# On-Chip Global Interconnect Optimization

*by*

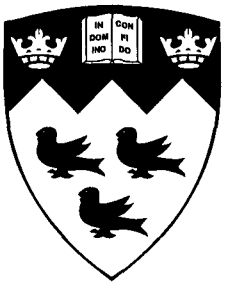
*Mourad Oulmane,*

*B.Eng. (Summa Cum Laude) 1995*

*M.Sc. in Physics 1997*

Department of Electrical Engineering

McGill University, Montréal




November 2001

---

**A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering**

© Mourad Oulmane, 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-79088-6

Canada

# Abstract

We present a practical yet accurate approach for dealing with the problem of inserting repeaters along on-chip interconnect lines to meet delay and transition time requirements. This approach is based on the fact that the transition time and the delay at the far end of an interconnect segment are, respectively, independent and linearly dependent on the driving repeater's input transition time, as long as the ratio of the output to input transition time does not exceed a pre-defined value. In this context, we first derive simple closed form expressions for the optimal repeater spacing and sizing. Then, we propose a bottom-up "pseudo" hierarchical quadratic programming method for inserting and sizing repeaters in RC interconnects. This method, unlike Van Ginneken's [30], although largely based on it, is able to account for transition times at every potential repeater insertion point along the RC line of interest. The resulting technique can be readily incorporated in a more general RC network optimization scheme (through repeater insertion) where, eventually, wire sizing can be formulated either as an objective or a constraint.

Accurate moment matching techniques for computing the RC delays and transition times are used in addition to an accurate CMOS inverter/repeater delay model that takes into account short channel effects that are prevalent in deep submicron (DSM) technologies. In particular, a new delay metric, based on the first two moments of the impulse response of the interconnect RC circuit, is derived. Also, a new empirical ramp approximation that takes into account the inherent asymmetry of signals in signal distribution networks in DSM technologies is presented.

# Résumé

Nous présentons une approche autant pratique que précise pour aborder le problème d'insertion de répéteur/inverseur CMOS tout au long d'une ligne d'interconnexion RC intégrée. Cette approche est basée sur le fait que le temps de transition et le retard a l'extrémité d'un segment de ligne RC sont, respectivement, indépendant et linéairement dépendants du temps de transition a l'entrée de l'inverseur conducteur, tant que le ratio de ces deux transitions n'excède pas une valeur prédéterminée. Dans ce contexte, nous dérivons une expression analytique pour l'espacement optimal des répéteurs ainsi que leurs tailles. Nous proposons également une nouvelle méthode "pseudo" hiérarchique pour l'insertion de répéteurs basée essentiellement sur une technique d'optimisation dite de programmation quadratique. Cette méthode, bien que largement basée sur celle, plus connue, de Van Ginneken [30], est différente dans le sens qu'elle tient compte des temps de transitions aux points d'insertion envisageables. La technique résultante peut être facilement incorporée dans un schéma plus général d'optimisation de réseaux d'interconnexions où, éventuellement, les dimensions de ces lignes peuvent être formulées tant en terme d'objectif que de contrainte.

Nous utilisons, tout au long de cette étude, les techniques les plus précises basées sur l'identification de moments pour le calcul de retard dans les lignes RC, ainsi qu'un modèle de retard d'inverseur CMOS tenant compte des effets de canaux courts prévalant dans les technologies submicroniques. Nous présentons en particulier un nouveau modèle pour le calcul de retard dans les lignes RC basée sur les deux premiers moments de la réponse a l'impulsion du circuit d'interconnections RC. Aussi, une nouvelle définition empirique de signaux en rampes pouvant approximer un signal réel est introduite. Le but est essentiellement de tenir compte des asymétries inhérentes au signaux dans les réseaux de distribution des signaux dans les technologies les plus récentes.

# Acknowledgments

First and foremost, I would like to express my deepest gratitude to Professor Nicholas Rumin who allowed me to be part of a wonderful group at McGill university. His continuous guidance, enthusiasm, and **patience** have been tremendous.

I would also like to acknowledge the contribution of a dear friend, Mohamed Hafed, for our numerous discussions that spanned a large array of subjects, including in particular, interconnect modeling and analysis.

Further, I wish to extend my thanks to friends and colleagues at the MACS lab for making my passage in this university one of the most memorables of my life. Without the pretension to be inclusive, I cite: Ahmed, Antonio, Arshan, Bardia, Boris, Christian, Daghan, Earl, Geoffrey, Ian, Lige, Mohamed, Mona, Nazmy, Ramez, Ziad. The same goes to my friends Madjid and Sarah. For all I wish the best.

Financial support by the Natural Science and Engineering Research Council of Canada is gratefully acknowledged.

Finally, my deepest gratitude to my parents, for the many sacrifices they made to provide the best education for my brothers and myself, their unconditional and unyielding support and encouragement throughout the years.

# Table of Contents

Abstract ..... I
Résumé ..... II
Acknowledgments ..... III
List of Figures ..... IV
List of Tables ..... XI

**1 Introduction** ..... 1

1.1 Technology Scaling Implications ..... 2
1.2 Overview of Interconnect Optimization techniques ..... 7
1.3 Motivation and Overview of the Thesis ..... 10

**2 Interconnect RC Effects, Model Order Reduction and Delay Metrics** ..... 12

2.1 Introduction ..... 12
2.2 Interconnect Model-Order Reduction ..... 14
2.3 Driving Point Impedance Approximation ..... 21
2.4 Analysis of Buffers with RC loads ..... 26
2.5 RC Interconnect Delay Metric ..... 32

**3 CMOS Inverter Delay and Current Model** 42

3.1 CMOS Inverter current model ..... 42
3.2 n-th. Power Law MOSFET Model ..... 50
3.3 CMOS Inverter Delay Model ..... 52
3.4 CMOS Inverter Driving RC Load ..... 58
3.4.1 Ramp Approximation for Signals in Repeated Interconnect Lines ..... 58
3.4.2 Relating the Ramp Approximation to Moments of the Impulse Response 63

**4 Repeater Insertion in RC Interconnects**

4.1 Preliminaries ..... 67
4.2 Repeater Insertion In Uniform Lines ..... 71
4.2.1 Optimal Number of Repeaters ..... 72
4.2.2 Repeater Sizing ..... 77
4.3 Bottom-Up Repeater Placement in RC Interconnect ..... 83

4.3.1 Base Technique ..... 85
4.3.2 The Proposed Method ..... 89
4.4 Results ..... 93
4.4.1 Repeater Insertion in Uniform Lines ..... 93
4.4.2 Repeater Insertion Considering Large Load Capacitance ..... 94

**5 Conclusion and Future Work** ..... 97

**References** ..... 99

# List of Figures

**Figure 1.1** Estimated and actual wire loads [5] ..... 3

**Figure 1.2** (a) Power dissipation, and (b) supply current for three successive generations of DEC Alpha microprocessors [8].. ..... 4

**Figure 1.3** Typical power breakdown in a high-performance Pentium microprocessor [9]. ..... 5

**Figure 1.4** Illustration of wire density (a), and the resulting appropriate interconnect load capacitance model. ..... 7

**Figure 1.5** 1997 NTRS projections on the maximum performances of optimized global interconnect and the required delay performance to attain the desired system operating clock cycle [4]. ..... 10

**Figure 2.1** (a) An inverter driving a number of gates through an interconnect network. (b) same inverter driving a reduced order load-model of the driven system in (a) to compute the voltage waveform at node A'. (c) A reduced order model of the path A'B driven by the voltage transition in A'computed in (b) to obtain the transition and the delay at fanout B ..... 14

**Figure 2.2** Example of RC tree modeling of interconnect networks or first order modeling of transistor clusters. ..... 18

**Figure 2.3** (a) An inverter driving a fanout trough an interconnect network.(b) Same inverter driving the three first reduced-order-models of the loading effect (driving point admittance) of the driven system in (a) ..... 23

**Figure 2.4** Four rules for upstream propagation of the driving point admittance Taylor expansion coefficients [12]. ..... 24

**Figure 2.5** Comparison between the loading effect of the actual interconnect (in term of the driving inverter 50% delay), and the first (...), the second (-.), and the third order (--) approximation of the driving point admittance. This, for both (a) semi-global and (b) global metal interconnects. ..... 25

**Figure 2.6** (a) Inverter driving a Π RC load. (b) Same inverter driving an equivalent lumped capacitance load. ..... 27

**Figure 2.7** Comparison of the output voltage waveform of an inverter loaded by Π load (-), the total load capacitance (--), and by an effective capacitance chosen to capture the 50% delay (-.). This capacitance is determined using HSPICE.....28

**Figure 2.8** (a) Inverter driving a Π RC load. (b) A lumped Ceff to capture the early portion of the inverter's output voltage (v2) waveform. (c) Equivalent linear circuit used to derive the tail portion of the inverter output voltage waveform. The charging transistor is modeled as a linear resistor. .....29

**Figure 2.9** Two poles reduced-order RC circuit. .....33

**Figure 2.10** Single time-constant-waveform capturing the 50% delay point (--) and the 70% delay point (-.) of a two-time-constant waveform (-). .....35

**Figure 2.11** The required Taylor's expansion order to capture the midpoint transition . .....36

**Figure 2.12** δ as an implicit function of the first two moments ratio. .....39

**Figure 2.13** (a) Two pole reduced-order linear circuit. The 50% (-) and 70% (--) delay points at the far end of the circuit (node *out*) for (b) $f$=0.96 and (c) $f$=0.78. The circles are the delays computed using HSPICE.....41

**Figure 3.1** (a) Transistor-level circuit model of a CMOS inverter, and (b) equivalent circuit model. .....43

**Figure 3.2** Typical Inverter short circuit current (--) and discharging current (-..) waveforms and their respective maximas. The inverter is driven by a typical rising CMOS signal.....45

**Figure 3.3** (a) The time $t_{st}$ as a function of the input transition time, and (b) $C_L/A$ as a function load capacitance, for an inverter with $W_n$ = 30μm (-), 15μm (--) and 5μm (...) The load capacitance in (a) is of 250 $f$F. Circles in (a) are the actual HSPICE simulated values of $t_{st}$ for each case. .....46

**Figure 3.4** (a) $B$ as a linear function of the load capacitance for various inverter sizes. (b) $w/E$ (-) is sufficiently modeled as having a linear dependence (-..)on inverter sizes. Circles in (a) are the actual HSPICE simulated values of $B$... .....47

**Figure 3.5** Piecewise linear model of the short-circuit current, $i_P$, in the case of a rising input .....48

**Figure 3.6** (a) The time $t_{pmax}$ as a function of the input transition time for an inverter with $W_p$ = 75μm, $W_n$ = 25μm, and $L_n$ = $L_p$ = 0.18μm. $C_L$ = 500*f*F (-) and 3*p*F(...). (b) $C_L/A$ as a function of $C_L$ for $W_p/W_n$= 15μm/5μm (-), 45μm/15μm (--), 75μm/25μm (...). Circles are the actual HSPICE simulated values for each case.....49

**Figure 3.7** Typical inverter output voltage waveform when driven by a ramp input (-) and the associated discharging current $i_N$ and short circuit current $i_P$ (--). $v_{cn}(t)$ is the output voltage $v_c(t)$ obtained by forcing $i_P$ = 0.....53

**Figure 3.8** Discharging transistor modeled as constant resistance while in triode...58

**Figure 3.9** Cascaded inverters driving an RC and pure capacitive load respectively. The first inverter is driven by a ramp signal.....59

**Figure 3.10** Applying a classic definition of the ramp input approximation (--) to a typical *RC* input signal at $v_1$ (-) results in (a) an underestimate of the delay and (b) overestimate of the supply current peak. $W_n$=15μm, $W_p$=45μm, $C_L$=1*p*F.. .....59

**Figure 3.11** comparison of the accuracy obtained in capturing the inverter's output voltage waveform at node (2) and the supply current in *inv2* when using (a) a small capacitance, $C_L$=300*f*F and (b) a large capacitance, $C_L$=1pF. Wn=15mm, Wp=45mm, R=200W and C=200fF. (-) Actual signals, (--) approximations using the saturated ramp, with slope given by the 50% & 70% of v1.....61

**Figure 3.12** Saturating the ramp approximation at a level below Vdd captures with good accuracy the effect of the actual non-symmetrical inverter input waveform on the output transition and the supply current. CL=1pF. $W_n$=15μm, $W_p$=45μm, $R$=200Ω and $C$=200*f*F. (-) Actual signals, (--) approximations using the saturated ramp. .....62

**Figure 3.13** The divergence metric between the actual inverter's input signal and its approximation is taken as the area between these two signal from the time β up to the time when the approximation saturates at $V_{dd}$.....63

**Figure 3.14** Correlation between the equivalent transition, the equivalent level of saturation and the defined divergence metric.....64

**Figure 3.15** $T_1$ dependence on $T_{in}$ for various values of $C$, the total value of the interconnect capacitance. *inv1* and *inv2* have $W_n/W_p$ =30μm/90μm, $R$ = 200 Ω.....66

**Figure 4.1** Delay dependence of M6 interconnect on its length when it is loaded and: repeated (...), not repeated (-). The line is of minimum width, and maximum lateral capacitance. Resistance per unit length $r$ is of 100Ω/mm and capacitance per unit length $c$ is of 280*f*F/mm. Repeaters are equidistant and equisized ($W_n$=10μm, $W_p$=30μm) .....69

**Figure 4.2** β20 circuit model approximation of the distributed RC line .....70

**Figure 4.3** Optimal number of inserted repeaters is uncorrelated with their size. ...71

**Figure 4.4** Assumed interconnect segment π model and notations. .....72

**Figure 4.5** (a) The resulting voltage waveforms using minimum-sized inverters are nearly symmetrical. (b) Using the ramp approximation that passes through the 50% and 70% of the actual input results in good accuracy in capturing the inverter's 50% delay for different segment lengths. Resistance per unit length is 100 Ω/mm and capacitance per unit length is 100 *f*F/mm. .....74

**Figure 4.6** Stage output waveform vout for (a) an Inverter, $W_n$=30μm $W_p$=90μm, driving a number of interconnect segment lengths and (b) a number of inverters, $W_n$=5μm $W_n$=15μm $W_n$=25μm driving a segment of optimal length, using HSPICE (-) and the model given by equation (4.19) (--)..77

**Figure 4.7** Stage's equivalent circuit when the discharging transistor is in (a) saturation and (b) triode [48]. .....78

**Figure 4.8** $\mathcal{D}_{stage}$ dependence on $T_i$ for various values of $C_{seg}$, the total capacitance of the interconnect segment. The inverter size is $W_n/W_p$ =30μm/90μm, $R_{seg}$ = 200Ω. .....79

**Figure 4.9** Procedure for obtaining $C_{eff}$ for a Π load. The Π load is decomposed into lumped capacitance $C_2$, and a Γ load $R_\Pi C_1$, whose effective capacitance is obtained using the model in Section 2.4 [48]. $v_{in}$ in this particular case is a step input. .....82

**Figure 4.10** RC routing tree with legal repeater insertion points represented by dashed inverter symbols. .....83

**Figure 4.11** Bottom-up pair generation using Elmore as the delay metric. .....86

**Figure 4.12** Bottom-up pair generation using a library of potential transition times at insertion points. .....87

**Figure 4.13** Handling of input transition time effects for required time computation in [32]. ..... 88

**Figure 4.14** Control-flow diagram of the proposed repeater insertion technique ..... 91

**Figure 4.15** Control-flow diagram of the proposed repeater insertion technique when considering the "no buffer" option represented by the symbol Φ. ..... 92

**Figure 4.16** Segmented interconnect line for repeater insertion. ..... 95

**Figure 4.17** Optimal position and size of inserted inverters/repeaters for (a) $C_{Load}$ =4*p*F, (b) $C_{Load}$ =2*p*F, (c) $C_{Load}$ =120*f*F and (b) $C_{Load}$ =10*f*F ..... 96

# List of Tables

**Table 2.1** δ Fitting parameters ..... 40

**Table 3.1** *C* and D parameters as function of inverter size. ..... 47

**Table 3.2** *M* parameter as a function of inverter size ..... 49

**Table 3.3** Optimized parameters for the n-th. power law model for the MOS transistors in the 1.8V 0.18μm CMOS process technology. ..... 51

**Table 3.4** Fitting parameters K ..... 64

**Table 4.1** Comparison of repeater insertion model with HSPICE results. ..... 93

**Table 4.2** Comparison of delay line using the proposed model and HSPICE. ..... 95

# Chapter 1

# Introduction

Until recently, the phenomenal advancement in integrated circuit technology has mainly been fueled by the continual shrinking of transistor feature sizes. Indeed, provided a proper logic design approach, integrating smaller and therefore faster transistors on consistently larger die areas[1] has long been a synonym of achieving higher system performance. However, as technology scales further down toward deep submicron (DSM) regions, system performance is no longer defined only by the underlying transistors or even individual logic functional blocks, but to a growing extent, by the overall wiring requirements of the system. As a result, the microelectronics industry and academia have recently directed a great deal of their attention toward the development of techniques for interconnect optimization, as well as circuit design methodologies that are specifically geared for efficient interconnect resource management.

In order to fully appreciate the growing importance of interconnect optimization for system performance enhancement, we present a review of the major implications of interconnect scaling that accompany the scaling of CMOS technology into DSM feature sizes. Then, a brief description of different existing techniques for interconnect delay reduction is presented. This chapter ends with the motivation and the outline of this thesis.

1. The dimensions of the die are both in terms of gate pitch and actual size, although the later is partly driven by the need for more I/O pads.

## 1.1 Technology Scaling Implications

The implications of interconnect scaling into deep submicron (DSM) are as complex as diverse, and present fundamental challenges to today's CAD tools and design flows. Probably the most problematic implication is the widening gap between gates and interconnect delay performances. That is, the delay incurred by interconnection wires increases and often dominates in datapaths. More importantly, the delay contribution of interconnect continues to increase dramatically in critical paths in IC circuits with each new generation, and it is projected that this contribution will reach 80% or even more in the near future [1]. From a pure performance measure perspective, this means that most of the clock cycle in high performance systems, such as microprocessors, is spent in carrying the data around functional blocks rather than processing it.

From an automated design process standpoint, the increasing contribution of interconnect to the delay of data paths will force a major overhaul of current design methodologies and paradigms. The reason is that, using conventional design processes, designers are able to fully design/synthesize and optimize high performance circuits prior to physical layout, largely because these processes neglect interconnect resistance. This has allowed a reliance on simple fanout-based models for interconnect. In these models, interconnect load is defined with respect to its associated fanout, rather than its actual characteristics, such as resistance and capacitance. That is, as illustrated in Figure 1.1, an interconnection having a certain fanout is modeled with a lumped capacitance whose value corresponds to a "statistical mean" of the capacitances of all the interconnections having the same fanout [2]. This value is determined by the means of some statistical analysis of the capacitive load associated with each fanout in previous designs.

As shown in the same figure, wire load for a particular fanout can span a large spectrum of values. While the error in estimating such loads is somewhat acceptable for short and medium wire lengths, it is quite large for the longest. Note that this figure applies qualitatively for both module circuit design and system designs comprising a large array of such modules. Therefore, as wires get longer, which means that interconnect load starts to dominate device load, these models become quite inappropriate because of their inconsistency with physical models. The resulting large misestimates of interconnect

delay make it quite arduous, if not impossible, to accurately identify potential critical paths, and fix all the related timing problems. These concerns are largely supported by the timing-convergence problems that are arising in today's high performance module-based IC designs, and the growing difficulty in making accurate top-down circuit performance predictions prior to physical layout.

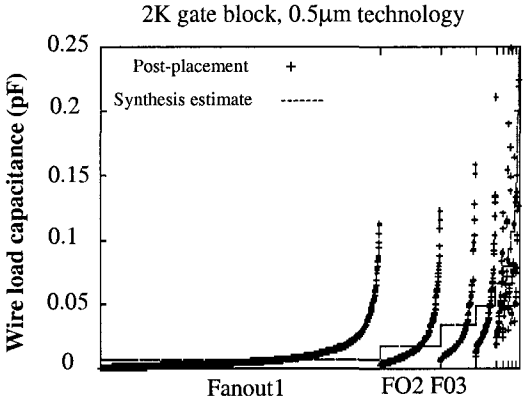



**FIGURE 1.1** Estimated and actual wire loads [2]

It is, however, quite interesting to note that, according to the study carried out by Sylvester and Keutzer [3][4], by defining an upper bound to the complexity of a functional block or module, it is possible to limit the number of relatively long wires within the block. By doing so, it becomes easier to identify potential critical paths and fix the associated timing problems manually. Note that this study is based on some conservative assumptions on the behavior of average length wires under scaling, and assumes adequate gates sizing. While this argument plays in favor of using the conventional methodologies mentioned earlier, for designing modules of typically 50000 gates [3], it completely disregards such methodologies for designing or synthesizing large systems. The reason is simply that, as the overall die capacity grows exponentially with each new technology, the number of modules[1], and thus the total number of inter-block wires, among which are long ones, will also increase exponentially. So, there is a need for new or improved CAD tools, incorporating accurate wire models, to handle long interconnection wire exceptions [2].

1. Here, it is assumed that module complexity will remain fairly constant over technology generations.

An equally important challenge that results from the scaling of CMOS technology into DSM, is the one associated with the continuing increase in supply current and power dissipation. Figure 1.2 shows the supply current and power dissipation over three successive generations of DEC Alpha miroprocessors [5]. As can be seen, despite the supply voltage reduction and the feature size scaling, power consumption has increased rapidly. The reason is that power supply currents drawn by these chips have grown at an even faster rate. Indeed, in addition to the integration of a greater number of devices operating at higher frequencies, advances in circuit design techniques and system architectures have resulted in a faster increase in switching activity, and therefore, power consumption, than would have been predicted by scaling alone. Such significant increases in power consumption and heat dissipation have make it quite difficult to provide appropriate cooling and packaging in DSM high performance systems. Moreover, the costs associated with packaging have become one of the primary concerns in integrated circuit industry. As a consequence, there are serious concerns about the fact that power consumption may set the limit on the amount of hardware that can be integrated on a single chip and the frequency at which it can be clocked.

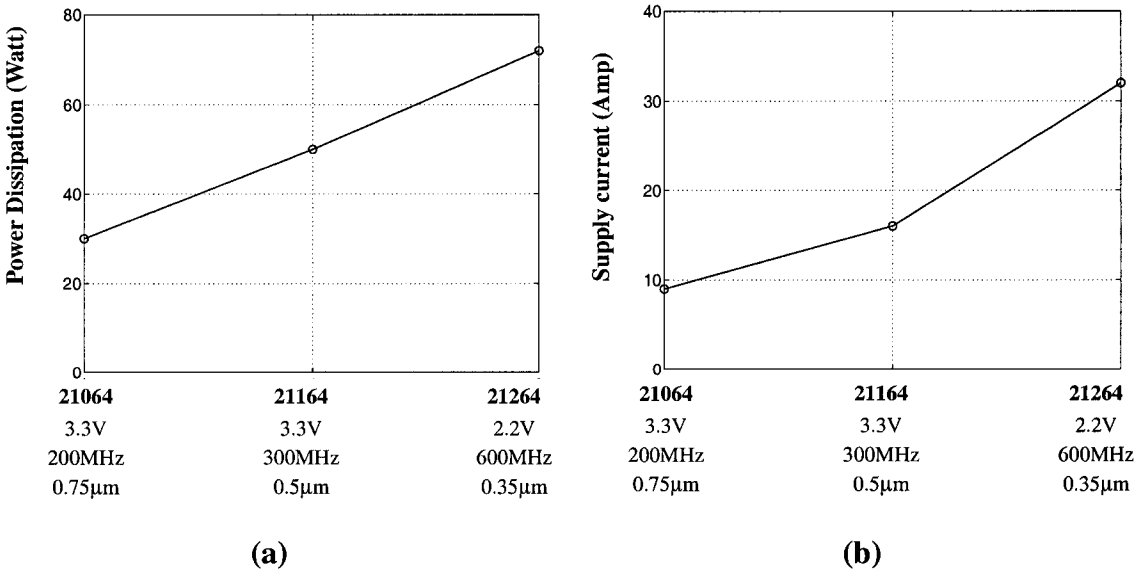



**FIGURE 1.2** **(a)** Power dissipation, and **(b)** supply current for three successive generations of DEC Alpha microprocessors [5].

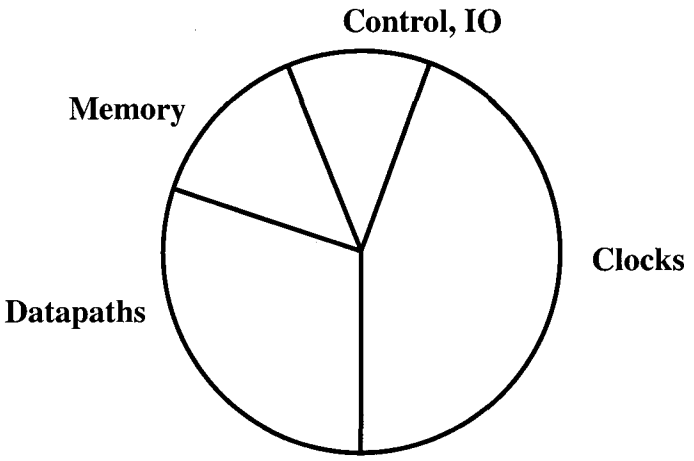



**FIGURE 1.3** Typical power breakdown in a high-performance Pentium microprocessor [6].

Figure 1.3 shows a typical breakdown of power consumption in a modern micoprocessors [6]. As illustrated, the clock is the largest power consuming component. This includes the clock generator, clock drivers, clock distribution trees, latches, and the loads that all the clocked elements represent. In any case, and irrespective of the criterion or the means utilized to determine the different sources of power dissipation, interconnect in general, whether it is for distributing power, clocks at different levels of hierarchy and signals in general, constitutes by far the largest source of power dissipation.

Based on these facts, new power efficiency-oriented design techniques have naturally focused on interconnect (especially clock networks and high capacitance data busses) to reduce the overall dissipated power, at both the wire and system architecture levels. For example, at an architectural level, power savings are realized through utilizing hierarchical clocking configurations comprising, in addition to a global clock, a large number of secondary conditional clocks. Such a clocking scheme enables the turn on of sections of the system only when needed, on a cycle-by-cycle basis, and also clock cycle stretching techniques for fast sections that are based on dynamic logic. Also, the recourse to elaborate schemes for data bus coding in patterns that lower the switching activity, have proven to be quite efficient for power savings [7]. At the wire level, however, properly sizing and spacing adjacent interconnects can contribute to reducing the effective wire load and therefore power.

Other important concerns that are arising as a result of interconnect scaling in DSM, are those associated with the increase in noise, especially for fairly long wires in data-busses, and electromigration hazards.

Basically, in order to achieve wiring density, and therefore circuit density (Figure 1.4(a)), wire pitches are dropping at the same rate as gate lengths [8], resulting in smaller interconnect cross-section and thus higher resistance values. However, in order to prevent the resistance from increasing too quickly, most processes are scaling up wire thickness, although at a slower rate[1] than the wire pitch, to limit the impact on coupling capacitance which is, by any measure, increasing dramatically. The resulting high aspect ratio wires have a pernicious side effect in that they result in a large amount of fringing and coupling capacitance to neighboring wires as illustrated in figure 1.4 (b). The latter dominates the overall wiring capacitance in lower interconnection levels (where circuit density is the highest), for technologies below the quarter micron feature size. Note that, the capacitors on the top and bottom are usually modeled as ground, since they represent orthogonally routed conductors that, averaged over the length of the wire, tend to maintain a constant voltage [8]. Such coupling produces noise that manifest itself in both crosstalk and delay deterioration. Delay deterioration is due to the fact that the interconnect load capacitance which a driver "sees" is no longer a constant value, but depends strongly on the switching pattern of neighboring wires. Since increasing the fundamental interconnect pitch is hardly an option, using low-k dielectric materials and decoupling with power and ground distribution networks, especially for clock supporting wires, are the major means to reduce such a noise.

Electromigration, on the other hand, is a reliability problem associated with large current densities. It occurs when large currents flow through a metal thereby displacing metal ions over time causing opens in the line or shorts to neighboring wire. Therefore, carefully sizing metal interconnections in the context of a trade-off involving reliability, performance and circuit density becomes imperative. Note that using copper (Cu) as the conducting material, would alleviate such problems, since it has a mean time to failure that

1. The thickness of each interconnect layer is one of the strong determinants of the total number of interconnect layers that can be afforded by the manufacturing process, and therefore the wiring resources of the corresponding technology.

is a hundred times longer than of Aluminum's. The problem, is that processes that include such materials are still beyond the reach of many designers.

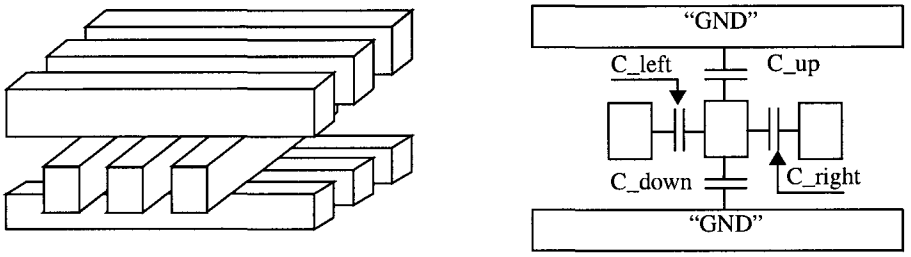



**FIGURE 1.4** Illustration of wire density (a), and the resulting appropriate interconnect load capacitance model.

## 1.2 Overview of Interconnect Optimization techniques

Because of the role of interconnect as a system performance determinant, significant research in the area of interconnect modeling and optimization has been conducted. Interconnect optimization, which encompasses a variety of techniques such as wire sizing, buffer insertion and buffer sizing, has become a crucial part in any modern design process of high performance VLSI system. The primary objective of such optimization techniques is delay reduction through interconnect lines. However, as technology scales down to DSM with gigahertz operating frequencies, targeting efficient management of the power budget, preserving the integrity of signals, and limiting the resulting impact on the additional silicon area, becomes more and more important.

Probably the most widely accepted and documented technique for global and semi global interconnect timing optimization is buffer insertion/sizing, which allows a significant reduction in propagation delay by mitigating the effects of resistance. Essentially, the delay of an uninterrupted wire[1] grows quadratically with wire length. By inserting and sizing repeaters periodically along the wire, such that the delay incurred by an individual repeater, roughly equals the delay through a resulting wire segment, the total delay

1. Here we assume uniform lines, and ignore possible via cuts from the upper metal levels down to the substrate

through the repeated line becomes almost linear [9]. In addition, wire sizing (widening), can contribute to reducing the delay by reducing the wire's resistance. However, this can also increase delay by increasing the wire's capacitance. Therefore, optimal wire sizing depends on various factors such as driver strength, interconnect topology, timing constraint and technology parasitics.

As mentioned earlier, methodologies for timing optimization have been extensively studied and reported in the literature. These methodologies make use of all possible combinations of delay reduction techniques. For example, in [10][11][12][13], repeater insertion/sizing techniques to reduce delay in uniform lines, while assuming fixed wire dimensions, are considered. Most of these studies resulted in elegant methods, where the problem of inserting and sizing repeaters is formulated as a constrained nonlinear mathematical program, typically quadratic or simplex, and switch-level models for interconnect and repeaters, typically CMOS inverters, are used to evaluate the delay constraints. In addition, in [11][12], for instance, the (delay) objective and constraints are formulated as convex functions, which guarantees the convergence to a global minimum. The problem with most of these approaches, however, is basically their reliance on crude first-order delay estimation models, such as the Elmore delay, whose inaccuracy in DSM technologies is notorious. Although interfacing accurate timing analysis tools or circuit simulators to these algorithms ultimately results in more accurate and optimal optimization solutions, it also leads, inexorably, to prohibitive computation costs.

