Propositions

accompanying the thesis

Trade-offs in Buffer Planning

by

Giuseppe Garcea

1. It is anachronistic to focus exclusively on speed optimization of chip interconnect without taking into account the actual design costs.
2. The final result of an optimization problem knows two critical moments that only depend on the experience and the ability of the designer. The first is the translation of a problem in a standard form solvable by computer, the second is the interpretation of the solution.
3. In selecting design points, simplicity prevails over accuracy.
4. Now that process technology has become very difficult to tune, robust design is the only way to reduce costs and increase yield.
5. The true scientist is similar to an artist. Their ideas gush from the same creative source and both aim at producing something inspiring. However, the radical difference between them lies in the fact that the first has the moral duty to produce something reproducible, while the second is only interested in making something unique.
6. Fatalism and sense of mystery, which are intrinsic to our lives, prevent our mind from falling asleep. In fact, if everything is always as foreseen and programmed, there can be no curiosity in life.
7. ”*Laws are like spider webs. If some poor weak creature comes up against them it is caught. But the bigger one can break through and get away*” (Solon, 594 B.C.). In older and in modern society authorities can circumvent the judicial system.
8. Communication anytime anywhere has changed our social relations. It brought faraway people closer together, and moved neighbors farther apart.
9. The business strategy to gain market by sinking the opponents using legal litigation wars will in the end have adverse results for the customers. In the long term, the investment in lawyer bills will consume precious resources that should be used to enhance the quality of the product.
10. The Dutch proverb *”Een goede buur is beter dan een verre vriend” (”A good neighbor is worth more than a distant friend”)*, confuses the concept of friendship with that of pragmatic usage of favors for convenience. In fact *”The true friendship is not slave of time and space, the material distance cannot separate us from our friends!”* (Richard Bach)

These propositions are considered defendable and as such have been approved by the supervisor, Prof. dr. ir. R.H.J.M. Otten.

Stellingen

behorende bij het proefschrift

Trade-offs in Buffer Planning

door

Giuseppe Garcea

1. Het is niet meer van deze tijd om zich bij het ontwerpen van chips alleen op snelheidsoptimalisatie van verbindingen te richten en geen rekening te houden met de actuele ontwerpkosten.

2. Het eindresultaat van een optimalisatieprobleem hangt af van twee kritieke momenten die uitsluitend worden bepaald door de ervaring en bekwaamheid van de ontwerper. Het eerste is de vertaling van een probleem in een standaard vorm die opgelost kan worden met een computer, en het tweede is de interpretatie van de oplossing.

3. Bij het selecteren van een ontwerppunt prevaleert eenvoud boven nauwkeurigheid.

4. Aangezien procestechnologieën steeds moeilijker op elkaar af te stemmen zijn, is robuust ontwerpen de enige manier om kosten te verminderen en opbrengst te vergroten.

5. Een echte wetenschapper lijkt op een artiest. Beiden putten hun ideen uit dezelfde bron van creativiteit en beiden proberen iets inspirerends te maken. Het grote verschil is dat de eerste de morele plicht heeft iets reproduceerbaars te maken, terwijl de tweede alleen genteresseerd is in het maken van iets unieks.

6. Fatalisme en het gevoel van mysterie horen bij het leven en zorgen ervoor dat onze hersenen niet in slaap vallen. Als alles altijd was zoals voorzien en geprogrammeerd, dan zou het leven oninteressant zijn.

7. *"Wetten zijn net spinnenwebben. Als een zwak wezen erin terecht komt, zit hij gevangen. Maar een sterker wezen kan eruit ontsnappen en wegvliegen"* (Solon, 594 B.C.). Zowel in het verleden als in de moderne maatschappij kunnen de autoriteiten het gerecht omzeilen.

8. Communicatie waar dan ook, en op welk tijdstip dan ook, heeft onze sociale contacten veranderd. Het heeft mensen van ver dichter bij elkaar gebracht en mensen van dichtbij van elkaar verwijderd.

9. De business-strategie om marktaandeel te winnen door de concurrentie uit te schakelen d.m.v. een juridisch gevecht, uit zich uiteindelijk niet in een positief resultaat voor de klant. Op de lange termijn zullen de advocaatskosten geput moeten worden uit de bronnen die eigenlijk bedoeld zijn voor het verbeteren van de kwaliteit van het product.

10. Het Nederlandse gezegde: *"Een goede buur is beter dan een verre vriend"*, verwart *vriendschap* met *baat hebben bij gebruik maken van elkaars gunsten*. Inderdaad, *"Echte vriendschap is niet onderhevig aan tijd en plaats. Geografische afstand kan ons niet scheiden van onze vrienden!"*. (Richard Bach)

Deze stellingen worden verdedigbaar geacht en zijn als zodanig goedgekeurd door de promotor, Prof. dr. ir. R.H.J.M. Otten.

Trade-offs in Buffer Planning

Trade-offs in buffer planning

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus, prof.dr.ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen

op maandag 27 june 2005 om 13.00 uur

door

Giuseppe Stefano Garcea

Ingegnere elettronico, Università degli Studi di Firenze, Italië
geboren te Taranto (Italië)

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. R.H.J.M. Otten

Samenstelling promotie commissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. R.H.J.M. Otten,	Technische Universiteit Delft, promotor
Dr. ir. N.P. van der Meijs,	Technische Universiteit Delft, toegevoegd promotor
Prof. dr. C.I.M. Beenakker,	Technische Universiteit Delft
Prof. dr. J.R. Long,	Technische Universiteit Delft
Prof. dr. ir. P.R. Groeneveld,	Technische Universiteit Eindhoven
Prof. dr. ir. D. Stroobandt,	Ghent University
Dr. K. Goossens,	Philips Research

Proefschrift. - ISBN 90-809648-3-2

Front cover: René Magritte *La reproduction interdite* 1937, Museum Boijmans van Beuningen, Rotterdam.
Back cover: Darlisa Riggs, *Balance*, 2002.

Bambino, se trovi l'aquilone della tua fantasia
legalo con l'intelligenza del cuore.
Vedrai sorgere giardini incantati
e tua madre diventerà una pianta
che ti coprirà con le sue foglie.
Fa delle tue mani due bianche colombe
che portino la pace ovunque
e l'ordine delle cose.
Ma prima di imparare a scrivere
guardati nell'acqua del sentimento.

Alda Merini

Child, if you find the kite of your fantasy
bind it to the intelligence of your heart.
You will see the rising of bewitched gardens
and your mother will become a plant
that will cover you with its leaves.
Make of your hands two white doves
that could carry everywhere the peace
and the order of the things.
But before learning to write
look at yourself in the water of the feelings.

Alda Merini

Ai miei genitori
Giuliano e Maria

Contents

Chapter 1

Introduction

Contents

MANY dissertations about digital microelectronics are introduced by a sentence like *"the increasing demand for . . .* ", when referring to complexity, speed, functionality, portability, etc., and support the statement by referring to the extraordinary evolution of the semiconductor industry in the last forty years. For already in 1963 Gordon Moore predicted that integration density on silicon was going to quadruple every four years, and until now the semiconductor industry has succeeded in fulfilling that prophesy, in spite of numerous discussions about when the evolution will deviate from this trend.

The very success of this industry is at the same time its burden: from a prediction Moore's law is, for decennia already, a goal setter, and reaching these goals nowadays is only possible with huge investments in production equipment, tight control of variability, and sophisticated design tools, while the market leaves only a narrow time window to compensate for these costs. This tremendous pressure can be illustrated by considering a product such as the Wi-Fi 802.11. Its market grew at an extreme pace, with many companies fighting for a share, but margins came down sooner than expected, making it very difficult to make a profit.

An average chip design in a $130nm$ process costs about ten million dollars, just to get the first chip out. A mask set for such a design could be 10% of that amount. It may have tens of millions of transistors. For 90-nm processes this may be a tenfold, with costs roughly doubled. These are staggering numbers! But can it be used? Can design methodology cope with those complexities? The gap between what is really manufacturable in silicon and our design capability is growing.

1.1 Performance characteristics

Performance has gotten a broader meaning: no longer is speed the only metric that counts. More an more products have demand for low power consumption and reliability. It is therefore of paramount importance in selecting a design point or identifying optimization problems, to understand that any decision is at best a compromise. The so-called *"design for ..."* approaches are pretty useless when they do not consider other characteristics. Making a circuit as small as possible, makes it often too slow, while optimizing speed may lead to unacceptable levels of power consumption. Modern IC-design is dominated by trade-offs, even when considering just a subproblem in the trajectory.

The smaller feature sizes will require new models to account for the new complex and no longer negligible physical effects. The growing design complexity as well as the complexity to have a reasonable representation of the physical implementation, requires the selection of the right level of abstraction and then making the right decision at that level. The lack of new physical models on one hand as well as the necessity to have right decisions at higher level of abstraction, open really challenging question on how to have a converging flow ready beyond $65nm$. In this context predictions tools become of utmost importance. The common denominator is to build a bridge between technologists and designers by incorporating in the system level prediction tools, also for the relevant physical effects. The decisions made on the basis of predicted metrics will then improve the correspondence with the real implementation.

A serious shortcoming of today's tools used for analysis and design is that everything is considered "deterministic", meaning that they are based on fixed models for devices and wires and do not consider statistical variations in the underlying silicon. Current methodologies can predict best-case, worst-case and nominal parameters sets, but there is a lack of stochastic computing strategies which can model the probabilistic nature of the variations.

We are on the verge of the era of probabilistic design [63, 30, 11]. Performance for example with this statistical approach will be not fixed to the nominal, best or worst case, but will be a statistical distribution. Engineers in optimizing their designs will ask not only whether the circuit has improved its performance, but also which percentage of its realizations presents a performance improvement. Performance estimations then will be more and more similar to the concept of acceptance already presented in the yield estimation.

Of course there is reluctance, both in EDA companies and with designers, to diverge from established design methodologies, but the investigation of trade-offs is gaining interest and statistical techniques are slowly getting accepted to significantly reduce the number of iterations in design flows. Vendors and academics predict that statistic timing analysis will hit the market in 2005 in order to address technologies of $65nm$ and below. That will require a lot of rethinking of the tools (extractions tools, technology files, libraries, place and route tools).

This thesis intends to represent a connection point between the new approach coming in the next years and the traditional "deterministic" approach. We propose an approach that uses traditional static performance prediction models and it estimates also its fluctuations due to random process parameters variations.

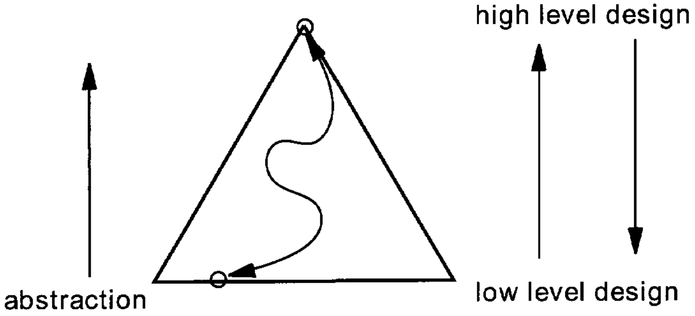


Figure 1.1. Traditional deterministic representation for the high level design abstraction

1.2 Scenarios for high-level design

The complexity of nowadays designs asks for a structured design flow with appropriate levels of abstraction for each step of that flow, such that the right decision and optimization can take place at that level assuring the final design convergence. Figure 1.1 illustrates that decisions taken at high levels using partial information about a system, can lead to several implementations at the low level showing more details. Any of the implementations at the low level on the bottom of the triangle could be reached by taking certain decisions at the various levels, thus exhibiting a specific path coming from a unique decision point taken at highest level represented by the top of the triangle. This traditional way to illustrate the importance of abstraction, perfectly valid in a deterministic framework, is now and for the future technologies much too simplistic.

Technology limitations used to get translated in *design rules* and following these rules enabled designers to make a design safe enough. As technology pushed further towards the physical limits the rules became stricter, constraining designers more and more, to the point where they can no longer comply without ending up with inadequate design, also where technology does support a solution.

Figure 1.2 illustrates the fact that process variations are going to affect the outcome of the design decisions. Indeed process variations will affect the design point chosen at low level design, and if we reconstruct the design path deterministically as in figure 1.1 we see that variability causes unpredictability at high level design.

The result is that decisions at high level design tend to be too conservative under a worst-case variability scenario, or too optimistic when no variations are considered. Moreover, because of the nature of random process parameter variations the same procedures and abstractions that make the decisions in a flow may result in different implementations. The intrinsic uncertainty introduced by process variations, makes that there is no longer a single path connecting high level decision points to real design implementation.

Therefore it is necessary to rethink high-level decision methods to account for process variations in a statistical way. Figure 1.3 shows that a probabilistic term p should be associated with each decision indicating how many systems can be really implemented on silicon with certain specifications. This probability propagates from high to lower level design through different step of decisions, and ultimately represents a figure of yield relative to a certain specification, also denoted as *parametric yield*. Adequate

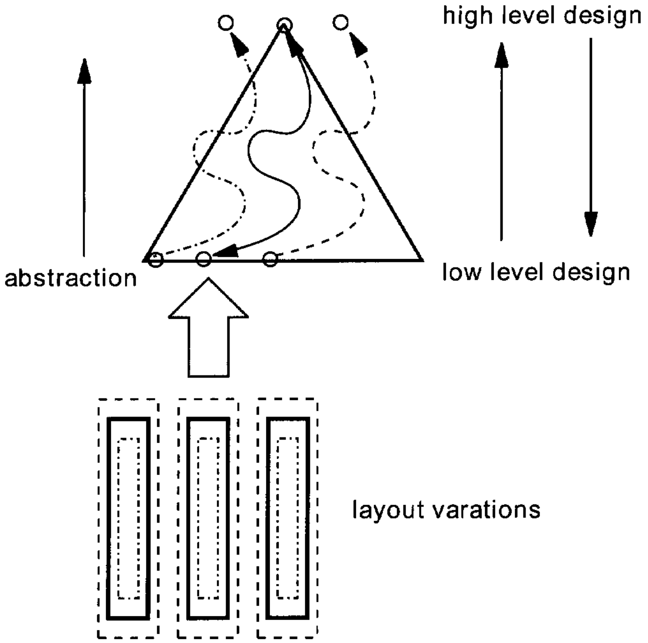


Figure 1.2. Impact of process variations on high level design

statistical modelling of the process parameters, translates the process variability in terms of parametric yield.

1.3 Global wires

Global wires are seen as a key potential bottleneck for achieving high performance in integrated circuits. Of course their cross-sectional dimensions scaled with feature size, shrinking the metal pitches to maintain sufficient routing density. But with a more or less unchanged die size, relatively longer wires start to appear. The delay of these wires is becoming multiples of gate delay, and it is not even possible to cross the chip within the clock periods of today's processors

The development in the dimensions of these wires also decreases the predictability. The high aspect ratio height to width and the decreased spacing make the coupling capacitance between neighboring wires dominant in the total wire capacitance. It is difficult to predict in early stages because neighbors are not known than, and wires may run next to each other over long distances. Real behavior, that is switching activity on wires,

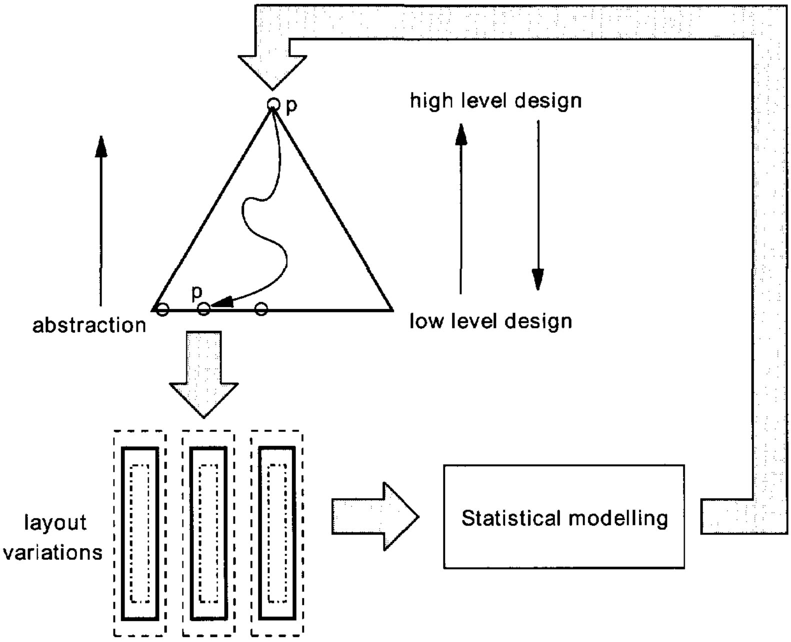


Figure 1.3. Statistical modelling to account for process variations at high level of design abstraction

increases this unpredictability even more. And since long wires will be buffered process variations affecting transistor properties also have an impact on long interconnect.

The performance degradation and growing uncertainty under process variations can be countered by technology improvements as well as by design solutions:

New materials:

- The use of copper wires significantly alleviates electro-migration, which was a significant problem for aluminum processing technologies. Being a lower resistivity material, copper was also expected to reduce the total resistance of the conductors when compared to an equivalent aluminum wire. However, due in part to the roughness induced scattering effects, caused by the intrinsic island-type morphology of the copper-films, its resistivity increases with thickness reduction. Thinner copper-films turn out to have very high resistivity (it is estimated that for a $40nm$

the resistivity value is twice the value obtained for broad lines ($2.2\mu\Omega cm$) [19]). This poses two problems: one is technology problem on how to control surface roughness, and the other is on how to have reasonable predictive models that account for this critical effect.

- Another way to reduce signal delay is by reducing the electrical permittivity, that is the k-value of the intermetal dielectric. The problem here is that low-k materials are really bad heat conductors and this implies complications in reliability.
- There are projects that already are investigating the possibility to have air gaps between the metal lines [19].

Design solutions:

- Buffer insertion can reduce delay by dividing a wire into smaller segments, making interconnect delay linear instead of quadratic in terms of the wire lengths. Optimal segmentation and sizing of buffers in point-to-point has been presented in [7, 21, 3, 15, 49], while buffer insertion in tree structures is subject of [62].
- Wire tapering, despite some theoretical benefits [24], remains problematic when it comes to integration into coherent a routing methodology and in generally it is not applied [4].
- Shielding of signal wires enhances predictability. Layout methodologies as the one proposed in [33] in which each signal has a constant effective capacitance achieved by alternating signal wires with power and ground on higher metal layers, are very effective. It frees up the resources usually taken by supply lines, but it probably will not fully compensate the decrease in wiring density.

Unconventional interconnect:

- 3D integration, that is the use of multiple active layers, for a lot of people seems to alleviate the delay problem [55, 52], without affecting the transistor packing and the chip area. This approach opens new degrees of freedom in system design, placement and routing. There are different techniques to use the different active layers. Several proposals for appropriate technologies were published (the processed wafer bonding [23], silicon epitaxial growth [44] and solid phase crystallization [35]), but film transfer [57] is the most promising. The main concerns are problems with heat dissipation and possibly an increase of coupling effects.

- Active top layers, a special case of 3D integration, were already proposed in [46] where it was stated that the essential interconnect complexity could not be curbed by adding more and more wiring layers. A top layer with optical receivers was suggested so that the clock can be flashed on the chip, reducing it claims on the wiring resource and solving skew problems. Later [48] the author added that the layer can also house the in-line buffers in long interconnect, avoiding via's through all layers to reach buffers in the bulk and the opportunity to reduce the critical delay.
- Photonic interconnects which will substitute polymers to the silicon for global wires. For the heterogeneous integration, wafer bonding will be used [45].

1.4 Buffer planning

With interconnect becoming a dominant factor in the total circuit performance, interconnect-centered synthesis techniques for performance optimization such as appropriate topology construction of the layer stack, buffer sizing, buffer insertion, wire and spacing sizing must be investigated. It has been estimated in [18] that appropriate buffer insertion and sizing and wire sizing can reduce the global interconnect delay (of a line of $2cm$) by a factor of five to six when compared to the use of nominal wire width without buffers in $0.07\mu m$ technology generation from [6].

Not surprisingly buffer planning got a lot of attention lately. We distinguish two main buffer strategies:

- *Uniform buffer insertion*

  Uniform buffer insertion reduces the quadratic increase with length observed in unbuffered wires, to linear if the ensuing segmentation is optimal. The optimum depends on the wire geometry, and is therefore layer dependent. It does not (or almost not) depend on the buffer properties.

  If we assume also that the buffers distributed along an optimally segmented wire are of equal size, then each wire segment between two of those buffers produces the same delay. Buffer planning is completely characterized by the number of buffers (or equivalently the buffer distance) and their size. That is, the result is independent of the length of the interconnect and one can focus on a single segment.

- *Buffer block planning*

  Because of the smaller feature sizes and therefore a relatively larger chip area, the number of nets that need buffering for high performance design will increase for future technology generations. The number of these buffers on a single chip in $50nm$ technology is estimated to be 800.000 according to [17]. These buffers require silicon area and power-ground connections, and may not be placed arbitrarily inside and between existing modules.

  Buffer block insertion during the floorplanning was proposed in [17]. Buffer blocks are constructed by using *feasible regions*. These are the biggest polygons in which buffers can be inserted satisfying the timing constraints for a specific net. One of the consequences of this approach is that the optima achievable with unconstrained buffer planning are out of reach.

This thesis aims at developing a decision making procedure for effective buffer planning in interconnect. The use of these procedures is useful for design engineers/methodologists of high-performance microprocessors and application specific integrated circuits. In particular we will address problems of multi-objective optimization in buffering for a point-to-point wires assuming that the wire geometry is given. At the end we will extend our results to typical unidirectional multiwires structures used in *Network-on-Chip* (NoC) applications.

The nature of the trade-offs we are going to study are: performance versus buffer area, performance versus power, uniform performance versus non uniform buffering in an unconstrained as well as in an area-constrained situation. Those kind of trade-offs assume that there are no process variations. However, process variations do influence the performance prediction and they invalidate the traditional solution for buffering in a deterministic case. This is due to the fact that performance value is not anymore deterministically found, but performance has fluctuations. This implies that the choice of optimal buffering has to account somehow for this performance uncertainty. Our contribution in this field it is to propose a buffering which enhances performance robustness.

It is important to stress that the solutions for these optimization problems are based on analytical models. Whenever possible closed-form expression for optimal buffer size and distance will be derived. Other solutions will be found by including analytical formulations in simple iterative schemes.

1.5 This thesis

The main contributions of this thesis are:

Analysis We propose an analytical model based on elmore approximation and calibrated by using spice simulations. The most important performance metric will be the signal propagation velocity. Impact of process variations on this performance metric will be estimated analytically. We will characterize the statistical behavior of the performance in terms of its expected value and in terms of its standard deviation.

Synthesis We derive buffer strategies in presence of area and power constraints and we show that the solution gives the best performance under those constraints. From this approach complete trade-offs between performance, area and power can be composed. Restrictions on buffer locations will require a resizing of each of the buffers. A complete algorithm for that, which works also for limited buffer area, is described. We propose two techniques for buffer planning in the presence of process variations. The first one is an analytical approach which solves the optimization problem for each of the samples in the variability space. The other one uses the statistical characterization of the performance to derive a yield-centric buffering.

NoC We present how to trade uniform buffering with wire density for maximum throughput of a channel of fixed width.

The organization of this thesis is as follows.
First we establish the important parameters of ic interconnect and derive suitable models for wire segments between buffers in chapter 2. In chapter 3 we summarize the well-known fundamental results using that delay model for the traditional analytical solution for unconstrained buffer insertion. We will explain how to calibrate the model aided by spice simulations. We define a performance metric and present analytical methods

for performance-versus-buffer area and performance-versus-power trade-offs. Moreover we compare the results for area-constrained problem with the power-constrained problem. Power and area constraint problems result in the same buffering provided that the short circuit power term can be neglected. Buffer planning under constrained allocation and the corresponding buffer resizing techniques are the subjects of chapter 4. Also resizing under area limitations is solved.

The chapters that follow discuss the impact of process variations on the performance. In chapter 5 an analytical method to predict performance variability is illustrated. In chapter 6, we propose buffering to enhance performance robustness and a parametric yield-centric buffering method is derived. We will show that optimal buffer planning in the deterministic case is not necessarily optimal in the presence of process variations. Starting from this consideration we maximize yield and we draw a complete trade-off between yield, buffer area and signal speed.

In chapter 7, we extend the results from the preceding chapters to unidirectional bus design, typically used for NoC applications. We show how to find the best wire density and buffering for maximum throughput of buses of fixed width. We also pay attention to parametric yield under process variations.

Chapter 8 illustrates two tools that where developed in this project. In particular, we present a tool for extracting accurate values for wire capacitance per unit length. A complete interactive demo suite which is web-based, has been written and its structure will be briefly explained.

Chapter 2

Delay modelling of IC interconnect

Contents

INTERCONNECT on a chip is realized by a conductive transmission medium surrounded by isolating material. It typically forms traces in layers in order to connect electrically terminals of components. Along such a trace the lateral dimensions are relatively small compared to the longitudinal "length". Several such traces may form topologically different structures, but for signal propagation they are almost exclusively trees. Some of these structures consist of a single trace and we will call them point-to-point connections.

It is safe to assume that the isolation is perfect, that is, there is no unintended conduction from the structures formed by these traces to other such structures or other conducting media. And also the intended return currents are typically not in the substrate. Ergo, a loss conductance will be insignificant, and on-chip interconnect can be characterized as a so-called RLC line with an in-line resistance r, a shunt capacitance c and an in-line inductance λ, all per unit length, until the cross-sectional dimensions become noticeable.

In this chapter we are interested in the delay, and preferably in such a form that we can account for it in the early stages of design. Repeated simulations for all affected interconnect with models full of non-lumped elements is out of the question. In section 2.1 we explore what models might be valid for IC interconnect. The important parameters are length and signal speed. Once the pertinent region is identified, we try to get simple closed-form expressions that are accurate enough to predict the delay to be expected, knowing that detailed layout information is not available. They should support early design decisions, based on optimizing aspects of interconnect.

Depending on whether r is relatively small or large, the propagation behavior will (approximately) be governed by the telegrapher's equations or the diffusion equation. The first case will give rise to a constant velocity of propagation, which we will denote as v_{TEM} since we will approximate this value by that of a (quasi-)TEM line (a line with a transverse electromagnetic field, essentially a 2-dimensional transmission line). The second case will give rise to an RC mode of propagation, with a quadratic dependence of propagation time on distance such that we cannot strictly speak about velocity.

After settling that question in section 2.1 we want to come to approximations of line delay, with and without buffer insertion, a topic we explore further in chapters 3 and 4. The basis of the discussion is the well-known delay formula from [22] and the derivations from [7, 49]. The analysis here is deterministic, that is without taking statistical variations into account. The latter will be taken up in chapter 5 and 6.

