AN ABSTRACT OF THE THESIS OF

Tim D. Low for the degree of Master of Science in Electrical and Computer Engineering presented on November 12, 1999. Title: Low Power, High Performance Pseudo-static D Flip-Flop.

Abstract approved: Redacted for Privacy

Shih-Lien Lu

Digital systems, in particular microprocessor, have recently experienced phenomena growth in performance. Both technology advancement and clever design have sustained this performance growth. As clock frequency heads into the Ghz range, new circuit design, for both logic and storage, are needed. Such new circuit technology must provide needed performance with minimum power consumption.

Flip-flops are essential elements of a digital system. They are used to hold both state information and results. As processor architecture such as superscalar becomes more advanced, the control logic grows more complex resulting in an increasing number of D flip-flops. These flip-flops are all driven by the global clock, which leads to higher power dissipation with increasing clock frequency. One way to reduce power consumption is to send the microprocessor into a sleep mode. Once in this mode, the clock is turned off (at logic low level), forcing the control logic to remain in a standby state. In this thesis, two D flip-flop designs are introduced and compared with conventional designs: dynamic NRC (no race condition) and pseudo-static cascode pull-down. Such design criteria comparisons include speed, power consumption, scaling, noise margin, and metastability.

Low Power, High Performance Pseudo-static D Flip-Flop

By

Tim D. Low

A Thesis Submitted
to
Oregon State University

In Partial Fulfillment of
the requirements for the degree of

Master of Science

Presented November 12, 1999
Commencement June 2000

Master of Science thesis of Tim Low presented on November 12, 1999

Approved:

Redacted for Privacy

Major Professor, representing Electrical and Computer Engineering

Redacted for Privacy

Chair of Department of Electrical and Computer Engineering

Redacted for Privacy

Dean of Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Redacted for Privacy

Tim D. Low, Author

Table of Contents

Table of Contents (Continued)

List of Figures

List of Tables

LOW POWER, HIGH PERFORMANCE PSEUDO-STATIC D FLIP-FLOP

CHAPTER 1: INTRODUCTION

With an increasing demand for higher performance and lower power dissipation in current microprocessor, new circuit design techniques are needed for both switching logic and storage devices. In a digital system, flip-flops are often thought of as memory devices, whose primary function is to store state information and data results. As complexity in microprocessor increases, both logic requirements and storage depth will also increase. This will lead to a larger number of flip-flops and may result in larger power consumption. In fact, the maximum speed of a flip-flop is directly proportional to the total power dissipated. In the mobile part used in today's computer notebooks, emphasis on power dissipation has been a major primarily design concern.

One way for a system to save power is to enter a sleep mode where the states of the logic remain saved until the system becomes active again. This is achieved by turning off the clock and forcing the system into a standby state. Once the system enters this state, the storage capacitance may leak over time resulting in a loss of stored information. To maintain the capacitive charge during sleep mode, a positive feedback inverter or level restorer is required. Such configurations are considered to be a pseudo-static design; a dynamic CMOS latch with feedback that refreshes itself to retain the stored content. The high gain from the cross-coupling inverter makes pseudo-static flip-flop ideally as signal driver. When the system revives into its normal state, the control

logic reinitializes and continues where it last left off. In this thesis, two designs are introduced and compared to the existing D flip-flop implementation.

1.1 Problem Statement

Static circuits are implemented in control logic over dynamic circuits primarily because of their ease of design and synthesis tool support. The biggest problem with dynamic design is its inability to retain capacitive charge. This may cause a problem when the system goes into sleep mode. As the capacitive charge leaks over time, the correctness of the logic deteriorates and become corrupted. When referring to leakage, the static component is dominant at low activity or standby operation. As core frequency linearly increases and transistor sizes decreases, smaller sub-micron technology introduces a greater leakage problem. By reducing the threshold voltage of transistors, leakage current rise exponentially due to the direct short-circuit path between source and drain. The correctness of the logic may also be affected by long or ill-defined transitions. This may cause undefined states in the clocked system and leads to processing error.

1.2 Summary of Criteria

Improvement has been made over the existing standard design in terms of cost and performance. As complexity of the processor increases, the number of flip-flops required in the logic operation will also increase. Reducing the transistor count in each flip-flop minimizes the total power consumed by the system. With a smaller flip-flop,

more area can be allocated to the logic depth. In addition to minimizing the number of transistors, direct source-to-drain current leakage is reduced with clock controlled transmission gates. For this design, the input appears directly at the output, delayed only by the propagation delay of the inverters. Since a flip-flop resembles a back-to-back inverter, Miller capacitance affects the resolution time and limits the maximum frequency of operation. The tradeoff between speed and power becomes the primary focus of this thesis.

1.3 Summary of the thesis

The main goal of this thesis is to develop a flip-flop that uses less area than existing designs, without compromising speed or power. This has been achieved in the NRC and cascode pull-down circuits where redundant transistors are examined and removed. In a way, the structure of the NRC circuit is quite similar to the low-area design but lacks the pull-up/pull-down networks needed to maintain the logic "0" charge. The speed of the low-area circuit is limited by the contention between the pull-up and pull-down networks and therefore, sampling data must overcome a non-transparent latch. However, the NRC design faces a similar problem. With the circuit configured as it is, logic "1" sees a transparent latch but, logic "0" must overcome a feeder in order for it to propagate into the output buffer. Therefore, the proposed design resembles a dynamic flip-flop. The NRC design offers increased speed while still managing to reduce total power dissipation. Because this design uses a single clocking scheme and lacks a feedback element for the logic "0" case, the flip-flop has difficulty storing a charge. Due to the storage limitation of the device, the NRC circuit cannot be

classified as a pseudo-static design. By replacing the single clocking scheme with two non-overlapping clocks and adding a clock controlled pull-down network into the slave latch, the dynamic implementation can be converted into a pseudo-static design. To eliminate the Miller effect, a cascode amplifier replaces the standard back-to-back inverter configuration. With this cascode configuration in the slave, current leakage is reduced and improvement has been made in the metastability. Overall, the pseudo-static cascode pull-down design offers high-speed, low power operation.

CHAPTER 2: DESIGN CRITERIA

In the past, the speed of the microprocessor has always been the key design criteria. As system features trend toward portability, speed is no longer the only design concern. In a portable notebook, power consumption becomes the main issue when comparing performance. Since D flip-flops make up the majority of control logic, the design of these circuits must consider the following issues: speed, scalability, and power consumption.

2.1 Speed

To prevent race-through conditions, the master-slave latch configuration is often used. Master-slave flip-flops are built by cascading two basic latches, with opposite clock phases. Each flip-flop design has three timing parameters: CLK to Q delay, setup time and hold time as stated by Unger and Tan in [1], where terminal C corresponds to Clk:

D_{CQ} propagation delay from the C terminal to the Q terminal, assuming that the D signal has been set early enough relative to the leading edge of the C pulse;

U *setup time*, the minimum time between a D change and the triggering (latching) edge of the C pulse such that, even under the worst conditions, the output Q will be guaranteed to change so as to become equal to the new D value, assuming that the C pulse is sufficiently wide;

H *hold time*, the minimum time that the D signal must be held constant after the triggering (latching) edge of the C signal so that, even under the worst case conditions, and assuming that the most recent D change occurred no later than U prior to the triggering (latching) edge of C, the Q output will remain stable after the end of the clock pulse (it is not unusual for the value of this parameter to be negative).

In reducing the clock cycle time, data evaluation is performed as close to the rising edge of the clock as the setup time permits. This way the propagation delay of a flip-flop or D-Q (Clk-Q delay + setup time in the stable region) becomes the optimum setup time. The optimum setup time presents the limit beyond for which, the performance of the latch is degraded and reliability becomes endangered. The cycle time however maybe reduced, if the change in data is allowed to arrive no later than the optimum setup before the trailing edge of the clock. The maximum speed of a flip-flop is defined such that, at a given environment, the current logic is guarantied to overcome the metastable region, consisting of the setup and hold time. The circuit noise margin and metastability can often affect this maximum speed.

2.2 Noise Margin

When noise acts against a stable logic level on a circuit node, it can transiently destroy logical information carried by the node. Ultimately, functional failure will result. Noise margin is defined to be the measurement of gate sensitivity to noise, and therefore a large noise margin is desirable. As long as the sampled signal is within the

boundary of the noise margin, the gate of the transistors will continue to switch correctly. Unfortunately, noise accumulates over time as it propagates from one device to another. To eliminate this problem, the gate must possess some form of a regenerative property. In Figure 2.1, the regenerative property allows a voltage signal to gradually converge to a nominal point where it can be defined as either a high or low transition. Assume that a voltage of V_0, deviating from the nominal voltage is applied into an inverter. The output of this inverter $V_1 = Finv(V_0)$ is fed into a secondary inverter where it gradually converges to the nominal signal after going through numerous of inverter stages.