In [14][15][16][17][18][19] the problem of optimizing the delay through interconnect by wire sizing alone is considered. In these studies, the wire is usually divided into small fixed-length segments, and the width of each segment optimized iteratively. Chen *et al* [14][15] considered another variant of this approach, by assuming a continuous set of possible wire widths, and concluded that, under the Elmore delay model, the optimal wire width is an exponential tapering function of the distance from the driver. The major problem with this method is that it requires quite sophisticated routers to avoid wasted area near the sinks and congestion near the drivers. As a consequence, designers have become reluctant to integrate such sizing approach into complete routing methodologies and tools. Moreover, these wire shapes are inherently more subject to electromigration failure due to

the increase in current densities as wires taper to narrower ones. Furthermore, it has been recently proved that the reduction in propagation delay when using wire sizing with tapering, as opposed to uniform wire sizing, in conjunction with buffer insertion, results in a marginal improvement in propagation delay [20]. This has risen serious questions about its relevance when considering long interconnect wires.

Since wire sizing and repeater insertion together affect timing, power consumption and overall silicon area, several works have studied their simultaneous optimization. Example of such studies are [12][21][22]. These are, typically, extensions of the algorithms for repeater insertion/sizing, and are based on the fact that the problem can be formulated as a convex quadratic program. A nearly comprehensive survey of all the algorithms mentioned so far can be found in [23].

Examples of techniques that target closed-form expressions of delay optimization solutions are [10][24][25]. In [10], for instance, the authors improved upon the buffer insertion methodology presented by Bakoglu in [9] by using more accurate models for CMOS inverters used as repeater elements. That is, MOS transistors were characterized using the Alpha-power law [26] to model the short channel effects which are prevalent in modern CMOS technologies, rather than using a simple current source driving an RC element. However, they assumed step input signals on the inverters, and made the approximation that transistors operate mostly in the triode mode, and therefore completely ignored saturation regime. Moreover, interconnection wires were modeled as $\Gamma$ lumped RC model, and the "screening" effect of a part of the total load capacitance by the series load resistance was ignored implicitly, thereby greatly overestimating inverter delay. In [25] the authors attempted to improve on that model by considering both the saturation and triode modes of operation of the transistors, and assuming saturated ramp approximation for the inverters input signals. The major problem with these techniques is that they ignore very important aspects of today's CMOS technologies, namely the typical signal waveform, especially in highly resistive interconnection wires, the distributed nature of interconnect, and their loading effect on driver gates in general.

Another class of optimization techniques or tools that are potentially applicable in timing analyzers, in that they can include very accurate models of both interconnect and

gates, are those based on the Van Ginneken method for buffer insertion[27]. Examples of these methods are [28][29][30]. A detailed discussion of these techniques will follow in Chapter 4, where we propose further improvements.

## 1.3 Motivation and Overview of the Thesis

Despite the growing importance of interconnect timing optimization in the design process of VLSI systems, as illustrated in Figure 1.5, there seem to be very little work on including the RC distributed nature of interconnects and accurate gate models in an effective way. Note that, although the techniques outlined in the previous section generally result in an appreciable delay reduction, the inherent inaccuracy of the models on which they are based, especially for deep submicron technologies, can simply lead to inferior optimization solutions. Moreover, using the techniques that rely on pure (general) non-linear mathematical formulations of the problem in conjunction with exact gate and interconnect delay models can result in prohibitive computation times. This can preclude a design from hitting a market window opportunity.

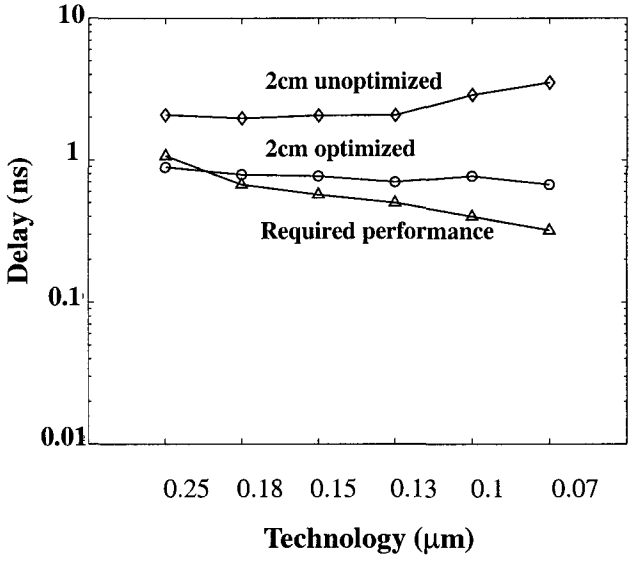



**FIGURE 1.5** 1997 NTRS projections on the maximum performances of optimized global interconnect and the required delay performance to attain the desired system operating clock cycle [1].

Most repeater insertion techniques that take into account interconnect resistive effects, utilize simplified and therefore inaccurate interconnect and gate delay models. That is, they use low order models for interconnect (see Chapter 2), do not take into account the operating regions of the repeater's constituent transistors in a satisfactory fashion, and completely ignore the inherently asymmetrical waveform shape of signals in repeated RC lines (see Chapter 3).

This thesis consists of three parts. The first part, Chapter 2, describes in some detail some of the relevant techniques for interconnect model-order reduction, and their loading effect on driver gates. A new, simple, and accurate interconnect delay metric is then derived. In the second part, Chapter 3, we start by presenting the modeling scheme of the CMOS inverter/repeater, used as the repeating element, which incorporates various RC interconnect effects. In parallel, we analyze the context in which interconnect wires and networks are designed and optimized. That is, we put forth one of the most important characteristics of optimized networks, namely "equalized" transition times at the end of each interconnect segment. This analysis leads to the third part, i.e Chapter 4, where techniques for inserting repeaters in uniform lines are presented. In particular, we present a technique, based on Van Ginneken's bottom-up approach for inserting repeaters, that takes transition time requirements at the interconnect's repeater insertion points into account. Finally, conclusions and comments are made in Chapter 5.

# Chapter 2

# Interconnect RC Effects, Model-Order Reduction and Delay Metrics

This chapter is mostly a review of the techniques that are utilized to speed up the delay and power analysis of interconnects and their associated buffers. We start by outlining interconnect model reduction techniques, such as moments matching, upon which VLSI interconnect analysis is based (section 2.2). Then, we show how moments matching is used to approximate the interconnect-load driving point admittance (section 2.3) by models that can be utilized in combination with advanced nonlinear CMOS gate macromodels (chapter3) to provide fast yet accurate timing simulation and optimization results. In Section 2.4, we present methods for the timing analysis of CMOS gate macromodels, which utilize the concept of an effective capacitance to model the driving point admittance of the interconnect load. Finally, in section 2.5, we introduce a new, simple, and accurate RC interconnect delay metric.

## 2.1 Introduction

The increasing speed and density of transistors in integrated circuits has made delays due to interconnect an increasingly important, if not dominant, factor in circuit and system performance. As a result, accurate interconnect system models have to be included

in circuit simulation. However, due to the distributed nature of interconnects, the poles of such systems are transcendental and theoretically infinite in number. Therefore simplified interconnect models are used instead. However, such models may contain a large number of energy-storage elements making subsequent circuit simulations very onerous in terms of computation time. One solution consists of using model-order reduction schemes to reduce the number of these poles without affecting the transient response properties of the model. Obviously, this has resulted in faster circuit simulations. Unfortunately, even when using very simple yet accurate interconnect models, dynamic simulators like spice may still prove to be inadequate for large circuits timing verification and optimization.

To bypass the inefficiency of electrical (analog) simulators, fast timing and power analysis techniques, in combination with efficient techniques for interconnect model-order reduction, have been developed and incorporated into elaborate logic simulators. The leitmotiv in developing all these techniques has obviously been to achieve reduced computational complexity, therefore computational speed, while retaining the maximum accuracy.

The basic operation of these simulators is hierarchical rather than simultaneous. That is, in contrast to electrical simulators such as SPICE, each element in a circuit, whether it is a gate or an interconnect segment, is analyzed and simulated separately based on upstream signal waveform and downstream load information. For instance, considering the simplified net in Figure 2.1 (a), where the delay from the input of the first (driving) gate at node A to the input of the fanout at node B, is defined as the sum of delays through the gate, i.e. path AA', and the interconnect path A'B. To compute the delay through the first gate (path AA'), a reduced order model of the loading effect of the entire interconnect network and fanout gates on the driving gate is first computed. Then, given this load model, the input transition at node A and the gate's physical parameters, the voltage waveform at node A' is determined (Figure 2.1 (b)). This computed waveform is fed through a reduced order model of the interconnect path A'B to compute the resulting waveform and delay at node B (Figure 2.1 (c)). We assume here that interconnects can be accurately modeled as RC tree networks with a capacitor from each node to ground, no floating capacitors, no resistor loops and no resistors to ground.

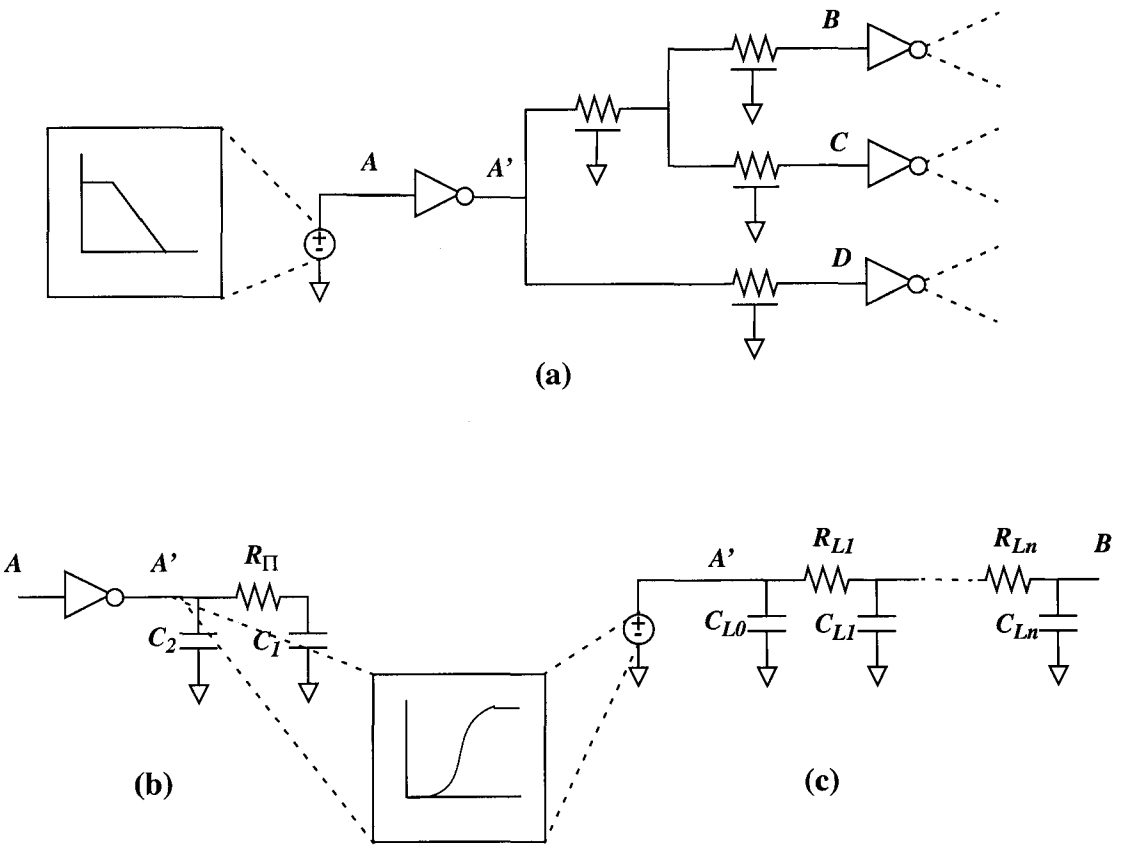



**FIGURE 2.1** (a) An inverter driving a number of gates through an interconnect network. (b) same inverter driving a reduced order load-model of the driven system in (a) to compute the voltage waveform at node A'. (c) A reduced order model of the path A'B driven by the voltage transition in A'computed in (b) to obtain the transition and the delay at fanout B.

## 2.2 Interconnect Model-Order Reduction

Timing simulators rely on models that capture the effect of interconnect on the delay and power to produce accurate yet fast timing verification and power analysis. Such models are based on the principle that, although a detailed linear or linearized circuit model of an interconnection may contain a large number of energy storage elements resulting in a comparable number of frequency domain poles, only a few of these poles dominate in the transient response of such a circuit. Therefore, capturing the dominant

poles, or an equivalent set of poles, leads to an excellent reduced-order interconnect model. Since, in general, the number of equivalent poles is very small, the resulting analysis is very fast.

Starting with a rational transfer function describing any node voltage or branch current we can write [31]:

$$\begin{aligned} H(s) &= \frac{1 + a_1 s + a_2 s^2 + \ldots + a_n s^n}{1 + b_1 s + b_2 s^2 + \ldots + b_m s^m} \\ &= \frac{k_1}{s - p_1} + \frac{k_2}{s - p_2} + \ldots + \frac{k_m}{s - p_m} \end{aligned} \tag{2.1}$$

where $p_i$ are the poles of the system and $k_i$ are the associated residues. $m$ is the order of the transfer function, where $m \geq n$. The techniques for interconnect model reduction comprise two successive phases; moments computation, and moments matching.

First, we define the moments of a circuit's impulse response $h(t)$. If $H(s)$ is the Laplace transform of $h(t)$, i.e the circuit's transfer function, we can by definition write:

$$H(s) = \int_0^\infty h(t) e^{-st} dt \tag{2.2}$$

Applying a Taylor series expansion of $e^{-st}$ about s = 0 yields:

$$\begin{aligned} H(s) &= \int_0^\infty h(t) \left\{ 1 - st + \frac{1}{2!} s^2 t^2 - \frac{1}{3!} s^3 t^3 + \ldots \right\} dt \\ &= \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} s^i \int_0^\infty t^i h(t) dt \end{aligned} \tag{2.3}$$

The $i$-th circuit response moment is defined as:

$$m_i = \frac{(-1)^i}{i!}\int_0^\infty t^i h(t)dt \tag{2.4}$$

Note that $\int_0^\infty t^i h(t)dt$ is nothing but the $i$-th time moment of the circuit impulse response $h(t)$[1]. Therefore the transfer function $H(s)$ can be expressed as

$$H(s) = m_0 + m_1 s + m_2 s^2 + m_3 s^3 + m_4 s^4 + \ldots \tag{2.5}$$

The earliest reduced-order models were based on a single dominant pole, and were first used to characterize linear or linearized circuits modeled as RC trees[2]. Such an approximation was considered accurate enough in estimating internal and output node voltage waveforms in RC interconnects in earlier process and design technologies. The approximate transfer function model of an RC circuit based on a single-time-constant reduces to:

$$\hat{H}(s) = \frac{1}{1 + \tau_D s} \tag{2.6}$$

where $\tau_D$ is the equivalent time constant, also known as the Elmore time constant [33]. Its value is determined by matching the first moment of $\hat{h}(t)$ to the first moment of the exact impulse response:

$$\int_0^\infty t\hat{h}(t)dt = \int_0^\infty th(t)dt \tag{2.7}$$

Therefore: $$\tau_D = m_1 \tag{2.8}$$

1. It is because of this analogy between the expressions of Taylor series expansion coefficients and time moments of $h(t)$ that $(m_i)$ are generally referred to as moments.

2. Historically, such a dominant pole approach, was used primarily in timing verification where transistor or gate clusters were modeled as RC trees, i.e, transistor channels were modeled as linear equivalent resistors. Interconnect effects were not significant [32].

The Elmore constant has been and for a long time the favorite measure for computing delays through RC trees, and it became the defacto standard delay metric for performance-driven design and optimization. The reason for its success resides in the fact that the mean and median of a typical CMOS waveform coincide[1] fairly well, given their near symmetry. Therefore, the mean value of a signal transition, i.e its first time moment, has been considered as an accurate approximation to the time at which the output response $v(t)$ of an interconnect to a step input reaches 50% of its final value:

$$\tau_D = \int_0^\infty t\dot{v}(t)dt \tag{2.9}$$

A path tracing algorithm was proposed by Rubenstein, Penfield and Horowitz to calculate the value of the first moment or Elmore constant in polynomial time[34]. This is done by traversing the tree topology (Figure 2.2), and summing up contributions of branches along a path $P_i$ from the root to the node of interest $i$. The contribution of a branch $j$ is the product of its resistance and the total downstream capacitance, $Cdsj$. Therefore the Elmore delay $\tau_D$ of path $P_i$ (from the root to node $i$) can be expressed as [3]:

$$\tau_D = \sum_{j \in P_i} R_j \cdot Cds_j = \sum_{j \in P_i} \left[ R_j \cdot \sum_{k \in ds(j)} C_k \right] \tag{2.10}$$

where $ds(j)$ is the set of capacitances downstream of $j$. For instance, the Elmore constant from the root of the RC tree in figure 2.2 to the positive node of the capacitance $C_4$ is expressed in terms of the circuit's constituent elements as:

$$\tau_D = R_1(C_1 + C_2 + C_3 + C_4) + R_2(C_2 + C_3 + C_4) + R_4C_4 \tag{2.11}$$

1. This is true for situations where devices present an output resistance higher than of the interconnect loads, which is typical for technologies anterior to the 0.25μm process.

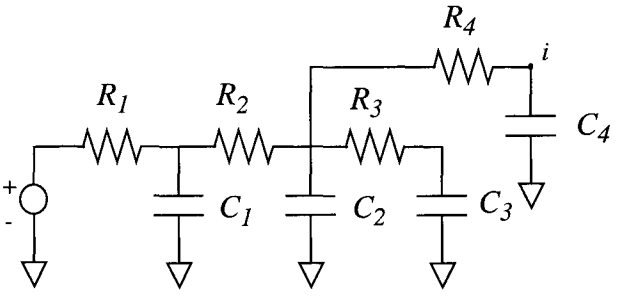



**FIGURE 2.2** Example of RC tree modeling of interconnect networks or first order modeling of transistor clusters.

Later, an attempt to characterize the transfer function using two poles was made by Horowitz in [32]. The proposed model attempted to define higher order moments of a transfer function in terms of circuit elements. In fact, no higher order moment than the second were computed, so the approximation to the transfer function was built using the two first moments, corresponding to a two pole and one zero approximation of the form:

$$\hat{H}(s) = \frac{k(1+s\tau_z)}{(1+s\tau_1)(1+s\tau_2)} \tag{2.12}$$

The values of the poles and zero are calculated by matching the first two moments of the circuit's exact impulse response to the first two moments of $\hat{h}(t)$, in addition to some heuristics to compensate for a necessary third equation, ideally matching for the third moment:

$$\int_0^\infty t\hat{h}(t)dt = \int_0^\infty th(t)dt \tag{2.13}$$

$$\int_0^\infty t^2\hat{h}(t)dt = \int_0^\infty t^2h(t)dt \tag{2.14}$$

$$H(s) = \frac{1+a_1s+a_2s^2+\ldots}{1+(\tau_1+\tau_2)s+b_2s^2+\ldots} \tag{2.15}$$

This technique was developed to model waveforms in RC trees that cannot be successfully modeled with a single time constant, especially charge sharing situations where the waveform usually needs multiple time constant modeling.

Model reduction using Asymptotic Waveform evaluation (AWE) [31] is a more recent and popular technique which attempts to compute and match an arbitrary number of higher order moments.The underlying idea is to approximate the transfer function of the linear(ized) network by a reduced set of $2q$ poles $p_i$ and residues $k_i$, corresponding to an approximate transfer function of the form:

$$\hat{H}(s) = \frac{b_{q-1}s^{q-1} + b_{q-2}s^{q-2} + \ldots + b_1 s + b_0}{a_q s^q + a_{q-1}s^{q-1} + \ldots + a_1 s}, a_q \neq 0 \tag{2.16}$$

which can be easily cast into the time domain equivalent:

$$\hat{h}(t) = k_1 e^{p_1 t} + k_2 e^{p_2 t} + \ldots + k_q e^{p_q t} \tag{2.17}$$

Applying the moments definition in equation (2.4) to $\hat{h}(t)$ in equation (2.17), the approximate moments to be forced to match the actual ones are of the form:

$$\hat{m}_i = \frac{k_1}{p_1^{i+1}} + \frac{k_2}{p_2^{i+1}} + \ldots + \frac{k_q}{p_q^{i+1}} \tag{2.18}$$

This explicit moment matching technique, also known as "Padé" approximation, has fundamental limitations. That is, the Padé approximation is accurate only near the point of expansion. As a result, moment matching techniques based on a single expansion (at the origin of the complex plane) often gives inaccurate[1] results. To overcome this problem, multiple expansion points are used. For example, in [35] the authors proposed a tech-

1. One of the main problems, is that Padé can produce unstable poles even when generated from stable RLC circuits, especially for high approximation orders.

nique, namely Complex Frequency Hopping (CFH) where they extended the process of moments matching to multiple expansion points. These points are determined via a binary search algorithm, starting at DC and a predefined maximum frequency of interest. In parallel, a new class of state-space-based algorithms for model-order reduction, often referred to as implicit moment matching, have been developed [36][37][38]. The major advantage of these techniques is their ability to guarantee the passivity of the reduced models, i.e their stability.

As for the actual time moments in linear circuits, $m_i$, AWE provides an elegant method to compute their values. This method requires a limited number of successive DC analyses of the circuit model, where capacitors are replaced by DC current sources. The computation process begins by replacing the input driver with a DC source set to the final value, and the capacitors with zero valued current sources. The first moment corresponding to the transfer function with node $i$ as the output node is the voltage across the capacitor at this node. Higher order moments are computed by setting the driver to zero and replacing each capacitor with a current source valued as the product of its previous moment and respective value of capacitance. As a result, computing the exact moments in an RC tree becomes trivial and of the form:

$$\hat{m}_i^{(n)} = -\sum_{j \in P_i} \left[ R_j \cdot \sum_{k \in ds(j)} C_k \hat{m}_k^{(n-1)} \right] \tag{2.19}$$

where $i$ is the node for which moments are computed, $n$ is the order of the moment and the other parameters are as defined for equation (2.10). For instance, recursive computation of the moments at node $C_4$ via path tracing in Figure 2.2 yields:

$$
\begin{aligned}
m_0 &= 1 \\
-m_{1-C_4} &= R_1(C_1 + C_2 + C_3 + C_4) + R_2(C_2 + C_3 + C_4) + R_4C_4 \\
m_{2-C_4} &= R_1(C_1m_{1-C_1} + C_2m_{1-C_2} + C_3m_{1-C_3} + C_4m_{1-C_4}) \\
&\quad + R_2(C_2m_{1-C_2} + C_3m_{1-C_3} + C_4m_{1-C_4}) + R_4C_4m_{1-C_4} \\
&\bullet \\
&\bullet \\
&\bullet
\end{aligned}
\tag{2.20}
$$

In fact there are various methods for computing the actual moments of a given transfer function, and an exhaustive survey of such techniques, with their points of strength and limitations, can be found in [39].

## 2.3 Driving Point Impedance Approximation

In this section we present the work published in [40] a decade ago, where an algorithm for estimating the driving point admittance, or equivalently the loading effect, of the driven interconnect on any arbitrary driving gate, is developed. The approach presented herein is analogous to those upon which the techniques for interconnect model reduction are based. That is, it is based on matching the coefficients of the Taylor series expansion of the exact driving point admittance $Y(s)$ with those of a predefined approximation $Y_{appi}(s)$ of $Y(s)$, where $i$ is the order of such an approximation. In other words, by expanding $Y(s)$ about $s = 0$, we can write:

$$
Y(s) = \sum_{i=1}^{\infty} y_i s^i \tag{2.21}
$$

If $y(t)$ denote the inverse Laplace transform of $Y(s)$, matching higher-order terms in (2.21) is mathematically equivalent to matching higher order time moments of $y(t)$ since, as can be easily demonstrated:

$$
y_i = \frac{(-1)^i}{i!} \int_0^{\infty} t^i y(t) dt \tag{2.22}
$$

Therefrom the analogy mentioned earlier.

As shown in Figure 2.3, the first order approximation (equivalently, first moment matching) is simply the total lumped capacitance downstream of the driving gate for which the loading effect of the RC network is computed, i.e $Y_{app1}(s) = C_{tot}s$, where $y_1 = C_{tot}$. Until recently, this approximation has been the most utilized in hierarchically based timing analyzers, rendering the delay analysis of on-chip gates quite practical. However, as circuit operating speeds increase, and interconnect resistance also increases, due to both increase in length and reduction in cross section, the resulting effects of the series resistance can no longer be ignored. Increasing the order of the approximation results in better accuracy in capturing the driving point admittance of the RC load.

The lumped Γ RC model approximation:

$$\begin{aligned} Y_{app2}(s) &= \frac{C_1 s}{1 + RC_1 s} \\ &= \sum_{i=1}^{\infty} (-1)^{i-1} R^{i-1} C_1^i s^i \end{aligned} \tag{2.23}$$

matches $Y(s)$ to the second order by setting:

$$\begin{cases} C = C_{tot} = y_1 \\ R = -y_2 / y_1^2 \end{cases} \tag{2.24}$$

The third order approximation, i.e the Π model approximation:

$$\begin{aligned} Y_{app3}(s) &= C_2 s + \frac{C_1 s}{1 + RC_1 s} \\ &= (C_1 + C_2)s + \sum_{i=2}^{\infty} (-1)^{i-1} R^{i-1} C_1^i s^i \end{aligned} \tag{2.25}$$

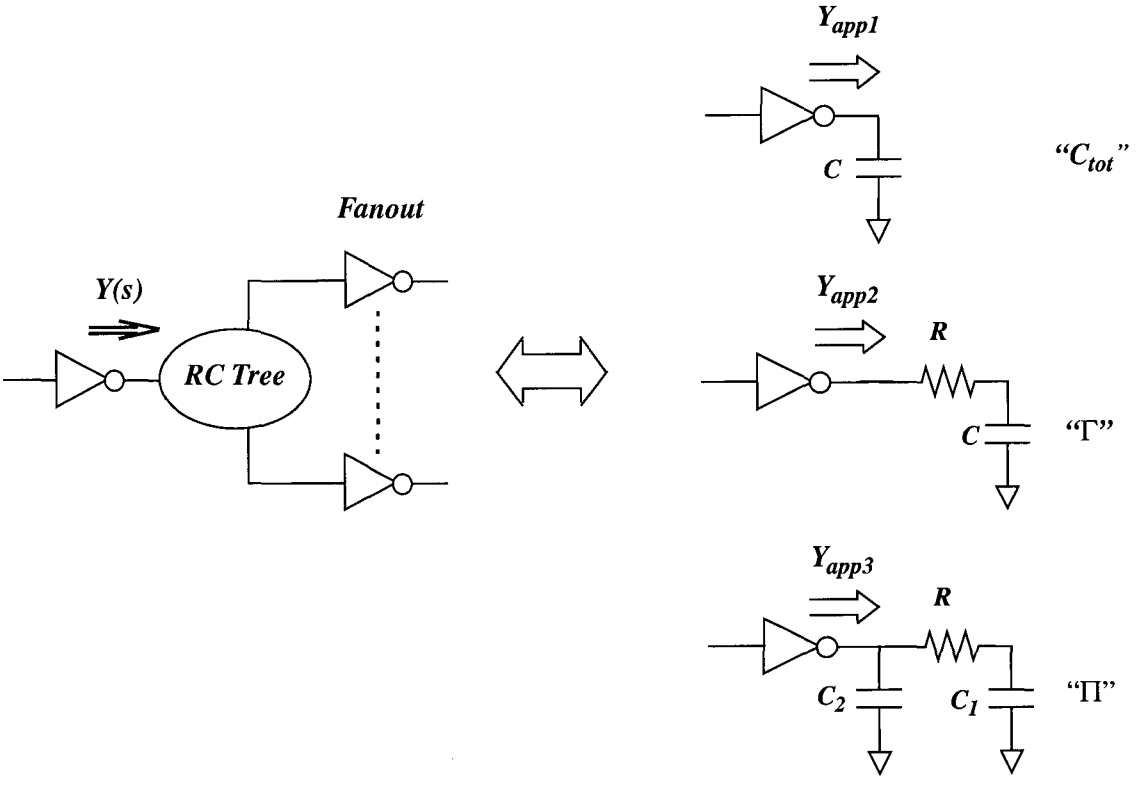
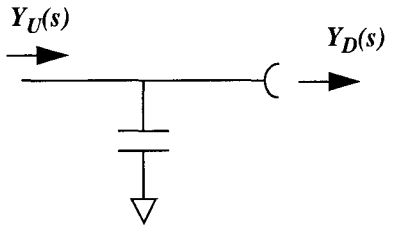



**FIGURE 2.3** (a) An inverter driving a fanout trough an interconnect network.(b) Same inverter driving the three first reduced-order-models of the loading effect (driving point admittance) of the driven system in (a).

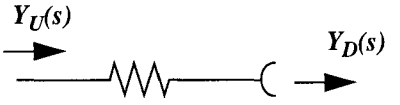
matches $Y(s)$ to the third order by setting:

$$\begin{cases} C_1 = y_2^2/y_3 \\ C_2 = y_1 - y_2^2/y_3 \\ R = -y_3^2/y_2^2 \end{cases} \tag{2.26}$$

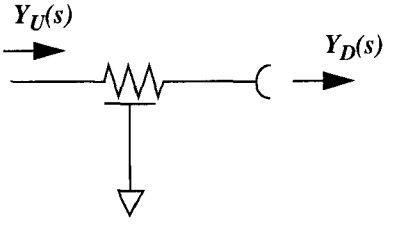
Note that so far the coefficients $y_1$, $y_2$ and $y_3$ are assumed to be known a priori. In fact all we know up to this point is that $y_1$ is equivalent to the total lumped capacitance $C_{tot}$. O'Brien and Savarino presented an algorithm that starts at the leafs of the RC tree and works back to the source of the tree. During this bottom-up tree traversal the coefficients of the driving point admittance looking downstream of a given point are propagated further upstream, following a set of rules (Figure 2.4).



