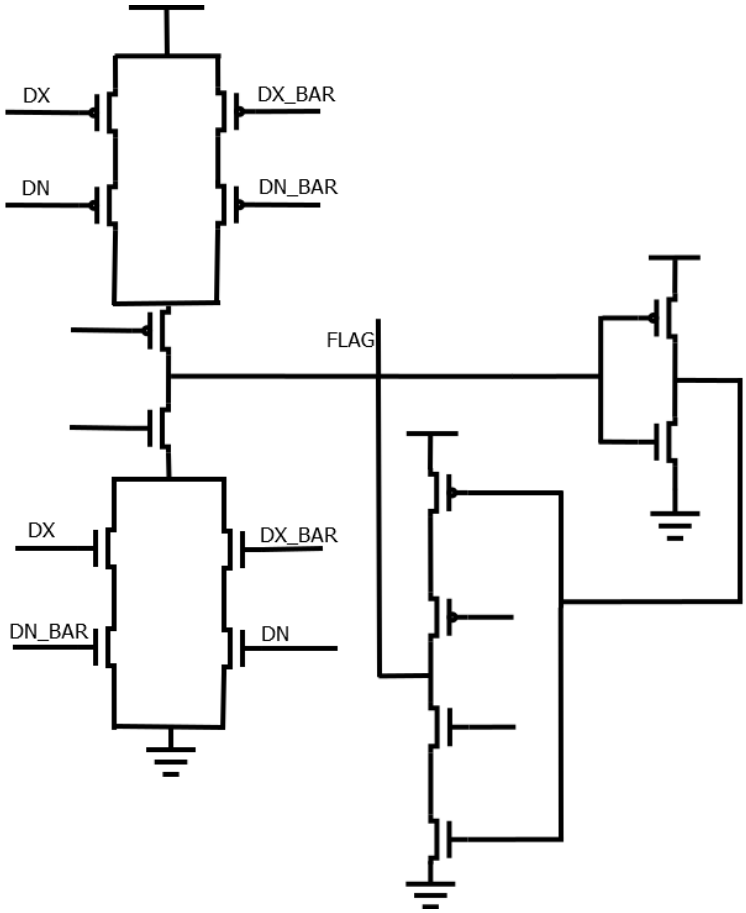

Figure 24: Transition Detector Block

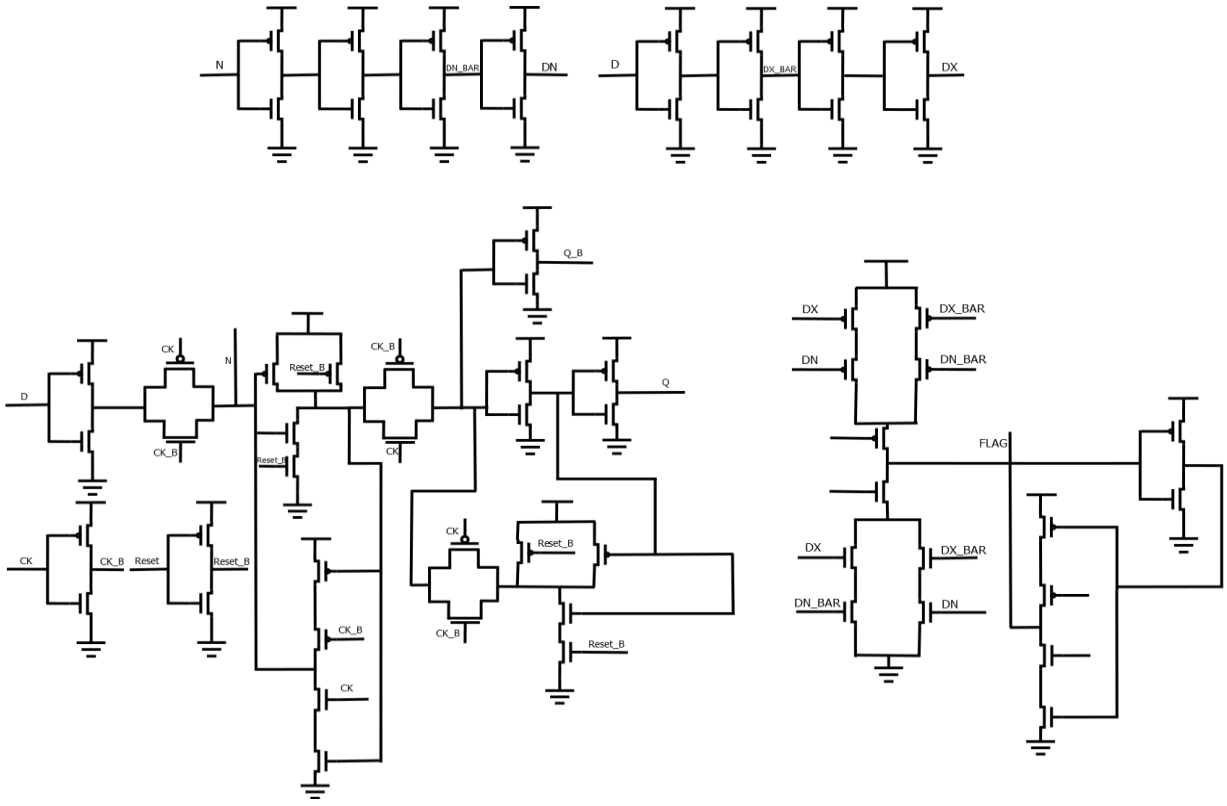

Figure 25: Schematic of the proposed razor flop

3.7 TESTING OF THE PROPOSED CIRCUIT

Design of the circuit is complete therefore it's testing should be done at various cases of input.