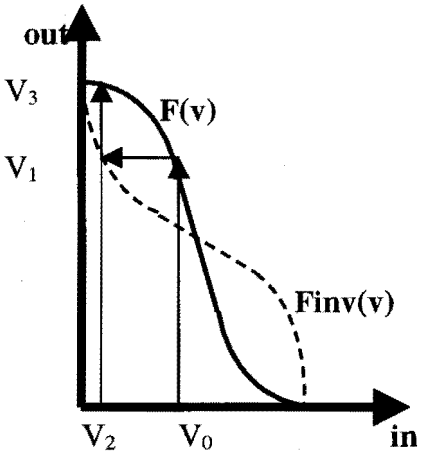

Figure 2.1: Regenerative property

2.3 Metastability

With small devices, noise injection can make a big impact on the reliability of the data. Basically, the hardness against the metastability is proportional to the static noise margin. Therefore, if the static noise margin of the latch is made higher, the flip-

flop will have a higher immunity against metastability. The static margin of the D flip-flops is configured such that it resembles a back-to-back inverter. Therefore, the static noise margin is dependent upon the diagonal length of the square created between the normal and mirrored transfer characteristic. At the intersection of the voltage-transfer characteristic curve and a given point such that Vout = Vin, the metastable voltage V_M, or switching threshold voltage, is defined. Figure 2.2 illustrates an ideal voltage-transfer characteristic curve. The D flip-flop in the metastable state is proportional to the slope of the DC voltage transfer curve at V_M. The V_{IH} and V_{IL} voltage levels delimit the regions of acceptable high and low voltages. These two values represents the points where the gain-bandwidth of the flip-flop *(dVout/dVin)* = -1. If the noise margin is made higher, both V_{IH} and V_{IL} are shifted to the middle region of the voltage transfer curve.

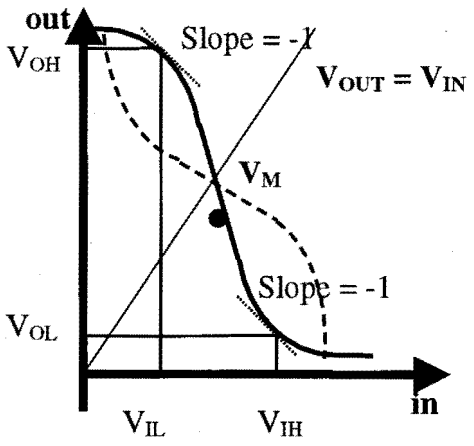

Figure 2.2: Voltage-transfer Characteristic curve

The logic failure in the CMOS latch/flip-flop is mostly due to the long decision time within the metastable state, which occurs for an indeterminate amount of time. As a result of this metastability problem, the flip-flop will end up having an unexpectedly long resolving time. Metastability is defined as the nondeterministic behavior such that the setup and hold times are violated and whereas the metastable state is an invalid state halfway between logic 0 and 1. When a bistable element requires an indeterminate amount of time to generate a valid output, nonbinary signal can propagate into the logic or storage elements, leading to intermittent error in the circuit. Figure 2.3 shows an idealized plot of the CMOS latch resolution time versus the data arrival time. The time indicated by t_{meta} is defined to be the separation point in determining whether data can be latched or not. The delay in the region close to t_{meta} blows up exponentially where the exponential time constant is, to the first order, the inverse of the gain-bandwidth product of the feedback element[2] and is defined as τ, or the resolution time constant. As the gain-bandwidth increases, the τ parameter becomes smaller, and therefore this constant is related to the latch's ability to resolve intermediate voltage level. The metastability window, δ, is defined as the region in which data can not be resolved within a given resolution time, t_r. The resolution time can also be considered as the CLK-Q delay of the flip-flop. The width of the metastability window can be calculated using the following equation: $\delta = T_O e^{-t_r/\tau}$ where T_O is the asymptotic width of the window with no resolution time. Miller capacitance is known to reduce an amplifier's bandwidth and as a result degrades the performances of small-signal linear amplifiers. Since a latch or flip-flop during the metastable state resembles a linear amplifier, biased at V_M, Miller capacitance is also a major factor in limiting the gain-bandwidth of the

back-to-back inverter positive feedback system. To improve the metastable hardness, maximizing the gain-bandwidth becomes an important issue. This can be done by removing the Miller capacitance loading effects[3].

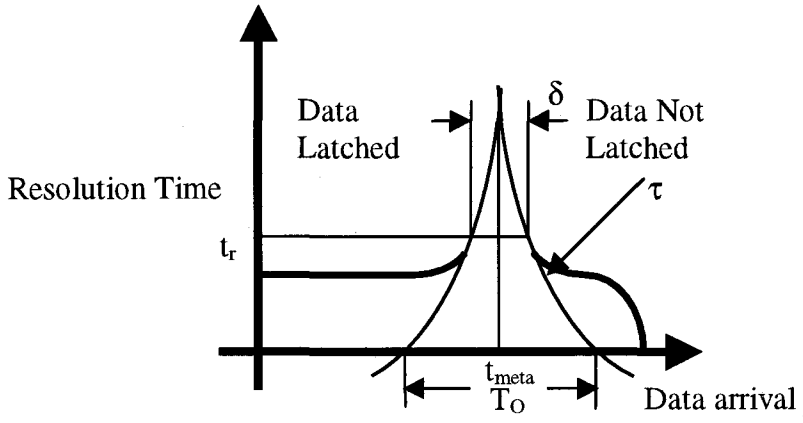

Figure 2.3: Output resolution time versus input arrival time for CMOS latch.

2.4 Scalability

The scalability of a flip-flop is often overlooked but should be considered a key criterion for advanced microprocessor design. By shrinking the die size of the microprocessor, higher speed and lower power dissipation can be achieved. Lowering the threshold voltage will allow transistors to switch faster but the processor will have to endure more leakage and noise problems. If the circuitry in a microprocessor is scaled properly, this flip-flop, without any changes to the circuit methodology, will still continue to operate correctly after a die shrink.

2.5 Power Consumption

For longer operation time, the power consumed by a portable system must be minimized. The power consumption of a gate determines how much heat a circuit dissipates and how much energy is consumed per operation. These properties, in turn, influence the supply-line sizing, power-supply capacity, and most importantly, the number of circuits that can be integrated onto a single chip. In supply-line sizing, the peak power P_{peak} becomes the most important factor: $P_{peak} = i_{peak}V_{supply} = \max[p(t)]$, where i_{peak} is the maximum current being drawn from the supply voltage V_{supply}. The total power dissipation is decomposed into two components: static and dynamic. The dynamic case only occurs during transients, when the gate is switching. This is due to charging capacitors and temporary current paths between the supply rails, and is, therefore proportional to the switching frequency. The higher the number of switching events, the higher the power consumption. As described by Vladimir Stojanovic and Vojin G. Oklobdzija [4], the power consumption of a circuit depends strongly on its structure and the statistics of the applied data. They claim that the data pattern "…. 010101010…" and α=1 would reflex the maximum internal dynamic power consumption, where the data activity rate α represents the average number of output transition per clock cycle. The static component is made up of three leakage currents; current flowing through the reverse-biased diode junctions of the transistors, sub-threshold current, and gate leakage. These are always present, even when the circuit is in stand-by. In the master-slave latch flip-flop, the power dissipated by the latch should also be considered. From [5], the three main sources of power dissipation in the latch are:

- Internal power dissipation of the latch, excluding the power dissipated for switching the output loads
- Local clock power dissipation, which presents the portion of the power dissipated in the local clock buffer driving the clock input of the latch
- Local data power dissipation, which presents the portion of the power dissipated in the logic stage driving the data input of the latch.

2.6 Previous Work

Illustrated in Figure 2.4, a standard pseudo-static D flip-flop CMOS design consists of two level sensitive latches. The design of this flip-flop is composed of two cascaded flip-flops: master and slave. When the clock is high, the master stage follows the D-input while the slave stage holds the previous value. When the clock changes from logic "1" to logic "0", the master latch ceases to sample the input and stores the D value at the time of the clock transition. At the same time, the slave latch becomes transparent, passing the stored master value, Q_M, to the output of the slave stage, Q_S. The input can no longer affect the output because the master stage is disconnected from the D input. Once the clock changes again from logic "0" to "1", the slave latch locks in the master latch output and the master stage starts sampling the input again. Using feedback inverters, the contents stored in the output capacitance node of both master and slave are maintained even with the removal of the clock. As indicated in Figure 4, transmission gates complete the feedback loop between the inverters, changing it into a bi-stable circuit that can store either a "0" or "1" stage. These transmission gates control the operation of the feedback such that any form of fighting node contention is eliminated. Thus, this circuit is a negative edge-triggered D flip-flop by virtue of the

fact that it samples the input at the falling edge of the clock pulse. The D flip-flop circuit can be seriously affected if the master stage experiences a set-up time violation. If the input D switches from "0" to "1" immediately before the clock transition occurs, the master stage fails to latch the correct value and the slave stage produces an erroneous output. Within a synchronously clocked system, as long as the clock-to-Q delays are longer than the setup times, synchronization failures can never occur. Two non-overlapping clocks are needed to prevent a race condition. The advantage of this D flip-flop is its simplicity and involves minimum design risk.