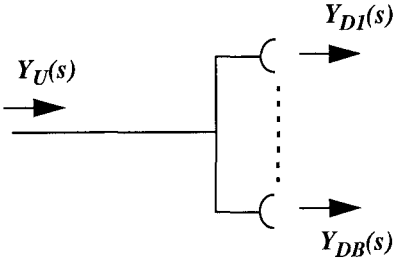

$$(y_U)_1 = (y_D)_1 + C$$
$$(y_U)_2 = (y_D)_2$$
$$(y_U)_3 = (y_D)_3$$




$$(y_U)_1 = (y_D)_1$$
$$(y_U)_2 = (y_D)_2 - R(y_D)_1^2$$
$$(y_U)_2 = (y_D)_3 - 2R(y_D)_1(y_D)_2 + R^2(y_D)_1^3$$




$$(y_U)_1 = (y_D)_1 + C$$
$$(y_U)_2 = (y_D)_2 - R\left[(y_D)_1^2 + C(y_D)_1 + \frac{1}{3}C^2\right]$$
$$(y_U)_3 = (y_D)_3 - R[2(y_D)_1(y_D)_2 + R^2(y_D)_1^3]$$
$$R^2\left[(y_D)_1^3 + \frac{4}{3}C(y_D)_1^2 + \frac{2}{3}C^2(y_D)_1 + \frac{2}{15}C^3\right]$$

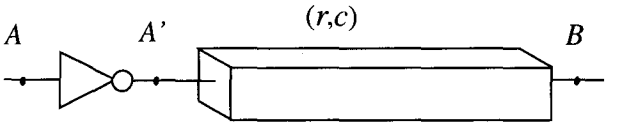



$$(y_U)_1 = \sum_{i=1}^{B} (y_{Di})_1$$
$$(y_U)_2 = \sum_{i=1}^{B} (y_{Di})_2$$
$$(y_U)_3 = \sum_{i=1}^{B} (y_{Di})_3$$

**FIGURE 2.4** Four rules for upstream propagation of the driving point admittance Taylor expansion coefficients [12].

In Figure 2.5, we compare the three approximations of the driving point admittance of a typical global and a semi-global uniform interconnect in a modern CMOS process as a function of length (or equivalently resistance). As can be seen, while the total capacitance $C_{tot}$ is a poor measure of delay, the $\Pi$ model on the other hand is an accurate enough measure of delay. Note that longer metal lengths require higher-order lumped cir-

cuit approximations to accurately model the resistive "shielding" of capacitance located far from the driver.



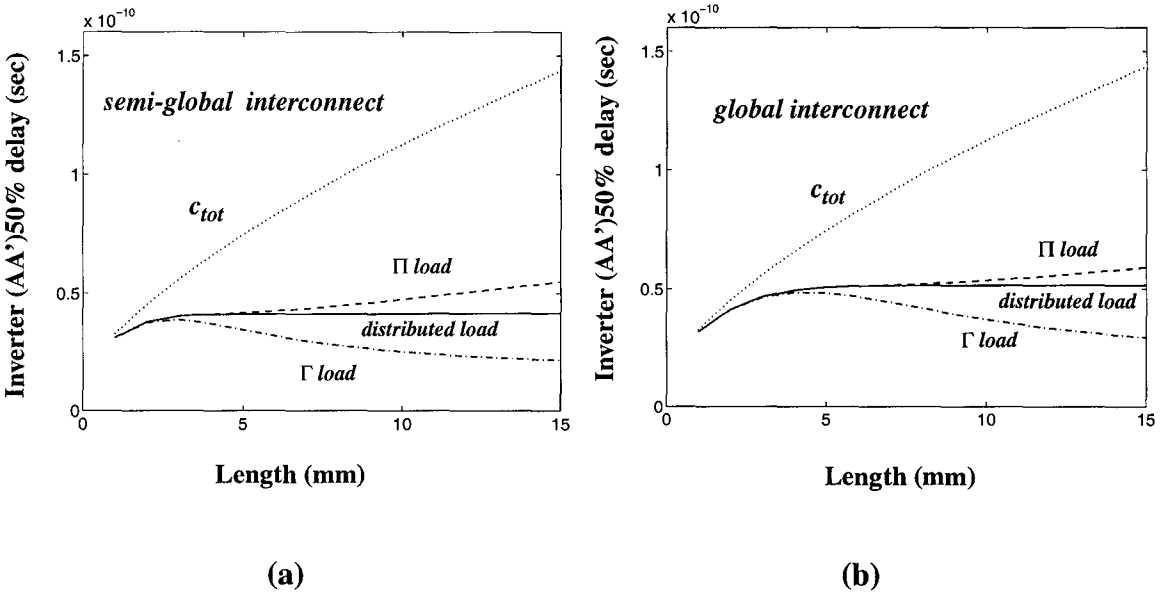




**FIGURE 2.5** Comparison between the loading effect of the actual interconnect (in term of the driving inverter 50% delay), and the first (...), the second (-.), and the third order (--) approximation of the driving point admittance. This, for both (a) semi-global and (b) global metal interconnects.

## 2.4 Analysis of Buffers with RC loads

The timing analysis of a CMOS gate driving a pure capacitive load has been extensively studied and various accurate delay models have been reported in the literature, such as in [41][53]. Surprisingly, including interconnect loading effects (i.e taking the resistive effect of interconnect into account), even if described by a simple equivalent Γ or Π *RC* load, proves to be considerably more difficult. As a result, a commonly used solution for obtaining a given delay point at the output of a driving gate, is to define an effective capacitance, that will result in the same delay as that due to the interconnect *RC* loading effect model. In this section, we consider the computation of the effective capacitance where the driving gate is a CMOS inverter and assume a falling ramp input signal.

In [43], the concept of an effective capacitance, $C_{eff}$, is used primarily to compute the delay through a CMOS gate, typically an inverter, and is defined as the single equivalent load capacitance that will result in the same 50% point delay as a Π model RC load (Figure 2.6). An expression for this capacitance is established by equating the average current drawn (i.e the total charge transferred) by both loads up to the midpoint transition of $v_{out}(t)$ at $t_D$:

$$\frac{1}{t_D}\int_0^{t_D} I_{\Pi}(t)dt = \frac{1}{t_D}\int_0^{t_D} I_{C_{eff}}(t)dt \tag{2.27}$$

The authors of [43] assumed the following waveshape for $v_{out}(t)$ to solve equation (2.27) for $C_{eff}$:

$$v_{out}(t) = \begin{cases} v_i - ct^2 & 0 \le t \le t_x \\ a + b(t - t_x) & t_x \le t \le t_D \end{cases} \tag{2.28}$$

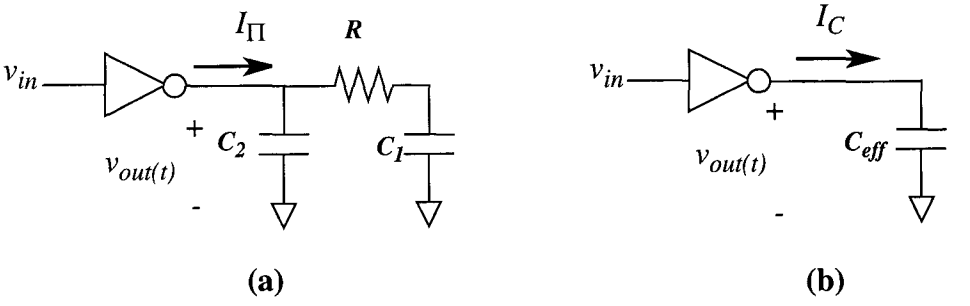



**FIGURE 2.6** (a) Inverter driving a Π RC load. (b) Same inverter driving an equivalent lumped capacitance load.

where $v_i$ is the initial output voltage and $t_x$ is the time at which the output completes 20% of its transition. Hence, it is assumed that the output voltage $v_{out}(t)$ follows a quadratic shape up to the 20% point $t_x$, then from the 20% point up to the midpoint $t_D$, where the transistors are assumed to be in saturation, the inverter output voltage is assumed to be linear. Thus, equating the average currents, using this waveform model results in the following expression of $C_{eff}$:

$$C_{eff} = C_2 + C_1\left[1 - \frac{RC_1}{t_D - t_x/2} + \frac{(RC_1)^2}{t_x(t_D - t_x/2)}e^{\frac{-(t_D - t_x)}{RC_1}}\left(1 - e^{\frac{-t_x}{RC_1}}\right)\right] \tag{2.29}$$

As can be seen, the effective capacitance is a function of the quantities sought; the delay $t_D$, and the time $t_x$. Therefore, the effective capacitance is calculated iteratively using a *k-factor* equations delay model. The computation procedure is detailed in [43].

The fundamental problem with this approach, in addition to its relative complexity[1], is that it does not take into consideration the conditions under which the driving point admittance of the *RC* load may be approximated by a capacitance. In fact, the output of a CMOS inverter driving an *RC* load is typically a two-time-constant waveform. As Figure 2.7 can suggests, while a single capacitance can accurately capture the lower (early) por-

1. The necessary curve fitting to compute the different parameters in equation (2.28) can result in a large look up table, especially if maximum accuracy, under this technique, is sought.

tion of the driver's output transition (p-channel transistor in saturation), it ultimately results in large errors in estimating the upper portion or tail (the p-channel transistor in triode). Therefore, the assumption that the definition of $C_{eff}$ is valid until the output voltage reaches the midpoint in its swing, not only can result in an inaccurate value of $C_{eff}$ (in situations where the charging transistor is in triode for a portion of this voltage swing) but also, inherently, yields a poor estimate of the entire output voltage waveform. Such an error tends to propagate and can cause large misestimates at output nodes of a subsequent inverter (considering, for instance, a repeated RC interconnect line) in which the current values through the discharging n-channel transistor can be significantly affected. Also, note that the quadratic output waveform model is not realistic in many cases, including overshoots/undershoots caused by Miller capacitances and very fast and slow output transitions [44].

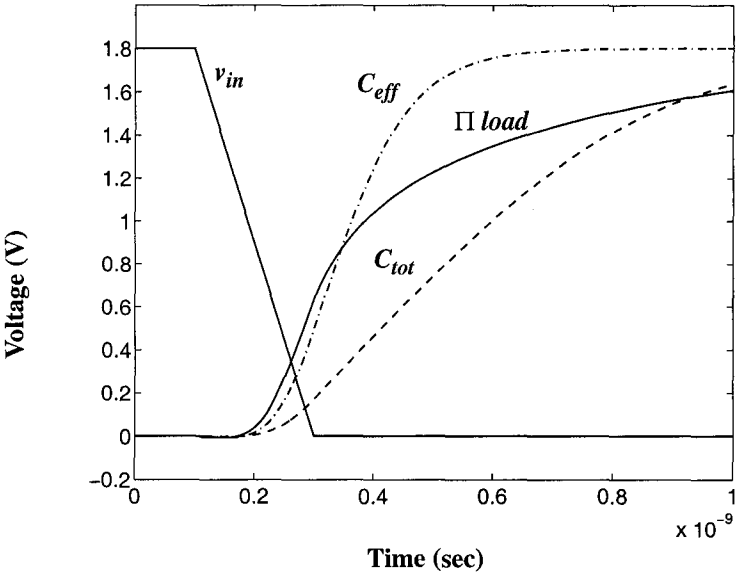



**FIGURE 2.7** Comparison of the output voltage waveform of an inverter loaded by Π load (-), the total load capacitance (--), and by an effective capacitance chosen to capture the 50% delay (-.). This capacitance is determined using HSPICE.

A more "natural" approach to estimating the output waveform of a CMOS gate is to use the effective capacitance to capture only the portion of the waveform during which the charging transistor is saturated, where it behaves, to a good approximation, as a current source [44][45]. Then, the remaining portion (tail) can be derived analytically, based on the analysis of the equivalent linear circuit in which the charging transistor can be accurately modeled as a linear resistor (Figure 2.8). The method in [45], in particular, builds on existing fast and accurate models, such as [53], for computing the peak supply current, its time of occurrence $t_m$, and delay of a CMOS inverter driving a capacitive load, thereby allowing and accurate estimation of the time at which the charging transistor leaves saturation and enters triode[1]. This method is adopted in this thesis.

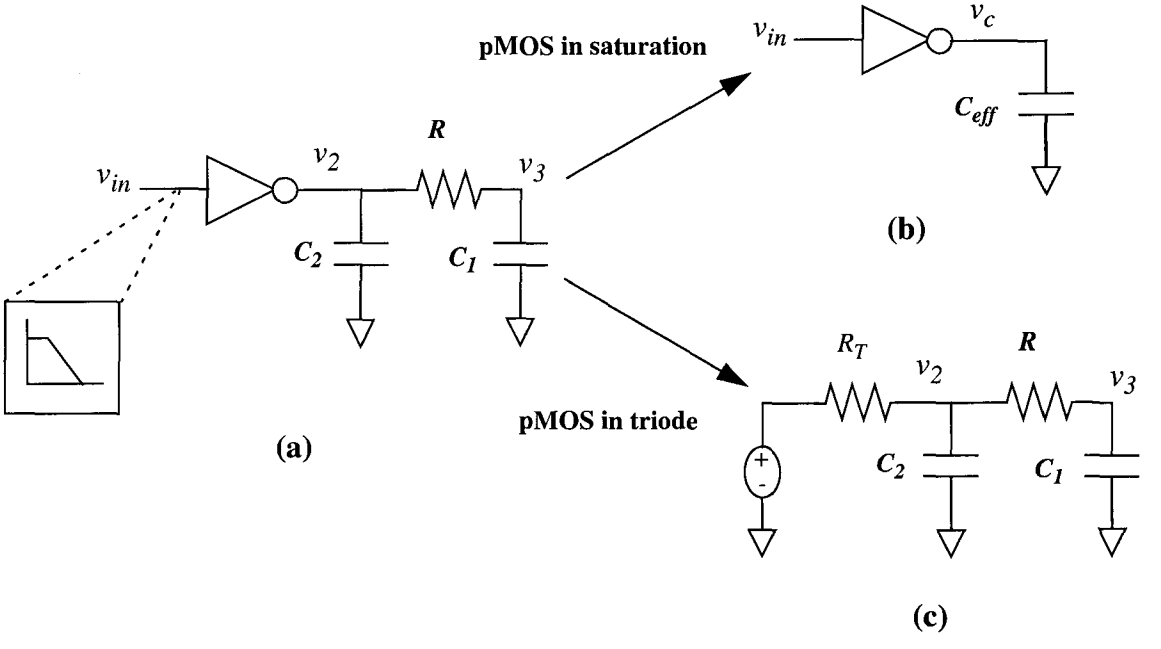



**FIGURE 2.8** (a) Inverter driving a Π RC load. (b) A lumped $C_{eff}$ to capture the early portion of the inverter's output voltage ($v_2$) waveform. (c) Equivalent linear circuit used to derive the tail portion of the inverter output voltage waveform. The charging transistor is modeled as a linear resistor.

1. In fact $t_m$ is used as good approximation of the time at which the charging transistor leaves saturation ($t_{st}$), as will be seen in the next chapter.

In this approach [45], the effective capacitance is also chosen in such a way that the average current that is drawn by the Π load model (or the total charge transferred) of the driving point admittance of the interconnect is equal to the average current drawn by $C_{eff}$ over the time interval of interest, namely between $t = 0$ and $t = t_m$. Therefrom:

$$\frac{1}{t_m}\int_0^{t_m} C_{eff}\frac{dv_c}{dt}dt = \frac{1}{t_m}\int_0^{t_m} C_2\frac{dv_2}{dt}dt + \frac{1}{t_m}\int_0^{t_m} C_1\frac{dv_3}{dt}dt \tag{2.30}$$

and therefore:

$$C_{eff} = C_2 + \frac{v_3(t_m)}{v_c(t_m)}C_1 \tag{2.31}$$

For practical reasons, the authors in [45] first considered the computation of the effective capacitance $C'_{eff}$ of the $RC_1$ Γ load, with the same conditions of input signal transition and transistors. Then, based on the linearity of the Π model and the fact that a transistor in saturation can be approximated by an ideal current source, the effective capacitance of the $C_2RC_1$ Π load can be computed by simple superposition, i.e:

$$C_{eff} = C_2 + C'_{eff} \tag{2.32}$$

$C'_{eff}$ is obviously computed by equating the total charge transferred by $C_1$ and $C'_{eff}$ over the time when the charging transistor is in saturation (note that $t_m$ in this case is the one associated with the Γ RC load):

$$\frac{1}{t_m}\int_0^{t_m} C'_{eff}\frac{dv_c}{dt}dt = \frac{1}{t_m}\int_0^{t_m} C_1\frac{dv_3}{dt}dt \tag{2.33}$$

This yields the following simple expression for the $RC_1$ equivalent capacitance:

$$C'_{eff} = \frac{v_3(t_m)}{v_c(t_m)}C_1 \tag{2.34}$$

As can be noticed, the value of this effective capacitance is actually dependent on the value sought for $v_c(t_m)$ (and therefore $v_3(t_m)$ which is dependent on the former). However, it is shown that by precharacterizing[1] the voltage waveform $v_3(t)$ as a function of the inverter's (driver's) input transition time and the $RC_1$ load, the effective capacitance can be accurately determined within an iterative procedure, which can be summarized as follows:

1. Chose an initial value for $C'_{eff}$.
2. Use an inverter model to compute $t_m$ and $v_c(t_m)$ with $C'_{eff}$ as the load capacitance.
3. Compute $v_3(t_m)$ (as precharacterized), and use equation (2.34) to compute the new value of $C'_{eff}$.
4. If the new $C'_{eff}$ differs from the old one by more than some specific error, then set $C'_{eff}$ to the value and go back to 2. Otherwise the iteration is terminated.

In addition, provided that the resistance of the load is small compared to the output resistance of the saturated pMOS transistor, the value of the effective capacitance is proven to be independent of the size of the driving inverter [45]. Therefore, using a reference inverter[2] allows the characterization of $v_3(t)$ only for this inverter, rendering the aforementioned procedure quite appealing. Further details of the computation procedure of the effective capacitance and its potential limitations can be found in [45].

For the case of a rising input signal, a similar analysis yields the following expression of $C'_{eff}$:

$$C'_{eff} = \frac{V_{dd} - v_3(t_m)}{V_{dd} - v_c(t_m)} C_1 \tag{2.35}$$

where $V_{dd}$ is the value of supply voltage.

---

1. $v_3(t_m)$ can be approximated by a linear ramp, whose slope and initial time can be formulated as a function of the inverter's input transition time and values of $R$ and $C_1$.
2. The independence of the effective capacitance from the driver size is affirmed as long as the driver's output resistance is larger than the load resistance. In situations where these resistances become comparable, which occur in DSM technologies, additional reference inverters might be required [45].

## 2.5 RC Interconnect Delay Metric

Over the years, the Elmore constant has been the delay metric "par excellence". Its simplicity resulted in fast computation speed, and allowed an intuitive understanding of the behavior of interconnect circuits, opening the way for a large number of interconnect optimization techniques. However, for the reasons mentioned earlier, this metric has become inappropriate[1] for timing analysis and optimization in DSM technologies. As a consequence, a number of interconnect delay metrics have been proposed. The goal has obviously been to identify an accurate metric that has, ideally, the simplicity of the Elmore delay, and more importantly, its intuitiveness.

Most of the proposed techniques are based on the assumption that matching the first three moments of the impulse response results in a circuit that can accurately describe the electrical behavior of the linear RC circuit at hand. In [46], for instance, Tutuianu *et al* proposed an "explicit" RC circuit delay approximation using the three first moments. These moments were used to compute a pair of stable dominant poles and their residues. Any delay point was then computed using a single Newton-Raphson iteration, where the delay considering only the most dominant pole was used as the initial guess. Later, in [47] Pileggi defined a new delay metric, where the impulse response at a node is compared to a probability distribution function[2]. A highly accurate distribution for this purpose was identified as the *h*-gamma distribution. The main problem with these techniques is that they result in somewhat complicated parametrized expressions that require large two dimensional look-up tables.

Another class of metrics is that proposed by Kahng and Muddu [48]. In fact the authors proposed two methods worth mentioning. The first method consists of simply choosing the dominant pole among the two poles computed using the three first moments of the impulse response. The second method consist of computing the pole, or the single pole circuit, that results in the same 3-dB attenuation as the circuit synthesized using the

1. Note that Elmore delay is still used as an effective means to validate timing optimization techniques because of its fidelity. That is a technique that is valid under Elmore delay is likely to be under more accurate delay metrics.
2. This was originally suggested by Elmore [33].

three first moments. The problem here, is that the authors tried to model a two-time constant waveform with a single time constant description, irrespective of the importance of each pole with respect to the other, or their ratio for instance. This is simply unreasonable.

In this section we propose a simple RC interconnect delay metric that is not only extremely accurate, but also nearly as simple as the Elmore delay. For this, consider the simple circuit in Figure 2.9. This circuit can be viewed as constructed from the three first moments of the impulse response at node *out* of a more complicated linear RC circuit. The transfer function of such a circuit can be easily determined:

$$H(s) = \frac{1}{R_1R_2C_1C_2s^2 + [R_1C_1 + (R_1 + R_2)C_2]s + 1} \tag{2.36}$$

Now let us write this transfer function in terms of the moments of the circuit as defined in the AWE technique (equation (2.19)):

$$H(s) = \frac{1}{(m_1^2 - m_2)s^2 - m_1 s + 1} \tag{2.37}$$

Note that the first two moments are sufficient to describe the transfer function of such a circuit. Therefore, a delay metric that takes into account only these two moments is perfectly justified. The poles of this circuit are therefore:

$$p_{1,2} = \frac{m_1 \pm \sqrt{4m_2 - 3m_1^2}}{2(m_1^2 - m_2)} \tag{2.38}$$

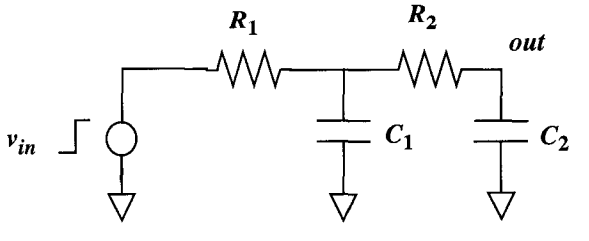



**FIGURE 2.9** Two poles reduced-order RC circuit.

Let us examine the two extreme cases. First, when we are in the presence of a really dominant pole, such that the second pole can be ignored altogether, i.e $p_1 \gg p_2$ (both are real negative), the step response at node *out* is simply:

$$v_{out} = 1 - e^{p_1 t} \tag{2.39}$$

and the 50% delay is therefore

$$t_{D0.5} = -\frac{1}{p_1}\ln 2 = -m_1 \ln 2 \tag{2.40}$$

Note that, in this particular case: $m_2 = m_1^2$ so (2.40) can also be expressed as :

$$t_{D0.5} = \frac{m_1^2}{\sqrt{m_2}}\ln 2 \tag{2.41}$$

The second extreme case is the one where $p_1 = p_2$. This occurs when the discriminator in equation (2.38) is nil. Therefore, in this case we have:

$$4m_2 = 3m_1^2 \tag{2.42}$$

Hence the double pole $p$, in terms of the first two moments, can be expressed as:

$$p = -\frac{4}{\sqrt{3}} \cdot \frac{\sqrt{m_2}}{m_1^2} \tag{2.43}$$

In this case the step response of the circuit (at node *out*) is:

$$v_{out}(t) = 1 - e^{pt} - pte^{pt} \tag{2.44}$$

Here, unlike [48], we do not try to define an equivalent pole that captures, approximately, the behavior of the entire waveform described in equation (2.44), from which the desired delay point would be computed. Rather, for each delay point of interest we compute the pole that results in exactly that same point, such as illustrated in Figure 2.10. The reason is that a two time constant behavior can not be approximated using a single pole unless one of the time constants is negligible with respect to the other.

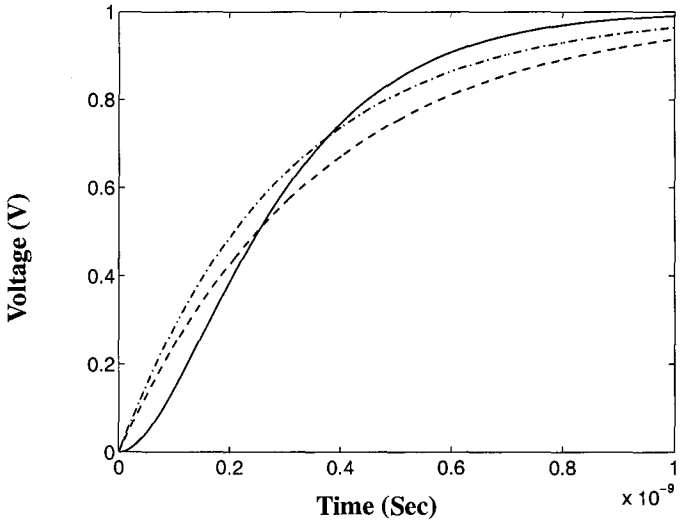



**FIGURE 2.10** Single time-constant-waveform capturing the 50% delay point (--) and the 70% delay point (-.) of a two-time-constant waveform (-).

Expanding the expression in equation (2.44) in terms of Taylor series about $t$=0, results in the following expression:

$$v_{out}(t) = \sum_{i=1}^{\infty} \frac{i}{(i+1)!}(pt)^i \tag{2.45}$$

As shown in Figure 2.11, an expansion to eighth order is necessary to capture the 50% delay point of $v_{out}$ with the highest accuracy. Now, since we are seeking an equivalent pole that captures the 50% point delay of $v_{out}$, we can write:

$$\sum_{i=1}^{8} \frac{i}{(i+1)!}\left(p\left(-\frac{1}{p_{eq}}\ln 2\right)\right)^i = 0.5 \tag{2.46}$$

Solving this equation numerically results in the following value of the ratio $p/p_{eq}$, where $p_{eq}$ is the pole of the first order RC circuit that captures 50% delay point of the second order circuit of interest.

$$\frac{p_{eq}}{p} \approx 0.413 \Rightarrow p_{eq} \approx -1.05 \cdot \frac{\sqrt{m_2}}{m_1^2} \tag{2.47}$$

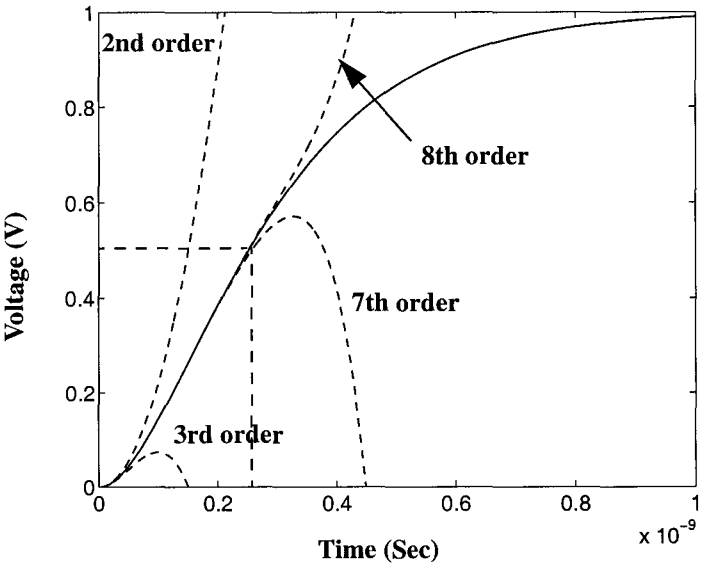



**FIGURE 2.11** The required Taylor's expansion order to capture the midpoint transition .

From here, the 50% delay of the second extreme case can be expressed as:

$$t_{D0.5} = 0.95 \cdot \frac{m_1^2}{\sqrt{m_2}} \ln 2 \tag{2.48}$$

As can be seen from (2.41) and (2.48), the delay expressions of the 50% delay for the two extreme cases differ only by a factor of 0.95. Hence, it is fair to assume that in a general case, for each ratio $m_1^2/\sqrt{m_2}$, there is a factor $\delta$ such that the 50% delay point of the step response of a 2-pole circuit can be expressed in the form:

$$t_{D0.5} = \delta \cdot \frac{m_1^2}{\sqrt{m_2}} \ln 2 \tag{2.49}$$

where

$$0.95 \leq \delta \leq 1 \tag{2.50}$$

Now, let us consider the general case of a system with two distinct poles. The general form of the step response of such a system is given by:

$$v_{out} = 1 - k_1 e^{p_1 t} - k_2 e^{p_2 t} \tag{2.51}$$

where

$$\begin{aligned} k_1 &= \frac{p_2}{p_2 - p_1} \\ k_2 &= \frac{-p_1}{p_2 - p_1} \end{aligned} \tag{2.52}$$

Now, in the general case one can assume that:

$$m_2 = f m_1^2 \tag{2.53}$$

Recall that in the case where one of the two poles totally dominates the behavior of the step response, $f$=1. On the other hand, if the two poles are coincident, $f$=3/4. Hence, in the general case we can assume that we always have:

$$\frac{3}{4} \le f \le 1 \tag{2.54}$$

Therefore, from equation (2.53)and (2.38), we can write

$$\begin{aligned} p_1 &= \rho_1 \cdot \frac{\sqrt{m_2}}{m_1^2} \\ p_2 &= \rho_2 \cdot \frac{\sqrt{m_2}}{m_1^2} \end{aligned} \tag{2.55}$$

where:

$$\begin{aligned} \rho_1 &= \frac{1 + \sqrt{4f - 3}}{2\sqrt{f}(1 - f)} \\ \rho_2 &= \frac{1 - \sqrt{4f - 3}}{2\sqrt{f}(1 - f)} \end{aligned} \tag{2.56}$$

Finally, expressing $p_1$ as a function of $p_2$:

$$p_1 = \frac{\rho_1}{\rho_2} p_2 = \rho p_2 \tag{2.57}$$

where $\rho = \rho_1 / \rho_2$

By replacing $k_1$ and $k_2$ in equation (2.51) by their expressions in equation (2.52), and $p_1$ by its expression (equation (2.57)) and then expanding $v_{out}$ in terms of Taylor series about $t$=0 results in:

$$v_{out}(t) = 1 + \sum_{i=1}^{8} \left( \frac{\rho^i - \rho}{1 - \rho} \right) \frac{(p_2 t)^i}{i!} \tag{2.58}$$

In order to determine the 50% delay point of $v_{out}$, the same procedure as before is applied. That is, we determine an equivalent single pole ($p_{eq}$) system that results in the same delay point. Therefore, if the 50% delay point is sought, we write:

$$0.5 = 1 + \sum_{i=1}^{8} \left( \frac{\rho^i - \rho}{1 - \rho} \right) \frac{(p_2 t_{D0.5})^i}{i!} = \sum_{i=1}^{8} \left( \frac{\rho^i - \rho}{1 - \rho} \right) \frac{\left( p_2 \left( -\frac{1}{p_{eq}^{(0.5)}} \ln 2 \right) \right)^i}{i!} \tag{2.59}$$

It is clear that for any given value of $\rho$, or $f$, the ratio $p_2 / p_{eq}$ can be computed numerically. Therefore, if the value of this ratio is $\sigma$ then:

$$p_{eq}^{(0.5)} = \frac{1}{\sigma} p_2 = \frac{\rho_2}{\sigma} \cdot \frac{\sqrt{m_2}}{m_1^2} = \delta^{(0.5)-1} \cdot \frac{\sqrt{m_2}}{m_1^2} = \frac{\ln 2}{t_{D0.5}} \tag{2.60}$$

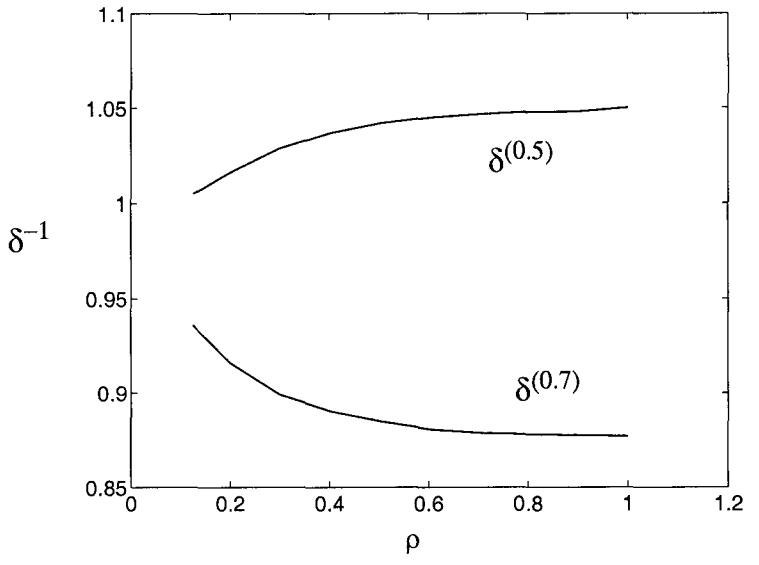



**FIGURE 2.12** $\delta$ as an implicit function of the first two moments ratio.

In Figure 2.12, we plot the coefficient $\delta^{(0.5)}$ and $\delta^{(0.7)}$ corresponding to the 50% and 70% delay points, respectively. As will be explained in the next chapter, these two point in particular are of special interest to us. Hence, from (2.60):

$$t_{D0.5} = \delta^{(0.5)} \frac{m_1^2}{\sqrt{m_2}} \ln 2 \tag{2.61}$$

By a similar procedure, the expression of the 70% delay point of the 2-pole RC circuit can be determined by computing the equivalent single pole system that results in the same 70% delay point. We find:

$$t_{D0.7} = \delta^{(0.7)} \frac{m_1^2}{\sqrt{m_2}} \ln 3.33 \tag{2.62}$$

Finally, by polynomialy fitting the curves in Figure 2.12 we end up with the following simple expressions:

$$\begin{aligned} \delta^{(0.5)} &= a_4\rho^4 + a_3\rho^3 + a_2\rho^2 + a_1\rho + a_0 \\ \delta^{(0.7)} &= b_4\rho^4 + b_3\rho^3 + b_2\rho^2 + b_1\rho + b_0 \end{aligned} \tag{2.63}$$

where

**TABLE 2.1** $\delta$ Fitting parameters

| $\delta^{(0.5)}$ | | $\delta^{(0.7)}$ | |
|---|---|---|---|
| a4 | 0.0039 | b4 | 0.2877 |
| a3 | 0.1132 | b3 | -0.853 |
| a2 | -0.286 | b2 | 0.964 |
| a1 | 0.2400 | b1 | 0.508 |
| a0 | 0.9791 | b0 | 0.985 |

This Metric, has very recently been reported in [49]. Our work was carried out independently from [49]. In fact, the authors of [49] provided no justification or derivation of the metric in question. In their case, it seemed purely empirical.

