A Methodology for Testing High-Performance Circuits at Arbitrarily Low Test Frequency

Muhammad Nummer and Manoj Sachdev
Department of Electrical and Computer Engineering,
University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1
mnummer@vlsi.uwaterloo.ca, msachdev@ece.uwaterloo.ca

Abstract

This paper presents a methodology for testing high-performance circuits with a low-speed clock in test mode. Using this technique, the frequency of the 50% duty cycle test mode clock can be reduced with virtually no lower limit. This poses very little requirements on automatic test equipment (ATE) and facilitates the testing process. A CMOS implementation that achieves 50ps accuracy is also presented. This technique targets designs using design for testability (DFT) and/or built-in self test (BIST) techniques.

Keywords: Delay-fault testing, high-performance testing, design for delay testability, built-in self test, controlled-delay flip-flop.

1. Introduction

The on-chip clock frequency of high-performance state-of-the-art VLSI CMOS circuits has surpassed 1 GHz. It is expected that the speed of such circuits will continue to increase for future technology generations. The 1999 edition of the International Technology Roadmap for Semiconductors (ITRS) predicts that the on-chip clock frequency will exceed 3 GHz by year 2005 and 13 GHz by year 2014 [1]. With smaller geometries, higher speeds, and increased interconnects, it is more likely for small imperfections in the fabrication process to cause device failure. Consequently, it is expected that the yield of semiconductor processing will deteriorate rapidly to reach 50% by 2014.

According to the ITRS, many of the technology problems causing yield losses and cost increases are related to the slower advances in ATE speeds versus the ever improving device speed [1]. In the past, accuracy of ATE used to be 5 times higher than the state-of-the-art ICs. However, in the last two decades, while the semiconductor speeds have improved at an average rate of 30% per year, the tester accuracy has improved at a rate of only 12%. If this trend continues, tester timing accuracy will soon approach the cycle time of advanced devices making at-speed test almost impossible. Table 1 shows the ITRS expected trends for yield, device speed, and tester accuracy and cost. It is clear from this data that even long before the tester timing accuracy reaches the cycle time of the devices, yield loss due to insufficient accuracy of testers will become unacceptably high.

The cost of ATE per pin for high-performance circuits has remained approximately constant for the past 20 year at around \$10-12K/pin. Recently, this value has begun to fall below \$8K/pin and is expected to continue to decrease in years to come (Table 1). Nevertheless, it may be difficult for high-pincount, 2 GHz ATE cost to fall below \$5K/pin in the future. This is attributed to the fact that the demand for higher speed, greater accuracy, more time sets, and larger vector memory will offset most of the gains seen for reducing ATE cost [1]. It is expected for the cost of state-of-the-art testers to rise significantly and to reach \$20M in less than a decade. According to the ITRS, it may cost more to test a transistor than it costs to manufacture it by 2014.

This discussion shows that, in the near future, it will be essential to include DFT/BIST techniques to reduce the semiconductor industry's reliance on traditional, high-cost, full-feature testers. The requirements for ATE designed to work with DFT/BIST techniques are much simpler than the traditional testers. This is expected to help reduce the cost of testing future high-performance devices considerably.

2. High-Performance Circuit Testing: A Review

The size of a defect determines whether the defect will affect the logic function of a circuit or not. Normally, smaller defects, which are likely to cause partial shorts or opens, have a high probability of occurrence. Such defects cause the circuit to fail to meet its timing specifications without altering the logic function of the circuit. A number of recent studies show concerns about new failure mecha-

Year	1999	2001	2003	2005	2008	2011	2014
Yield (%)	87	84	79	73	64	56	50
Device period (ps)	830	700	580	500	400	340	260
Overall ATE accuracy (ps)	200	160	130	100	100	100	100
Tester cost per pin ($)	8000	7000	6000	5000	4000	2000	1500

Table 1. ITRS Trends in yield, device speed, and tester accuracy and cost [1].

nisms in scaled geometries that may be harder to detect with conventional means. Nigh et al. [2] reported a significantly large number of timing only failures that did not affect the steady-state logic functionality. Similarly, for Intel's manufacturing processes, Needham et al. [3] reported an increasing shift towards soft defects as technology moved from 0.35 to 0.25μm. These defects do not always cause failures at all temperature and voltage conditions and are considered to be major long term reliability threats.