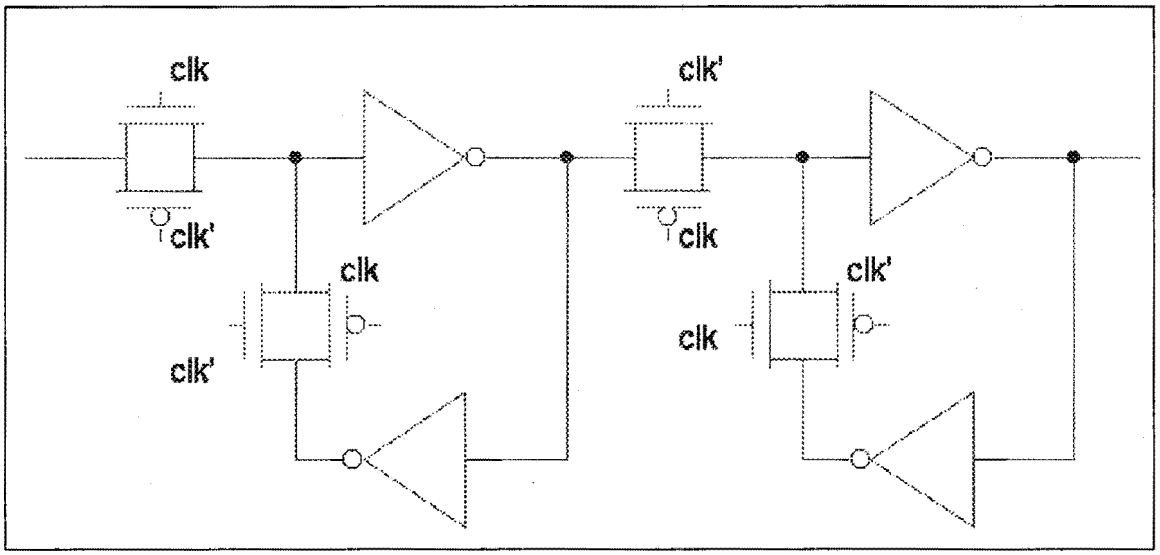

Figure 2.4: Standard Pseudo-static D flip-flop with transmission gate

To reduce the area overhead of the conventional design, the two feedback transmission gates can be eliminated as is shown in Figure 2.5. However, by doing so would increase the total power consumed and sacrifice performance dramatically. The

size of the feedback inverter is weakened to minimize short-circuit power dissipation due to voltage contention. Whenever data is sampled either in the master or in the slave, the input signal must fight the feedback inverter to overcome the previously stored value.

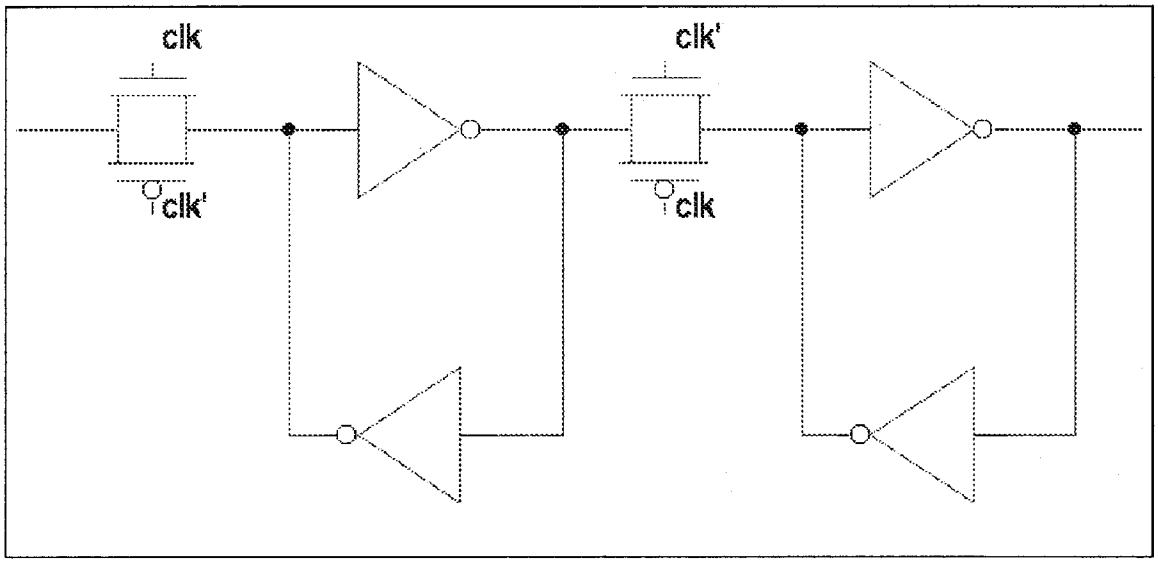

Figure 2.5: Low-area standard pseudo-static D flip-flop

CHAPTER 3: PSEUDO-STATIC DESIGN WITH OUTPUT LEVEL RESTORER FEEDBACK

To increase performance by eliminating voltage contention and still minimizing transistor count, a dynamic No Race Condition (NRC) flip-flop design may be implemented as illustrated in Figure 3.1. When "clk" signal is high, the master transmission gate becomes transparent and samples the data input into the output capacitance of the storage inverter, Q_M. Once "clk" becomes low, the data is transmitted and held in the output inverter buffer, Q_S while the master flip-flop disconnects from the D input. If the stored data in Q_S has logic "1", the output of the inverter buffer is fed back into a pull-down level restorer where it will pull the gate capacitance of the input to ground. Using this method, the output capacitance charge is held constant while repeatedly being refreshed. To prevent voltage contention, any form of feedback for the logic "0" case is eliminated. However, if the stored data has logic 0, the charge of the inverter gate capacitance leaks over time since no feedback is implemented. For high-speed operation, this leakage problem would not be an issue since the size of the gate capacitance would be sufficiently large to hold a charge. In sleep mode, this design may have a problem holding a charge. With the two level restorers implemented, logic "1" data passes through much faster than the logic "0" case. The advantages of this design are mainly of its high-speed operation, simplicity, and small transistor count, which result in low power dissipation. The major disadvantage of this design is its inability to retain data during sleep mode. Without feedback for the logic "0" case, the node of the inverter gate capacitor is left floating. With floating nodes, this dynamic NRC D flip-flop design faces many noise issues such

as cross-cap coupling. Another shortcoming is with lower clock frequencies, the D flip-flop must cope with increasing charge leakage. By having complementary transmission gates, only one clock is needed for operation. As shown in Figure 3.1, a single clock unit is implemented to reduce the local clock power dissipation. However, by using a PMOS over a NMOS device in the slave transmission gate, data passes through the latch much slower. This is mainly the result of its slower mobility property. Though this design may have its faults, it does contain some interesting features including "its incredibly small size and high speed operation" that should be considered noteworthy in future flip-flop design.

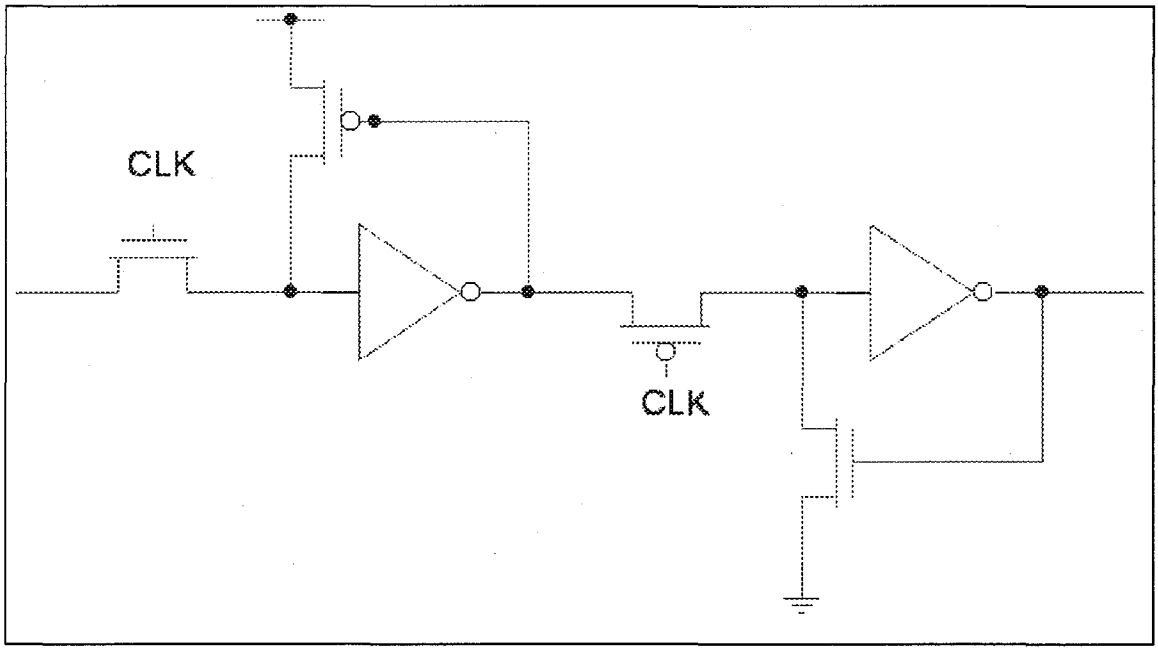

Figure 3.1: Dynamic No Race Condition D-flip flop

Resolving some of the previously stated problems, a pseudo-static version of this design is shown in Figure 3.2. Like the previous circuit, this design tries to maximize speed performance by eliminating voltage contention. Inside the master flip-flop, a pull-up transistor is used to maintain a capacitive charge on the input of the inverter. Conversely, no NMOS pull-down transistor is needed. When the device enters sleep mode with logic "0" stored in the inverter gate capacitor, the charge on the output node would remain high, unless cross-cap noise coupling forces the input to switch. A method of preventing voltage contention in the slave flip-flop is to have a NMOS pull-down transmission gate, controlled by "clk". This gate turns on only when data is no longer being transferred and therefore eliminates direct dc short-circuit path. The Miller effect is also reduced with the addition of the cascode pull-down network, which result in a higher gain-bandwidth product in the gate [5]. As stated by Lee-Sup Kim, "The greater the Miller effect is, the worse the metastable hardness becomes." As a result, by using the cascode configuration, the resolving time of the metastable operation becomes shorter.

