LOW OVERHEAD WARNING FLIP-FLOP BASED ON CHARGE SHARING FOR TIMING SLACK MONITORING

[1]Dr. John Paul Pulipati,Principal, [2]Dr.M.N.Yadav, [3]Dr.Purushotham Naik, [4]P.Venkatapathi, [5]M.Shiva Kumar,
[1]principal@mrce. in,Malla Reddy College of Engineering
[2]Assistant Professor, Dept. of ECE,Malla Reddy College of Engineering
[3]Assistant Professor, Dept. of ECE,Malla Reddy College of Engineering
[4]Assistant Professor, Dept. of ECE,Malla Reddy College of Engineering
[5]Assistant Professor, Dept. of ECE, Malla Reddy College of Engineering
1

Abstract— Timing error predictors have a strong potentialto reduce the worst case timing margins by monitoringtiming slack of a design. However, these timing error predictors incur substantial amount of silicon area and power which limit the overall benefits in the system level. This paper presents a low overhead warning flip-flop (FF), which predicts setup time violations. It consists of a delay buffer and a warning generator along with a conventional master-slave FF. Low overhead FF can be designed by exploiting the concept of charge sharing to implement the warning generator. As the warning generator requires only seven transistors to predict the timing violation, the proposed warning FF occupies 30% less area and consumes27% less power compared to the state-of- the-art timing error predictors. A test chip is fabricated using the proposed FF in a 130-nm CMOS technology to verify the functionality of the proposed warning FF in dynamic voltage and frequency scaling applications. Measurement results from the test chip show that a performance improvement of 44% can be achieved at a supply voltage of 0.9 V by employing the proposed technique compared to the worst case design. For a typical chip, the power consumption can be reduced by 36% compared to the worst case design.

Index Terms— Charge sharing, dynamic frequency scaling (DFS), dynamic voltage scaling (DVS), setup time violation, timing error predictors, timing margins.

I. INTRODUCTION

PROCESS, voltage, and temperature (PVT) variations in scaled technology nodes cause significant performance uncertainty in the digital designs. Timing or supply volt- age guard bands are added to maximum operating fre- quency (MOF) or minimum supply voltage to cope with PVT variations. However, these guard bands severely limit the performance and/or increase the power consumption of a design in the typical or best conditions. Moreover, transistor aging degrades the performance of a design with time. Hence, guard bands need to be added considering the life time of the design. As a result, traditional worst case design methodology is not suitable to implement energy-efficient designs as large guard bands are required in nanometer technology nodes. This motivates to implement design methodologies which can reduce the guard bands.

Traditionally, process monitors have been proposed to mon- itor the manufacturing process condition of a chip. In these techniques, body biasing is used to change the thresh- old voltage of the transistors based on process conditions. However, these techniques can address only global process variations but not local process and dynamic variations. Critical path replica circuits [1]–[4] have been explored to track the delay of the critical paths of a design. In this approach, a replica

DOI:10.21276/ijcesr.2017.4.12.29

circuit which is strongly correlated with the actual critical path of the design is monitored to observe delay variations. This method can address global varia- tions but not the local variations due to mismatch in the delay of the actual critical path and replica path. Moreover, the activation of critical path depends on the input data pattern.

On the other hand, in situ timing error monitoring tech- niques [5]–[8] can address both local and global variations by monitoring the timing slack. These techniques directly monitor the output of combinational logic using a specialized flip- flop (FF). An error signal is flagged in the case of timing violation. Hence, the supply voltage or frequency of design can be altered by monitoring the error signal. These techniques are mainly classified into two categories: error detectors [5],

[6] and error predictors [8], [9]. Error detectors such as Razor I [6] and Razor II [7] detect the timing errors after their occurrence and correct the timing violations using architectural replay mechanism [7]. However, error detectors introduce a significant minimum path delay constraint which causes large area overhead because of buffer insertion. Moreover, architectural replay mechanism employed to correct timing violations is available in high- performance processors but not in application-specific integrated circuits (ASICs). Bubble Razor [10] does not incur area overhead for short path padding. However, this technique uses latches in the pipelines instead of FFs.

Error predictors [11], [12] flag a warning signal before the occurrence of timing violations by monitoring the delayed data. As the output of FF is always correct, these techniques do not incur correction overhead. Error predictors are suitable to implement ASICs as they do not require a correction mechanism. However, these techniques can monitor only gradual change in the delay of the critical paths. Canary FF [8], [9], [11] falls into this category. Canary FF employs a double-sampling architecture to predict the timing violations. Canary FF incurs large area and power overhead because of the shadow FF and delay buffers. Moreover, output of the shadow FF may enter into metastable state.

In [13], an aging sensor has been proposed, which uses delayed clock to create a guard band interval before the rising edge of the clock. As the guard band interval is created using the delayed edge of the previous clock cycle, the design is a function of operating frequency. An alternative version is also proposed in [13], which uses double-sampling architecture similar to canary FF. A timing monitoring circuit presented in [14] uses a sensor to predict the timing violations and a warning window generator to create a detection window. In this approach, the detection window is generated by using a transition detector in the clock tree which makes the clock tree implementation difficult.

A warning detection sequential in [12] and [15] observesthe delayed data transition during a warning interval to predict timing violations. This FF is not area and power efficient as a large number of delay buffers are required to delay the input data and to implement the edge detector. In [16], a sensor has been proposed which monitors the delayed master output to predict the timing violations. It requires a large area and power as it uses double-sampling architecture. A pre-error FF in [17] monitors data transition during the negative half cycle of the clock cycle. As warning margin is determined by negative half cycle of the clock, the design requires a clock of more than 50% duty cycle for better performance of the design.

The pulse-based error predictor in [18] monitors delayed

master latch output in the high phase of the clock. However, edge detector used in this design requires more area and consumes more power. In situ monitors proposed in [19]–[21] observe master latch output and employ double-sampling architecture similar to the canary FF [8]. A timing error predictor in [22] employs double-sampling architecture similar to the canary FF except that a tunable delay buffer is used to delay the data signal. Double-sampling architectures require significant area and power overhead. A current-based timing error detector is proposed in [23]. A nine transistor transition detector [24] is designed to operate at supply voltage range 0.44–1.1 V for low-power applications.