1. Input is a falling pulse
2. Input is a rising pulse
3. Input is a combination of rising and falling pulse

Figures 26, 27 represent the outputs for falling pulse during normal cases (when there is no set up violation) and during set up violation. Similarly figures 28, 29 represent the outputs for rising pulse for the above mentioned cases. Figure 30 represents the outputs when input is combination of rising and falling pulse.

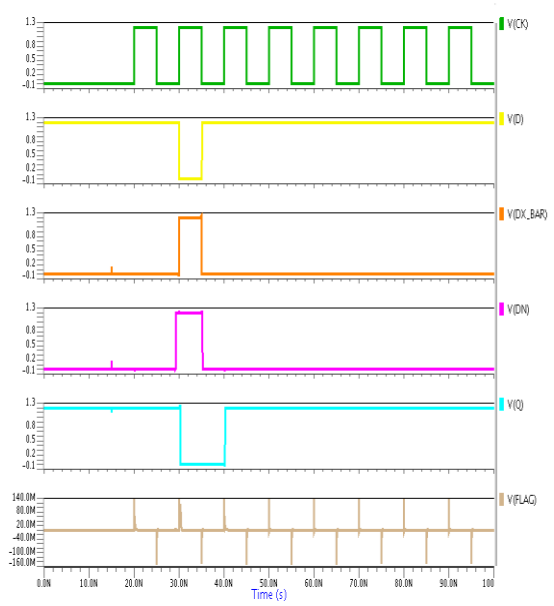

Figure 26: flag remaining low when there is no setup violation

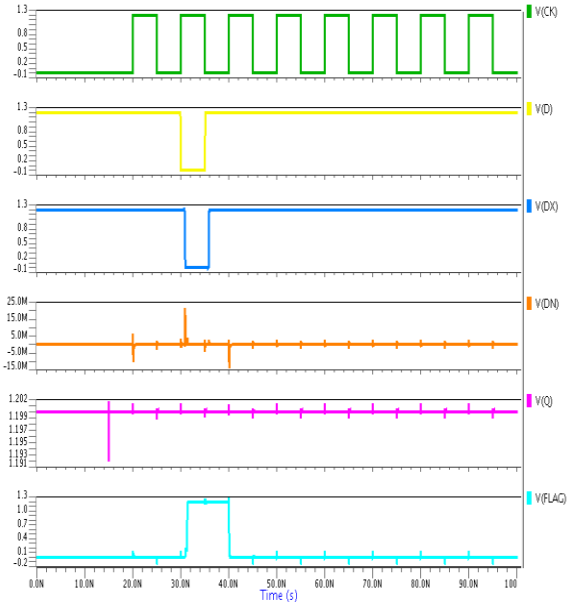

Figure 27: Setup Violation raising the output flag high

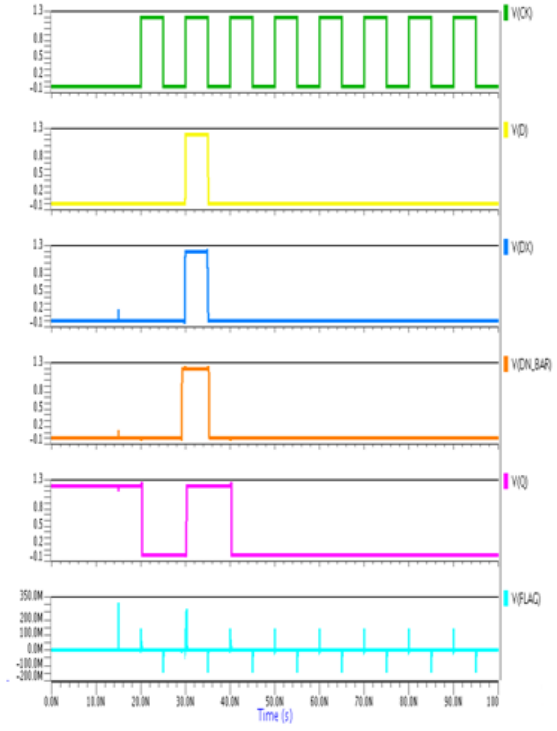

Figure 28: flag remaining low when there is no setup violation

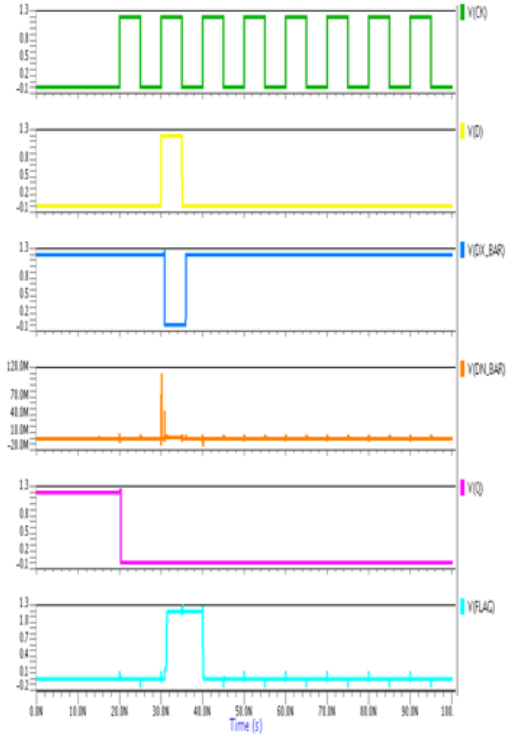

Figure 29: Set up Violation raising the output flag high

Figure 30: flag remaining low during normal cases and then going high for falling case and then going high for rising case