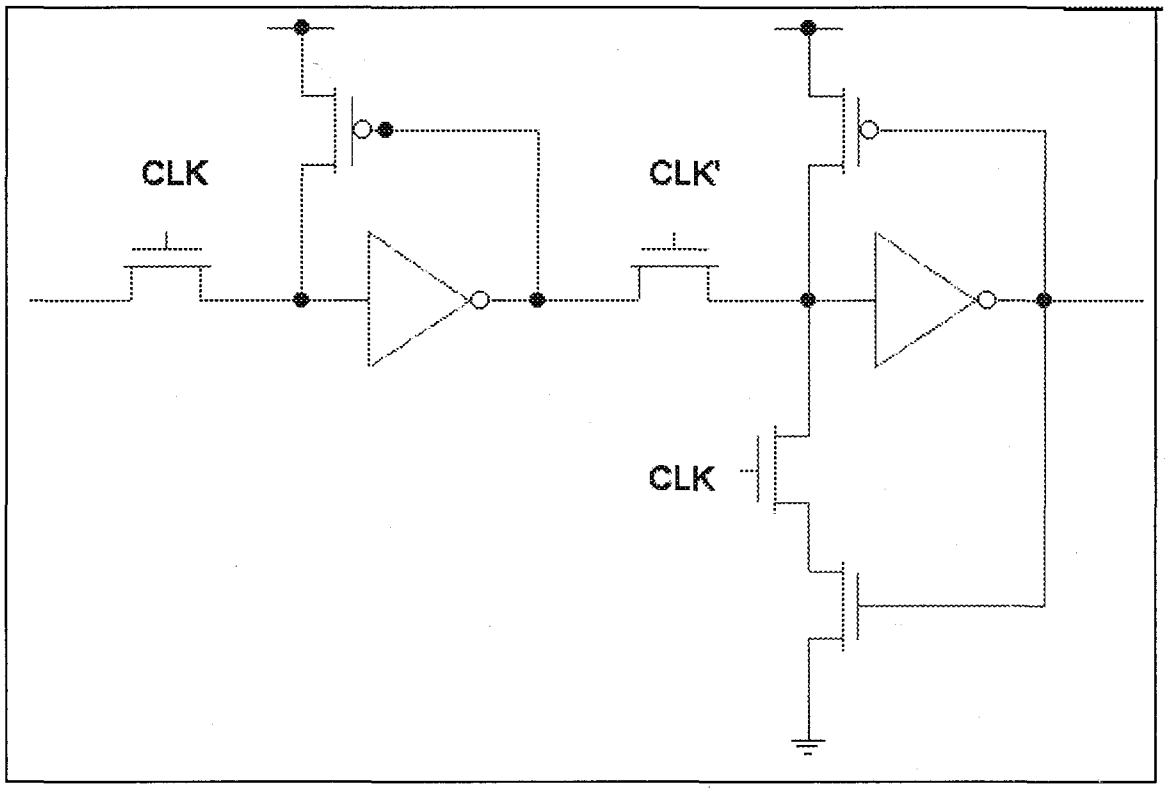

Figure 3.2: Pseudo-static cascode pull-down D flip-flop

CHAPTER 4: SIMULATION RESULTS

4.1 Maximum Frequency with Scaling of Supply Voltage

The four circuits are sized such that a maximum optimal speed can be achieved using a .35 micron technology. Figure 4.1 is a diagram that shows the setup structure used for these experiments. The maximum frequency is obtained by linearly increasing the clock rate until the flip-flop fails to latch data correctly.

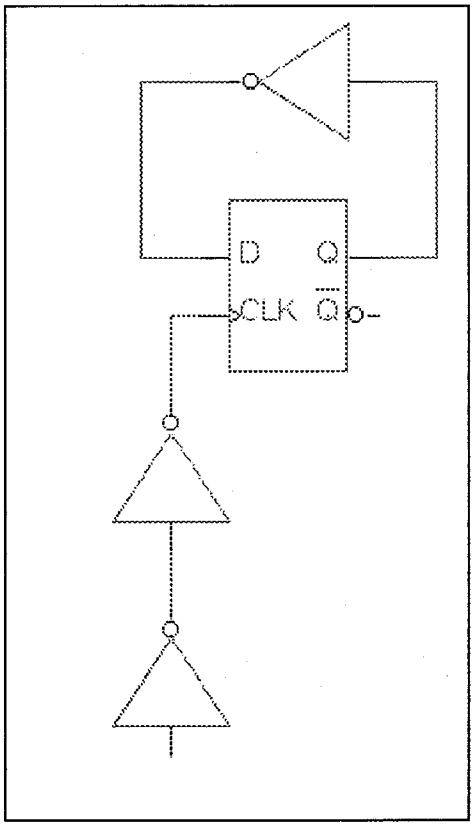

Figure 4.1: Simulation test circuit setup

To see how the flip-flop performance scales with supply voltage, Vcc is varied from 1.5V to 3.3V. The proposed pseudo-static cascode pull-down design (label D in Figure 4.2) achieves the highest frequency at 3.3V, but falls short when compared to the dynamic NRC design (label A) at lower Vcc. As expected, the performance of the low-area design (label C) is much lower than the previous three circuits. This is mainly due to the direct DC short-circuit path created by the feedback inverter. Ideally, the maximum frequency is linearly proportional to the supply voltage, as is apparent in the standard and low-area designs. However, this is not the case for the dynamic NRC and pseudo-static cascode pull-down circuits. Their maximum frequency is limited by the RC delay and therefore, saturates at 2V. The slope of the curves describes a great deal of how well a circuit scales with technology. As supply voltage increases, a steeper slope indicates a higher gain of achievable speed.

The standard and low-area circuits have symmetric structure that scales nicely with Vcc. The dynamic NRC and pseudo-static cascode pull-down circuits have a pull-up transistor in the master flip-flop and unfortunately do not scale with Vcc. This pull-up transistor is transparent to logic “1”, but not to logic ”0”. This imbalance in data resolving time reduces the scalability of the device.

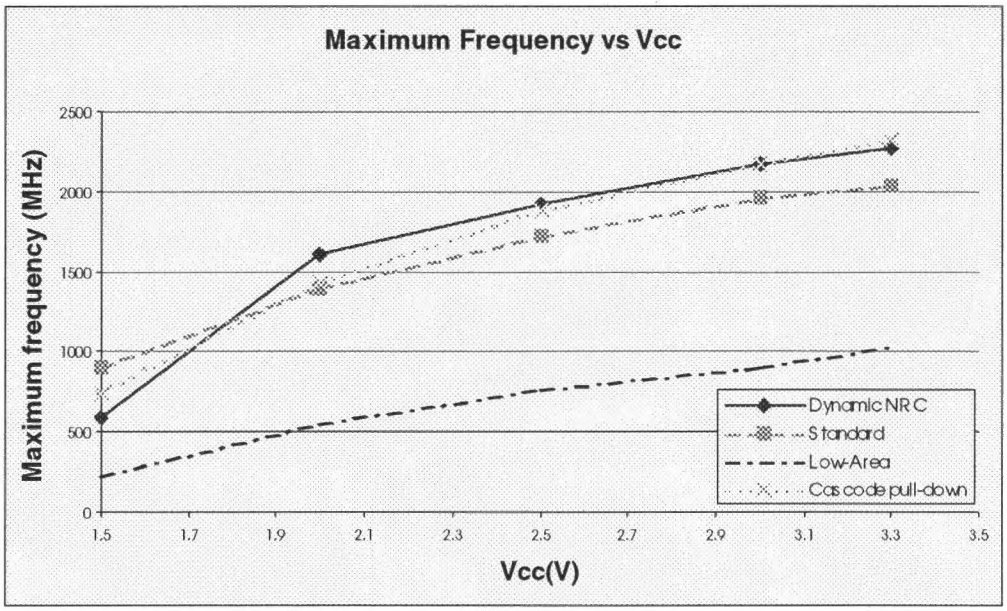

Figure 4.2: Illustrates the maximum frequency for a given supply voltage; a). Dynamic NRC, b). standard design with control gates, c). low-area and d). pseudo-static cascode pull-down.

4.2 Average Total Power Consumed with Scaling of Vcc

By keeping the frequency constant at 100 MHz and varying the supply voltage from 1.5V to 3.3V, the total average power consumed by each circuit are compared in Figure 10. To model the worst case scenario, the output of the D flip-flop is driven back into its input through an inverter and therefore, creating a test data pattern of "....01010101...". To include the power dissipated by the local clock, two additional inverters are added to the design as shown in Figure 4.3. As expected, the dynamic NRC design consumes the least amount of power, mainly because of its smaller transistor count and one less clock. However, even though the low-area design contains two more transistors than the dynamic design, it consumes 40% more power. This is the due to the fighting voltage nodes and direct short-circuit dc path. The pseudo-static

cascode pull-down design consumes 10% more power than the dynamic design mainly because of the additional pull-down/pull-up network in the slave flip-flop.