Critical path identification becomes difficult with increased variations in nanometer technology nodes [25]. Because of this, more number of critical paths have to be monitored using the timing error predictors to avoid functional failure. Hence, this paper presents a low area and power overhead timing error predictor for timing slack monitoring. In addition, the proposed warning FF can be used as an aging sensor similar to the work reported in [13] and [26].

The remainder of this paper is organized as follows.

The operation of the proposed warning FF is discussed in Section II. Chip implementation details are described in Section III. Measurement results are presented in Section IV. Section V discusses the comparison of timing error predictors and simulation results. Finally, this paper is concluded in Section VI.

II. PROPOSED CHARGE SHARING-BASED WARNING FLIP-FLOP

Warning FF is used to predict the setup time violation. Warning FF flags a warning signal, if data transitions at the input of FF happen during a timing window before the rising edge of the clock. Traditionally, either the input data of FF [11], [15] or master latch output [18], [20] is monitored to predict the setup time violations. In both approaches, the concept of delayed data is employed to predict the timing violations. Moreover, these approaches use either double- sampling technique or transition detection technique to flag warning signal. Based on these, warning FFs can be classified into four categories:

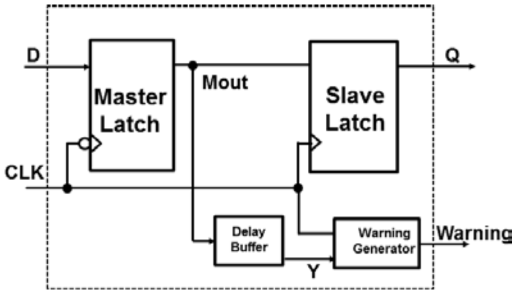


Fig. 1. Proposed warning FF.

1) samples at input data D of FF and uses double-sampling technique [11];

2) samples master latch output Mout and uses double- sampling technique [20];

3) samples at input data D of FF and uses transition detection technique [15];

4) samples master latch output Mout and uses transition detection technique [18].

If the input data of the FF is monitored to flag warning signal, the delay of buffer should account for both setup time of FF and warning margin. However, if the master latch output is monitored to flag warning signal, the delay of buffer should account for warning margin only as the master latch delay already accounts for the setup time of FF.

The proposed warning FF is based on monitoring the delayed data at the output of master latch. It consists of a delay buffer and a warning generator along with a conventional master-slave FF to predict the setup time violations. The delay of the buffer is the warning margin of the proposed FF. If a data transition happens during a timing interval equal to the warning margin before the setup time window of the master latch, the proposed FF flags the warning signal. The warning signal is an active low signal. In the case of early data arrival at the input of FF, the warning generator output is logic high. However, in the case of late data arrival, the warning generator output is logic low. The schematic of the proposed warning FF is shown in Fig. 1.

A. Operation of the Proposed Flip-Flop

A data transition at the input of the FF can happen in either of the four timing windows shown in Fig. 2. A data transition during the window 1 is treated as an early data arrival. In this case, the available timing slack is very high. A data transition during window 2 is treated as late data arrival. In this case, the available timing slack is less. A data transition in this window means that a timing violation may occur if the delay of the critical path is further increased due to PVT variations. A data transition in timing window 3 leads to setup violation,

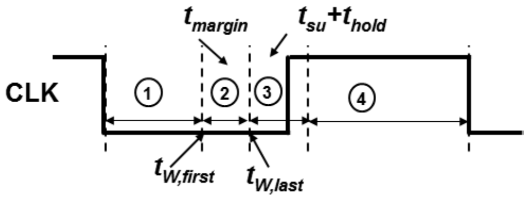


Fig. 2. Timing windows for possible data transition

TABLE I
OPERATION OF THE PROPOSED WARNING FF

Timing Window	Data Transition	Before the rising edge of clock			After the rising edge of clock			Q	Warning
		D	Mout	Y	D	Mout	Y		
1	Rise	1	1	1	X*	1	1	1	1
	Fall	0	0	0	X	0	0	0	1
2	Rise	1	1	0	X	1	1	1	0
	Fall	0	0	1	X	0	0	0	0
3	Rise or Fall	X	PS**	PS	X	PS/D	PS/D	PS/D	1/0
4	Rise or Fall	X	PS	PS	X	PS	PS	PS	1

and the output of master latch may enter into metastable state. So, the maximum delay constraint of the design should satisfy (1), so that the data transitions under PVT variations can happen before the timing window 3. The maximum delay constraint is given by

Tmax ≤ TCLK − tClktoQ − tsu − tmargin (1) where Tmax is the maximum combinational path delay,

TCLK is the clock period, tClktoQ is the previous stage FF clock to Q delay, tsu is setup time, and tmargin is warning margin. A data transition during the timing window 4 does not appear at the output of master latch as the master latch is opaque in this window.

The operation of the warning FF is given in Table I. If a data transition occurs during the timing window 1, the master latch output (Mout) and delayed master latch output (Y) (refer to Fig. 1) are the same in both low and high phases of the clock. However, if a data transition happens in the timing window 2, the delayed master output (Y) is different from Mout in low and high phases of the clock. In this case, the warning signal becomes low. However, in both the cases, the FF samples the correct value. When a data transition happens during the timing window 3, the master latch output (Mout) may enter into metastable state and it can resolve to either logic "0" or logic "1". Hence, the delayed master output (Y) may store the present data (D) or the previous state of FF. The conceptual timing diagram of the proposed warning FF is shown in Fig. 3.