One approach to tackle this problem is to employ correlation-based testing methods not requiring high-speed test equipment. Keshavarzi et al. [4] reported a strong correlation between I_{DDQ} and the maximum operating frequency of a 32-bit microprocessor. They argued that the two parameters are fundamentally related, as both are functions of the channel length. This information can be used as a mechanism for high performance binning. Another technique uses the correlation between supply voltage and device speed. Hao and McCluskey [5] suggested the use of very-low-voltage testing as a means for testing weak ICs. CMOS digital circuits exhibit an increasingly large switching delay as supply voltage is reduced. Supply voltage reduction causes the delay faults to be more noticeable. Hence, these faults can be detected easily at frequencies much lower than the operating frequency.

Another approach is based on the introduction of special techniques to enhance the capabilities of testers and/or the testability of the device under test (DUT). These techniques are generally preferred over correlation-based techniques [6]. One example of these techniques is the multiplexing of tester clock pins in order to extend the clock frequency range of the tester. This is a standard feature offered in most modern digital testers. Other techniques include the introduction of special structures and test signals resulting in a lower timing requirements for testers.

Agrawal and Chakraborty [7] proposed the creation of a low frequency test mode in digital circuits. A quantifiable, externally controlled delay is added such that high performance testing can be carried out with relatively slow-speed testers. They proposed a pulse-triggered flip-flop in which a dynamic latch is introduced inside a traditional master-slave flip-flop. The resulting three-latch structure has two modes of operation; normal mode and test mode. In test mode, flip-flop delay can be modulated by changing clock's pulse width. This allows for testing combinational logic and interconnects for delay faults with a lower clock frequency in test mode. Although the concept of adding delay in test mode is elegant, this implementation has some important shortcomings as the dynamic latch makes the flip-flop operation very sensitive and time critical.

Controlled delay flip-flop (CDFF) [6] is an alternative to pulse-triggered flip-flop that avoids many of its implementation problems. In this technique an additional test mode clock is used to control the delay of the flip-flop.

3. CDFF for Testing High-Performance Circuits at Low Speed

Figure 1 shows a gate level implementation of the CDFF. The transfer of data from the master latch to the slave latch is controlled through a control logic and depends on the relative timing of the clock (CLK) and the test clock (TCLK).

In Figure 2(a) a model for digital VLSI circuits; a combinational block between two sequential blocks (registers, flip-flops,...etc), is used to illustrate the operation of the CDFF. In normal mode, TCLK is kept high ensuring normal flip-flop operation (Figure 2(b)). Under this condition, the normal mode clock period (T_{NM}) is given by:

$$T_{NM} = t_{prop} + t_{comb} + t_{setup} \quad (1)$$

where t_{prop} is the propagation delay of the flip-flop, t_{comb} is the time slot given for the combinational block to evaluate its input, Q_1, and produce the input of the next sequential block, D_2, and t_{setup} is the setup time of the flip-flop.

In test mode, a tester programmed time offset of the clock is used to generate TCLK. This allows for the control of the data transfer from the master to the slave . The net effect is that, flip-flop output, Q, appears after an additional delay which is the time offset between the clock and the test clock. Figure 2(c) illustrates this scenario when the test clock is active. Under this condition, the test mode clock period is given by:

$$T_{TM} = t_{prop} + t_{comb} + t_{setup} + t_{offset} \quad (2)$$

where t_{offset} is the time offset between the clock and the test clock. The test mode clock period should be large enough to accommodate all delay terms in Equation 2. It