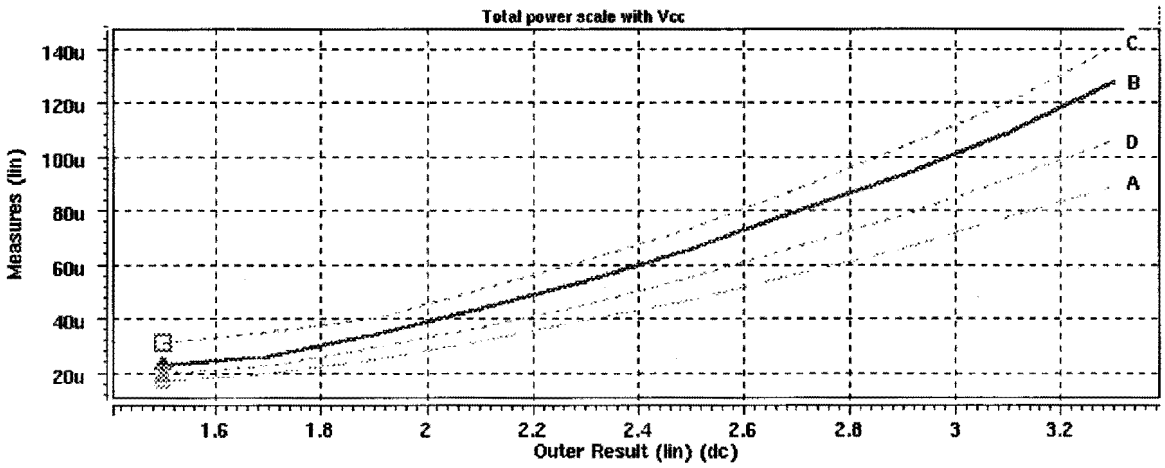

Figure 4.3: Average total power consumed by each circuit; a). Dynamic NRC, b). standard design with control gates, c). low-area and d). pseudo-static cascode pull-down.

4.3 Power-Delay Product

A tradeoff between speed and power has always been a big design concern. In high-performance and low-power applications, both features are equally important. The point of minimum power-delay product (PDP) is the point of optimal energy utilization at a given clock frequency. The power-delay product is what its name describes, the product of the delay and total power parameters. The propagation delay is determined by the speed at which a given amount of energy can be stored on the gate capacitors. The faster the energy transfer (or higher the power consumption), the faster the gate.

Therefore, the product of the delay and power will remain as a constant and can be considered as a quality measure of merit for switching devices.

To compare the optimal energy utilization of the four designs, the clock frequency is left constant at 100 MHz. As shown in Figure 4.4, the optimal energy consumed by the gate per switching event is scaled with the change in supply voltage. At Vcc < 2.0V, the power-delay product or PDP remains fairly the same for all cases. By increasing the supply voltage, the PDP asymptotically increases. However, both standard designs (B and C) consumes more energy than the two proposed designs while the pseudo-static design label D, uses slightly more energy than the dynamic design label A.

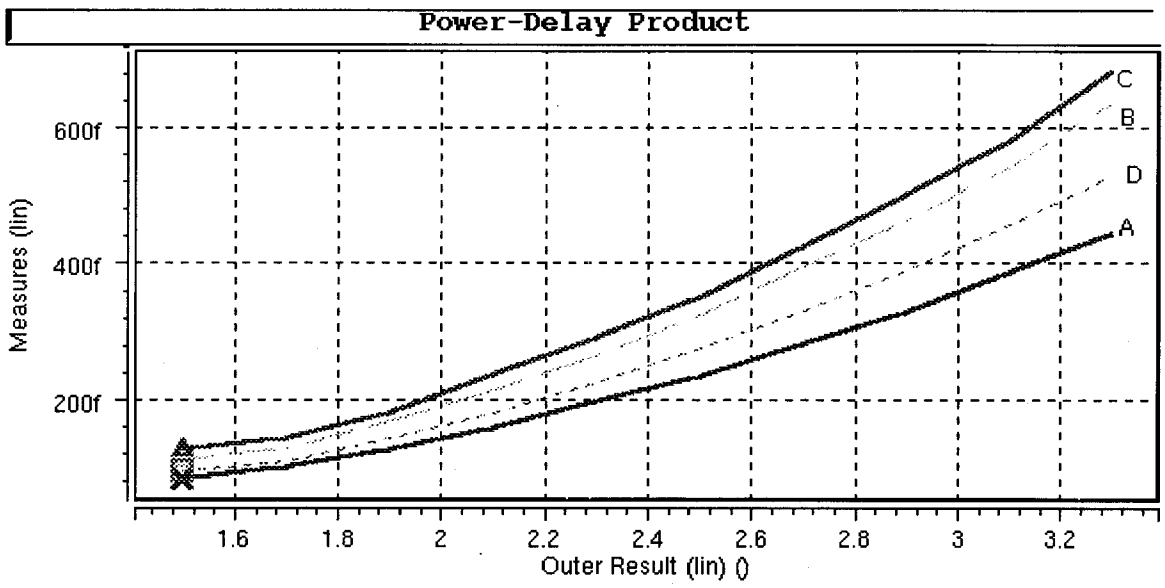

Figure 4.4: A tradeoff comparison can be made between speed and power; a). Dynamic NRC, b). standard design with control gates, c). low-area and d). pseudo-static cascode pull-down.

4.4 Effects of Scaling Vt

A way to increase performance without having to change architecture is to shrink the die size of the processor. In a shrink, the size of the transistors and interconnections are reduced and therefore, lowering the threshold voltage. This may result in higher speed and lower power dissipation. To model this scaling of threshold voltage, batteries are attached to the gate of the transistors. This will provide a larger current drive that would resemble a lowering of threshold voltage. Two cases are considered in this simulation: change due to scaling of threshold voltage and the effect of scaling down Vcc. As shown in Figure 4.5 and 4.6, the maximum frequency of the two standard designs scales linearly with change in threshold voltage. In fact, this pattern holds for the following two cases: Vcc=3.3 V and Vcc=1.5V. As discussed before, the dynamic NRC and pseudo-static cascode pull-down designs have problems with lower supply voltage, and this becomes apparent in Figure 13. Even at 3.3V, the two proposed designs will not scale with Vt. Again, this is the result of the dominated RC delay.

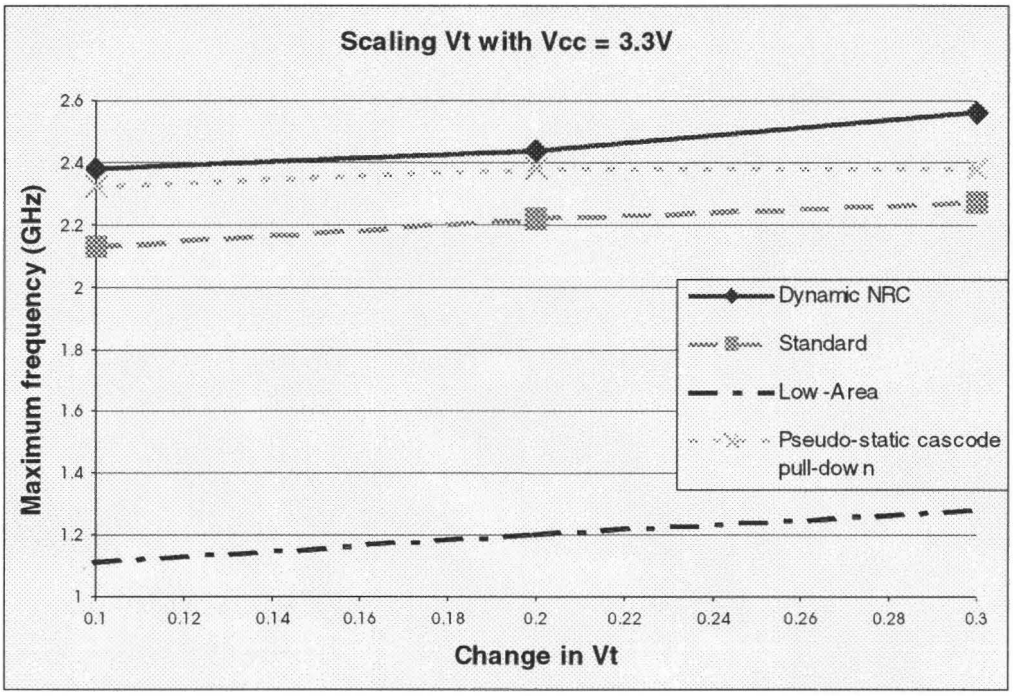

Figure 4.5: Describes how the four device scales with change in Vt at 3.3V; a). Dynamic NRC, b). standard design with control gates, c). low-area and d). pseudo-static cascode pull-down.

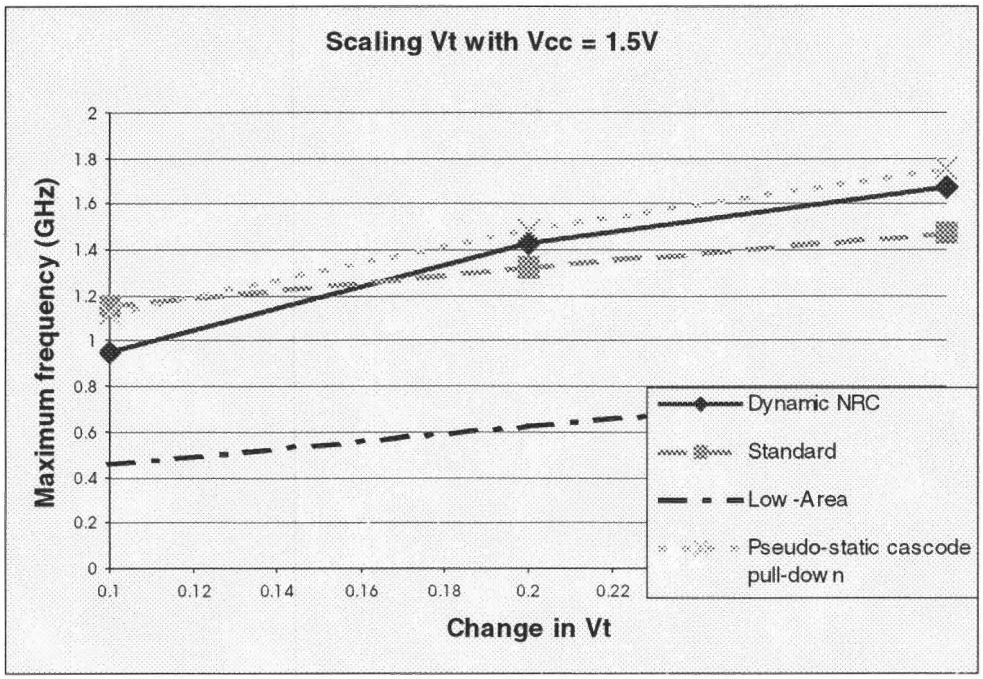

Figure 4.6: Describes how the four device scales with change in Vt at 1.5V. The two standard designs seem to scale linearly with change in Vt whereas the two proposed designs will not; a). Dynamic NRC, b). standard design with control gates, c). low-area and d). pseudo-static cascode pull-down.