In clock cycle I, the data transition happens early, before the warning margin. The master latch samples the correct value. The delayed master latch output (Y) makes a transition during the negative half cycle of the clock represented as A1. In this case, delayed master output (Y) is the same in both the low and high phases of the clock. Hence, the warning generator output is logic high signal. The delay between the input of FF (D) and the master latch output (Mout) is the setup time of master latch. In clock cycle II, data transition happens during the warning margin interval, the master latch samples the valid data represented as A2. However, the delayed master latch output makes transition during the high phase of the clock represented as A3. In this case, the delayed master output is different in high and low phases of the clock. So, the warning generator output becomes low represented as A4. Even the warning signal is flagged, the output of FF is always correct represented as A5. So, there is no correction overhead with this approach. In clock cycle III, the data transition happens during the setup time window of the master latch. This data transition does not appear at the output of master latch

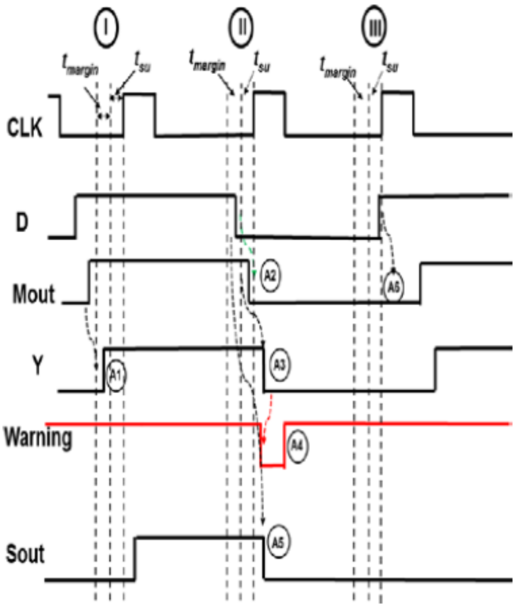


Fig. 3. Conceptual timing diagram of the proposed warning FF.

represented as A6. Hence, the proposed FF does not detect data transitions happened during this window. Similarly, data transitions during the high phase of clock will not appear at the master latch output as it is opaque during the high phase of the clock. Hence, the proposed FF does not suffer from short path problem. The minimum delay constraint of the design is given by

Tmin ≥ thold − tClktoQ (2)

where Tmin is the minimum combinational path delay, tClktoQ is the minimum clock-to-Q delay of the previous stage FF, and thold is the hold time of FF.

The detectable slack of the proposed warning FF is defined as the difference between the point-of-first-warning (POFW) (tW,first) and the point-of-last-warning (tW,last) (refer to Fig. 2). Ideally, the detectable slack is the delay of buffer, which is used to delay the master latch output. Hence, the detectable slack is given by

td,slack = td,buffer (3)

where td,slack is the detectable slack and td,buffer is the delay of the buffer.

B. Warning Generator

The warning generator is designed using the concept of charge sharing. It requires only seven transistors to detect a transition during the high phase of the clock. The schematic of the warning generator is shown in Fig. 4. The delayed master output drives two cascaded-clocked inverters. During the low phase of the clock, these inverters behave like conventional inverters and the intermediate node X is discharged to logic zero. However, during the high phase of the clock, node X becomes floating as the transistor M1 is OFF. The warning signal is at logic high state. Whenever the input of warning generator is different in low and high phases of the clock, the intermediate node X charges from the nodes Z1 and Z2 for rise and fall data transitions, respectively. However, the charged voltage at intermediate node X is less than supply voltage for rise and fall transition at node Y. The inverter INV1 should treat the charged voltage at node X as logic high-input voltage. For this, the switching threshold of inverter should be less than charged voltage at node X. The inverter INV1 is skewed to change the switching threshold. The switching threshold of the inverter is less than half of the supply voltage.

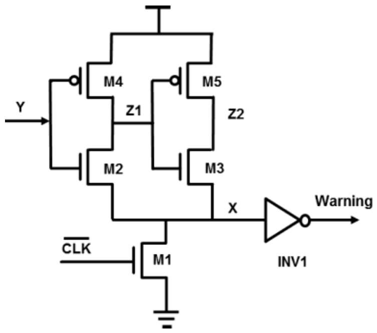


Fig. 4. Schematic of the warning generator

TABLE II

OPERATION OF WARNING GENERATOR

Timing Window 1	Data Transition	CLK=0				CLK=1				Warning
		Node Y	Node Z1	Node Z2	Node X	Node Y	Node Z1	Node Z2	Node X	
2	Rise	1	0	1	0	1	0	1	0	1
	Fall	0	1	0	0	0	1	0	0	1
3	Rise	0	1	0	0	1	Dischraged	Charged	Charged	0
	Fall	1	0	1	0	0	Charged	Dischraged	Charged	0
4	Rise	0	1	0	0	0	1	0	0	1
	Fall	1	0	1	0	1	0	1	0	1
	Rise	0	1	0	0	0	1	0	0	1
	Fall	1	0	1	0	1	0	1	0	1

The operation of the warning generator is given in Table II. The delayed master output (Y) depends on timing window (refer to Fig. 2) in which input data (D) transition happens. An input data (D) transition may lead to: 1) a transition at Y during the low phase of the clock; 2) a transition at Y during the high phase of the clock; and 3) no transition until the next low phase of the clock. A transition at Y during the low phase of the clock occurs, if the data (D) transition happens in timing window 1. A transition at Y during the high phase of the clock occurs, if the data (D) transition happens in timing window 2. For a data transition in timing windows 3 and 4, no data transition happens at Y until the next low phase of the clock. The detailed operation of the warning generator is discussed in Sections II- B.1–II-B.3.

1) Data (Y) Transition During Clock Low Phase: If a signal transition at Y occurs during the low phase of the clock, the state of the node Y is the same in both low and high phases of the clock. So, the intermediate nodes X, Z1, and Z2 do not get affected after the rising edge of the clock. When CLK = 0, the intermediate node X is discharged. Hence, the warning signal is high.

2) Rise Transition (Y) During Clock High Phase: For a rise transition at the input of the

warning generator (Y) in the high phase of clock, the warning generator input is initially logic zero. A logic zero input during the low phase of clock, charges the intermediate node Z1 and discharges the intermediate node Z2. The initial states of nodes X, Z1, and Z2 are shown in Fig. 5(a). The transistor M1 is ON, during the low phase of the clock. So, the node X is discharged to logic zero. In addition, the nodes Z1 and Z2 are logic high and low, respectively. After a rise transition at the input of warning generator in the positive half cycle of clock, the states of nodes X, Z1, and Z2 are shown in Fig. 5(b). In this case, the transistor M1 is off. So, the intermediate node X is charged through the transistor M2. Because of charge sharing between nodes X and Z1, the voltage at node X starts increasing toward logic high. If this voltage is more than the switching threshold of the inverter, the warning signal becomes low. In this way, a rise transition at delayed master output is detected.