In the case of a ramp input of duration $\tau$, the Laplace transform of $v_{out}$ can be written as:

$$V_{ramp}(s) = H(s)\left(\frac{1}{s^2\tau}(1-e^{-s\tau})\right) = \sum_{i=0}^{\infty} m_i s^i \cdot \left(\frac{1}{s^2\tau}(1-e^{-s\tau})\right) \tag{2.63}$$

By expanding the exponential term in terms of Taylor series about $s$=0 and multiplying the resulting terms with the circuit's moments, we end up with:

$$V_{ramp}(s) = \frac{1}{s}\left(1 + \left(m_1 - \frac{\tau}{2}\right)s + \left(\frac{\tau^2}{6} + \frac{m_1\tau}{2} + m_2\right)s^2 + \ldots\right) = \frac{1}{s}H_{eq}(s) \tag{2.64}$$

As can be seen, the delay metric for a step input can be used in the case of a ramp input by replacing the moments $m_1$ with $m_1$-$\tau$/2 and $m_2$ with $\tau^2/6+m_1\tau/2+m_2$ [50].

The accuracy of the delay metric developed in this section is shown in Figure 2.13. Note that this metric can not be used to determine delay points at node *c*, in general, since the transfer function at this particular node may comprise a low frequency zero. That is, the transfer function form in equation (2.36) will be inappropriate in this case, unless of course, such a zero is a high frequency one compared with the circuit's poles.

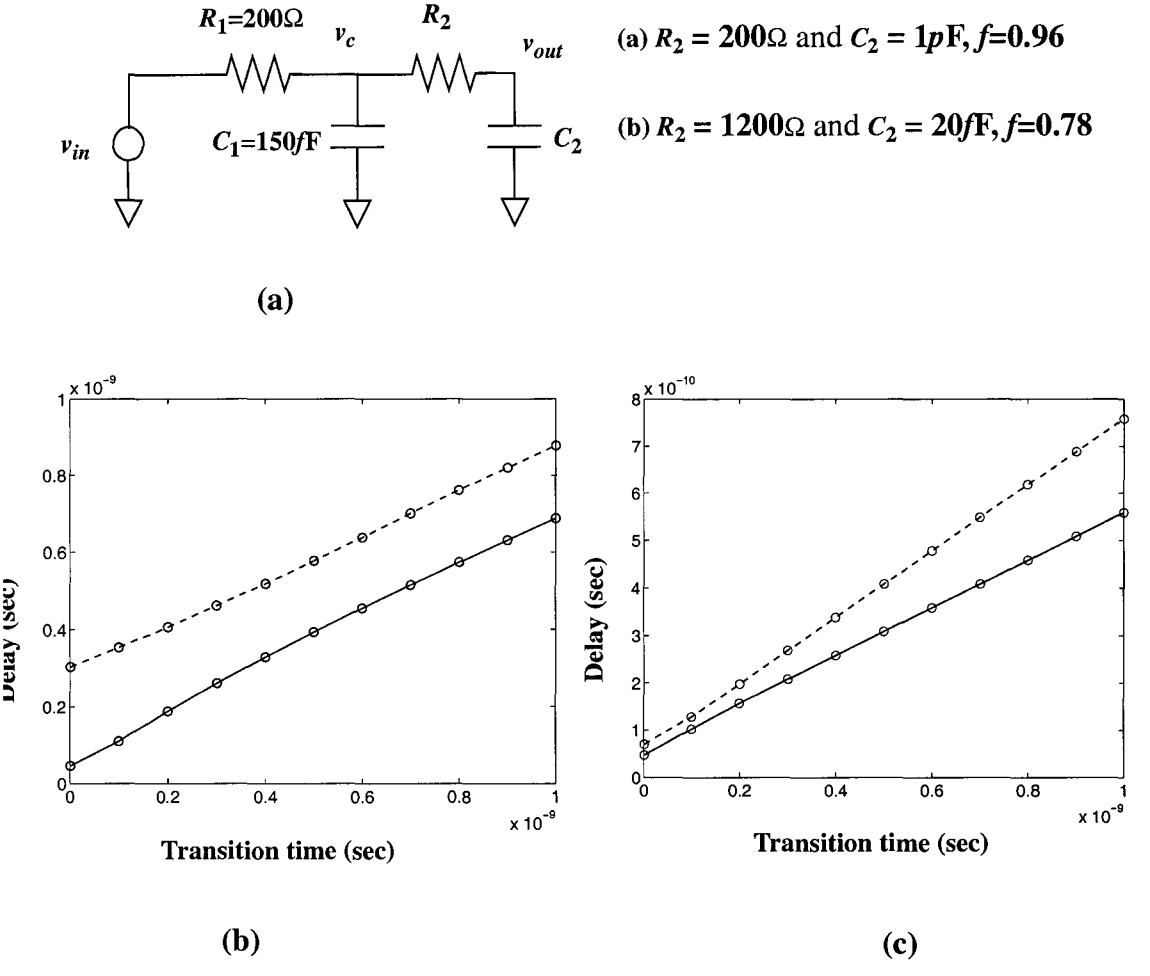



**FIGURE 2.13 (a)** Two pole reduced-order linear circuit. The 50% (-) and 70% (--) delay points at the far end of the circuit (node *out*) for **(b)** $f$=0.96 and **(c)** $f$=0.78. The circles are the delays computed using HSPICE.

# Chapter 3

# CMOS Inverter Delay and Current Model

In this chapter, we introduce the current and delay model of the CMOS inverter used as the repeater element in signal distribution networks. First, we describe the operation of the CMOS inverter based on the description of the various parasitic capacitances that are accounted for in the computation of the delay and the various currents (short-circuit and discharging/charging currents). Then, an accurate and fast technique for computing the inverter output waveform is presented under the assumption of a ramp input signal and a purely capacitive load. In section 3.4, the model is extended to incorporate some interconnect effects. In particular, the concept of an effective input ramp signal that accounts for the non-symmetry in interconnect signals is introduced in subsection 3.4.1. The validity of such a concept in the particular context of optimized RC interconnects is also discussed. Throughout this chapter, the emphasis is primarily on the case of a rising input (falling output), the case of a falling input (rising output) being similar.

## 3.1 CMOS Inverter current model

In this section, the general scheme for computing the inverter delay and the charging or discharging current is presented. Only the discharging operation is discussed based on a transistor-level model of the inverter. Figure 3.1 (a) illustrates such a model where all the parasitic capacitances that are accounted for in the computation are shown explicitly.

$C_{GP}$ ($C_{GN}$) accounts for the gate-to-source capacitance $C_{GSP}$ ($C_{GSN}$), and the gate-to-bulk capacitance $C_{GBP}$ ($C_{GBN}$) of the pMOS (nMOS) transistor. In this thesis, the inverter input capacitance, $C_i$, is defined simply as the sum:

$$C_i = C_{GP} + C_{GN} = C_{GBP} + C_{GSP} + C_{GBN} + C_{GSN} \tag{3.1}$$

$C_P$ ($C_N$) includes both the drain-to-bulk capacitance of the pMOS (nMOS) transistor and the effect of the loading capacitance of the fanout gate(s). Therefore the effective total load capacitance at the inverter output is

$$C_L = C_P + C_N = C_{DBP} + C_{DBN} + C_{fanout} \tag{3.2}$$

Finally, $C_M$ is the inverter input-output coupling capacitance, also known as "Miller capacitance". It consists of the gate-to-drain capacitances of both the nMOS and the pMOS transistors.

$$C_M = C_{GDP} + C_{GDN} \tag{3.3}$$

The circuit in Figure 3.1(a) can therefore be reduced to the circuit shown in Figure 3.1 (b).

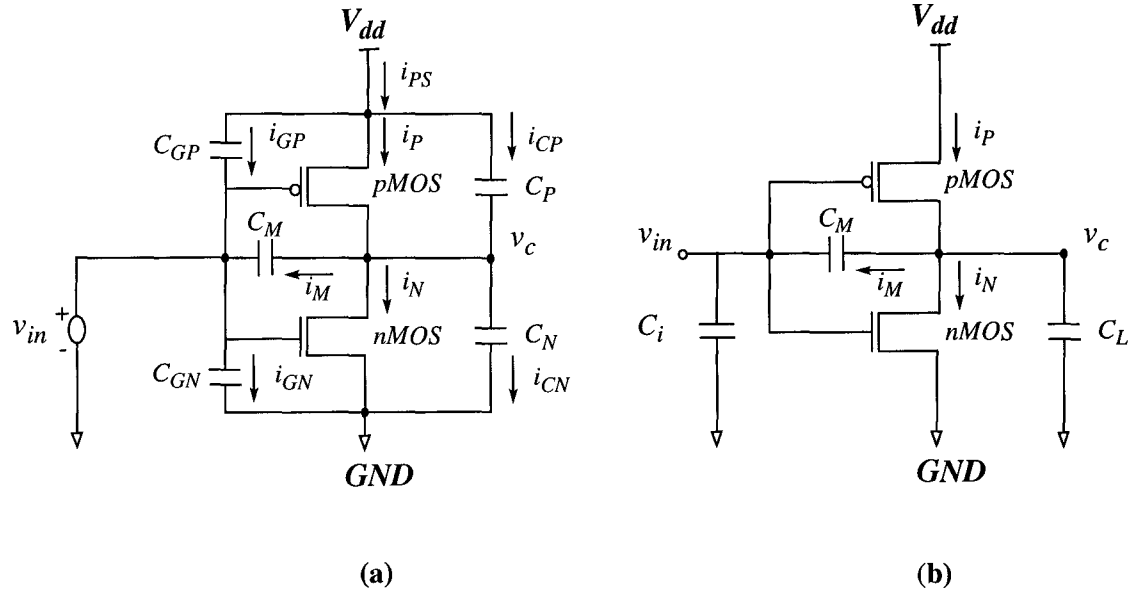



**FIGURE 3.1** **(a)** Transistor-level circuit model of a CMOS inverter, and **(b)** equivalent circuit model.

Note that the gate capacitances $C_{GB}$, $C_{GD}$ and $C_{GS}$ are lumped MOS capacitances that account for the actual non-linear, voltage-dependent components of the gate capacitance and include the overlap capacitancein the case of $C_{GD}$ and $C_{GS}$. $C_{DBP(N)}$ describes the lumped drain-to-bulk p-n junction capacitance. Due to the highly nonlinear and strong voltage-dependence nature of this capacitance, an equivalent constant capacitance is used instead.

If we consider that the input voltage $v_{in}$ is a rising signal, the nMOS transistor is turned on when $v_{in}$ reaches the threshold voltage $V_{tn}$, thereby allowing a discharging path to ground for the load capacitance $C_L$. However, until $v_{in}$ reaches the threshold voltage of the complementary transistor, i.e the pMOS transistor, turning it off, a conducting path persists between the power and ground rails, allowing the flow of the pMOS drain-source current, $i_P$, known as short circuit current, to ground. As a result the discharging of the load capacitance is slower given that the effective current available for the discharge is reduced.

During the input signal rising transition, the Miller capacitance $C_M$ allows a current $i_M$ to flow from the input to the output node, forcing the output voltage above $V_{dd}$. This is known as the feedforward effect.

Figure 3.2 shows HSPICE simulations of both the discharging current ($i_N$) and the short circuit current ($i_P$) of a CMOS inverter driven by a rising CMOS input signal, where the maximas and minimas are labeled *NSC*, *PSC* and *PD*. The *negative short circuit* peak *NSC* at the early stage of the input transition is due to the feedforward effect. That is, forcing the output voltage above $V_{dd}$ results in a pMOS current ($i_P$) flowing from the output node to the power supply rail. Once the input signal reaches the threshold voltage of the nMOS transistor, thereby providing a discharging path to ground, $i_P$ starts to reverse its course to the ground rail. Therefore, one can reasonably assume that *NSC* occurs when the nMOS transistor turns on. The positive *short circuit current PSC*, on the other hand, occurs when the pMOS transistor leaves triode regime and enters saturation. The *positive* (*supply*) *discharging* current peak *PD* occurs before the nMOS transistor enters the triode

region at time $t_{st}$, but no later than the time when the its gate-to-source voltage $V_{GS}$ reaches its maximum, i.e $V_{dd}$ at $T_i$. Therefore, the time of occurrence of $PD$, $t_m$, can be expressed as:

$$t_m = min(t_{st}, T_i) \tag{3.4}$$

In other words, for relatively fast inputs $t_m = T_i$ and for slow inputs $t_m = t_{st}$.

In the following subsections, the aforementioned current maximas and minimas are determined in terms of their time of occurrence.

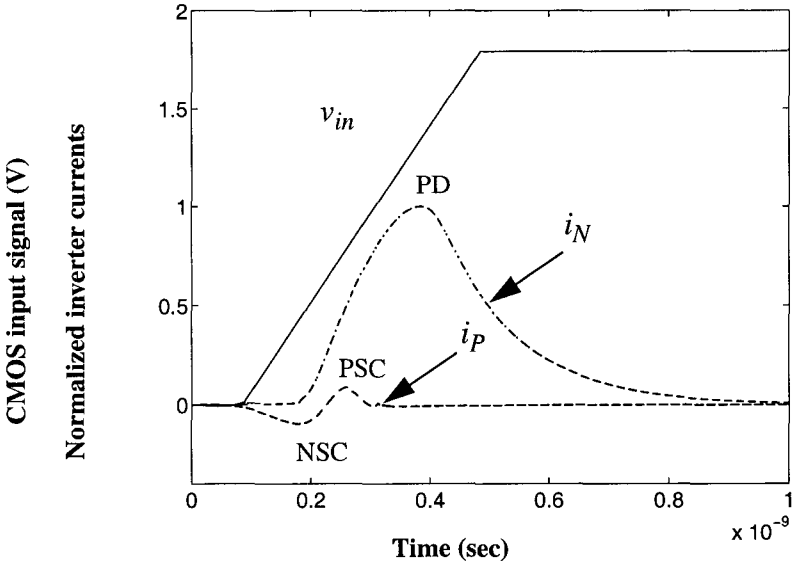



**FIGURE 3.2** Typical Inverter short circuit current (--) and discharging current (-..) waveforms and their respective maximas. The inverter is driven by a typical rising CMOS signal.

### 3.1.1 Maximum Supply Current

Here, we present a very empirical model for computing $t_{st}$ in the context of a symmetrical CMOS inverter and relatively slow inputs. That is $t_{st}$ will only be computed when its value coincides with $t_m$ in the case of slow inputs. The case of fast inputs will be explicitly discussed in section 3.3 where an accurate description of the operation of the CMOS inverter is presented.

Figure 3.3(a) shows the variation of $t_{st}$ as a function of the input transition time for different sizes of inverters. As shown, $t_{st}$ is a linear function of $T_i$, and can therefore be expressed as:

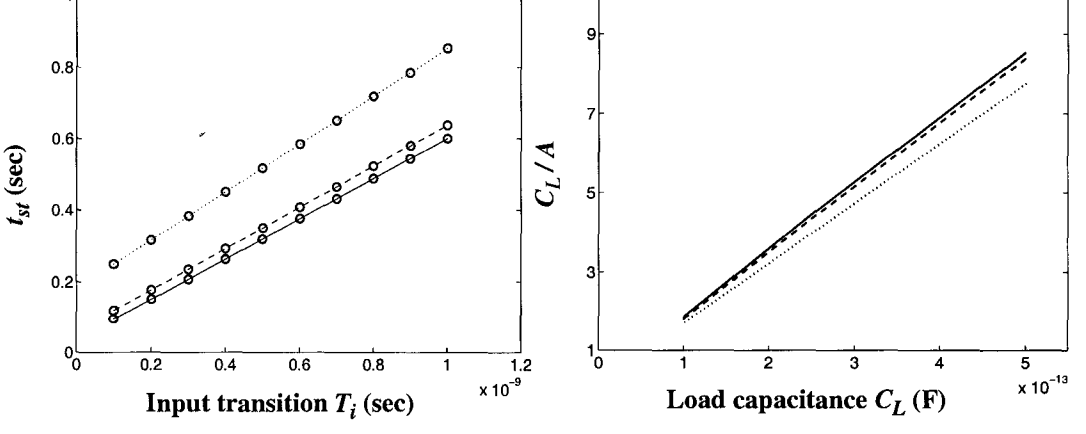
$$t_{st} = AT_i + B \tag{3.5}$$

In order to completely specify $t_{st}$ for any inverter, all that is needed is to express the slope and "y-intercept" terms of equation (3.5), i.e $A$ and $B$ respectively, in terms of the load capacitance and size of the inverter.

By looking at Figure 3.3 (b), it is evident that the proportionality parameter $A$ is a load-dependent parameter that can be expressed as follows:

$$A = \frac{C_L}{D \cdot C_L + C} \tag{3.6}$$

where $C$ and $D$ are empirical technology-dependent parameters which are weakly dependent on the inverter size. As a consequence, each one of these parameters can be sufficiently approximated by a constant value over a specific range of inverter sizes, resulting in the following look up table (in the 0.18μm technology), where **W** is the size




**(a)** **(b)**

**FIGURE 3.3 (a)** The time $t_{st}$ as a function of the input transition time, and **(b)** $C_L/A$ as a function load capacitance, for an inverter with $W_n$ = 30μm (-), 15μm (--) and 5μm (...) The load capacitance in **(a)** is of 250 *f*F. Circles in **(a)** are the actual HSPICE simulated values of $t_{st}$ for each case.

of the nMOS transistor:

**TABLE 3.1.** *C* and D parameters as function of inverter size

| $W_{min}$ / $W_{max}$(μm) | *C(F)* | *D* |
|---|---|---|
| **0.5 / 5** | $1.88\ 10^{-14}$ | 1.483 |
| **5 / 10** | $1.92\ 10^{-14}$ | 1.542 |
| **10 / 20** | $2.05\ 10^{-14}$ | 1.644 |
| **20 / 100** | $2.48\ 10^{-14}$ | 1.665 |

On the other hand, the y-intercept $B$ is a linear function of the load capacitance $C_L$ (Figure 3.4(a)) and is therefore expressed as:

$$B = E \cdot C_L + F \tag{3.7}$$

where $F$ is a technology-dependent parameter, with a value of 20 psec. The proportionality factor $E$ is a driver size-dependent parameter which, by looking at figure Figure 3.4 (b), can be conveniently modeled as follows:

$$E = \frac{w}{G \cdot w + H} \tag{3.8}$$

where $G$ and $H$ are also empirical technology-dependent parameters with values of 0.0075 and -0.0685 respectively. $w$ is the relative size of the actual inverter with respect to the minimum size.

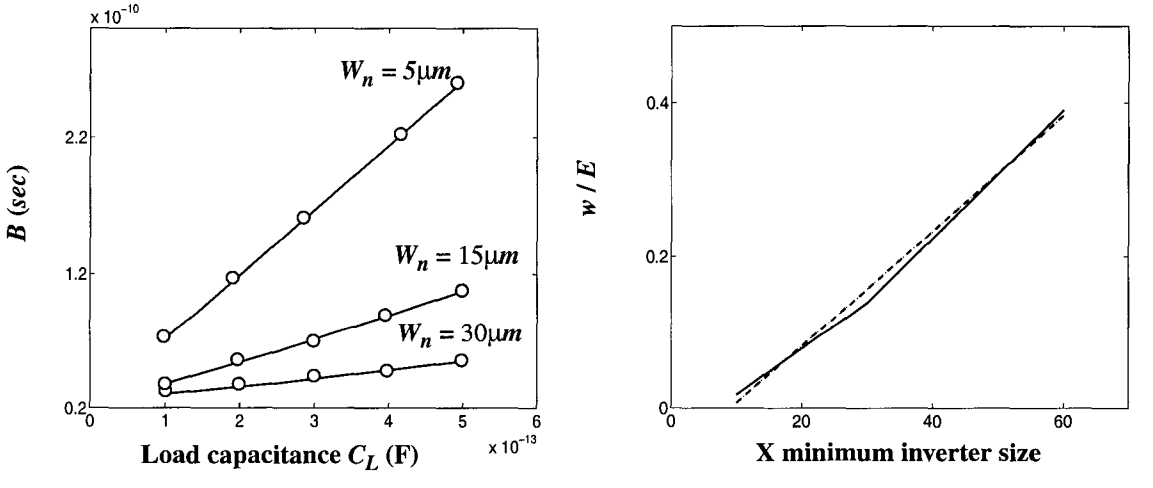



**FIGURE 3.4 (a)** *B* as a linear function of the load capacitance for various inverter sizes. **(b)** *w/E* (-) is sufficiently modeled as having a linear dependence (-..)on inverter sizes. Circles in **(a)** are the actual HSPICE simulated values of *B*.

### 3.1.2 Short Circuit Current

The short circuit current in CMOS inverters has been extensively investigated, given its importance in estimating the power consumption, for instance, as a function of the switching activity in integrated circuits. The reported models for computing the short circuit current are mostly based on the charge conservation principle [42][51]. Although, these models are quite accurate, their inherent complexity does not allow them to be incorporated efficiently in CMOS delay and power models that are to be used in timing optimization routines. In this thesis, we assume a piecewise linear approximation of the short circuit $i_P$ [52] as shown in Figure 3.5. $t_{pmin}$ is defined as the time when the inverter input signal reaches the threshold voltage of the nMOS transistor. $t_{pmax}$ is the time when the bleeding transistor, pMOS, changes its mode of operation from triode to saturation. Finally $t_0$, is the time when the pMOS transistor turns off. While $t_{pmin}$ and $t_0$ are trivially determined, $t_{pmax}$ has yet to be computed.

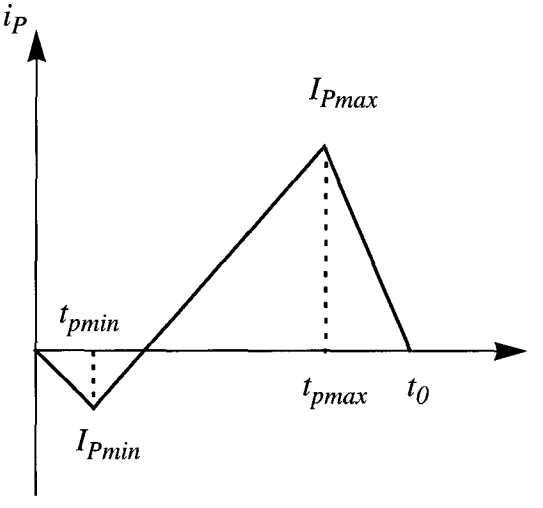



**FIGURE 3.5** Piecewise linear model of the short-circuit current, $i_p$, in the case of a rising input

Figure 3.6 illustrates the variation of $t_{pmax}$ as a function of the input transition time $T_i$ for an inverter driving different capacitive loads (in this case, we used a saturated ramp signal as an input to the inverter). As shown on the same figure, $t_{pmax}$ is linearly proportional to $T_i$ and can be expressed as follows:

$$t_{pmax} - t_{pini} = P \cdot (T_i - T_{ini}) \tag{3.9}$$

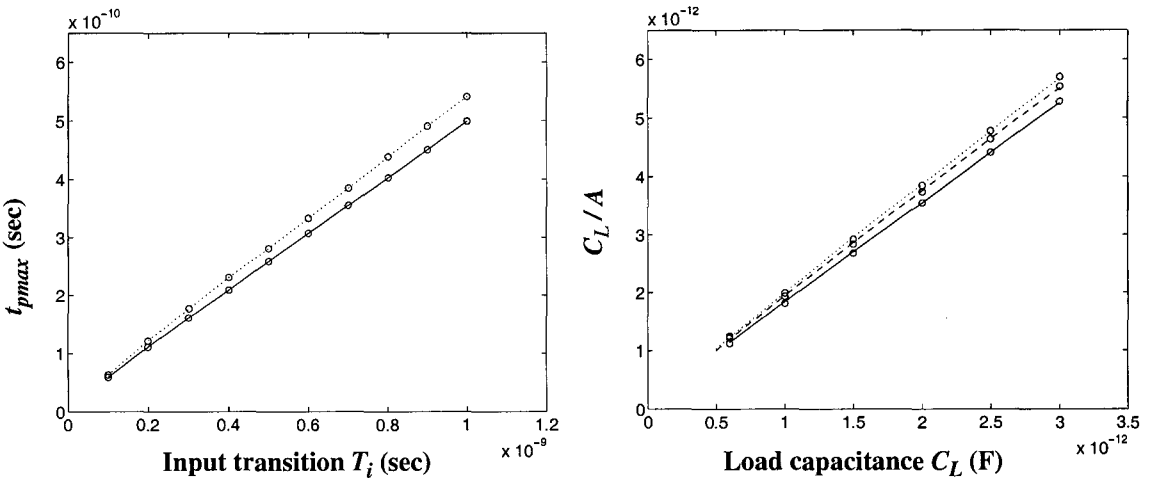



**FIGURE 3.6** **(a)** The time $t_{pmax}$ as a function of the input transition time for an inverter with $W_p$ = 75μm, $W_n$ = 25μm, and $L_n = L_p$ = 0.18μm. $C_L$ = 500*f*F (-) and 3*p*F(...). **(b)** $C_L/A$ as a function of $C_L$ for $W_p/W_n$= 15μm/5μm (-), 45μm/15μm (--), 75μm/25μm (...). Circles are the actual HSPICE simulated values for each case.

where $P$, $t_{pini}$ and $T_{ini}$ are empirical parameters. $t_{pini}$ and $T_{ini}$ are only technology dependent parameters and are equal to 60psec and 100psec respectively for the 0.18μm technology. On the other hand, $P$ depends linearly on the load capacitance and, to a lesser extent, on the inverter size as illustrated in Figure 3.6 (a) and 3.6(b) respectively. Therefore:

$$P = \frac{C_L}{M \cdot (C_L - C_{Lini}) + N} \tag{3.10}$$

where $C_{Lini}$ and $N$ are empirical technology-dependent parameters, with values of 0.5pF and 1pF, respectively. $M$ is an empirical parameter whose value depends on the inverter size used. However this dependence turns out to be quite weak. In fact, very good accuracy in computing $t_{pmax}$ is achieved using just three values of $M$ corresponding to three ranges of inverter sizes as shown in the table below

**TABLE 3.2.** *M* parameter as a function of inverter size

| $W_{min}/W_{max}$(μm) | *M* |
|---|---|
| **0.5 / 5** | 0.481 |
| **5 / 20** | 0.502 |
| **20 / 100** | 0.519 |

## 3.2 n-th. Power Law MOSFET Model

An accurate empirical model for the MOS transistors, namely the Sakurai-Newton n-th. power law model [26], is used here to accurately model the delay and the output voltage waveform of a CMOS inverter. This model has two main advantages over "exact" transistor models, such as the BSIM (Berkeley Short channel IGFET Model) implemented in circuit simulators such as HSPICE. First, it involves a smaller number of parameters to describe the operation of MOS transistors, resulting in analysis techniques that accurately predict circuit behavior at speeds of two to three orders of magnitude faster than HSPICE. Second, given the reduced number of parameters, it allows an intuitive understanding of the electrical behavior of the device from such parameters [53].

The Sakurai-Newton power law model is, on the other hand, more accurate than the traditional Shichman-Hodges model, in that it accounts for various short channel effects such as velocity saturation. The model can be formulated as shown below:

- $V_{DS} \leq V_{Dsat}$ (Triode region):

$$I_{DS} = I_{Dsat}\left(2\frac{V_{ds}}{V_{Dsat}} - \frac{V_{ds}^2}{V_{Dsat}^2}\right)(1 + \lambda V_{DS}) \tag{3.11}$$

- $V_{DS} \geq V_{Dsat}$ (saturation region):

$$I_{DS} = I_{Dsat}(1 + \lambda V_{DS}) \tag{3.12}$$

where

$$I_{Dsat} = \frac{W_{eff}}{L_{eff}}\mathcal{B}(V_{GS} - V_t)^n \tag{3.13}$$

$$V_{Dsat} = \mathcal{K}(V_{GS} - V_t)^m \tag{3.14}$$

$L_{eff}$ and $W_{eff}$ are the effective channel length and width, respectively. $V_t$[1] is the effective MOSFET threshold voltage, and $\lambda$ models the channel length modulation effect. $m$, $n$, $\mathcal{K}$ and $\mathcal{B}$ are empirical technology-dependent parameters used to model the transistor short channel effects.

It is a well known fact, that in short channel transistors, the threshold voltage is a decreasing function of the drain-to-source voltage as a result of the DIBL effect (Drain Induced Barrier Lowering). Moreover, the threshold voltage is also a function of the dimensions of the transistor gate, following the short-length and the narrow-width effects. These effects, and others inherent to deep submicron technologies, are implicitly taken into account by optimizing the set of parameters ($V_t$, $m$, $n$, $\mathcal{K}$, $\mathcal{B}$, $\lambda$)[2] over a specific range of transistor widths. As it turns out, the operation of the MOSFET can be fully described over the whole range of transistor widths (0.5μm up to 100μm) using only three sets of parameters. The results of such optimization for a 1.8 *V*, 0.18μm CMOS process are shown in the table below.