3.8 COMPARISON OF THE PROPOSED DESIGN

	RAZOR-1[9]	RAZOR-II [10]	DSTB [11,12]	RAZOR-LITE [13]	Proposed Design
TYPE	Flop	Latch	Latch	Latch	Flop
No of Transistors	44	31	26	8	32
Clock Path Loading	Yes	Yes	Yes	No	Yes
Data Path Loading	Yes	Yes	Yes	No	Yes
Usage of circuitry to detect Timing Violation	Sequential	Apart from transition detector the entire design is sequential	Sequential	Combinational	Apart from transition detector the entire design is sequential

Although the Razor-Lite design has advantages of clock path loading, data path loading, less no of transistors incorporated in design but the biggest disadvantage of the design is that entire circuitry has dependency on threshold voltage and the design is not successful for lower supply voltages.

On the other hand DSTB design of Razor flop incorporates sequential circuitry to detect timing violation. Since sequential circuits are itself triggered by clock there is high probability of occurrence of timing violations in theses circuits and output getting corrupted

Although our design requires extra transistor to detect timing violations but the design is free from the limitations mentioned above.

4 ON CHIP DELAY MEASUREMENT

In the last chapter discussion was about the transistor level implementation of Razor Flop. This chapter will discuss how Razor flop will acts as an important building block in on chip delay measurement

4.1 NECESSITY OF ON CHIP DELAY MEASUREMENT

In high speed systems delay is one of the most important parameters that can affect system performance. There are many off chip measurements like spectrum analyzer, sampling oscilloscope, time interval analyzer that can measure delay but as the complexity of the system and chip integration increases these off chip instruments offers many demerits.

1. Accessibility: The main requirement of these off chip instruments is that signal which is under test has to accessible to the test probe. However now days this cannot be always guaranteed due to the highly integrated SOC's.
2. Loading: The signals which we are using are high frequency signals which are extremely sensitive. The parasitic capacitance and inductance associated with the test probe may introduce distortion in the measurement.
3. Test Cost Challenge: These off chip instruments are really expensive and they also need longer time for setup and calibration. These disadvantages may limit their effectiveness for high volume production.

Currently On chip delay measurement is done by using three techniques [14]

1. Vernier Delay Line: Based on Vernier concept. Offers good resolution and is robust. Main Disadvantage is that while measuring delays of higher order area overhead will be large. [15]-[17]
2. Ring Oscillator: Here, delay to be measured is made part of ring oscillator and frequency of oscillations is measured with or without it. Resolution is limited here [18]-[20]
3. Time to Voltage Conversion: Here delay to be measured is converted to a voltage level which in turn is converted to a digital output using ADC etc. Calibration and robustness are issues here [21]-[23]

4.2 PROPOSED DELAY MEASUREMENT TECHNIQUE

In the proposed delay measurement technique razor flop is being used as an important building block to measure delay. Testing of the circuitry is done using NVM macro. Access time of the NVM macro is being measured using the proposed circuitry. In flash memories access time is defined as that time after which data becomes reliable for the next cycle or access time is the time that can be tolerated without causing setup violation for the next cycle.

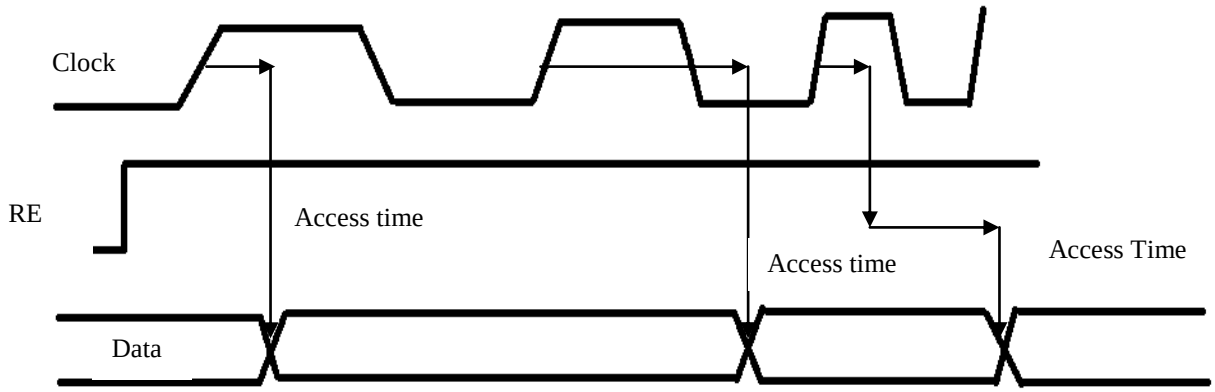

Figure 31: Proposed Technique

In the figure 31 shown above time between the rising edge of the clock and the corresponding edge (rising / falling) of the data needs to be measured. Let‘s say initially at the start clock period is 100ns and time (need to measure) after which the data comes is 37ns. So for the next clock cycle there will be no violation. Then shrink the clock period to say 50ns and again for the next clock cycle there will be no violation. Now again shrink the clock period to 37ns then there will be a timing violation which would be detected by the razor circuitry attached to the flop. Then that clock period will be access time.