3) Fall Transition (Y) During Clock High Phase: For a fall transition at Y in the high phase of the clock, the delayed master output should be logic high in the negative half cycle of the clock. The initial states of nodes X, Z1, and Z2 are shown in Fig. 5(c). The transistor M1 is ON during the low phase of the clock, so the warning signal is high during this interval. The intermediate nodes Z1 and Z2 are logic low and high, respectively. After a fall transition in the high phase of the clock, the node X is charged from node Z2 through transistor M3. This charge sharing leads to a rise transition at node X. The output of warning generator becomes low. The states of nodes X, Z1, and Z2 in the high phase of the clock are shown in Fig. 5(d).

III. CHIP IMPLEMENTATION DETAILS

A two-stage pipelined design is implemented to verify the functionality of the proposed warning FF. The internal blocks of the test chip is shown in Fig. 6. The implemented blocks are divided into three sections: 1) input section; 2) pipelined design; and 3) output section. The input section consists of a serial-in parallel-out (SIPO) shift register and a toggle circuit. The pipelined design section consists of two- stage pipelined circuit with the proposed warning FF and an AND tree to group the warning signals. The output section consists of a parallel-in serial-out (PISO) shift register.

A 13-bit SIPO shift register is used in the implementation of the design. Out of 13 bits, 8 bits for input data of the pipelined design, 1 bit for enable signal of the toggle circuit, 2 bits for selecting the critical path delay, and 2 bits for selecting the warning margin of the warning FF are assigned. To check the functionality of the proposed warning FF on chip, the critical path needs to be activated in each clock cycle, which can be done by changing the inputs of the pipelined design using a toggle circuit. The data input to the design is given from an SIPO shift register through the toggle circuit. Toggle circuit has an enable signal. If the enable signal is high, the inputs of the design will change in every clock cycle which is required to evaluate the warning FF in the pipelined design. In this case, the input data pattern stored in the SIPO register can be inverted in each clock cycle using the togglecircuit and is provided to the pipelined design. As all the paths in the pipelined design are made critical at the design time, all the paths have 100% activity during run time.

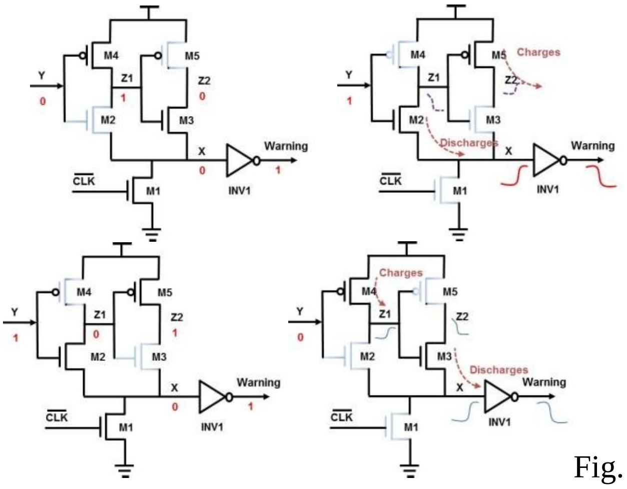


Fig. 5. Operation of warning generator for data transitions in timing window 2. (a) Before rise transition. (b) After rise transition. (c) Before fall transition. (d) After fall transition.

Fig. 6. Details of chip implementation.

Each stage of the pipelined design has eight paths which are designed using inverter, NAND

and NOR chains. The proposed warning FF is inserted at the end points of all the critical paths. The length of the critical path can be modified inside the chip using a configurable delay chains with a 4:1 multiplexer in the critical path. Hence, this scheme allows operating the same design at different clock frequencies. The warning signal from all the warning FFs are grouped into a single warning signal using an AND tree. An inverter converts the AND-tree output to an active high warning signal WARNING as the warning signal in the FF level is active low. The delay of the AND tree should be less than the minimum pulsewidth of the warning signal of the proposed FF and is given by

td,AND-tree = TPW,min (4)

where td,AND-tree is the delay of AND tree and TPW,min is the minimum pulsewidth of the warning signal of the proposed FF.

An 8-bit PISO shift register is used to load the outputs of the pipelined design. The PISO shift register has two external inputs: one is the clock (OSR_CLK) to the shift register and the other is a control signal (LDEN) to enable the shift register to load from the pipelined design and to shift the data of the shift register. If the LDEN is high, the outputs of the pipelined design are loaded into PISO shift register in every clock cycle.

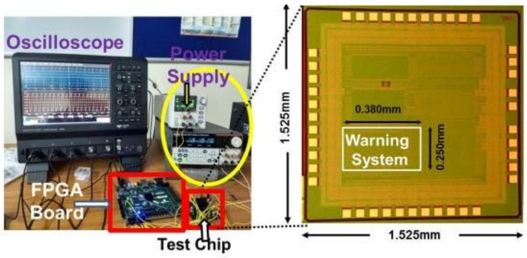


Fig. 7. Experimental setup.

If the LDEN is low, the stored data of PISO shift register is serially shifted out in every clock cycle.

IV. MEASUREMENT RESULTS

The experimental setup to measure the test chip is shown in Fig. 7 which shows a test chip implemented in 130-nm CMOS technology, mounted on QFN-48 socket, a field-programmable gate array (FPGA) board and an oscil- loscope. The SIPO shift register inside the chip is programed through the FPGA Board (ZedBoard) to provide the input data. The serial output of the PISO register is monitored continuously. In the case of warning, all the outputs of the design are verified by shifting the PISO register.

The frequency of baseline or worst case design with

10% voltage drop, a temperature of 85 °C, and 2-sigma process variations is obtained by using the method of

[22].

