A Low Power, Reduced Swing Global Clocking Methodology

by

Farhad Haj ali asgari

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2003

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Farhad Haj ali asgari

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Farhad Haj ali asgari

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

As feature sizes become smaller, global interconnect wires dominate the overall delay of a chip. Unlike the transistor parasitics that become smaller in each new technology, global wires get longer due to increased chip complexity. Charging and discharging the large capacitance of these lines is a potential source of power consumption.

Clock interconnect wires in particular, are one of the most important parts of the global interconnections. They contribute to the length of these wires since the clock signal must be carried to most parts of the chip to synchronize all activities. The fact that the clock signal must be active constantly results in a major power consumption in the clocking networks. Depending on the type of the VLSI chip, from 20% to 50% of the power dissipated is consumed by the clock network.

This thesis presents the interconnection and clock distribution network related issues. The possibility of building a low power clock tree by using a low swing clock signal is examined. The clock swing is reduced only in global clock distribution network. A level converter buffer is introduced to serve this purpose and is compared with a conventional buffer.

The various stages of designing a clock distribution network are studied. This clock distribution network is the testbench for comparing the proposed levle converter buffer circuit with the conventional buffer. A power saving of $30 - 40$ % was achieved in 0.13 μm CMOS technology.

Acknowledgments

I would like to thank my supervisor, Professor Manoj Sachdev for his invaluable advice and help. His support and patience were the main reasons behind this accomplishment. His great sence of understanding is greatly appreciated.

I would also like to thank my friends Andrei Povlov and Ricky Yuen for their valuable advice and suggestions. And I would like to thank whoever I don't remember and in different times saved my time even small helps that as a matter of fact were great to me. I am always graceful to such people that save my day and making it invaluable with teaching me something.

Dedication

This thesis is dedicated to my grandmother who was the best grandmother in the world, in every aspect.

And to my parents that without their support and advises I could never have approached to my goals.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With decreasing minimum feature sizes in submicron technologies, device sizes are becoming much more smaller and this leads to the possibility of higher integration density. Integrating more devices in a specific area decreases the cost per gate. On the other hand, to achieve higher performance, microelectronic components and subsystems are moving towards ever increasing miniaturization. System on chip (SOC) design is becoming more and more desirable and semiconductor manufacturers continuously provide faster, more cost effective and more highly integrated devices. Examples of SOC include portable and wireless applications such as PDAs[1] or digital camera chips. According to international technology roadmap for semiconductors (2001 edition), die size of low power, low cost SOCs which are among the densest applications are expected to increase on average by 20% per node[2]

[1] A Personal Digital Assistant (PDA) is a small, mobile computing device providing services tailored to people "on the go". The most familiar of these applications have included personal organizer, calculator, alarm, and notepad.

[2] Each process technology is called a node. For instanse 180 nm process technologies called 180nm node.

through 2016 to accommodate increased functionality [1].

Ever so smaller feature sizes result in shorter local wires, but this is not the case for global wires. The length of the global wires increase with a factor of $S_C = \sqrt{A_c}$, where A_c is the area of the die. Long and wide global wires have especially large parasitic capacitances and, with current technology trends, they become more of a significant concern. Once considered to be electronically negligible, interconnections are becoming the dominant part of the overall delay. With smaller feature sizes, gate parasitics capacitances are smaller for new technologies so gate delay becomes smaller as well. At the same time, with larger die sizes, we have longer global wires, which in turn results in longer wire delays.

Table 1.1 shows that by linear conformal scaling global interconnect delay increases by a factor of $S^2 S_C$, while local interconnects RC delay remains unchanged [2]. S is the feature size scaling factor and is larger than 1. Figure 1.1 shows the delay for local and global wiring versus feature size [1]. This delay can be decreased by breaking down the interconnect line and using buffers, repeaters, or a combination of both. Figure 1.1 shows the effectiveness of using repeaters in decreasing the global delay. In addition to the propagation delay, the large interconnection parasitic also add to the total power consumption of the chip. This power, which is a dynamic power, is spent to charge and discharge the large parasitic capacitors of the global interconnections.

Clock distribution network interconnect wires form the major part of the global interconnections of the modern synchronous VLSI chips. They distribute the clock signal to almost all points of the chip to synchronize activities. With increasing parasitic wire capacitance and resistance due to larger chip areas, clock networks consume a significant part of the chip power. According to the research done, high performance clock networks might dissipate 20 – 50 % of the total power on

Table 1.1: Scaling of Local and Global Interconnections

Parameter	Scaling factor
Cross sectional dimensions	$1/S$
Resistance per unit length	S^2
Capacitance per unit length	1
RC constant per unit length	S^2
Local interconnection length	$1/S$
Local interconnection RC delay	1
Die size	S_C
Global interconnection length	S_C
Global interconnection resistance	$S^2 S_C$
Global interconnection RC delay	$S^2 S_C^2$

a complex chip [3] depending on different applications. This does not seem very attractive in low-power portable applications.

1.1 Thesis Outline

In this thesis different topics in signal distribution and interconnections are studied in general at first. There will be more emphasis on clock distribution networks as we proceed. Different ways of reducing the parasitics will be studied and buffering as a circuit solution will be discussed in particular. In addition to delay, the major concern in clock distribution network is power consumption by large parasitic capacitances of the global wires. The use of level conversion of the clock signal will be studied in detail as a solution for reducing power dissipated by the clock network. Robustness of the proposed circuit and the variations in power, transient

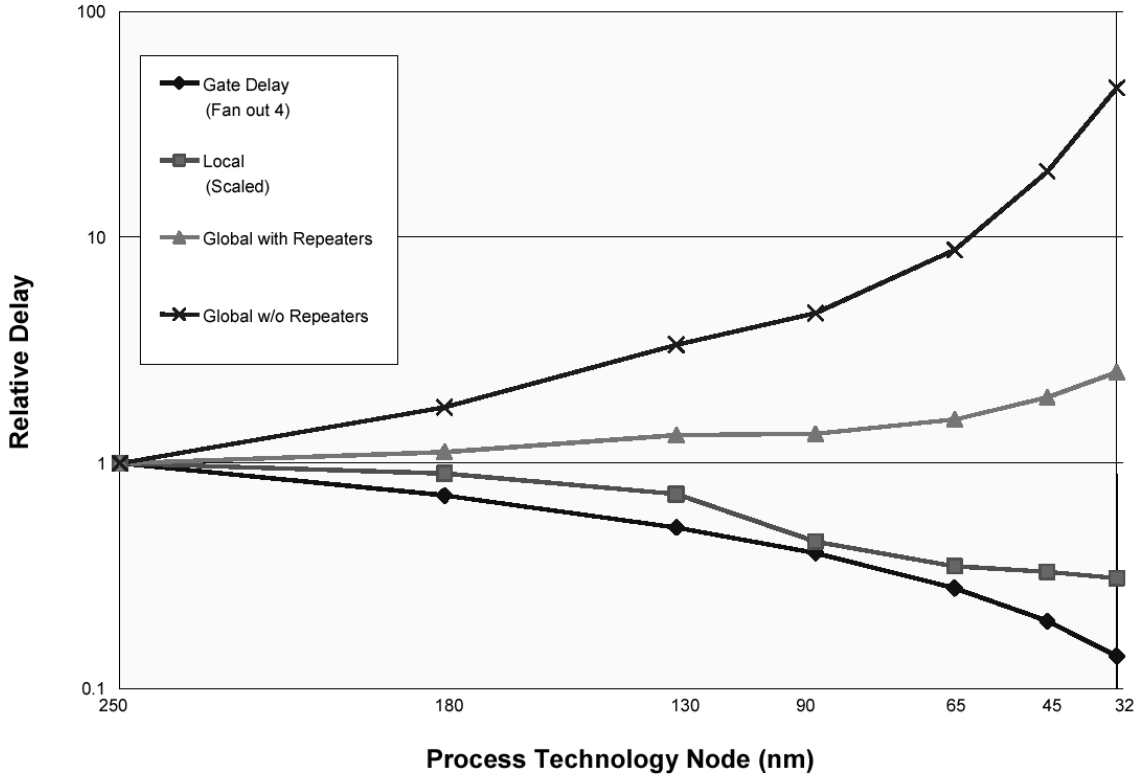

Figure 1.1: Delay for Local and Global Wiring versus Feature Size.

times, etc. due to changes in different circuit and environmental parameters will be examined. Following is the content of different chapters:

Chapter 2 is dedicated to the background topics and theories that are necessary for the following chapters. Interconnection impedances and the influence of technology scaling on these elements, clock distribution networks, and power related materials are among these topics.

Chapter 3 introduces a circuitry that is used as a buffer in the designed clock distribution network to save power consumption of the chip. This circuit is a level converter and reduces voltage swing to decrease the power dissipation. The

evaluation of the cell performance is presented in the following chapter but the operation of it is studied in this chapter. Other circuits that are used for reducing signal voltage swing are discussed briefly at the beginning of the chapter.

Chapter 4 is dedicated to the evaluation of the proposed level converter circuit. First the proposed level converter circuit is compared with the conventional buffer. A simple testbench is used in this part as a single cell driving different loads. Then a clock distribution network is designed with consideration of more practical issues. Both pre-layout and post-layout simulation results are presented for the latter case.

In chapter 5 as a conclusion, the advantages and disadvantages of the circuit designed is presented.

Chapter 6 mentions work that can be done in the future to study the behaviour of the circuit used in clocking network of a test chip and the limitations that are ahead of us to implement such a testbench.

Chapter 2

Background

In this chapter, the theory required to understand the rest of the thesis is discussed. First, different scaling trends and their impact on global delay are studied. Components of parasitic capacitances and solutions to reduce them are discussed next. Finally the theory of clock distribution networks and H-tree in particular is presented.

2.1 Interconnection Resistance and Capacitance

Once considered electrically negligible, interconnection delays are now a dominant factor in overall delay. Although with scaling, transistor capacitances decrease, however global wire capacitances increase as chip size increases. With smaller transistors, millions of transistors can be integrated on the same chip. With this possibility, designers want to integrate a complete system on a single chip to reduce the cost. Therefore, the die sizes increase, rather than decrease, as a consequence of scaling. The average interconnection length also increases since it is proportional

to the chip size. With increasing die sizes, long wires are required to connect subsystems and sub-blocks to each other or to a reference point like a system clock.

The 50% propagation delay between two gates connected by an interconnect wire can be simplified to the following equation:

$$t_{d_{50\%}} = 0.69 R_{tr} \left(C_{int} + C_{gate} \right) \tag{2.1}$$

where R_{tr} is the on-resistance of the driving gate, C_{int} is interconnection capacitance, and C_{gate} is the input capacitance of the receiver. With S and S_C as the scaling factors for the minimum feature size and chip size, the scaling properties of the delay components are [2]:

$$R_{tr} \simeq \frac{1}{\frac{W}{L} \mu C_{gox} (V_{DD} - V_{th})} \propto 1 \tag{2.2}$$

$$C_{gate} = \varepsilon_{ox} \frac{W_n L_n + W_p L_p}{t_{gox}} \propto 1/S \tag{2.3}$$

$$C_{int} = \varepsilon_{ox} \frac{W_{int} L_{int}}{t_{ox}} \propto S_C \tag{2.4}$$

W_n, L_n,W_p, and L_p are the width and length of the transistors, μ is the mobility of the carriers, V_{DD} and V_{th} are the power supply and threshold voltage of a transistor, C_{gox} is the gate oxide capacitance per unit area, t_{gox} and t_{ox} are the gate oxide and dielectric thicknesses, and W_{int} and L_{int} are the length and width of interconnection.

S and S_C are both greater than one. So, on one hand, device capacitances decrease and, on the other hand, interconnect capacitances increase. In contemporary chips designed in submicron technologies, wires with lengths as short as even

a fraction of millimeter have capacitances comparable to device capacitances. This fact prevents designers from neglecting these wires.

The interconnection capacitance has three components (Figure 4.4):

- parallel plate capacitance or area capacitance (C_{p-p} or C_a)
- fringing capacitance (C_f)
- wire-to-wire coupling capacitance (C_c)

Capacitance of a single wire can be modeled by a parallel plate capacitor with a width $W_{int} - \frac{H_{int}}{2}$ and a cylindrical wire with a diameter of H_{int} [4]. Based on this model, the interconnection capacitance per unit length can be calculated as:

$$C_{int} = \frac{\varepsilon_{ox}\left(W_{int} - \frac{H_{int}}{2}\right)}{t_{ox}} + \frac{2\pi\varepsilon_{ox}}{\ln\left[1 + \frac{2t_{ox}}{H_{int}}\left(1 + \sqrt{1 + \frac{H_{int}}{t_{ox}}}\right)\right]} \tag{2.5}$$

The first expression is the equation for a parallel plate capacitor in the model and the second one is for the capacitance of the cylindrical wire. Figure 2.1 plots C_{int} based on equation 2.5 as a function as W_{int}/t_{ox} for two H_{int}/t_{ox} ratios [5]. H_{int} and t_{ox} are wire and dielectric thicknesses, respectively. As can be seen, as W_{int} approaches values close to t_{ox} or, in fact, H_{int}, C_{int} stops decreasing and reaches a constant value of about 1 pF/cm.