Thus measurement of access time requires a clock whose period can be controlled. Thus, the implementation of above mentioned technique requires construction of ring oscillator generating clocks of different period.

4.3 DELAY GENERATING CLOCK

In figure 32 shown below two types of delay cells are being used: D1 and D2. D1 is introducing a delay of 1ns and D2 is introducing a delay of 40ps. 8x1 and 4x1 are the multiplexers that are

being used to generate proper amount of delay using the delay cells D1 and D2. Minimum and maximum clock period which can be generated is 3.54ns (all the select line are zero) and 43.847ns (all the select lines are one). Minimum delay which can be added is 40ps by making one of the select lines 1 (3.40ns). Thus the resolution of circuit is 40ps

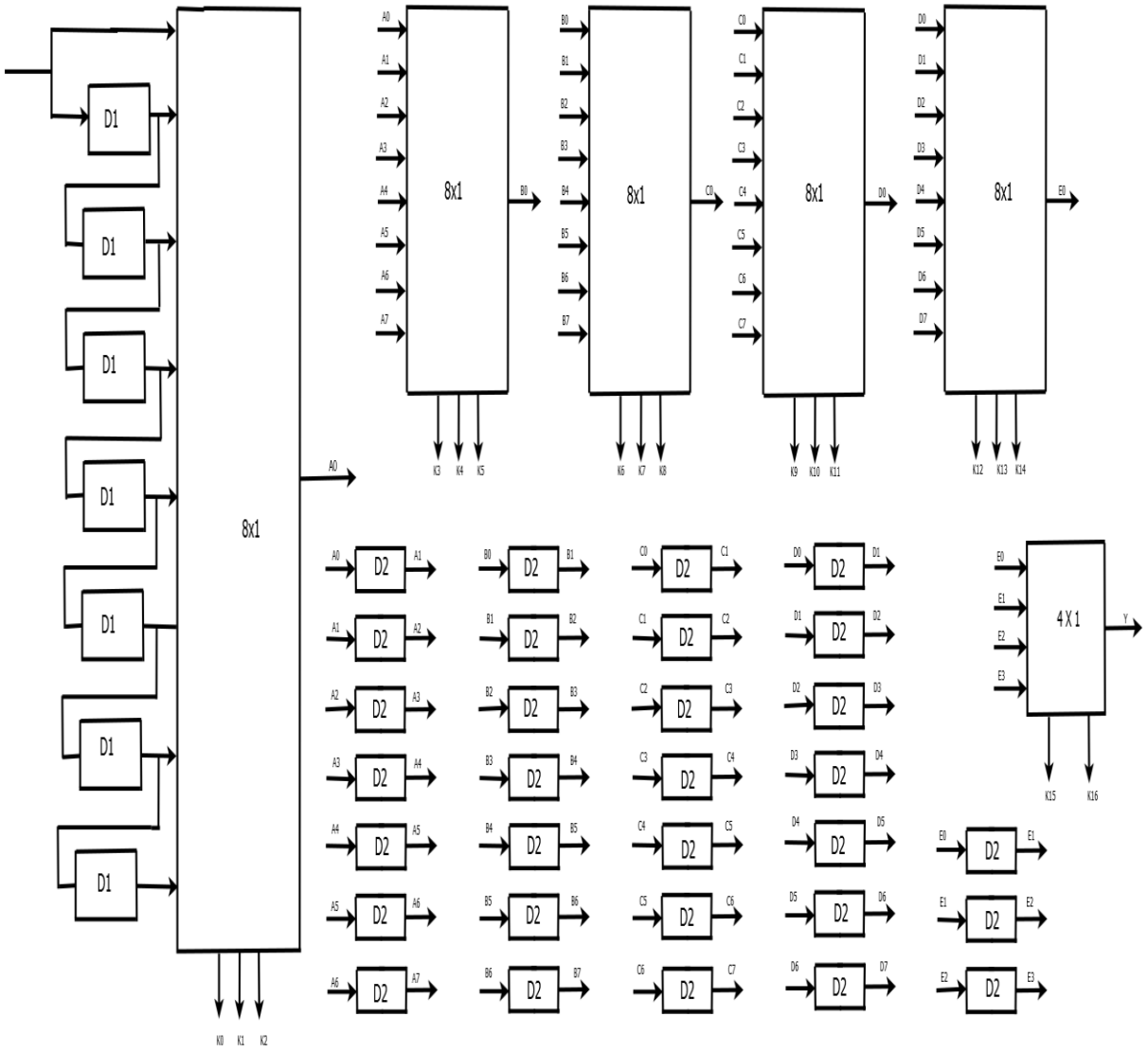

Figure 32: Delay generating block

Figure 33, 34, 35 shows the clocks of different period generated using delay generating block.