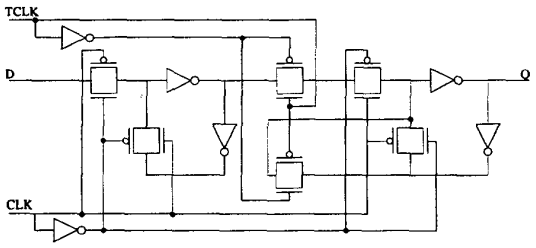

Figure 1. Controlled delay flip-flop.

is clear from this equation that increasing t_{offset} allows the circuit to be tested at a lower frequency than the normal mode frequency. In other words, clock frequency can be reduced while the combinational circuit delays are tested with the same delay margins.

4. Using CDFF to Arbitrarily Reduce Test Clock Frequency

In this section, we present a methodology for generating the clock and the test clock of the CDFF in a way that allows the test mode clock frequency to be reduced arbitrarily. This technique is targeted for on-chip implementation as part of a DFT/BIST design.

When generating the clock and the test clock for a circuit using CDFF to improve testability, one has to take into consideration the timing requirements for correct operation. For the CDFF to function properly, the timing of the clock and the test clock must be carefully adjusted to accommodate both the setup time (t_{setup}) and the propagation delay (t_{prop}) of the flip-flop. For the combinational block it is necessary to have the flexibility to change the value of t_{comb} so as to determine exactly the delay through this block and test the circuit for delay faults.

4.1. Reducing Test Mode Clock Frequency

In test mode, reducing clock frequency while maintaining correct operation for all parts of the circuit means that, if the clock frequency becomes very low, t_{offset} has to be extremely large. As suggested in [6], the test clock can be generated as a delayed version of the clock with a delay of t_{offset}. Normally, it is difficult for a low-frequency tester to provide very large time offset with reasonable accuracy. H. Speek et al. [8] suggested the use of two programmable duty-cycle controllers and a programmable delay line to generate the clock and the test clock in test mode. Using their design, reducing the test clock frequency to a very small value requires a large delay line to generate the required delay with appropriate resolution.

Careful examination of the timing diagram in Figure 2(c) shows that, instead of generating the test clock by delaying

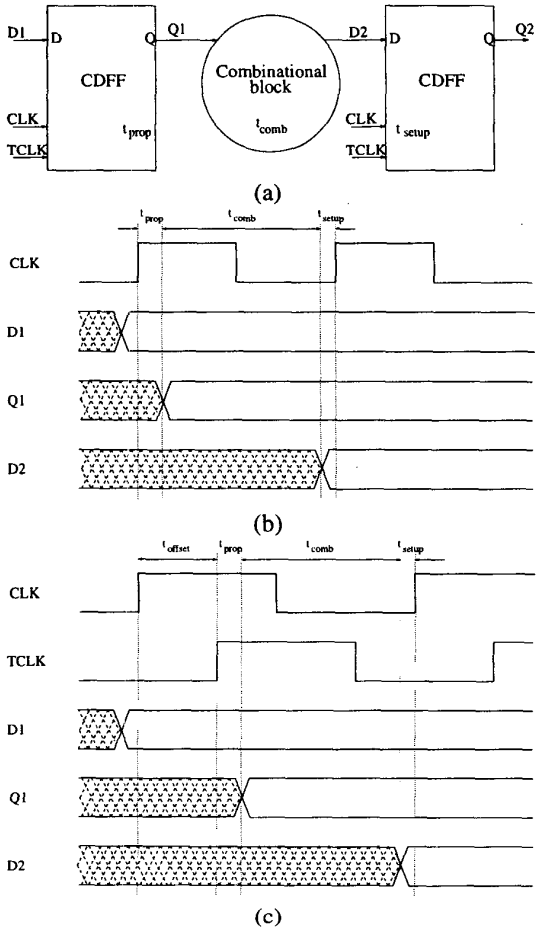

Figure 2. CDFF operation. (a) Circuit model. (b) Normal mode. (c) Test mode.

the clock in test mode, the clock can be generated by delaying and inverting the test clock. Generating the clock this way makes t_{offset} (which is the key factor in reducing the test mode clock frequency) independent on the relative timing of the two clocks and allows its value to be increased arbitrarily. Increasing t_{offset} while keeping all the other terms in Equation 2 unchanged implies a reduction in test mode clock frequency. It is clear that by doing this, the test mode clock frequency can be reduced with no lower limit.