To improve integration density and, at the same time, keep propagation delays constant, it is desirable to keep the conductor and oxide thicknesses constant and to reduce the wire width and spacing. In these fat wires, fringing and coupling capacitances together are much larger than parallel plate capacitance and are the dominant components of the total wire capacitance. Figure 2.2 plots the computed

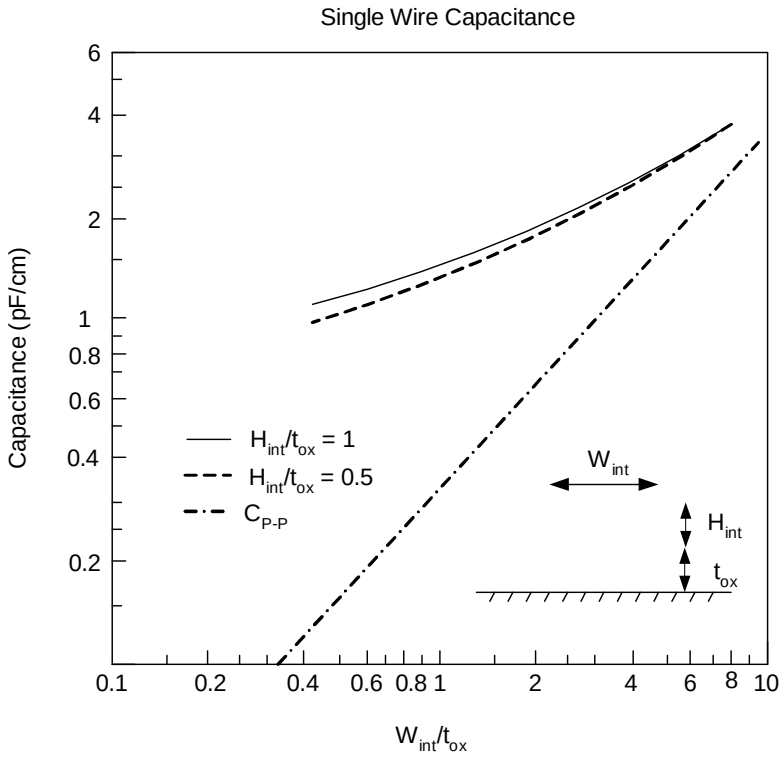

Figure 2.1: Capacitance of a single wire including firinging capcitance. The total capacitance for two different T/H is compared with the parallel plate cpacitor alone.

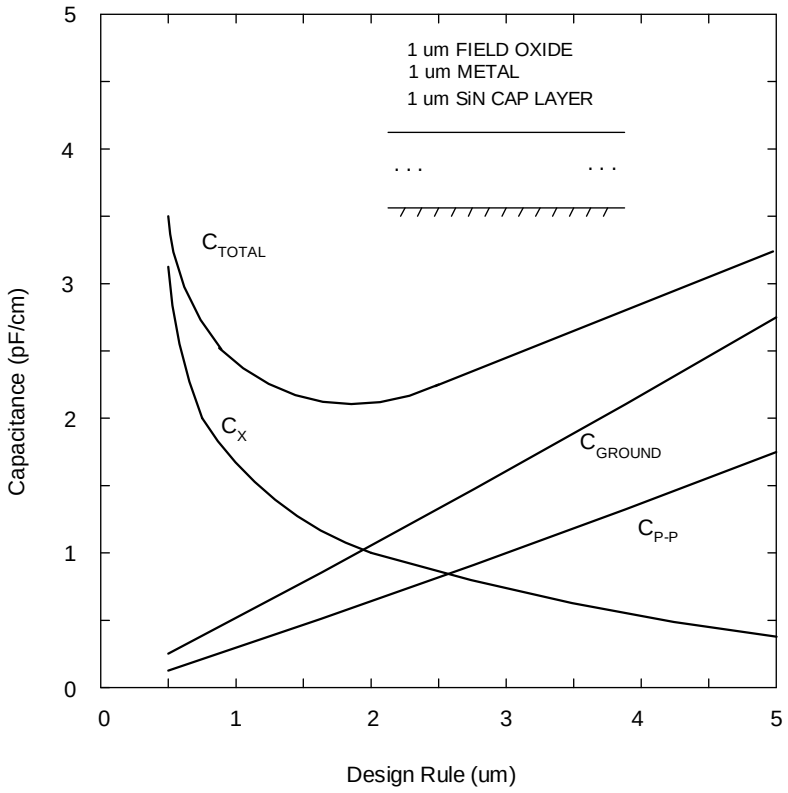

Figure 2.2: Plots of the total capacitance and the components of it for a metal line running in an array of lines with spaces equal to their width.

capacitance of a wire running in parallel with an array of wires with the same width and spacing [5]. Capacitance values are plotted for wire widths from 0.5 to 5 μm. As can be seen, the minimum value of the total capacitance is about 2 pF/cm.

From the previous two paragraphs a few points can be concluded. One way to reduce the wire capacitance is to make wires narrower, but as the wire width gets comparable to dielectric thickness, the capacitance approaches to a constant value. At such dimensions fringing and coupling capacitances are dominant components of the total wire capacitance which, are independent of the wire width.

Table 2.1 shows the impact of scaling factors on different components of VLSI chips. As can be seen, interconnection resistance increases with a factor of S^2S_C. This increased resistance causes an increase in propagation delay and power dissipation of the wires. Power dissipation of this increased resistance is crucial if we have a low power application. With wire parasitic resistances comparable to R_{tr}, the on-resistance of the drivers, the global interconnection delay becomes the dominant component. Therefore, the propagation delay increases proportionally to the square of the interconnection length, as we have $S^2S_C^2$.

Table 2.1: Global Interconnections Scaling.

Parameter	**Scaling factor**
Cross sectional dimensions	$1/S$
Resistance per unit length	S^2
Capacitance per unit length	1
RC constant per unit length	S^2
Local interconnection length	$1/S$
Local interconnection RC delay	1
Global interconnection length (l_{int})	S_C
Global interconnection capacitance (R_{int})	S_C
Global interconnection resistance (C_{int})	S^2S_C
Global interconnection RC delay	$S^2S_C^2$

2.2 Solutions for Reducing Propagation Delays

Long wires with large distributed resistance and capacitance produce large propagation delays in contemporary VLSI chips. There are different approaches to solve

this problem. The first step is to reduce parasitic elements of the interconnect wire. Another solution is dividing the wire into smaller sections and insert buffers between them that reduces the delay by N, the number of buffers.

2.2.1 Process Technology Solutions

To reduce propagation delay, we need to reduce parasitic capacitance and resistance. Increasing dielectric thickness helps reduce the capacitance to some extent. But, from the previous section, we know that as the dielectric thickness becomes comparable to the interconnection width and thickness, this method does not help any more. This is because of fringing capacitance becomes the dominant component of the total capacitance. The other approach is to reduce the capacitance using low-K dielectric materials.

To reduce the propagation delay of interconnect wires, the other component that can be reduced, is the resistance of the line. This can be accomplished by following two different techniques: using wires with lower resistances or increasing the number of metal layers.

As chip sizes become larger and the number of transistors on VSLI chips increase, we need more metal layers to connect these transistors. Multilayer interconnection can also be considered a technology solution to reduce the propagation delay in global interconnections. First, more metal layers present wires with lower resistances than that of polysilicon or diffusion wires. Second, upper metal layers with thicker wires have optimum RC delays that can be used to route global wires. Lower level wires can be used for local interconnections. As already mentioned, multilayers of metal make interconnections more efficient and denser and, as a result, the chip size becomes smaller. Since the average interconnection length is

proportional to the chip size this further improves the propagation delay. Figure 2.3 shows a seven layer interconnection scheme. The typical usage of different layers are shown as well. As can be seen, two upper level layers that are thicker are used for global routing.

Figure 2.3: A mutilayer inteconnection scheme with seven metal layers.

In $0.13\mu m$ process technology, there are eight metal layers. Metal layers seven and eight are almost three times as thick as the lower layers, so their resistance is three times smaller. This makes them very convenient for routing global wires. Figure 2.4 is a microphotograph of interconnection layers with dielectric layers removed. The upper layers that are wider and thicker are used for global wires. As can be seen, the local interconnections at the bottom are much narrower because they connect transistors together and to adjacent units.

2.2.2 Circuit solution

Consider an interconnect wire with distributed parasitics C and R. The time constant of this wire will be $0.4RC$:

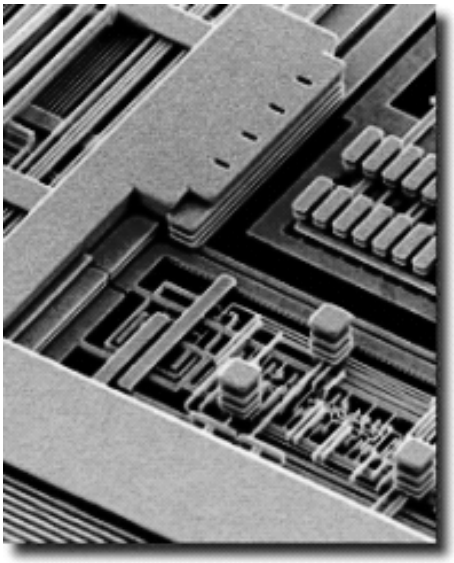

Figure 2.4: A photograph of the interconnection wires on a chip with the dielectric etched away.

$$t_d = 0.4RC \tag{2.6}$$

Now if we break this wire down into n segments and isolate these sections from each other with buffers, the new parasitics of each section are $\frac{R}{n}$ and $\frac{C}{n}$. The new propagation delay will be:

$$t_{d_{buffered}} = n\left(\frac{0.4RC}{n^2} + t_{buffer}\right)$$

or

$$t_{d_{buffered}} = \frac{0.4RC}{n} + n \times t_{buffer} \tag{2.7}$$

Comparing Equations 2.6 and 2.7, it can be seen that the propagation delay decreases by an order of n using n buffers. Of course the delay of the buffers should be taken into account and this is a factor used to calculate the optimum value for n. The added area and power dissipation of the buffers are not significant in CMOS technology.

Repeaters and cascaded drivers are usually used as buffers in VLSI circuits. Cascaded drivers are suitable for large capacitive loads and repeaters are useful for RC loads. Cascaded drivers are usually used to drive off-chip capacitances.

Figure 2.5 shows the configuration of a set of repeaters used to reduce the propagation delay. The interconnection has been divided into k sections. C_{gate} and R_{tr} are input capacitance and output resistance of the buffer, and R_{int} and C_{int} are the resistance and capacitance of the interconnection. The optimum value for k can be calculated as [2]:

$$k = \sqrt{\frac{0.4 R_{in} C_{int}}{0.7 R_{tr} C_{gate}}} \tag{2.8}$$

In order for this method to be applicable the delay of each wire segment must be at least seven times as large as the delay of the buffer.

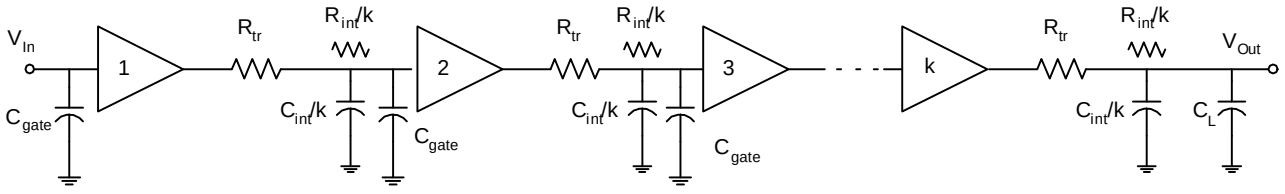

Figure 2.5: Repeaters configuration for driving an interconnection line with total resistance and capacitance of Rint and Cint.

2.3 Clock Distribution Networks

Even a moderately complex design has several thousand clock-driven transistors. In clock distribution networks, closely spaced points are usually grouped together and we concentrate on driving the groups instead of the individual points. These subgroups are called sub-blocks. A clock signal must be distributed to the sub-blocks on a chip. The sub-blocks are designed small enough so that the skew within the block is tolerable. The clock signal might be buffered in a sub-block and distributed by a clock distribution network i.e. a grid-based network.

The clock distribution network that distributes the clock signal to the sub-blocks is called the global clock network. The wires in the global clock distribution network are longer and we need a clock distribution topology to minimize the total capacitance and the clock skew. The clock distribution network in sub-blocks is known as the local clock network. This network has higher fan-out and larger capacitance. The clock skew between the nodes should also be small. The local clock buffers are more susceptible to noise.

One of the most useful distribution geometries in high-density VLSI is a H-tree. Figure 2.6 shows the fundamental unit of a H-tree clocking distribution network. The distance from point A to every corner is constant. So for a signal sent from point A, the arrival time to every corner is the same. In other words, all of the received signals are in phase. A clock network can be constructed by adding H structures to the end nodes repeatedly, as shown in Figure 2.7. In this structure the clock signal is delayed by an equal amount for each sub-block since all blocks are equidistant from the clock source. Therefore the clock skew is theoretically zero. More material relating to H-tree clock distribution networks is discussed in chapter four when a clock distribution network is designed as a testbench.

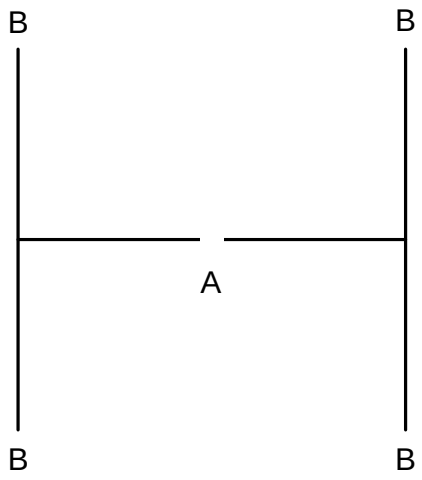

Figure 2.6: Basic unit of H-tree clock distribution.

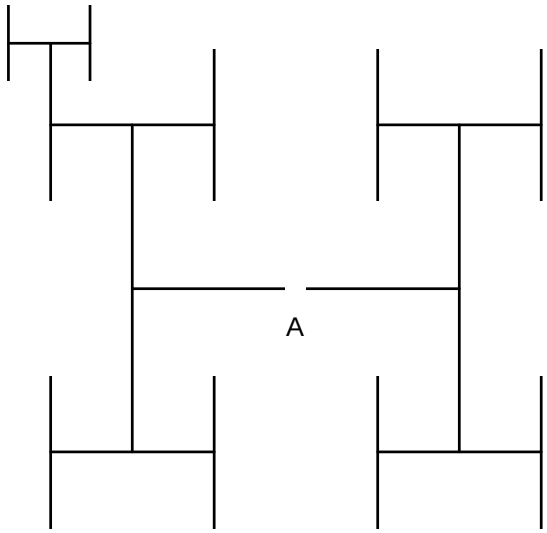

Figure 2.7: By adding smaller H-trees to the end nodes of the previous H geometry a clock tree with minimum skew can be constructed. Only one of the third possible extensions of the H-tree is shown in this figure for the sake of simplicity.