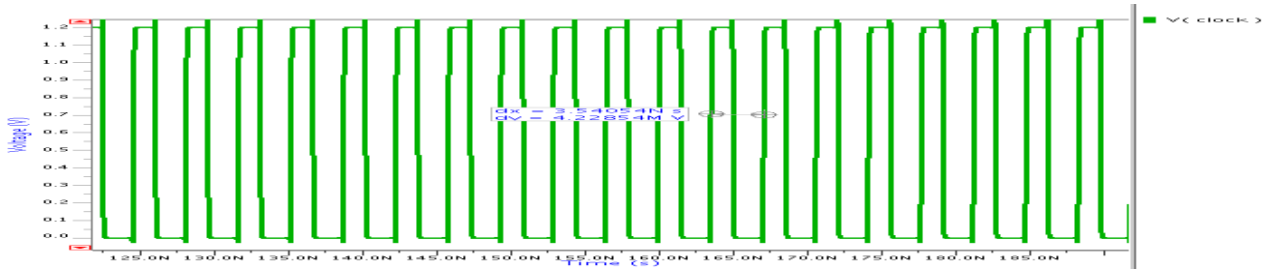

Figure 33: Minimum period clock

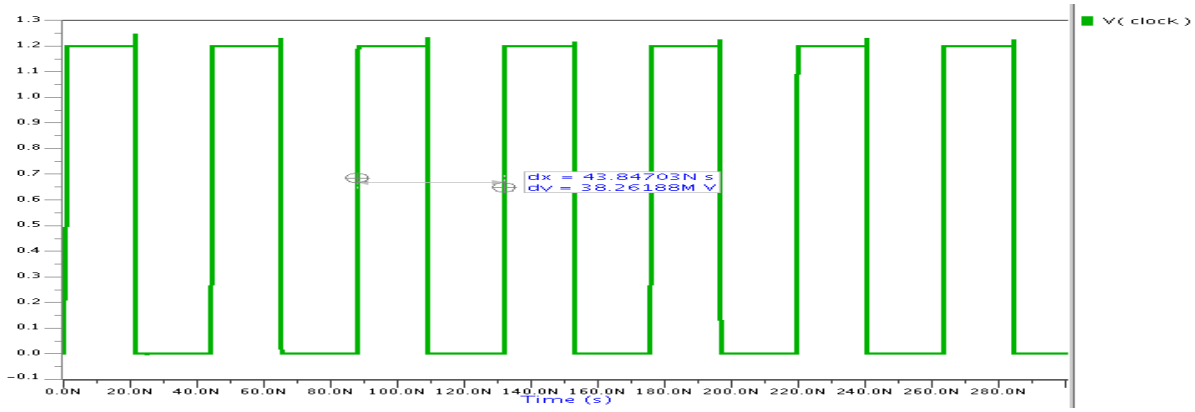

Figure 34: Maximum period clock

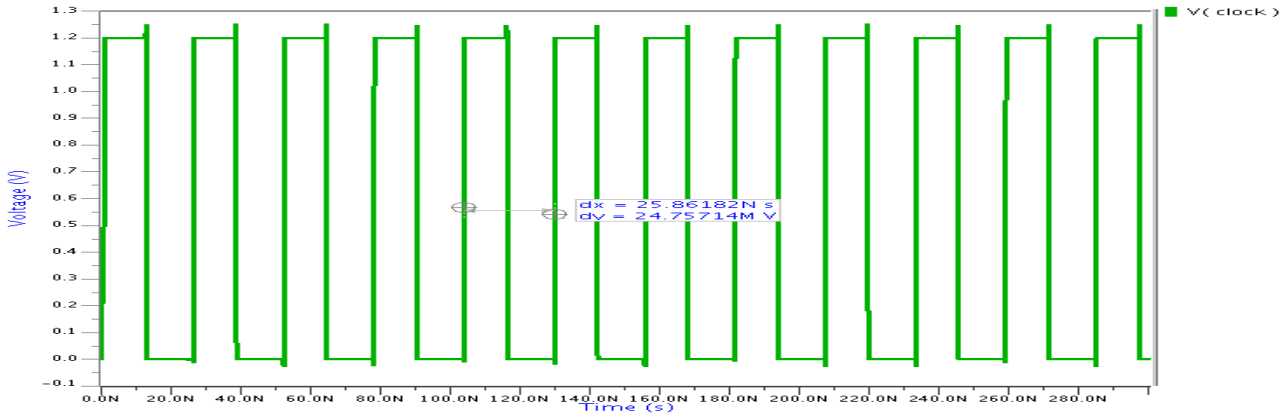

Figure 35: Intermediate period clock

So far clocks of different periods have been generated; the only thing left is the construction of the delay measurement circuitry and the measurement of the access time of the NVM Macro.

4.4 DELAY MEASUREMENT BLOCK

Delay Measurement Circuitry consists of an ring oscillator, delay generating block, setup violation detector (flop and razor circuitry), 2x1 multiplexer and another flop connected in toggle mode for the on chip delay measurement. The clock for the NVM macro is generated by the ring oscillator and output of the macro is connected to input of D flop. If the timing violation takes place in the circuit the output of the razor circuitry will go high which in turn is connected to the select line of the multiplexer. If the select line of the multiplexer is high clock corresponding to the setup violation will go into the clock input of the flip flop and the output will of the flip flop will be a signal having twice the period of the clock for which timing violation have taken place.