**TABLE 3.3.** Optimized parameters for the n-th. power law model for the MOS transistors in the 1.8V 0.18μm CMOS process technology.

| $W_{min}/W_{max}$(μm) | | $V_t(V)$ | $m$ | $n$ | $\mathcal{K}$ | $\mathcal{B}$ | $\lambda$ |
|---|---|---|---|---|---|---|---|
| **0.5/5** | **NMOS** | 0.61 | 0.4463 | 1.069 | 0.58755 | $7.8\text{x}10^{-5}$ | 0.052 |
| | **PMOS** | 0.63 | 0.51 | 1.15 | 0.62 | $2.9\text{x}10^{-5}$ | 0.107 |
| **5/20** | **NMOS** | 0.58 | 0.52 | 1.1 | 0.6 | $6.7\text{x}10^{-5}$ | 0.051 |
| | **PMOS** | 0.6 | 0.53 | 1.18 | 0.625 | $2.6\text{x}10^{-5}$ | 0.11 |
| **20/100** | **NMOS** | 0.5116 | 0.5 | 1.145 | 0.56866 | $5.8\text{x}10^{-5}$ | 0.07 |
| | **PMOS** | 0.53 | 0.5 | 1.145 | 0.628 | $2.5\text{x}10^{-5}$ | 0.113 |

1. Note that $V_t$ in our case is not necessarily the normally defined gate-source voltage corresponding to the onset of strong inversion under the transistor's gate oxide. $V_t$ is optimized along with other parameters to achieve the best fit of the transistor's I-V characteristics.
2. These parameters can be indexed either n for nMOS transistor or p for pMOS transistor

## 3.3 CMOS Inverter Delay Model

In this section, the model for computing the output voltage waveform of a CMOS inverter and the propagation delay, in the case of a pure capacitive load and rising input, is presented. The input $v_{in}$ is assumed to be a saturated ramp signal:

$$v_{in}(t) = \begin{cases} S_i t & 0 \le t \le T_i \\ V_{dd} & t \ge T_i \end{cases} \tag{3.15}$$

where $S_i$ is the slope of the input ramp. Based on Figure 3.1 (b), we write the differential equation that governs the discharging operation of the CMOS inverter:

$$\frac{dv_c}{dt} = -\frac{(i_n - i_p)}{C_L + C_M} + c_m S_i \tag{3.16}$$

where $c_m = \dfrac{C_M}{C_L + C_M}$.

Note that in most cases of clock distribution networks in VLSI systems, where an inverter can be used as a repeater element, the transition time at the output of the inverter is generally shorter than the input transition, if one includes the effect of the resistance of the interconnect, which shields a part of the load capacitance. Generally, in such a case the short circuit current cannot be neglected altogether.

Figure 3.7 illustrates the typical inverter output voltage waveform in these cases. $v_c(t)$ is the actual output voltage and $v_{cn}(t)$ is the output voltage in the case where the short circuit current is neglected. As shown, the short circuit current $i_p$ has a little effect on the global aspect of the inverter output voltage waveform, but rather it sets the initial condition on the output voltage when the load capacitance starts discharging through the nMOS transistor. Thence, neglecting $i_p$ altogether results ultimately in underestimating the propagation delay of the inverter.

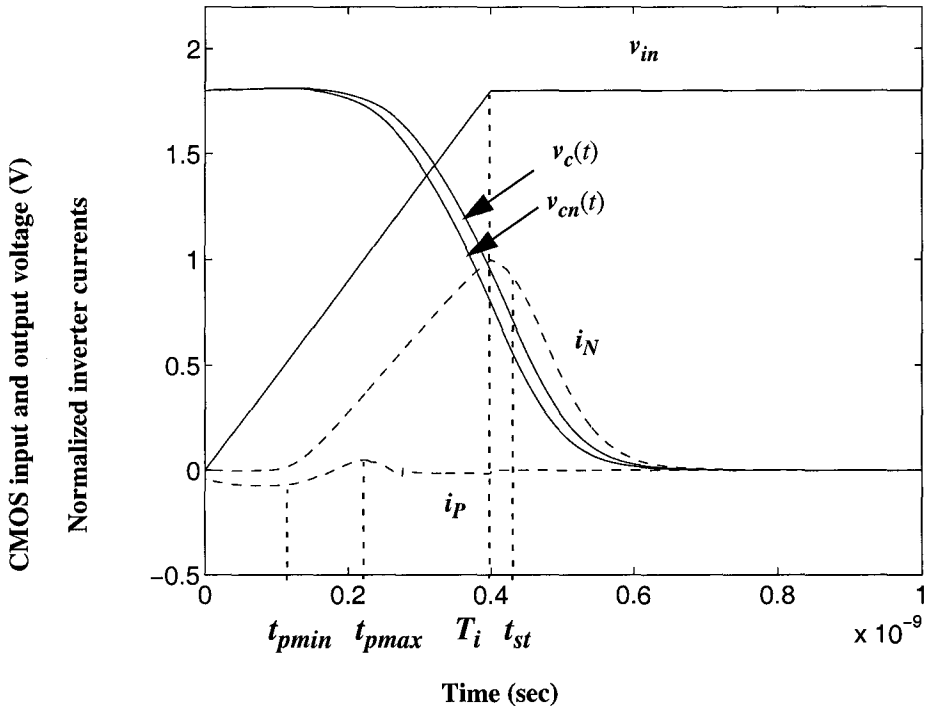



**FIGURE 3.7** Typical inverter output voltage waveform when driven by a ramp input (-) and the associated discharging current $i_n$ and short circuit current $i_p$ (--). $v_{cn}(t)$ is the output voltage $v_c(t)$ obtained by forcing $i_p = 0$.

Let us examine the falling output voltage waveform at different time intervals, corresponding to different modes of operation for both devices, as well as the state of the input signal (rising or saturated).

• $0 \le t \le t_{pmin}$: the pMOS transistor is in triode regime while the nMOS transistor is off. In this region, the inverter output voltage is around $V_{dd}$, and therefore the source-to-drain voltage of the pMOS transistor, $V_{SDp} = V_{dd} - v_c$, is negligible $(1 + \lambda_p V_{SDp} \approx 1)$ as is its quadratic term in equation (3.11). Thus, the short circuit current, $i_p$, can be approximated by

$$i_P = 2\frac{W_p B_p}{L_p K_p}(V_{dd} - V_{tn} - |V_{tp}|)^{n_p - m_p}(V_{dd} - v_c) \tag{3.17}$$

Substituting $i_p$ by its above expression in equation (3.16)and solving for $v_c$ with the initial condition $v_c(0)=V_{dd}$, results in the following expression for the inverter output voltage waveform in this region:

$$v_c(t) = V_{dd} + \frac{c_m}{\beta_p} S_i \left( 1 - e^{\frac{\beta_p}{C_L + C_M} t} \right) \tag{3.18}$$

where $\beta_p = 2\frac{W_p B_p}{L_p K_p}(V_{dd} - V_{tn} - |V_{tp}|)^{n_p - m_p}$.

The minimum short circuit current $I_{pmin}$ is therefore computed using equation (3.11), where $V_{SDp} = V_{dd}\text{-}v_c(t_{pmin})$ and $V_{SGp} = S_i t_{pmin}$.

• $t_{pmin} \le t \le t_{pmax}$: the pMOS transistor is still in triode and the nMOS transistor turns on at $t_{pmin}$ and is in saturation for all the time interval. In this region, the inverter output voltage is in its early stage of falling transition. The drain-to-source voltage of the nMOS transistor $V_{DSn}$ is close to $V_{dd}$, and therefore we can make the approximation: $1 + \lambda_n V_{DSn} = 1 + \lambda_n V_{dd}$. The discharging current, $i_n$ is therefore

$$i_n = \frac{W_n}{L_n} \mathcal{B}_{n1} (S_i t - V_{tn})^{n_n} \tag{3.19}$$

where $\mathcal{B}_{n1} = \mathcal{B}_n(1 + \lambda_n V_{dd})$.

From Figure 3.5 the short circuit current can be expressed as follows:

$$i_p = S_{ip}(t - t_{pmin}) + I_{Pmin} \tag{3.20}$$

where $S_{ip}$ is the slope of the *PWL* approximation of the short circuit current in this region.

$$S_{ip} = \frac{I_{Pmax} - I_{Pmin}}{t_{pmax} - t_{pmin}} \tag{3.21}$$

By substituting the expressions of $i_n$ and $i_p$ in the differential equation (3.16) and solving for $v_c$ with the initial condition $v_c(t_{pmin})$, calculated using equation (3.18), yields the following expression for the inverter output voltage in this region:

$$v_c(t) = v_c(t_{pmin}) + \left(\frac{I_{Pmin}}{C_L + C_M} + c_m S_i\right)(t - t_{pmax}) + \frac{S_{ip}}{2}(t - t_{pmax})^2 + V_{dd} - \frac{W_n \mathcal{B}_{n1}}{((n_n + 1)(C_L + C_M))L_n S_i}(S_i t - V_{Tn})^{n_n + 1} \tag{3.22}$$

At this point, we distinguish between two cases. The first is the one in which $t_{st}$ occurs later than $T_i$. The second is the one in which it occurs earlier. Whether we are in the first situation or the second is trivially established by comparing $t_{st}$ and $T_i$ using the empirical model for computing $t_{st}$ developed in subsection 3.1.1.

1. $t_{st} \geq T_i$:

Since the nMOS is in saturation, from equation (3.14) we can write:

$$v_c(t_{st}) = V_{d0} = \mathcal{K}_n(V_{dd} - V_{tn})^{m_n} \tag{3.23}$$

• $t_{pmax} \leq t \leq T_i$, the short circuit current $i_p$ and the Miller current $i_M$ are ignored altogether. The linear differential equation (3.16) reduces to:

$$\frac{dv_c}{dt} = -\frac{i_n}{C_L + C_M} \tag{3.24}$$

The discharging transistor in this case is still in saturation, and the output voltage is on average very close to $V_{dd}/2$. Hence we can make the approximation in the expression of the drain to source current $1 + \lambda_n V_{DSn} = 1 + \lambda_n V_{dd}/2$. Solving the differential equation (3.24)for $v_c$ with the initial condition being $v_c(t_{pmax})$ calculated using equation (3.22), yields the expression

$$v_c(t) = v_c(t_{pmax}) - \frac{W_n \mathcal{B}_{n2}}{(n_n+1)(C_L+C_M)L_n S_i}(S_i t - V_{tn})^{n_n+1} \tag{3.25}$$

where $\mathcal{B}_{n2} = \mathcal{B}_n(1+\lambda_n V_{dd}/2)$.

• $T_i \le t \le t_{st}$, the input voltage completes its rising transition at $V_{dd}$. Solving equation (3.24) for $v_c$ with $v_c(T_i)$ as the initial condition yield the expression of the output voltage in this region:

$$v_c(t) = v_c(T_i) - \frac{W_n \mathcal{B}_{n2}}{(C_L+C_M)L_n}(V_{dd} - V_{tn})^{n_n}(t - T_i) \tag{3.26}$$

The time when the discharging transistor enters triode, $t_{st}$, is therefore

$$t_{st} = T_i + \frac{v_c(T_i) - V_{d0}}{\frac{W_n \mathcal{B}_{n2}}{(C_L+C_M)L_n}(V_{dd} - V_{tn})^{n_n}} \tag{3.27}$$

Finally, for times later than $t_{st}$ the inverter output voltage is modeled as a decaying exponential. In fact, the pull-down transistor is modeled as a simple resistance as shown in Figure (3.6). Computing the effective switching transistor's resistance is quite straightforward:

$$R_n = c_R \frac{v_c(t_{st})}{i_n(t_{st})} \tag{3.28}$$

In the equation above, $c_R$ is an empirical averaging constant that has a value of 0.85 for the process we are using, i.e the CMOS 0.18μm. This constant is used because in reality the nMOS transistor resistance varies with time as the inverter output completes its transition. Therefore the value of $c_R$ is chosen to obtain the best fit for typical configurations of input transition time, inverter size and capacitive load.

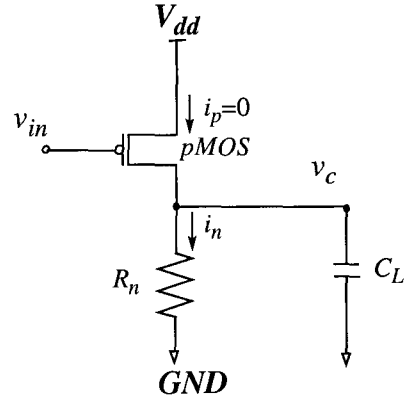



**FIGURE 3.8** Discharging transistor modeled as constant resistance while in triode.

The output voltage can now be expressed as follows

$$v_c(t) = v_c(t_{st})\ e^{-\frac{(t-t_{st})}{\tau_c}} \tag{3.29}$$

where $\tau_c = C_L R_n$.

2. $t_{st} \leq T_i$:

Here $t_{st}$ coincides with $t_m$ and can therefore be expressed using the empirical model of section 3.1.1. Proceeding in a similar manner, within the time interval $[t_{pmax}, t_{st}]$, $v_c$ is expressed using equation (3.25). For times later than $t_{st}$, $v_c$ is a decaying exponential following equation (3.29), where:

$$v_c(t_{st}) = \mathcal{K}_n(S_i t_{st} - V_{tn})^{m_n} \tag{3.30}$$

## 3.4 CMOS Inverter Driving RC Load

In this section, we introduce some RC effects into the inverter delay and current model. In particular, we investigate the effect of the resulting asymmetry in signal waveforms on these models.

### 3.4.1 Ramp Approximation for Signals in Repeated Interconnect lines

For simplicity and efficiency, most inverter delay models use a saturated ramp to approximate gate input voltage waveforms. However, in signal distribution networks, where inverters/buffers drive RC interconnect segments, defining the ramp that best approximates the actual input signal is not straightforward. Ideally, applying such an approximation must yield the same inverter's output voltage and supply current waveforms as in the case where the actual input is applied, thereby allowing an accurate estimate of the actual delay and power dissipation. The problem is even harder for finer feature size technologies, such as the one we are considering here. That is, due to their inherently more pronounced RC effects, interconnect resistance is usually comparable to the gate's output resistance in triode, and even in the saturation regime in some cases. This case occurs especially for large transistors driving a highly resistive load. The higher channel length modulation coefficient for pMOS transistor results in even lower output resistance as opposed to an nMOS transistor of equivalent drive. As a result, voltage waveforms no longer have the symmetry that used to characterize older CMOS technologies.

To fully appreciate the importance of an accurate approximation definition to input signals, consider the system illustrated in Figure 3.9. This system can be viewed as the last two stages of a repeated interconnect line. The first inverter, $inv_1$, drives both a lossy interconnect segment, modeled by a distributed RC load, and the input capacitance of the second inverter, $inv_2$. This inverter in turn, drives an arbitrary capacitive load $C_L$. Here we will attempt to define the best ramp approximation of the voltage waveform at node (1) such that it accurately captures the supply current in $inv_2$ and the voltage transition at node (2).

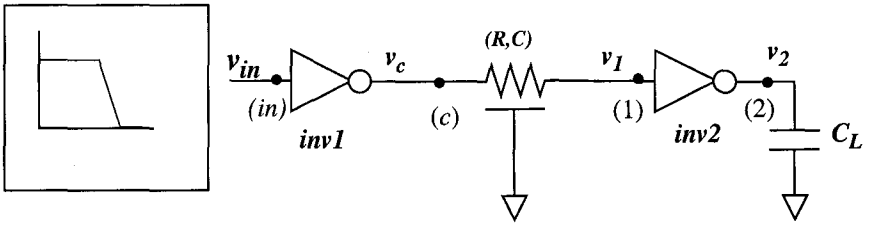



**FIGURE 3.9** Cascaded inverters driving an RC and pure capacitive load respectively. The first inverter is driven by a ramp signal.

Adopting the classic[1] approach for approximating voltage waveforms at inputs to CMOS gates (Figure 3.10) can result in large errors in estimating the inverter's delay and peak supply current. Note the resulting asymmetry of the signal at node (1). As can be seen, such an approach can result in a large divergence between the slowly saturating original signal and its faster approximation. Note that, because almost all the excursion of the inverter's output from high to low occurs during the upper portion of the input signal, the emphasis will be on that region.

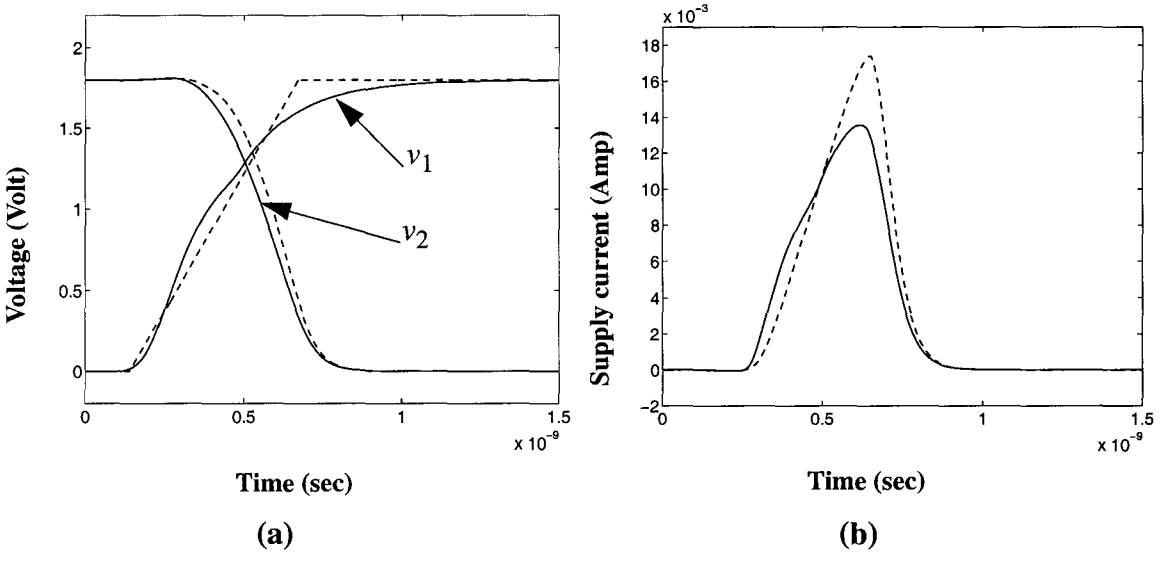



**FIGURE 3.10** Applying a classic definition of the ramp input approximation (--) to a typical RC input signal at v1 (-) results in (a) an underestimate of the delay and (b) overestimate of the supply current peak. Wn=15μm, Wp=45μm, $C_L$=1pF.

1. The classic method for approximating CMOS waveforms has been to use a saturated ramp obtained by passing a straight line through the 20% and 80% points of the actual waveform.

As expected, faster approximations, by providing an "additional drive", can cause underestimates of the delay and, equivalently, overestimates of the peak supply current.

To overcome the obvious inadequacy of the ramp definitions mentioned above, we propose in this thesis a new definition that, not only considers the appropriate transition time of the ramp approximation, $T_i^{eq}$, but also its "equivalent level of saturation" $V_{dd}^{eq}$. As will be shown, in the context of repeated interconnect RC trees, these two quantities can be uniquely correlated with the first moments of the impulse response of the system driving the gate in question (or its dominant poles). This correlation is later expressed through the use of empirical expressions.

The criteria that we applied for determining $V_{dd}^{eq}$ and $T_i^{eq}$, which achieve the best accuracy in the approximation of the RC input signal are both empirical and "objective". The empirical criterion is based on some trials using HSPICE to best define the slope of the input approximation. This trial and error procedure over a wide range of inverter sizes and capacitance loads, $C_L$, resulted in choosing the *PWL* input that passes through the 50% and 70% of the actual waveform, at times $\alpha$ and $\beta$, respectively. Note that, for small values of $C_L$, most of the inverter's output voltage transition occurs within the time frame [$\alpha$, $\beta$]. In such a case, as shown in Figure 3.11 (a), our definition seems to be sufficient for capturing the transition at (2) as well as the supply current pulse of $inv_2$. However, if $C_L$ is large, this definition falls short in providing the same precision (Figure 3.11 (b)). That is, most of the inverter's output voltage transition and supply current pulse occur beyond time $\beta$, coinciding with the "problematic" region where the divergence between the input signal and its approximation is far more pronounced.

By looking at Figure 3.11 (b), the "objective" solution that comes to mind is to saturate the initial ramp approximation at a level below $V_{dd}$. This is equivalent to truncating part of the discharging current, such that the resulting inverter's output voltage would match the one due to the slower exponential tail of the actual input.

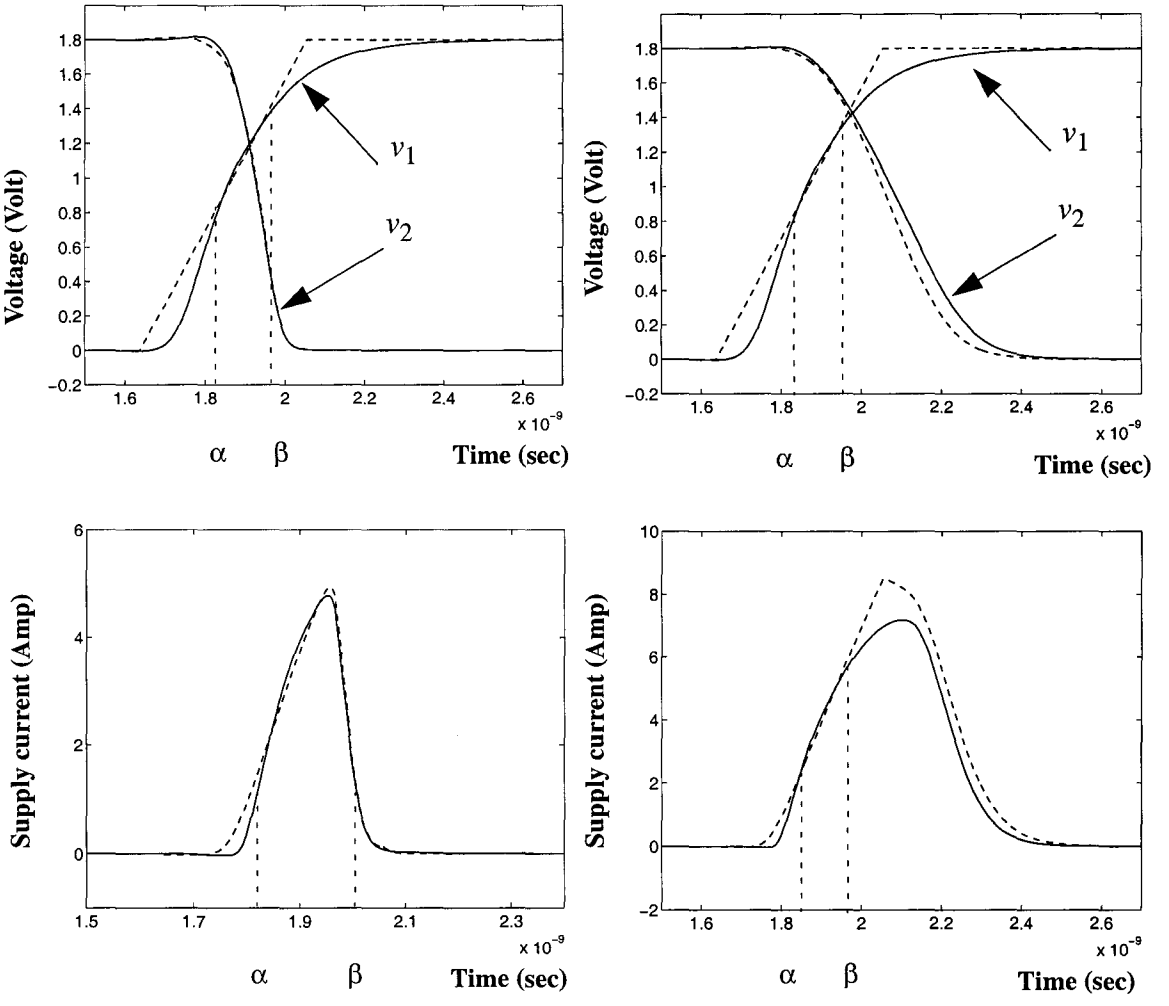



**FIGURE 3.11** comparison of the accuracy obtained in capturing the inverter's output voltage waveform at node (2) and the supply current in $inv_2$ when using (a) a small capacitance, $C_L$=300fF and (b) a large capacitance, $C_L$=1pF. $W_n$=15μm, $W_p$=45μm, $R$=200Ω and $C$=200$fF$. (-) Actual signals, (--) approximations using the saturated ramp, with slope given by the 50% & 70% of $v_1$.

The efficiency of such an approach is shown in Figure 3.12. Note that the validity of our approach is limited to the situations where the inverter's output voltage transition and the corresponding current pulse do not occur in times much later that β. That is, for large loads, the peak supply current and its time of occurrence is largely dependent on the value of $C_L$ and the input's upper exponential tail, and therefore the definition of $V_{dd}^{eq}$ or $T_i^{eq}$ should be tuned to capture the current pulse. But with an understanding of the context

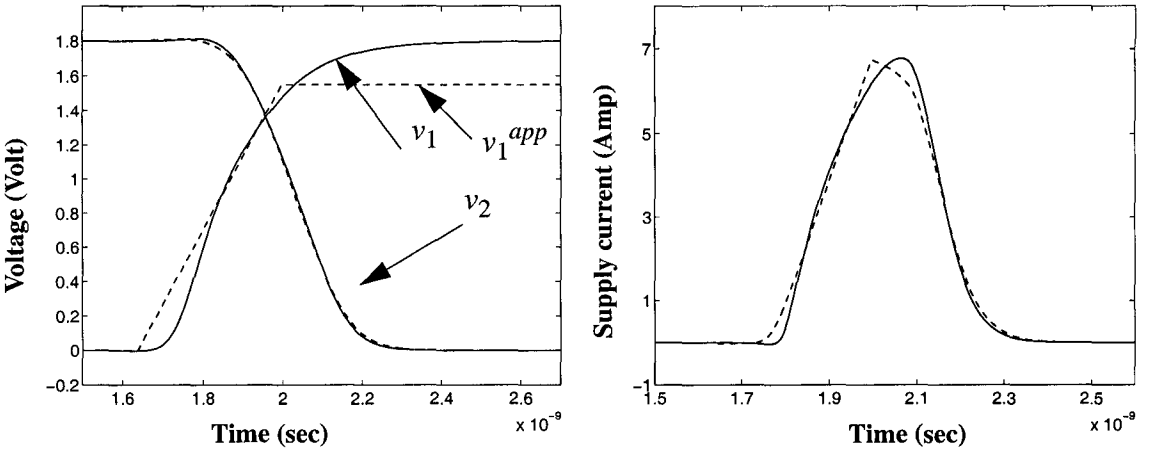



**FIGURE 3.12** Saturating the ramp approximation at a level below Vdd captures with good accuracy the effect of the actual non-symmetrical inverter input waveform on the output transition and the supply current. CL=1pF. Wn=15µm, Wp=45µm, R=200Ω and C=200fF. (-) Actual signals, (--) approximations using the saturated ramp.

in which our modeling scheme is developed, we make the reasonable assumption that, for a given inverter size, there is a limit on the "effective" load capacitance one may consider. In fact, an extensive analysis of practical situations in repeated interconnect lines reveals that this capacitance is usually such that:

$$T_2 \leq 0.9 \cdot T_1 \tag{3.31}$$

where $T_2$ and $T_1$ are the transition times at nodes (2) and (1) respectively. Note that this relation is completely empirical and describes only practical situations in the defined context of repeated RC trees. Also, as will be shown shortly, this relation ensures the uniqueness of the of the definition of both $V_{dd}^{eq}$ and $T_i^{eq}$.

As for the computation of $V_{dd}^{eq}$ and $T_i^{eq}$, we do simply the following. For a number of predefined waveforms, for which we know the exact characteristics (in terms of moments or poles for instance), and given arbitrary reference inverters. $V_{dd}^{eq}$ and $T_i^{eq}$ are extracted using HSPICE while considering a load capacitance that satisfies the relation in equation (3.31).

### 3.4.2 Relating the Ramp Approximation to Moments of the Impulse Response

Once $V_{dd}^{eq}$ and $T_i^{eq}$ are determined, they have yet to be related to a quantity that bears enough information on the transient properties of the actual input signal. In our case, we consider the quantity $\Pi$, representing the area between the actual input signal at node (1) and our defined ramp approximation when it saturates at $V_{dd}$, as illustrated in Figure 3.13. The reason for this choice is that obviously $\Pi$ constitutes an excellent "divergence metric" between the input $v_1$ and its approximation $v_1^{app}$ at their upper portion. Therefore we can write:

$$\Pi = \int_a^b (v_1^{app} - v_1)\ dt \tag{3.32}$$

In Figure 3.14 we plot the resulting extracted $V_{dd}^{eq}$ and $T_i^{eq}$ versus $\Pi$ for a set of input signals with different transient characteristics. As can be seen, the slope of the *PWL* approximation is proportional to the inverse of $\Pi$, and the equivalent input transition is linearly proportional to $\Pi$.

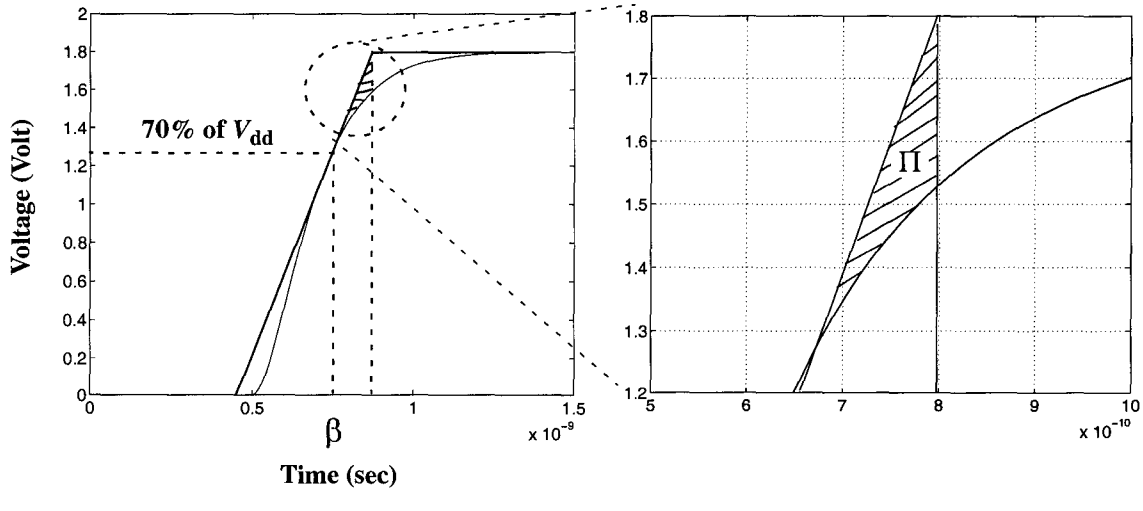



**FIGURE 3.13** The divergence metric between the actual inverter's input signal and its approximation is taken as the area between these two signal from the time β up to the time when the approximation saturates at $V_{dd}$.