2.4 Power Dissipation

Building a clock distribution tree is imperative to ensure zero skew and a sharp slew rate for the clock edge. This requires inserting buffers within the clock network to isolate the downstream capacitance, thus reducing the transition times. The clock distribution network is spread all over the chip and long wire lengths are unavoidable.

Large capacitances of the long global interconnect wires dissipate considerable power [6], especially when they carry a clock signal. Activity factor of the clock signals is practically one unless clock gating is used to decrease it. So charging and discharging of clock distribution wires all the way could be a major source of power consumption of a chip.

Studies have shown that high performance clock networks may dissipate 20-50% of the total power on a complex chip [3]. In this context, the growing importance of low power designs for portable electronics, makes it necessary to develop strategies to reduce the power dissipation of the clock network .

The total power dissipation in a clock network consists of three components:

- static power dissipation,
- dynamic power dissipation, and
- power dissipation due to the leakage current.

The leakage current is dependent on the technology and is a relatively small component in a clock network. Similarly, keeping proper rise and fall times throughout the clock tree may also minimize static power component. Therefore, the clock network power consumption can be approximated as:

$$P = fC_L V_{DD} V_{sw} \tag{2.9}$$

where f, C_L, V_{DD}, Vsw are respectively the clock frequency, the load capacitance of a given node, the supply voltage, and the output swing of the buffer. If the output of the buffer swings from *ground* to V_{DD}, then $V_{sw} = V_{DD}$, and the formula reduces to:

$$P = fC_L V_{DD}^2 \tag{2.10}$$

Often f cannot be changed without significant architectural changes. Therefore, the power dissipation of the clock network can only be reduced by:

1. Reducing the total load capacitance, C_L, on all nodes.
2. Reducing V_{DD}, which creates a quadratic reduction. If V_{sw} is also simultaneously reduced by the same factor.
3. Reducing V_{sw}, without reducing V_{DD}, which corresponds to a linear reduction in the power dissipation.

C_L, a combination of interconnection capacitance and receiver gate input capacitance, is out of our hands and has already been reduced using techniques in process technology. In order to achieve power savings, we choose reducing Vsw without reducing V_{DD}. It does not require multiple power supplies, and hence, it is easier to implement. Implementing a second voltage power supply for driving a clock network with a large current drive would be quite challenging.

In the following chapter, a level converter circuit is introduced that can be used to reduce the voltage swing across the interconnect wires and reduce the power

consumption of the network. This level converter is used to reduce the power dissipation of the global clock distribution network. A low-swing clock signal in the local clock distribution network can be degraded as local wires are more susceptible to noise. Large clock fan-out and need to redesign flip-flops to accept low-swing clock input are the other reasons which, make the idea of reducing swing in the local clock distribution networks undesirable.

Chapter 3

Level Converter buffers

Reducing signal swing helps to reduce the power consumed by the parasitic capacitance of the interconnect wires. There are different methods and each method uses different circuits. A short list of different methods includes:

- Cutting either tail or head of the signal
- Changing both high and low logic levels
- Using differential driver/receiver circuit

The last method is commonly used in memories by implementing sense amplifiers. Using this method for clock distribution networks causes an increase of about 30% in chip area, because it requires double rails for clock transmission. This disadvantage makes differential circuits undesirable for clock distribution networks.

There are many circuits that use the first method[3]. The major challenge here is the fact that gaining the desirable reduction in voltage swing requires cutting the signal at the chosen level aggressively. This, in addition to increasing static

power dissipation, makes recovering full swing clock signals at the receiver difficult and adds to the complexity of the receiver. In this study, we try to stick to simple receivers as the number of end node receivers is much higher. This approach could save the power consumption by saving a little power in many cells which, leads to a substantial saving.

The second method seems the most promising and a convenient one for clock distribution networks. In the following section, different ways of implementing a circuit that reduces voltage swing symmetrically is discussed briefly. The reason the proposed circuit was selected is also discussed.

3.1 Symmetrical Level converters

Previous reduced swing clocking schemes can be categorized into two groups [7]:

1. dual power supply voltage, and
2. regular power supply voltage.

On one hand, the first group achieves the reduced swing by using a separate reduced supply voltage, usually generated on chip. This adds extra area and complexity to the overall chip layout, which is not desirable [8]. The advantage of having a separate power supply for generating the reduced voltage swing is in the reduced number of the overhead transistors, which leads to improved power saving [9]. On the other hand, the second group attempts to achieve the reduced swing voltage with circuit methods. However, the design of reduced swing buffers becomes challenging in the absence of the second power supply.

3.1.1 Level converters with Double Power Supply Voltage

Figure 3.1, shows a non-inverting buffer that drives an interconnect line with line capacitance and resistance of respectively R_{int} and C_{int}. The reason that two inverters are used in the buffer is explained in section 4.1. In clock distribution networks, we need to distribute the second power supply to all points that need a buffer. As such, additional area is added to the chip layout in this method.

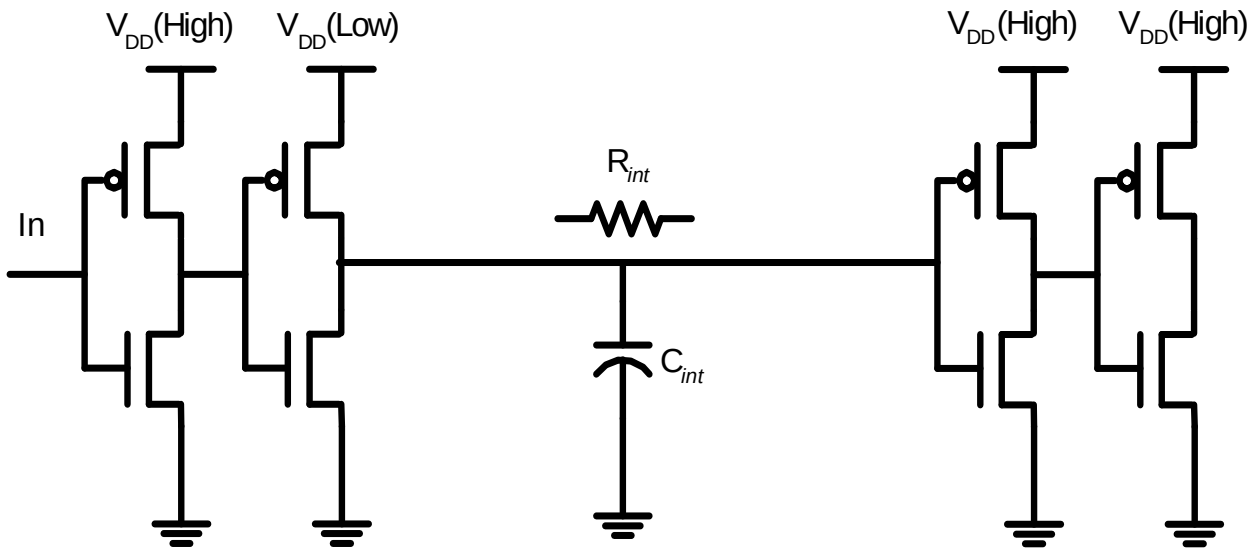

Figure 3.1: A pair of driver and receiver for driving an interconncet line using dual power supply

3.1.2 Source Follower Repeater, Single Power Supply Voltage Level converter

Most single voltage-supply reduced-swing buffers utilize a pMOS transistor for passing a low logic level and a nMOS transistor for passing a high logic level [3],[10].

This configuration is called a source follower buffer since it is similar to the source follower amplifier.

Figure 3.2, shows a transmitter/receiver pair using source follower configuration as a level converter. The two converters in the first stage are for restoring full swing voltage as is necessary for a repeater. One of the most serious drawbacks of this technique is the poor performance of reduced swing buffer [11]. Unlike the conventional inverter here the nMOS transistor and the pMOS transistor are connected to V_{DD} and *ground* respectively. nMOS is very good at passing V_{SS}, but not as good at passing V_{DD}. Figure 3.3 shows a nMOS transistor with its gate connected to V_{DD} and its input, V_A, to V_{DD}. This circuit resembles a source follower configuration with a nMOS transistor connected to V_{DD}. In this circuit, the transistor turns off when the voltage across C_L reaches $V_{DD} - V_{th_n}(V_B)$ in which V_{th_n} is a function of V_B, the substrate voltage. So we can get a lower voltage for high logic level.

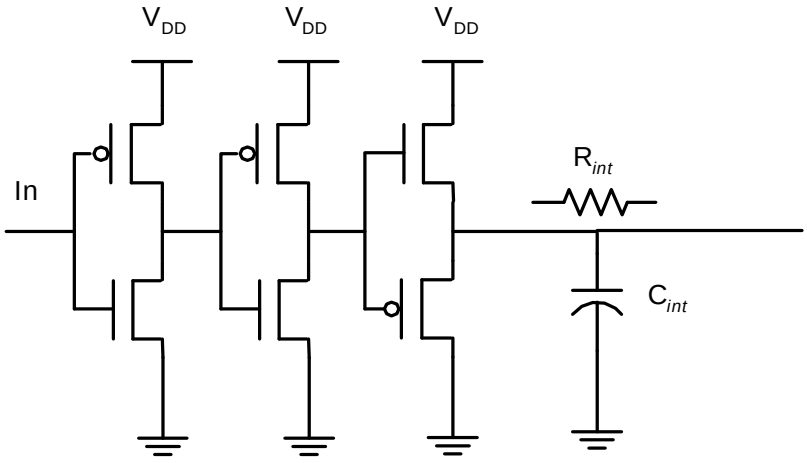

Figure 3.2: Level convertor using a source follower buffer

The disadvantage of this circuit is that as V_B reaches to voltage $V_{DD} - V_{th_n}(V_B)$,

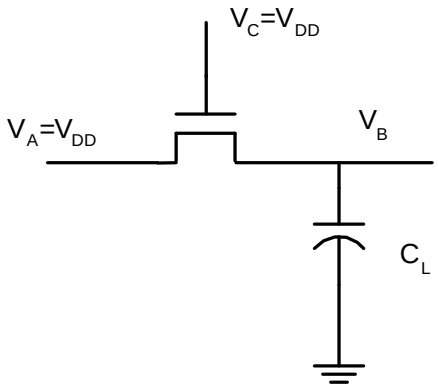

Figure 3.3: An nMOS transistor when passing level logic one.

the current through C_L decreases. The nMOS transistor operates in the linear region in this conditions. Equation (3.1) shows the drain-source current, I_{DS}, of an nMOS transistor in the linear region:

$$I_{DS} = \mu C_{gox} \frac{W}{L} \frac{(V_{GS-} V_{th})^2}{2} \tag{3.1}$$

Here V_{GS} equals $V_C - V_B$ and as V_B reaches its final value V_{GS} gets smaller and this reduces current drive ability of the circuit in Figure 3.3. As a consequence, if slows down the waveform as it approaches V_{DD}. With the same analogy, we can conclude that the pMOS transistor in source follower cuts off the low logic level from 0 to V_{th_p} and similarly has poor current drive when the output reaches the new low logic level.

Therefore in a source follower buffer where, the nMOS/pMOS transistor is connected to the $V_{DD}/ground$, current-drive is very weak and even large transistors result in very slow waveforms at the output. Figure 3.4 shows the output waveform of a source follower buffer. As can be seen, the output changes rapidly at the beginning of the switching of the input, but slows down as output reaches its

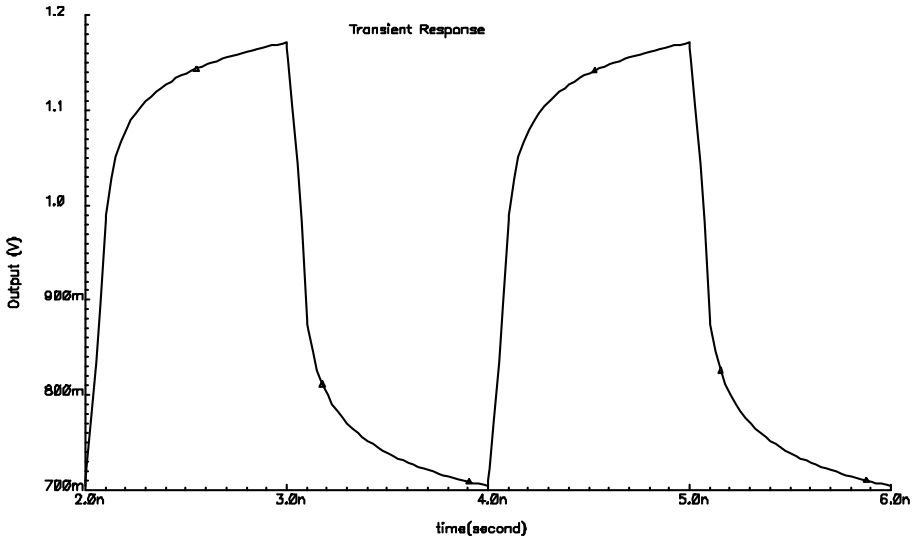

Figure 3.4: Output waveform of the source follower buffer.

final value. So, the source follower buffer has a long propagation delay and poor transient times. The slow rise/fall time results in a larger direct path[1] static power consumption as both the pull-up and pull-down are simultaneously on for a longer period of time. As a consequence, such buffers are not suitable for high performance clock distribution networks.

In a clock distribution network, the number of downstream buffers are significantly larger than the number of upstream buffers. Therefore, a RS-FS (reduced-to-full swing) restorer with a simple design becomes critical in reducing power consumption. More power efficient RS-FS buffers (simple with less area and transistor count) save power due to their large population.