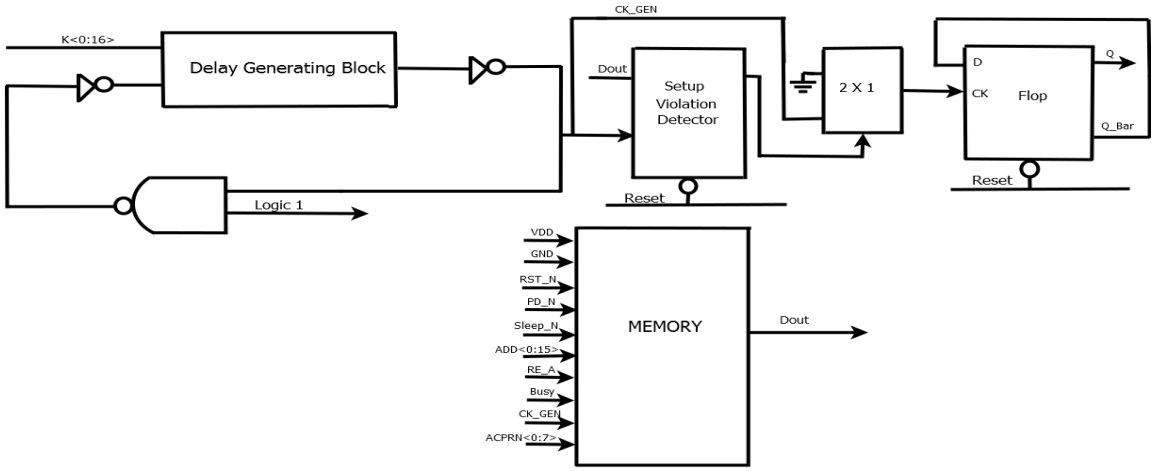

Figure 36: Delay Measurement Circuitry

As mentioned before access time of the non volatile memory is equal to the time period of the clock. So the relation between the access time and output signal will be

$$Toutput = 2 * access\ time$$

Maximum error in the measurement is equal to the: $(Time\ at\ which\ Clock\ edge\ appears - Setup\ Time\ of\ the\ flop) + Error\ due\ to\ the\ toggling\ flop$

Error due to Toggling flops

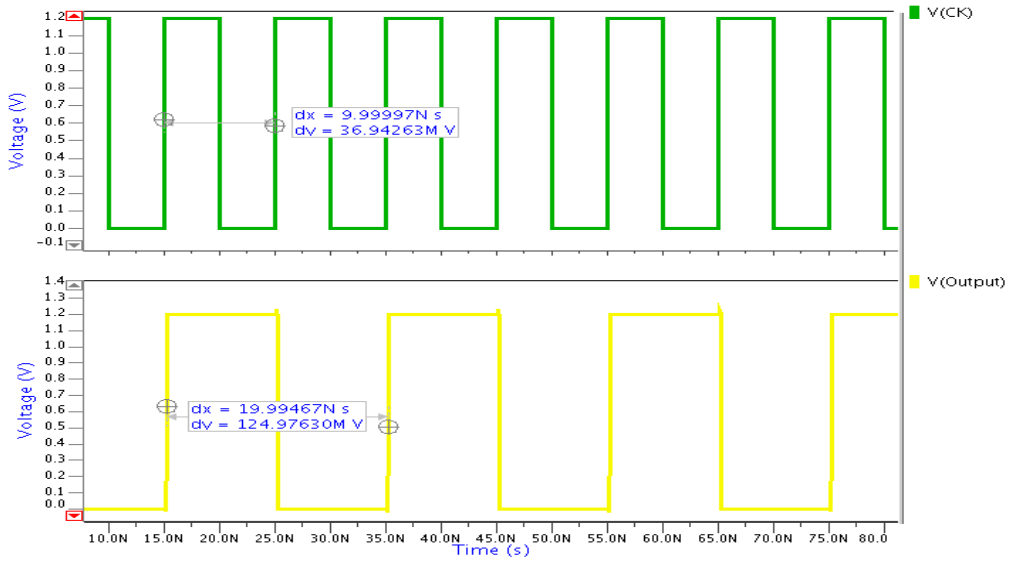

FIGURE 37: REPRESENTING ERROR IN THE TOGGLING FLOP

4.5 Delay Measurement

Figure 38 and 39 show two different cases of on chip delay measurement. Data is arriving at 25.05 ns. In fig 28 since clock period is greater than 25.05ns no timing violation occurred that ‘s why flag remained low and no on chip delay measurement took place.