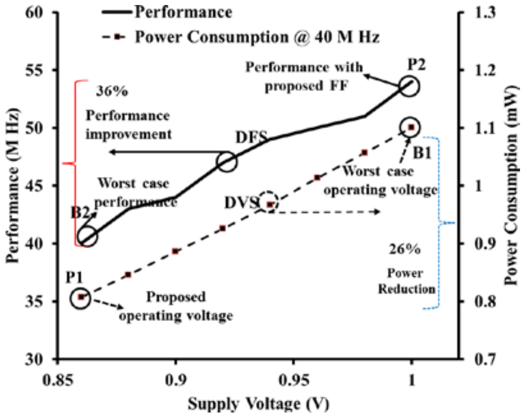


Fig. 8. Performance and power consumption of the design.

The frequency of the baseline design is 40 MHz for a supply voltage of 1 V. The conventional operating voltage and baseline frequency of the design are represented as B1 and B2, respectively, in Fig. 8. However, by employing the proposed FF in the pipelined design with dynamic voltage scaling (DVS), the pipelined design can operate at a reduced voltage represented as P1 for the same performance which results in 26% savings in the power consumption. For dynamic frequency scaling (DFS), the supply voltage is kept constant and frequency of the design is increased. The pipelined design with the proposed FF and DFS can operate at an increased frequency represented as P2 which leads to 36% performance improvement at a supply voltage of 1 V. This shows the advan- tage of using the proposed FF in reducing the timing or voltage guard bands.

A. Impact of Supply Voltage Scaling

Conventionally, the design will operate at a supply voltage of 1.2 V considering worst case margins. However, the pro- posed warning FF can detect the actual operating condition which in turn can reduce the worst case guard bands. Initially, the designs are operated at a supply voltage of

1.2 V with an operating frequency of 50 MHz. The supply voltage of the design is reduced until the warning signal is asserted. The supply voltage for which warning signal becomes high is defined as POFW voltage. If the supply voltage of the design is further reduced, the warning signal is still high and a functional error may occur in the design. A system failure is verified by monitoring the outputs of the design using PISO shift register. The voltage for which a functional failure occurs is known as point-of-first-failure (POFF) voltage. The difference between the POFF and POFW voltages is the available margin for warning. This warning margin is a function of number of delay buffers used to delay the master latch output of the proposed warning FF.

The POFW voltage is measured for 26 chips by reducingthe supply voltage manually in steps of 1 mV until the warning signal becomes high. As the implemented test chip does not include an adaptive controller, the supply voltage is reduced manually while monitoring the warning signal.

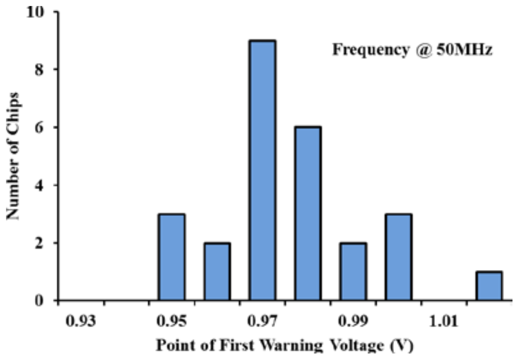


Fig. 9. POFW voltage of measured chips.

However, an external controller used in [15] or on- chip controller proposed in [22] can be employed to change the voltage or frequency of the design. Fig. 9 shows histogram of the POFW voltage for 26 measured chips operating at 50 MHz and at a temperature of 25 °C. It is evident from the POFW voltage of measured chips that the design can operate at reduced supply voltage for a given performance compared to a worst case design. For a typical chip among the measured chips, the supply voltage can be reduced to 0.97 from 1.2 V. However, the minimum voltage required for the worst chip among the measured chips is

1.015 V. Among 26 chips,14 chips can operate at a supply voltage less than 0.97 V for the same performance. Moreover, the best chip among the measured chips requires a supply voltage of 0.94 V, which is 75 mV less than that of the worst chip. This reduction in sup- ply voltage results in lower power consumption, and it shows the significance of warning FF for low-power applications.

To measure the dynamic power consumption of the design (PDesign), the following steps are used. First, the power consumption of the design along with input- output (IO) cells (PD_IO) is measured. Then, the power consumption of only IO cells (PIO) is measured by not applying the clock to design. However, three IO cells (clock, output, and warning signals of design) are active, only when the design is working. So, the power consumed by IO cells (PIO) does not include the dynamic power of these three IO cells. So SPICE simulations are carried out to find the dynamic power of these IO cells (PSIM_IO). To increase the accuracy in the measurement of power consumption, the dynamic power of three IO cells (PSIM_IO) is subtracted. The dynamic power consumption of the design is given by

PDesign = PD_IO − PIO − PSIM_IO. (5)

The abovementioned method is used to measure the power consumption in this paper as the same supply voltage is con- nected for IO cells and the proposed design. The implemented design is having a dedicated clock signal. The histogram of dynamic power consumption of the design for 26 measured chips is shown in Fig. 10. The chips are operated at the POFW to measure the power consumption of the design. The maximum power consumed by the design is

1.35 mW, whereas the minimum power consumed by the design is 1.20 mW.

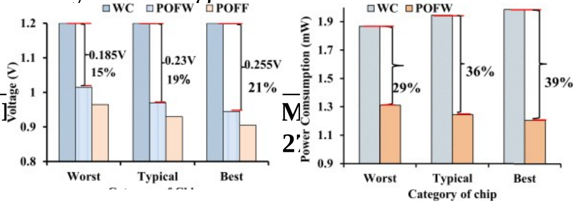


Fig. 11. Comparison of worst, typical, and best chips among the 26 measured chips. (a) POFW voltage. (b) Power consumption.

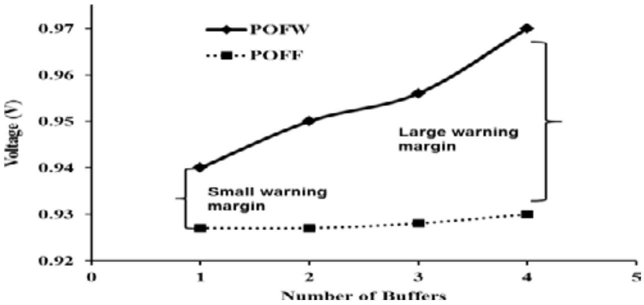
The power consumption of a typical chip among the measured chips is 1.26 mW. Among 26 measured chips, 15 chips can work with a power consumption less than 1.26 mW.