2.1 Interconnect parameters

Only when traces are "short" connections can be seen as equipotential areas, or nodes in a lumped network. And what "short" is, depends on how fast the transients are. Depending on length-speed combinations different approximations allow simplifications of the analysis. In figure 2.1 regions for approximations are indicated. For slow transients only the dc-resistance and the capacitance matter. Signal components with radial frequency less than r/λ, where r is the resistance per unit length and λ is the inductance per unit length, are not significantly affected by the inductance. At higher frequencies the inductive reactance becomes noticeable, and when the internal inductance of the conductors becomes comparable to the dc-resistance, a redistribution of current within the conductors takes place and we enter a region with skin effect. Modelling interconnect as a linear circuit then becomes more difficult. The behavior becomes more intricate with increasing frequencies, and when the corresponding wavelength gets close to cross-sectional dimensions of the line waveguide techniques have to be applied to analyze behavior. One of the effects for speed of operation in that region is the dispersion of the rising and falling edges.

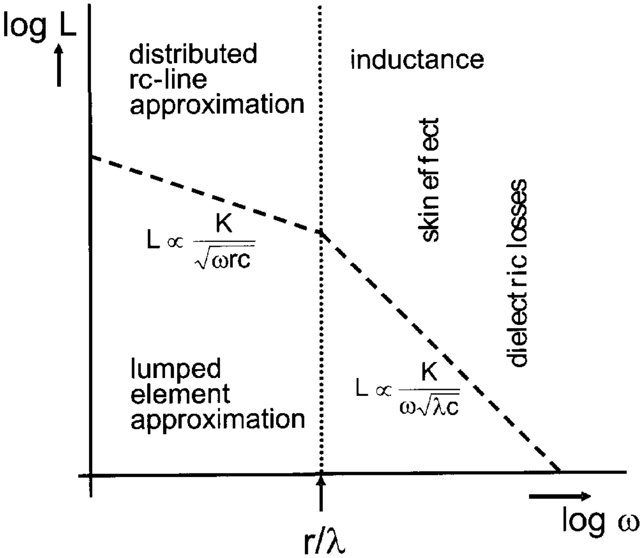

log L
distributed rc-line approximation
inductance
skin effect
dielectric losses
$L \propto \frac{K}{\sqrt{\omega rc}}$
lumped element approximation
$L \propto \frac{K}{\omega\sqrt{\lambda c}}$
r/λ
log ω

Figure 2.1. Approximation regions (after [31])

The value of λ can be estimated from transmission line theory, using the model for the so-called TEM regime. The velocity of propagation along a TEM line is given by $v_{TEM} = \sqrt{1/\mu\epsilon} = c_0/\sqrt{\epsilon_r}$ with μ the permeability of the medium (which in our case is equal to μ_0, the permeability of vacuum), ϵ the permittivity of the medium, c_0 the speed of light in vacuum and ϵ_r the relative permittivity of the medium. For the case of SiO$_2$ with $\epsilon_r = 3.9$ as the dielectric medium, we arrive at $v_{TEM} \approx 1.5c_0 = 1.5 \cdot 10^8 m/s$. For a TEM line we also have $\lambda c = \mu\epsilon = 1/v_{TEM}$ leading to $r/\lambda = rcv_{TEM}^2$. In the sequel of this thesis, we will present actual values of r and c for realistic technologies. Typically, they are around $c = 25fF/\mu m$ and $r = 25m\Omega/\mu m$. Using these values, we arrive at $r/\lambda = 14 \cdot 10^{12} rad/s \approx 2THz$.

This means that only very fast components in the signal will be affected. For digital applications, and delay analysis in particular, these components are of no interest. Although much of the analysis are done under the assumption of square waveforms and step responses (and therefore with fast signal components present), such waveforms cannot be maintained in today's digital circuitry, and concern is rather in passing given voltage levels than in the exact waveform (i.e. distortion is not a real concern).

This analysis suggests that it is justified to model interconnects as RC lines rather then LC(R) lines, which subsequently will lead us to the conclusion of a delay growing with the square of the length of the interconnect (see also equation 2.4). This quadratic dependence is traditionally subdued by segmentation of the line, by inserting buffers. Therefore, apart from the interconnect line itself, we also need to consider the driver and receiver models and their effect on the propagation delay. Without developing suitable models here, we note that the delay can only increase. In particular, the TEM velocity will form an upper bound for the actual effective velocity. Furthermore, we will in the sequel of this thesis develop accurate models for the RC mode of propagation including driver and receiver. Hence, if the effective RC propagation velocity including driver and receiver is small compared to the LC (TEM) propagation velocity, the RC mode of propagation prevails and our conclusion still stands.

Thus, as long as the effective velocities that we can achieve using our modeling and optimization procedures are small compared to v_{TEM} as derived above, $1.5c_0 = 1.5 \cdot 10^8 m/s$, our models are justified. This will be the case for all our results, as can readily be concluded in what follows. The results in section 3.2 and table 3.1 in particular, confirm this as the velocity turns out to be around $0.35 \cdot 10^8 m/s$ for all considered process nodes.

It must be noted, however, that the above analysis does not imply that inductive effects cannot occur for on-chip interconnect systems. These can occur if the total series resistance would be small enough. This will happen with very wide wires and very large drivers, in regions of the design space that are not considered in this thesis.

2.2 Elmore time constants

Delay is mostly analyzed by studying the response at an output to a step-like stimulus at the input. Unfortunately, even after simplifying interconnect to an RC line, there is no analytic solution, except for some special cases, not of interest for delay characterization on complex chips. However, useful approximations can be derived.

If an interconnect line can be modelled as resistance the delay can be measured by the time constant τ of the RC low-pass filter consisting of the line resistance and the capacitive load at the destination. There are several ways of defining this constant, but convenient for the derivations that follows is to set τ equal to the time it takes to let a non-zero voltage over the capacitor drop by a factor e (e is the natural number: 2.718...) when the source end is grounded. We denote time constants thus defined by τ_E. It is easy to verify that for the RC low-pass filter $\tau_E = RC$, and therefore equal to usual time constant for that circuit.

The question we want to answer is what RC low-pass filter has the same τ_E as an interconnect line that cannot be modelled by a single RC section, but at best as an RC chain, if not as an inhomogeneously distributed resistance and capacitance. For short lines terminated by a capacitance considerably larger than the total line capacitance this is possible by simply taking the total resistance and total capacitance. That is how delay was modelled for complex chips even for the larger part of the nineties. Let us assume that it is possible for any line of length up to L by some resistance R_L and some capacitance C_L and determine the time constant τ_E for a line of length $L + \Delta$ where Δ is short and has total resistance R_Δ and total capacitance C_Δ.

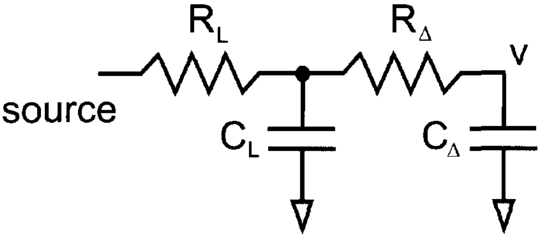


Figure 2.2. Model of a line of length L+Δ.

With the line of length L modelled as a section with R_L and C_L, and the additional piece as a section with R_Δ and C_Δ an RC chain of two sections is obtained (figure 2.2). The voltage v over C_Δ has to obey

$$\frac{d^2v}{dt^2} + \frac{R_LC_L + R_LC_\Delta + R_\Delta C_\Delta}{R_LC_LR_\Delta C_\Delta}\frac{dv}{dt} + \frac{v}{R_LC_LR_\Delta C_\Delta} = 0$$

with solution

$$v = Ae^{-\alpha t} + Be^{-\beta t}$$

if

$$\alpha + \beta = \frac{R_LC_L + R_LC_\Delta + R_\Delta C_\Delta}{R_LC_LR_\Delta C_\Delta} \qquad \alpha\beta = \frac{1}{R_LC_LR_\Delta C_\Delta}.$$

$A + B$ is the voltage at $t = 0$, the time when we ground the source end. Further, because the derivative of v with respect to time has to be 0 we also have the initial condition $A\alpha + B\beta = 0$. This yields

$$v(t) = (A+B)\left(\frac{\beta}{\beta-\alpha}\,e^{-\alpha t} - \frac{\alpha}{\beta-\alpha}\,e^{-\beta t}\right)$$

and shows that τ_E has to satisfy by definition

$$\frac{\beta}{\beta-\alpha}\,e^{-\alpha\tau_E} - \frac{\alpha}{\beta-\alpha}\,e^{-\beta\tau_E} = e^{-1}. \tag{2.1}$$

α and β are the solutions of the characteristic equation

$$p^2 + \frac{R_L C_L + R_L C_\Delta + R_\Delta C_\Delta}{R_L C_L R_\Delta C_\Delta}p + \frac{v}{R_L C_L R_\Delta C_\Delta} = 0$$

which can also be written as

$$p^2 + \left(\frac{1}{R_L C_L} + \frac{1}{R_\Delta C_\Delta}\right)p + \frac{1}{R_L C_L}\frac{1}{R_\Delta C_\Delta} = -\frac{1}{R_\Delta C_L}p$$

or

$$\left(p + \frac{1}{R_L C_L}\right)\left(p + \frac{1}{R_\Delta C_\Delta}\right) = -\frac{1}{R_\Delta C_L}p.$$

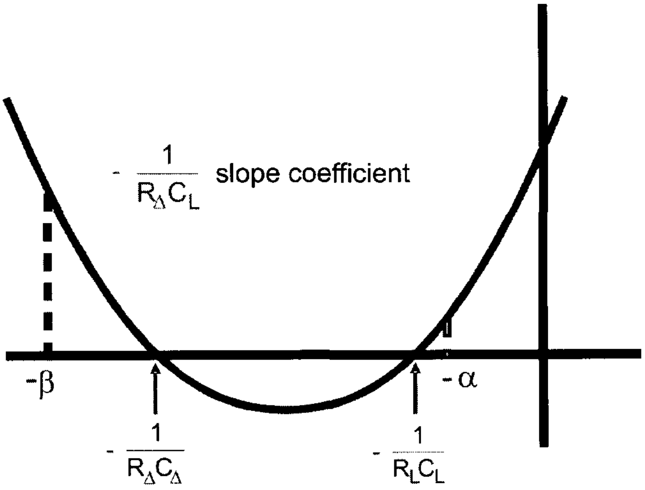

$-\frac{1}{R_\Delta C_L}$ slope coefficient
-β
-α
$-\frac{1}{R_\Delta C_\Delta}$
$-\frac{1}{R_L C_L}$

Figure 2.3. Solutions of the characteristic equations.

A glance at figure 2.3 shows that, since $R_L C_L \gg R_\Delta C_\Delta$, we have $\beta \gg \alpha$. Going back to equation 2.1 we see that in that case

$$\frac{\beta}{\beta-\alpha}\,e^{-\alpha\tau_E} \approx e^{-1}$$

and from that follows

$$\alpha\tau_E \approx 1 - \ln\left(1 - \frac{\alpha}{\beta}\right) \approx 1 + \frac{\alpha}{\beta}.$$

Accuracy is better than 1%, even when $R_L = R_\Delta$ and $C_L = C_\Delta$.

This is in essence the formula first derived, from a different insight, by W.C. Elmore in 1948 [22], because

$$\tau_E \approx \frac{\alpha + \beta}{\alpha\beta} = (R_L + R_\Delta)C_\Delta + R_L C_L.$$

We therefore call τ_E the *elmore time constant.* The usual interpretation is that the elmore time constant of an RC chain equals the sum of the time constants of single RC sections, obtained by combining every capacitance with the total resistance that separates it from the source. It has been demonstrated for two sections above, and the general result follows by induction. So, for an an RC chain with n sections, indexed from the source, we have

$$\tau_E = \sum_{i=1}^{n} \left[C_i \sum_{j=1}^{i} R_j \right] \tag{2.2}$$

Elmore delay estimation applies to well-damped circuits composed of any number of series resistances and shunt capacitances. It does not apply (in its original form) to circuits involving inductance, resonance, overshoot, or any form of poorly damped or non-monotonic behavior.

When the input rise time significantly exceeds the response time of the circuit (that is, τ_E is much less than the time the input takes to reach 90% of the swing), the output waveform tends to look like a strict delay of the input waveform and the elmore delay becomes exact. When the input rise time is shorter than τ_E, the output waveform becomes distorted and the elmore delay serves as an estimate of delay.

To apply the elmore delay concept to a tree-structured circuit, one must identify all the resistors R_i that lie along a path from the input to the output under analysis. Then, for each series resistor R_i one must identify the total downstream capacitance C_i charged by that resistance and sum all these R_iC_i time constants. The caption of figure 2.4 gives the the elmore estimate for the destination with index 9 in the schematic of that figure.

In practice, the elmore delay estimate may be successfully applied to well-damped RC networks, but not for lines having appreciable amounts of inductance or for non-monotonic inputs. In [28] an attempt is described to extend elmore delay estimation to account for line inductance.

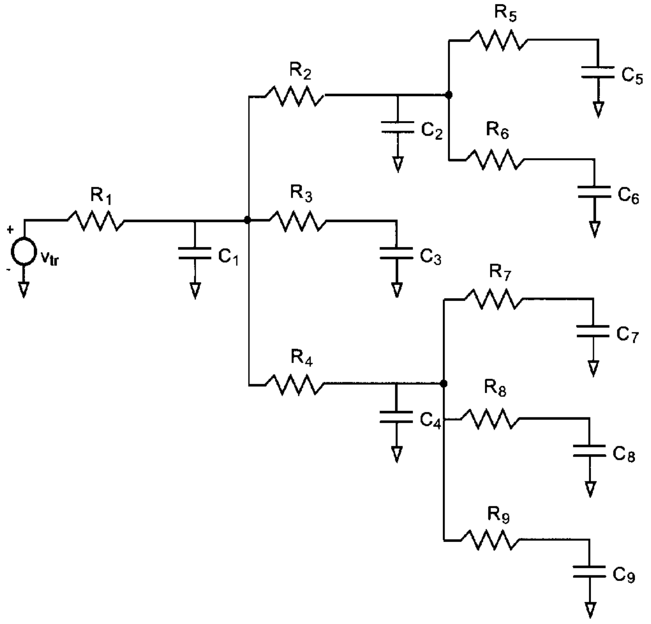


Figure 2.4. $\tau_{E,9} = R_1 \sum_{i=1}^{9} C_i + R_4 \left(C_4 + \sum_{i=7}^{9} C_i \right) + R_9 C_9$.

2.3 Lumped and distributed RC configurations

Cases where all sections are identical, except possibly the last one with capacitance C_d, are of special interest. In that case

$$\tau_E(n) = \frac{1}{2} n(n-1) RC + nRC_d. \tag{2.3}$$

If $R = rL/n$ and $C = cL/n$ are chosen, it models a homogeneous line, and taking the limit gives the famous result

$$\lim_{n \to \infty} \tau_E(n) = \frac{1}{2} rcL^2 + rLC_d, \tag{2.4}$$

showing that the delay of an homogeneous line increases quadratically with its length.

The elmore time constant was chosen to be the time for a step response to complete the $1 - e^{-1}$ part of the swing, that is about 63%. That is not always convenient, especially if we want to cascade interconnection structures. Figure 2.5 compares the step responses of an RC low-pass filter and a homogeneously distributed line of the same total resistance and capacitance. It also marks the points where 50%, 63.2% and 90% of the swing

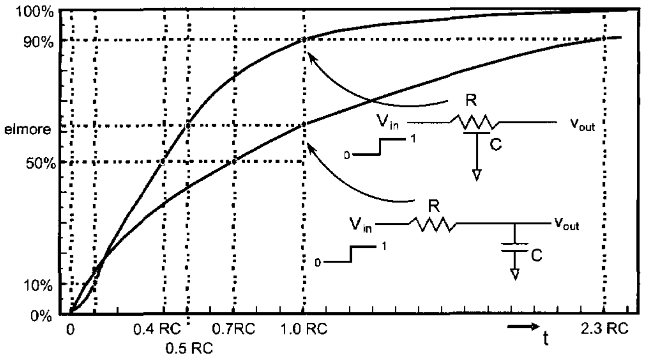


Figure 2.5. Step responses of an RC section and a distributed line after [7].

is completed. Obviously, 63.2% is reached after $1.0RC$ for the section and after $0.5RC$ for the distributed line. It is the elmore time constant for these configurations. The 50% marks are at $t = 0.69RC$ and $t = 0.38RC$ respectively, while 90% requires $2.3RC$ and $1.0RC$ respectively. In general we will use aRC for distributed lines and bRC for lumped RC section, and calibrate the values for the situation at hand as for example in section 3.2.

2.4 Modelling in-line buffers

Elmore time constants are also useful for obtaining computationally convenient models for in-line buffers (or repeaters), as Sakurai showed in [53]. His buffer model is a simple inverter driving a line of length l, terminated with C. Although different delay models have been proposed in the literature, this one still seems the one which can guarantee a good trade-off between accuracy and simplicity. Figure 2.6 shows the model for a single segment.

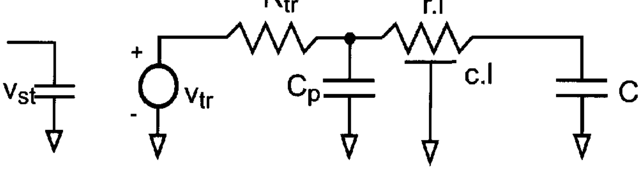


Figure 2.6. Generic restoring buffer model

The repeater driving the line is represented as a voltage source controlled by the voltage v_{st} at the input capacitance. This voltage source switches instantaneously when the fraction denoted by x, $0 \leq x \leq 1$ of the total swing has been reached. The switching at the voltage source is a perfect step. R_{tr} defines the equivalent transistor resistance which represents the sum of the channel resistance, the via resistances, and the source and drain resistance, which all scale inversely proportional with the transistor size. C is the capacitance at the end of the line, collecting the input capacitances of gates connected there. C_p stands for the drain capacitances in case of a CMOS-inverter. It scales with the size of the repeater.

Using the elmore formula (equation 2.2) yields the delay:

$$\tau = b(x)R_{tr}(C + C_p) + b(x)(R_{tr}c + rC) * l + a(x) * rcl^2 \qquad (2.5)$$

where a and b now also depend on the switching model being employed. Thus, this RC representation offers simple closed-form approximation for the delay of the circuit very useful in optimization problems of chapters 3 and 4. For another benefit provided by this kind of approximation is interconnect delay can be easily incorporated. The simplest metric of performance evaluation for IC interconnect is when interconnects are modelled in terms of RC circuits. A complete RC network will be then sufficient to describe repeater and line delay in combination.

Chapter 3

Uniform buffer planning

Contents

A WELL known technique to reduce the performance degradation for global wires consists in using repeaters inserted along their length. The segmentation of these wires due to those repeaters is done ideally by regenerating the signal once the delay along the wire approaches the gate delay. The general problem here is to obtain the type of buffers, their number, their size and their positions in the line such that the delay is minimized. The optimal segmentation as well as the optimal buffer size can be determined for a given technology and for a defined interconnect architecture. [7] and [49] present analytic, closed-form formulas for optimal buffering without constraints on the area needed for repeaters.

In this chapter we extend these results by constraining the buffer area and finding the best performance under constraint. These results enable us to draw the full trade-off curve for buffer area versus performance in global interconnects. Such curves show among other things that sacrificing 15% of the performance is paid off with a 70% reduction of area. The analytic results closely match spice simulation results.

The extension is relevant, because designers must always seek trade-offs between performance characteristics and because unconstrained buffering might not be sustainable into the future. In [47] it has been shown that, accepting some assumptions for the rent exponent in statistical models for wire length distribution, unconstrained optimal buffering might require up to 80% of the total silicon area in imminent semiconductor technologies. Our current results show that this might be reduced significantly with only a slight performance degradation.

We will investigate in this chapter a deterministic uniform buffer planning, that is, we assume that there are no restrictions in placing the buffers, such that a perfect uniform distance between buffers is always possible. This provides a simplification to our problem, in part because it can become independent of the total length of the wire to be buffered.

This chapter has the following structure: after summarizing the results on unconstrained repeater insertion, we use them to devise a method for calibrating the parameters of the segment model of section 2.4. Section 3.3 is about reducing area at the cost of speed, while section 3.4 brings power into play, leading to a complete power-area-delay trade-off.

3.1 Optimal unconstrained repeater insertion

The repeaters, both driving and loading a segment, can be characterized by a linear output resistance R_{tr}(defined as the equivalent transistor resistance), input gate capacitance C and parasitic output capacitance C_p, mainly originating from the drain capacitance. The delay is given by:

$$\tau = b(x)R_{tr}(C + C_p) + b(x)(R_{tr}c + rC) * l + a(x) * rcl^2. \tag{3.1}$$

Defining the buffer size as w_b, its relation with the buffer parameters will be $C = c_0 w_b$ and $R_{tr} = r_0/w_b$ and $C_p = c_p w_b$. Moreover we will define a device dependent parameter τ_0 as $\tau_0 = R_{tr}(C + C_p) = r_0(c_0 + c_p)$.

If a wire of total length L is regularly segmented into n cells such that $n = L/l$, then the total delay along the wire will be:

$$T = n\tau = L\left(\frac{b\tau_0}{l} + b(\frac{r_0}{w_b}c + rc_0 w_b) + a\tau_i l\right) \tag{3.2}$$

A remarkable property of equation (3.2) is that it only contains terms with either l (the segment length) or w_b (the buffer size). Therefore, the optimum segment length (obtained via differentiation to l) and buffer size (obtained via differentiation to w_b) are independent! In particular, both the l and w_b dependencies are of the form $f(x) = A/x + Bx$, with x_{min} equal to $\sqrt{A/B}$.

Thus, taking the derivative of equation (3.2) with respect to l and setting it to zero shows that the insertion of repeaters at a fixed optimal distance $l{=}l_{crit}$ is found from the expression $b\tau_0/l^2 = a * \tau_i$ and it will reduce the delay on wire length to a linear function. This optimal segment length will be

$$l_{crit} = \sqrt{\frac{\frac{b}{a}\tau_0}{\tau_i}} \qquad (3.3)$$

The optimal buffer size $w_b = w_{opt}$ can be obtained from taking the derivative of (3.2) with respect to w_b and setting it to zero:

$$w_{opt} = \sqrt{\frac{r_0}{c_0}}\sqrt{\frac{c}{r}} \qquad (3.4)$$

Equations (3.3) and (3.4) clearly show that l_{crit} and w_{opt} are independent. An optimal segment will be defined as a segment with $l = l_{crit}$ and $w_b = w_{opt}$ and its delay τ_{crit} will not be dependent on the geometry of the wire but only on the technology parameters. We will show in the next section a method to find these parameters. The value of τ_{crit}, the delay of one segment of an optimally buffered line, can be expressed by:

$$\tau_{crit} = 2b\tau_0\left(1 + \sqrt{\frac{b}{a}\frac{r_0c_0}{\tau_0}}\right) \qquad (3.5)$$

and is a characteristic for a given technology.

The delay of an optimal segmented line of length L with optimal buffer size will then be

$$T = \frac{L}{l_{crit}}\tau_{crit} = L\tau_{crit}\sqrt{\frac{a}{b}\frac{\tau_i}{\tau_0}} \qquad (3.6)$$

In the rest of this thesis, it will prove useful to work with the reciprocal velocity of signals along the line. Reciprocal velocity is more convenient than velocity because it is additive. From (3.2) we can derive

$$v^{-1} = \frac{T}{L} = \frac{\tau_0}{l} + a\tau_i l + \frac{br_0c}{w_b} + brc_0w_b \qquad (3.7)$$

This reciprocal velocity is a good characteristic for the performance, and it is independent of the total line length L (and the delay along equidistantly buffered lines is linear in L). In particular we will in section 3.3 minimize v^{-1} under area constraints.

For completeness, we also give the unconstrained minimum reciprocal velocity v^{-1}, to be derived from (3.6):

$$v_{min}^{-1} = \frac{\tau_{crit}}{l_{crit}} = \tau_{crit}\sqrt{\frac{a}{b}\frac{\tau_i}{\tau_0}} \tag{3.8}$$

3.2 Calibrating the in-line buffer delay model

The delay of a single optimally buffered segment, given in (3.5) does not depend on the wire geometry, but exclusively on the technology parameters. Of course, if r_0 includes also the via resistance, this independence is not completely true. However, it is shown in [49] that the effect of via resistance remains negligible. Another characteristic of the uniform line segmentation is that the optimal distance between buffers is independent of the buffer size. This implies that l_{crit} and w_{opt} can be found using two decoupled interactive simulations loops.

The circuit that we are simulating using hspice is a ring oscillator with an odd number of identical inverters connected by uniform wires. The wires are modelled with a certain number of lumped π-models. A ring oscillator, after a transient, has the property of oscillating with a certain a natural frequency. This frequency is a function of the number of buffers, but, more importantly, it is a function of the wire length and inverter size. The delay of a single segment τ is related to the natural frequency of the ring oscillator $1/T_{ring}$ by the relation $\frac{T_{ring}}{2} = N\tau$.

Our objective is to find the values of technology parameters by simulating a ring oscillator, and using reverse engineering to find r_0 and c_0, which are properties of the process technology, through the analytical model presented in the previous section.

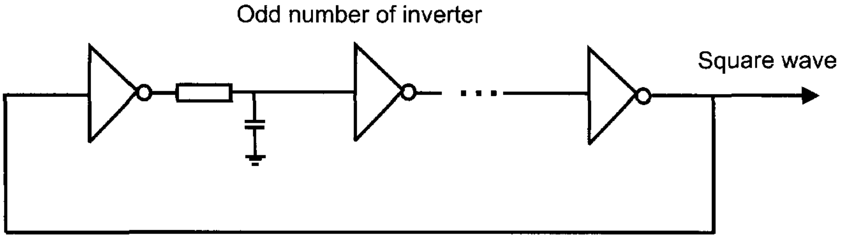


Figure 3.1. Ring oscillator with RC interconnect parasitics

First, l_{crit} is derived by changing the wire length of each of the segments in the ring oscillator and assuming that the buffer sizes are sufficiently large so that the contribution of the gate delay is negligible when compared to the wire delay. The optimal value l_{crit} is found once τ/l is minimum. Once l_{crit} is obtained, the optimal size of the inverters w_{opt} is found by simulating ring oscillators with inverters located at distance l_{crit}. Again the minimum $v^{-1} = \tau/l$ gives the actual value of optimal inverter size w_{opt}. Finally for the value of l_{crit} and w_{opt} we can compute τ_{crit} from simulations. The value of $\tau_0 = \tau_{l=0}/b$ is retrieved by imposing length 0 on the wires between the inverters in the ring oscillator.