[1]The finite slope of the input signal causes a direct current path between V_{DD} and *gnd* for a short period of time during switching, while the nMOS and the pMOS transistors are conducting simultaneously.

Since in any realistic clock distribution network, the number of downstream buffers increases significantly, a simple design becomes critical in achieving power and energy savings. More power efficient receivers (simple with less area and transistor overhead) save power due to their larger population. The reason that there are more leaf-node buffers is partly because downstream branches of a clock tree at the end of the tree have smaller lengths. Capacitances are smaller and most of the buffers are concentrated at the end nodes.

The output stage of the circuit in Figure 3.2, the source-follower, has slow transient time due to the reduced over drive voltage ($V_{GS} - V_t$). Therefore, in the receiver side, a complex circuit is necessary to convert the reduced-swing to full-swing. This extra complexity is necessary to speed up the transition time and reduce direct-path current time.

Since the transition time in a source follower buffer is slow, the receiver side in a low swing clock tree needs to be more complex to restore the low-swing slow waveform. This is in contrast with our goal of having simple receivers. We need to design a faster transmitter and repeater that let us use smaller and more efficient receivers. In the next chapter we will describe a level-converter circuit that generates low-swing output with sharp transient times.

3.2 Proposed circuit

We need an output waveform with a faster transition time than the source follower to have a simpler receiver. The first thing that was tried was to improve the rise time of the source follower circuit using an additional circuit in series with it. To implement such a circuit, we need a mechanism to detect rise time of the input signal and implement some control pulse to drive our helping circuit.

The proper circuit to make the required pulse, in the event of a transition in clock signal, is a monostable circuit. A monostable circuit is a circuit that generates a pulse with a specific width whenever the circuit is triggered by a rising edge. We call it monostable because it has only one stable state. A trigger event makes the circuit go to a quasi-stable state. It eventually returns to its original state after a delay determined by some mechanism in the circuit.

Figure 3.5 shows a transition-triggered one-shot monostable circuit. Signal at point B is a delayed inverted version of the input signal, the delay element can be implemented using an odd number of inverters. Figure 3.6 shows the input and output waveforms of this circuit.

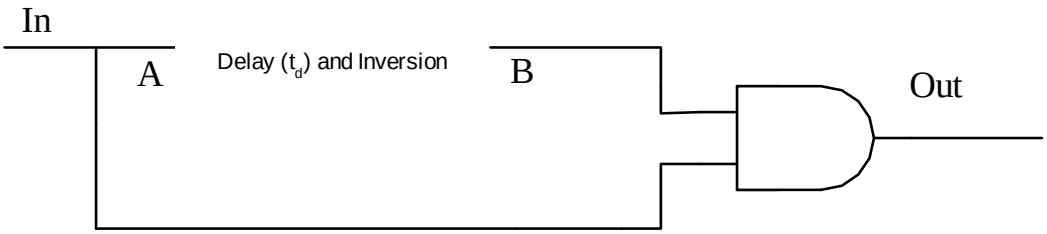

Figure 3.5: One-shot monostable circuit

The main waveform, and its delayed inverted copy are input to an AND gate so at the output we have high logic level whenever both inputs to the AND gate are high. We can detect the falling edge of the input signal using an OR gate with the same analogy. These signals can be used to turn on a circuit that can improve the transition time of the source follower circuit.

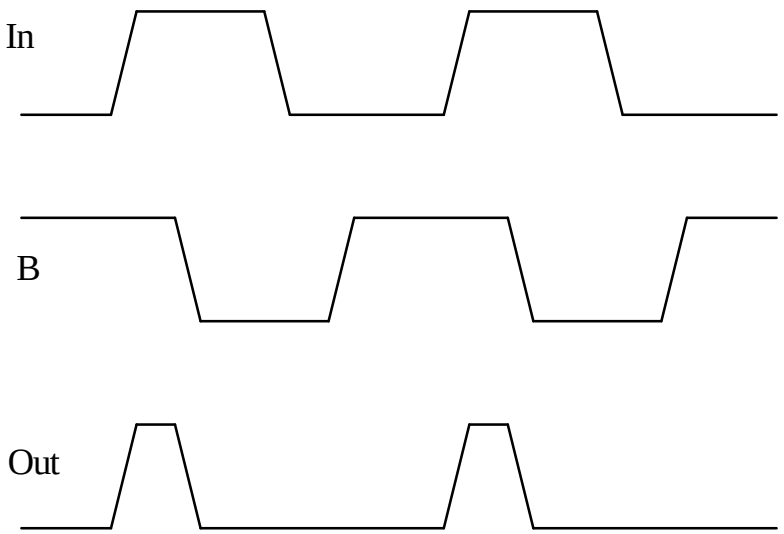

Figure 3.6: monostable circuit waveforms

3.2.1 Source-follower Buffer with Helping Circuitry Added

Figure 3.7 shows the block diagram of a potential circuit that can improve the rise time of the source-follower circuit. This is a general block diagram for both transmitter and repeater circuits. The first stage in the main path of the signal is a regenerator and its function is to recover the reduced-swing signal. Other parts of the system need a full-swing signal to be able to function properly so we need to recover the signal first. This part of the circuit is also necessary in a conventional source follower repeater.

As already mentioned, we need two monostable elements to generate control signals for a helping circuit. The output stage of the circuit is named *helper* since it helps the output rise time of the source-follower be improved. As can be seen from the block diagram the helping circuit consists of two arrays of pMOS and nMOS transistors such that their drains are connected directly to the output of the

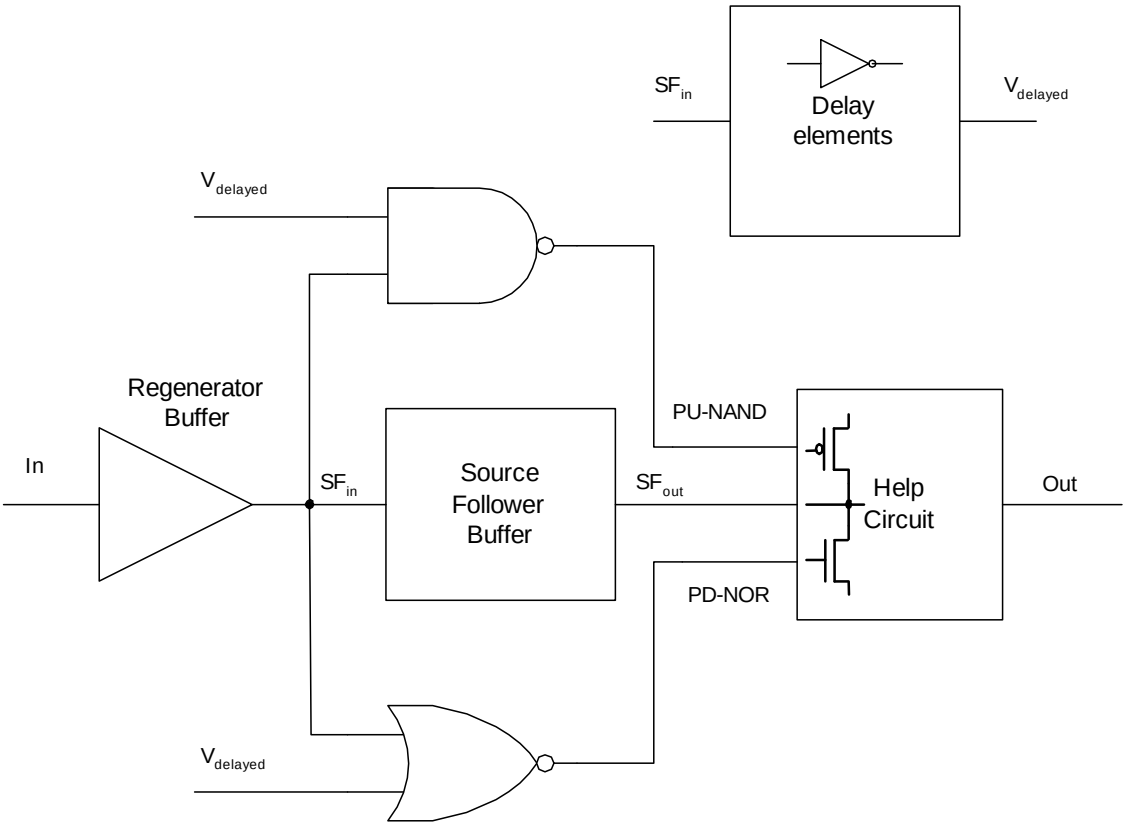

Figure 3.7: Block diagram of the proposed circuit for improving the poor transition time of the source-follower buffer.

source-follower. The pMOS transistor must be turned on during the transition of SF_{out} from low to high to accelerate the charging of the output load. The required control pulse must be "0" for this time period.

This pulse is generated using a NAND gate whose inputs are the input of the source-follower and its inverted delayed copy. The output of the NAND gate is zero only when both inputs are one, which happens after SF_{in} rises to one. The condition remains valid until $V_{delayed}$ falls down to zero. Therefore, after SF_{in} switches from zero to one, and for a period of time that can be adjusted with the delay time, we have a negative control pulse. Since pMOS needs voltage zero to be turned on, pMOS helper turns on during this period and accelerates the charging of the output. Figure 3.8 shows the different waveforms of the important points of the circuit. As is shown, PU_{NAND} is zero at the rising edge of SF_{in}. The transition time and maximum voltage that the output can reach can be adjusted using pMOS sizing and the width of the control pulse. It is obvious that if the control pulse lasts for a long time then the output reaches V_{DD}. Practically, this never happens as we want to reduce the voltage swing and considering that the typical propagation delay of the delay chain is in a range from 100 to 150 ps.

We need a pulse that is high during the falling edge of the output of the source-follower circuit to turn the nMOS transistor on. nMOS transistor turns on when its input is high. The two last waveforms in Figure 3.8 show a PD_{NOR} pulse that is used to turn the nMOS transistor on and SF_{out} or, in fact, the output of the circuit. As is clear, the output has sharper rise time than that of the source-follower buffer (Figure 3.4).

Figure 3.9 shows a circuit based on the above block diagram. The terminal names are the same as the block diagram in Figure 3.7. Three inverters are used to generate the required delay. The size of N_2 and P_2, the helper transistors, and

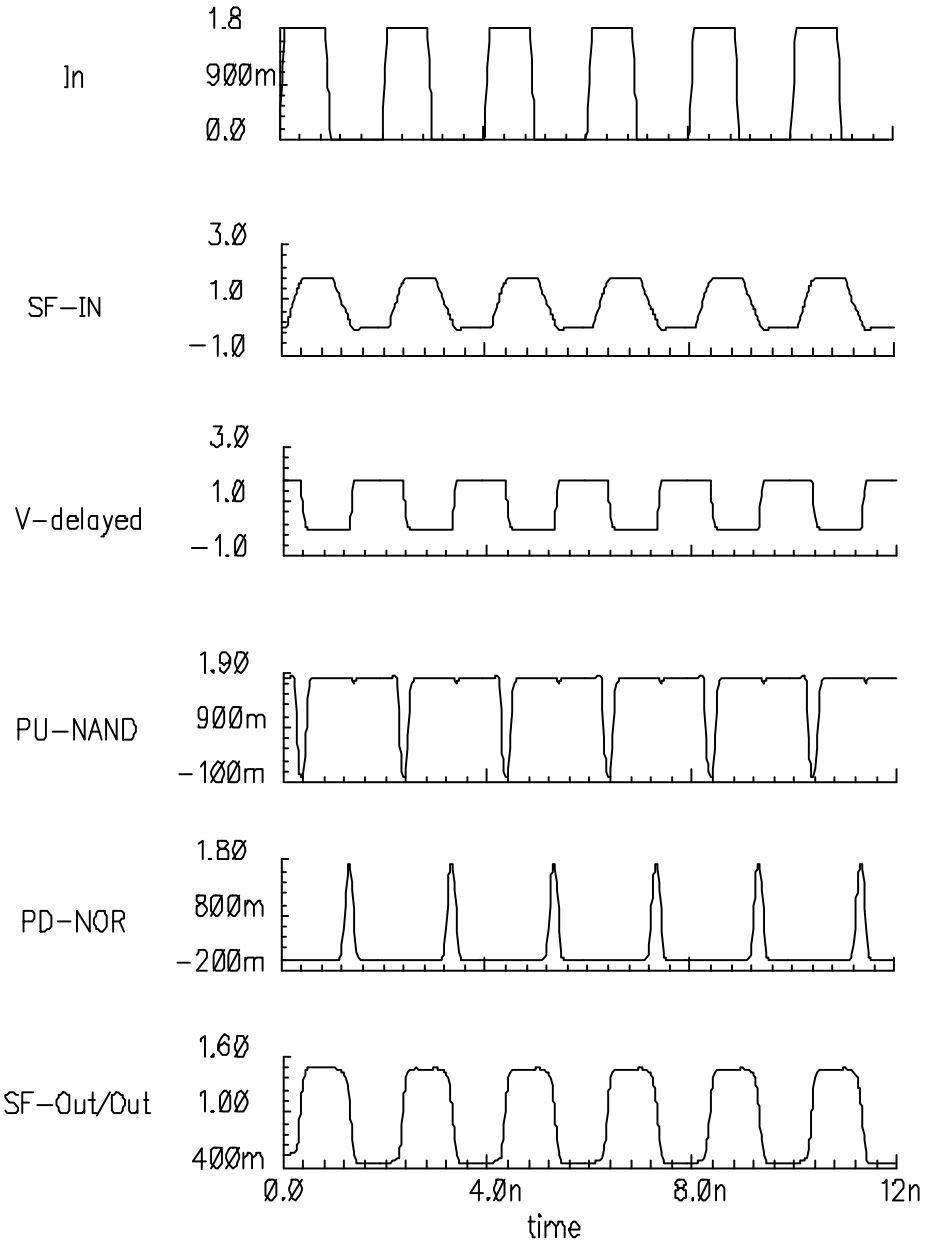

Figure 3.8: Waveforms of the helped source-follower circuit. Control pulses PU-NAND and PD-NOR are synchronous with SF-IN

the ratio $\frac{P_2}{N_2}$ must be adjusted properly to get the desired signal swing. As already mentioned, the buffer in the first stage, the regenerator buffer, converts the reduced swing signal back to full swing. As it drives the input of the source follower it is a small buffer. But since NAND and NOR gates, used to generate control signals, also load this buffer it must be capable of driving the input capacitance of these gates as well.