The POFW and POFF voltages for three different categories of chips are shown in Fig. 11(a). The POFW voltage of the worst and typical chips, out of measured chips is 185 and230 mV less than the nominal supply of 1.2 V, respectively. The power consumption of the proposed design with DVS is measured by operating the design at a POFW voltage. For a typical chip, the power consumption is reduced by 36%, whereas for the best chip, the power consumption is reduced by 39% using the proposed warning FF with DVS. The power consumption of three different categories of chips is shown in Fig. 11(b).

B. Impact of Delay Buffers on POFW Voltage

In the proposed approach, a tunable delay buffer used in [15] is employed to change the warning margin of the proposed FF. This tunable delay buffer can be used to provide different warning margins for varying process conditions. A process monitor can be used to detect global process conditions. The warning margin of the proposed FF is varied using the select lines of the multiplexer. In this case, a 4:1 multiplexer is employed for selecting the warning margin.

For a select input of 00, the critical path delay is minimum and POFW occurs at a voltage of 0.94 V for a typical chip. In this case, maximum power savings can be achieved. However, the design can tolerate less delay variations as the warning margin is small. For a select input of 11, the warning margin is maximum and the POFW occurs at a voltage of 0.97 V for a typical chip. However, in this case, the design can tolerate large delay variations as the warning margin is maximum. In all the cases, the POFF voltage is the same, as it is determined by the setup time of the FF. Fig. 12 shows the POFW and POFF voltages for a typical chip as a function of delay buffers.



C. Impact of Frequency Scaling

The frequency of baseline or worst case design at a supply voltage can be obtained by using the method of [22]. First, the MOF of each chip is measured at room tem- perature with 10% drop in the supply voltage. Second, Gaussian fitting of measured MOF of all chips is used to find the mean (μ) and standard deviation (σ). Then,2-sigma frequency (2σ) and 5% of the mean value are reduced from the mean value to account for process variations and temperature margin of 60 °C, respectively. As the design is measured at room temperature, the timing margin to account for the temperature of 85 °C is found by SPICE simulations. It is found from the simulation results that the performance of the design at a temperature of 85 °C is degraded by 6.4% and 5% at a supply voltage of

1.05 and 0.9 V, respectively, compared to performance at room temperature (25 °C). For an operating voltage of 1 V, the MOF of the measured chips at room temperature with 10% voltage drop is shown in Fig. 13. The mean and standard deviation are found to be 45 and1.7 MHz, respectively. The baseline frequency is 40 MHz atan operating voltage of 1 V. Similarly, the baseline frequencies for supply voltage of 0.9, 0.95, and 1.05 V are 31, 35, and44 MHz, respectively.

The MOF of three categories of the chips (worst, typical, and best) is shown in Fig. 14.

Out of 26 measured chips, the chip which flags a warning signal for the lowest frequency at a fixed voltage is treated as the worst chip. Similarly, the chip which flags a warning signal for the highest frequency at a fixed voltage is treated as the best chip. The best chip can perform 23% better than a worst chip at a supply voltage of 0.9 V. The same chip can perform 14% better than a worst chip at a supply voltage of 1.05

V. This shows the advantage of using warning FF at lower supply voltage. The proposed warning FF can improve the performance of the design by 44% and 31% compared to baseline design at a supply voltage of 0.9 and 1.05 V, respectively, under typical operating conditions.

Fig. 14. Performance improvement of worst, typical, and best chips at room temperature.

. SIMULATION RESULTS

Section IV discussed the advantages of using the proposed warning FF. The implementation details of warning FF, such as the maximum charged voltage (MCV) at the intermediate node, and sizing of the transistors in the warning generator are discussed in this section. Moreover, the FF-level comparison is discussed to quantify area and power savings with respect to the warning FFs available in the literature. The proposed FF is implemented in industrial 130-nm CMOS technology. Postlayout simulations on parasitic extracted netlist have been carried out using HSPICE.

A. Warning Generator

The warning generator (refer to Fig. 4) is designed using the concept of charge sharing. The MCV at node X of the warning generator and switching threshold of the skewed inverter are critical factors for the proper functionality of warning FF. To verify the MCV at node X for data transitions at the input of warning generator, 10 000 Monte Carlo simulations have been performed with 3-sigma process variation for a supply voltage range 0.7–1.2 V at a temperature of 25 °C. The minimum value of MCV is used to determine the switching threshold of the skewed inverter INV1 (refer to Fig. 4). The MCV at node X and switching threshold of the skewed inverter (INV1) are shown in Fig. 15.

Post layout simulations on parasitic extracted net list have been carried out to verify the operation of the proposed

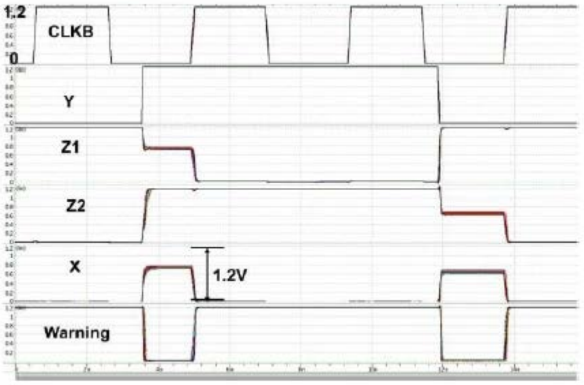
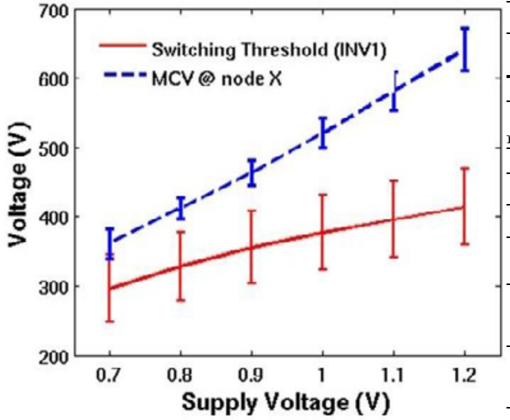


Fig. 16. Timing diagram of the warning generator.