The inverter parameters r_0 and c_0 are found using the equations from the previous sections. The results can be summarized as follows:

$$r_0 = \sqrt{\frac{r}{c}\frac{\tau_0}{(1+\alpha_p)}} * w_{opt} \tag{3.9}$$

and

$$c_0 = \sqrt{\frac{c}{r}\frac{\tau_0}{(1+\alpha_p)}} * \frac{1}{w_{opt}} \tag{3.10}$$

where $\alpha_p = c_p/c_0$. It can be considered constant through technology scaling. The explicit expression for α_p as function of the quantities derived from simulation is:

$$\alpha_p = \frac{1}{a\tau_0\left(\frac{\tau_{crit}}{2b\tau_0} - 1\right)^2} - 1 \tag{3.11}$$

Table 3.1 summarizes for different technology nodes, the value of the device parameters used in elmore delay model calibrated with the spice simulations. Wire geometry values are taken from [1], wire capacitance values are obtained by using the analytical model presented in [65, 10] and the spice technology files used are available in [10].

Experimentally, it can be confirmed that w_{opt} is only a weak function of l and that l_{crit} is only a weak function of w_{opt}. This confirms the validity of the model represented by equation (3.7).

3.3 Area-delay trade-off

Area can be reduced by increasing the segment length l over l_{crit} and/or reducing the buffer size w_b to values smaller than w_{opt}, where l_{crit} and w_{opt} are defined in (3.3) and (3.4) respectively. It is indeed to be expected that some combination of both measures gives optimal results, where optimality

Process (μ)		0.18	0.13	0.10	0.07
V_{DD} (V)		1.8	1.5	1.2	0.9
L_{eff} (nm)		100	70	50	35
t_{ox} (Å)		45	35	30	20
V_{TH} (V)		0.37	0.27	0.22	0.16
number of levels		6	7	8	9
Global wires parameters	h (μ)	1.2	1.15	1.15	1.15
	w (μ)	0.7	0.5	0.4	0.45
	s (μ)	0.7	0.5	0.4	0.35
	t_{ox} (μ)	1	0.7	0.6	0.5
	ρ	2.2	2.2	2.2	2.2
	κ	3.5	3.2	2.8	2.2
τ_0 (ps)		17.8	14.6	11.6	10.3
τ_{crit} (ps)		78.9	69.6	46.9	44.6
l_{crit} (mm)		2.9	2.2	1.5	1.7
$w_{opt}(\mu)$		60	50	40	40
r_o $(k\Omega \cdot \mu m)$		2.9	2.6	1.94	1.8
c_o $(fF/\mu m)$		8.09	7.84	7.45	7.1
$\alpha_p = c_p/c_0$		0.12	0.15	0.17	0.14

Table 3.1. Calibrated parameters

is defined as the best possible performance for a given total buffer area. It has to be noted again that we only consider uniformly segmented lines.

So, let $l = \alpha l_{crit}$ and $w_b = \beta w_{opt}$, where l and w_b are our area-constrained segment length and buffer size respectively. We will consider $\alpha \geq 1$ and $0 < \beta \leq 1$, assuming that the optimum always lies in that range. They occur when $l > l_{crit}$ and $w_b < w_{opt}$. Also, let $\gamma = A/A_{opt}$, that is, γ $(0 < \gamma \leq 1))$ is the ratio of buffer area in the area-constrained case to that in the area-unconstrained case. We will in sequel refer to γ as the normalized total buffer area, to β as the normalized single buffer area and to α as the normalized buffer distance.

The area A for buffering a line of length L, with segments of length l, is given by the number of segments times the area of one buffer, denoted by A_b. This area is proportional to w_b, say $A_b = \frac{L}{l}.L_{\text{eff}}.w_b$. The total buffer area normalized to the area for optimal buffering, denoted by γ, is given by:

$$\gamma = \frac{A}{A_{opt}} = \frac{L.L_{\text{eff}}.w_b}{l}.\frac{l_{crit}}{L.L_{\text{eff}}.w_{opt}} = \frac{\beta}{\alpha} \tag{3.12}$$

Substituting (3.3) and (3.4) yields

$$\gamma = \frac{\beta}{\alpha} = \sqrt{\frac{b}{a}\frac{\tau_0 c_0}{r_0}\frac{1}{c}\frac{w_b}{l}} \tag{3.13}$$

Because we are considering uniformly segmented lines, this expression is independent of L. The area expressed by γ is then proportional to the total area of the complete line if the number of segments is large enough such that it can be approximated by the number of buffers. (Always, the number of buffers is one plus the number of segments.) This assumption sensibly simplifies our analytical approach and gives a worst-case solution. That is, the simulated performance for a certain area is always better than the performance predicted by the model. Later in section 3.3, we will show that if we compensate for this difference, the simulated and the predicted results show excellent agreement.

Now, substituting the values $w_b = \beta w_{opt} = \gamma\alpha w_{opt}$ and l in the reciprocal velocity, equation (3.7), we obtain

$$v^{-1}(\alpha, \gamma) = \frac{\sqrt{\tau_i}(\sqrt{ab\tau_0} + \frac{b\sqrt{r_0 c_0}}{\gamma})}{\alpha} + \sqrt{\tau_i}(\sqrt{ab\tau_0} + b\sqrt{r_0 c_0}\gamma)\alpha \tag{3.14}$$

This equation is a function of α, with the same structure of dependency as (3.7) on both w_b and l. It reaches a minimum when the derivative w.r.t. α

equals zero, resulting in:

$$\alpha_{opt}(\gamma) = \sqrt{\frac{\sqrt{ab\tau_0} + \frac{b\sqrt{r_0c_0}}{\gamma}}{\sqrt{ab\tau_0} + b\sqrt{r_0c_0}\gamma}} \tag{3.15}$$

and the buffer size ratio becomes:

$$\beta_{opt}(\gamma) = \alpha_{opt}(\gamma).\gamma \tag{3.16}$$

Note that we see α as a function of γ, that is, for each value of γ, equation (3.15) gives the value of α that minimizes the reciprocal velocity (and thus maximizes the performance) for that γ. Because γ is defined as the normalized buffer area, equations (3.15) and (3.16) present the optimal trade-off between segment length and buffer size versus area. These equations are our principle sizing results for area constrained repeater insertion.

The key to this behavior is in equation (3.12), where it is shown that $\gamma = \beta/\alpha$. Although both α and β can be chosen independently, this equation establishes an elegant and useful connection to the ratio of the total buffer area, γ, such that the results (3.15) and (3.16) will follow.

If we substitute $\alpha_{opt}(\gamma)$ into (3.12) and (3.14) and we take the ratio, we will find the normalized performance:

$$\frac{v^{-1}_{min}(\gamma)}{v^{-1}_{min}} = \frac{\sqrt{(\sqrt{ab\tau_0} + \frac{b\sqrt{r_0c_0}}{\gamma})(\sqrt{ab\tau_0} + b\sqrt{r_0c_0}\gamma)}}{\sqrt{ab\tau_0} + b\sqrt{r_0c_0}} \tag{3.17}$$

This equation, unlike the equation for v^{-1}_{min} alone, but like the evaluation for τ_{crit}, is independent of the wire geometry. We will use this fact in section 7.2 on throughput driven repeater insertion.

The delay of a segment of the wire can be expressed by:

$$\tau(\gamma) = 2b\tau_0\left(1 + \sqrt{\frac{b}{a}\frac{r_0c_0}{\tau_0}\frac{1}{\gamma}}\right) \tag{3.18}$$

which only differs form (3.5) in the $1/\gamma$ term.

Figure 3.2 plots the normalized performance versus the normalized area. The solid line, labelled with $A(\gamma)$, is the result of the equation (3.17), where the normalized area is given by $\gamma = \beta/\alpha$. This graph really confirms that the ultimate performance is expensive in terms of area. For example, 50% of the total buffer area is needed for 95% of the absolute maximum speed and 20% of total area for 75% of the speed and only 10% of the total

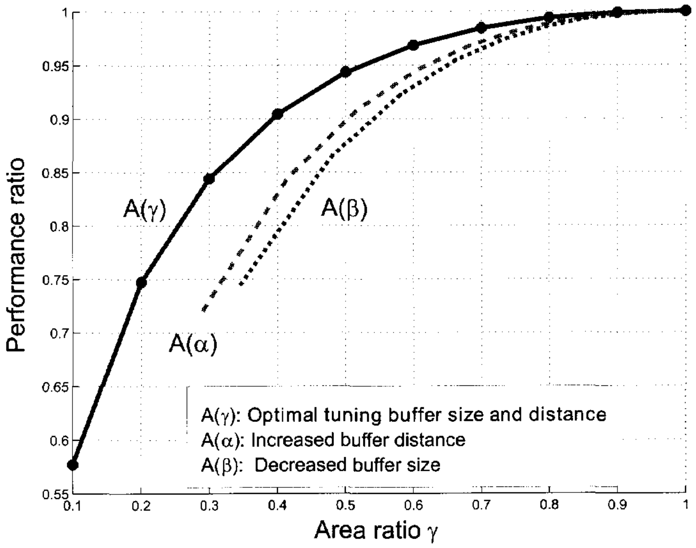


Figure 3.2. Normalized relation between total buffer area and performance

area for 57% of the speed. It follows from (3.17) that the curve presented in figure 3.2 is independent from the geometry of the wire and partly dependent on the technology through the ratio c_p/c_0. However this ratio tends to remain pretty constant through technology scaling and then the figure 3.2 tends to be same.

For comparison, the graph also shows two other area reduction scenarios. The dashed line, labelled with $A(\alpha)$, corresponds to the case with optimal buffer size but increased segment length and the dotted line, labelled with $A(\beta)$, corresponds to the case with optimal segment length but reduced buffer size.

Like equation (3.12), we can derive

$$\frac{A(\alpha)}{A_{opt}} = \frac{L.L_{\text{eff}}.w_{opt}}{l} . \frac{l_{crit}}{L.L_{\text{eff}}.w_{opt}} = \frac{l_{crit}}{l} = \frac{1}{\alpha} \tag{3.19}$$

and

$$\frac{A(\beta)}{A_{opt}} = \frac{L.L_{\text{eff}}.w_b}{l_{crit}} . \frac{l_{crit}}{L.L_{\text{eff}}.w_{opt}} = \frac{w_b}{w_{opt}} = \beta \tag{3.20}$$

and the normalized performance follows analogously to (3.17). We can conclude that the $A(\gamma)$ solution indeed presents a better trade-off than the $A(\alpha)$ and $A(\beta)$ solutions.

Figure 3.3 shows the optimal trade-off relation between buffer size and segment length, as derived from (3.12). The point $(1,1)$ corresponds to the unconstrained optimum, increasing the segment length goes together with reducing the buffer size. The dots along the curve correspond to $\gamma = 1, 0.9, \ldots, 0.1$

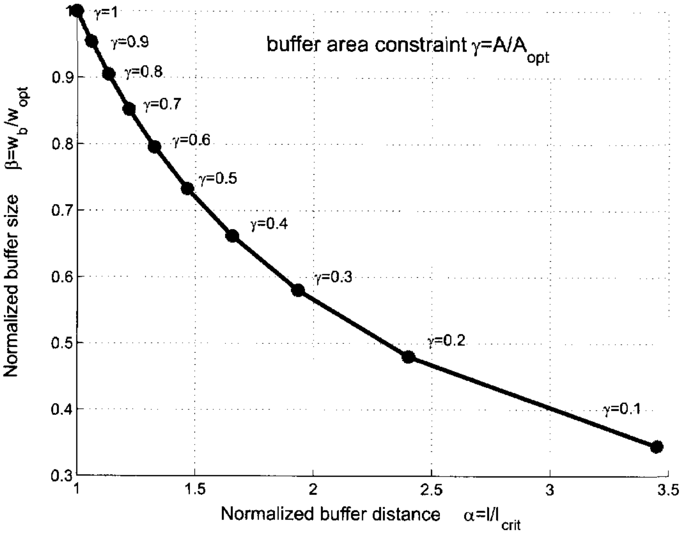


Figure 3.3. Normalized relation between buffer size and segmentation length

A graphical confirmation that $A(\gamma)$ is indeed the best possible trade-off, is presented in figure 3.4. It plots the normalized total buffer area (γ) and the normalized single buffer area (β) along the two horizontal axes and the normalized performance along the vertical axis. Via (3.12), each point on the horizontal (β, γ) plane also corresponds to a value of α, i.e. $\alpha = \beta/\gamma$. The surface mesh in the graph shows the normalized performance, as obtained by substituting α and β in 3.14.

However, not all (γ, β) pairs satisfy our optimality relation (3.15). Those that do, are highlighted on the surface mesh with the solid line. Now, by traversing the horizontal plane along the lines of constant area (i.e. γ), we can see that the highest performance indeed occurs at the highlighted line corresponding to (3.15). Again this confirms the fact that repeater insertion using (3.15) and (3.16) presents the best trade-off between area and performance.

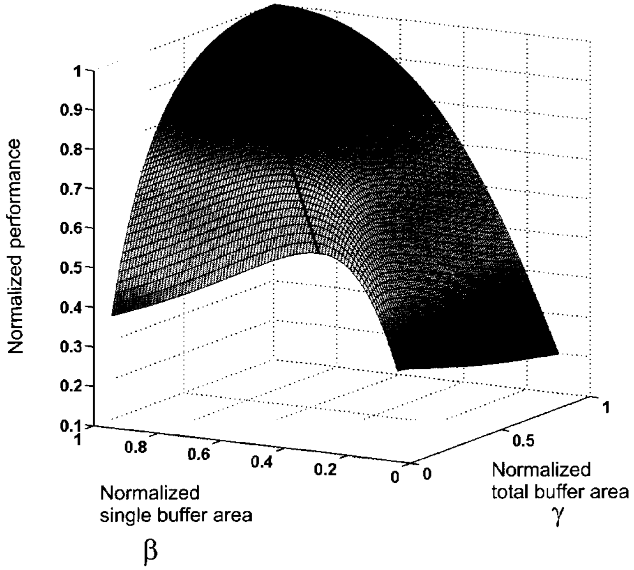


Figure 3.4. Optimal performance by assessing single buffer size for constrained area

Figure 3.5 shows for a line of $2cm$ the result from our area-constrained repeater insertion as a function of the number of buffers. For this graph, we are assuming a typical $0.18\mu m$ technology from the MOSIS website [40] and we consider the uppermost metal layer. Both the wire width and spacing are $0.7\mu m$ and the dielectric has $k = 3.5$ with a $1\mu m$ thickness to the metal layer beneath. We are considering a buffered line that is in between two other lines, assumed grounded. Its capacitance was derived from closed-form models presented in [65] and included in the Berkeley Prediction Model [10]. The transistors were characterized using the procedure in section 3.2. The results are:

$r_0 = 3k\Omega.\mu m$	$c_0 = 8.44fF/\mu m$	$\alpha_p = 0.12$
$r = 23.8m\Omega/\mu m$	$c = 23.7fF/\mu m$	

Table 3.2. Typical parameters for $0.18\mu m$ technology

Given the data, (3.3) and (3.4) with $a = 0.38$ and $b = 0.69$ (see section 2.3) will give $l_{crit} = 2.82mm$ and $w_{opt} = 60\mu m$. Rounding the results to 7 repeaters of size w_{opt}, the optimal performance is $v = 3.51e^{+13}\mu m/s$.

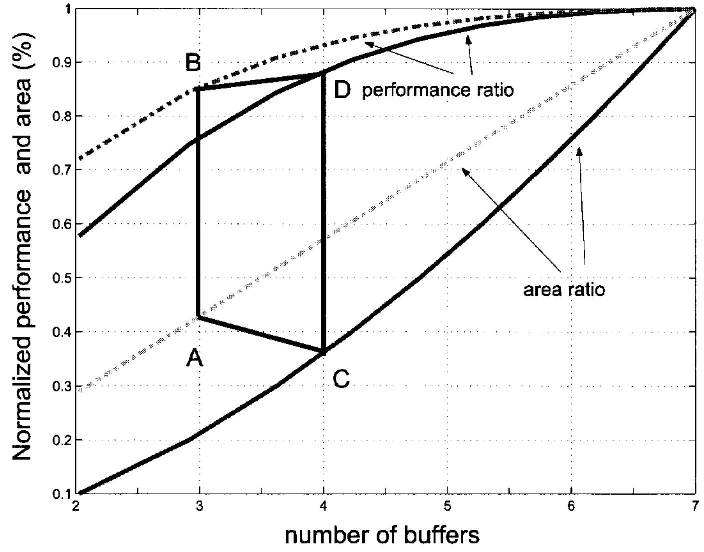


Figure 3.5. Normalized relation between performance and buffer area as function of the number of buffers

Figure 3.5 has along the horizontal axis the number of buffers and along the vertical axis the normalized performance (w.r.t. the unconstrained optimum performance) as well as the normalized total buffer area (with respect to the area for unconstrained optimum performance). The graph has two sets of curves, one for performance and one for the area ratio. The solid curves are from our area-constrained buffer insertion, while the dashed curves belong to the case where buffers keep their unconstrained optimum size but with the segment length increased. This will indeed give a straight line for the area ratio, since, with fixed buffer size, the total area is proportional to the number of buffers.

If we look at the case of three buffers with size w_{opt} (point A in the graph), and the associated performance (point B), then we see that, using the area-constrained repeater insertion results, better performance (point D) is achieved with less area (point C) though four buffers have been inserted.

Finally, we compare our analytic model with results from spice simulations for the same $0.18\mu m$ technology as above. The simulation served to determine the period of the same 7-stage ring oscillator with a 7-stage π-model for interconnect as was used for calibration earlier in this section. Using a ring oscillator automatically produces realistic waveforms at the input of each inverter, which is an advantage over using synthetic step or ramp inputs. For each value of γ, the transistors and interconnect in the oscillator were sized according to (3.15) and (3.16). The performance was obtained by dividing the period of oscillation by 14.

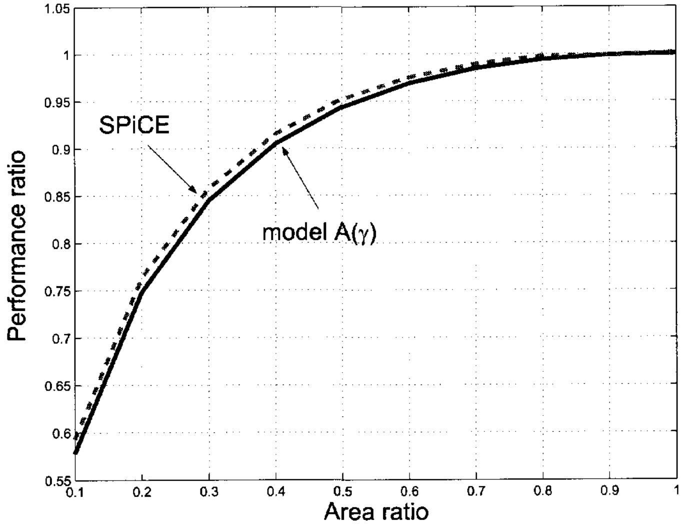


Figure 3.6. Comparison between spice and the proposed model

The results are depicted in figure 3.6, again with area ratio against performance ratio. The solid curve presents the results of (3.17) for that specific technology. The dashed curve presents the spice results, with the same normalization as for the analytic results. In point $(1, 1)$ both results match precisely, of course, since this point was used for calibration of the model. However, the results also match very well when the area is reduced, confirming the validity and fidelity of our results.

3.4 Area-power-delay trade-offs

We adopt the repeater power dissipation as proposed in [8]:

$$P_{repeater} = P_{switching} + P_{leakage} + P_{short-circuit} \tag{3.21}$$

where the terms are respectively the *switching power* $P_{switching}$, the *leakage power* $P_{leakage}$, and the *short-circuit power* $P_{short-circuit}$, explained below.

Switching Power: The switching power is the power needed to charge the total fanout capacity. For an inserted buffer the latter consists of wire capacitance, parasitic capacitance and input gate capacitance. Consequently,

$$P_{switching} = \alpha_1(w_b(c_p + c_0) + lc)V_{DD}^2 f_{clk} \tag{3.22}$$

where α_1, the *activity factor*, is the fraction of repeaters that are on average, switching in a clock cycle. This number can typically be taken 0.15. The precise value does not influence the validity of the results to be developed. V_{DD} is the supply voltage and f_{clk} is the clock frequency.

Leakage Power: The leakage power is due to the current flowing in the reverse-biased diode junctions of the transistors, located between source or drain and the substrate. The leakage power can be modelled as:

$$P_{leakage} = \frac{1}{2}V_{DD}(I_{\text{offn}}W_n + I_{\text{offp}}W_p) \simeq \frac{3}{2}V_{DD}I_{\text{offn}}w_b \tag{3.23}$$

The assumptions are here that $I_{\text{offn}} \simeq I_{\text{offp}}$ and $W_p = 2W_n = 2w_b$. Leakage current is caused by the thermally generated carriers and the generation of these carriers grows exponentially with the temperature. At 85^0C, the leakage current increases by a factor of 60 over the room-temperature values. Since the temperature is a function of the dissipated heat, it becomes important to use efficient heat removal mechanisms. According to some authors [8] this term is expected to become significant and even dominant compared with the other terms in (3.21).

Short circuit Power: The finite slope of the input signal causes a situation where both nmos and cmos transistors in an inverter are conducting for a certain time t_r. There is then a current path from V_{DD} to GND which causes power consumption. Short-circuit power can be modelled as follows:

$$P_{short-circuit} = \alpha_1 t_r V_{DD} I_{peak} f_{clk} \frac{3}{2}V_{DD}I_{\text{offn}}w_b \tag{3.24}$$

It depends on the rise time t_r as shown in figure 3.7. It has been demonstrated, by using spice simulations, that $I_{peak} \simeq w_b I_{short-circuit}$ where $I_{short-circuit}$ is fairly constant (and equal to $65\mu A/\mu m$ across all technologies).

If we assume that the input voltage waveform can be approximated by a single time-constant exponential and that $V_{t_n} = V_{t_p} = 1/4V_{DD}$ we can

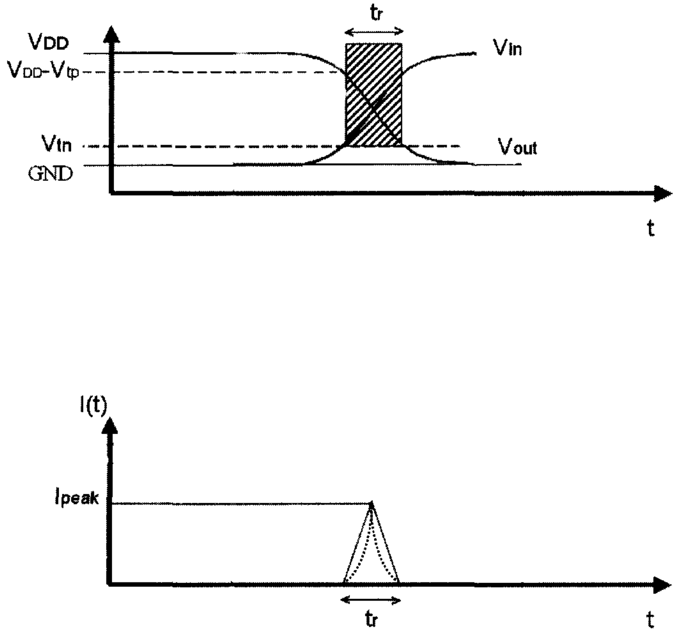


Figure 3.7. Voltage and approximated current waveforms for a CMOS inverter

relate the rise time t_r to the elmore delay of a buffered segment τ by the following equation:

$$t_r = \tau \ln\left(\frac{V_{DD} - V_{t_p}}{V_{t_n}}\right) = \tau \ln 3 \tag{3.25}$$

The contribution of this term remains relatively small [8]. Moreover, for smaller technologies it becomes less and less significant when compared to the increased leakage current. If we remain close to the unconstrained optimal buffer insertion, τ and thus t_r will be small and then so is short-circuit power. However, in the next section we will show that for certain non-uniform area-constrained buffer insertions the relative contribution of the short-circuit power is no longer negligible.

Substituting the individual contributions in equation 3.21 yields for a single segment of length l and a repeater with buffer size w_b

$$\begin{aligned} P_{repeater} &= P_{switching} + P_{leakage} + P_{short-circuit} \\ &= k_1((c_0 + c_p)w_b + cl) + k_2 w_b + k_3 w_b \tau \end{aligned} \tag{3.26}$$

where $k_1 = \alpha_1 V_{DD}^2 f_{clk}$, $k_2 = \frac{3}{2} V_{DD} I_{\text{offn}}$ and $k_3 = \alpha_1 V_{DD} I_{sc} f_{clk} \ln 3$ are constants that do not depend on buffer sizing. When the short-circuit power is neglected, (3.26) can be rewritten as:

$$\begin{aligned} P_{repeater} &\approx P_{switching} + P_{leakage} \\ &= k_1((c_0 + c_p)w_b + cl) + k_2 w_b \end{aligned} \tag{3.27}$$

For a line with n buffers uniformly inserted we can consider the power p per unit length: it

$$p = \frac{nP_{repeater}}{L} = (k_1(c_0 + c_p) + k_2)\frac{w_b}{l} + k_1 c \tag{3.28}$$

which can be rewritten by using $l = \alpha l_{crit}$ and $w_b = \beta w_{opt}$ as:

$$p = (k_1(c_0 + c_p) + k_2)\frac{\beta w_{opt}}{\alpha l_{crit}} + k_1 c = \frac{k'}{c}\frac{\beta}{\alpha} + k_1 c \tag{3.29}$$

where $k' = c(k_1(c_0 + c_p) + k_2)w_{opt}/l_{crit}$. Substituting the results for w_{opt} and l_{crit} yields:

$$k' = \frac{k_1(c_0 + c_p) + k_2}{\sqrt{\frac{a}{bc_0(c_0+c_p)}}} \tag{3.30}$$

Because k' and k_1 only depend on the buffer and c is the line capacitance per unit length, we may conclude from (3.29) that for minimum power per unit length β/α has to be minimal. Since (3.12) says that $\beta/\alpha = \gamma = A/A_{opt}$, we have – in violation of Murphy's law – that area minimization for a certain performance simultaneously minimizes power. In other words, (3.15) and (3.16) not only minimize area, but also power (provided that short-circuit power can be neglected).