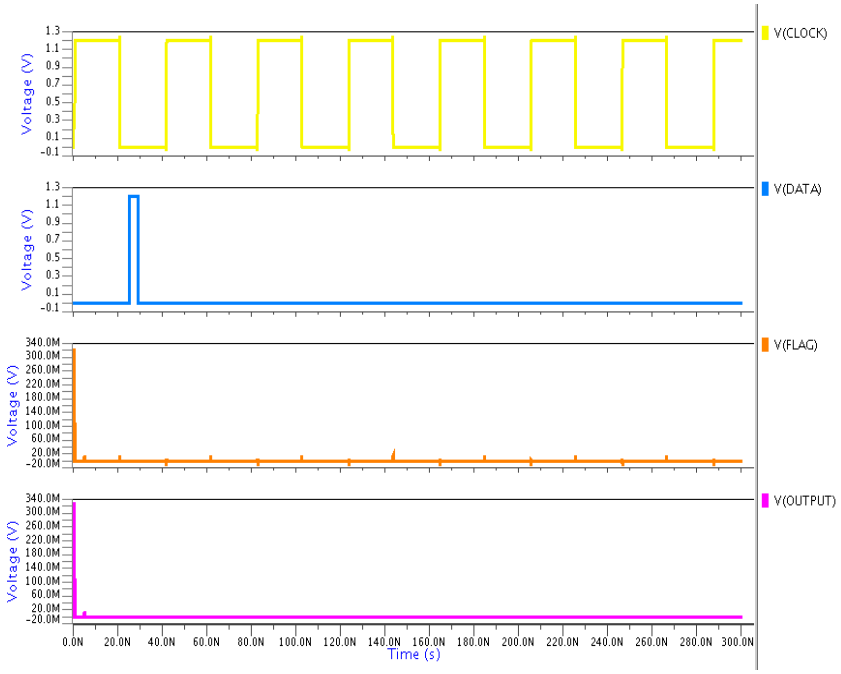

Figure 38: case of no timing violation and no delay measurement

But in figure 39, with reduction in clock period, a setup violation, caused by early clock arrival causes a setup violation. As a result output flag was raised and consequently delay was measured at the output of the toggling flop.

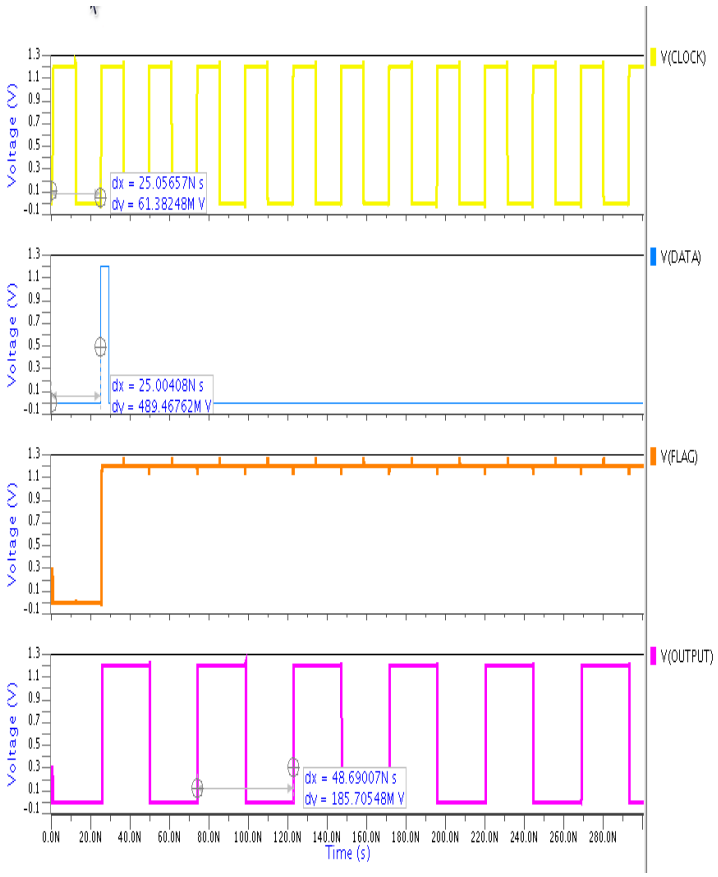

Figure 39: ON Chip Delay Measurement

5 CONCLUSIONS AND FUTURE WORK

The concept of Power dissipation in the CMOS circuits has been discussed in detail in this dissertation. Various techniques to minimise static and dynamic power have also been briefly reviewed. Timing violations have been briefly described and popular circuits to detect them have been reviewed. Problems with these circuits have forced designers to implement novel solutions with regard to area and power conservation. A new Razor flop utilising transition detector and buffer cell to detect timing violation has been proposed in this dissertation.

The necessity of " On chip Delay Measurement" in wake of problems such as accessibility, loading, manufacturing cost with regard to off chip measurements, has forced designers to look for techniques to utilize on chip circuit components for delay detection. In this regard, a new technique based on Razor flop has been proposed. This technique uses setup time failures to calculate memory access time. The technique has been implemented using basic combinational and sequential circuits in 90nm technology. The entire set up starting from clock generation to on chip delay measurement has been verified and results have been found to be reasonably accurate.

Future work may relate to improvements in resolution, accuracy oscillator and razor flop design

REFERENCES:

[1] N.H.E.Weste, D.M.Harris, "Cmos VLSI Design", 4th edition, Pearson.

[2] N.Azemard and L.Svensson, "Integrated Circuit and System Design", In Proc 17th International Workshop on Power and Timing modelling Optimization and Simulation (PATMOS), Sep. 2007

[3] T.Sato and I.Arita, "Constructive Timing Violation for Improving Energy Efficiency" In: Benini.L, Kandemir.M., Ramanujam.J "Compilers and Operating Systems for Low Power", Kluwer Academic Publishers, Dordrecht (2003)

[4] Sato.T, Kunitake.Y, "A Simple Flip-Flop Circuit for Typical-Case Designs for DFM," 8th International Symposium on Quality Electronic Design (ISQED), pp: 539-544, March 2007.