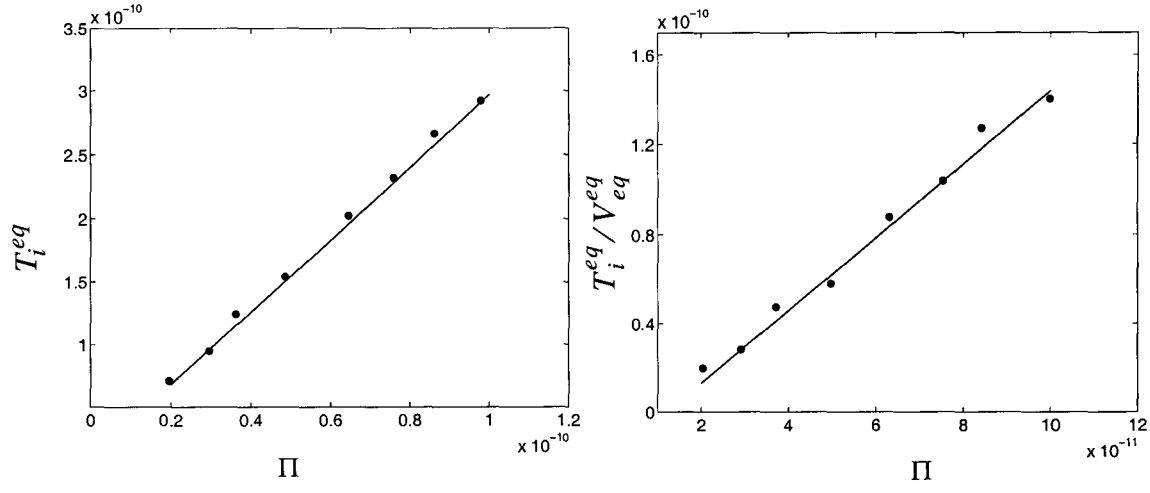



**FIGURE 3.14** Correlation between the equivalent transition, the equivalent level of saturation and the defined divergence metric.

Therefore after curve fitting we can write:

$$\begin{aligned} T_i^{eq} &= K1 \cdot \Pi + K2 \\ T_i^{eq}/V_{dd}^{eq} &= \frac{K1 \cdot \Pi + K2}{K3 \cdot \Pi + K4} \end{aligned} \tag{3.33}$$

where *K1 K2 K3* and *K4* are technology-dependent empirical parameters, whose values are shown in Table 3.4:

**TABLE 3.4.** Fitting parameters K.

| | |
|---|---|
| K1 | 51.51 |
| K2 | -24.5 psec |
| K3 | 29.5 $V^{-1}$ |
| K4 | 1.18 psec.$V^{-1}$ |

Note that, by assuming that the upper half portion of the input signal can be modeled by a single exponential, it can be trivially shown that, following our definition of the ramp approximation:

$$\Pi = 0.705 \cdot \tau \tag{3.34}$$

where $\tau$ is the exponential time constant. As seen in Chapter 2, since the 50% and 70% delay of an RC circuit can be computed accurately, $\tau$ can very well be the time constant of a first order RC circuit that passes through these points. More explicitly:

$$\tau = \frac{t_{70} - t_{50}}{0.336} = \frac{\beta - \alpha}{0.336} \tag{3.35}$$

Recall that, as seen in Chapter 2, $t_{50}$ and $t_{70}$ are directly related to the two first moments of the system that generates the actual input signal.Therefore, by replacing $\Pi$ in equation (3.33) by its expressions given by equation (3.34) and (3.35) we get an explicit relationship between $V_{dd}^{eq}$, $T_i^{eq}$ and the first moments of the impulse response of the system that generates the signal waveform at node (1).

At this point, what remains to be done is to further justify, always in the context of a repeated line, the assumption made in equation (3.31). For that, reconsider the system illustrated in Figure 3.9. The propagation delay trough $inv_2$, obviously depends on the transition time of the input signal at node (1). This input transition in turn depends on the characteristics of $inv_1$, the value of the RC (distributed) load, and, under certain conditions, on the transition of the ramp input at node *in*. However, as will be explained, in the context of an optimized interconnect line, or more generally RC trees, the transition time of the signal at node *in* should have very little effect on that of node (1).

To make our point, it is important to stress that, among the very important objectives when designing and optimizing any signal distribution network, is the one to limit transition times all over the net. Such a limitation not only allows the preservation of signal integrity, but more importantly ensures the necessary bandwidth for transmitting the clock signal, for instance. Consequently, a high performance design measure is the ability to achieve, within a certain "tolerance", the same transition at the input of every repeating element, such as the CMOS inverter.

In Figure 3.15, we plot the transition time of the signal at node (1) ($T_1$), defined as the transition time of the ramp that passes through the 50% anf 70% points of $v_1$, versus

the transition time of the input signal at node *in*, i.e $T_{in}$. As can be seen, for $T_{in} \leq T_1$, $T_1$ is independent of $T_{in}$. Moreover, even if $T_{in}$ is 20% or even 30% longer than $T_1$ the resulting increase in this transition is on average on the order of 1.3% and 2.5% respectively. This means that the voltage waveform at node (1) does not differ from the waveform at the same node obtained by applying a step input at node *in*, as long as:

$$T_1 \geq 0.7 \cdot T_{in} \tag{3.36}$$

It is interesting to note that this analysis, indeed, justifies the assumption made in equation (3.31), that the inverter's output transition is faster than the input transition, in the context of a repeated interconnect line. Furthermore, the quasi- independence of $v_1$ from $v_{in}$, as will be discussed in Chapter 4, has a tremendous impact in the optimization process of RC trees, especially when considering a bottom-up procedure for post-layout optimization.

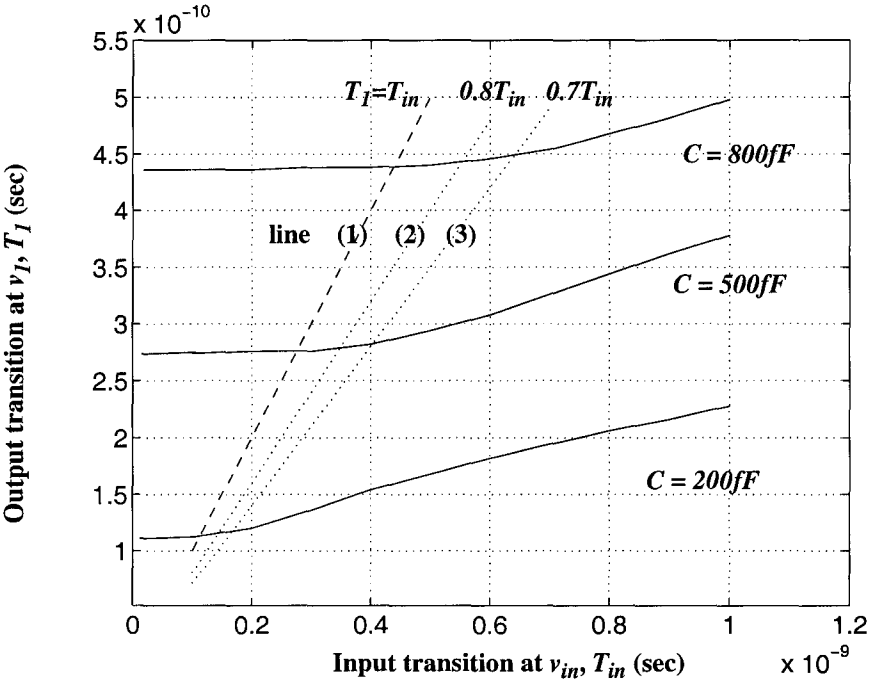



**FIGURE 3.15** $T_1$ dependence on $T_{in}$ for various values of C, the total value of the interconnect capacitance. *inv1* and *inv2* have Wn/Wp =30μm/90μm, R = 200 Ω.

# Chapter 4

# Repeater Insertion in RC Interconnect

In this chapter, we begin by examining the problem of repeater insertion in deep sub-micron technologies, using accurate CMOS inverter and interconnect delay models and metrics, as detailed in the previous chapters. In Section 4.3 we address the problem of inserting and sizing repeaters in on-chip interconnection wires terminated with an arbitrary capacitive load using a quadratic programming, bottom-up approach. Throughout this chapter, we only consider the discharging operation of the symmetrical inverter, the charging operation being similar. Also, for convenience, and without loss of generality, the polarity inversion resulting from a signal passing through an odd number of inverters is ignored.

## 4.1 Preliminaries

In Chapter 1, a number of timing optimization techniques, such as wire sizing/spacing, buffer insertion and gate sizing were reviewed. Repeater insertion has proven to be the "technique par excellence" to minimize propagation delay in interconnect and to control slew, while limiting impact on area and power. However, given the implications of technology scaling on clock, power and, more generally, signal distribution networks[1], interconnect tuning, concurrent to repeater insertion, becomes increasingly a performance

1. Here, we refer essentially to the decreasing interconnect fundamental pitch over technology generations (to achieve routing density) accompanied with an increase in global and semi-global interconnect lengths.

determinant in VLSI systems. That is, the optimal interval at which the repeaters should be inserted, and their sizes, depends largely on the loading by the interconnect, i.e its total capacitance and resistance. This load, of course, depends strongly on the interconnect's aspect ratio and, in some situations[1], on the space between adjacent wires, considering that lateral capacitance tends to dominate the overall interconnect capacitance in modern technologies. On the other hand, one has to know the maximum run made by an interconnect line without an intervening repeater, in order to properly size and space these wires for reliability and signal integrity purposes [54].

In the present work, we are primarily concerned with developing a practical yet accurate method to determine the optimal location and size of repeaters in global and semi global metal layers that support clock and inter-block signal runs. In other words, the question that we will attempt to provide an answer to in this chapter is: given a particular interconnect pitch and/or a certain allocation of width to the interconnect, what is the optimal location or interval at which repeaters should be inserted, and their size, to maximize performance. The resulting technique can be readily incorporated into a more general interconnect optimization scheme that takes conductor sizing and spacing into account.

The problem of optimal repeater placement for a minimum propagation delay target, under different constraints of performance, such as area and power, has been extensively studied, and a variety of techniques have been reported in the literature. These techniques are the result of different approaches to the problem. The most common approach is based on the assumption that a repeater, whose size can span a continuum range[2] of values, can be inserted at any location along an interconnect path [9][10][25], although some restrictions on the repeater placement can be formulated so as to avoid potential wire congestion and via blockages. Other more realistic techniques, such as in [27][29], are based on the assumption that repeaters are organized into blocks, also called feasible location regions. These regions are generally planned in the early stages of the

1. We refer mostly to busses, where lines run parallel for relatively long distances.
2. A large number of post layout optimization tools integrate a buffer library that contain a limited number of inverting and/or non-inverting buffers.

design process, i.e during floorplanning. While the interconnect topology in the first approach is assumed fixed, it is not necessarily the case for the second one.

Repeater insertion is primarily intended to break the quadratic dependence of wire delay on the wire length [9]. The delay in one of the resulting segments is still quadratically dependent on its length, but the total wire delay (delay of the segments plus the inserted repeaters) is nearly linear with total length. However, these repeaters must be optimally spaced and sized to yield optimum results. Figure 4.1 illustrates the efficiency of optimal repeater insertion as a timing optimization technique. This figure may suggest that under the 0.18μm CMOS technology, repeater insertion is not necessary if there is no significant capacitive load at the end of the interconnect wire. In other words, the more loading at the end of the interconnect line, the more relevant the technique of inserting repeaters is. This "relevance" is translated by the rapid movement of the intersection point between "repeated" and "non-repeated" graphs to small lengths as the "output capacitive" load at the end of the interconnect increases [55].

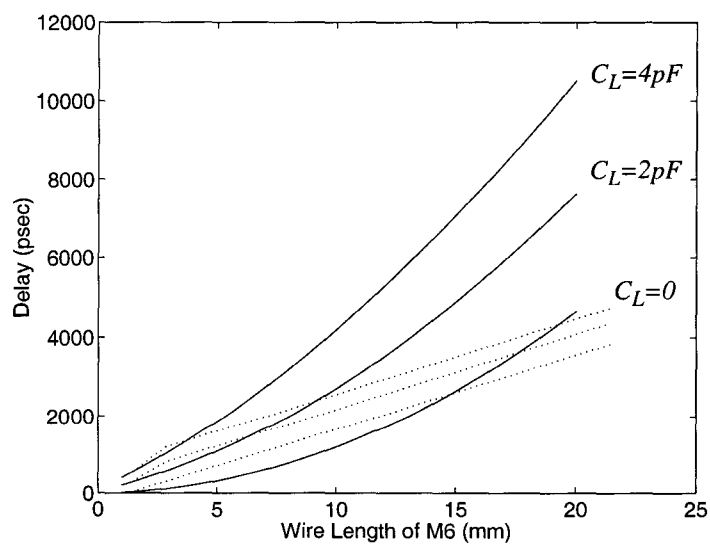



**FIGURE 4.1** Delay dependence of M6 interconnect on its length when it is loaded and: repeated (...), not repeated (-). The line is of minimum width, and maximum lateral capacitance. Resistance per unit length $r$ is of *100Ω/mm* and capacitance per unit length $c$ is of *280fF/mm*. Repeaters are equidistant and equisized ($w_n$=*10μm*, $w_p$=*30μm*)

As discussed in Chapter 2, due to the trends in modern technologies, accurate modeling of the RC distributed nature of interconnect as well as their driving gates is of paramount importance. That is, using simplified models may lead to inferior solutions, since the objective is only an approximation of the delay. Under the 0.18μm CMOS technology process that we use to develop our buffering scheme, the interconnect network is represented as an RC tree where each wire segment is modeled as a π RC circuit. Moreover, the loading effect of such a network on driver gates is modeled by constructing a reduced order π-model approximation based on a frequency analysis of the interconnect network where the first three moments of the driving point admittance are computed.

In this thesis, the CMOS inverter is chosen as the repeater element, since most CMOS gates can be reduced to equivalent inverters [41]. The accuracy of the inverter delay and current model is established with respect to the exact model implemented in a dynamic simulator such as HSPICE. In such a simulator, the distributed nature of interconnect wire is approximated by a β20 model, independently of its length as shown in Figure 4.2.

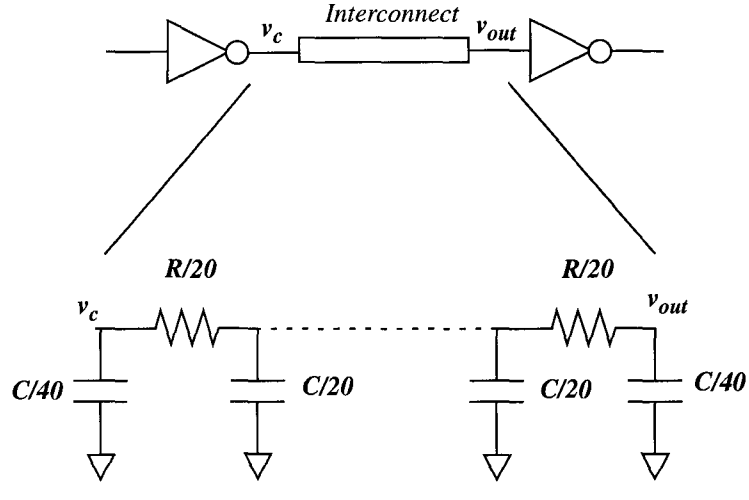



**FIGURE 4.2** β20 circuit model approximation of the distributed RC line

## 4.2 Repeater Insertion In Uniform Lines

It has long been recognized, especially since [9], that under pure delay performance requirements, repeaters in uniform lines should be inserted at uniform intervals and be equally sized [10][11]. Moreover, under the same considerations, these studies show implicitly that while the optimal number of inserted repeaters depends on the total resistance and capacitance of the line, and the technology process under consideration, it is independent of the repeater size. This suggests that optimal repeater placement and sizing in uniform lines are two uncorrelated problems that can be solved separately.

Figure 4.3 shows the optimal number of equisized inverter-like repeaters, when inserted uniformly in a $1cm$ long line, as a function of their size, using HSPICE. This shows that one can use a "reference" inverter in order to determine the optimal interval between adjacent inverters. This information can then be utilized to properly size these repeaters. In our case, such a "reference" is the minimum sized inverter. The reason resides in the fact that such an inverter exhibits a very large output resistance that dominates the interconnect's, and results in very long transition times at the interconnect nodes. As a result, the loading effect of the interconnect segment can be accurately modeled by its total capacitance. Moreover, the segment's interconnect delay reduces to its Elmore delay, thereby greatly simplifying the calculations in this particular context.

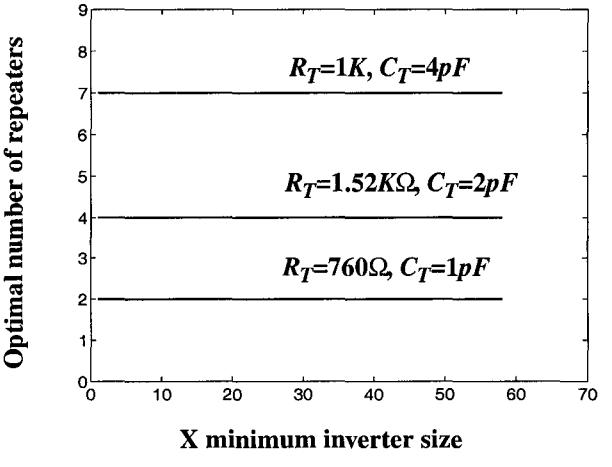



**FIGURE 4.3** Optimal number of inserted repeaters is uncorrelated with their size.

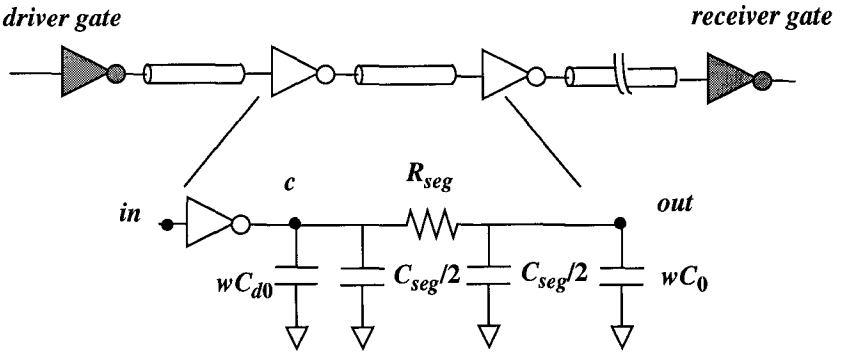



**FIGURE 4.4** Assumed interconnect segment π model and notations.

One important consequence is that the signal transitions at both the repeater's output node (*c*) and the segment's output node (*out*) (Figure 4.4) are identical (Figure 4.5 (a)) and, as will be shown shortly, can be determined as a function of the inverter's total loading capacitance. Note also that in the context of equisized and equidistant symmetrical inverters, the transition characteristics at node (*in*) are identical to those of node (*out*), irrespective of the size of the inserted repeaters.

In this section, we further assume that the inserted repeaters have the same driving capabilities and input capacitance as the driver gate and the receiver gate, respectively, as shown in Figure 4.4. In the same figure, $C_{d0}$ is the sum of the drain-to-substrate equivalent capacitances of both nMOS and pMOS transistors in a minimum size inverter. $C_0$, on the other hand, is the minimum size inverter's input capacitance as expressed in Equation (3.2).

### 4.2.1 Optimal Number of Repeaters

Following the observations made above, the 50% propagation delay through a single stage (from node *in* to node *out*), where the inverter is of minimum size, is:

$$\mathcal{D}_{stage} = \mathcal{D}_{inv} + R_{seg}\left(\frac{C_{seg}}{2} + C_0\right) \tag{4. 1}$$

where $\mathcal{D}_{inv}$ is the propagation delay through the inverter. Recall that due to the very long input transition times resulting from inserting very small inverters along the

interconnection line, the loading effect of the interconnect segment can be modeled by its total capacitance. Moreover, it has been observed that the 50% delay point of $v_c$ occurs always later than $T_i$, i.e the time when the PWL approximation of $v_{in}$ completes its transition to $V_{dd}$. Note also that, given the technology at hand (pronounced velocity saturation limitation), $v_c(t_{st})$ in this case is lower that $V_{dd}$ / 2. Therefore, from equation (3.26) we can write:

$$\mathcal{D}_{inv} = T_i + (v_c(T_i) - 0.5V_{dd})\frac{C_{tot}}{I_{d0}} - \frac{T_i}{2} \tag{4.2}$$

where:

$$I_{d0} = \frac{W_{nmin}\mathcal{B}_{n2}}{L_n}(V_{dd}-V_{tn})^{n_n} \tag{4.3}$$

and

$$C_{tot} = C_{seg} + C_0 + C_{d0} \tag{4.4}$$

At this point we make the following observations. First, in the case of very small inverters, and therefore high output resistance, the resulting voltage waveforms are almost symmetrical (Figure 4.5 (a)). Second, as it turns out, in this same context, $V_{dd}^{eq}$ is always very close to $V_{dd}$. In this case, $T_i$, the transition time of the piecewise-linear approximation obtained by passing a straight line through the 50% and 70% points of $v_{in}$ can be expressed as follows:

$$T_i = \frac{(t_{70} - t_{50})}{0.2} = 5(t_{70} - t_{50}) \tag{4.5}$$

Finally, given the observation we made that, in the particular case of inserting very small inverters, the transition characteristics at nodes *in*, *c* and *out* are identical, $T_i$ can be determined from knowing $t_{50}$ and $t_{70}$ of the inverter's output waveform, i.e at node *c*.

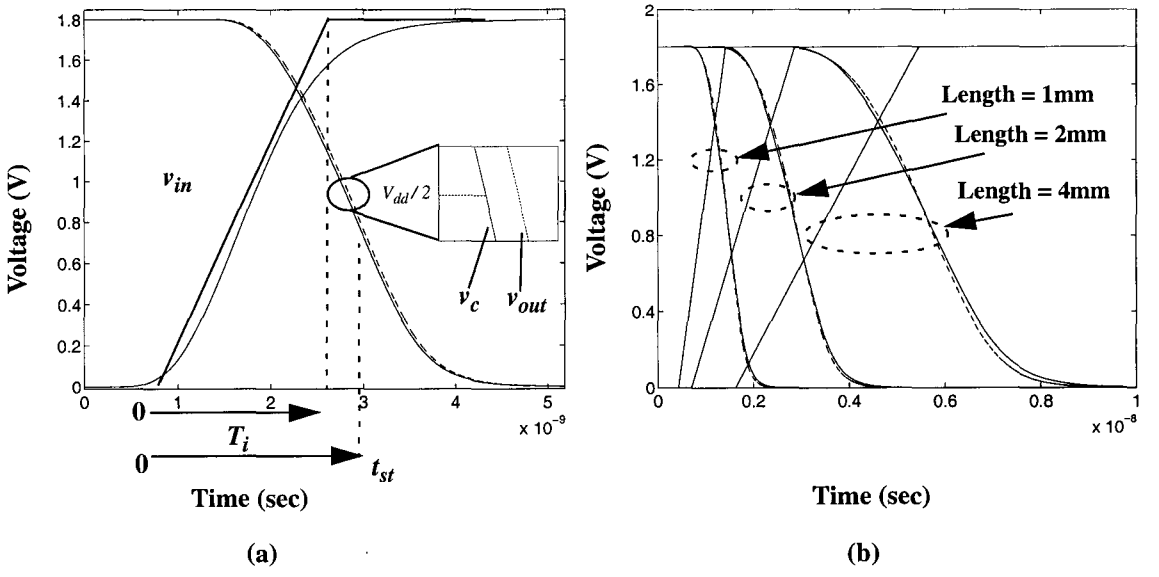



**FIGURE 4.5** **(a)** The resulting voltage waveforms using minimum-sized inverters are nearly symmetrical. **(b)** Using the ramp approximation that passes through the 50% and 70% of the actual input results in good accuracy in capturing the inverter's 50% delay for different segment lengths. Resistance per unit length is 100 Ω/mm and capacitance per unit length is 100 *f*F/mm.

At times $t_{50}$ and $t_{70}$, the discharging transistor is, respectively, in saturation and triode. The complementary transistor on the other hand is off. Therefore, from equation (4.2):

$$t_{50} = T_i + (v_c(T_i) - 0.5V_{dd})\frac{C_{tot}}{I_{d0}} \tag{4.6}$$

and from equation (3.29):

$$t_{70} = t_{st} + R_n C_{tot} \ln\left(\frac{0.3V_{dd}}{V_{d0}}\right) \tag{4.7}$$

where, from equation (3.28):

$$R_n = 0.85\frac{V_{d0}}{I_{d0}} = 0.85R_0 \tag{4.8}$$

and from equation (3.23):

$$V_{d0} = \mathcal{K}_n(V_{dd} - V_{tn})^{m_n} \tag{4.9}$$

From equation (3.27) and (3.25) respectively, $t_{st}$ and $v_c(T_i)$ can be expressed as:

$$t_{st} = T_i + (v_c(T_i) - V_{d0})\frac{C_{tot}}{I_{d0}} \tag{4. 10}$$

$$v_c(T_i) = 0.95V_{dd} - \frac{I_{d0}(V_{dd} - V_{tn})^{n_n+1}}{C_{tot}(n_n+1)V_{dd}}T_i \tag{4. 11}$$

Here $0.95V_{dd}$ is the average value of $v_c(t_{pmax})$ over a large spectrum of (practical) inverter's load capacitance values. In fact, it turns out that $v_c(t_{pmax})$ is a slowly varying quantity as a function of $C_{tot}$, and using equation (3.22) to compute its value will result in highly complicated expression of the inverter's delay with a marginal effect on the delay computation accuracy.

Substituting $t_{st}$ and $v_c(T_i)$ in equations (4. 7) and (4. 6), respectively, by their expressions in equations (4. 10) and (4. 11), results in the following expression of the transition time $T_i$:

$$T_i = 5(t_{70} - t_{50}) = 5\left[\frac{0.5V_{dd} - V_{d0}}{I_{d0}} + R_n \ln\left(\frac{0.3V_{dd}}{V_{d0}}\right)\right]C_{tot} \tag{4. 12}$$

By replacing $T_i$ in equation (4. 2) by its expression in equation (4. 12), the inverter's 50% delay reduces to:

$$\mathcal{D}_{inv} = \eta_{min}C_{tot} \tag{4. 13}$$

where:

$$\eta_{min} = 5\left[\frac{0.5V_{dd} - V_{d0}}{I_{d0}} + R_n \ln\left(\frac{0.3V_{dd}}{I_{d0}}\right)\right]\left(\frac{1}{2} - \frac{V_{dd} - V_{tn}}{(n_n+1)V_{dd}}\right) + \frac{0.45V_{dd}}{I_{d0}} \tag{4. 14}$$

Considering $n$ to be the number of interconnect segments resulting from inserting $K$ equidistant repeaters between the driver and the interconnect load, the total resistance and capacitance seen by each repeater is:

$$\begin{cases} R_{seg} = R_T/n \\ C_{tot} = C_T/n + C_0 + C_{d0} \end{cases} \tag{4. 15}$$

where $R_T$ and $C_T$ are the interconnect total resistance and capacitance, respectively. The total line delay $\mathcal{D}_{line}$ is hence $n$ times the delay of a single stage given by equation (4. 1). Thus $\mathcal{D}_{line}$ has the form:

$$\mathcal{D}_{line} = n\left(\eta_{min}\left(\frac{C_T}{n} + C_0 + C_d\right) + \frac{R_T}{n}\left(\frac{C_T}{2n} + C_0\right)\right) \tag{4. 16}$$

Setting $\frac{d\mathcal{D}_{line}}{dn_{stage}} = 0$ results in the optimum number of stages:

$$n_{opt} = \sqrt{\frac{R_T C_T}{2\eta_{min}(C_0 + C_{d0})}} \tag{4. 17}$$

Therefore the optimum number of inserted repeaters $K_{opt}$ is:

$$K_{opt} = \sqrt{\frac{R_T C_T}{2\eta_{min}(C_0 + C_{d0})}} - 1 \tag{4. 18}$$

This expression resembles the one presented by Bakoglu in [9] and by the authors in [25], where the optimal number of repeaters inserted along an interconnect is a direct function of the product of the total resistance and capacitance of such an interconnect. However, unlike [9], the expression in equation (4. 18) takes into account explicitly the inverter's drain and source diffusion capacitance as well as the actual transition time at the input of the repeaters. Moreover, the transistors are modeled using the n-th power law as opposed to the switch-level model utilized in [9]. Also, since $K_{opt}$ is not necessarily an integer number, using Bakoglu's less accurate expression, $K_{opt}$ may be rounded off to the wrong integer resulting in overestimating the required number of repeaters.

### 4.2.2 Repeater Sizing

Once the load, or the length, of a single segment has been determined, the optimal size of the repeater can now be computed. For this, we start by modeling the stage's (from node *in* to node *out*) output voltage waveform $v_{out}$. For times prior to $t_{st}$, the time when the discharging transistor leaves saturation and enters triode, this waveform can be simply modeled using a PWL approximation. For times later than $t_{st}$, however, for which most of the $v_{out}$ excursion from high to low occurs, the waveform is accurately described by a decaying exponential. More explicitly, $v_{out}$ can be expressed as follows:

$$\begin{aligned} v_{out}(t) &= V_{dd} - S_{out}(t - t_{start}) && t \le t_{st} \\ v_{out}(t) &= v_{out}(t_{st}) e^{-\frac{(t-t_{st})}{\tau_{out}}} && t \ge t_{st} \end{aligned} \tag{4. 19}$$

where $S_{out}$ is the slope of the PWL approximation, $t_{start}$ is the time when $v_{out} = V_{dd}$, and $\tau_{out}$ is the fall time-constant. Therefore, we can write:

$$t_{start} = t_{st} + \frac{V_{dd} - V_{out}(t_{st})}{S_{out}} \tag{4. 20}$$

The relevance (and accuracy) of such an approach is shown in Figure 4.6.

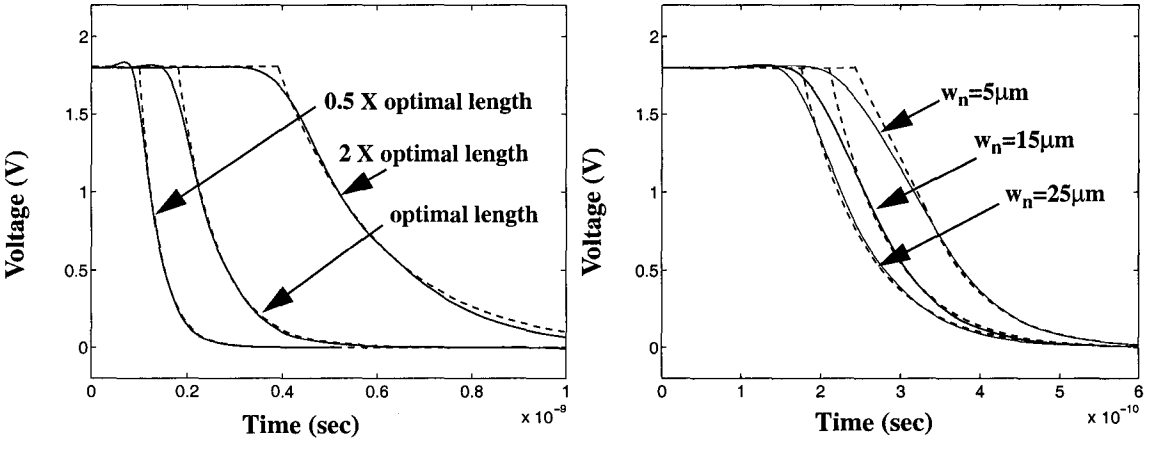



**FIGURE 4.6** Stage output waveform $v_{out}$ for (a) an Inverter, $W_n$=30μm $W_p$=90μm, driving a number of interconnect segment lengths and (b) a number of inverters, $W_n$=5μm $W_n$=15μm $W_n$=25μm driving a segment of optimal length, using HSPICE (-) and the model given by equation (4.19) (--).