4.2. Clock and Test Clock Generation

The block diagram of the system used to generate the clock and the test clock is shown in Figure 3(a). The input clock, IPCLK, is a rated frequency signal in normal mode and a low frequency, 50% duty cycle signal in test mode.

A multiplexer (MUX) is used to select the mode of op-

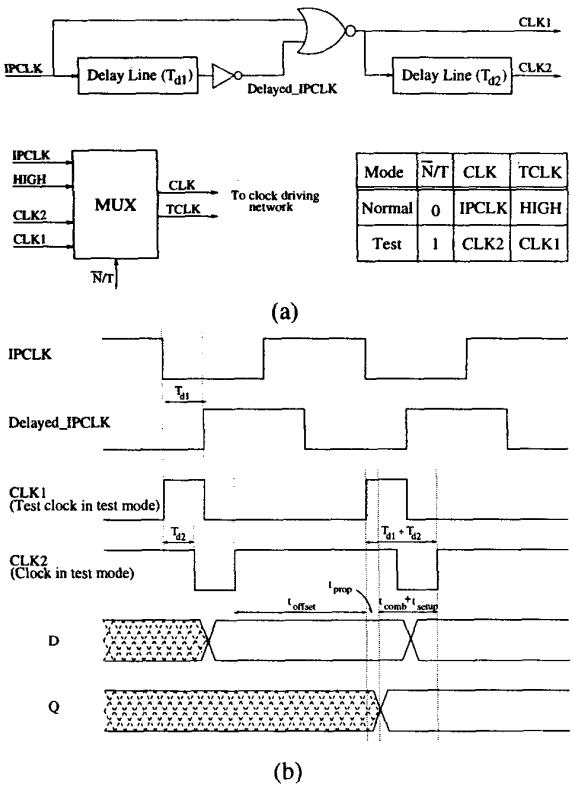

Figure 3. Generating clock and test clock. (a) Block diagram. (b) Timing diagram.

eration through the mode select input ($\overline{N}/T$). For normal mode operation ($\overline{N}/T$=LOW), IPCLK passes through the MUX to the clock driving network and TCLK is set to HIGH. In test mode ($\overline{N}/T$=HIGH), two delay lines are used to generate both CLK and TCLK. This is illustrated by the timing diagram in Figure 3(b). The system consists of two sections. The first section is used to generate a pulse whose width is equal to the delay of the first delay line, T_{d1}. The output of this section (CLK1) is selected through the MUX to be the test clock, TCLK. In the second section, CLK1 passes through the second delay line (T_{d2}), resulting in CLK2 that is used as the clock, CLK. Figure 3(b) also shows the D and Q signals of the CDFF to illustrate the relationships between the various timing factors of the system in Figure 2(a) on one side and the delays T_{d1} and T_{d2} and the frequency of IPCLK, f, on the other side . These relationships can be expressed by the following two equations.

$$T_{d1} + T_{d2} = t_{prop} + t_{comb} + t_{setup} \quad (3)$$

$$1/f = T_{d1} + T_{d2} + t_{offset} \quad (4)$$

These equations suggest that, for constant f, a change in either T_{d1} or T_{d2}, or both will lead to an equivalent change in t_{comb}. This allows the combinational block to be tested for delay faults by changing the time slot allowed for the evaluation of its inputs. For constant T_{d1} and T_{d2}, changing f will cause only t_{offset} to change without affecting the operation of neither the flip-flop nor the combinational block.