Figure 3.8 presents power versus performance results using (3.15) and (3.16) and all three power terms from (3.26). This figure was derived from the $0.18\mu m$ technology from section 3.3. For this technology we can see that at 85% of the performance requires only 67% of the power is required. However, whereas the area optimizations were exact, the power minimization was performed under the assumption that the short circuit power term could be neglected. The following case study should indicate whether this

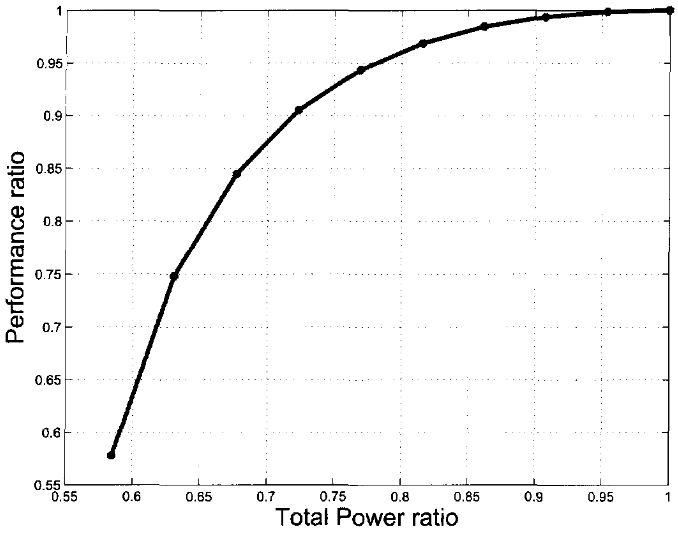


Figure 3.8. Normalized relation between power and performance

assumption was justified. We use once more the assumptions and technology data from section 3.3 that lead to table 3.2). And of course, just as in section 3.3 we get l_{crit}=$2.82mm$, w_{opt}=$60\mu m$ and seven $60\mu m$ repeaters for an optimal performance of v=$3.51e^{+13}\mu m/s$. For the repeater power model we take I_{offn}=$0.2\mu A/\mu m$, f_{clk}=$1.2Ghz$, V_{DD}=$1.8V$, I_{sc}=$65\mu A/\mu m$.

We will now combine area, power and performance results into one graph which gives figure 3.9. The axes are the same as those in figure 3.3 (but with a different range), i.e. segment length l and buffer size w_b normalized with l_{crit} and w_{opt} from (3.3) and (3.4). The optimal performance is then reached in the point (1,1). The closed contour shows all normalized segment length and buffer size combinations resulting in the same performance, in this example 15% worse than the optimal performance. All design points inside this contour exhibit better performance, and all outside points exhibit worse performance. The dashed line is the same as the curve from figure 3.3, showing the combinations of l and w_b that are optimal with respect to the area-delay trade-off. It passes through the (1,1) point by construction.

The solid curves starting at (0,0) are lines of constant normalized power, including all three power terms. They are plotted using (3.26), with parameters appropriate for the $0.18\mu m$ technology. Normalization was done using l=l_{crit} and w_b=w_{opt}. The labels identify the normalized power associated with the curve and the curve labelled "1" passes through the (1,1)

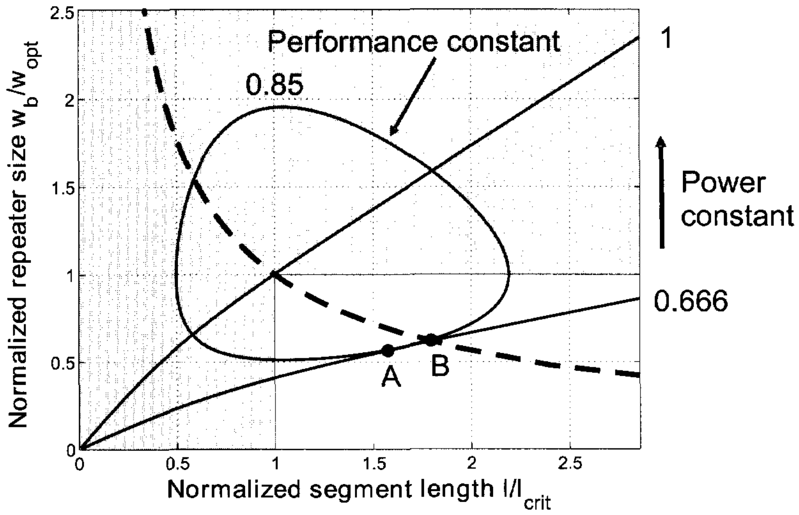


Figure 3.9. Minimization of total buffer area and power for a 15% of performance degradation for uniform buffer insertion

design point by construction. All curves for lower power will lie below the "optimal" curve, because lower power can only be achieved by smaller w_b for the same l.

Now, the optimal power for a certain performance is achieved in the point where a constant power curve is tangential to the performance contour. In figure 3.9 (for a 15% performance penalty), this is point A. This is approximately the same as the point for optimal area, which is the intersection of the dashed line for optimal area with the 15% performance penalty contour and marked as point B. The difference is due to the short-circuit power term in (3.26).

These results are further illustrated in figure 3.10, showing that the discrepancy between area and power optimization grows slightly with increasing performance penalties. However, for more advanced technologies the relative contribution of the short-circuit power is reduced [8], and our analytical results (3.29) show that in that case area and power optimization are exactly equivalent.

Table 3.3 quantifies optimal power solution and optimal area solution for different performance penalties. The values in the table are normalized with respect to the unconstrained buffering for maximum performance. For a certain performance loss (of 5%, 10%, 15%, 20% as presented in the table) the buffering which minimizes the area is compared to the buffering for

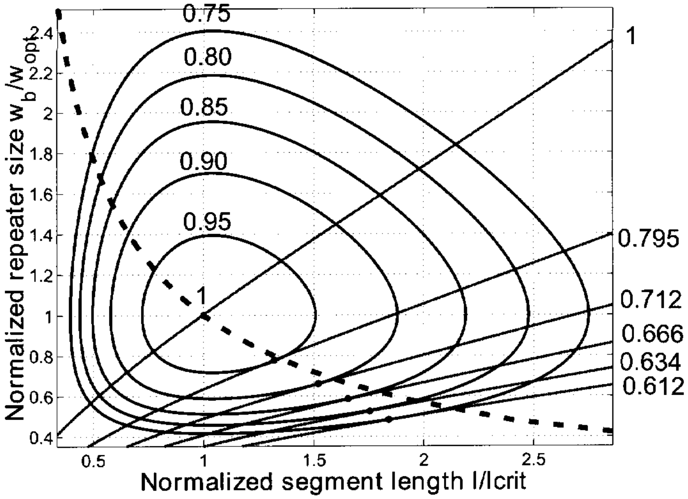


Figure 3.10. Area and Power versus Performance degradation

the minimization of power. For the case of area optimization, the values of the buffer size and distance are analytically obtained as explained in section 3.3. For power optimization , those values are found by an iterative procedure which detects the tangent point of power constant curve with the contour of the performance in the space (l, w_b), see figure 3.10

	Trade-off terms			Buffering	
	v^{-1}/v_{min}^{-1}	P_{rep}/P_{rep_opt}	A/A_{opt}	$\frac{w_b}{w_{opt}}$	$\frac{l}{l_{crit}}$
Power Optimization	0.95	0.795	0.54	0.71	1.3
	0.90	0.712	0.41	0.62	1.5
	0.85	0.666	0.31	0.57	1.65
	0.80	0.634	0.3	0.51	1.7
Area Optimization	0.95	0.795	0.53	0.76	1.41
	0.90	0.715	0.41	0.67	1.63
	0.85	0.670	0.32	0.61	1.85
	0.80	0.642	0.29	0.57	1.95

Table 3.3. Power optimal versus Area optimal solution

We can notice that power optimization and area optimization lead to similar trade-off values, but the gap between the two optimizations widens for lower performance. This is due the fact that short-circuit power for lower performance becomes relatively more important. In fact, when the short-circuit power is of the same order as the other terms of the total power budget, its minimization is influenced by this term. Then a buffering which can ultimately reduce buffer size and their distance help in minimizing specifically the short-circuit term.

Figure 3.11 further displays the effectiveness of our trade-off model in practical situations, where a discrete number of buffer has to be used for a given global wire. It is an extension to figure 3.5. In particular, figure 3.11 shows along the horizontal axis the number of buffers and along the vertical axis the normalized performance (with respect to the unconstrained optimum performance), the normalized total buffer area (with respect to the area for unconstrained optimum performance) and the normalized total buffer power (with respect to the power for unconstrained optimum performance). The graph shows three sets of curves, for performance ratio, area ratio and power ratio. The solid curves are obtained under the area-delay-power trade-off of this section, and the dashed curves belong to the case where the segment length is increased without changing the buffer sizes from their value under unconstrained optimal buffering.

Again consider the case of three buffers of size w_{opt}. This corresponds to a point A in the graph, the associated performance is given by B and the power is given by E. Using our new area-constrained repeater insertion results, one could use four buffers with less area (point C) while performance (point D) and power (point F) are reduced. Clearly, our new repeater insertion method gives not only better performance with less area, but also uses less power.

3.5 Discussion

This chapter has presented theoretically optimal results for uniform buffer insertion for point-to-point connections. In cases where buffer locations are constrained, these results can provide a lower bound on area and power and can be used as a preconditioner for optimization based buffer planning algorithms. The lower bound property can be used for screening of critical segments that are amenable to further optimization.

We further studied, as an extension of previous results on repeater sizing and line segmentation for optimal speed [7, 49], the problem of optimal repeater insertion in global interconnects when only limited total buffer area is available and/or the total repeater power is bounded. This allows the designer to limit the total buffer area or the repeater power, and obtain the reduced repeater sizes and increased separations belonging to the best performance under those constraints.

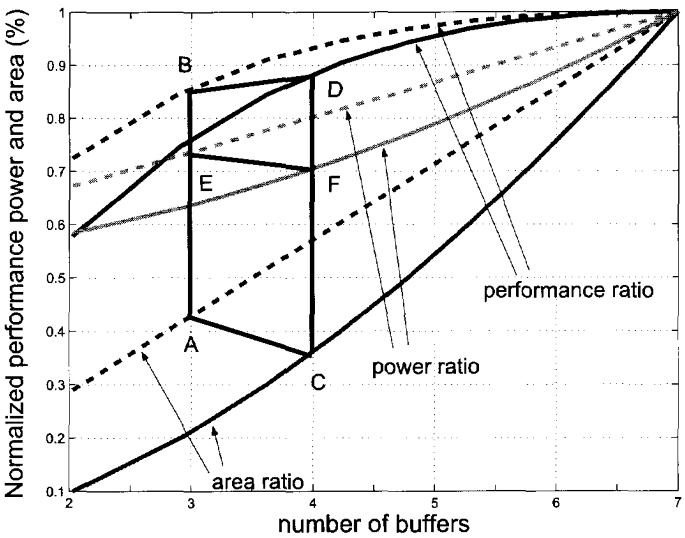


Figure 3.11. Normalized relation between performance, buffer area and power as function of the number of buffers

Our result can be summarized in a precise trade-off curve for area, power and performance in a global interconnect line with repeaters. The curve of area versus performance does not depend on the interconnect characteristics and only mildly on the transistor parameters. The curve of power versus area depends on the wire capacitance. Compared to the unconstrained result, only 50% of the total buffer area and 77% of the total buffer power is needed for 95% of the absolute maximum speed and 20% of the total area and 63% of the total power for 75% of the speed and only 10% of the total area and 58% of the total power for 57% of the speed.

Chapter 4

Non-uniform buffering

Contents

UNIFORM buffer insertion, whether area-constrained or not, can provide tight upper bounds on the performance of a given interconnection. The results of chapter 3 may therefore be useful also when there are constraints on the placement of these buffers. Such constraints generally shut the door on optimal segmentation without changing the length. And they are very common due to numerous layout rules, constrained positions of IP-blocks and blockage issues. So, buffers can in general be placed only in certain locations and buffer allocation in the available locations is most likely not going to be uniform.

In non-uniform buffering segments may be longer than l_{crit}, in which case we speak of an underdriven segment, and shorter than l_{crit}, in which case we say that the segment overdriven. In both cases the signal velocity propagation is lower than with all segments equal to l_{crit} and optimal buffer sizes. In [13] a study on the worst-case buffer placement (obtained by alternating overdriven and underdriven buffer segments) shows a marginal effect on the total delay, but a considerable effect on peak noise. Peak noise is a specific function of a single segment and not of the total path length. However, in [13] no buffer resizing is considered and the size of devices driving and ending the total path length are not specified.

Other works do focus on buffer sizing, but they do not consider location constraints. In [66] the authors propose an algorithm for timing-driven buffer sizing. Hier critical path delays can be controlled by using new adequate buffer cells, and static timing analysis is used to verify whether the path with the new buffers respects the specifications. [16] addresses the problem of buffer sizing in combination with wire sizing which aims at reducing power while achieving better delay. In another work the dynamic programming algorithm from [62] is extended with simultaneous buffer sizing for a given Steiner tree [2]. Here, the authors come to the conclusion that the buffer sizes should not be too large, because otherwise the delay penalty on the upstream part of the tree becomes significant.

In this chapter we focus on the problem where a point-to-point connection is given together with the location of its buffers, and we want know the sizes of the buffers that minimize delay over that connection. Also the size of the driver at the beginning and of the receiver at the end are included in the optimization. First we assume no limitation on the buffer area. The second part of this chapter is devoted to area-constrained non-uniform buffering.

4.1 Sizing in-line buffers

The size of the devices driving and loading the wire, called driver and receiver, are denoted by w_{TX} and w_{RX} respectively. The number of in-line buffers is $N-1$ and their position on the wire(s) routed from the buffer w_{TX} to the buffer w_{RX} is given. The distance between buffers is not necessarily uniform and not necessarily equal to l_{crit} (equation 3.3). The wire is thus divided in N segments. A generic wire segment between the buffer j with size w_j, driving the segment, and buffer $j+1$ with size w_{j+1}, loading it, is $l_{j,j+1}$ and for such segment the delay is denoted by $\tau_{j,j+1}$.

For this fixed number of buffers with assigned buffer locations, we want to tune the buffer sizes such that the total delay of the wire is minimized. Since the delay T is after calibration assumed to be additive in terms of the delay contribution of each of the buffered segments, we can write this optimization problem as:

$$minimize \ : \ T(w_1, w_2, w_3 \dots w_{N-1}) = \sum_{j=1}^{N-1} \tau_{j,j+1}(w_j, w_{j+1}) \qquad (4.1)$$

where $w_0 = w_{TX}$ and $w_N = w_{RX}$, and $w_1, w_3 \dots w_{N-1}$ are the sizes of the $N-1$ buffers to be tuned.

Again we start from the segment delay formula at the end of chapter 2. For the segment before buffer j its input capacitance is the load. The delay for that segment is then:

$$\tau_{j-1,j} = br_0(\frac{w_j}{w_{j-1}}c_0 + c_p) + arcl^2_{j-1,j} + \left(\frac{br_0c}{w_{j-1}} + bc_0rw_j\right)l_{j-1,j} \qquad (4.2)$$

For the segment after buffer j which is driven by a driver of size w_j we get:

$$\tau_{j,j+1} = br_0(\frac{w_{j+1}}{w_j}c_0 + c_p) + arcl^2_{j,j+1} + \left(\frac{br_0c}{w_j} + bc_0rw_{j+1}\right)l_{j,j+1} \qquad (4.3)$$

Taking the derivative with respect to w_j of the sum of the expressions in 4.2 and 4.3, and solving $\frac{\partial(\tau_{j-1,j}+\tau_{j,j+1})}{\partial w_j} = 0$ for w_j gives its optimal value for given w_{j-1} and w_{j+1}:

$$w_j = \sqrt{\frac{r_0}{c_0}\frac{cl_{j,j+1} + c_0w_{j+1}}{rl_{j-1,j} + \frac{r_0}{w_{j-1}}}} \qquad (4.4)$$

But w_{j-1} and w_{j+1} are not given! Actually we only solved, in closed-form, a 2-segment problem. In other words, what size a buffer should have, positioned on a given location in a line, that is not further segmented. For an arbitrary number of buffers along such a line we do not have a closed-form solution like (4.4). But (4.4) can be useful for that problem. If, for example, all buffers are given an initial size and on the basis of those sizes we calculate for every buffer a new size, using (4.4) and the initial values for the buffer before and after that one, we get a new series of sizes. They can be used to update all sizes once again. Clearly, we have the inner works of an iterative procedure. What is missing is an initialization and a stopping criterion, for at the i-th iteration the sizes are updated by

$$w_j^{(i)} = \sqrt{\frac{r_0}{c_0}\frac{cl_{j,j+1} + c_0w_{j+1}^{(i-1)}}{rl_{j-1,j} + \frac{r_0}{w_{j-1}^{(i-1)}}}}. \qquad (4.5)$$

If, as we expect, convergence in absolute value is smooth, a sensible stopping criterion can be when all changes in buffer size over one iteration are below the feature size. As for initialization, a case can be made for $w_j = w_{opt}$ $1 \le j \le N-1$ when the number of segments is close to L/l_{crit}, because though optimal segmentation might be impossible we still want to stay close to it. And if we are successful the buffer sizing will be close to the optimum as well. Consequently, the buffer sizes will stay close w_{opt} when we are not too far from the unconstrained optimal segmentation.

As a first test however the algorithm was run from a random initialization for an optimally segmented wire to see whether it converges to the well-known solution of unconstrained optimization. And indeed for a long wire divided uniformly in segments of length l_{crit} as in (3.3), the sizes of the buffers on this wire converge to w_{opt} as in (3.4).

In addition to this specific case, we have tested the convergence of the algorithm for different non-uniform buffer locations. The assigned buffer locations are assigned randomly for each of the experiments. We also used different driver and receiver sizes. Typically 5 to 15 iterations are sufficient to obtain stable values for the each of the buffer sizes.

An example of the number of iterations needed to achieve stable values of the buffer sizes is shown in figure 4.1. We have plotted the sum of the normalized squared difference between the buffer sizes at two consecutive iterations.

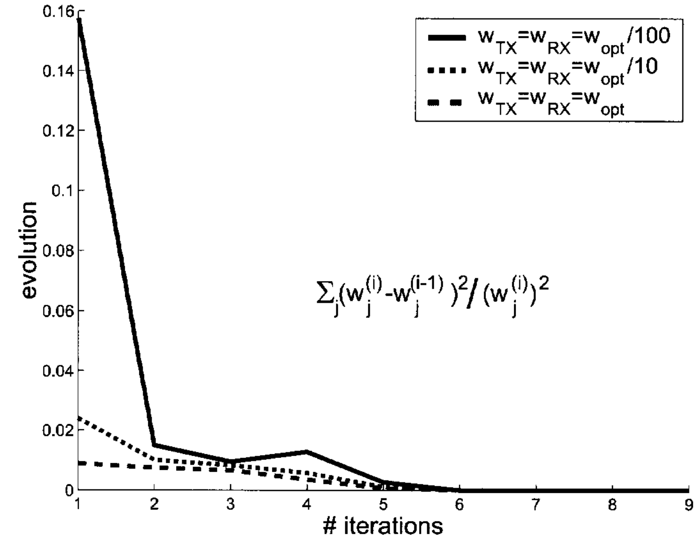
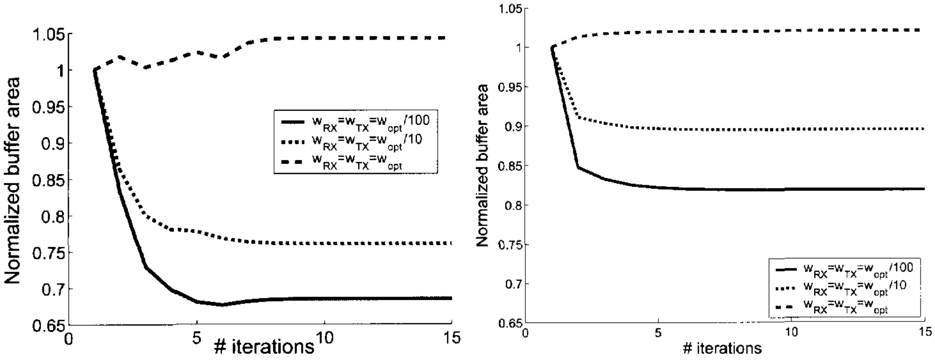


Figure 4.1. Convergence of the area-unconstrained buffer sizing algorithm

Figure 4.2 shows the number of iterations needed to reach a constant value of the total buffer area in two specific cases: a wire of length $L = 7l_{crit}$ with 6 buffer to size, and a wire with $L = 15l_{crit}$ and 14 buffers to size. In both cases the buffer locations are assigned randomly. The convergence for the longer wire length seems to be faster, but this is due to the averaging effect introduced by evaluating the total buffer area instead of single buffer sizes. In fact for longer wires, which require more buffers, the total buffer area becomes less sensitive to the non-uniform assigned buffer locations. We also note the impact of the driver and receiver size.



(a) wire length $L = 7l_{crit}$

(b) wire length $L = 15l_{crit}$

Figure 4.2. Convergence of the area-unconstrained buffer resizing algorithm for different wire lengths

We have also studied performance convergence over the iterations. We skip a detailed description of those results, but we have observed that the convergence of the performance to the optimal value tends to be faster than the convergence of the total buffer area.

Now for results obtained by using the tuning of the buffer sizes we consider once again technology and the device parameters and interconnects parameters of table 3.2. Optimal uniform buffering will have $l_{crit} = 2.8mm$ and $w_{opt} = 59\mu m$. It is assumed that the total length of the line is an integer multiple of l_{crit}, namely $L = 7l_{crit}$. and that the number of buffers to be inserted between driver and receiver is $L/l_{crit} - 1$. Each of those buffers is inserted in an arbitrary but given position and consequently the buffer sizes have to be tuned depending of where those buffers are placed. The segments lengths, relative to l_{crit}, are given in table 4.1.

We size the buffers for different values of the driver and receiver size. They are assumed to be equally sized, although that is not necessary for the algorithm. First the driver and receiver size are set to the value w_{opt} and

l_{12}	l_{23}	l_{34}	l_{45}	l_{56}	l_{67}	l_{78}
$0.3l_{crit}$	$1.7l_{crit}$	$1l_{crit}$	$0.4l_{crit}$	$1.6l_{crit}$	$1l_{crit}$	$1l_{crit}$

Table 4.1. Example of buffer resizing: assumed buffer locations

this choice helps in seeing the impact on performance of buffer displacement from their optimum positions. The second situation is with driver and receiver size equal to $w_{opt}/10$. And the last set of results are for driver and receiver size set at $w_{opt}/100$. (Typically the driver and receiver are smaller than the in-line buffers.)

TX/RX	Buffer sizes (μm)						**Normaliz. I**		**Normaliz. II**	
$w_{TX} = w_{RX}$							$w_j = w_{opt}$		$w_j = w_{opt}$	
							w_{opt}		design spec.	
	w_2	w_3	w_4	w_5	w_6	w_7	Perf.	Area	Perf.	Area
w_{opt}	89	51	51	79	54	58	0.97	1.091	1.02	1.09
$\frac{w_{opt}}{10}$	31	43	49	78	51	45	0.92	0.80	1.06	0.87
$\frac{w_{opt}}{100}$	10	30	44	75	51	43	0.57	0.65	1.7	0.78

Table 4.2. Buffer resizing and relative performance and area for different sizes of TX and RX

The results are summarized in the table 4.2 using two kinds of normalization.

I : performance and buffer are expressed as ratio with respect to the situation in which every buffer, including driver and receiver are set to w_{opt}. In particular if we assume uniform buffer distance $l = l_{crit}$, then we have the absolute maximum performance. The effect of non-uniform buffer distribution and driver and receiver sizes are compared directly to the absolute maximum performance.

II : compares the performance and area obtained by tuning the buffer sizes to the scenario where every internal buffer is $w_j = w_{opt}$ and where driver and receiver are differently specified. Thus, this normalized point depends on the driver and receiver size selected. We use this normalization to verify that our resizing method gives performance improvement (70% better performance with 22% less area) especially for small drivers and receivers.

The results in table 4.2 induce some general observations:

- A single buffer, depending also on its assigned location, can be upsized to a value larger than w_{opt}. In table 4.2 the buffer w_5 is always larger than w_{opt}, and regardless of driver and receiver size.
- Displacement of buffers has a marginal effect on the sizing, certainly compared to changing the driver and receiver sizes. Indeed, the first data row in the table 4.2 shows that when driver and receiver are set to w_{opt} and the buffer resizing is only due to the non-uniform distribution, the performance is not significantly different from uniform buffering (3%). If we consider the buffer resizing for smaller drivers and receivers, we notice that the performance degrades but at the same time the total buffer area is reduced.
- The buffer resizing does not improve the performance significantly in the situation when driver and receiver equal w_{opt}. This is because having all buffers including driver and receiver equal to w_{opt}, we are already quite close to the absolute optimal value of performance curve. Moreover the performance curve as function of the buffer size in the region close to w_{opt} is quite flat. Different is the situation when driver and receiver are smaller than w_{opt}. Here buffer resizing can help improving the performance considerably. The smaller the driver and receiver, the more buffer resizing is required to maintain performance. We can gain 70% in performance by resizing buffers when driver and receiver are $w_{opt}/100$.

4.2 Sizing in-line buffers under area constraints

Let A_{max} be the total buffer area on the line after sizing them according to section 4.1. That is, all buffers have been tuned by the algorithm described there and obtained sizes $w_1, w_2, \ldots w_{N-1}$. Suppose that that area is not available, and the total buffer area has to be reduced with a factor γ. Of course, this reduction can be achieved by scaling all buffers with that factor γ. This is unlikely to achieve the lowest delay possible by spending γA_{max} in buffer area. The problem we want to consider here is to distribute the total area penalty over each of the buffers along the path such that the best performance under this constraint is guaranteed. In other words we have to scale each of the buffer sizes with a factor γ_i such that the total area is equal to γA_{max} and delay is minimized:

$$\begin{aligned} minimize \quad &: \quad T(\gamma_2, \gamma_3 \ldots \gamma_N) = \sum_{j=1}^{N} \tau_{j,j+1}(\gamma_j, \gamma_{j+1}) \\ subject \quad to \quad &: \quad w_1 + \sum_{j=2}^{N-1} \gamma_j w_j + w_{N+1} = \gamma A_{max} \end{aligned} \tag{4.6}$$

Note that an individual scaling factor in (4.6) depends on all other scaling factors and that it is not possible to derive a closed-form expression for such a factor in terms of the factors of the buffers just before and after it, as we could do for the buffer size in section 4.1.