3.2.2 Omitting Source-follower

In the circuit in Figure 3.9, the only advantage of keeping the source-follower block is to protect the output from discharging due to the noise when helping transistors are off. But this block has very low performance and it is in the critical path of the signal. We will see later that a circuit with implicit NAND/NOR gates is independent of the quality of the input waveform and only needs a correct timing.

Long interconnect wires and clock distribution wires in particular are laid all over the chip, and, unlike applications like bit-lines in memory arrays the effects of noise can not be controlled. At first the source-follower circuit was used to make sure outside the time periods that neither of the helping transistor are on the output is not floating. During the rest of the time the output is not driven by the helping block and the output would be floating if the source follower was not there.

In order for the source-follower block in the proposed circuit to keep the output level constant it should be as large as the one in a stand-alone source-follower in which there is no helping circuit. This fact brings doubts about the use of this block in the proposed circuitry.

Most of the noise problems related to discharging nodes increase when the node capacitances are not large enough. With technology scaling, device parasitic capac-

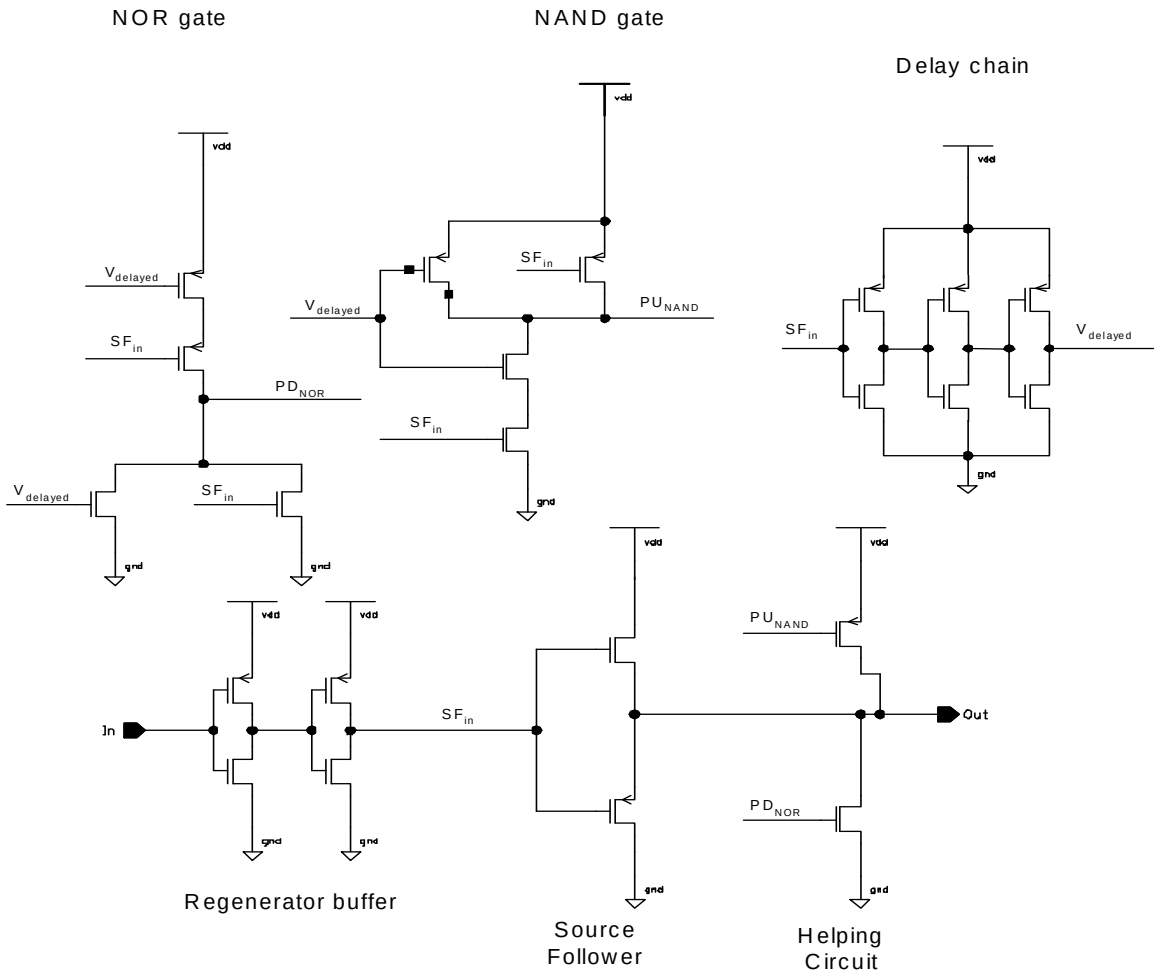

Figure 3.9: Circuit implementation of the helped lever convertor. It consists of helping, delay and pulse generation circuitries. Control pulses are generated using the input of source-follower buffer and its inverted/delayed version using the delay chain.

itances become smaller. This is especially a concern with regards to noise problems in dynamic circuits and dynamic memories. In these circuit categories, capacitances are used for storing previous states and memory contents.

However, the application of the proposed circuitry is in driving large capacitances of interconnect wires. Here with technology scaling, global interconnects are getting longer and parasitics are getting larger. These large distributed capacitances help the output of the circuit remain considerably unaffected by the noise sources. Therefore, there is no need to use another circuit to do this and we can leave the source-follower block behind.

3.2.3 A Circuit with Implicit NAND and NOR Gates

The circuit described improves the rise time of the source-follower buffer, but increases the number of transistors. Logic gates used for control pulse generation add about eight transistors to the circuit. This overhead adds to the area and at some point for smaller loads to the power dissipation and makes the solution inefficient. We have already concluded that the source-follower is not necessary anymore and so we can reduce the critical path delay significantly. In an attempt to reduce the transistor overhead without deteriorating the performance a new circuit is introduced here. This circuit also eliminates one of the inverters in the critical signal path.

Figure 3.10 shows the modified level converter circuit. Transistor pairs P_2/P_3 and N_2/N_3 function as output drivers and the output is driven only when both transistors are on. In fact, they function as two intrinsic logic gates in addition to driving the interconnect line. Transistors P_3 and N_3 receive the clock after a single inversion. At the same time, transistors P_2 and N_2 receive a delayed and inverted

copy of the signal at point A. Following is a description of the operation of these transistors.

Figure 3.11 shows the timing diagram of the signals at different points of the circuit. Before t_1, the input of P_3 is low, so this transistor is on. But, since the input of P_2 is still high, this transistor is off and the output is floating. At t_1, the input of P_2 changes from high to low and this turns the transistor P_2 on. Since transistors P_2 and P_3 are connected in series, the load starts getting charged toward V_{DD} at this time. At t_2, after a delay of t_d , V_B, input of P_3, changes from low to high turning this transistor off and again the output is floating. Since this time period is short the output does not fully reach V_{DD} and the maximum voltage level is a function of the driving transistor sizes and the delay t_d. During this period the input voltage of transistors N_2 and N_3 is low so both transistors are off and there is no direct path from V_{DD} to *ground*.

Transistors N_2 and N_3 operate the same way from time t_3 to t_4. During this time period, which starts from the falling edge of the voltage at point A, output is pulled down to a minimum voltage level larger than zero. By carefully controlling the transistor sizes in the delay chain the desired delay can be achieved which in turn helps in realizing an appropriate output swing.

3.2.4 Taking Advantage of Low V_{th} and High V_{th} Transistors available in 0.13 Process Technology

Using high V_{th} transistors in the regenerator buffer to recover the low-swing waveform in the proposed circuit while using low V_{th} transistors to reduce the output voltage swing helps reduce the total power consumption. This also helps reduce the static power. In fact, using low threshold voltage transistors, to reduce the output

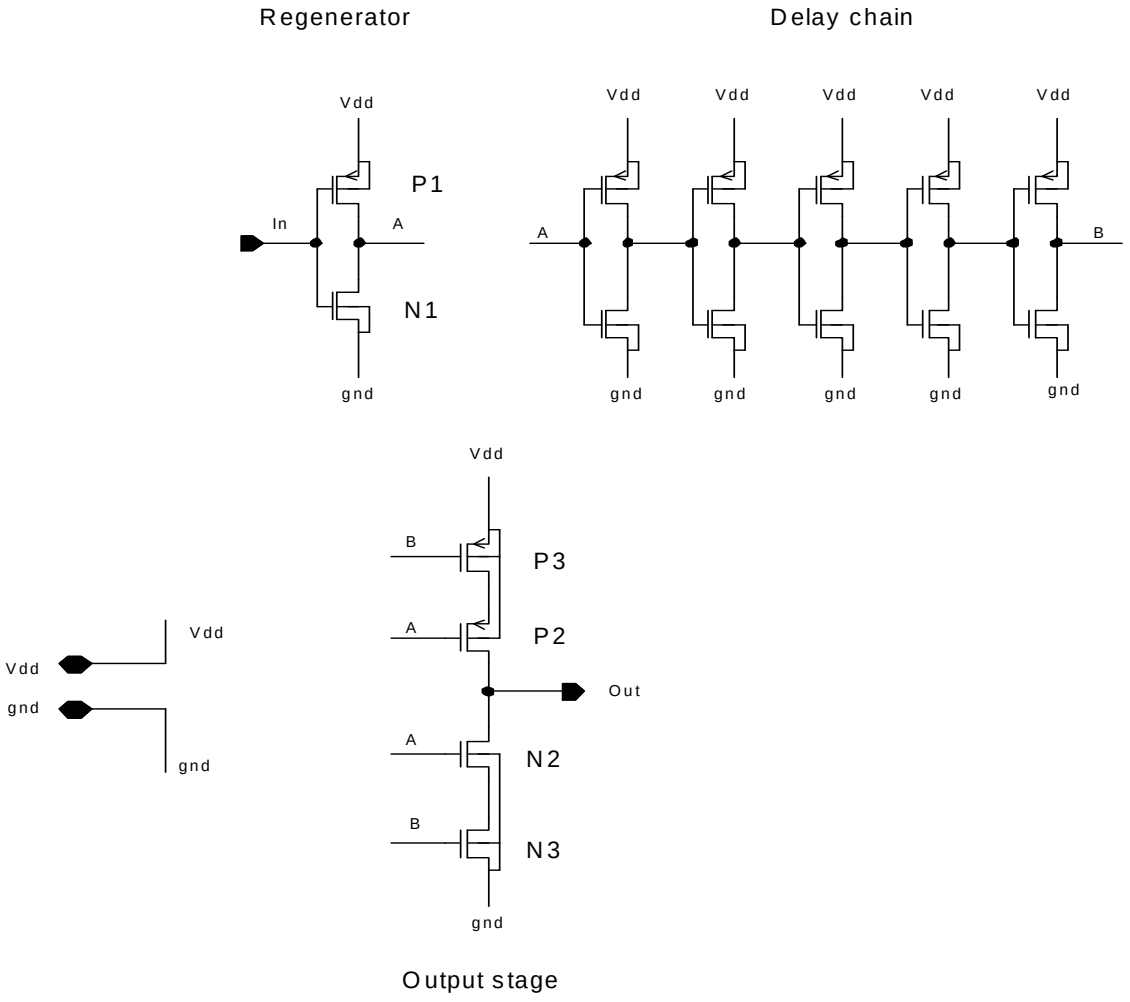

Figure 3.10: Level converter circuit with implicit NAND and NOR gates. Transistor pairs (P2, P3) and (N2, N3) function as elements of tow implicit logic gates and output driver at the same time.

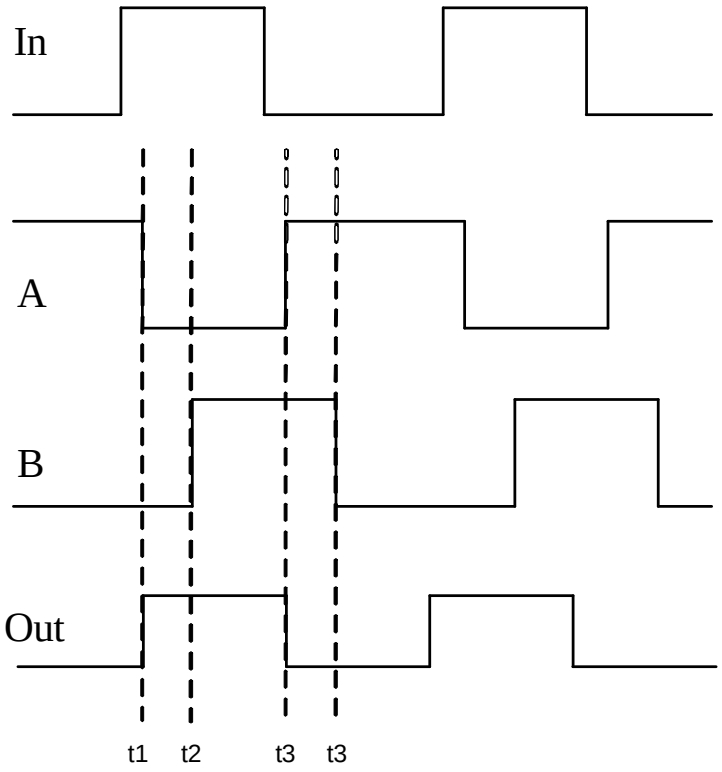

Figure 3.11: Timing diagram of the low-swing circuit with implicit control gates

voltage swing, causes the input stage transistors of the next repeater to turn on and off sooner and the period that both transistors are on reduces substantially. Unfortunately there are no low V_{th} transistors available in 0.18 μm process technology, although, results from simulations in 0.18 technology show there is still substantial power saving.