As discussed before, the generation of CLK and TCLK must be in a way that facilitates the testing of the combinational block and maintains correct operation of the flip-flop. It is clear that the first objective can be achieved by changing T_{d1}, T_{d2}, or both. To achieve the second objective, the limiting values of T_{d1} and T_{d2} necessary for correct flip-flop operation have to be determined. This is done with the help of the model given Figure 2(a) where the combinational block is replaced by a chain of inverters and used to characterize the performance of the flip-flop. To do this, T_{d1} is held constant at a relatively high value while T_{d2} is reduced gradually. For the flip-flop used in our study, simulations show that when T_{d2} falls below 122ps, the flip-flop ceases to function properly. This is attributed to the fact that T_{d2} has to be large enough to allow the propagation of the data from the master to the slave. This value of T_{d2} is equal to the worst case propagation delay of the flip-flop. With this value of T_{d2}, T_{d1} is reduced gradually till the flip-flop failed to function properly. The limiting value in this case is 53ps more than the propagation delay of the inverter chain which is equal to the setup time of the flip-flop.

The following section presents a CMOS implementation of the proposed system in 0.18μm technology.

5. Implementation

The objective of our design is to have the capability of testing high speed combinational blocks working at speeds as high as 2.5 GHz for delay faults. To achieve this, the design allows t_{comb} to be varied from 400ps to 1150ps. Referring to Equation 3 and considering the limiting values of T_{d1} and T_{d2}, the minimum and maximum values of $T_{d1} + T_{d2}$ should be 575ps and 1325ps, respectively. As stated before, T_{d1}, T_{d2}, or both can be varied to achieve these requirement. It is clear that keeping one of them constant while changing the other should save a lot of hardware required for programmable delay lines.

When choosing the values of T_{d1} and T_{d2}, two factors should be taken into consideration. First, due to interconnect delays, the propagation of extremely small pulses might be difficult to achieve. And second, it might be difficult to maintain a very small phase difference between the two clocks due to clock skew across the chip. Our implementation is designed such that T_{d1} can be varied from 275ps to 1025ps, while T_{d2} is held constant at 300ps. Although a 275ps pulse width might seem very small, propa-

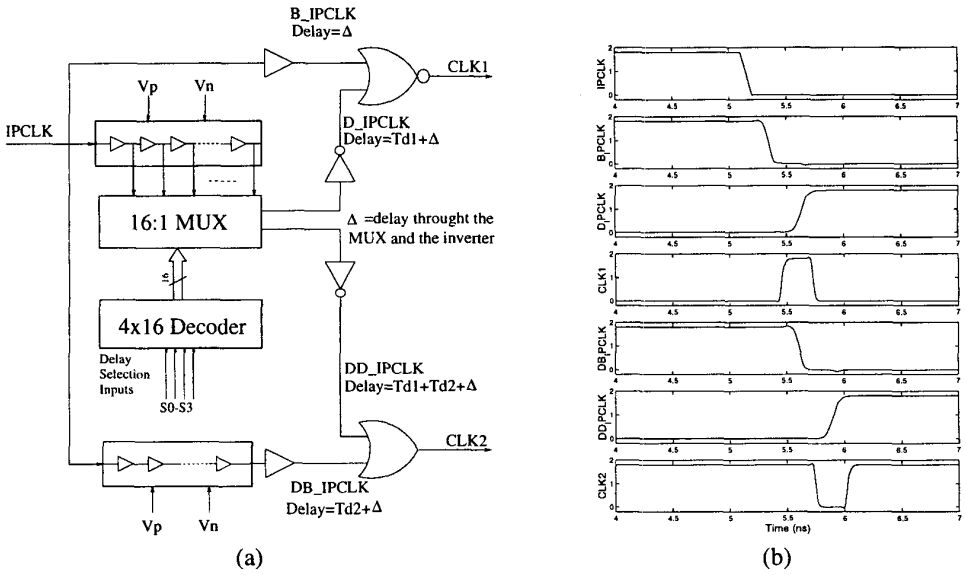

Figure 4. (a) Circuit used to generate CLK1 and CLK2. (b) Signals at different points for T_{d1}=275p.

gating such a small pulse is within the capability of a circuit working at 2.5 GHz. Building the TCLK driving network as a replica of the CLK driving network should help minimize the skew between the two clocks.