A standard way to solve the minimization problem (4.6) is by using the lagrange method of undetermined multipliers [5], [68]. It works as follows. We define

$$F = \sum_{j=1}^{N-1} \tau_{j,j+1} + \lambda(\gamma A_{max}) \ ,$$

and solve the non linear system of $N - 2$ equations

$$\frac{\partial F}{\partial \gamma_2} = 0, \ldots, \frac{\partial F}{\partial \gamma_j} = 0, \ldots, \frac{\partial F}{\partial \gamma_{N-1}} = 0.$$

Instead of solving this system of equations, we propose an approximation by locally optimizing a single buffer at a time, based only on the information from the segment preceding and following the buffer to be sized. Of course this will introduce an error, an error that we hope to reduce by subsequent resizing of the buffers based on this error. We will show that the error thus introduced is progressively reduced at each iteration while each of the buffer sizes on the long wire converges to a constant value. We first explain how to solve the minimization problem by scaling a single buffer size under an area constraint. This will be used then in the buffer scaling for a long wire containing N buffers.

Area-constrained tuning of a single buffer We consider the particular case in which we have only one buffer j between the "driver and receiver" with w_{j-1} and w_{j+1} respectively. So, we might say $j = 2$, but we keep the notation with j for generalization to N later. The three buffers are to be scaled such that

$$\gamma A_{max} = \gamma_{j-1} w_{j-1} + \gamma_j w_j + \gamma_{j+1} w_{j+1} \tag{4.7}$$

where w_{j-1}, w_j, w_{j+1} are given. They might have been determined by the procedure of section 4.1. The ”driver and receiver” are scaled by the same factor ($\gamma_{j-1} = \gamma_{j+1}$) which compromises the optimization, but allows us to solve the problem analytically. We will show, in the generalized context, that the error thus introduced, will be reduced over a series of iterations. The delay of each of the two segments in this configuration can be written as

$$\begin{aligned} \tau_{j-1,j} &= \tau_{j-1,j}^{min} + \Delta\tau_{j-1,j}(\gamma_{j-1}, \gamma_j) \\ \tau_{j,j+1} &= \tau_{j,j+1}^{min} + \Delta\tau_{j,j+1}(\gamma_j, \gamma_{j+1}) \end{aligned} \tag{4.8}$$

where $\tau_{j-1,j}^{min}$ and $\tau_{j,j+1}^{min}$ are the delays obtained by tuning without area constraint (section 4.1), and $\Delta\tau_{j-1,j}(\gamma_{j-1}, \gamma_j)$ and $\Delta\tau_{j,j+1}(\gamma_j, \gamma_{j+1})$ denote the delay penalties due to the area constraint for the respective segments.

We change for convenience the notation with $\gamma_{j-1} = \gamma_{j+1} = \tilde{\gamma}$ and $\gamma_j = \alpha\tilde{\gamma}$ so that the problem can be written as

$$\begin{aligned} minimize: \quad \Delta T(\gamma_{j-1}, \gamma_j, \gamma_{j+1}) &= \Delta\tau_{j-1,j}(\gamma_{j-1}, \gamma_j) + \Delta\tau_{j,j+1}(\gamma_j, \gamma_{j+1}) = \\ &= \Delta\tau_{j-1,j}(\tilde{\gamma}, \alpha\tilde{\gamma}) + \Delta\tau_{j,j+1}(\alpha\tilde{\gamma}, \tilde{\gamma}) \end{aligned} \tag{4.9}$$

We use (4.2) and (4.3) expressed in $\tilde{\gamma}$ and α to explicitly solve the minimization in (4.9). The performance degradation due to the area constraint is:

$$\Delta\tau_{j-1,j}(\tilde{\gamma}, \alpha\tilde{\gamma}) = br_0c_0\frac{\alpha w_j}{w_{j-1}} + br_0c\frac{l_{j-1,j}}{\tilde{\gamma}w_{j-1}} + brc_0\alpha\tilde{\gamma}w_jl_{j-1,j} \tag{4.10}$$

$$\Delta\tau_{j,j+1}(\tilde{\gamma}, \alpha\tilde{\gamma}) = br_0c_0\frac{w_{j+1}}{\alpha w_j} + br_0c\frac{l_{j,j+1}}{\alpha\tilde{\gamma}w_j} + brc_0\tilde{\gamma}w_{j+1}l_{j,j+1} \tag{4.11}$$

Notice that the sum $\Delta\tau_{j-1,j}(\tilde{\gamma}, \alpha\tilde{\gamma}) + \Delta\tau_{j,j+1}(\tilde{\gamma}, \alpha\tilde{\gamma})$ has once more the form $f/\alpha + g\alpha$ where f and g are two positive functions α independent. This implies that we can expect an absolute minimum at $\alpha = \sqrt{f/g}$ which is obtained by setting the sum's derivative with respect to α equal to zero. This means that $\Delta\tau_{j-1,j}(\gamma_{j-1}, \gamma_j)$ has to satisfy the following two equations:

$$\alpha_{min}^2(\tilde{\gamma}) = \frac{r_0c_0\frac{w_{j+1}}{w_j} + r_0c\frac{l_{j,j+1}}{w_j}\frac{1}{\tilde{\gamma}}}{r_0c_0\frac{w_j}{w_{j-1}} + rc_0w_jl_{j-1,j}\tilde{\gamma}} \tag{4.12}$$

$$\gamma A_{max} = \tilde{\gamma}(w_{j-1} + w_{j+1}) + \tilde{\gamma}\alpha_{min}w_j \tag{4.13}$$

Generalization to N buffers In essence we solve N systems of two equations as described in (4.12) and (4.13). The solution of each system will give buffer j the size $\tilde{\gamma}\alpha_{min}w_j$. This is only an approximated solution, because the mutual dependence of individual scaling factors is neglected. In fact we compute scaling factor "locally" by imposing $\gamma_{j-1} = \gamma_{j+1} = \tilde{\gamma}_j$ and then finding the value of $\gamma_j = \alpha_{min}\tilde{\gamma}_j$ from (4.12) and (4.13). After scaling we set the total buffer area obtained equal to $\hat{\gamma}A_{max}$. This may differ from the target scaling γA_{max} so that an error $\epsilon = \hat{\gamma} - \gamma$ is introduced.

This error will be reduced when we apply this "local buffer scaling" iteratively. If i is the iteration index, the iterative scheme adopted updates the value of $\gamma^{(i)}$ based on the local approximation error and the value of $\gamma^{(i-1)}$ from the previous iteration as follows:

$$\gamma^{(i)} = \gamma^{(i-1)} - \epsilon^{(i)} \tag{4.14}$$

where $\gamma^{(0)} = \gamma$. To get an idea about the convergence of this scheme we look at figure 4.3 where the squared sum of the differences between the buffer sizes at two consecutive iteration steps defined as $\sum_j (w_j^{(i)} - w_j^{(i-1)})^2/(w_j^{(i)})^2$ is depicted. We present in this figure only the results for $\gamma = 0.3$, but similar convergence trends can be observed for other values of γ.

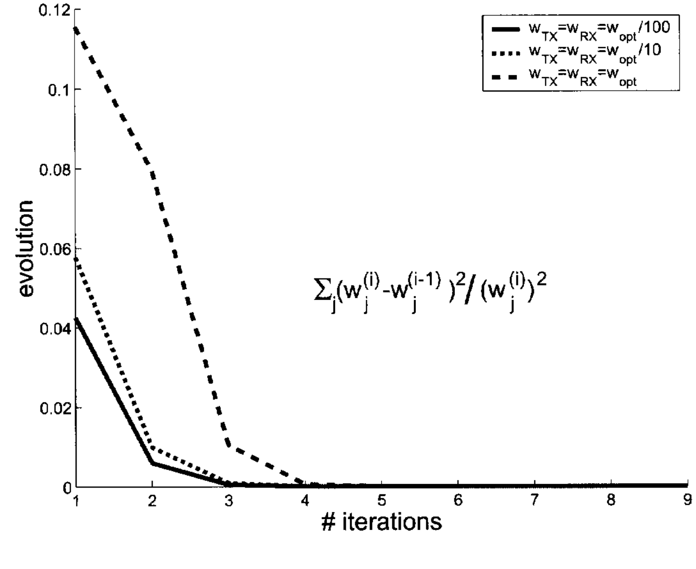
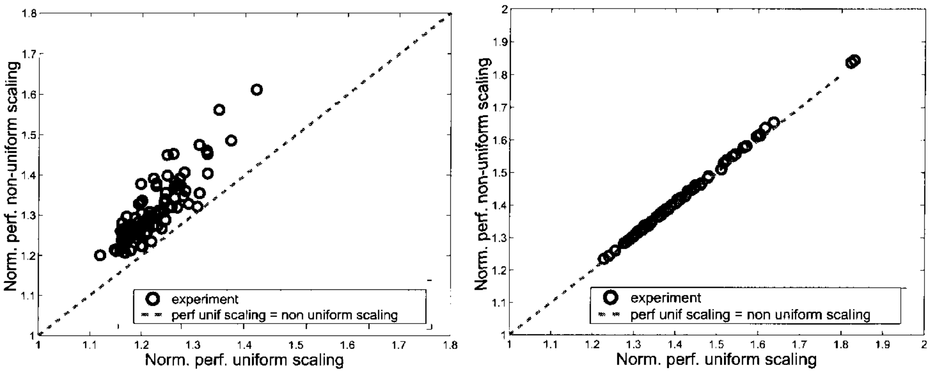


Figure 4.3. Convergence of non-uniform scaling of under $\gamma = 0.3$

4.3 Results

Figure 4.4 summarizes the performance values obtained by making 100 experiments which use every time a randomly buffer locations, where the number of buffers is fixed at 6, and where the size of the driver and the receiver is set to $w_{opt}/100$. We evaluate the improvement achieved by using our scaling algorithm over uniform buffer scaling. Results are presented for aggressive and moderate area constraint, $\gamma = 0.3$ and $\gamma = 0.7$ respectively. Improvements in performance due to our ad hoc scaling of the buffer sizes can amount to more than 20%.



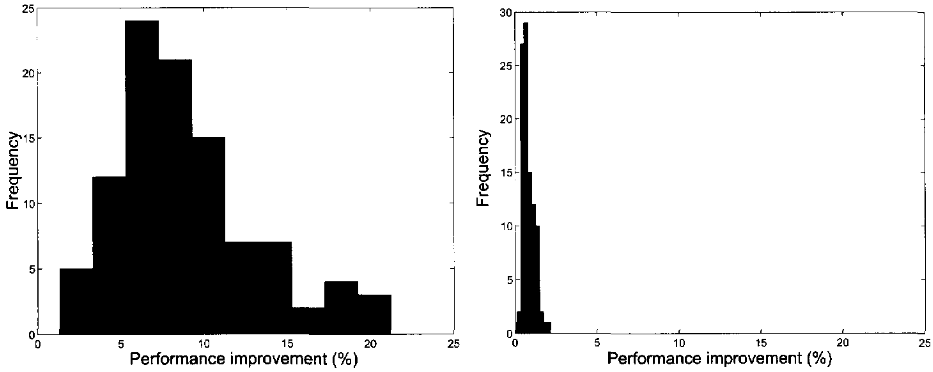
(a) $\gamma = 0.3$

(b) $\gamma = 0.7$

Figure 4.4. Improvement in performance due to our ad hoc scaling over the uniform buffer size scaling in presence of area constraints for 100 experiments with randomly assigned locations

Frequency diagrams show the performance more clearly. They are presented in figure 4.5 for aggressive and moderate area constraints. For each of the experiments we select the buffer locations randomly, while the number of buffers is kept at 6 and the total length of the wire is $7l_{crit}$; driver and receiver size are set to $w_{TX} = w_{RX} = w_{opt}/100$.

Note that for $\gamma = 0.3$ the performance gain is significant, with a maximum value of 20% and an average of 7% when compared to the case $\gamma = 0.7$ with less of 3% in performance improvements. This can be generally observed. It seems therefore recommendable to use uniform buffer scaling for moderate constraints, while for aggressive area constraints the scaling



(a) $\gamma = 0.3$

(b) $\gamma = 0.7$

Figure 4.5. Frequency diagrams of the performance improvement due to buffer sizing of our ad hoc scaling over uniform scaling. Assumptions: 100 experiments with 6 buffer locations randomly assigned and wire length $L = 7l_{crit}$

technique section 4.2.

In figure 4.6 the sizes of each of the buffers are plotted for each experiment. These results are generated again by making 100 experiments, where buffer locations are randomly generated, the length of the wire is fixed, the number of buffers is fixed to 6. Figure 4.6 summarizes the results and the difference with respect to the uniform buffer scaling for $\gamma = 0.3$. Analogously to what has been presented in figure 4.6, figure 4.7 presents the results for $\gamma = 0.7$.

The gain that we obtain in performance by adopting our ad hoc buffer scaling is almost totally due to the fact that the first buffer in the path is oversized with respect to the uniform scaling case. To respect the area constraint the other buffers are consequently slightly downsized.

Consider now the buffer distribution as the one presented in table 4.1. Once more we compare the improvement in performance of ad hoc buffer scaling over the uniform scaling as dependent on the area constraint. Figure 4.8 shows the improvement in performance due to the tuning algorithm which resizes continuously the buffers given their locations and the driver and receiver size. Driver and receiver sizes are set to $w_1 = w_8 = w_{opt}/100$ and the total length of the wire is $2cm$.

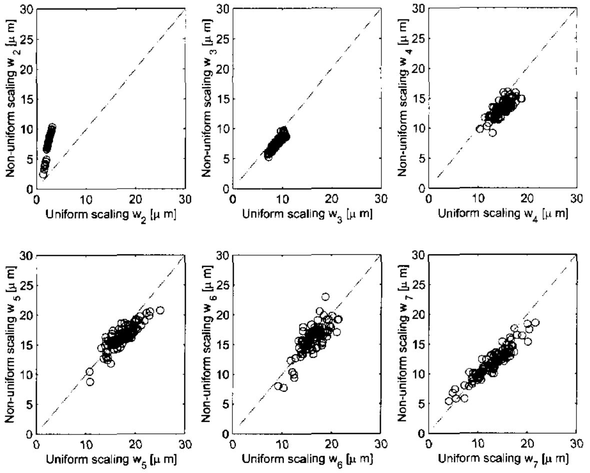


Figure 4.6. Details of the buffer sizes for uniform scaling versus optimal ad hoc scaling experiment with random assigned locations and $\gamma = 0.3$

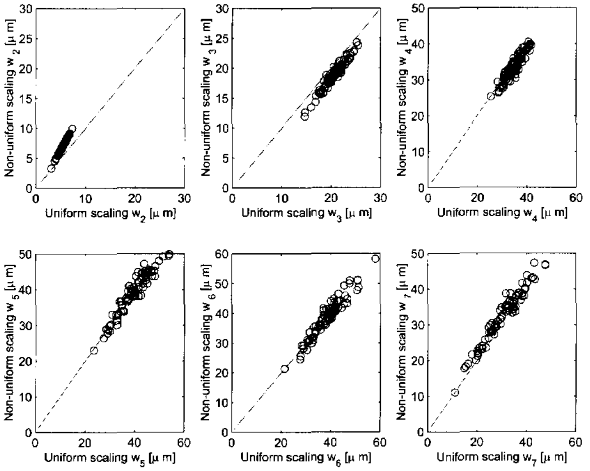


Figure 4.7. Details of the buffer sizes for uniform scaling (dashed) versus optimal ad hoc scaling experiment with random assigned locations and $\gamma = 0.7$

First we assume that the buffer sizes are all equal to w_{opt}. This corresponds in figure 4.8 to the point of performance $1/T(w_{opt})$ and total buffer area $A(w_{opt})$, and this is assumed to be the normalized point. The normalization adopted is the same *Normalization II* in table 4.2. Because of the non-uniform locations (assumed to be as in table 4.1) and because the driver and the receiver are smaller than w_{opt} ($w_{TX} = w_{RX} = w_{opt}/100$) an optimal resizing can be achieved by applying the algorithm in section 4.1. This will give a substantial better performance (70% better) and will use less total buffer area (30% smaller). This point will have the area A_{max} and the maximum performance for a given driver and receiver sizes and for the given buffer positions.

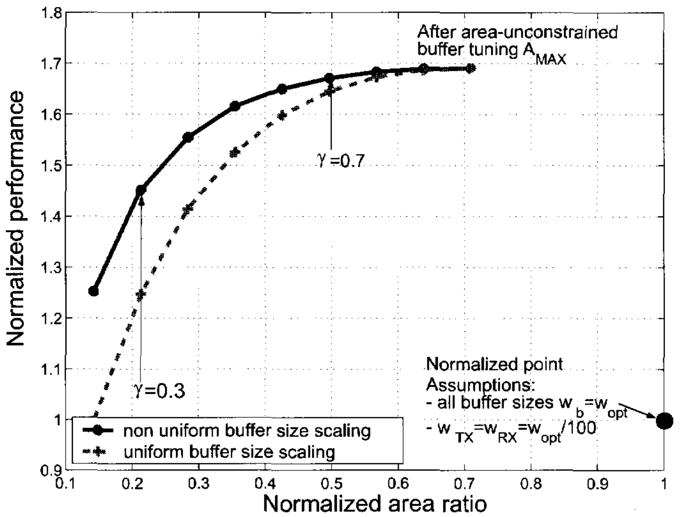


Figure 4.8. Comparison performance versus area for uniform scaling of buffer sizes and for non-uniform distribution of the scaling factor for each of the buffers

Suppose now that A_{max} is still too large and we have to resize the buffers. We have now in the figure 4.8 a comparison between uniform buffer size scaling (dashed line) against the proposed non-uniform scaling of buffers (solid line) as function of the normalized area constraint. Our resizing technique results in significantly better performance over uniform scaling. For example, if the area available is γA_{max} with $\gamma = 0.3$, by applying our resizing method the value of normalized performance is 1.45 which, compared to an uniform sizing with a normalized performance of 1.25, leads to a gain of performance is of about 20%.

Finally we show in figure 4.9 the size of each buffer, for uniform and for non-uniform scaling due to the area constraint γA_{max}, where $\gamma = 1$ (i.e. no constraint) and in the situation where $\gamma = 0.7$ and $\gamma = 0.3$. Also presented in figure is the situation where all the buffers have size w_{opt}, the reference point in figure 4.8.

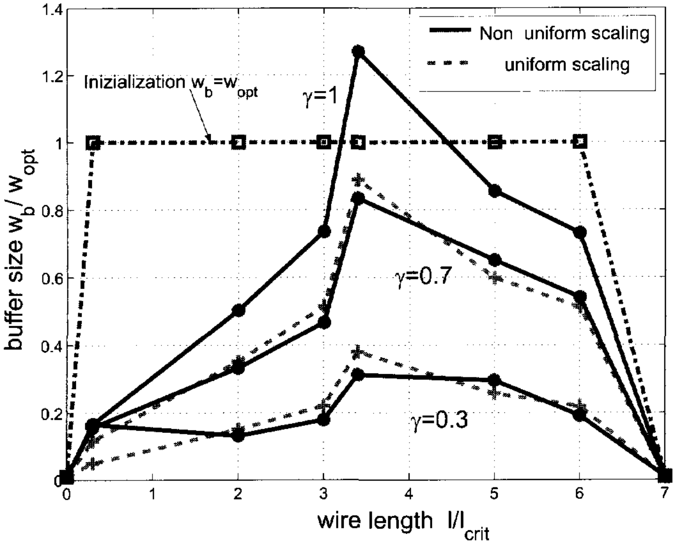


Figure 4.9. Normalized buffer sizes and for non-uniform distribution of the scaling factor for each of the buffers

Note that buffers tend to be bigger in the middle of the line, while their sizes approach the driver and receiver size for the buffers towards the ends. In this example $w_{TX} = w_{RX} = w_{opt}/100$.

In particular the gain in performance that we obtain by using our buffer size scaling with respect to the uniform scaling is determined by selecting an appropriate value for the first buffer.

4.4 Summary

We have presented two algorithms which are used to resize the buffers distributed along a wire where buffer locations are restricted and where driver and receiver (respectively driving and loading the wire) are fixed.

The first algorithm gives the optimal solution for minimum delay, provided that there are no constraints on area usage by buffers. The buffer sizing has a marginal effect in the situation where driver and receiver size are set to w_{opt}. However, the resizing method presented gives significant performance improvements (70% better then performance obtained by using all internal buffer of sizes w_{opt}) and uses less area (30% less when compared to the case when all internal buffer sizes are w_{opt}) when driver and receiver are small. In the latter situation the optimization resizes the buffers in such a way that the buffers closer to the middle of the wire assume sizes close to w_{opt} while the buffer near the receiver and driver are much smaller and similar to the receiver and driver size values.

The second algorithm allows optimal buffer sizing with limits on the total buffer area. It takes care of size buffers such that their collective area stays under the limit and the delay is as low as possible. The performance obtained by using this algorithm are significantly better (more that 20% in some cases and 8% on average for random assigned locations) than when using uniform buffer sizing. We have observed that our algorithm gives larger performance improvements for aggressive area-constraints than for moderate area-constraints cases.

Those two algorithms use simple iterations schemes which iterate with analytical expressions for the size updates. Both algorithms show fast convergence.

Chapter 5

Impact of process variations

Contents

APPROACHING deep-submicron dimensions causes the gap between the intended (or designed) layout and what is really manufactured in silicon to increase enormously, with the consequence that the performances predicted at design time will diverge from the results after real manufacturing on silicon, unless we learn how deal with the complications of modern processing. Aggressive technology scaling also introduces new variation sources and at the same time the process control during manufacturing becomes more difficult and critical to tune.

It is important to understand the impact of the relevant process variations on the performance quality metric. This can give guidelines not only on how to tune the processes, but it can also suggest alternative smart design solutions that compensate for the variations.

Coping with variations at design time has potentially big advantages in terms of time-to-market as well as in terms of costs in process control. The former because taking the right decisions early, even at system level, reduces the number of design flow iterations. As far as the latter is concerned, it may be that the reduction of the variations by process control requires too expensive machineries. Thus, the impact of variability should be compensated by innovative design tools due to the very costly nature of the process innovative techniques.

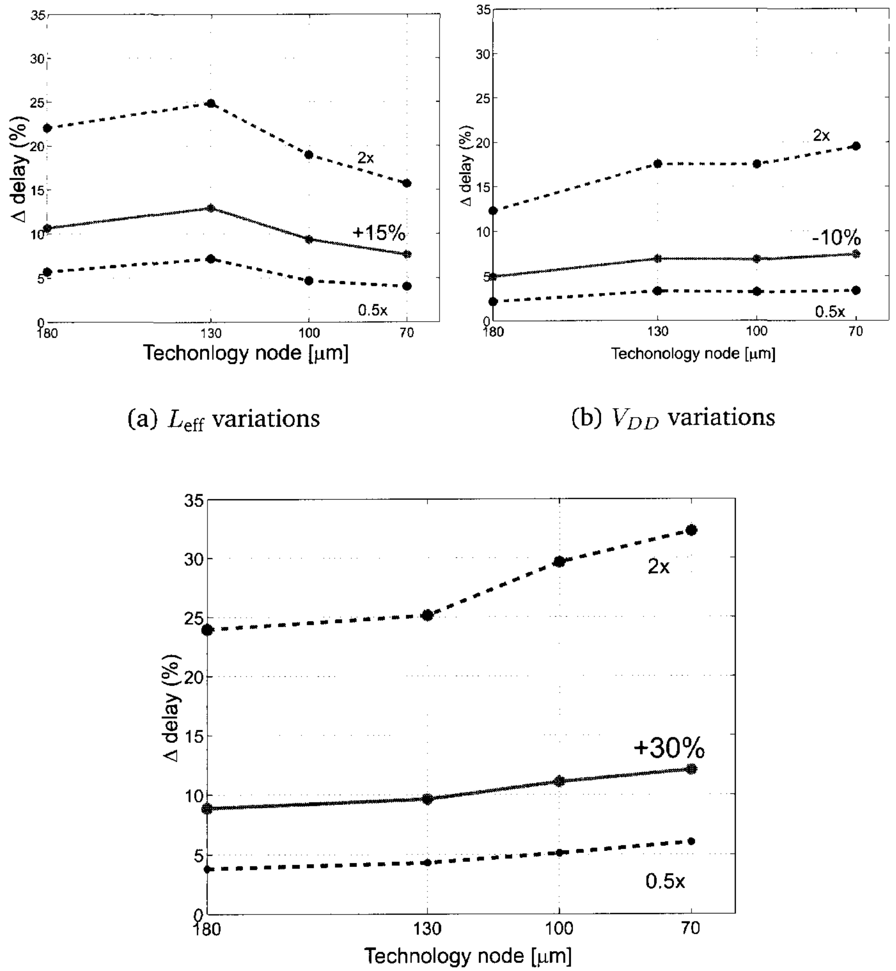
General considerations about the trend of process variations across technology scaling:

- variations (for example metal line widths, layer thicknesses and transistor minimum dimensions) while continually shrinking in absolute terms, are growing as percentage of the increasingly minuscule dimensions they affect.
- variations grow in number as processes become more complex. Correlations between different sources and a general quality figure of the process is becoming more difficult to predict.

We show in figure 5.1 an example of a sensitivity analysis of the impact of relevant device parameters on performance through different technology nodes. For this sensitivity analysis, nominal values according [1] are assumed and the delay is obtained by using SPICE simulations. The SPICE technology used for our simulations can be found in [10].

We extrapolate through technology nodes the variability value measured for typical mature $0.18\mu m$ technology ($\Delta L_{\text{eff}} = -10\%$ $\Delta V_{DD} = -10\%$, $\Delta V_{TH} = +30\%$), comparing a uniform scaling (where the variability is assumed to remain constant in percentage also for smaller feature sizes) with a 2x scaling and a 0.5x. It is possible to conclude from the figure 5.1 that the relative variability grows in percentage and it can become quite critical. In particular, we have to focus especially on keeping the V_{TH} variations under control.

It is very clear that the performance will be so dramatically affected by the variation sources, that for the designs based exclusively on optimization of the nominal process parameters, the analysis will become inaccurate and synthesis will possibly leads to wrong decisions when we deviate from the nominal situation. We can overcome this problem by choosing to design in the worst case scenario. The worst-case method identifies the extreme (i.e. worst) values of the performance resulting form the variations in its component values. Since only the extreme performance values are of interest, the detailed nature of the probability density functions neither of the performance nor of the single parameters is necessary for this analysis. When using this static time analysis two levels of conservatism are introduced. The first conservatism is introduced by the worst-case modelling in the case all parameters are simultaneously set to their worst-case



(a) L_{eff} variations

(b) V_{DD} variations

(c) V_{TH} variations

Figure 5.1. SPICE simulated sensitivity analysis of relevant device parameters for different scaling scenarios

values. By taking the extreme or corner value for each of the electrical parameters leads to overpessimistic estimation of the performance. This is essentially due to the fact that no correlation between electrical parameters is included. In other terms the scenario with all the parameters considered in their extreme worst values has really lower probability to happen or really can not happen at all. The other level of conservatism concerns the non-probabilistic way of computing the delay, i.e. the analysis is done for the single worst-case scenario, and there are possible problems of accuracy introduced on the intrinsic approximation in computing the the delay using an analytical model. The fundamental problem is that standard timing techniques have in general non-probabilistic formulations.