In the next chapter we use a few testbenches to compare the proposed level converter presented in this chapter with a conventional buffer designed for this purpose. The conventional buffer is discussed in next chapter.

Chapter 4

Measurements and Observations

In this chapter the efficiency of the proposed circuit is compared with a conventional buffer using different testbenches and platforms. First we will look at the circuitry of the conventional buffer.

In all cases, cells are sized so that the rise and fall times in all circuit segments are about 100 ps. In practical situations this is not necessary. In cell library design, especially, usually there are generic cells including generic buffers used for driving different but very close loads. The generic cells function with any capacitive load at their inputs and outputs. Therefore, there will be a range for parameters like propagation delay and transit time, and this fact is considered in different design phases.

A H-tree is designed to have a more realistic comparison. The design of the clock tree includes extracting parasitic elements, metal layer selection, choosing clock topology, and placement of different buffers. This clock tree is used to examine the impact of different parameters on the circuit performance.

4.1 Conventional Buffer

To measure the performance of the proposed circuit it is compared with a conventional buffer. This buffer is used to implement a full voltage swing H-tree clocking network that is compared to the reduced voltage swing H-tree clocking network.

This buffer consists of two series inverters that increase gradually in size to drive the output load. Cascaded drivers are useful in driving large capacitive loads. For our case two stages are enough to drive the typical interconnect loads. However, for driving off chip loads more stages are required. The second inverter is large enough to drive the load and the first inverter only drives the second inverter's input capacitance. Using an even number of inverters results in the effect of variation of process corner on the rise/fall time being minimized. If due to process variations (especially, SF and FS) one of the transient times becomes sharper than the other one, the second inverter compensates the effect. In theory the optimum size ratio between the two stages is $e = 2.7182$, which in the designed circuit was set to 3. Figure 4.1 shows the schematic diagram of the conventional buffer.

4.2 Evaluation of Single Cell

The test bench in Figure 4.2 was used to measure the different parameters of the proposed circuit and to compare it with the conventional buffer. The operating frequency was 500 MHz. CMOS 0.13 μm process technology parameters were used for simulations. HSPICE version 98.2 was used as the simulator.

Initially it was necessary to measure the efficiency of the level converter cell for different loads. This observation was necessary since it is expected to have different results for different parasitic capacitances that are to be driven. Typical capacitive

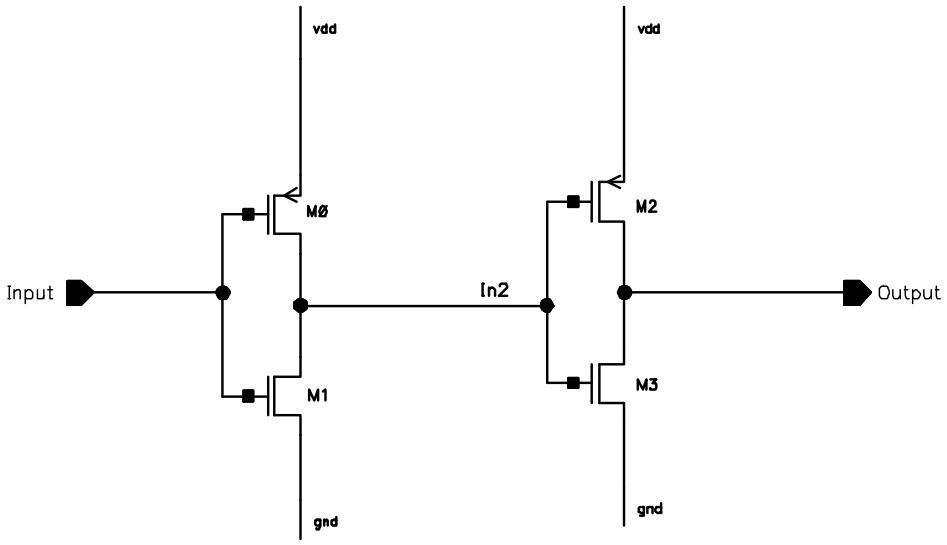

Figure 4.1: Conventional buffer that is used to evaluate the performance and other paremeters of the propised circuit

loads are already in the range of a few hundred femto Farads in the submicron technologies. A range of loads from 20-600 fF was chosen for the simulations.

Figure 4.3 shows the PDP (Power Delay Product) and power efficiency of the proposed circuit over the conventional buffer. As can be seen for loads greater than 300 fF, PDP and power are less by at least 18% and 60%, respectively. The reason that the efficiency of PDP is smaller, is that the delay of the proposed buffer is 30% longer than the conventional buffer on average. Considering the fact that capacitive interconnect wire loads less that 400 fF are not buffered, the proposed circuit looks quite efficient in saving power and even PDP in our application.

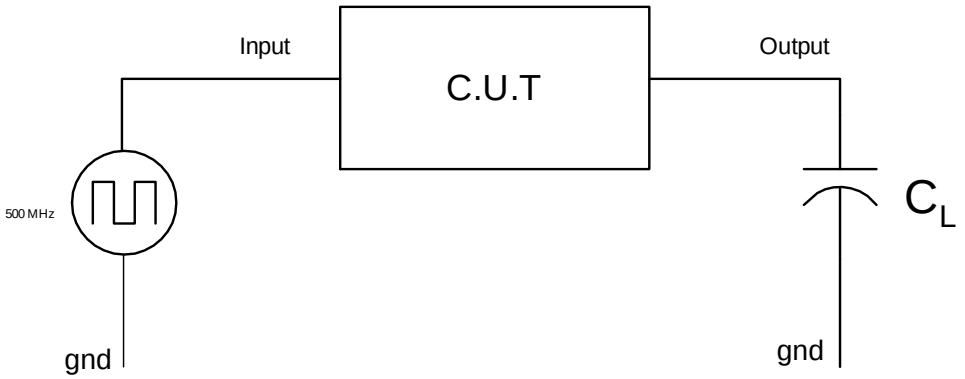

Figure 4.2: Test bench for evaluating proposed circuit performance

4.3 Clock Distribution Network Simulation Results

It can be predicted that the results from single cell simulations are not realistic and could be a little optimistic. It seems necessary to test the proposed circuit in a more realistic testbench; a clock distribution network.

As we move into submicron technologies and high clock frequencies, it has become much more difficult to model, analyze, and predict the interconnections. It is hard to find a model that matches a real design. In addition, layout and placement of gates and interconnections are not known until the final stages of the chip design. We have tried to design a model that best satisfies our purpose.

In following section a H-tree clock distribution network is introduced which, is used to model the global clock distribution network.

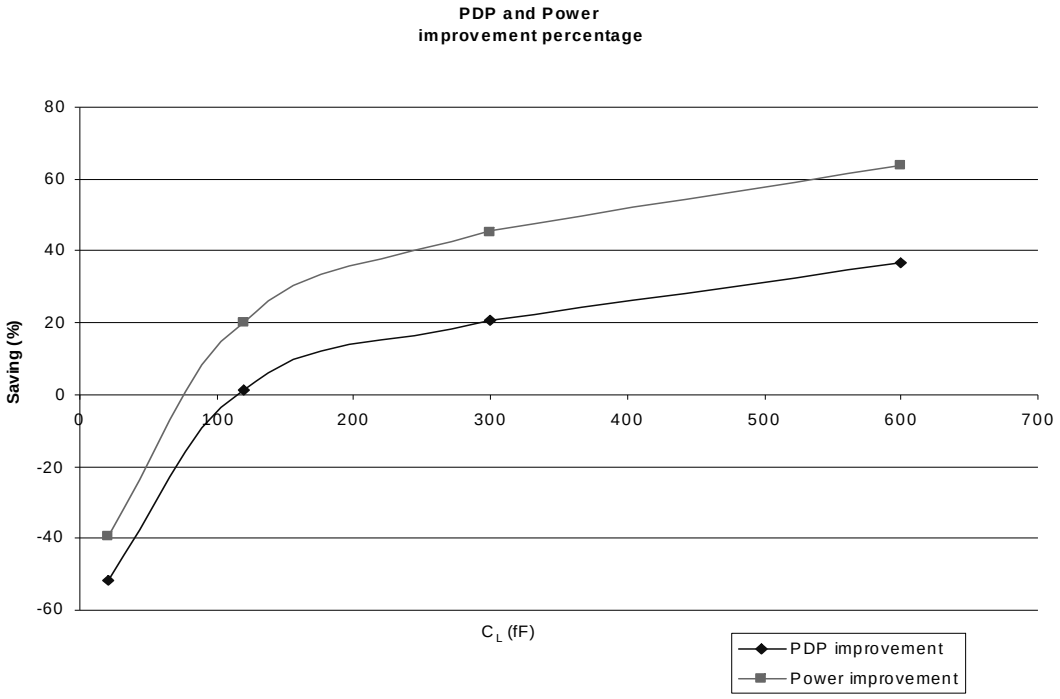

Figure 4.3: Efficiency of the proposed circuit over conventional buffer.Plots show the percentage of improvement in PDP and Energy

4.3.1 H-tree as a Testbench

The most common approach for distributing clock signals in a complex VLSI circuit is to insert buffers at the clock source and along the clock path forming a tree structure. The clock source is the root of the tree, the initial portion of the tree is the trunk, the individual paths driving each register are the branches and the driven registers are the leaves of the clock tree. These buffers restore the degraded clock signal and isolate the local clock from the upstream load. Also, clock trees can be designed to minimize the clock skew. The number of the buffer stages between the clock source and the final clocked registers depends on several factors, such as:

(i) the total load capacitance comprising the combined interconnect and buffer input capacitance,

(ii) the clock skew tolerance of the system,

(iii) the latency requirements of the clock network, and

(iv) the nature (e.g., asymmetrical/symmetrical) of clock distribution network.

H-tree is a popular configuration for implementing a clock distribution network. In a properly designed H-tree clock network, the clock skew can be minimized to a fraction of the clock period. In the H-tree clock network, the primary clock driver is connected to the main "H".

In the reduced swing H-tree clock distribution network the global clock is reduced at the clock source by the proposed RS-RS buffer. The same circuit is used to buffer the reduced swing over the interconnect lines. The reduced voltage swing is converted to full voltage swing before it reaches the local clock distribution network.

Clock Structure

As a testbench a clock network was designed for a hypothetical chip with an area of $1.5 \times 1.5\ cm^2$(Figure 4.5). With one main "H" and two sub-H's we can get a sub-block with an area of $3.516\ mm^2$. A load capacitance of 25 pF based on the typical data for existing VLSI chips was used. We assume wire segments one and two belong to global wires and wire segments three to six are part of semi-global wires. Based on this assumption, metal layer eight was selected for wire segments one and two and metal layer six for the remaining wires.

The interconnect delay can be lowered using wider wires. But, increasing wire widths does not decrease wire delay or the RC time constant beyond some limits [2]. Tapering wires in clock trees reduces the overall Elmore delay [12]. Following the structure of the H-tree that satisfies the tapering strategy [13] and based on layout design rules, wires with a width of 16 μm were used for the first wire segments and halving the width of each stage as we go downstream. The width of the last stage wire was calculated as 0.5 μm. The information for the width and length of each stage is given in Table 4.1. This information is used for calculating parasitic elements of the wires.

Calculating Parasitic Resistance and Capacitance

The most common method to calculate parasitic elements of wires is to use tables presented in the corresponding technology information collected as Process Interconnect Capacitance Model. The information in these tables are very precise and measured for certain wire models. The following model was used for calculating wire parameters for metal layers six and eight. The model is shown in Figure 4.4. The total capacitances is:

Table 4.1: H-tree dimensions and calculated parasitics

Segment #	**Length** (μ**m**)	**Width** (μ**m**)	C_a (fF)	C_f (fF)	C_{total} (fF)	$R_\square$ $m\Omega/sq$	R_{total} Ω
1	3750	16	401.68	83.66	569.00	26	18.281
2	3750	8	200.83	83.66	368.17	26	36.562
3	1875	4	92.258	47.01	186.46	78	36.562
4	1875	2	46.129	47.01	140.326	78	73.125
5	937.5	1	11.532	23.55	58.631	78	73.125
6	937.5	0.5	5.7661	23.55	25.865	78	146.25

$$C_{total} = C_a \times W \times L + 2 \times C_f \times L + 2 \times C_c \times L \tag{4.1}$$

Where:

W : Conductor width (unit: μm)

L : Conductor length (unit: μm)

S : Conductor spacing (unit: μm)

C_a : Parallel plate area capacitance (unit: $fF/\mu m^2$)

C_f : Fringing capacitance to bottom plate (unit: $fF/\mu m/side$)

C_c : Coupling capacitance to adjacent conductor (unit: $fF/\mu m/side$)

The information in Table 4.1 was calculated from the values for the substrate as the bottom plate. In this model the coupling capacitances were neglected. The reason for this is that we do not have any idea about the adjacent wires passing by the clock wires. The value of C_c is very difficult to be computed accurately since

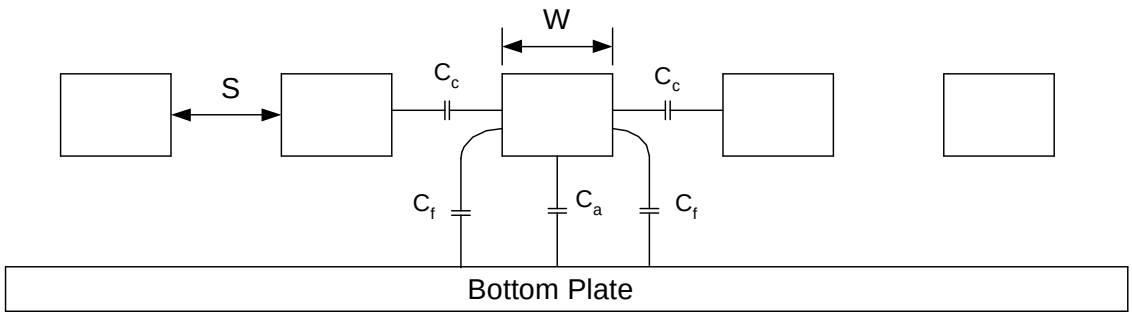

Figure 4.4: Wire model chose for calculating parasitic capacitance

we need to consider both the spacial locations of all neighboring wires in the three dimensional structure and the temporal relations between the signals on these wires. Such information needs a complete knowledge of the structure of the designed chip, which was not available. This approximation does not effect our results because the performance of the proposed level converter improves as its load increases.