5.1. Circuit Design

Figure 4 shows the circuit used to generate CLK1 and CLK2 (TCLK and CLK in test mode), as well as the signals at different points in the circuit when T_{d1} is set to 275ps. Two delay lines are used to generate CLK1 and CLK2. The basic component in these delay lines is the delay element.

5.1.1. Delay Element. Each delay line consists of a chain of delay elements each having a delay of 50ps. The design of the delay element is very crucial to ensure accurate delays regardless of process, temperature, and supply voltage variations. The delay element used in our design is shown in Figure 5. It consists of two inverters with current control transistors M1 and M6. Referring to Figure 4(b), it can be shown that only the delay for the negative edge of IPCLK is critical for correct timing of CLK1 and CLK2. Therefore, the delay element is designed such that only the delay for negative going input is set to 50ps. This makes the sizing of transistors M3 and M4 not critical for the design. These two transistors should be small to minimize the loading of the previous stage and consequently to help reduce the delay for the negative going edge of the input. The only effect this has on the operation of the circuit is that it might limit the maximum test mode frequency to a value much lower than the rated normal mode frequency. This is because of

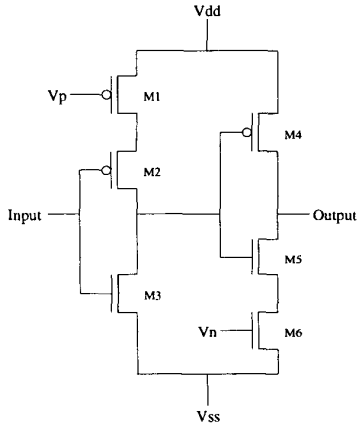

Figure 5. Schematic diagram of the delay element used for the delay lines.

the unbalanced delay for rising and falling edges. However, this is not an issue as reducing the test mode frequency is the basic objective of the design.

The delay of the delay element is controlled by controlling the currents through transistors M1 and M6. This is done by two control voltages, V_p and V_n. If V_p and V_n are set to V_{ss} and V_{dd} respectively, currents through M1 and M6 will be maximum causing minimum delay for the delay element. Alternatively, if V_p and V_n are set to $V_{dd} - V_{th}$ and $V_{ss} + V_{th}$ respectively, where V_{th} is the highest of the PMOS and NMOS threshold voltages, currents through M1 and M6

will be small causing a very high delay. The sizes of M1 and M6 should be large enough to provide currents sufficient to achieve the required delay. Area overhead due to large control transistors can be reduced considerably by sharing the current control transistors between many delay elements. In our design, only three PMOS (for M1) and three NMOS (for M6) transistors are used for all 32 delay elements. Optimum sizing of transistors M2 and M5 is very important to provide sufficient currents without excessively loading the previous stage. For $V_n = V_{dd}$ and $V_p = V_{ss}$, transistor sizes are selected such that the delay element will have the required delay (50ps) under worst case conditions (slow-PMOS and slow-NMOS transistor models, $T = 100°C$, and V_{dd} is 10% less than its nominal value).

5.1.2. Programmable Delay Line. As shown in Figure 4(a), a programmable delay line is used to generate two delayed signals, D_IPCLK and DD_IPCLK. The delay line consists of a chain of delay elements whose outputs are tapped and fed to a multiplexer. Two of these signals are selected by a decoder to be the outputs of the multiplexer. The delay of D_IPCLK (used to generate CLK1) has to be set to the required value of T_{d1}, while DD_IPCLK (used to generate CLK2) always has a fixed phase shift of 300ps (T_{d2}) with respect to D_IPCLK. This means that the delay chain has to be designed to have a minimum delay of 275ps ($T_{d1_{min}}$) and a maximum delay of 1325ps ($T_{d1_{max}} + T_{d2}$). Having a fixed phase shift between the two outputs allows the same decoder to be used to select both signals simultaneously.