warning generator. Fig. 16 shows the timing diagram of CLKB (inverted clock), input of warning generator (Y), warning signal, intermediate nodes X, Z1, and Z2 for rise and fall transitions. For both rise or fall transitions at node Y, when CLKB = 0, the intermediate node X is charged to voltage greater than half of the supply voltage. For a rise transition, the intermediate node Z1 is discharged to intermediate value, which charges the node X and the node Z2 is charged to supply voltage. Similarly, in the case of fall transition, node Z2 is discharged to an intermediate value which charges the node X and node Z1 is charged to supply voltage. In both cases, the warning signal is active low pulse. The delay of the warning generator is increased by 5× at a supply voltage of 0.7 V compared to the delay of



Parameter	Proposed	Ref [11]	Ref [15]	Ref	Ref [18]	DFF
			3.1×		2.81×	1×
Nu N			3	1	2	NA *
			FF input	Master latch output	Master latch output	NA
		ng	Transition detection	Double sampling	Transition detection	NA
C			1.01×	1.03×	0.97×	1×
			77	117.8	187.76	76
			1.9	-13.2	-17	2
Cl			4.05	8.49	4.04	3.97
/ (N			15.6	19.2	15.8	7.32
/ (			17.1	19.2	17.3	NA
Peak Power (No warning)	1.14×	2.45×	1.12×	2.42×	1.82×	1×
Peak Power (Warning)	1×	1.66×	1.52×	1.81×	1.33×	NA

warning generator at 1.2 V. The variability of the delay of warning generator is 8.7% and 32% at supply voltages of 1.2 and 0.7 V, respectively, with 3- sigma process variations (including both within-die variation and die-to-die variation).

B. Flip-Flop-Level Comparison

To quantify the area and power savings of the proposed warning FF compared to existing FFs, the canary FF [11], warning detection sequential [15], in situ monitor [20], and pulse-based timing error predictor [18] are implemented in industrial 130-nm CMOS technology. The comparison of different warning FFs is given in Table III.

1) Area: The proposed warning FF requires 1.95× area compared to the conventional FF. However, the proposed FF occupies 30% less area compared to [18]. This area savings are due to the requirement of less transistors to implement the warning generator. Moreover, the approach in [18] requires two delay buffers, one for warning margin and the other for transition detection unlike the proposed warning FF which requires only one delay buffer to provide warning margin. The area of canary FF [11], warning detection sequential [15], in situ monitors [20] and pulse-based predictor [18] is 3.45×,

×, 3×, and 2.81×, respectively, compared to that of conventional FF as given in the second row of Table III. The large area overhead in [11] and [20] is due to the usage of double-sampling architecture. Moreover, the canary FF

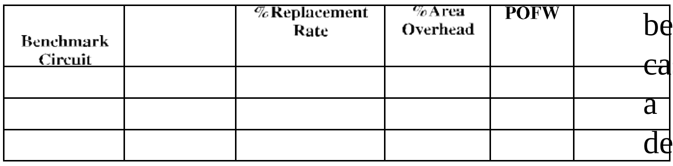
[11] requires extra delay buffers to account for the setup time of the main FF as it monitors at the input of FF.

2) Average Power: The proposed warning FF consumes 1.57× more power compared to the conventional FF. However, the proposed FF consumes 27% less power compared to the

pulse-based timing error predictor [18] with 50% data activity at an operating frequency of 250 MHz. The reduction in power consumption is due to the less number of the transistorsin the warning generator. Moreover, clock power of the proposed warning FF is just 1.16× to that of the conventional FF with 0% data activity. However, Canary FF [11] and in situ monitor [20] consume 2.14× and 2.13× more power, respectively, compared to the conventional FF with 0% data activity. This increased power consumption in [11] and [20] is due to the usage of more number of clocked transistors in the double-sampling architectures. In the proposed warning FF, the power consumption is more with the warning signal asserted compared to no warning condition. However, Canary FF consumes less power with warning signal asserted because of the reduced data activity at shadow FF compared to no warning condition. The power consumption due to the warning window generator of [15] is not taken into account for the comparison as it can be shared among warning FFs.

3) Peak Power: The peak power of the proposed FF is measured for different data to clock durations. When the data transition happens during the warning margin window, the peak power is higher compared to peak power due to a data transition before the warning margin window. The peak power of the different FFs is given in Table III. The double-sampling techniques [11], [20] consume higher peak power compared to transition detection approaches [15], [18]. The proposed FF consumes minimum peak power, out of the compared FFs, when the warning signal is flagged. When there is no warning signal, the proposed FF consumes 14% extra peak power compared to the conventional DFF. The warning detection sequential of [15] consumes 12% extra power compared to the conventional DFF.

TABLE IV
POWER SAVINGS OF ISCAS89 BENCHMARK CIRCUITS



Benchmark Circuit		%Replacement Rate	%Area Overhead	POFW	

C. ISCAS89 Benchmark Circuits

The ISCAS89 benchmark circuits [27] are implemented to operate under worst case conditions with slow process corner, 125 °C temperature, and 10% drop in the supply voltage. The critical paths with slack less than 20% of time period are monitored with proposed FFs. The replacement rate is the ratio of number of proposed warning FFs to the total number of FFs in the design. Initially, the designs are simulated using a gate-level simulator (Synopsys VCS) under typical operating conditions with a nominal supply voltage of 1.2 V at a temperature of 25 °C. The timing libraries for different supply voltages are created using Cadence Liberate. The supply voltage of the design is reduced in steps of 20 mV until a warning signal is flagged. The POFW voltage is found by simulations for each design. From the simulation results, it is evident that by using the proposed FF, the designs can operate at reduced voltage under typical conditions. Power savings up to 22% can be obtained by employing the proposed FF in ISCAS89 benchmark circuits. The power savings obtained using the proposed FF in ISCAS89 benchmark circuits are given in Table IV.

VI. CONCLUSION

DVS and DFS with timing error monitors have become an effective way to reduce the worst case timing guard bands. This paper presented a low overhead warning FF, which monitors delayed master latch output to predict the timing vio- lations. The proposed warning FF consumes 27% less power at 50% activity factor and requires 30% less area compared to the timing error predictors available in the literature. A test chip is fabricated in industrial 130-nm CMOS technology to verify the effectiveness of the proposed warning FF in reducing the timing margin. Measurement of the test chip demonstrates that a design with the proposed warning FF can operate at

44% and 31% higher clock frequency at a

supply voltage

of 0.9 and 1.05 V, respectively, compared to the traditional worst case design in typical conditions. For a typical chip, the power consumption can be reduced by 36% compared to the conventional worst case design.