4.5 Noise Margin

A simple method of obtaining the noise margin is to ramp up an input signal to the master latch and wait for the device to fail. By leaving the input transmission gate on and predefining the output node of the master inverter to high, V_{IH} is defined to be the point in which the ramping input signal overcomes either the level restorer or feedback inverter to switch the logic of the output node. V_{IL} is obtained through a similar method with the output node of the master inverter is predefined low while the input signal is ramped high. In this simulation, the voltage supply is fixed at 3.3V, and the frequency is standardized at 100 MHz. Table 4.1 compares the noise margin of the

four designs. As expected, the dynamic NRC and the pseudo-static cascode pull-down designs both sacrifice noise margin for speed, whereas the standard design possess the best noise margin property. The low-area design has good noise margin in the low region but poor in the high region.

Table 4.1: Static Noise Margin Comparison

	Dynamic NRC	Standard Design with control	Low-area	Pseudo-static Cascode Pull-down
NMH = VOH - VIH	1.99	2.6	1.8	2.1
NML = VIL - VOL	1.51	2.4	2.4	1.31

4.6 Metastability

In a typical design, the metastable level will exist for a short period of time and then resolve as the output moves to one of the stable states. If the flip-flop drives succeeding circuitry such as a gate or another flip-flop input, the metastable state can lead to a fault in the system. This peculiar behavior of clocked flip-flop with asynchronous inputs is a problematic event that depends on the clock frequency, the input data frequency, and the design of the flip-flop. There is no known method of constructing a

single flip-flop to avoid metastable levels when asynchronous inputs are present. An appropriate approach to this metastability problem is to recognize its existence and design systems that are unaffected by the occurrence of non-binary state.

The metastable region is made up of three timing parameters: Clk-Q delay, setup and hold time. The Clk-Q delay or the resolution time (t_r) measures the propagation delay from the falling edge of the driving clock input to the Q output signal. The setup time is the minimum time between a change in the data input to the triggering edge of the input clock signal such that, even under the worst conditions, the output Q will be guaranteed to change to the content represented by the stored data in the input. The hold time is the minimum time required such that, after the triggering of the latch, the data input signal must continue to remain constant so that the Q output remains stable by the end of the clock pulse. In terms of design criteria, "The question arises: how much can we let the Clk-Q delay be degraded in the metastable region and still benefit from the increase in performance (due to the decreased D-Q) while maintaining the reliability of operation?" [4]

In comparing the metastability of the four flip-flop designs, the mean time between failures, MTBF, is a quality figure of merit, which can be described by the equation:

$$MTBF = \frac{1}{f_D f_{CLK} T_o e^{-tr/\tau}}.$$

where f_D is the average frequency of the data and f_{CLK} is the frequency of the clock. Since the frequency of the clock is constant and so as the average frequency of the data, only the resolution time constant τ and the zero resolution time window T_o will be considered in the analysis. Again the simulation is standardized such that frequency of the clock is 100 MHz and the voltage supply is 3.3V. The frequency of the data is held constant since the result of the output is fed right back into the input. Figure 4.7 and 4.8 depicts the output high-to-low (hl) and low-to-high (lh) resolution time (t_r) versus the

setup time respectively. As discussed earlier, when the input data approaches the metastability point t_{meta}, the resolution time explodes exponentially. This phenomenon is shown in the provided plots.

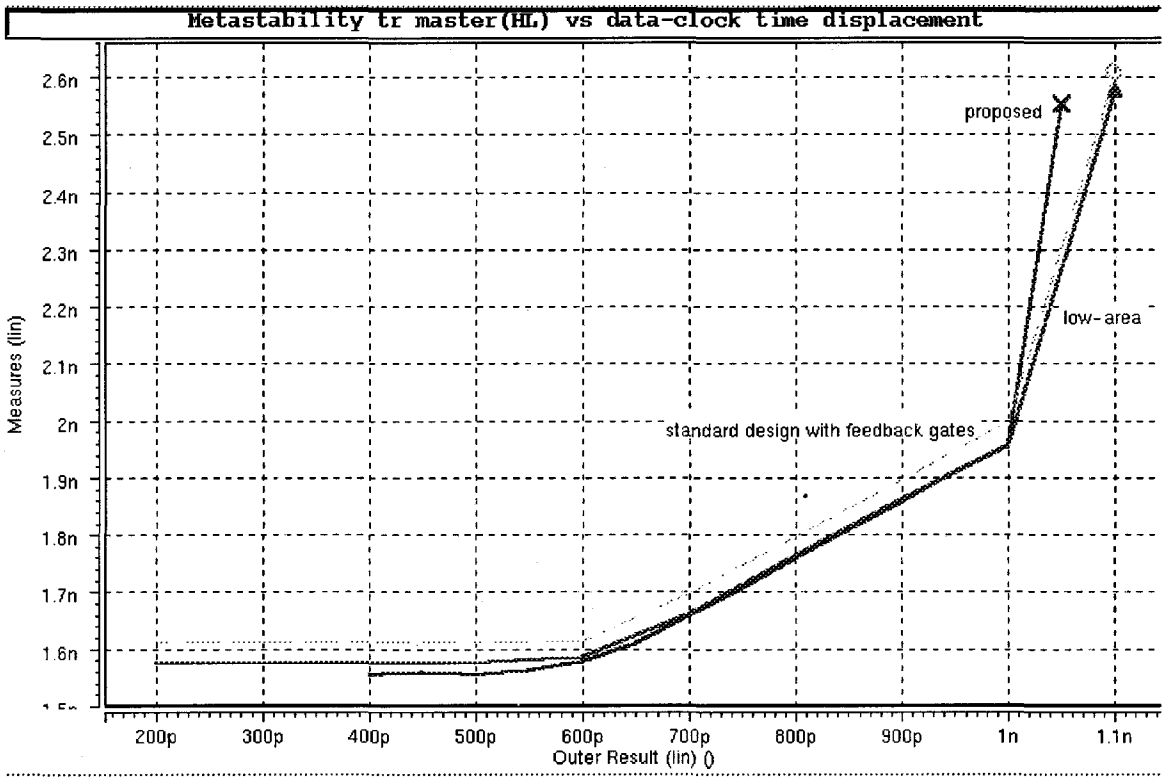

Figure 4.7: Master Clk-Q vs data-clock time displacement (High-to-Low transition)

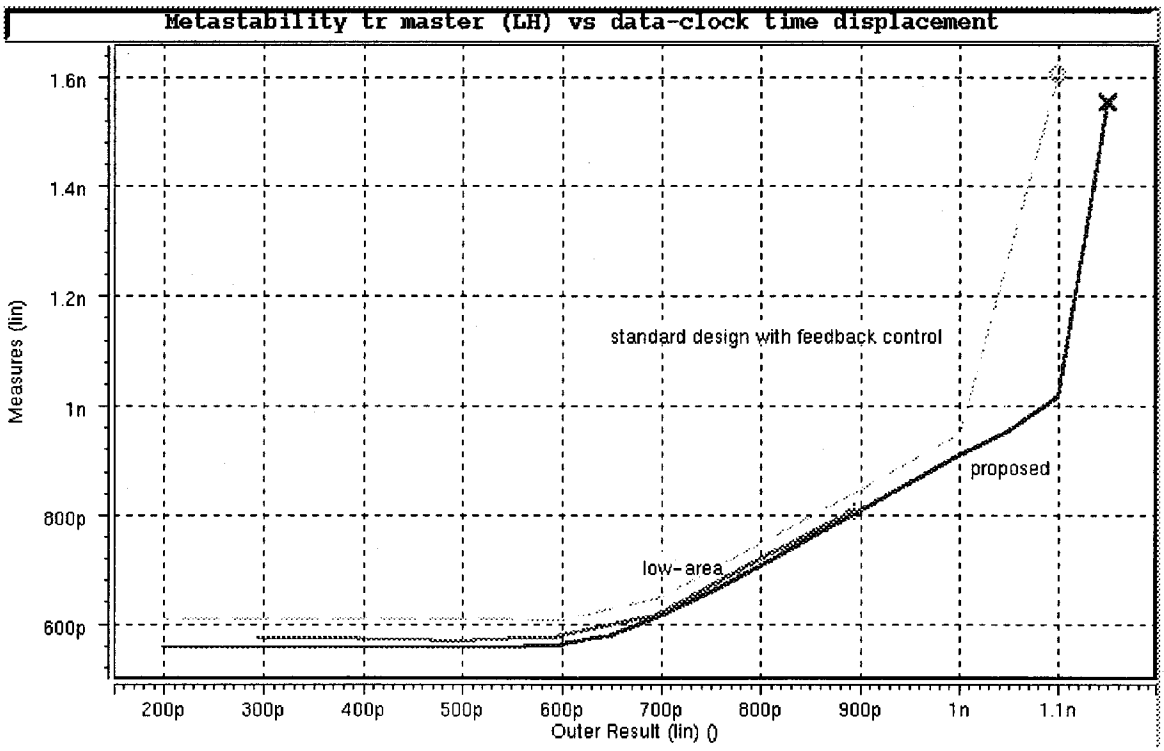

Figure 4.8: Master Clk-Q vs data-clock time displacement (Low-to-High transition)