The multiplexer is designed to have minimum delay. This is very important to minimize its effect on the delays of its outputs with respect to IPCLK. The decoder is built with minimum size transistors as its switching speed is not crucial for the operation of the circuit.

5.1.3. Buffers, Gates, and Fixed-Delay Delay Line. As shown in Figure 4(a), IPCLK is buffered to generate B_IPLK which is then NORed with D_IPCLK to generate CLK1. A fixed-delay delay line is used to generate DB_IPCLK to have a delay of 300ps (T_{d2}) with respect to IPCLK. CLK2 is generated by ORing DB_IPCLK and DD_IPCK. To balance and minimize their effect, the OR and NOR gates are designed to have equal delays as well as small rise and fall times. Proper sizing of the buffers is important to compensate for the delays through the multiplexer and the inverters ($\triangle$ in Figure 4(a)). Furthermore, this is crucial to help adjust the minimum delay of the programmable delay line ($T_{d1_{min}}$).

5.2. Simulation Results

Several simulations were made to test our design and to verify the possibility of reducing test mode clock frequency.

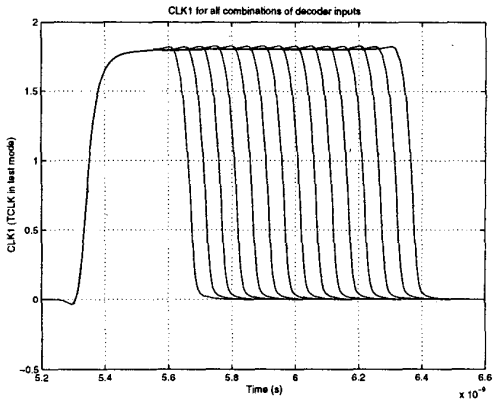

Figure 6. CLK1 for all possible values of S0-S3.

Figure 6 shows CLK1 for all possible combinations of decoder inputs. Simulation results show that our design is capable of accurately controlling the value of T_{d1} to achieve the required design goals. The maximum error in the delays of the delay line was found to be $\pm 15.8\%$ which is highly acceptable regarding the small delay values. For all values of T_{d1}, simulations show that T_{d2} is always constant at 298ps.

Signals generated from our circuit were used to test the CDFF at a wide range of frequencies. These results are shown in Figure 7 for T_{d1}=275ps and at frequencies of 100MHz and 100kHz. These results show that any increase in the period of IPCLK appears as a similar increase in the CLK-Q delay of the flip-flop. This means that the time slot allowed for the evaluation of the combinational block remains constant with reducing test mode clock frequency.

5.3. Implementation Issues

One of the important implementation issues associated with our design is the number of extra pins to be added to the chip. Although, the design requires 7 extra pins for $\overline{N}/T$, V_n, V_p, and S0-S3, this number can be reduced to three or two. The number of pins associated with the decoder can be reduced to only one by storing the data in a shift register. V_n can always be set to $V_{dd} + V_{ss} - V_p$ such that one of the two control inputs can be generated from the other. The need for external control signals can even be eliminated using techniques like that presented in [9], in which a delay-locked loop (DLL) is used to generate the control signals.