ACKNOWLEDGMENT

The authors would like to thank Europractice IC service for chip fabrication.

REFERENCES

[1] T. Kuroda et al., “Variable supply-voltage scheme for low-power high- speed CMOS digital design,” IEEE J. Solid-State Circuits, vol. 33, no. 3, pp. 454–462, Mar. 1998.

[2] J. W. Tschanz et al., “Adaptive body bias for reducing impacts of die-to- die and within-die parameter variations on microprocessor frequency and leakage,” IEEE J. Solid-State Circuits, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.

[3] J. T. Kao, M. Miyazaki, and A. P. Chandrakasan, “A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias archi- tecture,” IEEE J. Solid-State Circuits, vol. 37, no. 11, pp. 1545–1554, Nov. 2002.

[4] A. Drake et al., “A distributed critical-path timing monitor for a 65 nm high-performance microprocessor,” in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2007, pp. 398–399.

[5] D. Ernst et al., “Razor: A low-power pipeline based on circuit- level timing speculation,” in Proc. Int. Symp. Microarchitecture, Dec. 2003, pp. 7–18.

[6] S. Das et al., “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.

[7] S. Das et al., “RazorII: In Situ error detection and correction for PVT and SER tolerance,” IEEE J. Solid-State Circuits, vol. 44, no. 1, pp. 32–48, Jan. 2009.

[8] T. Sato and Y. Kunitake, “A simple flip-flop circuit for typical- case designs for DFM,” in Proc. Int. Symp. Quality Electron. Des. (ISQED), Mar. 2007, pp. 539–544.

[9] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, “Adaptive

performance compensation with in-situ timing error prediction for subthreshold circuits,” in Proc. Custom Integr. Circuits Conf. (CICC),2009, pp. 215–218.

[10] M. Fojtik et al., “Bubble Razor: Eliminating timing margins in an ARM Cortex-M3 processor in 45 nm CMOS using architecturally independent error detection and correction,” IEEE J. Solid-State Circuits, vol. 48, no. 1, pp. 66–81, Jan. 2013.

[11]H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, “Adaptive performance compensation with in-situ timing error predictive sensors for subthreshold circuits,” IEEE Trans. Very Large Scale Integr. Syst., vol. 20, no. 2, pp. 333–343, Feb. 2012.

[12] B. P. Das and H. Onodera, “Warning prediction sequential for transient error prevention,” in Proc. IEEE Int. Symp. DFT VLSI Syst., Oct. 2010, pp. 382–390.

[13]M. Agarwal, B. C. Paul, M. Zhang, and S. Mitra, “Circuit failure prediction and its application to transistor aging,” in Proc. VLSI Test Symp. (VTS), May 2007, pp. 277–286.

[14] . Rebaud et al., “Timing slack monitoring under process and envi- ronmental variations: Application to a DSP performance optimization,” Microelectron. J., vol. 42, no. 5, pp. 718–732, May 2011.

[15]B. P. Das and H. Onodera, “Frequency-independent warning detection

sequential for dynamic voltage and frequency scaling in ASICs,” IEEE Trans. Very Large Scale Integr. Syst., vol. 22, no. 12, pp. 2535–2548, Dec. 2014.

[16]T. Azam and D. R. S. Cumming, “Efficient sensor for robust low-power

design in nano-CMOS technologies,” IET Electron. Lett., vol. 46, no. 11, pp. 773–775, May 2010.

[17]M. Wirnshofer, L. Heiß, G. Georgakos, and D. Schmitt-Landsiedel, “A

variation-aware adaptive voltage scaling technique based on in-situ delay monitoring,” in Proc. IEEE Int. Symp. Design Diagnostics Electron. Circuits Syst. (SDDECS), Apr. 2011, pp. 261–266.

[18]L.-Y. Chiou, C.-R. Huang, and M.-H. Wu, “A power-efficient pulse- based in-situ timing

error predictor for PVT-variation sensitive circuits," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), Jun. 2014, pp. 1215–1218.

[19]V. Huard et al., "Adaptive wearout Management with in-situ aging monitors," in Proc. IEEE Int. Rel. Phys. Symp., Jun. 2014, pp. 1–11.

[20]M. Saliva et al., "Digital circuits reliability with in-situ monitors in 28 nm fully depleted SOI," in Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE), 2015, pp. 441–446.

[21] A. Benhassain et al., "Timing in-situ monitors: Implementation strategy and applications results," in Proc. Custom Integr. Circuits Conf. (CICC), 2015, pp. 1–4.

[22] W. Shan, X. Shang, L. Shi, W. Dai, and J. Yang, "Timing error prediction AVFS with detection window tuning for wide-operating-range ICs," IEEE Trans. Circuits Syst. II, Exp. Briefs, to be published.

[23]Y. Zhang et al., "iRazor: 3-transistor current-based error detection and correction in an ARM Cortex-R4 processor," in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2016, pp. 160–162.

[24]X. Shang, W. Shan, L. Shi, X. Wan, and J. Yang, "A 0.44 V-1.1 V 9-transistor transition-detector and half-path error detection tech- nique for low power applications," in Proc. A-SSCC, Nov. 2017, pp. 205–208.

[25]I. Kwon, S. Kim, D. Fick, M. Kim, Y.-P. Chen, and D. Sylvester, "Razor-lite: A light-weight register for error detection by observing virtual supply rails," IEEE J. Solid-State Circuits, vol. 49, no. 9, pp. 2054–2066, Sep. 2014.

[26]M. Omaña, D. Rossi, N. Bosio, and C. Metra, "Low cost NBTI degradation detection and masking approaches," IEEE Trans. Comput., vol. 62, no. 3, pp. 496–509, Mar. 2013.

[27](Nov. 2017). ISCAS'89 Benchmarks. [Online]. Available: http://www.pld.ttu.ee/maksim/benchmarks/iscas89/verilog/