The values for the parasitic resistors was calculated from Equation 4.2. $R_\square$ is the sheet resistance of the metal layer. The calculated values for R, the resistance of different wire segments, are shown in Table 4.1 on page 47 as well.

$$R = R_\square \frac{L}{W} \tag{4.2}$$

Buffer Placement

As can be seen from Table 4.1, wire segments close to end nodes have small parasitic capacitances. So, as we get closer to the end nodes there is no need to use repeaters. Since we want to use the same transistor sizes in both repeaters, they must see equal parasitic capacitances. Following this algorithm the buffers are placed at the nodes shown in Figure 4.5.

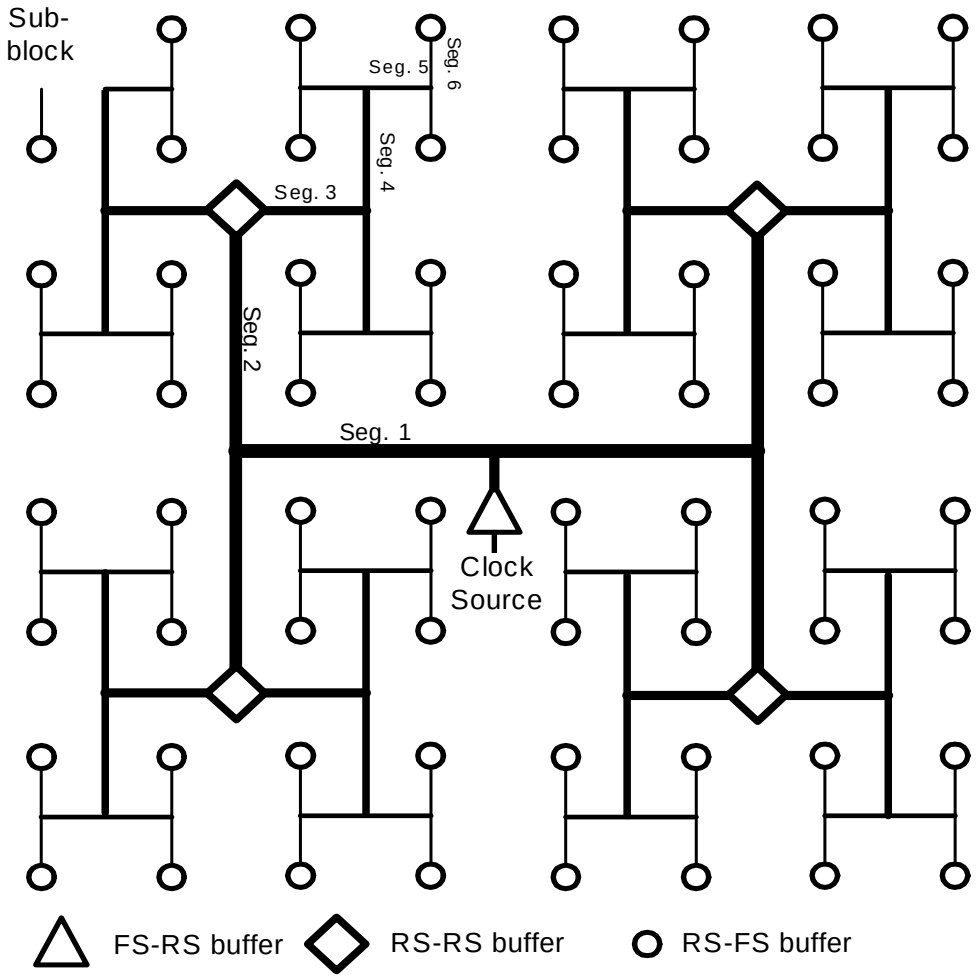

Figure 4.5: H-tree clock distribution network with three level of hierarchy. Different wire segments, and buffers in a typical signal path from clock generator to an end node that covers a sub-block is shown.

4.3.2 Pre-layout Simulation Results

The simulations were performed in a dual threshold 0.13 μm CMOS technology. The frequency of simulation is 500 MHz with 50% duty cycle and a rise time of 100 ps. The value of distributed interconnect impedance was calculated from the technology data sheets. As already mentioned, only the area and fringe capacitances for the clock network was considered . The mutual coupling capacitance was ignored. Although this component can be significantly large, it is very difficult to model such couplings. Besides, for this comparative study, our conclusions are not changed if we ignore the mutual coupling capacitance. Furthermore, the clock can be routed so as to reduce/minimize this component. The distributed impedance of the interconnecting wires was modeled using a π model. Metal layers eight and six were used for interconnect wires.

The topology and buffer placement of the H-tree is shown in Figure 4.5. With the proposed buffer placement structure, the loading capacitance of the first and second stage buffers is approximately 2500 fF. The third buffer drives a local clock distribuiton network of a sub-block.

Power and PDP saving

Figure 4.6 plots the power and energy savings of the low voltage swing clock distribution network with respect to the conventional full voltage swing clock distribution network, that uses conventional buffers, as a function of the end (leaf) node load. The simulations were done for an end node load from 40 fF to 120 fF.

The simulation results show considerable power and energy saving for the proposed circuit compared to the network with conventional buffer. The power (energy) efficiency varies from 32% (22%) to 19% (10%) for different load conditions.

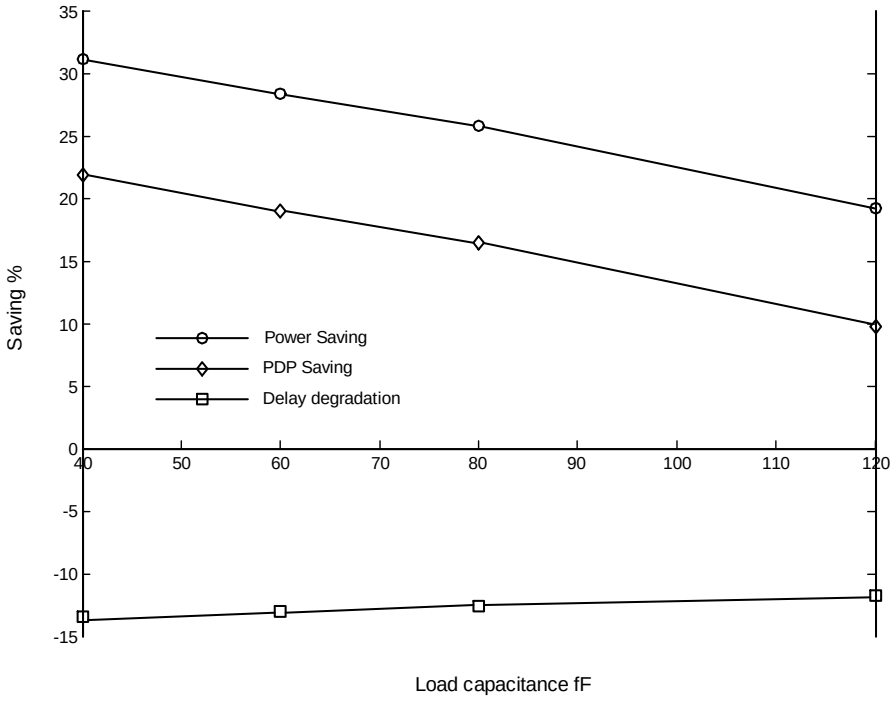

Figure 4.6: Proposed structure's relative efficiency as a function of output load capacitance of H-tree

For larger loads the efficiency of the clock network with proposed buffer declines. This is attributed to the fact that the direct path current of the reduced swing buffer is larger than that of the conventional buffer. To drive larger load capacitances we need larger buffers at the final stage. In the proposed circuit the direct path current is more of a concern as larger transistors draw more current during the short period that both transistors in the inverter are on. This explains why we have less saving for larger end node load. This has been already minimized by using dual V_{th} transistors. Using low V_{th} transistors for reducing the clock swing and reconstructing to full swing using high V_{th} transistors in the first inverter reduces the direct path current substantially. But, as it can be seen, it still impacts the

circuit operation.

Sensitivity to Process Corner Variations

Two H-tree testbenches were also simulated for different process corners to estimate the influence of process corner variations on the robustness of the circuit. All the parameters were stable for different corners except the pulse-width of the waveform on the wires and at the output

Figure 4.7 and Figure 4.8 show the output signal pulse-widths of the conventional and proposed H-tree networks for different process corners and temperature, respectively.

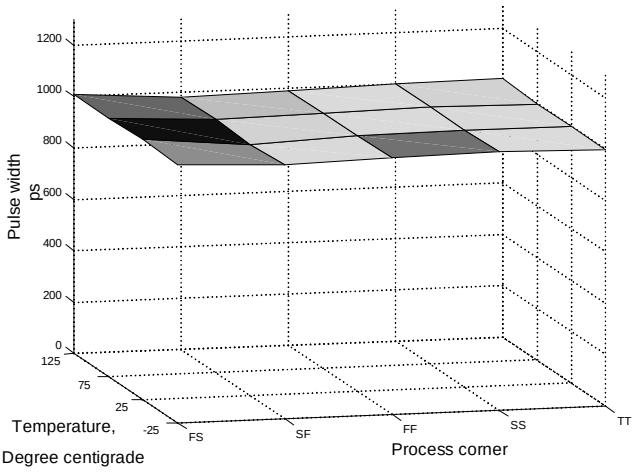

Figure 4.7: The output pulse width of the conventional H-tree as a function of process and temperature

In general, the proposed H-tree exhibits increased sensitivity with process variation. For process corners FS and SF there is about 70ps decrease and 89ps increase

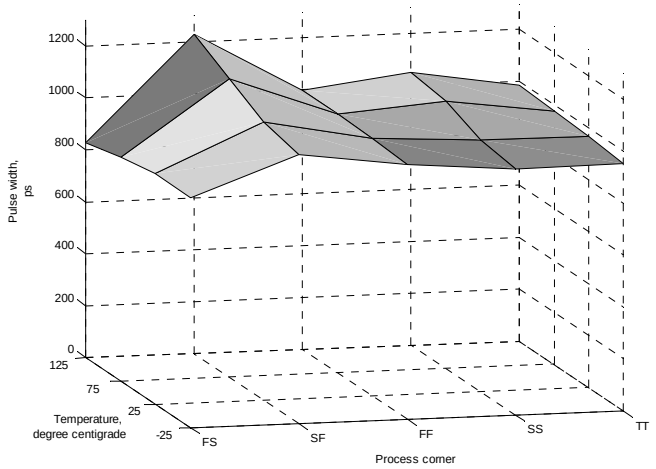

Figure 4.8: The output pulse width of the proposed H-tree as a function of process and temperature

in pulse-width at room temperature, respectively. This deviation is always a fraction of the clock period and this fact makes it tolerable for timing requirements as it decreases with clock period and is not a constant value.

4.3.3 Post-layout Simulation Results

For post-layout simulation it is necessary to have a model for the local clock distribution network driven by the end nodes. As already mentioned, based on the data [14] from typical chips, a load capacitance of 25 pF for an area of 3.516 mm^2 can be estimated. This capacitance can typically be from 15 to 35 pF. The local clock distribution network model for the sub-block is shown in Figure 4.9. The sub-block is divided into ten segments. Each segment, with 2.5 pF capacitance, is driven by a two-stage buffer. The input capacitance of each such buffer is 120 fF. The total

input capacitance of all buffers, 1.2 pF, is buffered by another two-stage buffer. the input capacitance of this buffer is about 50 fF.

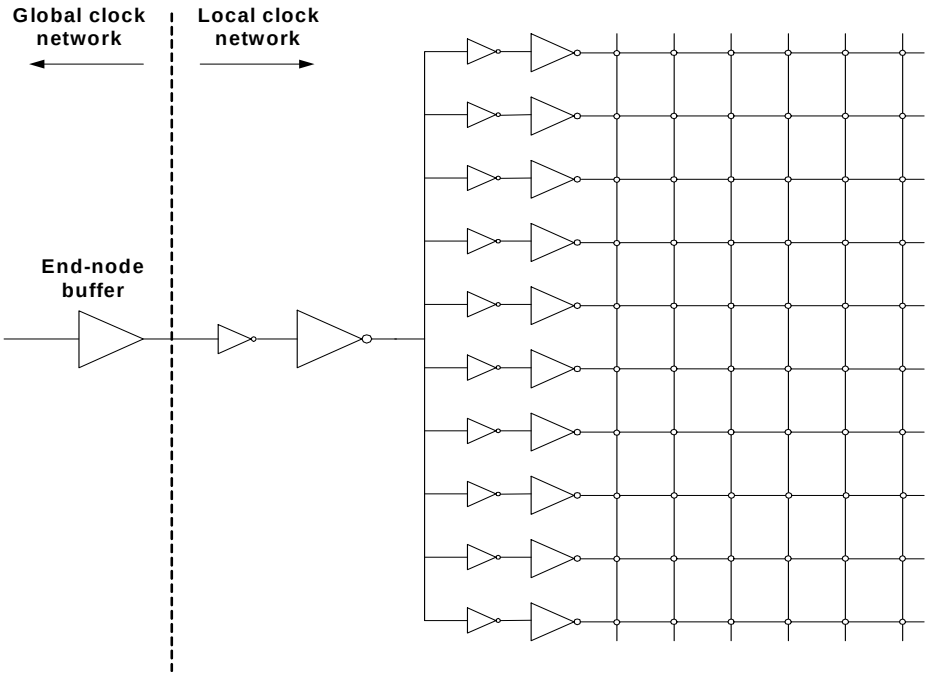

Figure 4.9: Sub-block clock distribution scheme

Power Efficiency vs. Changes in Temperature and Voltage Supply

A power saving of 27% was achieved from post-layout simulations, which is close to what was expected. To study the behaviour of the circuit, with changes in power supply, a parametric post-layout simulation was performed. Figure 4.10 depicts the normalized power, t_{d_r} and t_{d_f}, for changes in V_{DD} from 1.08 to 1.32. Changes are within $\pm 10\%$ variations in V_{DD} and this results in at most 8 ps variation in the delay of the main path which, is about 400 ps. This shows that the proposed circuit is very stable against the variations in V_{DD} and can tolerate typical power

supply noises.