Now consider the circuit in Figure 4.7. In Chapter 3 we showed that the transition time at node *out*, $T_{out}$, is independent of the one at node *in*, $T_i$, as long as $T_{out} \geq 0.7 \cdot T_i$. In this case, the transition at *out* does not differ from the one obtained at the same node when applying a step input at node *in*.This is assumed as always true in the case of repeated interconnects or RC trees. $T_i$, as also mentioned, determines only the delay at *out*.

As can be seen in Figure 4.8, as in the case of transition times, as long as $T_{out} \geq 0.8 \cdot T_i$, the stage delay is a linear function of input transition time $T_i$. Based on this observation the stage delay can therefore be expressed as follows:

$$\mathcal{D}_{stage} = \mathcal{D}_{stage}^{step} + \zeta \cdot T_i \tag{4. 21}$$

where $\mathcal{D}_{stage}^{step}$ is the 50% stage delay when a step input is applied at *in*. $\zeta$ is an empirical technology-dependent parameter whose value in our case is $\zeta$=0.15. In this case $\mathcal{D}_{stage}^{step}$ can be trivially computed. For this particular case, from equation (4. 19):

$$\mathcal{D}_{stage}^{step} = t_{st} - \tau_{out} \ln\left(\frac{0.5V_{dd}}{v_{out}(t_{st})}\right) \tag{4. 22}$$

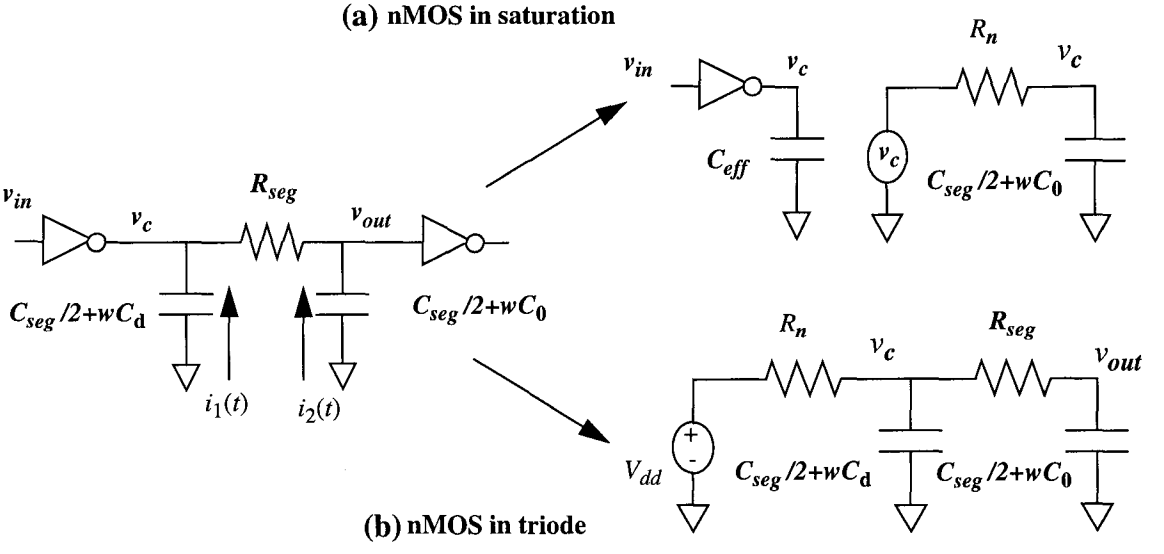



**FIGURE 4.7** Stage's equivalent circuit when the discharging transistor is in **(a)** saturation and **(b)** triode [45].

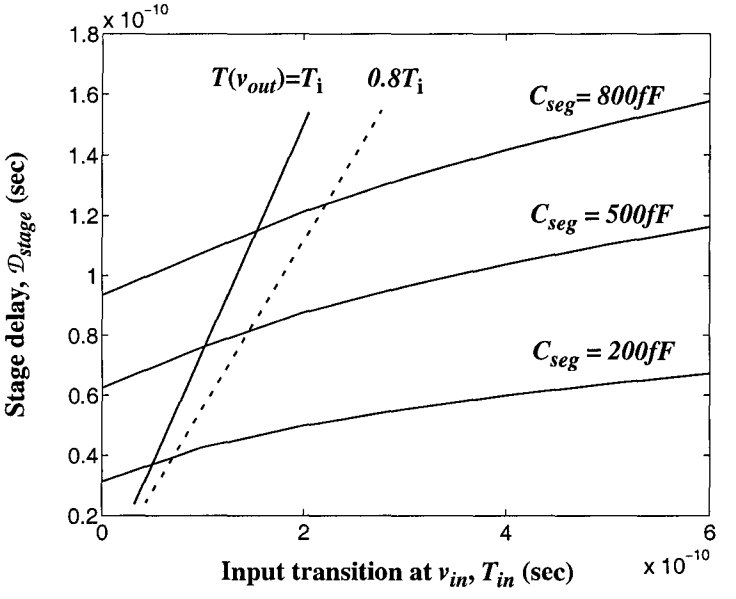



**FIGURE 4.8** $\mathcal{D}_{stage}$ dependence on $T_i$ for various values of $C_{seg}$, the total capacitance of the interconnect segment. The inverter size is $W_n/W_p$ =30μm/90μm, $R_{seg}$ = 200Ω

Note that in the case of relatively large inverters, most of the transition at node *out* occurs while the discharging transistor is in the triode regime. This justifies the preceding delay expression of the 50% delay point of $v_{out}$, or $\mathcal{D}_{stage}^{step}$.

Now, to sort $\mathcal{D}_{stage}^{step}$ we need to compute $v_{out}(t_{st})$ and $t_{st}$. Applying *KVL* on the circuit of Figure 4.7 we have:

$$\begin{aligned} v_{out}(t_{st}) &= R_{seg} i_2(t_{st}) + v_c(t_{st}) \\ &= R_{seg}(wI_{d0} - i_1(t_{st})) + V_{d0} \end{aligned} \tag{4.23}$$

where:

$$\begin{aligned} i_1(t_{st}) &= (C_{seg} + wC_d) \cdot \left.\frac{dv_c}{dt}\right|_{t = t_{st}} \\ &= \frac{(C_{seg} + wC_d)}{C_{eff}} \cdot wI_{d0} \end{aligned} \tag{4.24}$$

Therefore:

$$v_{out}(t_{st}) = wR_{seg}I_{d0}\left(1 - \frac{(C_{seg} + wC_d)}{C_{eff}}\right) + V_{d0} \tag{4. 25}$$

Here $w$ is the relative size of the repeaters with respect to the minimum sized one.

$t_{st}$ on the other hand, following equation (3.27), can be expressed as:

$$t_{st} = \frac{\Psi_w}{w} \tag{4. 26}$$

where:

$$\Psi_w = \frac{V_{dd} - V_{d0}}{\frac{W_{min}\mathcal{B}_{n2}}{L_n}(V_{dd} - V_{tn})^{n_n}} C_{eff} \tag{4. 27}$$

Recall that the input signal ramp approximation, as defined in the previous chapter, is the one that passes through the 50% and 70% of the actual input. Moreover, as mentioned earlier, since the inserted repeaters along the line are equidistant and symmetrical, both the input waveform at node *in*, and stage output voltage waveform at *out* exhibit the same transition time characteristics in their rise and fall respectively. Thence using (4. 5) we can write:

$$\frac{T_i}{V_{dd}} = 5\frac{\tau_{out}}{V_{dd}} = 2.55\tau_{out} \qquad (V_{dd} = 1.8V) \tag{4. 28}$$

Therefore, from (4. 21) and (4. 22) the stage delay is:

$$\begin{aligned} \mathcal{D}_{stage} &= t_{st} - \tau_{out}\ln\frac{0.5V_{dd}}{v_{out}(t_{st})} + 2.55\zeta V_{dd}\tau_{out} \\ &= \frac{\Psi_w}{w} + \chi_w\tau_{out} \end{aligned} \tag{4. 29}$$

where:

$$\chi_w = 2.55\zeta V_{dd} - \ln\left(\frac{0.5V_{dd}}{v_{out}(t_{st})}\right) \tag{4. 30}$$

Since we are in the presence of uniform interconnect segments, $\tau_{out}$ can be very well approximated by the first moment of the impulse response of the circuit in Figure 4.7(b).

$$\tau_{out} = (wC_{d0} + C_{seg}/2) \cdot R_n + (C_{seg}/2 + wC_0) \cdot (R_{seg} + R_n) \tag{4. 31}$$

where from equation (3.28)

$$R_n = 0.85 \cdot \frac{V_{d0}}{wI_{d0}} \tag{4. 32}$$

Note that in the expression above, $V_{d0}$ replaces $v_c(t_{st})$ and $wI_{d0}$ replaces $i_N(t_{st})$ since we are computing $\mathcal{D}_{stage}^{step}$, and therefore in this case we are in the situation where $T_i < t_{st}$.

Now, by setting $\frac{d\mathcal{D}_{stage}}{dw} = 0$, the optimal size of the inserted repeaters resulting in the minimum propagation delay along the interconnection line $W_{opt} = w_{opt}W_{min}$ is expressed as:

$$w_{opt} = \frac{W_{opt}}{W_{min}} = \sqrt{\frac{\Psi_w + \chi_w R_0 C_{seg}}{\chi_w R_{seg} C_0}} \tag{4. 33}$$

Note that, from equations (4. 25) (4. 27) and (4. 30), both $\Psi_w$ and $\chi_w$ are dependent on the quantity $w_{opt}$ we are seeking. The following iterative procedure is therefore used to compute accurately the value of $w_{opt}$:

1. Chose an initial value of $w_{opt}$.
2. Compute $\Psi_w$ and $\chi_w$ given the background developed in Chapter 2 and Chapter 3. More explicitly, from (2.32) and the expression of the effective capacitance in the case of a falling output in (2.35) we have (considering a step input):

$$\begin{aligned} C_{eff} &= \left(\frac{V_{dd} - wR_{\Pi}I_{d0} - V_{d0}}{V_{dd} - V_{d0}}\right)_{\Gamma} C_1 + C_2 \\ &= \left(1 - \frac{wR_{\Pi}I_{d0}}{V_{dd} - V_{d0}}\right)_{\Gamma} C_1 + C_2 \end{aligned} \tag{4. 34}$$

Then $C_{eff}$ in equation (4. 25) is replaced with its expression above to yield $v_{out}(t_{st})$. For convenience, the procedure for computing the effective capacitance [45] is reproduced in Figure 4.9.

3. Compute $w_{opt}$ using equation (4. 33).
4. If the new $w_{opt}$ differs from the old one by an error of more than 5%, then set $w_{opt}$ to the new value and go back to step 2. Otherwise, terminate iteration.

This iterative procedure usually converges in two or three steps. Note that the expression in equation (4. 33) resembles somewhat the one reported in [25]. However, in our case $w_{opt}$ is not a direct function of the ratio of the interconnect's total capacitance to its total resistance given the additional term ($\Psi_w$) in the numerator. In fact, the authors in [25] did not take resistance shielding into account.

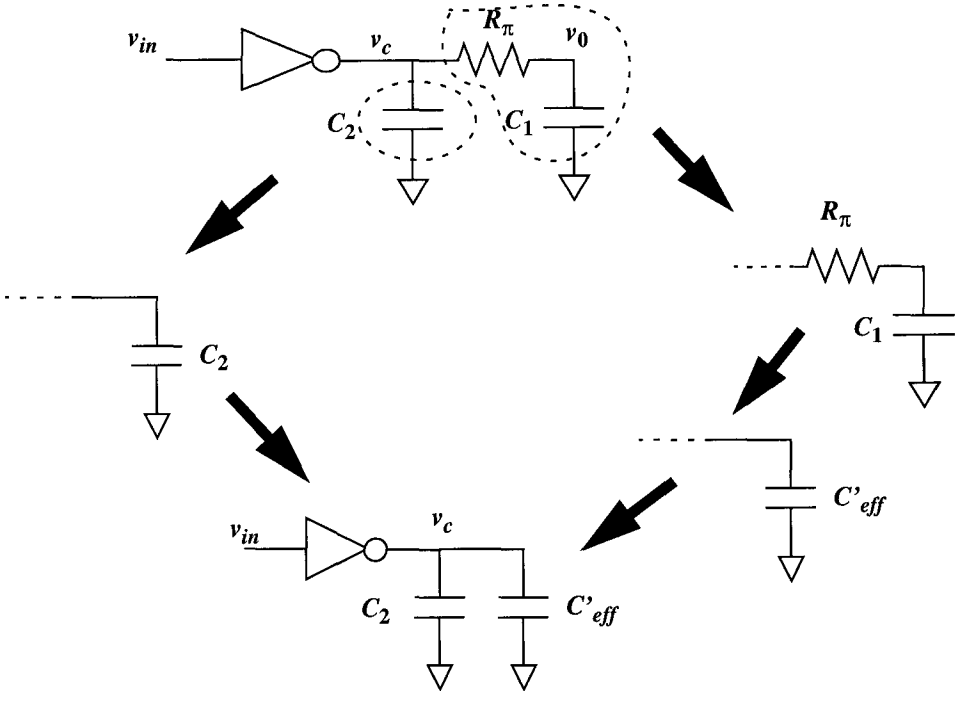



**FIGURE 4.9** Procedure for obtaining $C_{eff}$ for a Π load. The Π load is decomposed into lumped capacitance $C_2$, and a Γ load $R_\Pi C_1$, whose effective capacitance is obtained using the model in Section 2.4 [45]. $v_{in}$ in this particular case is a step input.

## 4.3 Bottom-Up Repeater Placement in RC Interconnect

Algorithms for post-layout repeater insertion have been an extremely active area of study in the past decade. An early work on buffer insertion in RC trees was presented by Van Ginneken [27], who proposed a dynamic programming algorithm that solves for minimum Elmore delay from the source to the leaves of a given wiring tree. That is, as illustrated in Figure 4.10, given the required arrival times $t_i$ at the leafs $S_i$ (also known as signal sinks) of the RC network, the algorithm finds the best insertion solution, i.e repeater locations and sizes, that results in the latest signal "departure" time, $t_{source}$, at the root of the tree $S_0$. This is equivalent to defining the repeater insertion configuration that ensures that:

$$t_{source} = min_i(t_i - D_i) \tag{4.35}$$

where $D_i$ is the delay between the source of the signal $S_0$ and the sink $S_i$.

Essentially, the formulation assumes that the possible repeater locations (called legal positions) in the RC tree are known, as well as all possible allocations of repeater and wire segment sizes. This method comprises two distinct phases. In the first phase, the algorithm exploits the hierarchical nature of the Elmore delay to construct, in a bottom-up fashion, a set of feasible partial solution consisting of $(c_k, t_k)$ pairs, called pair options, at a potential insertion point $k$ represented by dashed repeater symbols in Figure 4.10. $t_k$ is the

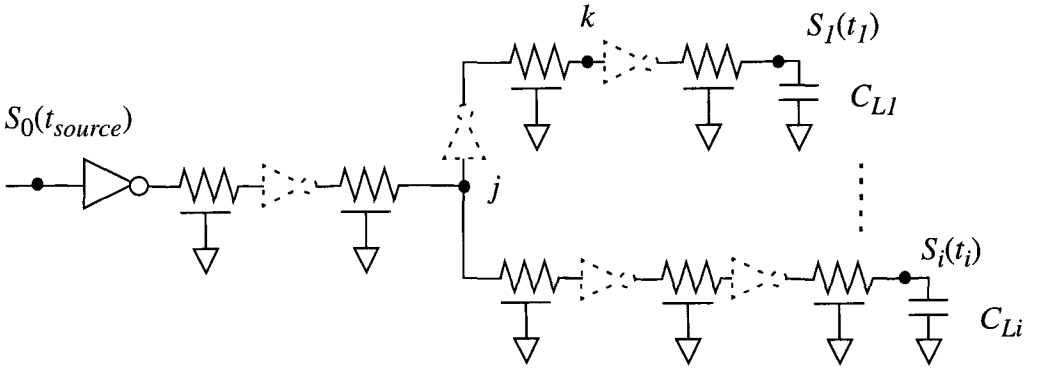



**FIGURE 4.10** RC routing tree with legal repeater insertion points represented by dashed inverter symbols

required arrival time at point $k$, and $c_k$ is the total capacitance "seen" downstream from

that point. It is clear that these pairs carry new load information and timing effects resulting from inserting a repeater at a particular node. For instance, consider the wire segment $[k, S_1]$ in Figure 4.10. Let us assume that this wire is of length $l$ and has a capacitance and resistance of $c_w$ and $r_w$ per unit length, respectively. Using Elmore as a delay metric, and inserting a repeater of input capacitance $c_{buf}$ and output resistance $r_{buf}$ at node $k$, the partial insertion solution $(c_k, t_k)$ at this point is therefore such as:

$$\begin{aligned} c_k &= c_{buf} \\ t_k &= t_1 - D_w - D_{buf} \end{aligned} \tag{4. 36}$$

where $D_w$ and $D_{buf}$ are the wire and buffer delays respectively. Therefore:

$$\begin{aligned} D_w &= r_w l(0.5 c_w l + C_{L1}) \\ D_{buf} &= (c_w l + C_{L1}) r_{buf} \end{aligned} \tag{4. 37}$$

If one possesses a library of buffers of different sizes, a pair $(c_k, t_k)$ is generated for every possible size at node $k$. In the resulting set, inferior pairs are then pruned using the following rule:

*For any two pairs,* $(c_1, t_1)$ *and* $(c_2, t_2)$ *part of a partial solution set, if* $c_1 < c_2$ *and* $t_1 > t_2$, *then the pair* $(c_2, t_2)$ *can be removed from the set.* (4. 38)

At an insertion point $j$, for instance upstream of $k$, a pair is generated for every possible repeater size. The first element in the pair is obviously the input capacitance of the inserted repeater. The required time, on the other hand, is the latest time that can be achieved when considering every option pair in the partial solution set at node $k$. The pair in the set at node $k$ that led to the one in $j$ is then tagged to it using a general data structure. These option pairs are then in turn propagated backwards toward the root of the tree.

It is easy to see that $t_{source}$ can be easily computed recursively using this method, and corresponds to the option with the best required time at node $S_0$. The second phase traces back the computation of the first phase that led to this option, and places buffers.

Later, several variants to Van Ginneken's algorithm were proposed [28][30]. Especially in [30], Lillis *et al* presented an elegant dynamic programming approach to concurrent repeater insertion and wire sizing. In the following subsection we describe more formally such an approach since it constitutes the basis for our repeater insertion scheme.

### 4.3.1 Base Technique

For convenience we reproduce the notation often used for post layout repeater insertion technique:

- $T$: A routing tree with $n$ branches and a set of $m$ sinks $\{S_1, S_2, ..., S_m\}$.
- $e_i$: $i$-th branch of the tree, $1 \leq i \leq n$.
- $l_i$: Length of $e_i$.
- $Children(e_i)$, set of immediate children of $e_i$.
- $rcdelay(e_i, c)$: The RC delay of branch $e_i$ when loaded by a capacitive load $c$.
- $cap(e_i)$, $res(e_i)$, capacitance and resistance of branch $e_i$.
- $B$: Library of repeaters.
- $delay(b, c)$: Delay of repeater $b$ when loaded by a capacitive load $c$.
- $t_i$: User-specified maximum delay constraint at sink $S_i$.
- $C_{Li}$: Capacitance load at sink $S_i$.

At each sink $S_i$ with load capacitance $C_{Li}$, the partial solution set at this point, $\Omega_i$, obviously consists of a single pair $(C_{Li}, t_i)$, where $t_i$ is the user-specified required time for the sink $S_i$. Working back toward the root of the tree, for each potential insertion point a partial solution set is constructed based on the solution set of the insertion point immediately downstream. As illustrated in Figure 4.11, assuming that the solution set of pairs $\Omega_j$ at a point $j$ is known, the solution set $\Omega_i$ at a point $i$ upstream of $j$ can be constructed. That is, for every repeater $b \in B$ potentially inserted at node $i$, a pair of the set $\Omega_i$ is generated. The capacitance of the deduced pair in the set $\Omega_i$ is the repeater's input capacitance $c_b$. The required time of the pair, on the other hand, is the latest time that can result from considering every pair in $\Omega_j$. In other words one has to find the pair $(c_j, t_j) \in \Omega_j$ such that:

$$t_j - delay(b, c_j + cap(e_i)) - delay(e_i, c_j) > t' - delay(b, c' + cap(e_i)) - delay(e_i, c') \qquad \forall (c', t') \in \Omega_j \tag{4.39}$$

The generated pairs are stored in a global data structure and the pair $(c_j,t_j)$ of the set $\Omega_j$ indexes the pair of the set $\Omega_i$ that it had led to. For the pairs at $S_0$ the option with the best $t_{source}$ is chosen. From there it is easy to trace back the option pairs that led to the optimal solution and therefore to place and size repeaters optimally. Note that this method can lead to more than one solution, i.e placement configuration. Choosing the right configuration would depend on eventual constraints such as wire congestion, via blockage, area and power. Note also that the pair at node $i$+ is the pair for which no repeater is inserted.

The main drawback of the techniques mentioned above is their reliance on the Elmore constant as a delay measure. In fact, in addition to the inherent inaccuracy of the Elmore delay in modern technologies, it is unable to account for the transition time which is quite an important determinant of performance in terms of delay, power and noise immunity. Moreover, as mentioned in Chapter 2, one of the inputs to the effective capacitance model is the driver's input transition time. Therefore, while inserting a repeater at point $i$, the input transition time has to be taken into consideration. Surprisingly, very few

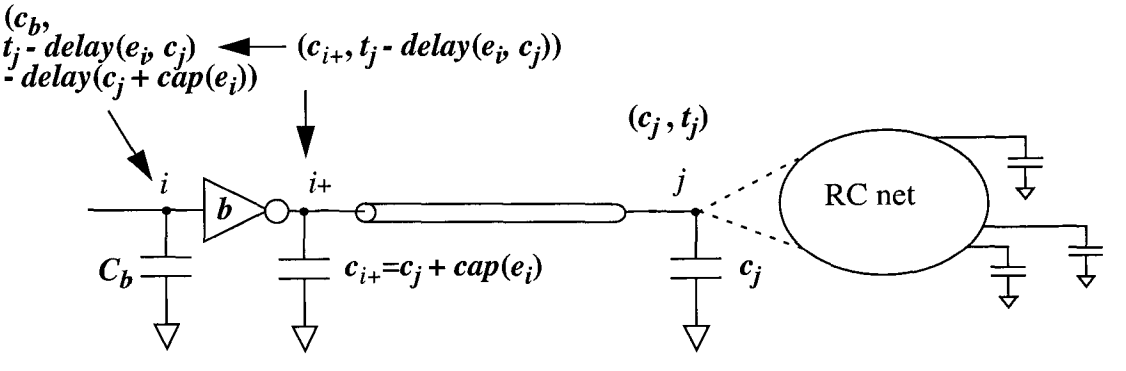



**FIGURE 4.11** Bottom-up pair generation using Elmore as the delay metric.

studies take this parameter into account in their timing analysis models and therefore, by extension, into their timing optimization routines.

A first attempt to address some of these issues was made by Alpert *et al* [28] who presented an extension to the Van Ginneken technique that uses both accurate interconnect and gate delay models for inserting repeaters and sizing wires. That is, they used moments matching and propagation techniques. The problem, however, is that, given the bottom-up nature of Van Ginneken's algorithm, they assumed fixed input transition times at the input of every inserted buffer. In [29] Menezes and Chen, on the other hand, presented an accurate technique for handling transition times at the input of the inserted repeater. This technique is elegant and interesting enough to deserve further examination.

Consider the system illustrated in Figure 4.12. As seen, at the signal sink, both the "arrival" and the corresponding transition times are specified. Also, for every repeater in the buffer library, a pair is generated at node $k$. This pair consists of the input capacitance of the inserted buffer $c_k$ and a table of (input-transition-time, required arrival time), i.e $(c_k, \{(tr_1, t_{k1}), (tr_2, t_{k2}),..., (tr_n, t_{kn})\}$ for $n$ possible transition times, $tr_1$ $tr_2$, ..., $tr_n$. When inserting a repeater $b$ at node $i$ is considered, the best pair $(c_b, t_b)$ for this repeater that satisfies the condition in equation (4. 39) is extracted. Then, for each value of the input transition time $tr_i$ (at node $i$), the delay from node $i$ to node $k$, $t_{bk}(tr_i)$, is computed. During the pro-

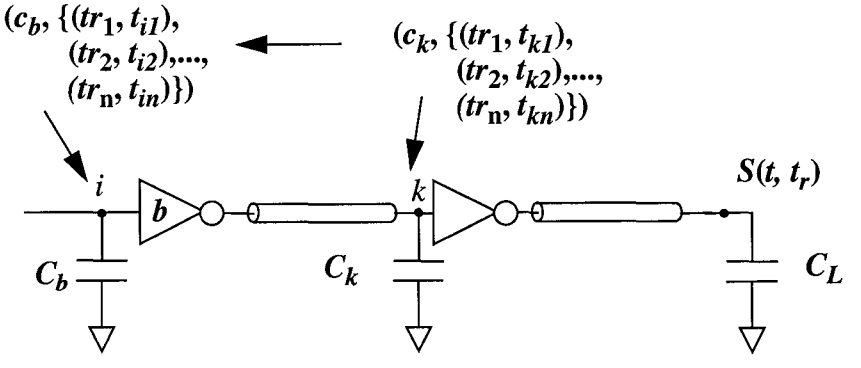



**FIGURE 4.12** Bottom-up pair generation using a library of potential transition times at insertion points.

cess, the transition at node $k$, $tr_{bk}(tr_i)$, is also computed. Now, to compute the required

time at node $i$ when applying the input transition $tr_i$, one has to know the "resulting" required time at $k$. In other words, one has to know the required time at $k$ that leads to the one at $i$ when applying the transition $tr_i$. Given the information in the table (transition time, required arrival time) at node $k$ and the transition resulting from applying $t_{ri}$, i.e $tr_{bk}(tr_i)$, the authors of [29] constructed the graph shown in Figure 4.13 from the table at node $k$ by linear extrapolation.

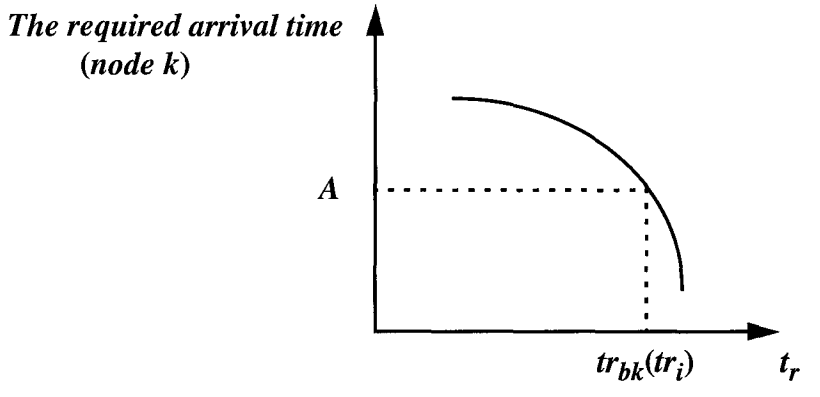



**FIGURE 4.13** Handling of input transition time effects for required time computation in [29]

Therefore the required time at node $i$ resulting from inserting buffer $b$ and applying transition $tr_i$ at node $i$ $t_i(tr_i)$ is:

$$t_i(tr_i) = A - t_{bk}(tr_i) \tag{4.40}$$

Each entry of the resulting ($t_r$, required-time) is tagged with the output transition time, such that one would be able to trace back the optimal solution during the second phase of the algorithm. From here, the transition time constraints are handled in a straightforward manner. That is, when a table is created for the first repeater before a sink, if the output transition time is less than the user specified constraint, the corresponding entry is eliminated.

### 4.3.2 The Proposed Method

In the previous chapter we mentioned that among the very important objectives when designing and optimizing any signal distribution network, is the one to limit transition times all over the net. Therefore, a high performance design measure is the ability to achieve, within a certain "tolerance", the same transition at the input of every repeating element along any interconnect. As shown in Figure 3.15, implementing such an objective results in the property that the transition time at the far end of an interconnect segment (resulting from inserting repeaters along the line) is independent from the one at the input of the segment's driving repeater. Therefore, it is very well justified to assume that in an RC signal distribution tree, the transition at node $k$ is independent from the one at node $i$ in Figure 4.12. Based on this crucial observation, we developed a repeater insertion scheme that has the accuracy of the technique presented by Menezes *et al* [29] and nearly the simplicity of Lillis's [30] while taking transition times into account. The insertion scheme is primarily concerned with inserting repeaters in global interconnects terminated with an arbitrary capacitive load.

Consider the system illustrated in Figure 4.14. At the sink, we specify both the load capacitance and the arrival time. The pair options at node $k$ are generated based on the following procedure:

1. At the sink, the partial solution set, $\Omega_S$, consists of a single pair $(C_L,t)$.

2. For every repeater $b \in B$ , we generate a pair consisting of the input capacitance of the inserted repeater, $c_b$, and a table of ("a priori" repeater at node $i$, arrival time at node $i$, required time at node $k$) triplets, $(c_b, \{(c_{b1},t_{i1}, t_{k1}), (c_{b2}, t_{i2}, t_{k2}),..., (c_{bn}, t_{in}, t_{kn})\}$. The arrival time at node at node $i$ is computed assuming a step input at $i$. Note that this arrival time, as shown in Figure 4.14, is not the required time part of the solution set $\Omega_i$. On the other hand, given the assumption that the transition at $k$ is independent from the one at $i$, the required times at node $k$ in the table are the actual ones. This results in a set of pair options, $\Omega_k$, that constitutes evidently a partial solution set at node $k$.

3. We prune inferior solutions in the set $\Omega_k$. That is for every "a priori" repeater (at node $i$), we search for the repeater at node $k$ (i.e the first element in the pair) that corresponds to the latest arrival time at node $i$. The other pairs with the same "a priori" repeater are eliminated from the set. The remaining pair is then tagged with the repeater at node $i$.

4. Finally, we prune obvious inferior solution pairs in $\Omega_k$ following the criterion in equation (4. 38).

Transition time requirements can be trivially handled. When we generate the options at node $k$, the corresponding transitions at this node are naturally computed since used for delay computation. If a transition exceeds a pre-specified upper bound on transition times, the corresponding pair is eliminated from the solution set. At the root node ($S_0$), in addition to the latest arrival time, other constraints, or requirements, can be taken into account, such as area.

For example, assume that library $B$ contains three repeaters $b_1$, $b_2$, $b_3$ (Figure 4.14). We want to generate the partial solution set at node $k$ by assuming a priori buffers at node $i$. From Figure 4.14, generating solution sets at nodes upstream of $k$ becomes a trivial task.

The special case of no buffer inserted in a candidate insertion location is handled differently. That is, a "no buffer" can not be considered as a "a priori" buffer. For this reason all the pairs in the set $\Omega_k$, for instance, are tagged to the "no buffer" option when considered in insertion at node $i$. Figure 4.15 illustrates the insertion procedure when considering the "no buffer" option is represented by the symbol $\Phi$. Here, for convenience, we assume that the buffer library contains only one repeater of input capacitance $c_b$.