5.1 Statistical performance modelling

As semiconductor technology is aggressively scaled, the ”time constant” of VLSI circuits is more and more dominated by interconnect. Thus, interconnect is a key concern in the current design flows in order to improve performance and to achieve timing closure. However, interconnect resistance and capacitance in deep-submicron technologies are more and more affected by manufacturing variations. Thus, interconnect is not only limiting the performance but also increasing the performance variability [12], [43], [27]. For this reason, a new branch of research has started in statistical modelling for design optimization, taking the impact of variability on interconnect performance into account. It strives not only for optimal performance, but also for greatest tolerance to process variations [67], [63].

Since real process measurements, especially for the state of the art technologies, are not easily available, there is a wild use of TCAD-based interconnect and circuit simulators. However, because these numerical simulators are often computationally too expensive for any exhaustive search for the optimum design point, sampling techniques for the design parameter space are adopted. Then, the choice of the experimental points or in other terms the space-filling strategy is crucial, which gives rise to different techniques [20], [67] but they all remain relatively slow.

A completely different approach, which is convenient especially if the number of parameters is significant and the correlation between them is very complex to characterize, is the use of analytical models together with monte carlo methods to analyze process variations [12]. The computational costs to consider all the parameter combinations then can be compensated by the fact that design outputs can be obtained by using analytical models.

It is the objective in this chapter, to seek analytic expressions that describe the statistical properties of the design under process variations. An analytic approach could even directly produce the sensitivity of (the statistical properties of) the performance to (the statistical properties of) the design and technology parameters. Compared to the monte carlo approach, such an approach would be much faster, and would allow a much tighter design optimization loop. It can also provide a better insight in the factors involved.

In particular, we will use analytical models to estimate a quality figure of the design, based on some key technology and design parameters. When studying their statistical properties, we are dealing in general with non-linear functions of random variables. Our approach will be valid and applicable when the variability among these variables can be described by their mean and (co)variance. This is a common situation for modeling manufacturing variability, especially applicable when the process is reasonably mature. In many cases, more detailed information on the statistical distribution is not even available.

We will show that this approach can be useful for the problem of uniform buffer insertion for a point-to-point connection. As already investigated in chapter 3, there is a well defined tradeoff between performance and power consumption versus silicon area, the latter actually being proportional to the buffer size and distance ratio w_b/l. In this chapter we will show that the conclusions derived in the deterministic contents are not completely true in the presence of statistical performance variations.

Possible variation sources are:

- *spatial variations*
  We can make a first distinction of the variations on the base of the impact they have on a die, on a wafer, on a lot. We will distinguish then between within-die, intra-die variations. For circuit design optimization purposes, sources of variations can be separated in intra-wafer as well as inter-die variations [39, 56]. With the advance of the deep-submicron technologies, intra-die variations has been increasing. This is due to various processing and physics factors such as random dopant placement in the channel, spatially correlated, proximity-gate L_{gate} variation and interconnect metal thickness variation. The intra-die variations contribute significantly in the uncertainly and degradation of the circuits, requiring then a new design approach which can guarantee a performance prediction as close as possible to the silicon behavior.

- *temporal variations*
  We can distinguish between *environmental variations* and *physical variations*. The first one arise the operational behavior of the circuit subject to deviation of factors like power supply, temperature

and switching activity for nominal values. The second are related to the manufacturing process then essentially to the limited accuracy in the lithography process and to the masking limitations. In this respect we can talk also about *intrinsic variations* that are caused by the fabrication of the integrated circuits, it means that they are inherent variations of the process parameters. This definition is done in order to distinguish from the *dynamical variations* which arises only the sources that play a role only when the circuit is operational.

- *nature of variations*
  In this framework we can talk about systematic variations, or random variations.

It essential to understand that electrical and manufacturing variability cause design performance fluctuations [42] and there are variations that, even if not catastrophic for the circuits, they degrade their performance. By modelling the impact of process variations on performance we can enhance performance predictability and propose also possible solutions to control the performance variability.

More specifically, we address the problem of statistical performance prediction by proposing a practical approach to quantify the total variability on the signal propagation velocity induced by process variations.

Our framework is illustrated in figure 5.2

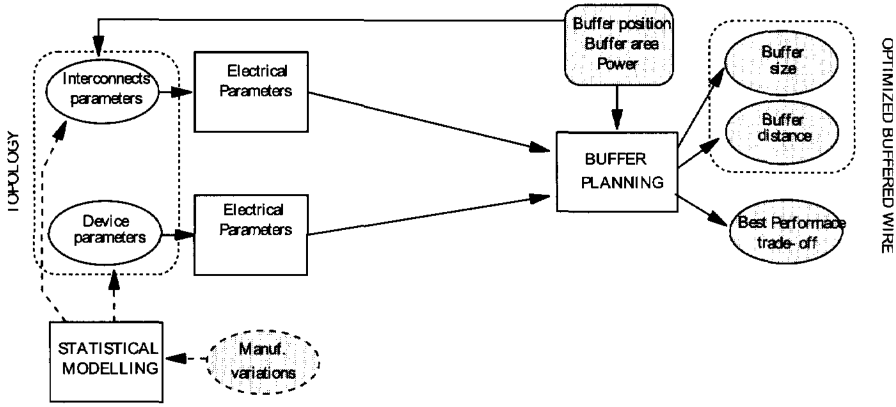


Figure 5.2. Integrate statistical modelling in buffer planning

Buffer planning for uniform point-to-point wires was up to now described in terms of size of and distance between the buffers. Possible constraints (buffer area, power, wire density) required redefining buffer planning. Most of the material presented in this thesis offers analytical solutions to satisfy those constraints with minimum performance loss. We will now propose opportune buffering which also accounts for process variation. In practice we use the buffer size and distance as a control parameter to enhance performance predictability. By combining the deterministic buffer planning with the statistical model for the process variations, new interesting trade-offs can be found which can suggest new buffer insertion strategies.

5.2 Modelling performance variability

The approach proposed [38] uses the 2^{nd} order taylor expansion of a function ϕ, which represents a generic quality figure. This is a non-linear function of the vector of random variables $P = (p_1, p_2, \ldots, p_n)$. The nominal value of the parameters is denoted by $P^0 = (p_1^0, p_2^0, \ldots, p_n^0)$. The taylor expansion around the point P^0 is then:

$$\begin{aligned}\phi(P) &= \phi(P^0) + \sum_i \frac{\partial \phi}{\partial p_i}(p_i - p_i^0) \\ &+ \frac{1}{2!}\left(\sum_i \frac{\partial^2 \phi}{\partial p_i^2}(p_i - p_i^0)^2 + 2\sum_{i>j} \frac{\partial^2 \phi}{\partial p_i \partial p_j}(p_i - p_i^0)(p_j - p_j^0)\right) \\ &+ \ldots \end{aligned} \tag{5.1}$$

where all the partial derivatives are evaluated at the point P^0. Thus, we have reduced our function to a 2^{nd} order expansion of random variables which can be treated much easier. In fact we are able for a function represented as in (5.1) to estimate its statistical behavior in the presence of random variations of its parameters.

The average (expected) value is given by $E(\phi) = \phi(P^0) + bias(E(\phi))$. A general first order expression for $bias(E(\phi))$ is from (5.1)

$$bias(E(\phi)) = \frac{1}{2}\sum_i \frac{\partial^2 \phi}{\partial p_i^2}(\sigma_{p_i})^2 + \sum_{i>j} \frac{\partial^2 \phi}{\partial p_i \partial p_j} cov(p_i, p_j) \tag{5.2}$$

where $\sigma(p_i)$ is the standard deviation of p_i and $cov(p_i, p_j)$ is the covariance between p_i and p_j. Thus, the bias can be in general non-zero, except for linear models without correlation among the parameters.

For a simple function $z = ax + by$ where a, b are constants and x, y are two independent statistical variables with means μ_x and μ_y and variances σ_x^2 and σ_y^2, the total variance of z becomes $\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2$ [38]. This result can be generalized for a generic number of random variables which together with (5.1) gives:

$$\sigma^2(\phi(P)) = \sum_i (\frac{\partial \phi}{\partial p_i})^2 (\sigma_{p_i})^2 \tag{5.3}$$

If the variables are not independent, this expression can be modified to include a term for covariances, but that will not be necessary for our current purpose. In fact we will evaluate in our study performance subject to process variations, by propagating primitives for which the parameters can be considered independent.

In this chapter we present the results in normalized form as the ratio between the standard deviation (square root of the variance) and the expected value. The notation $\sigma_n(\phi(P))$ will denote the following ratio:

$$\sigma_n(\phi(P)) = \frac{\sqrt{\sigma^2(\phi(P))}}{E(\phi(P))} \tag{5.4}$$

The choice of this estimation of the variability is convenient since often also the parameters variations are measured in terms of their deviation from their nominal value.

In the next section, details about the modelling of the interconnects and the device are given. We will focus on the effect of variability on the performance, i.e. the signal velocity propagation. The standard deviation will then be estimated using (5.4). This variability estimation can indicate a new approach to uniform buffer planning for point-to-point connections which aims at minimizing this performance variability.

5.3 Impact of interconnect variability

Our interconnect model, see section 3.1, is basically characterized by r and c, the resistance and capacitance per unit length. They exhibit statistical variation due to process variations, and we could perform our analysis using resistance and capacitance as statistical variables if we would know their mean and (co)variance. We cannot neglect in our analysis resistance and capacitance correlations. For this reason we have decided to model r and c in terms of more primitive parameters that are directly linked to the geometry and the fabrication process. Those primitives can be considered as first-order approximation statistically independent or they have a well-determined correlation. The idea is to propagate the statistical behavior of those process parameters according to the method presented in section 5.2, to predict their effect on the performance estimation.

We will use a particularly simple model to relate r and c to more primitive parameters. The only requirement is that this model is analytic, so that the necessary derivatives to perform the steps of section 5.2 exist. *A more complex model, instead of the parallel plate model below, could somewhat improve the numerical accuracy of this method but will not change the trends and the insights that can be obtained.* Our interconnect model and the definition of some basic parameters are illustrated in figure 5.3.

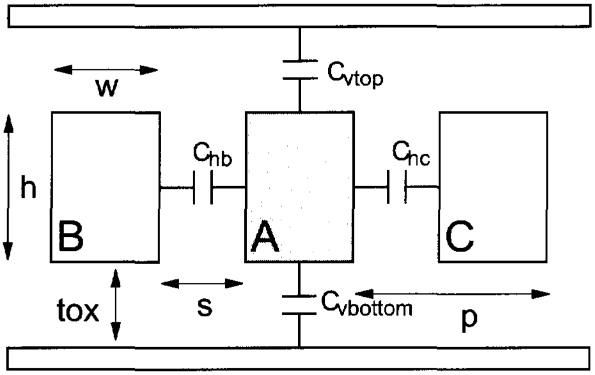


Figure 5.3. Interconnect parameters

In this model, we calculate interconnect resistance per unit length as $r = \rho/(w * h)$. We calculate the interconnect capacitance per unit length as

$$\begin{aligned} c &= c_{vbottom} + \Gamma c_{vtop} + SF(c_{hb} + c_{hc}) \\ &= \epsilon((1+\Gamma)\frac{w}{t_{ox}} + 2SF\frac{h}{s}) \end{aligned} \qquad (5.5)$$

where SF is the so-called switching factor [32] and $\Gamma = 0$ if the wire is on the top interconnect layer and $\Gamma = 1$ if this wire is on an intermediate layer. The switching factor is used to model the Miller effect due the lateral capacitance. This switching factor depend on the signal alignment between adjacent parallel wires and its value can vary between 0 and 2 d if the transition time of the signals are ideally zero. If also transition time are considered then it has been shown that this range of variations become larger, typically between 0 and 3.

In the rest of this chapter, we set $\Gamma = 0$ (top layer) and we set to a constant value $SF = 1$. However, it is possible to consider the switching factor as a statistical variable by itself when desired.

The parameters that we model as statistical variables are in general correlated. In fact, the complexity of the actual fabrication processes make those correlation values strongly process dependent and very difficult to extract. The experiential measures of those parameters give the exact value of the parameter variations, but cannot relate with sufficient accuracy those variations to the set of specific lithography steps which have generated them. This consideration shifts actually the problem of statistical analysis to the correlation estimation. We are not going to address the problem on how to derive correlations, because we are not able to have sufficient process data to perform this kind of analysis. We assume only that if we have a correlation estimation we are able to deal with it to have more accurate performance estimation. Our method focus on the propagation of statistical variables to achieve performance variability estimation.

Under the assumption of constant wire pitch p, we can easily identify the correlation between wire width w and wire spacing s such that $w = p - s$ and then they have correlation equal -1. This special case allows for elimination of s from the set of parameters. We consider the other correlations between parameters as second order effects, and thus negligible in this study. This is in general not true and can be an arbitrary choice since not detailed process data where available. However, once the correlations among these low level parameters are known, they can be included if necessary, by including the appropriate covariance term in (5.3) [38].

We use typical parameters for a $0.18\mu m$ technology. Nominal values for interconnect parameters with the relative standard deviation used in this chapter, are summarized in table 5.1.

Table 5.1. Interconnect parameters

Parameter	**Nominal value**	σ_n
ρ	$2.2\mu\Omega.cm$	0.2
ϵ	3.5*8.85pF/m	0.2
w	$0.7\mu m$	0.2
h	$1.2\mu m$	0.2
t_{ox}	$1\mu m$	0.2
SF	1	0
p	$1.4\mu m$	0

The assumption that all parameters have the same normalized standard deviation $\sigma_n = 20\%$, permits to classify the parameters in order of importance for their contribution to the overall standard deviation. It is important to stress that this assumption is made only in order to give a quantitative example, but the framework that has been created can be used for each standard deviation derived from realistic process measurements. See [64] for our online version of the model that allows to adjust the different parameters of the model.

Now, we begin our statistical analysis by expressing our performance metric, v^{-1} in (3.7), in terms of the parameters from table 5.1. Subsequently, we evaluate (5.1) and we collect terms to make the dependence on the controllable design parameter space (w_b, l) explicit. A partial result is as follows (the complete expressions (including the 2^{nd} order derivatives) have been omitted for brevity):

$$k_\rho = \frac{\partial v^{-1}}{\partial \rho} = a_\rho l + b_\rho w_b \tag{5.6}$$

$$k_\epsilon = \frac{\partial v^{-1}}{\partial \epsilon} = a_\epsilon l + \frac{b_\epsilon}{w_b} \tag{5.7}$$

$$k_w = \frac{\partial v^{-1}}{\partial w} = -a_w l + \frac{b_w}{w_b} - c_w w_b \tag{5.8}$$

$$k_h = \frac{\partial v^{-1}}{\partial h} = -a_h l + \frac{b_h}{w_b} - c_h w_b \tag{5.9}$$

$$k_{tox} = \frac{\partial v^{-1}}{\partial tox} = -a_{tox} l + \frac{b_{tox}}{w_b} \tag{5.10}$$

$$k_{SF} = \frac{\partial v^{-1}}{\partial SF} = a_{SF} l + \frac{b_{SF}}{w_b} \tag{5.11}$$

Here, k_x is the sensitivity of v^{-1} to parameter x, and a_x, b_x and c_x are the coefficients for this sensitivity with respect to l and w_b. The expressions for these sensitivity coefficients a_x, b_x and c_x are presented in detail in appendix B.

Subsequently, we can apply (5.2), (5.3) and (5.4), which we do first for each parameter separately. The result is in figure 5.4 which quantifies the impact each of the interconnect parameters on the standard deviation of v^{-1}, as a function of the normalized buffer size with the normalized segmentation length as a parameter. These latter normalizations are with respect to the optimal unconstrained scenario with w_{opt} and l_{crit} according to (3.3)-(3.4). We assume that the segmentation length satisfies $l_{crit} < l < 2l_{crit}$ and that the buffer size satisfies $0.1w_{opt} < w_b < 3w_{opt}$.

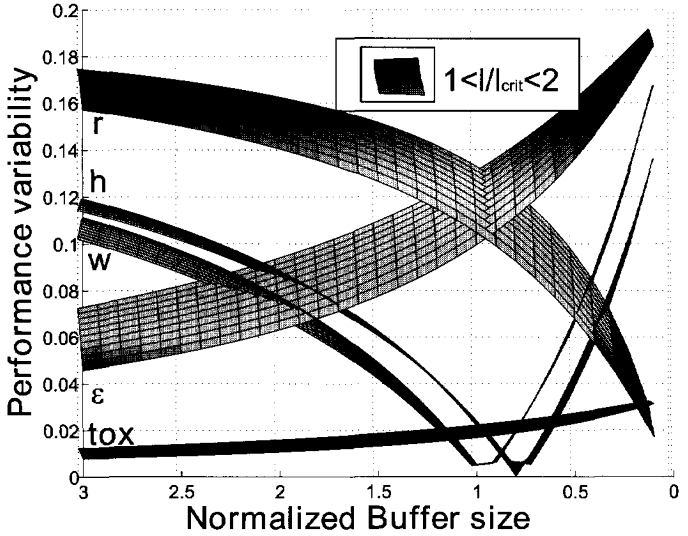


Figure 5.4. Performance variability for each of the interconnect parameters as a function of the buffer size

The standard deviation of the performance under the interconnect variations according to (5.4) is then given by

$$\sigma_n(v^{-1})(l, w_b) = \frac{\sqrt{\sum_i k_i^2(l, w_b)\sigma_i^2}}{v^{-1}(l, w_b)} \tag{5.12}$$

The total contribution of the interconnect variations on the standard deviation of v^{-1} is presented in figure 5.5.

The standard deviation, presented in figure 5.5, appears to be a monotonic function which decreases slowly moving from larger to smaller buffer distances. This is understandable, because with short segmentation lengths the delay is dominated by the delay of the buffers, and buffer delay variation is not modelled here. Later in this chapter we will include in the estimation of the performance variability also the device variations.

The dependence of the standard deviation on the buffer size seen in figure 5.5 is more interesting from an optimization point of view. In fact, the performance variability is considerable for large buffer sizes, then decreases till reaching a minimum point and then again increases fast for small buffer sizes. This opens the question on where to find the size value which gives the minimum performance variability or in other terms the

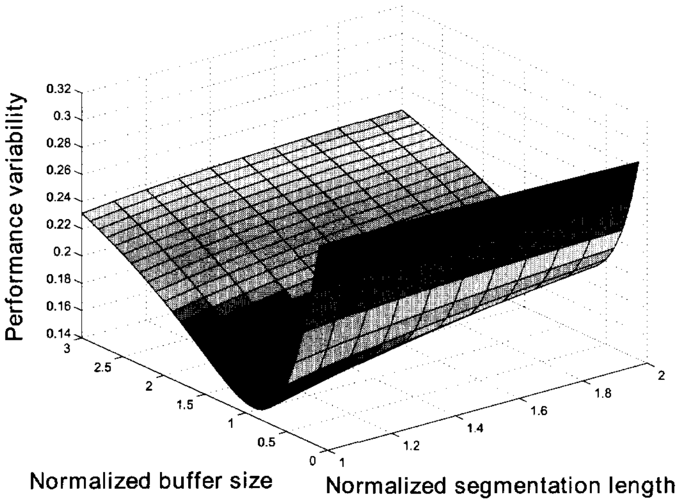


Figure 5.5. Total impact of interconnect variability on performance

higher performance prediction.

The elmore delay model that we use, indicates that for large buffer or for a fixed buffer distance the performance variability depends almost exclusively on the wire resistance variability. For small buffer sizes, however, the variability values are dominated by the capacitance. The exact shape of figure 5.5 (and the relative magnitudes in figure 5.4) depends strongly on the assumed standard deviations of each of the parameters, and also on the nominal value of the switching factor, (see table 5.1). We are not able to provide a closed-form solution to find the buffer size which minimizes the performance variations, but the whole model is analytic and the optimal value can be found by evaluating the performance variability in the space (w_b, l) and choosing the minimum. However, in general this minimum of performance variability does not correspond to the point which maximizes the deterministic performance.

As already mentioned, the values in table 5.1 only serve to illustrate the methodology. With the parameters from table 5.1, the tuning of the buffer size to a value of $0.8w_{opt}$ corresponds to a v^{-1} standard deviation of 15%. Here it must be noted that large buffers of size $w_b > w_{opt}$ (normalized buffer sizes > 1) are never advantageous when variability is neglected [25]. This analysis shows that also when interconnect variability is considered, such large buffers only reduce the predictability. The useful range of w_b thus seems to remain restricted to $w_b < w_{opt}$. However, for really small buffer sizes the performance variability may also become unacceptably large. A validation of our statistical model for performance against monte carlo analysis is shown in figure 5.6. The number of samples used in this example is 1000 for each of the parameters and the performance variability is estimated as a function of the normalized buffer size and for a buffer distance of $l = l_{crit}$. We have also studied other comparisons for different buffer distances, and they show similar correspondence.

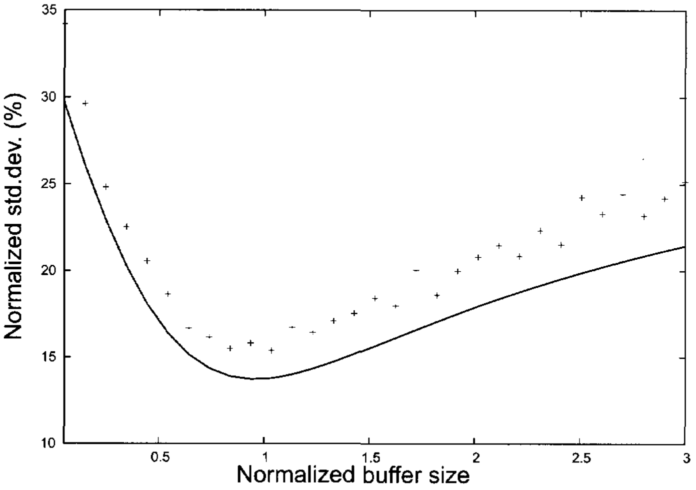


Figure 5.6. Comparison between our analytical model for variability (solid line) and a 1000 samples monte carlo analysis

Because we neglect higher order terms in the taylor expansion of the performance, it is expected that our model underestimates the variability when compared to monte carlo.

Our next step will be to study performance variability resulting from device parameter variability. When we subsequently combine the results of this and the next section, the overall conclusion will be similar to the one based on the analysis above.

5.4 Impact of device variability

The delay model from section 3.1 needs r_0, c_0 and c_p as basic device parameters. Like r and c in section 5.3, we will express them in terms of the technology and design related parameters using a simple first order model.

The input capacitance per transistor width, c_0, can be written as

$$c_o = c_{ox}.L_{\text{eff}} \tag{5.13}$$

where c_{ox} is the gate capacitance per unit area and L_{eff} is the effective channel length. Thus, we have c_{ox} and L_{eff} as two parameters for our model. We could further decompose c_{ox} into thickness and permittivity, but since the latter is usually very well controlled this is unnecessary.

We will eliminate c_p from the statistical model, by assuming that it is perfectly correlated (proportional) to c_0.

A first order approximation for the on-resistance of a mos transistor is derived in the situation where the n-mos is in linear operation. In this case $V_{GS} > V_{TH}$ while V_{DS} is small. Then the current $I_{DS} = K_n W/L_{\text{eff}}((V_{GS} - V_{TH})V_{DS} - V_{DS}^2/2)$ and for small values of V_{DS} the quadratic term can be neglected. Thus, $I_{DS} \propto V_{DS}$ which indicates indeed that we are in the linear region. The parameter $K_n = c_{ox}\mu$ is process dependent. Assuming $V_{GS} = V_{DD}$, the resistance is:

$$R_{tr} = \frac{V_{DS}}{I_{DS}} \approx \frac{L_{\text{eff}}/W}{\mu c_{ox}(V_{DD} - V_{TH})} \tag{5.14}$$

If the n-mos is considered to be in saturation, then $V_{DS} > V_{DSAT}$ with $V_{DSAT} = V_{GS} - V_{TH}$. The saturation current is $I_{DSAT} = K_n W/2L_{\text{eff}}(V_{DD} - V_{TH})^2$ and it does not depend on V_{DS}. The on-resistance will be given by $R_{tr} = V_{DSAT}/I_{DSAT}$.
The value of r_0, is

$$r_0 = k_{r_0} R_{tr} * W \tag{5.15}$$

where k_{r_0} is a calibration parameter which can be extracted using spice, it is typically smaller than 2. It accounts for the non-linearity of the device, as well as the contact resistances. The nominal values, assuming a $0.18\mu m$ technology, for the resulting set of statistical variables with their standard deviation are presented in table 5.2.

Table 5.2. Device parameters

Parameter	**Nominal value**	σ_n
c_{ox}	$7.342e - 14F/\mu m^2$	0.2
V_{DD}	$1.8V$	0.2
V_{TH}	$0.35V$	0.2
L_{eff}	$0.115\mu m$	0.2

The derivatives of v^{-1} with respect to the device statistical variables can also be rearranged as an explicit function of the controllable design parameter space (w_b, l) as follows

$$k_{c_{ox}} = \frac{\partial v^{-1}}{\partial c_{ox}} = -\frac{a_{c_{ox}}}{w_b} + b_{c_{ox}} w_b \tag{5.16}$$

$$k_{V_{DD}} = \frac{\partial v^{-1}}{\partial V_{DD}} = -\frac{a_{V_{DD}}}{l} - \frac{b_{V_{DD}}}{w_b} \tag{5.17}$$

$$k_{V_{TH}} = \frac{\partial v^{-1}}{\partial V_{TH}} = \frac{a_{V_{TH}}}{l} + \frac{b_{V_{TH}}}{w_b} \tag{5.18}$$

$$k_{L_{\text{eff}}} = \frac{\partial v^{-1}}{\partial L_{\text{eff}}} = \frac{a_{L_{\text{eff}}}}{l} + \frac{b_{L_{\text{eff}}}}{w_b} + c_{L_{\text{eff}}} w_b \tag{5.19}$$

Again, k_x, a_x and b_x are the appropriate sensitivity coefficients, similar to those of (5.6)-(5.11). The expressions for these coefficients are presented in appendix B. The effect of each of these parameters separately on the standard deviation of v^{-1}, as function of w_b with l as a parameter, is presented in figure 5.7. The total contribution of the device variations on the standard deviation of v^{-1} is presented in figure 5.8.

Small buffer sizes lead to large performance variability, while the selection of buffer distance does not influence it significantly.

5.5 Combined device and interconnect variability

We can evaluate the impact on the inverse of velocity of the device and interconnect variability together and plot the standard deviations in the design parameter space composed by normalized buffer size and segmentation length. The result is shown in figure 5.9.