[5] Shih-Lien Lu, "Speeding up processing with approximation circuits," *Computer* , vol.37, no.3, pp: 67,73, March 2004

[6] N.R.Shanbhag, "Reliable and efficient system-on-chip design," *Computer* , vol.37, no.3, pp.42,50, Mar 2004

[7] A.K.Uht, "Going beyond worst-case specs with TEAtime," *Computer* , vol.37, no.3, pp.51,56, Mar 2004

[8] Ernst, D.; Nam Sung Kim; Das, S.; Pant, S.; Rao, R.; Pham, T.; Ziesler, C.; Blaauw, D.; Austin, T.; Flautner, K.Mudge, T., "Razor: a low-power pipeline based on circuit-level timing speculation,"In Proc 36th Annual IEEE/ACM International Symposium on Microarchitecture,(MICRO-36) , pp.7-18, Dec. 2003

[9] S.Das, D.Roberts, S.Lee, S.Pant, D.Blaauw, T.Austin,K.Flautner, and T.Mudge,"A self-tuning DVS processor using delay error detection and correction," IEEE J. Solid-State Circuits, vol. 41, no. 4, pp. 792–804, Apr. 2006.

[10] S.Das, C.Tokunaga, S.Pant, W.-H.Ma ,S.Kalaiselvan ,K.Lai , D.M. Bull,and D.T.Blaauw, "RazorII: Insitu error detection and correction for PVT and SER tolerance,"IEEE J. Solid-State Circuits,vol.44,no. 1, pp. 32–48, Jan. 2009

[11] K.A.Bowman, J.W.Tschanz, N.S.Kim, J.C.Lee, C.B.Wilkerson, S.-L. L. Lu, T. Karnik, and V. K. De, "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance,"IEEE J.Solid-StateCircuits,vol.44,no.1,pp.49–63,Jan.2009

[12] K.A.Bowman, J.W.Tschanz, N.S.Kim, J.C.Lee, C.B.Wilkerson, S.-L. L. Lu, T. Karnik, and V. K. De, "Energy-efficient and metastability immune resilient circuits for dynamic variation tolerance,"IEEE J.Solid-State Circuits,vol.44,no.1,pp.49–63,Jan.2009.

[13] Inyong Kwon,Seongjong Kim, Fick.D, Myungbo.Kim, Yen-PoChen,Sylvester.D., "Razor-Lite: A Light-Weight Register for Error Detection by Observing Virtual Supply Rails," *Solid-State Circuits, IEEE Journal of* , vol.49, no.9, pp.2054,2066, Sept. 2014 C. Stroud, A Designer's Guide to Built-In Self-Test. Kluwer Academic Publishers, 2002

[14] Jain, A.; Veggetti, A.; Crippa, D.; Rolandi, P., "On-chip delay measurement circuit," *Test Symposium (ETS), 2012 17th IEEE European* , vol., no., pp.1,6, 28-31 May

[15] R.Tayade, J.A.Abraham, "On-chip Programmable Capture for Accurate Path Delay Test and Characterization", Proceedings of International Test Conference, 2008

[16] R. Datta, A. Sebastine, A. Raghunathan and J.A. Abraham. "On-chip Delay Measurement for Silicon Debug", Proceedings of GLSVLSI, 2004, pp. 145-148.

[17] Ming-Chien Tsai, Ching-HwaCheng, Chiou-Mao Yang, "An All- Digital High-Precision Built-In Delay Time Measurement Circuit" Proceedings of 26th IEEE VLSI Test Symposium, 2008, pp. 249-254.

[18] Ming-Jun Hsiao, Jing-Reng Huang,Shao-Shen Yang Tsin-Yuan Chang, "A Built-in Timing Parametric Measurement Unit", Proceedings of International Test Conference, 2001,pp. 315-322.

[19] M. Collins, B.M. Al-Hashimi and P.R. Wilson, "On-chip timing measurement architecture with femtosecond resolution", IEEE ELECTRONICS LETTERS 27th April 2006 Vol. 42 No. 9.

[20] Abhishek Jain, Andrea Veggetti, Dennis Crippa, Pierluigi Rolandi. An On-Chip Flip-Flop Characterization Circuit. In Proceedings of PATMOS'2010. pp.41-50

[21] Poki Chen, Shen-Iuan Liu, and Jingshown Wu "A CMOS PulseShrinking Delay Element For Time Interval Measurement", IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: ANALOG AND DIGITAL SIGNAL PROCESSING, VOL. 47, NO. 9, SEPTEMBER 2000, pp 954-958

[22] Collins, M., Al-Hashimi, B.M., and Ross, N.: 'A programmable time measurement architecture for embedded memory characterization'. Proc. of 10th IEEE European Test Symp. (ETS'05), 2005, pp. 128–133.

[23] S. Ghosh, S. Bhunia, A. Raychowdhury and K. Roy, "A Novel Delay Fault Testing Methodology Using Low-Overhead Built-In Delay Sensor", IEEE Transactions on Computer-Added Design of Integrated Circuits and Systems, Vol.25, No.12, pp. 2934-2943, Dec. 2006