Finally, It is easy to see that this method can be readily extended to handle more complex signal distribution networks modeled as RC trees, where a required arrival time is specified for each sink.

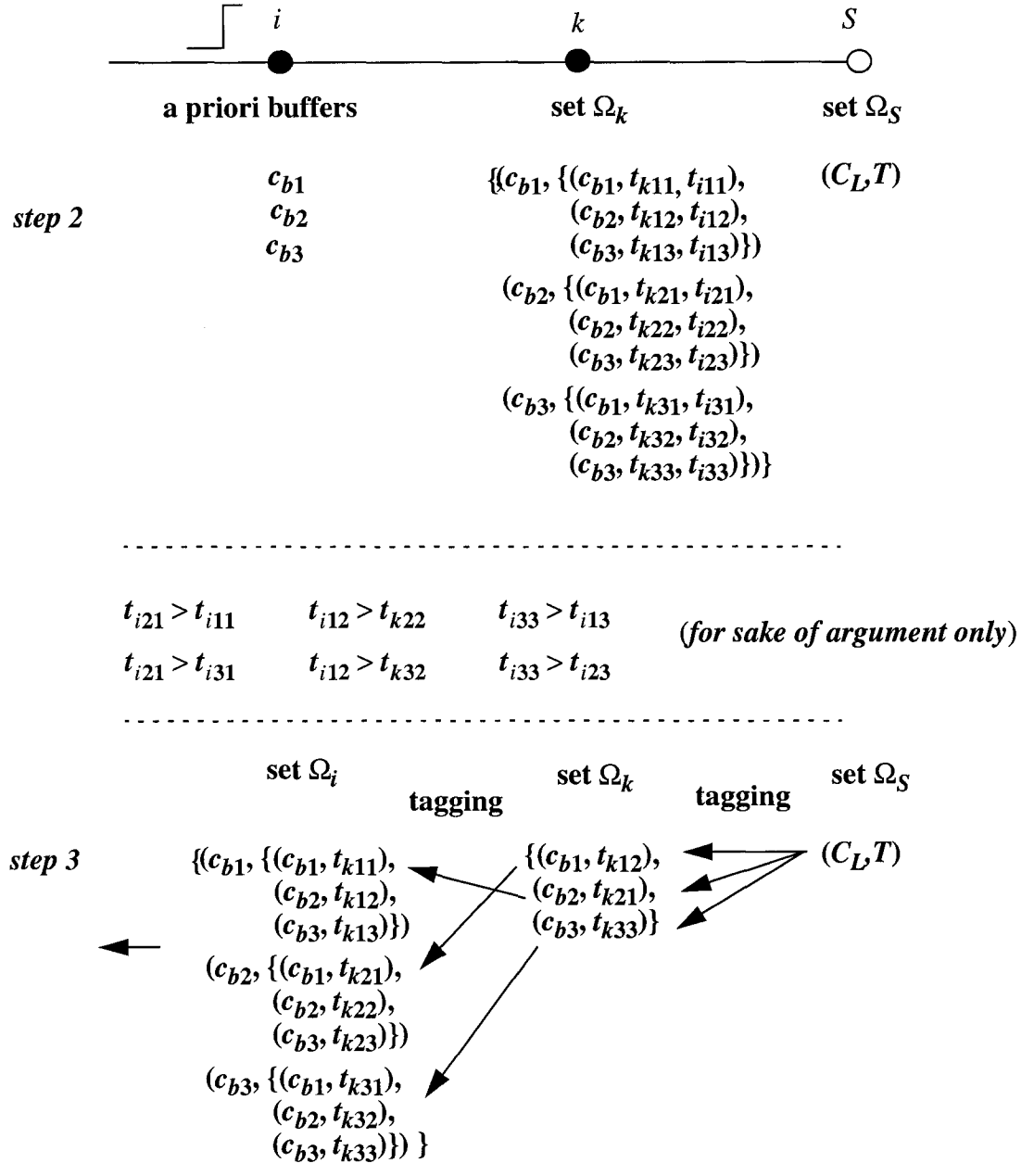



**FIGURE 4.14** Control-flow diagram of the proposed repeater insertion technique

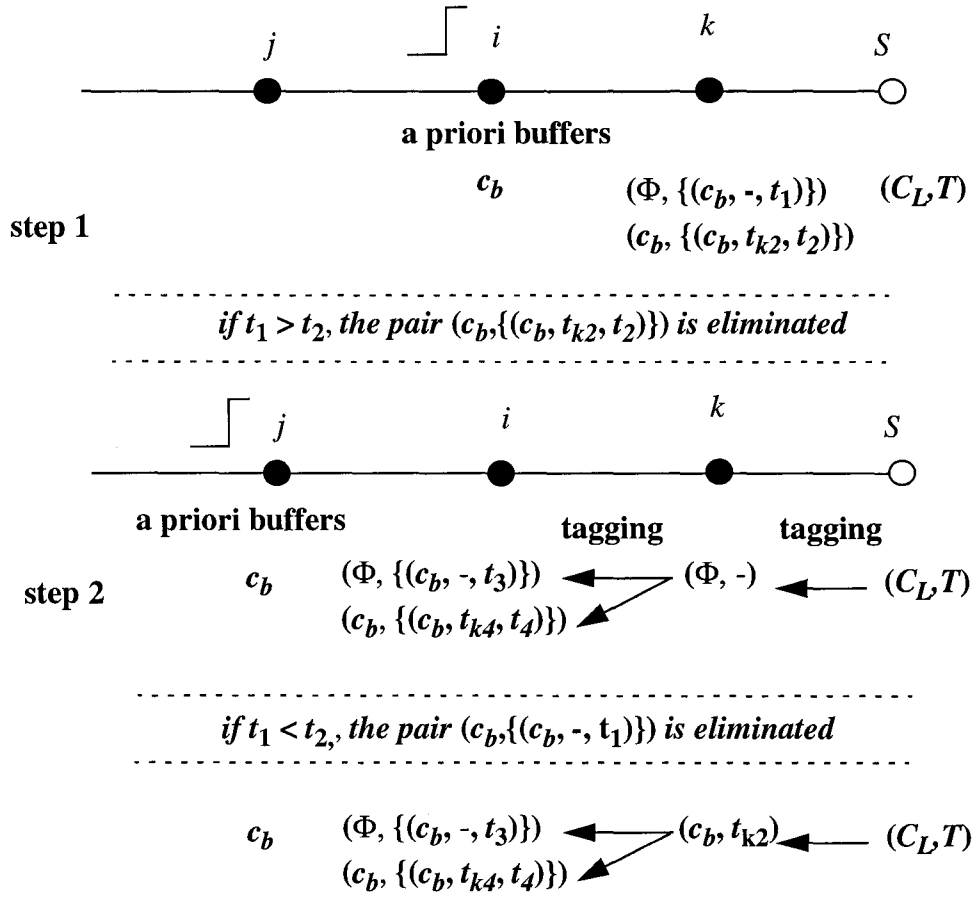



**FIGURE 4.15** Control-flow diagram of the proposed repeater insertion technique when considering the "no buffer" option represented by the symbol Φ.

# 4.4 Results

## 4.4.1 Repeater Insertion in Uniform Lines

For an RC interconnect line of, say $1cm$ length, we determine the optimal number of inserted repeaters and their size using equation (4. 18) and (4. 33), respectively, as a function of its total resistance $R_{tot}$ and total capacitance $C_{tot}$. The size of these repeaters and the resulting line delay is compared with what is obtained using HSPICE. In this particular case, as mentioned in section 4.2, we assume that the driver gate and the load gate have the same driving capabilities and input capacitance, respectively, as the inserted repeaters. The accuracy in computing the optimum repeater size and the line's overall delay is shown in Table 4.1. As reported in previous studies, such as in [9] and [25], the optimal interval between adjacent repeaters and the overall optimized interconnect's delay is a function of the product $R_{tot}C_{tot}$. On the other hand, the optimal size of a repeater is rather a function of the ratio $C_{tot}/R_{tot}$. Note that, from Table 4.1, for a given $R_{tot}C_{tot}$ product, the estimated line delay remains the same independently from the actual value of $R_{tot}$ or $C_{tot}$. For instance, the delay for a line with $R_{tot}$ of 3KΩ and $C_{tot}$ of 1pF with an optimum repeater size of 12μm is 708 psec. Similarly, the delay for a line with $R_{tot}$ of 1KΩ and $C_{tot}$ of 3pF with an optimum repeater size of 35μm is also 708 psec.

**Table 4.1** Comparison of repeater insertion model with HSPICE results.

| $C_{tot}(pF)$ | $R_{tot}$ (KΩ) | $n_{opt}$ | $W_{opt}$ (μm) (model) | $W_{opt}$ (μm) (HSPICE) | Delay (ps) equ. (4. 21) | Delay (ps) (HSPICE) | Error (%) |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 4 | 20 | 20.2 | 431 | 411 | 4.8 |
| 3 | 1 | 7 | 34.6 | 35 | 731 | 708 | 3.1 |
| 1 | 3 | 7 | 11.5 | 12 | 732.5 | 708 | 3.2 |
| 3 | 3 | 13 | 20 | 20.2 | 1276 | 1225 | 4.18 |
| 1.5 | 1 | 5 | 24.5 | 24 | 526 | 502 | 4.7 |
| 1 | 1.5 | 5 | 16.32 | 16 | 526 | 502 | 4.7 |
| 2 | 2 | 8 | 20 | 21 | 843 | 815 | 3.43 |

Note that, in our study, even when using accurate interconnect and gate delay models, computing the optimum spacing and size of the repeaters was virtually as simple as using interconnect and gate switch level models as in [9]. Needless to say that the accuracy in our approach is much higher than in [9].

### 4.4.2 Repeater Insertion Considering Large Load Capacitance

In this section we consider inserting repeaters in long uniform interconnect lines terminated with a large capacitive load using the quadratic programming algorithm presented in section 4.3.2. Note that the general problem of inserting buffers in RC trees for minimum (or for a specific delay) at the net's leaves has long been recognized as NP hard and therefore does not have an analytical solution. At least, to our knowledge, no exact analytical solution has ever been reported.

The proposed insertion method, implemented in MATLAB, has been tested on a typical long ($1cm$) interconnect wire located in the upper metal layer, i.e M6 in the CMOS 0.18µm process. The total resistance $R_{tot}$ in this case is of 1KΩ and the total capacitance $C_{tot}$ is of $1pF$. As illustrated in Figure 4.16, the wire has been segmented into 10 segments of $1mm$ long for each segment, which is half the optimal segment length in the previous sub-section. The size of the driver gate is set to $w_{\text{opt}}$ and no repeater is allowed immediately upstream of the load capacitance $C_{Load}$, i.e at distance of 10 $mm$ from the driving gate (at $S_0$). The resulting insertion sheme is compared with the result of an exhaustive search of optimum location and size of repeaters using HSPICE. The Buffer library used contains 8 symmetrical inverters of different sizes:

$$w = [70, 60, 55, 50, 45, 40, 35, 30] \tag{4.41}$$

where $w$ is a multiple of the minimum sized inverter.

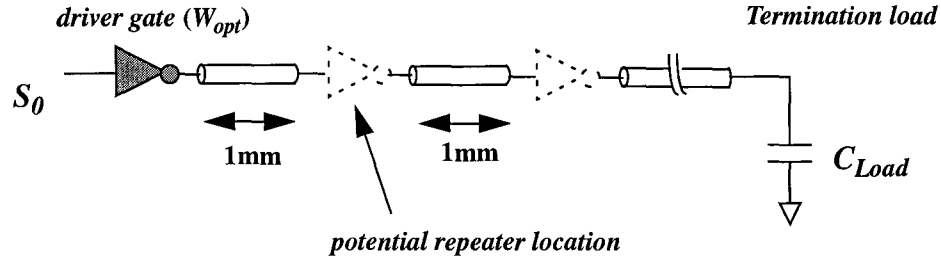



**FIGURE 4.16** Segmented interconnect line for repeater insertion.

In Figure 4.17 we compare the positions and sizes of the inserted repeaters resulting from running our repeater insertion method with those obtained by running an exhaustive search using HSPICE. As can be seen, for large $C_{\text{Load}}$ (Figure 4.17 (a) and (b)), at a certain point on the line, the repeater's size tends to grow progressively as the distance to the load decreases. This particular point becomes closer to the load if the latter is smaller (Figure 4.17 (b)). Note that this progressive increase in repeater size is due to the fact that the problem of inserting repeaters near $C_{\text{Load}}$ can be viewed as the classical problem of driving pure capacitive loads [56]. Note also that our scheme finds the right position for the repeaters, and, nearly, their optimal sizes. Note also that a small discrepancy between their sizes and what is actually computed using HSPICE is observed. It is, however, very important to mention that around the optimal solution, or insertion configuration, the delay characteristics are flat. That is, the discrepancy mentioned above does not result in a large discrepancy in terms of overall interconnect line delay as shown in Table 4.2.

**Table 4.2** Comparison of delay line using the proposed model and HSPICE.

| CLoad (pF) | Line delay (Model) (psec) | Line delay (HSPICE) (psec) |
|---|---|---|
| 4 | 828 | 791 |
| 2 | 631 | 605 |
| 0.12 | 446 | 411 |
| 0.01 | 389 | 367 |

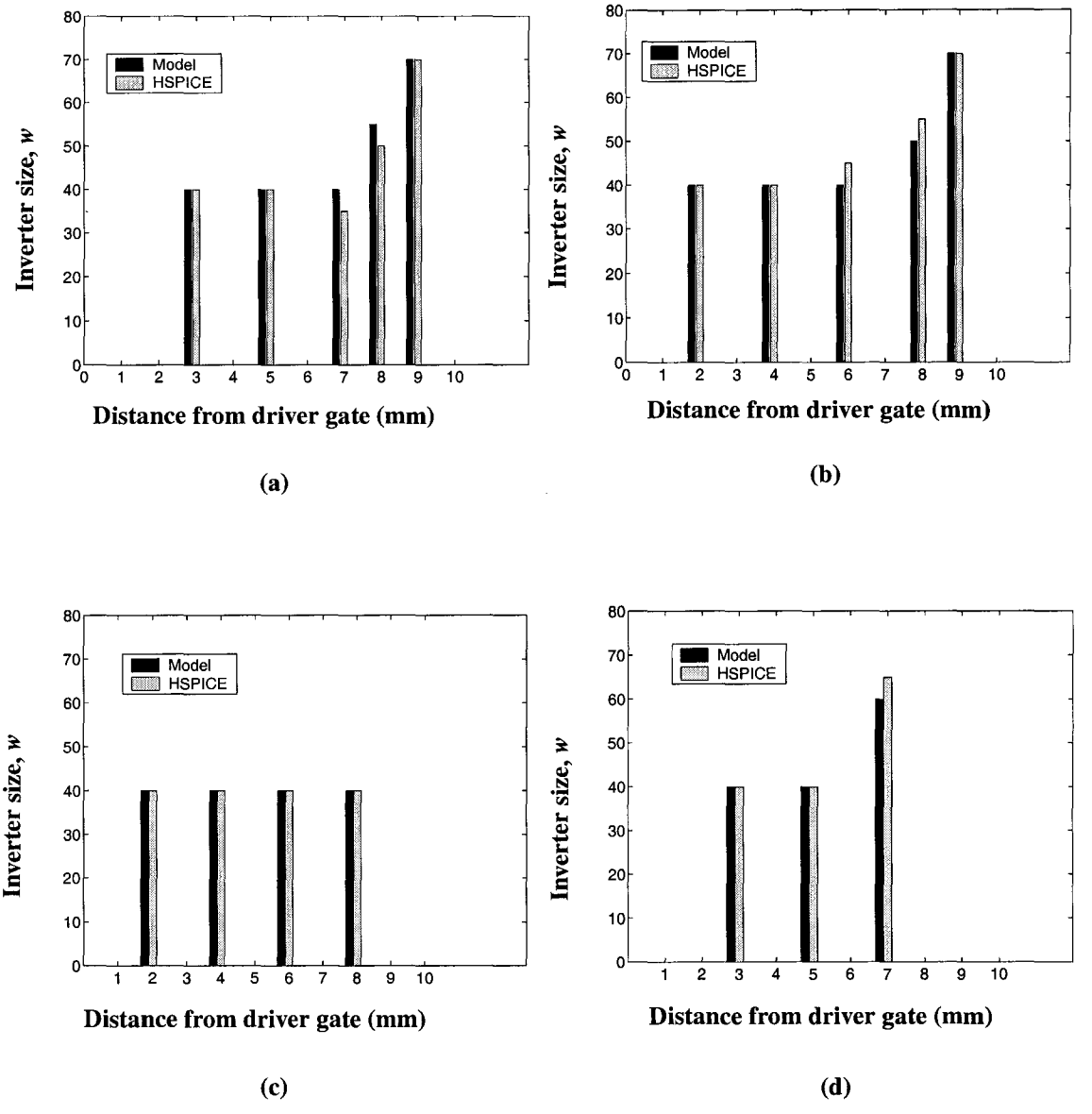



**FIGURE 4.17** Optimal position and size of inserted inverters/repeaters for **(a)** $C_{Load}$ =4pF, **(b)** $C_{Load}$ =2pF, **(c)** $C_{Load}$ =120fF and **(d)** $C_{Load}$ =10fF.

For a load equivalent to the input capacitance of the driving inverter at the source of the interconnect line (Figure 4.17 (c)), the insertion solution reduces to the one described by equations (4. 17) and (4. 33). In other words, the inserted inverters are equisized and equidistant. Finally, if the load $C_{Load}$ is negligible (Figure 4.17 (d)), less repeaters are required to drive the line. Again, our buffering scheme is able to determine the appropriate insertion configuration.

# Chapter 5

# Conclusion and Future Work

In this thesis, a practical yet accurate approach for dealing with the problem of inserting repeaters along on-chip interconnect lines for delay and transition time requirements has been presented. Such an approach builds on the fact that the transition time and the delay at the far-end of an interconnect segment are, respectively, independent and linearly dependent on the driving repeater's input transition time as long as the ratio of the two transition times does not exceed a pre-defined value. Fortunately, practical situations occur, generally, below this value. This corresponds to the situation where transition times at each insertion point are limited in value, which, by any means, is a high performance design measure.

In this context, we derive simple closed form expressions for the optimal repeater spacing and sizing using accurate RC interconnect and CMOS repeater delay models. That is, we use moment matching techniques for computing the RC delays and transition times in addition to an accurate CMOS inverter/repeater delay model that takes into account short channel effects that are prevalent in deep submicron (DSM) technologies. In particular we develop an accurate delay metric based on the first two moments of the impulse response of the interconnect RC circuit. in addition, we present a new empirical ramp approximation definition that takes into consideration the inherent asymmetry of signals in RC networks in DSM technologies. This approximation is based on allowing the ramp's

saturation level to be below the maximum saturation one, i.e $V_{dd}$. A multi-ramp approximation approach would have been generally more accurate, especially for situations where the transition time at the input of the driving gate is faster than the one at the far-end of interconnect segment of interest. However, this would have resulted in a more complicated inverter delay model.

The insertion solution, in terms of spacing and sizing, is similar in form to what has been reported by Bakoglu in [9]. However, unlike the previously reported solutions, the interconnect and CMOS inverter delay models used exhibit errors less than 5% and 8%, respectively, using the CMOS 0.18μm technology. Also, the delay models used were compared with HSPICE where the distributed nature of interconnections is modeled using a β20 circuit approximation (Chapter 2).

Finally, we extended and improved upon Van Ginneken's bottom-up technique [27] for inserting repeaters in signal distribution networks. That is unlike [27], we developed a method that, although based on a bottom-up-like approach, takes into account the transition times at each potential insertion point along the RC line of interest. The techniques has proven to result in an optimal repeater insertion configuration that is very close to what can be obtained running an exhaustive search using HSPICE. In fact, our algorithm yields delay estimates within 8% of exhaustive circuit simulation results.

The future work on this technique could be to incorporated it in a more general multi-sink RC network optimization scheme that takes into account potential branching points. Also the technique could take into consideration, in conjuction with repeater insertion, wire sizing which can be formulated either as an objective or a constraint.

# References


[1] The 1997 National Technology Roadmap for Semiconductors, Semiconductor Industry Association, san Jose, California, 1997.

[2] H. Kapadia and M.Horowitz, "Using Partitioning to Help Convergence in the Standard Cell Design Automation Methodology," DAC 1999.

[3] D.Sylvester and K.Keutzer, "Getting to the Bottom of Deep Submicron," ICCAD, presentation slides, 1998.

[4] D.Sylvester and K.Keutzer, "Getting to the Bottom of Deep Submicron," ICCAD, 1998.

[5] P. E. Gronowski *et al*, "High-Performance Microprocessor Design," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 5, pp 676-686.

[6] V. Tiwari, "Reducing Power in High-Performance Microprocessors," *Design Auto mation Conference*, pp. 726-731, 1998.

[7] P. P. Sotiriadis and A. Chandrakasan, " Low Power Bus Coding Techniques Considering Inter-Wire Capacitances," *IEEE Custom Integrated Circuits Conference*, CICC, pp. 507-510, 2000.

[8] R Ho, R. K. Mai, H. Kapadia, M. Horowitz, "Interconnect Scaling Implication for CAD," *IEEE/ACM Int. Conference on Computer-Aided Design*, 1999, pp. 425-429.

[9] H.B. Bakoglu, Ciruits, Interconnections, and Packaging for VLSI. Reading, MA : Addison-Wesley, 1990.

[10] V. Adler and E. G. Friedman, "Uniform Repeater Insertion in RC Trees," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, Vol.47, No. 10, pp. 1515-1523, October 2000.

[11] S. Dhar and M. A. Franklin, "Optimum Buffer Circuits for Driving Long Uniform Wires," *IEEE J. Solide-State Circuits*, Vol. 26, No. 1, pp.32-40, January 1991.

[12] Chris C. N. Chu and D. F. Wong, " A Quadratic Programming Approach to Simultaneous Buffer Insertion/Sizing and Wire Sizing," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, Vol. 18, No. 6, pp 787-798, June 1999.

[13] C. ALpert and A. devgan, "Wire Segmenting for Improved Buffer Insertion," *Proc. ACM/IEEE Design Automation Conference*, pp. 588-593, 1997.

[14] C. P. Chen, Y. P. Chen, and D. F. Wong, "Optimal Wire-Sizing Formula Under The Elmore Delay Model," *Proc. ACM/IEEE Design Automation Conference*, pp. 487-490, 1996.

[15] C. P. Chen and D. F. Wong, "Optimal Wire-Sizing with Fringing Capacitance Consideration," *Proc. ACM/IEEE design Automation Conference*, pp. 604-607, 1997.

[16] J. Cong and K. S. Leung, "Optimal Wiresizing under the Distributed Elmore Delay Model," *Proc. IEEE International Conf. Computer-Aided Design*, pp. 634-639.

[17] S. S. Sapatnekar, "RC Interconnect Optimization under the Elmore Delay Model," *Proc. ACM/IEEE Design Automation Conference*, pp.387-391, 1994.

[18] J. Cong and L. He, "Optimal Wiresizing for Interconnects with Multiple Sources," *ACM Trans. Design Automation of Electron. Syst.*, Vol. 1, No. 4, October 1994.

[19] C. P. Chen and D. F. Wonf, "A Fast Algorithm for Optimal Wire-Sizing under Elmore Delay Model," *Proc. IEEE International Symposium on Circuits and Systems*, pp. 604-607, 1996.

[20] C. J. Alpert, A. Devgan, S. T. Quay, "Is Wire Tapering Worthwhile?," *IEEE/ACM International Conference on Computer-Aided Design*, pp 430-435, 1999.

[21] N. Menezes, R. Baldick, L. T. Pileggi, "A Sequential Quadratic Programming Approach to Concurrent Gate and Wire Sizing," *Proc. IEEE Int. Conf. Computer-Aided Design*, pp. 144-151, 1995.

[22] J. Cong, C. K. Koh, and K. S. Leung, "Simultaneous Buffer and Wire Sizing for Performance and Power Optimization," *Proc. Int. Symp. Low-Power Electronics and Design*, pp.271-276, August 1996.

[23] J. Cong, L. He, C. K. Koh and P. Madden, "Performance Optimization of VLSI Interconnect Layout," *Integration, the VLSI Journal*, vol. 21, pp. 1-94, 1996.

[24] S. Muddu, E. Sarto, M. Hofmann, A. Bashteen, "Repeater and Interconnect Strategies for High-Performance Physical Designs," *Proc. XI Brazilian Symposium on Integrated Circuit Design*, pp. 226-231, 1998.

[25] A. Nalamalpu and W. Burleson, "Repeater Insertion in Deep Sub-Micron CMOS: Ramp-Based Analytical Model and Placement Sensitivity Analysis," *Proc. International Symposium on Circuits and Systems*, pp. 766-769,2000.

[26] T.sakurai and A.R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," *IEEE J. Solid-State Circuits*, Vol. 23, pp. 548-594, Apr. 1990.

[27] L. P. P. P. van Ginneken, "Buffer Placement in Distributed RC-Tree Network for Minimal Elmore Delay," *Proc. International Symposium on Circuits and Systems*, pp. 865-868, 1990.

[28] C. J. Alpert, A. Devgan, S. T. Quay, "Buffer Insertion With Accurate Gate and Interconnect Delay Computation," *36th Proc., Design Automation Conference*,1999, pp. 479-484.

[29] N. Menezes and C. P. Chen "Spec-Based Repeater Insertion and Wire Sizing for On-chip Interconnect," *12th Int. Conf. VLSI Design*, Jan. 1999, pp. 476-482.

[30] J. Lillis, C. K. Cheng, T. T. Lin, "Optimal and Efficient Buffer Insertion and Wire Sizing," *Proc. Custom Integrated Circuits Conference*, pp.259-262, May 1995.

[31] L. T. Pillage, R. A. Rohrer, "Asymptotic Waveform Evaluation for Timing Analysis," *IEEE Trans. Computer-Aided Design*, Vol. 9, No. 4, pp.352-366, April 1990.

[32] C. Y. Chu, M. A. Horowitz, "Charge-sharing models for switch-level simulation," IEEE Trans. Computer-Aided Design, Vol. CAD-6, No. 6, pp.1053-1061, November 1987.

[33] W.C. Elmore, "The Transient Response of Damped Linear Networks with Practical Regard to Wideband Amplifiers," *J. Appl. Phy.*, Vol. 19, no. 1 pp. 55-93, Jan. 1948.

[34] J. Rubinstein, P. Penfield, M. A. Horowitz, "Signal Delay in RC Tree Networks," *IEEE Trans. Computer-Aided Design*, Vol CAD-2, No. 3, pp.202-211, July 1983.

[35] E. Chiprout, M. Nakhla, "Analysis of Interconnect Networks Using Complex Frequency Hopping (CFH)," *IEEE Trans. Computer-Aided Design*, Vol. 14, No.2, pp.186-199, February 1995.

[36] K. J. Kerns, I. L. Wemple, A. T. Yang, "Stable and Efficient Reductionof Substrate Model Using Congruence Transforms," *IEEE/ACM Int. Conf. Computer-Aided Design*, pp.207-214, November 1995.

[37] J-R. Li, F. Wang, J. K. White, "An efficient Lyapunov Equation-Based Approach for Generating Reduced-Order Models of Interconnect," *Proc. 36th Design Automation Conference*, pp. 1-6, 1999

[38] L.M. Silveira, M. Kamon, I. Elfadel, J. K. White, "A Coordinate-Transformed Arnoldi Algorithm for Generating Guaranteed Stable Reduced-Order Models of RLC Circuits," *IEEE/ACM Int. Conf. Computer-Aided Design*, pp.288-294, 1996.

[39] C.-K. Cheng, J. Lillis, S. Lin, N. Chang, Interconnect Analysis and Synthesis, Wiley Inter-Science

[40] P. R. O'Brien and T. L. Savarino, "Modelig the Driving-Point Characteristic of Resistive Interconnect for Accurate delay Estimation," *IEEE/ACM/ICCAD*, 1989, pp. 512-515.

[41] A. Nabavi-Lishi and N. Rumin, "Inverter Models of CMOS Gates for Supply Current and Delay Evaluation," *IEEE Trans. Computer-Aided Design*, Vol. 13, pp. 1271-1279, October 1994.

[42] A. Embabi, R. Damodaran, "Delay Models for CMOS, BiCMOS, and BiNMOS circuits and their Applications for Timing Simulations," *IEEE Trans. Computer-Aided Design*, Vol.13, pp. 1132-1142, September 1994.

[43] J. Qian, S. Pullela, L. Pillage, "Modeling the Effective Capacitance for the RC Interconnect of CMOS Gates," *IEEE Trans. Computer-Aided Design*, Vol.13, pp. 1526-1535, December 1994.

[44] J. T. Kong and D. Overhauser, "Combining RC-Interconnect Effects with Nonlinear MOS Macromodels," *IEEE Int. Symp. Circuits Systs.*, 1995, pp. 570-573.

[45] M. Hafed, M. Oulmane, N. Rumin, "Delay and Current Estimation in a CMOS Inverter with an RC Load," *IEEE Trans. Computer-Aided Design*, Vol.20, No. 1, pp. 80-89, January 2001.

[46] B. Tutuianu, F. Dartu, and L. Pileggi, "An explicit RC-Circuit Delay Approximation Based on the First Three Moments of the Impulse Response," *Design Automation Conf.*, 1996, pp. 611-616.

[47] T. Lin, E. Acar, L. Pileggi, "h-gamma: an RC Delay Metric Based on a Gamma Distribution Approximation of the Homogeneous Response," *IEEE/ACM Int. Conf. Computer-Aided Design*, pp. 19-25, 1998.

[48] A. B. Kahng, S. Muddu, "Accurate Analytical Delay Models for VLSI Interconnects," Univ. California, Los Angeles, UCLA CS Dept. TR-950034, September 1995.

[49] C. Alpert, A. Devgan, C. K ashyap, "RC Delay Metrics for Performance Optimization," *IEEE Trans. Computer-Aided Design*, Vol. 20, No. 5, pp. 571-582, May 2001.

[50] A. J. Bhavnagarwala, A. Kapoor, J. Meindl, "Generic Models for Interconnect Delay across Arbitrary Wire-tree Networks," *Proceedings of the IEEE 2000 International conference on Interconnect Technology*, pp. 129 -131.

[51] L. Bisdounis, O. Koufopavlou, "Short-Circuit Energy Dissipation Modeling for Submicrometer CMOS Gates," *IEEE Trans. Circuits and Systems-I*, Vol. 47, No. 9, September 2000.

[52] A. Hirata, H. Onodera, K. Tamaru, "Estimation of Propagation Delay Considering Short-Circuit Current for Static CMOS Gates," *IEEE Trans. Circuits and Systems-I*, Vol. 47, No 11, November 1998.

[53] A. Hamoui, N. C. Rumin, " An Analytical Model for Current, Delay, and Power Analysis of Submicron CMOS Logic Circuits," *IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing*, Vol. 45, Issue 10, October 2000.

[54] A. B. Kahng, S. Muddu, E. Sarto, "Interconnect Optimization Strategies for High-Performance VLSI Designs," *12th International Conference on VLSI design*, pp. 464-469, Jan. 1999.

[55] R. Ho, K. W. Mai, M. A. Horowitz, "The Future of Wires," *Proceedings of the IEEE*, Vol. 89, Issue 4. pp. 490-504, April 2001.

[56] J. M. Rabaey, "Digital Integrated Circuits, a Design Perspective," Englewood Cliffs, NJ: Prentice-Hall, 1996.