This figure still shows a weak dependence of the variability on the length of the segment. This is desirable because by placing a suboptimal number of buffers on the global wire, the variability of v^{-1} is only marginally affected.

Performance variability then seems to depend mainly on the value of the buffer size. In particular for really small buffer sizes, the variability can become really large. We can estimate at least for this example that for $0.5 < w_b/w_{opt} < 1.5$ the increased standard deviation actually remains

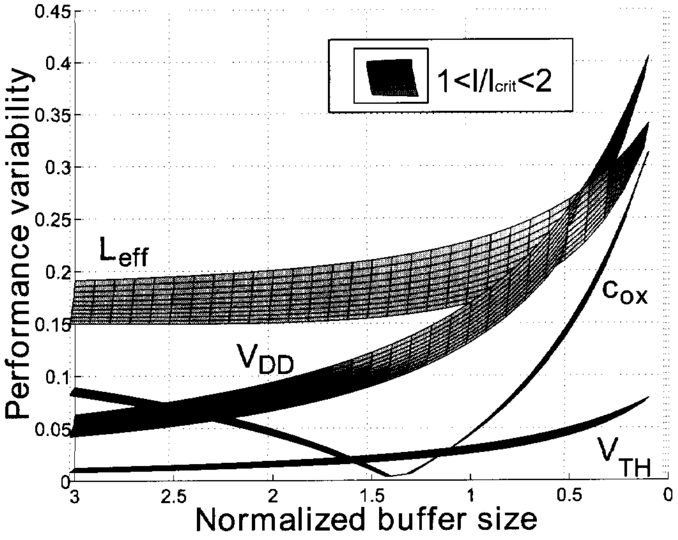


Figure 5.7. Performance variability for each of the device parameters as a function of the buffer size

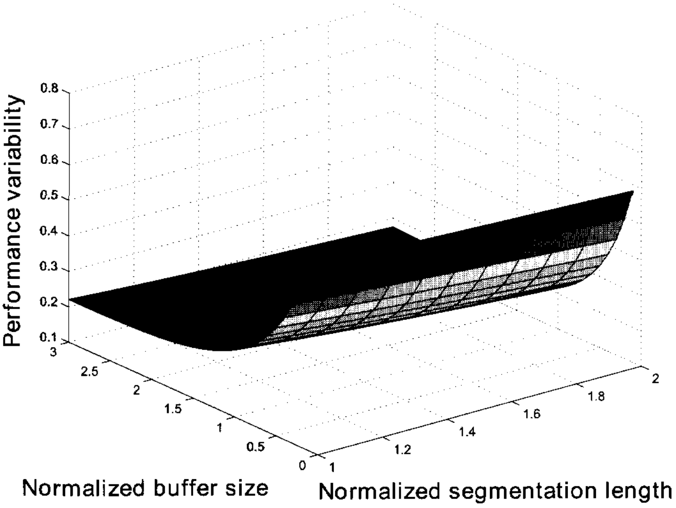


Figure 5.8. Total impact of device variability on performance

below 10%, but for still smaller sizes the standard deviation is increasing very rapidly.

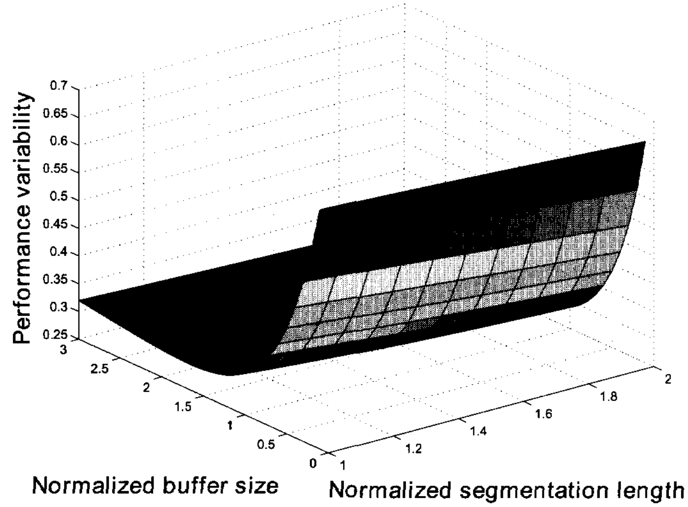


Figure 5.9. Total effect of variability

Thus, figure 5.9 indicates that small buffer sizes are very unfavorable from a predictability point of view. This actually could augment the results from [25]. For example, that paper presents a design point corresponding to 57% of the maximum performance that is obtained using only 10% of the silicon area necessary for maximum performance. This design point corresponds to $w_b = 0.34w_{opt}$ and $l = 3.4l_{crit}$. This point is not present in figure 5.5, because the normalized segmentation length (l/l_{crit}) only runs through 2. However, in the point $(0.34, 2)$ the standard deviation of v^{-1} is already 42%, and it would be about as high in the $(0.34, 3.4)$ point. It is clear that this point, under the assumption of the standard deviations from tables 5.1 and 5.2, is unacceptable for high parametric yield.

5.6 Conclusions

We have addressed the problem of statistical performance prediction by proposing a practical approach to quantify the total variability on the signal propagation velocity induced by process variations. This will give the possibility to develop a "variation aware" buffer insertion methodology for global wires, but we refer for this to chapter 6. By combining the deterministic buffer planning with the statistical model for the process variations, new interesting trade-offs can be found which can suggest new buffer insertion strategies. For example, we show that in a statistical setting the cost of reducing die area is not only in terms of performance but also in terms of increased variability. We have developed a general analytic approach for estimating the statistical properties of a uniformly buffered uniform RC line for global interconnect. The model takes as input the mean and (co)variance of the controllable design and technology parameters. It was concluded that a fully deterministic model to tune the buffer size and segment length might cause yield problems because of excessive variability.

In this work, several simplifications and assumptions have been made. For example, a first order parallel plate model was used for the interconnect capacitance and we did not consider inductive effects nor correlation among the input parameters. However, these are non-essential simplifications and can be alleviated when necessary for a particular purpose. All that is required is that the model is analytic, such that the 2^{nd} order taylor expansion exists.

Of course, this approach looses its numerical validity when the models are too rough or when the variability becomes so large that the taylor expansion becomes inaccurate. In the range of validity of this model, its advantages include a possibly tight and fast design optimization loop and effective 'what-if' analysis.

Furthermore, this approach can be extended to directly give design-for-yield solutions. For example, we can find the minimum buffer area for a certain performance under yield constraints.

Chapter 6

Robust buffer planning

Contents

IN this chapter we apply robust design methodologies for buffer planning in presence of process variations. Those techniques for design optimization have already been used in different manufacturing industries, not necessarily operating in the semiconductor field, to enhance quality of the product. Robust design aims at reaching the design point and the process conditions which are minimally sensitive to the various causes of variation, such that a high quality product with low development and manufacturing costs can be produced. In this context taguchi's parameter design [60, 59] is an important tool for robust design and it has been used successfully by many companies in Japan, USA and elsewhere. The major goal of parameter design is to reduce quality variation, by minimizing the influence of noise. All the undesirable and uncontrollable sources that can cause deviation from the target values in a product's functional characteristics are called noise. The overall quality optimization should be robust with respect to all noise factors.

According to Taguchi, no matter how the quality of the product is measured, the quality characteristics are divided into the following three categories:

- *Nominal-is-best*: if there is finite target value to achieve.
- *The-smaller-the-better*: if results become worse as the measured value increases.
- *The-larger-the-better*: as the value becomes larger, the quality increases.

For each of those characteristic Taguchi has introduced specific loss functions, which combine opportunely function specification to financial loss. The minimization of those functions contributes to the quality improvement.

We use the taguchi approach to find buffer insertion schemes robust to process variations. In particular, we focus on how to determine the best design point in buffering a point to point connection which is able to meet both the performance specifications and remains robust against process variations. In this context, buffer size and segmentation are used as control parameters to enhance variability tolerance.

We will show that reducing of performance variability by buffering degrades the average performance. In particular, buffering which has as objective the performance maximization is not the same as buffering to enhance robustness. Dependent on the design specification we have to decide which optimization is more important and to trade the cost/benefit in selecting the buffering. This chapter address this trad-off issue.

Two techniques are introduced in this chapter. The first one is based on sampling the variability space to obtain a population of samples and on performing a design optimization specific for each of those samples. This optimization for each of the samples will be done analytically, which will keep the computational effort acceptable. Of course, the accuracy of the results depends intrinsically on the accuracy of the calibrated delay model used. We will enhance the performance variability tolerance by selecting the buffering with the higher probability (or higher frequency to occur) among the set of samples considered.

The second technique uses statistical characteristics (expected value and variance) of the performance, to derive guidelines for buffering considering parametric yield. Yield-centric buffering has the advantage that it does not need any sampling and does not have to repeat the optimization procedure for each of those samples. We will show how the yield can be trade for total buffer area. A comparison between buffer planning for yield under process variations and buffer planning for maximum speed without any process variations suggests that buffering guidelines for the deterministic case are no longer valid under process variations.

6.1 Enhancing robustness

We propose a model-based optimization which provides the most suitable buffer insertion for a given variability scenario.

Process variations have effect on the performance predictability. To improve tolerance against process variations we want to tune buffer size and distance in such a way that performance variation is minimized. However, the best solution to enhance tolerance against process variations is not necessarily the one maximizing nominal performance. In fact the absolute maximum performance is achieved when buffer size and distance are w_{opt} and l_{crit} as presented in (3.3) and (3.4), and any other buffer planning leads to performance degradation. Of course, we would like to remain with our average performance as close as possible to this maximum performance.

Buffer size and the length of the segments are used as control parameters to improve the robustness against electrical parameter variations. We consider electrical parameter variations, because those parameters directly influence the performance metric that we choose. Those variations originate from process variations (geometries) and environmental variations (T, V_{DD}), (see chapter 5). Obviously, electrical parameter variations exhibit correlations with each other.

The problem of process parameter correlations can be approximated analytically by propagating the first-order term of the expressions containing uncorrelated process parameters. Electrical parameters are then obtained by using analytical expressions in process parameters as shown in the chapter 5.

Each set of random values of the parameters constitutes a sample in the space of variations, and for each sample we will have to evaluate the resulting performance and its deviation from the nominal situation. Buffer insertion is planned for each sample. This implies that the value of buffer distance and size depends on the sample chosen.

We will derive the optimal buffering for each sample in the variation space and we will decide which is the most suitable buffering, that is the buffer distance and size having the higher probability of occurring.

6.2 The parameter variation space

In this section we model the performance deviation from the nominal situation. Suppose to perform a single experiment. Because of the process variations we will have electrical parameters which will differ from the nominal values and consequently also the estimated performance will differ from the one computed in the nominal situation when no variations are present. We compensate for this performance deviation by tuning buffer size and distance.

We describe in figure 6.1 the performance degradation in presence of electrical parameter variations. d^*/l_{crit} is the absolute minimum delay per unit length or in other terms the inverse of signal propagation velocity. This absolute minimum is achieved by using buffer distance l_{crit} and buffer size w_{opt} as explained already in chapter 3. In presence of electrical parameter variations this performance degrades such that $d/l > d^*/l_{crit}$, and buffer distance and size have to be tuned. Using these new values of size and buffer to estimate the delay d_{NOM}/l when all the electrical parameters are set to their nominal values, we see that this delay value increases when compared to d^*/l_{crit} since buffer distance and size differ from l_{crit} and w_{opt}.

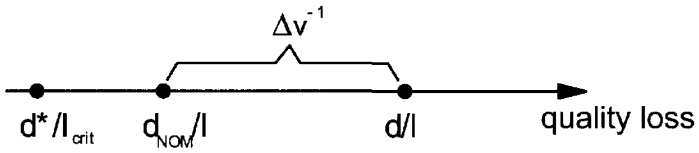


Figure 6.1. Quality loss in presence of electrical parameter variations

Figure 6.1 shows that minimization of the delay per unit length does not correspond to the minimum of its nominal value.

Let us define a nominal point in our space as (r_0, c_0, r, c) and a sample which is a (r_0', c_0', r', c') due to process variations. The deviation of each of those electrical parameters from the nominal value will be denoted by $(\Delta r_0, \Delta c_0, \Delta r, \Delta c)$.

We define as $v_{NOM}^{-1} = v^{-1}(r_0, c_0, r, c)$ as the nominal performance and we denote its deviation exclusively due to r_0 variation with

$$\Delta v^{-1}(r_0) = v^{-1}(\Delta r_0, c_0, r, c) - v_{NOM}^{-1} \tag{6.1}$$

By adopting a similar notation for the other electrical parameters and assuming that the effect of the manufacturing variations on the performance can be modelled in terms of the sum of the contribution of each electrical parameter variations, the total performance variation is:

$$\Delta v^{-1}(r_0, c_0, r, c) = \Delta v^{-1}(r_0) + \Delta v^{-1}(c_0) + \Delta v^{-1}(r) + \Delta v^{-1}(c) \tag{6.2}$$

This model is a first order approximation to estimate performance deviation form the nominal situation. It is valid for small variations of the electrical parameters. Moreover, since the electrical parameters are linear in Elmore's delay formula, the delay variation can be expressed as a linear combination of the electrical parameter variations (i.e. Δr_0, Δc_0, Δr, Δc) once (3.7) is used in (6.2). Using the latter substitution and rearranging the terms, we can expose the performance variation explicitly dependent on the buffering by w_b and l:

$$\begin{aligned}
&\Delta v^{-1}(r_0, c_0, r, c) = && (6.3)\\
&\quad = (ac\Delta r + ar\Delta c)l + b(\Delta r_0 c_0(1+\alpha_p) + \Delta c_0 r_0(1+\alpha_p))\frac{1}{l} && (6.4)\\
&\quad + (bc_0\Delta r + br\Delta c_0)w_b + (br_0\Delta c + b\Delta r_0 c)\frac{1}{w_b} = && (6.5)\\
&\quad = a_1 l + \frac{a_2}{l} + b_1 w_b + \frac{b_2}{w_b} && (6.6)
\end{aligned}$$

$a_1 = ac\Delta r + ar\Delta c$	$a_2 = b(\Delta r_0 c_0(1+\alpha_p) + \Delta c_0 r_0(1+\alpha_p))$
$b_1 = bc_0\Delta r + br\Delta c_0$	$b_2 = br_0\Delta c + b\Delta r_0 c$

Note that the latter equation has the same dependence on w_b an l as the inverse of velocity, see (3.7), and that it is the sum of two functions of one variable where each of the function has an absolute minimum. The effect of the parasitic capacitance is considered to be negligible and thus $\alpha_p = c_p/c_0 = 0$. This is of course only an approximation and it in general not true. However it allows us to determine analytically the values for a buffer size and distance that minimize performance deviation. We are not interested in an accurate analysis and estimation of the performance deviation, but we are looking for a practical way to tune buffer size and distance that reduces the performance deviation. If we assume that Δr_0, Δc_0, Δr and Δc are positive then so are a_1, a_2, b_1 and b_2. This assumption implicitly means that we assume a worst-case situation where all the electrical parameter deviations contribute positively to the total performance deviation.

We can derive an analytical expression for buffer size w_b^Δ and segmentation length l^Δ by taking the first derivative equal zero of (6.6) with respect to w_b and l and solving them in w_b and l such that:

$$w_b^\Delta = \sqrt{\frac{b_2}{b_1}} \qquad l^\Delta = \sqrt{\frac{a_2}{a_1}} \tag{6.7}$$

In the next sections we focus on the study of the differences of the buffer planning obtained by using different optimization objectives. Those objectives are:

- *Minimum* v^{-1}: minimization of the inverse of performance $v^{-1}(r_0', c_0', r', c')$ in the worse case situation. The worse case situation is achieved when, for a given value of the buffer size and distance, each of the electrical parameters increases to the total performance deviation.

- *Minimum* Δv^{-1} minimization of the performance deviation respect to v_{NOM}^{-1}. Nominal situation occurs when no electrical parameter variations are considered.

- Minimum $v_{NOM}^{-1}.\Delta v^{-1}$ minimization of an empirical cost function, combination of performance and its variability.

We will see that two first cases are extreme cases. In fact, the buffering which minimizes only the nominal performance leads to high value of variation, affecting then the quality of the performance prediction. On the other hand if we plan a buffering to minimize performance variation this will result in high nominal performance loss. It exists then a trade-off between the performance loss and the tolerance to variation. The third criteria proposes a buffering which is an intermediate solution between those two extremes.

Minimum v^{-1}

This optimization objective aims at maximize performance when the combination of the electrical parameters causes the maximum deviation from its nominal value. This corresponds to a "minimum-the better" problem in the traditional classification of quality problems. For the reasons already explained about the Elmore's delay model used in chapter 2, it is convenient to solve the minimization of the inverse of performance (i.e.

delay per unit length) instead the maximization of performance. Thus, we will minimize $v^{-1}(r_0', c_0', r', c')$ as in (3.7) where the electrical parameters are $r_0' = r_0 + \Delta r_0$, $c_0' = c_0 + \Delta c_0$, $r' = r + \Delta r$ and $c' = c + \Delta r$. This can be considered as the worse case situation, or in other terms the situation which leads to the larger performance deviation in presence of electrical parameters variations. This situation, according to (6.2) occurs when Δr_0, Δc_0, Δr and Δr_0 are assumed to be positive. Note that this kind of situation is pessimistic and has low probability to occur. Analogously to the minimization of v^{-1} presented in section 3.1, the optimal buffer size and segment with w_b^{wc} and l^{wc} are:

$$w_b^{wc} = \sqrt{\frac{(r_0 + \Delta r_0)(c + \Delta c)}{(r + \Delta r)(c_0 + \Delta c_0)}} = w_{opt}\sqrt{\frac{(1 + \sigma_{r_0})(1 + \sigma_c)}{(1 + \sigma_r)(1 + \sigma_{c_0})}} \tag{6.8}$$

$$l^{wc} = \sqrt{\frac{b/a(r_0 + \Delta r_0)(c_0 + \Delta c_0)(1 + \alpha_p)}{(r + \Delta r)(c + \Delta c)}} = l_{crit}\sqrt{\frac{(1 + \sigma_{r_0})(1 + \sigma_{c_0})}{(1 + \sigma_r)(1 + \sigma_c)}} \tag{6.9}$$

where l_{crit} and w_{opt} result from minimizing v_{NOM}^{-1}, and σ_x denotes the relative deviation $\Delta x/x$ of the parameter x respect to its nominal value.

Minimum Δv^{-1}

The minimization of the performance deviation for given electrical parameter values can be analytically found by taking the first derivative equal zero in w_b and in l of Δv^{-1} of (6.6)

$$w_b^{\Delta} = \sqrt{\frac{b_2}{b_1}} = w_{opt}\sqrt{\frac{\sigma_c + \sigma_{r_0}}{\sigma_r + \sigma_{c_0}}} \tag{6.10}$$

$$l^{\Delta} = \sqrt{\frac{a_2}{a_1}} = l_{crit}\sqrt{\frac{\sigma_{c_0} + \sigma_{r_0}}{\sigma_r + \sigma_c}} \tag{6.11}$$

where σ_x is the relative deviation respect to the nominal value of the electrical parameter x. Buffer distance and size which minimize performance deviation are presented in terms of the l_{crit} and w_{opt}, buffer distance and size respectively resulting from the minimization of v_{NOM}^{-1}. This problem can be seen as "nominal the best" approach in robust design.

Minimum $v_{NOM}^{-1}.\Delta v^{-1}$

We know how to select the buffering which minimizes delay per unit length and we know also how to buffer if the performance deviates from the nominal situation because under process variations the electrical parameters differ from their nominal values. Thus, maximization of nominal performance has to be distinguished from the minimization of the performance deviation and those optimizations result in optimal buffer planning with optimal buffer size and distance values which are specific of the optimization objective selected.

If we want to decide about an optimal buffer strategy, should the buffer planning be optimal for the performance of for its deviation? The answer is strongly dependent on the real entity of the variations involved and on the target specifications of the design.

In both the optimizations shown in the previous subsections the expressions for segment lengths and for the buffer sizes are analytically found and they are independent from each other.

In the nominal situation where no process parameters are considered and consequently the electrical parameters have their nominal values, optimal buffer planning for performance is given by l_{crit} and w_{opt} which represents the absolute maximum performance achievable. In presence of process variations, the minimization of the performance deviation requires the buffering l^{Δ} and w_b^{Δ} as in (6.11) which differs from l_{crit} and w_{opt}. The use l^{Δ} and w_b^{Δ} degrades the nominal performance. We would like then to find a buffer planning which can maximize the nominal performance and it keeps as small as possible the performance deviation.

In the field of quality engineering undesirable and uncontrollable sources that can cause deviation from the target values in product's functional characteristic are denoted typically as *noise*. The overall quality system should be designed such that the product derived results robust respect to noise factors. A formalization that was introduced by Taguchi which aims at optimization in design for quality is actually borrowed from the field of communication engineering uses the *signal to noise (SN) ratio* as quality characteristic of choice. Taguchi adapt this concept to the field of design of experiments. In particular this kind of quality metric is used in improving quality by variability reduction. In figure 6.2, to achieve the goal intended in communication equipment, signals enter into the system and an output accordingly results. However the noise interferes with the system and the intended goal is not usually achieved.

The ratio between the power of the signal and the power of the noise is

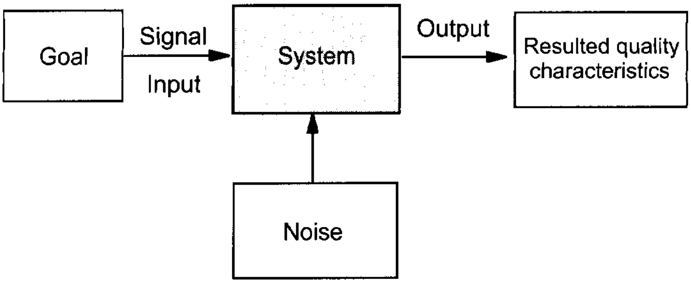


Figure 6.2. Functional relation in communication equipment

called signal-to-noise ratio:

$$SN = \frac{power\ of\ signal}{power\ of\ noise} = \frac{\mu^2}{\sigma^2} \tag{6.12}$$

The larger is the value of SN, the more desirable the system is.

Coming back to our buffer planning problem, we identify as noise the performance deviation, and as goal the nominal performance. The real goal to achieve is not to be as close as possible to the nominal performance itself but also to keep its value as high as possible. In other terms we want to be as close as possible to the absolute maximum achievable. The choice of our control parameters influence both nominal performance as well as the nominal value. We use then the following quality characteristic figure:

$$\frac{nominal\ performance}{performance\ deviation} = \frac{v_{NOM}(r_0, c_0, r, c)}{\Delta v(\Delta r_0, \Delta c_0, \Delta r, \Delta c)} \tag{6.13}$$

The larger is the value of this quality characteristic, the more desirable the system is. In practice we will solve a dual problem which uses the inverse of the signal propagation velocity (i.e the delay per unit length) v^{-1}. We minimize $\Delta v / v_{NOM} = v_{NOM}^{-1} * \Delta v \simeq v_{NOM}^{-1} * \Delta v^{-1}$. We solve then a minimization problem which is formally described by:

$$minimize\ :\ v_{NOM}^{-1}(w_b, l).\Delta v^{-1}(w_b, l) \tag{6.14}$$

This is a problem of unconstrained optimization which can be solved by taking the gradient of the above equation and letting it be zero. We can obtain then:

$$v_{NOM}^{-1}\nabla(\Delta v^{-1}) + \Delta v^{-1}\nabla(v_{NOM}^{-1}) = 0 \tag{6.15}$$

or equivalently:

$$\frac{\partial v_{NOM}^{-1}}{\partial w_b}\Delta v^{-1} + \frac{\partial \Delta v^{-1}}{\partial w_b} v_{NOM}^{-1} = 0 \tag{6.16}$$

$$\frac{\partial v_{NOM}^{-1}}{\partial l}\Delta v^{-1} + \frac{\partial \Delta v^{-1}}{\partial l} v_{NOM}^{-1} = 0 \tag{6.17}$$

Multiply $\partial v^{-1}/\partial l$ and $\partial v^{-1}/\partial w_b$ respectively the above two equations and subtracting them, we will arrive to:

$$\frac{\partial \Delta v^{-1}}{\partial w_b}\frac{\partial v^{-1}}{\partial l} - \frac{\partial \Delta v^{-1}}{\partial l}\frac{\partial v^{-1}}{\partial w_b} = 0 \tag{6.18}$$

We will denote the solutions of this minimization problem with w_b^c and l^c. The (6.18) can be rewritten explicitly in terms of w_b and l as follows:

$$c_3 w_b^2 + c_2 l^2 + c_1 (l.w_b)^2 + c_4 = 0 \tag{6.19}$$

$c_1 = arcb_1 - brc_0 a_1$	$c_2 = br_0 ca_1 - arcb_2$
$c_3 = brc_0 a_2 - \tau_0 b_1$	$c_4 = \tau_0 b_2 - br_0 ca_2$

where a_1, a_2, b_1, b_2 area already defined above in this section.

The equation (6.19) represents the trajectory in the space (w_b, l) of minimum values for the optimization objective $\min v^{-1}\Delta v^{-1}$. This trajectory will be strongly dependent on the particular sample. Solving (6.19) with respect to w_b will give:

$$w_b^2 = \frac{-c_4 - c_2 l^2}{c_3 + c_1 l^2} \tag{6.20}$$

This value of the buffer size, that now is a function on the wire segment length, can be used to find the minimum of the optimization objective of (6.14). Even if we cannot provide a close form expression for the optimum buffer planning for this objective we have demonstrated analytically that exists a trajectory in the 3D space $(l, w_b, v^{-1}.\Delta v^{-1})$ which is suboptimal and a point on this trajectory will give the absolute optimum for this objective. This point is found by searching iteratively in the set of all the possible segment lengths. Each segment length will have a correspondent unique buffer size according to (6.20). For each couple $(l, w_b(l))$ we evaluate the equation $v^{-1}.\Delta v^{-1}$ and the buffer planning which minimize $v^{-1}.\Delta v^{-1}$ will be denoted by (l^c, w_b^c).

It is possible to demonstrate that equation (6.19) represents an implicit relation between buffer size and distance not only for the problem as in (6.14), but also for other two constrained optimization problems.

The first one can be expressed by is searching for a buffering which minimizes the inverse of performance for a target performance variation.

$$\begin{aligned} minimize &: v_{NOM}^{-1}(w_b, l) \\ subject \;\; to &: \Delta v^{-1} = \Delta v_{target}^{-1} \\ & \quad w_b > 0, l > 0 \end{aligned} \tag{6.21}$$

where Δv_{target}^{-1} is the maximum deviation form the nominal performance value allowed. The problem (6.21) the constraint $\Delta v^{-1} = \Delta v_{target}^{-1}$ is equivalent to solve $\Delta v^{-1} \leq \Delta v_{target}^{-1}$ if $\Delta v^{-1} > 0$. The demonstration can be found in [41].