Although this technique adds to the cost of the chip in terms of area and power consumption, this cost can be acceptable as the design is expected to reduce the cost of testing and manufacturing. Other implementation issues asso-

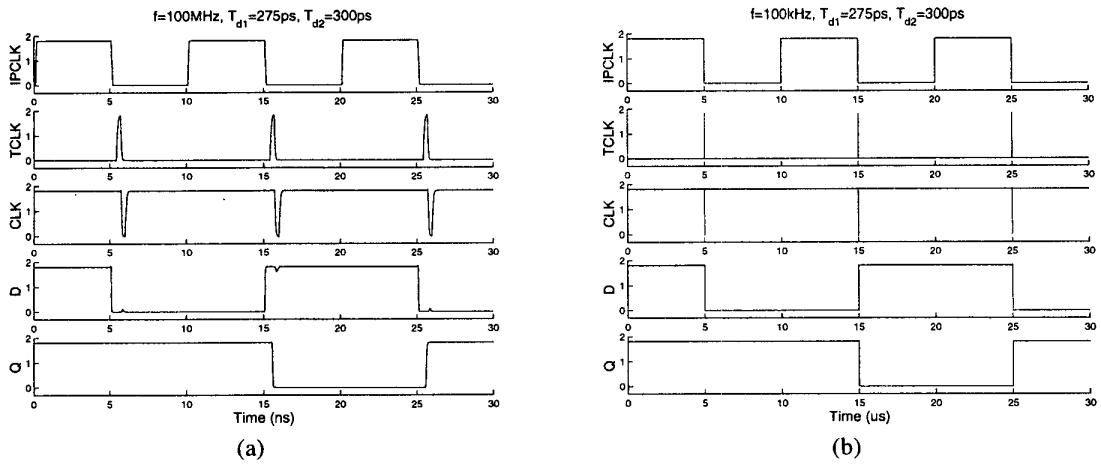

Figure 7. CDFF simulation results for T_{d1}=275ps. (a) f=100MHz. (b) f=100kHz.

ciated with the use of CDFF are discussed in [6].

6. Conclusions

We have presented a methodology for testing high-performance circuits with very little requirements on ATE. The technique uses CDFF to control the delay of the DUT to facilitate its testing. This technique allows such circuits to be tested with arbitrarily low-frequency, 50% duty cycle input clock.

Circuit used to generate CLK and TCLK in test mode have been presented along with various design and implementation issues. Simulation results show that using this design, pulse widths as small as 275ps for TCLK can be efficiently achieved. The accuracy of the design is 50ps which is 3-4 times better than the state-of-the-art ATE.

Simulations of the CDFF with clocks from our design show the validity of this methodology. These simulations prove that the test mode clock frequency can be reduced with no lower limit while keeping the time slot allowed for DUT evaluation constant.

Acknowledgment

The authors gratefully acknowledge support of Communications and Information Technology Ontario (CITO) operating grant for this work.

References

[1] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors, 1999 Edition", 1999.

[2] P. Nigh. et al., "So What is an Optimal Test Mix? A Discussion of The Sematech Methods Experiment", *Proc. of International Test Conference*, pp. 1037–1038, 1997.

[3] W. Needham, C. Prunty and E. H. Yeoh, "High Volume Microprocessor Test Escapes, An Analysis of Defects Our Tests are Missing", *Proc. of International Test Conference*, pp. 25–34, 1998.

[4] A. Keshavarzi, K. Roy and C. F. Hawkins, "Intrinsic Leakage in Low Power Deep Submicron ICs", *Proc. of International Test Conference*, pp. 146–155, 1997.

[5] H. Hao and E. J. McCluskey, "Very-Low-Voltage Testing for Weak CMOS Logic ICs", *Proc. of International Test Conference*, pp. 275–284, 1993.

[6] M. Shashaani and M. Sachdev, "A DFT Technique for High-Performance Circuit Testing", *Proc. of International Test Conference*, pp. 267–285, 1999.

[7] V. D. Agrawal and T. J. Charaborty, "High-Performance Circuit Testing with Slow-Speed Testers", *Proc. of International Test Conference*, pp. 302–310, 1995.

[8] H. Speek et al., "Bridging the Test Speed Gap: Design for Delay Testability", *Proc. of European Test Workshop*, pp. 3–8, 2000.

[9] Gary C. Moyer et al., "The Delay Vernier Pattern Generation Technique", *IEEE Journal of Solid-State Circuits*, vol. 32, n. 4, pp. 551–562, April 1997.