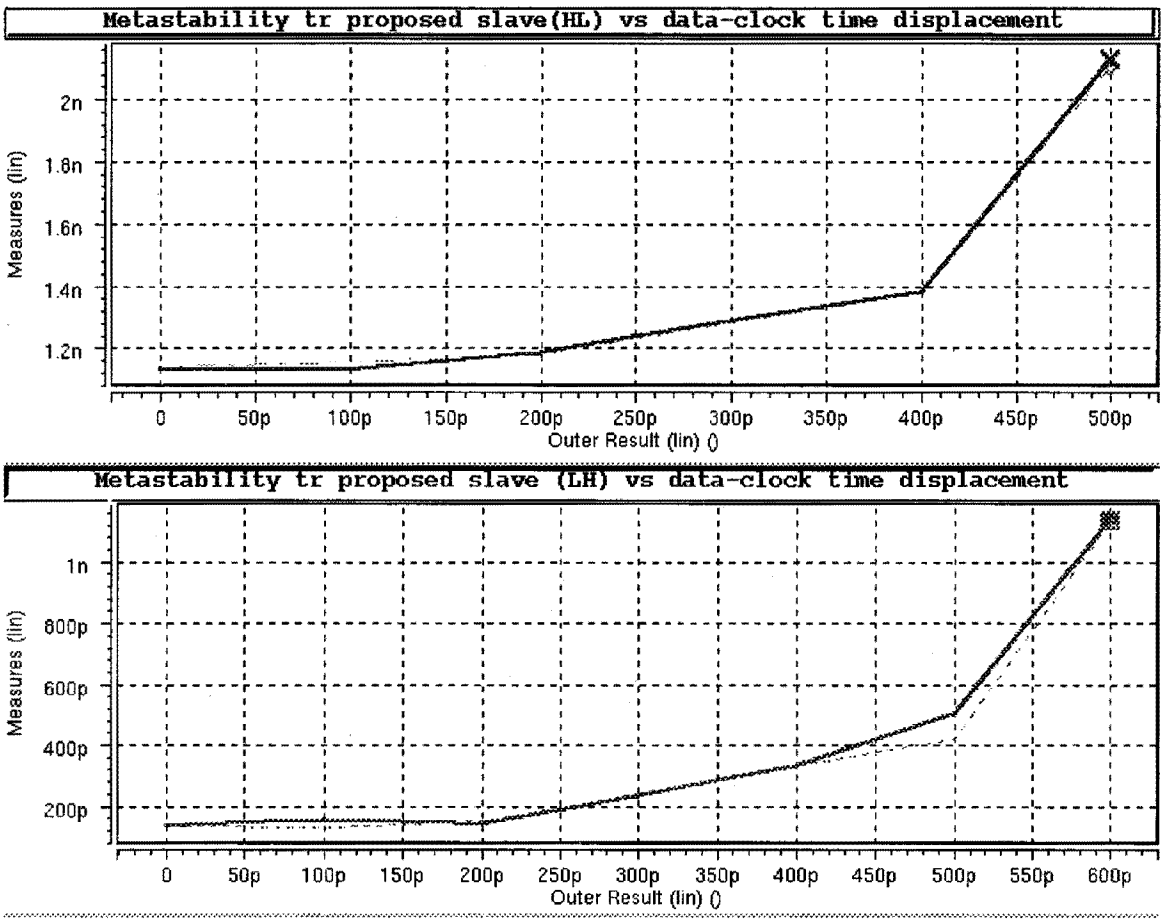

Figure 4.9: Slave Clk-Q vs data-clock time displacement

In the standard and low-area designs, the master and slave are symmetric to one another and therefore, only the master portion is considered. However, as shown in Figure 16, the metastability does not fair well for the master latch of the two proposed designs. This becomes more apparent as the resolution time constant and T_o are extracted. By plotting the resolution time over the log scale of the difference between t_{meta} and clock displacement, τ is obtained from the slope of the plot and $\ln(T_o/2)$ is the point in which the plot intersect with the X-axis. These plots are shown in Figure 4.10 and 4.11. Table 4.2 lists out the results of the two extracted parameters.

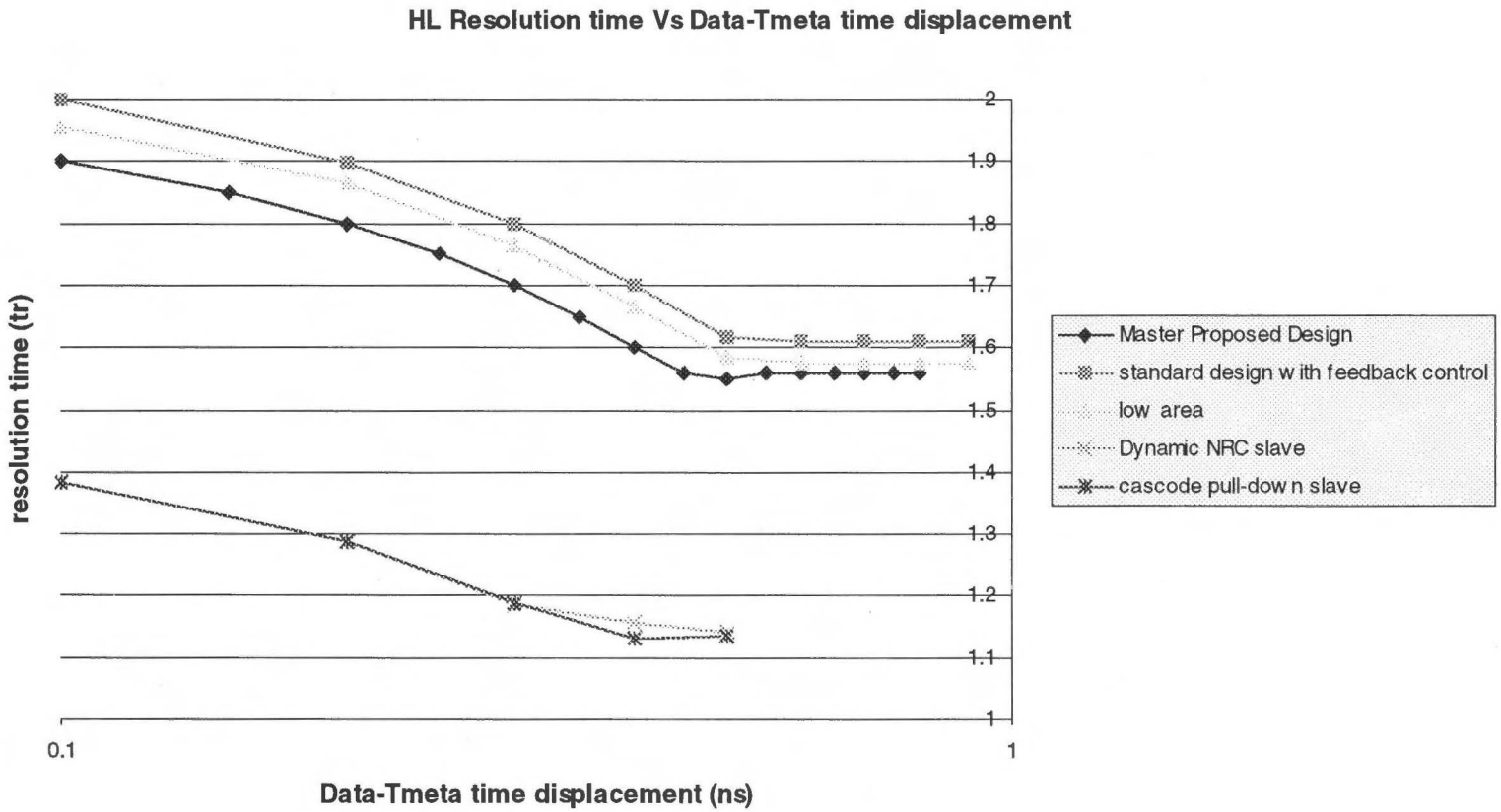

Figure 4.10: Resolution time vs. Data-t_{meta} time displacement for High-to-Low transition. Note, the lower performances of the metastability in the slave latches of the proposed design