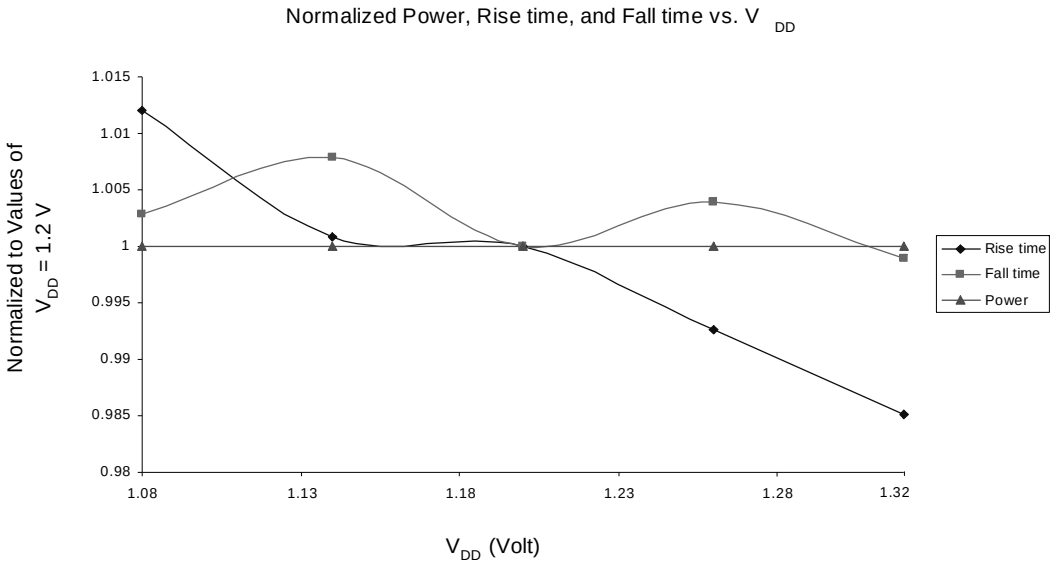

Figure 4.10: Normalized Power, Rise time, and Fall time versus 10% changes in power supply voltage

Simulation results for different working temperatures are shown in Figures 4.11, 4.12, and 4.13. Figure 4.11 depicts power, rise time delay, and fall time delay normalized with respect to the corresponding values at $25°C$ for conventional clock tree. Figure 4.12 shows the same results for the proposed clock network. For power the range of variations is a little smaller for the conventional than that of the proposed clock network (about $[-2\%, +4\%]$ and $[-4\%, +11\%]$, respectively) but it is not substantial. Instead, the range of changes in propagation delays is less in the proposed H-tree. For the conventional clock distribution network the range is almost twice as much.

Figure 4.13 plots the efficiency of the proposed structure over the conventional one. From Figure 4.13 it can be concluded that power efficiency is higher for lower

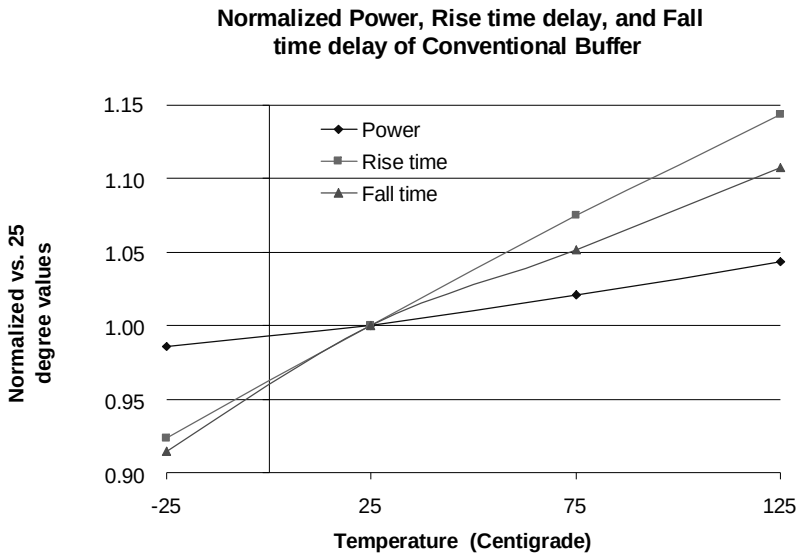

Figure 4.11: Change of power and rise time delays vs. temperature in the conventional network (The values are normalized to the value at $25°C$).

temperatures. For temperatures from $-25°C$ to $125°C$ the efficiency decreases from 30% to 18% with a value of 26% at room temperature. There is only a 3% drop in efficiency percentage for temperatures as high as $75°C$ though.

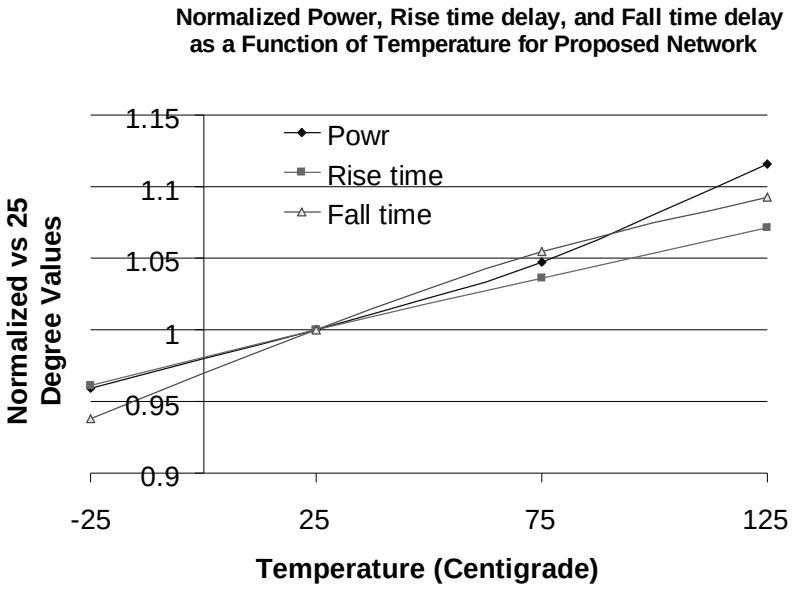

Figure 4.12: Change of power and rise time delays vs. temperature in the proposed network (The values are normalized to the value at $25°C$).

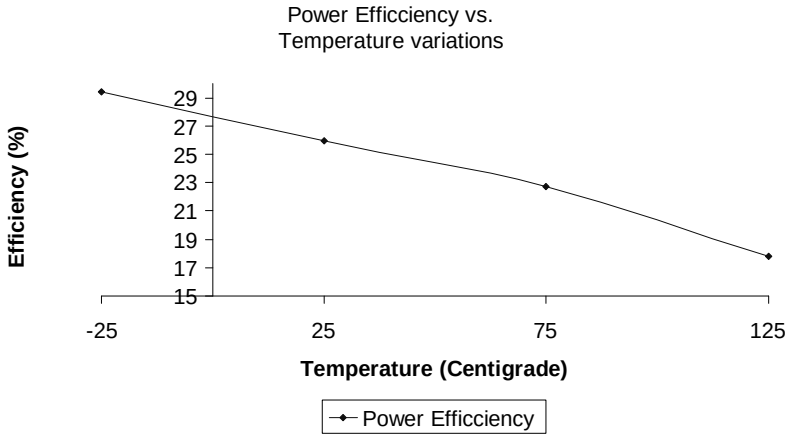

Figure 4.13: Variation of power efficiency for different ambient temperatures in the conventional network.

Chapter 5

Conclusion

5.1 Summary

The goal of this research was to study the potential of reduced-swing clock signal distribution scheme in reducing the power consumption of the clock distribution network. The clock network, based on the wire lengths and clock fan-out, can be divided into: (i) a global clock distribution network (ii) a local clock distribution network. The local clock distribution network is more susceptible to noise and has larger fan-out. In addition, we did not want to change the flip-flops to make them operate with low-swing clock inputs. Therefore, the swing was only reduced in the global clock distribution network. The global clock swing reduced at the source by a novel "full-swing to reduced-swing" buffer. The same circuit is used to buffer the reduced-swing over the interconnect lines. The reduced voltage swing is amplified to the full-swing by a simple buffer before it reaches the local clock distribution network.

The global clock network delivers the clock signal to the local clock distribution

network that was described in section 4.3.3. The typical capacitance of each such sub-block can typically be between 15 to 35 pF. This load is equivalent to the input capacitance of between 4,000 to 9,000 flip-flops. However, several stages (3-5) of local clock buffers can be employed to drive the load. Therefore, loading of the global clock distribution network is not an issue.

As a level converter circuitry, the proposed buffer seems to have substantial power saving over conventional circuits. None of the circuits presented before have transient parameters even close to this one. In other applications it is necessary to use structures similar to schmit-trigger in the receiver side to improve poor transient times. This adds to the transistor count and complexity in the receiver side.

The only disadvantage of the proposed buffer is that for some process corners the pulse width changes significantly. Unfortunately, despite all efforts no solution was found for this problem as it is something inherent to the structure of the circuit. However, if the stable rising edge of the signal is used as a reference, this problem only has an impact on timing of the circuit that if is considered, in calculating the minimum chip clock period, poses no serious problem to the operation of the system. In a system that uses this circuit there is going to be a trade-off between power and performance.

The proposed circuit's power dissipation and dynamic parameters are independent of the power supply variations as much as the conventional circuit is. For variations in ambient temperature there are some changes in parameters for the proposed circuit but the efficiency is still good enough. Even for low temperatures the power efficiency increases for the proposed level converter and the high temperatures are the only concern in this regard.

Chapter 6

Future Work

Since clock distribution is a concern in large die sizes it will be more precise to study the circuit in such environment. Unfortunately, the die size of the chips that could be submitted to CMC can not be that large. Besides, designing a complete system is not possible as the cell library of the 0.13 technology is not available yet.

Future work includes designing a test chip that can serve our goal in evaluating the proposed circuit and at the same time can be small enough in size to be fabricated.

Bibliography

[1] International Technology Roadmap for Semiconductors, *Semiconductor Industries*, 2001 edition.

[2] H. B. Bakoglu, "Circuits, Interconnections, and Packaging for VLSI", *Addison-Wesley Publishing Company*, 1995.

[3] H. Zhang, V. George, J.M. Rabaey, "Low-Swing On-Chip Signaling Techniques: Effectiveness and Robustness", *IEEE Transactions on Very Large Scale Integrated (VLSI) Systems*, vol. 8, pp. 264-272, June 2000.

[4] C. P. Yuan and T. N. Trick, "A simple formula for the estimation of the capacitance of two-dimensional interconnects in VLSI circuits", *IEEE Transactions on Electron Device Letters*, vol. EDL-3, pp. 391-393, Dec. 1982.

[5] L. W. Shaper and D. I. Ameym, "Improved electrical performance requierd for future MOS packaging," *IEEE Transactions on Components, Hybrids, and Manufacturing Technology,* vol. CHMT-6, pp. 282-289, Sep. 1983.

[6] D. Duarte, V. Narayanan, M.J. Irwin, "Impact of technology scaling in the clock system power", *VLSI on Annual Symposium, IEEE Computer Society ISVLSI 2002* , pp. 52 -57, 2002.

[7] J. Pangjun, S. Sapatnekar, "Low-power clock distribution using multiple voltages and reduced swings", *IEEE on Very Large Scale Integrated (VLSI) Systems*, Vol. 10, No. 3, June 2002

[8] R. Golshan and B. Haroun, "A novel reduced swing CMOS BUS interface circuit for high speed low power VLSI systems", *Proc. IEEE Symp. Of Circuits and Systems*, vol. 4, pp.351-354, May 1994.

[9] Y. Moisiadis, I. Bouras, A. Arapoyanni, "High performance level restoration circuits for low-power reduced-swing interconnect schemes", *The 7th IEEE International Conference on Electronics, Circuits and Systems*, vol.1, 2000, pp. 619 -622.

[10] Y. Nakagome, K. Itoh, M. Isoda, K. Takeuchi, M. Aoki, "Sub-1-V swing internal bus architecture for future low power ULSIs", *IEEE J. Solid-State Circuits*, vol. 28, pp. 414-419, April 1993.

[11] A. Rjoub and O. Koufopavlou, "Efficient drivers, receivers and repeaters for low power CMOS bus architectures", *Proceedings of ICECS'99 6th IEEE International Circuits and Systems*, pp. 789-794, 1999.

[12] J. P. Fishburn and C. A. Schevon, "Shaping a distributed- RC line to minimize elmore delay", *IEEE Transactions on CAS-I 42*, December, 1995 pp. 1020-1022.

[13] W. Wolf, "Modern VLSI design: system-on-chip design", *Printice Hall PTR*, 2002.

[14] P. J. Restle, T. G. McNamara, D. A. Webber, P. J. Camporese, K. F. Eng, K. A. Jenkins, D. H. Allen, M. J. Rohn, M. P. Quaranta, D. W. Boerstler, C. J.

Alpert, C. A. Carter, R. N. Bailey, J. G. Petrovick, B. L. Krauter, and B. D. McCredie, "A clock distribution network for microprocessors ", *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 792 -799, May 2001.