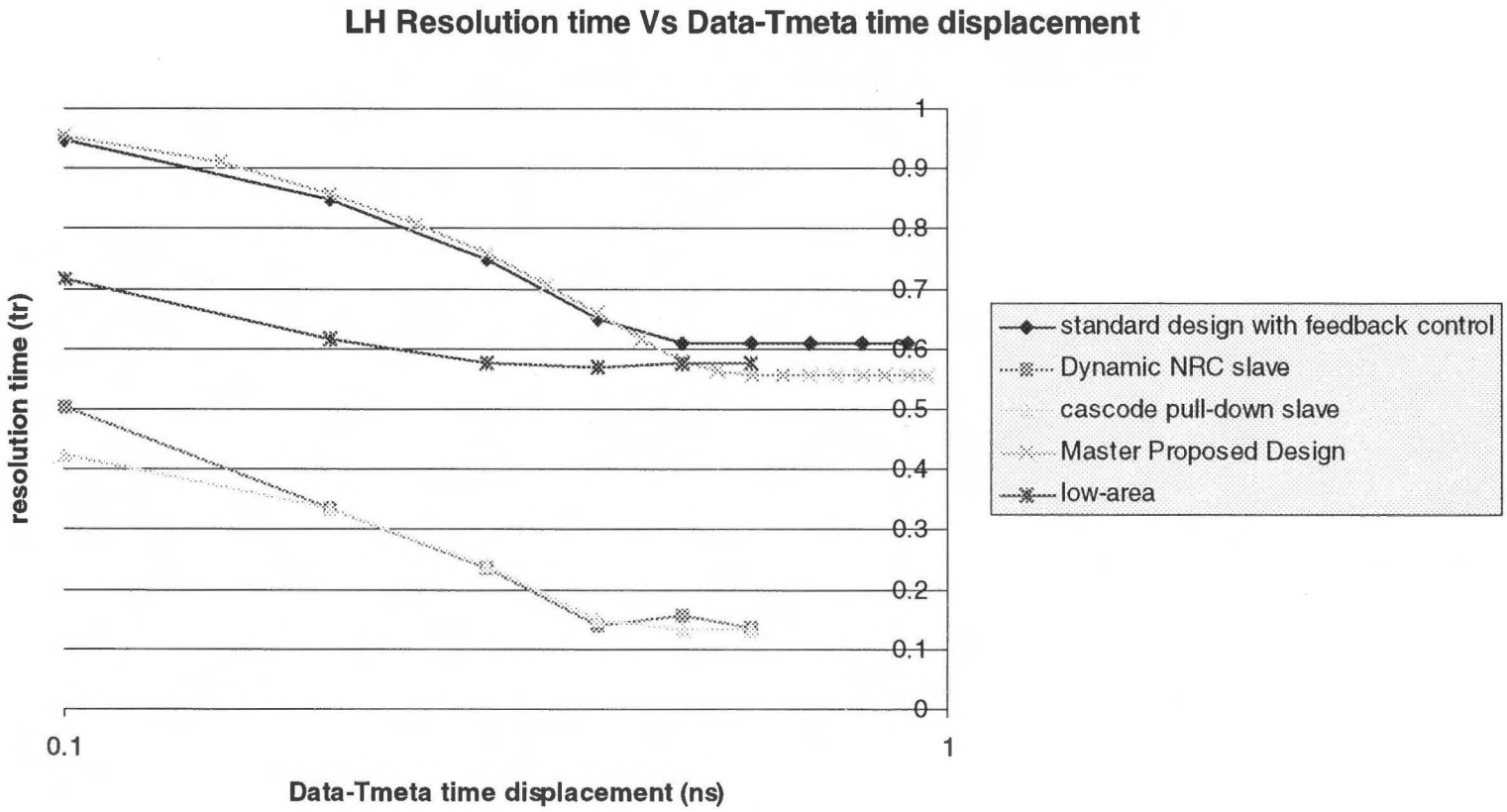

Figure 4.11: Resolution time vs. Data-t_{meta} time displacement for Low-to High transition.

Table 4.2: Metastability Parameter Results

Master (HL)	Resolution time constant	ln(To/2)	To
Proposed	1	2	14.77811
Standard Design with feedback control	0.913	2.2672	19.30467
low-area	1.0065	2.052276	15.5712

Slave (HL)	Resolution time constant	ln(To/2)	To
Dynamic NRC	1.607457143	1.1062	6.045699
Cascode pull-down	0.8579	1.705	11.00277

Master (LH)	Resolution time constant	ln(To/2)	To
Proposed	0.99986	1.0575	5.758328
Standard Design with feedback control	0.99425	1.0534	5.734767
low-area	0.705	1.1016	6.017953

Slave (LH)	Resolution time constant	ln(To/2)	To
Dynamic NRC	1.3411	0.4671	3.190722
Cascode pull-down	0.92995	0.5594	3.499245

To analyze the metastability of a circuit, only τ and T_o are considered. Referring back to the MTBF formula, smaller τ and T_o are much desired. The smaller the τ, the faster the binary signal is resolved. From Table 4.2, in comparing the metastability parameters of the four designs, the cascode pull-down design holds the best metastability properties. By using a single pull-up transistor to maintain data as shown in the proposed master flip-flop, little drop in metastability can be seen. This master latch uses a single transistor to maintain the capacitive charge, whereas the low-area uses two and even worse, the standard design requires four. In review of the slave latch

of the Dynamic NRC circuit, it becomes quite apparent that the pull-down network driven by a NMOS device may cause some problems for data transition and this is illustrated through poor metastability parameters. The difference in the metastability parameters depends on how hard a sampled signal must drive to overcome the pull-up or pull-down transistor. Improvement in the slave latch has been achieved by replacing the single pull-down transistor with a cascode amplifier as is in the Cascode pull-down circuit. As shown in Table 4.2, with this configuration, both τ and T_o are improved dramatically.

CHAPTER 5: CONCLUSION

The four designs mentioned in this thesis hold distinct properties that are each unique. When designing logic circuits for a cost efficient microprocessor system, scalability becomes an important issue. To maintain proper functionality, control logic must scale with process technology. If changes were to be made onto the die shrink factor, current logic implementations may generate false data and therefore, will require new modification. Since frequency is linearly proportional to supply voltage, scalability can be defined as whether or not a circuit maximum frequency of operation saturates at a given supply voltage. This reduction of performance is due to the dominated RC delay. When speed and power becomes nonessential, and noise immunity and scalability are emphasized, both standard and low-area designs pose to be the best solution. The standard design is commonly implemented in current microprocessor mainly because of their balanced structure. By having clock-controlled transmission gates to manage the pull-up and pull-down networks of the feedback element, contention between nodes is eliminated and therefore provides a faster data resolving time. However, in the low-area design, speed and power are drastically sacrificed to compensate for its reduced area. Without the clock-controlled transmission gates, sampling data must overcome either the pull-up or pull-down networks. Each time data is sampled, a direct short-circuit path exists between source and ground. This lead to higher power dissipation in the circuit. The symmetric balance of these two designs provides larger noise immunity and better scalability. For high speed, low power operations in an area-limited microprocessor, the Dynamic No Race Condition design is

the logical solution. The dynamic NRC design sacrifices its robustness in maintaining capacitive charge to maximize its operating frequency. A better answer to speed and low power is the Cascode pull-down flip-flop. This design has the advantage of being able to retain data in sleep mode with little loss in performance. Since the proposed circuit structure is no longer symmetric, noise margin and scalability are compensated for its high-speed low power operation. Because of its non-symmetric property, RC delay dominates at higher supply voltage, which causes frequency to flatten out and saturate. Overall, when choosing a flip-flop design, it all comes down to what the architecture of the system requires.

BIBLIOGRAPHY

[1] S. H. Unger and C. Tan, "Clocking schemes for high-speed digital systems," *IEEE Trans. Comput.*, vol. C-35, pp 880-895, Oct. 1986.

[2] Fred U. Rosenberger and Charles E. Molnar, "Comments on Metastability of CMOS Latch/Flip-Flop," *IEEE Journal of Solid-State Circuits*, vol. 27, no.14 pp. 128-130, January 1992.

[3] C.L. Portmann and T. H. Y. Meng, "Metastability in CMOS Library Elements in Reduced Supply and Technology Scaled Applications," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 1 pp. 39-46, January 1995.

[4] Vladimir Stojanovic and Vojin G. Oklobdzija, "Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 4 pp. 536-548, April 1999.

[5] Lee-Sup Kim and Robert W. Dutton, "Metastability of CMOS Latch/Flip-Flop," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 4 pp. 942-951, August 1990.

[6] Uming Ko and Poras T. Balsara, "High Performance, Energy Efficient D Flip-flop Circuits," September 1998.

[7] Kenneth L. Shepard and Vinod Narayanan, "Conquering Noise in Deep-Submicron Digital ICs," IEEE Design and Test of Computers, pp 51-62, January-March 1998.

[8] David Harris and Mark A. Horowitz, "Skew-Tolerant Domino Circuits," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 11 pp. 1702-1711, November 1997.

[9] Fred Rosenberger and Thomas J. Chaney, "Flip-Flop Resolving Time Test Circuit," *IEEE Journal of Solid-State Circuits*, vol. SC-17, no. 4 pp. 731-738, August 1982.

[10] L. R. Marino, "General Theory of Metastable Operation," *IEEE Trans. Computers*, vol. C-30, pp. 107-115, Feb. 1981.

[11] A. Chandrakasan *et al.*, "Low Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4 pp. 473-484, April 1992.

[12] T.J. Chaney and C.E. Molnar, "Anomalous Behavior of Synchronizer and Artiber Circuits," *IEEE Trans. Computers*, vol. C-22, no. 4 pp. 421-422, April 1973.

[13] G. R. Couranz and D. F. Wann, "Theoretical and Experimental Behavior of Synchronizers Operation the Metastable Region," *IEEE Trans Computers*, vol. C-24, no. 6 pp. 133-139, Feb. 1976.

[14] T. Karprzak and A. Albicki, "Analysis of Metastable Operation in RS CMOS Flip-flops," *IEEE Journal of Solid-State Circuits*, vol. SC-22, no. 1 pp. 57-64, Feb. 1987.

[15] Thomas J. Chaney and Fred U. Rosenberger, "Charactericzation and Scaling of MOS Flip Flop Performance in Synchronizer Application," *Process Caltech Conference Very Large Scale Integration (Pasadena, CA), 1979, pp 357-374.*