The second constraint optimization problem minimizes the impact of process variation by having a maximum tolerable performance loss denoted by $v_{NOM-target}^{-1}$

$$\begin{aligned} minimize &: \Delta v^{-1}(w_b, l) \\ subject \;\; to &: v_{NOM}^{-1} = v_{NOM-target}^{-1} \\ & \quad w_b > 0, l > 0 \end{aligned} \tag{6.22}$$

It possible to demonstrate [41] that the constrained $v^{-1} = v_{target}^{-1}$ is equivalent to solve the same problem with $v^{-1} \leq v_{NOM-target}^{-1}$.

We show that solving the problem (6.21) leads to the (6.19). A similar demonstration holds also for the problem (6.22) and we will skip it in this thesis.

Using the Lagrange multiplier method, the problem (6.21) becomes:

$$\nabla(v_{NOM}^{-1}) + \lambda \nabla(\Delta v^{-1} - \Delta v_{target}^{-1}) = 0 \tag{6.23}$$

or

$$\begin{aligned} \frac{\partial v_{NOM}^{-1}}{\partial w_b} + \lambda \frac{\partial \Delta v^{-1}}{\partial w_b} &= 0 \\ \frac{\partial v_{NOM}^{-1}}{\partial l} + \lambda \frac{\partial \Delta v^{-1}}{\partial l} &= 0 \end{aligned} \tag{6.24}$$

where λ is the Lagrange multiplier. Replacing λ, by solving the coupling equation we obtain the (6.18) which is equivalent to (6.19). We can observe that solution found is independent on the vale of the constraint. This let us conclude that nominal performance optimization or performance variation optimization are particular cases of the problems (6.21) and (6.22) respectively.

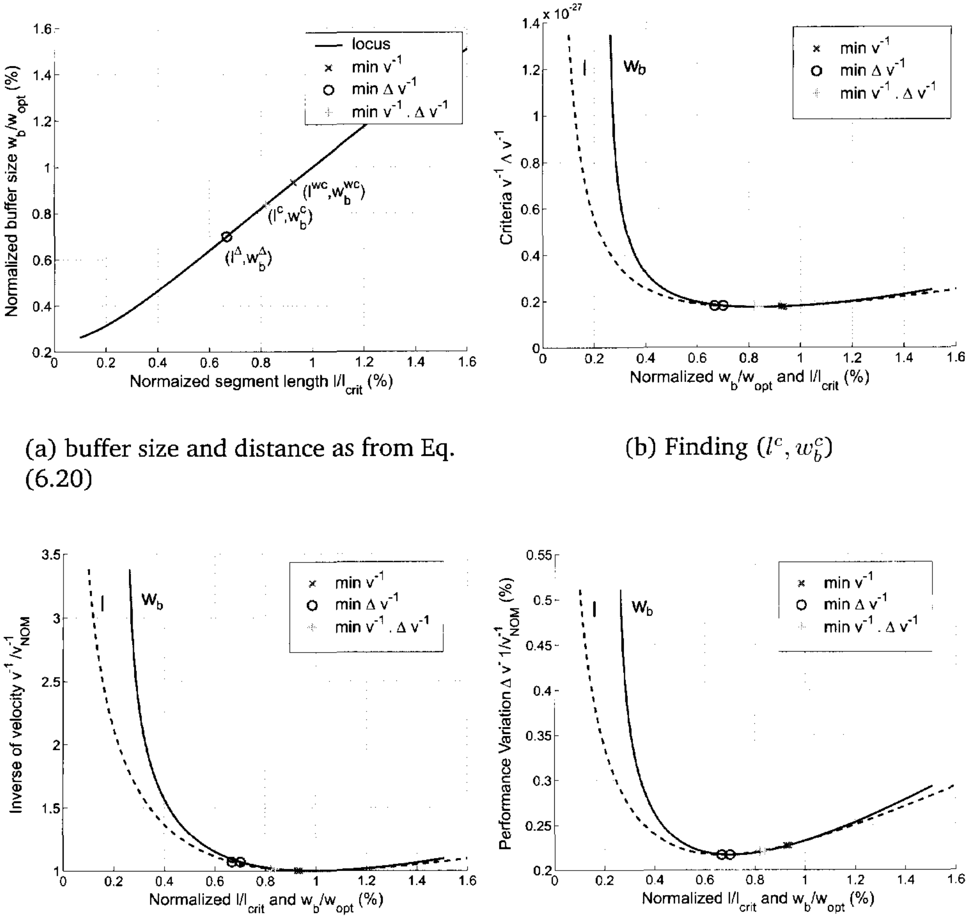
We have shown analytically that all the optimization problems till now studied between nominal performance and performance variation can be solved by a buffering which is on the curve described by (6.19). In order to understand better how the designers can decide about the buffering in presence of process variation, we study an example of a single performance sample. We assume that the deviation form the nominal value are assigned $(\Delta r_0/r_0 = 8\%,\ \Delta c_0/c_0 = 6\%,\ \Delta r/r = 25\%,\ \Delta c/c = 7\%)$.

Figure 6.3(a) describes the trajectory (6.19) of the suitable couples w_b and l and from figure 6.3(b) the value l^c which corresponds to the absolute minimum of the cost 6.14 and the correspondent value of w_b^c, normalized with respect to optimal unconstrained buffer insertion without any process variation, is derived. Analogously, the value of performance, in figure 6.3(c), and the value of its variation, in figure 6.3(d), are estimated.

Each mark reported in each of those figures represents a different optimization objective and we can compare directly the choice that a certain buffering has on the performance and on its variation. Figure 6.3(c) and figure 6.3(d) show that the performance variation optimization differs from nominal performance optimization and in particular buffering for one of those objective penalize the other.

It is interesting to note that in all the three optimization objectives proposed, the optimal values are on the the trajectory described by (6.19) with the consequence that we have somehow restrict our buffering space to a single pattern. It means that we can find an analytical buffering which privileges singularly or deterministic performance or performance variation but we know with this pattern (6.19) all the sub-optimal solutions in between.

The figure 6.4 illustrates the consequence that the choice of a certain buffer strategy has on the performance and on its deviation. Each point of the curve corresponds to a specific buffer size and length value. The different marks in the figures indicate the specific buffer planning optimal for the specific optimization objective. We can notice that for a buffer planning which maximizes the performance, the performance variation will be negatively affected by this buffering choice. On the contrary, by selecting a buffer planning which reduces exclusively the impact of process variation



(a) buffer size and distance as from Eq. (6.20)

(b) Finding (l^c, w_b^c)

(c) Normalized Performance

(d) Normalized Performance Variation

Figure 6.3. $\Delta r_0 = 8\%, \Delta c_0 = 6\%, \Delta r = 25\%, \Delta c = 7\%$

on the performance (and consequently the performance variation), we will have considerable performance loss.

Summarizing, *a buffer planning which enhances the variability tolerance can be obtained at the cost of reducing the nominal performance.* This remains valid whatever variability sample we consider. For a different sample, the value of the maximum nominal performance remains the same (because it does't depends on the process variation) but the performance loss which follows to the reduction of variability depends considerably on the sample selected.

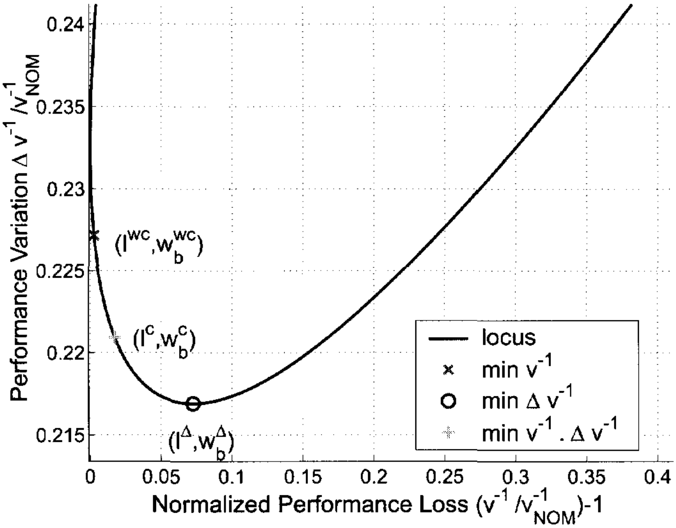


Figure 6.4. performance loss vs variation

The optimization exclusively for performance or for its variability constitute two extreme scenarios, but in general the designer has to know how to select a buffer planning which is an intermediate solution. The curve plotted in figure 6.4 indeed describes the optimal trade-off between performance loss and performance variability. The point which represents the buffering for $\min v^{-1}\Delta v^{-1}$ constitutes an example for trading performance vs variability.

6.3 The process variation space

The analytical approach to buffer planning for a fixed variational scenario presented till now, can be extended to a statistical formulation. Suppose now that different samples, corresponding to different variational scenarios, are originated form random process variables whose statistical proprieties are known. Without loosing in generality we assume that those process parameters have Gaussian distributions centered at the their nominal value.

For each experiment, we consider different variability random samples, where each sample is independent form the others while the distribution of each of the process variation parameters remains the same. Each sample of the process variability can be mapped on a buffer size and distance which is optimal with respect to the selected optimization objective. The mapping of those samples onto the (l, w_b) space, will give a population of possible buffer planning candidates. Each population obtained is relative to an particular optimization objective.

In figure 6.5 is given an example of normalized distribution for buffer size and distance for different optimization objectives. We will assume, similarly to the example in the previous chapter, that each of the process parameters has a standard deviation $\sigma_n = 0.2$. We consider 1000 samples for each experiment and for each sample optimal buffer planning, which is relative to a certain objective, is computed, see section 6.1. Such buffer sizes and distances are collected in a frequency diagram, plotted in figure 6.5. The values presented are normalized with respect the unconstrained buffer planning without any process variations.

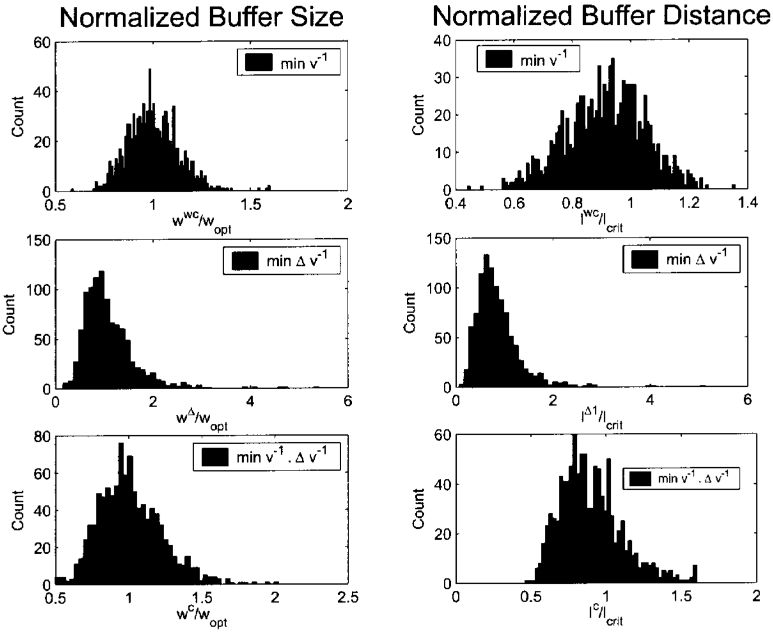


Figure 6.5. Buffer size and segmentation distribution for different optimization criteria and for a single experiment

We note form figure 6.5 that the distributions obtained for buffer size and segment result to be unimodal (single peaked) and for the objective $\min v^{-1}$ it is almost symmetrical while for the criteria $\min \Delta v^{-1}$ it becomes more asymmetrical.

The selection of the buffering under statistical variations will be made on the base of the most commonly occurring value or in other terms we will buffer by using the value of buffer length and size that occur with the greatest frequency. An empirical model which gives the values of the mode for distributions that have moderate departure from symmetry is:

$$Mo = \bar{X} - 3(\bar{X} - Me) \tag{6.25}$$

where $\bar{X}$ is arithmetic mean and Me is the median (it is the 50% in the cumulative distribution) [51].

Each experiment of 1000 samples from the process variation space corresponds then to an unique most probable buffering. Figure 6.6 summarizes the resulted most probable buffer sizes and distances for 13 experiments.

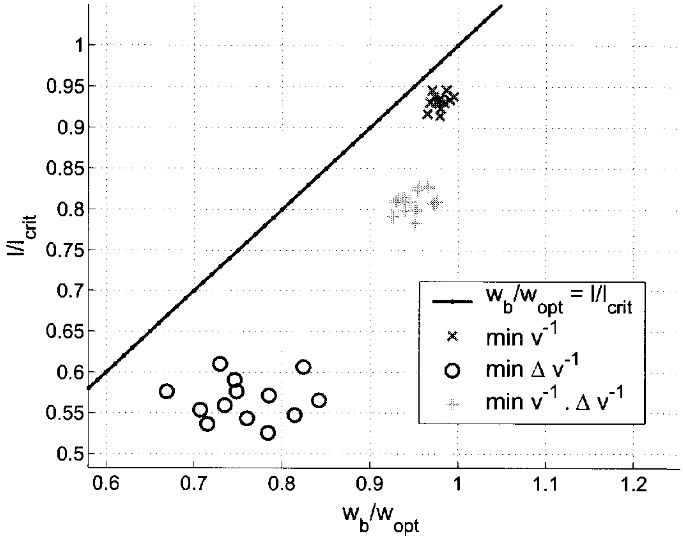


Figure 6.6. Best possible buffer planning for different optimization objectives and for different experiments

The different marks refer to different optimization objectives. We note in this example that, independently form the optimization objective selected, the best buffering privileges more a reduction of the buffer distance over a reduction of buffer size. Indeed the the buffering results always to be $\frac{l}{l_{crit}} < \frac{w_b}{w_{opt}}$. Note also that the buffering for $min\ v^{-1}$ is quite concentrated around a unique point, while we have much more dispersion of values for

$min\ \Delta v^{-1}$. This has to do with the fact that process variations influence considerably performance variation and weaker its nominal value. Again, the objective $min\ v^{-1}.\Delta v^{-1}$ results in an trade-off solution between performance and its variation. An unique value for the buffering relative to a specific criteria can be found by using a minimum distance technique.

We have then proposed a methodology to find unique buffering for a certain optimization objective under process variation. Moreover, by comparing different objectives we have demonstrated that is possible to enhance variability tolerance by using opportunely the buffering but this will cost in terms of nominal performance.

6.4 Yield-centric buffering

We have used until now an iterative optimization procedure which is based on averaging the buffering values obtained for each sample. A unique buffer size and distance "robust" against process variation is then the result of this procedure.

In this section we use the analytic method proposed in chapter 5 to estimate the impact of process variations on the performance for buffered wires. Indeed, by using nominal value and a (co)variance of the process parameters and the appropriate design and technology parameters requested by the performance metric, the statistical performance behavior is approximated by a performance mean and performance variance. Our analytic method shows a good agreement with Monte Carlo based approaches as shown in chapter 5. Moreover it allows a much tighter design optimization loop and provides a better insight in the factors involved (sensitivity analysis is straightforward).

The variability of v is not Gaussian, even if the the variability of the input parameters would be (but which is also not Gaussian in practice). This is because of the nonlinear dependence of v on the input. However, in the sequel we will nevertheless treat v as a Gaussian quantity. It is mainly justified from practical considerations, as no more information than mean and standard deviation of the distribution of v is usually known. In doing so, we can analytically predict trends and understand behavior. Otherwise, numerical statistical techniques can be used [63], [30].

We use this statistical characterization of the performance to describe the performance parametric yield and to derive appropriate buffering for minimizing it.

We define the *parametric yield* as the number of buffered wires of which the performance exceeds a certain threshold. Because of increased generality of the results, we will consider again performance and buffer area being normalized to the case of the absolute maximum value of the performance achievable in a certain technology when unconstrained buffer insertion is performed [7, 49]. Also, the performance threshold is defined using this normalization. Thus, a performance threshold of $x\%$ means that percentage of the absolute maximum performance for normalized buffer area equal to 100%. By actually performing one unconstrained optimal sizing for a certain technology, these normalized values can be translated into absolute sizes and performances.

As the traditional defect process yield is based on the statistical analysis of defect density distribution on the die, analogously we can determine our performance yield using the statistical characteristics of the performance distribution. Our parametric yield will be calculated as

$$Y = \int_{v \geq v_{min}} pdf(v)dv \tag{6.26}$$

where v_{min} is the lower performance limit of acceptability. All the wires above this limit will then meet the specifications. As stated above, we assume that the distribution density function of our performance *pdf(v)* is Gaussian, characterized by the deterministic value of performance as mean (this assumes the bias, see (5.2), is zero) and by the standard deviation calculated according to (5.4). The computation of the yield then involves the mean, the standard deviation of the performance and the design goal v_{min}, and can easily be performed analytically [38].

Thus, we can evaluate the parametric yield in the case of the buffer insertion for the optimal area-constrained buffer insertion as summarized in section 3.3 and [25]. Our analysis is based on the technology data with variability from section 5.3 and 5.4.

In figure 6.7, the solid line presents the optimal trade-off between performance and buffer area from section 3.3. The normalized effective buffer area and the normalized performance are on the horizontal and the vertical axis, respectively. The point $(1, 1)$ corresponds to the absolute maximum performance for a given technology under our deterministic delay model. Further increasing the buffer area/or decreasing the segment length would give a lower performance. The dashed curves present the parametric yield Y (vertical axis) as a function of the normalized buffer size for a specific minimum performance as design goal, identified by v_{min} in (6.26). This design goal is translated into a percentage of the maximum achievable performance for a the particular technology, and is annotated with the curves.

It runs in steps of 10% from 100% to 30%.

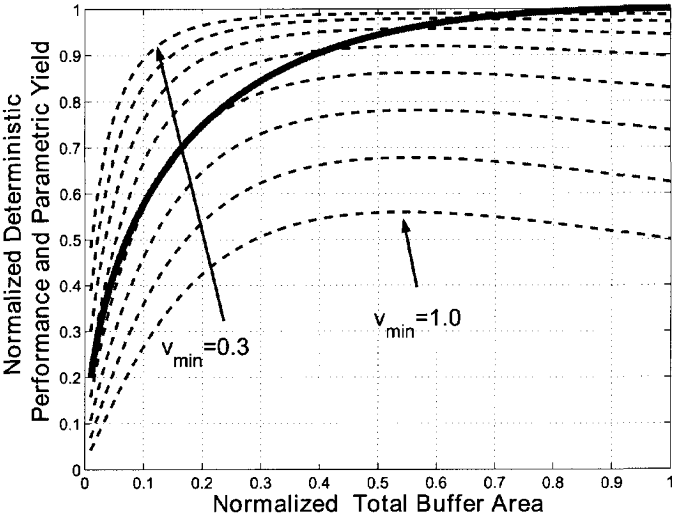


Figure 6.7. Best deterministic normalized performance (solid line) and parametric yield with design goal v_{min} as a parameter (dashed lines) as function of the total buffer area

It is intuitive to understand that a higher-performance design goal v_{min} corresponds to a lower yield. Thus, the high performance curves are running below the low performance curves.

One way to read the figure 6.7 in detail is by first considering a certain minimum performance v_{min}, say 50%. This corresponds to the 0.5 point on the vertical axis. Then, find the corresponding normalized deterministic area by the intersection with the solid curve, which is actually equal to 0.07 on the horizontal axis. Finally, find for this buffer area and design goal the corresponding parametric yield by interpolating the 50% yield curve. In this case the result would be a yield of 0.67. Thus, deterministically optimal buffer planning could lead to a 67% parametric yield, assuming the process and variability data from section 5.3 and 5.4.

Figure 6.7 also allows us to make the following observations:

- For high performance design targets, the optimal yield is achieved with smaller buffer area than suggested by deterministic buffer planning.

- Low performance design targets can be realized with high yield, but using a buffer area that is considerable larger than suggested by deterministic buffer planning.

- Small buffer area leads to low yield, more strongly for high performance design targets but also for low performance targets.

Let us compare now two extreme optimization scenarios. The first one finds the smallest possible area A_{min} for different values of v_{min}. This is equivalent to the deterministic optimization already illustrated in chapter 3. The correspondent value of the yield is found by interpolating, in the figure 6.7, the value of A_{min} with the curve at yield constant which has minimum allowed performance v_{min}. We can summarize then the minimum buffer area for a given performance as:

$$\textit{Min Buffer Area Opt.} \qquad v_{min} \longrightarrow A_{min} \longrightarrow Y(A_{min}) \qquad (6.27)$$

Note that the last optimization is equivalent to find the max performance for a given area as demonstrated in chapter 3.

Suppose now that we want to optimize for maximum yield. From figure 6.7, for a given value of v_{min} we fond on the correspondent curve at yield constant the maximum value of the yield Y_{max}. This maximum yield will correspond to a total buffer area value $A(Y_{max})$, on the x-axis in the figure. In summary:

$$\textit{Max Yield Opt.} \qquad v_{min} \longrightarrow Y_{max} \longrightarrow A(Y_{max}) \qquad (6.28)$$

Note that this maximum yield found refers for the specific buffer insertion proposed for the deterministic case with $w_b = \beta w_{opt}$ and $l = \alpha l_{crit}$ with $\gamma = \frac{\beta}{\alpha}$. However, this does not show that by using a different segment length versus buffer size trade-off, better value of yield are achievable. We will discuss later in this section how good is the assumption to limit ourself to this specific buffering.

We apply the two optimization problem just mentioned in (6.27) and (6.28) to different values of v_{mim} and we collect the results in figure 6.8. In solid lines we have the values of area and yield optimal for the minimum buffer area usage (6.27). In dashed we have the values of area and yield optimal for the maximum yield achievable(6.28). In both cases, the area is denoted by the "o" markers and the yield by "□".

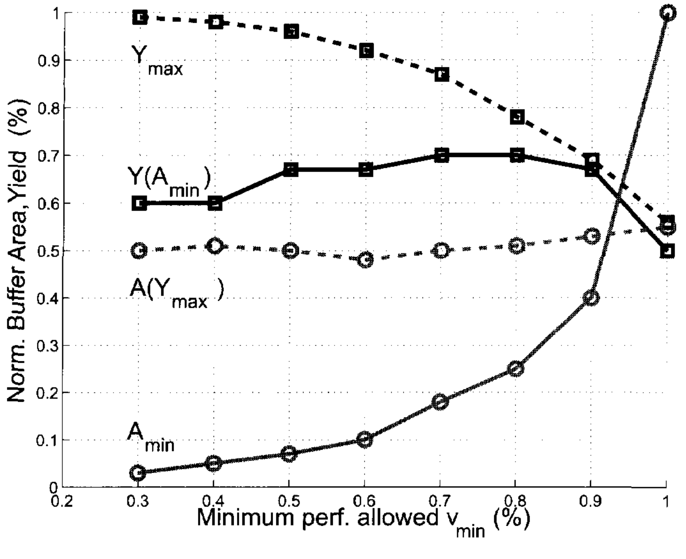


Figure 6.8. Optimal Area vs Optimal Yield

The choice to go for minimum buffer area comes together with a degradation of the yield which remains pretty constant (between 60% and 70%) for different design specifications. And while this yield is acceptable for very high design specifications, for low design specifications this is not acceptable anymore. Thus, the question naturally arises if the method for optimal deterministic buffer planning yields the optimal results when considering variability. We would really want to optimize the yield for a certain performance under area constraints or performance under yield and area constraints, or similar. It is not true that these results are obtained using the same trade-off between buffer size and segment length as in the deterministic case. This is clear from figure 6.9, which displays the yield as a function of the segment length l and the buffer size w_b for $v_{min} = 80\%$. In particular, figure 6.9 suggests that small buffers are more strongly decreasing yield than long segments.

A complete overview of the absolute maximum yield as function of the performance target and its relative area is shown in solid line in figure 6.10. Those values in the figures are compared to the suboptimal solution obtained by searching the maximum yield in the limited area-constrained deterministic buffering $w_b = \beta w_{opt}$ and $l = \alpha l_{crit}$ with $\gamma = \frac{\beta}{\alpha}$ and presented dashed in figure.

The gain obtained by searching the absolute maximum yield in the

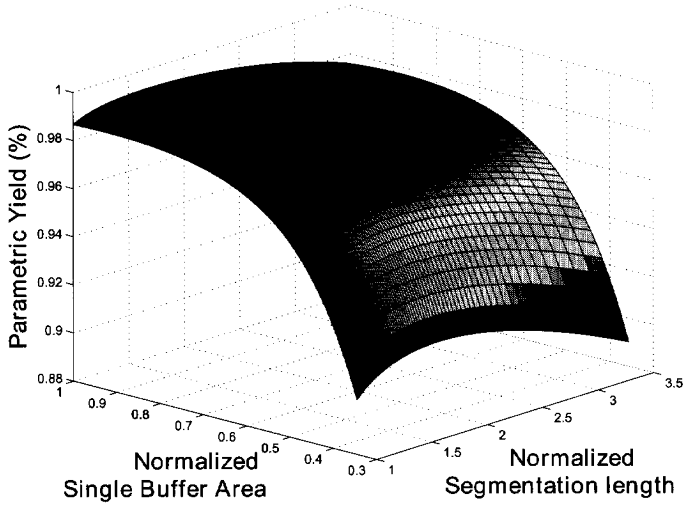


Figure 6.9. Parametric yield in the normalized space (w_b, l) for $v_{min} = 0.8$

whole buffer size and distance space instead of limiting to the $w_b = \beta w_{opt}$ and $l = \alpha l_{crit}$ with $\gamma = \frac{\beta}{\alpha}$, determines only a limited yield improvement (at most 5% for high design performance specification to decrease to almost zero for low specifications) at the cost however of almost 50% of area indifferently for high and low design specifications. Thus, our suboptimal solution should be preferred then in situation when buffer area is constrained.

The actual buffer size and distance value for absolute max performance is in figure 6.11. For high design specifications the maximum yield requires a buffer planning which gives also the absolute maximum deterministic performance. If the target performance is more relaxed, the buffering for maximum yield requires less and bigger buffers. This trend can be explained by the fact that in our example devices variations dominates interconnect variations. Then a reduction of the number of buffers and a simultaneously increasing of their sizes are used to compensate for the variability (and consequently it recovers yield).

If we use the maximum yield solution for deterministic buffer planning the total normalized buffer area will be 50% with a buffer size $w_b = 0.72w_{opt}$ and $l = 1.4l_{crit}$.

Choosing to go for maximum yield is paid in terms of area, which re-

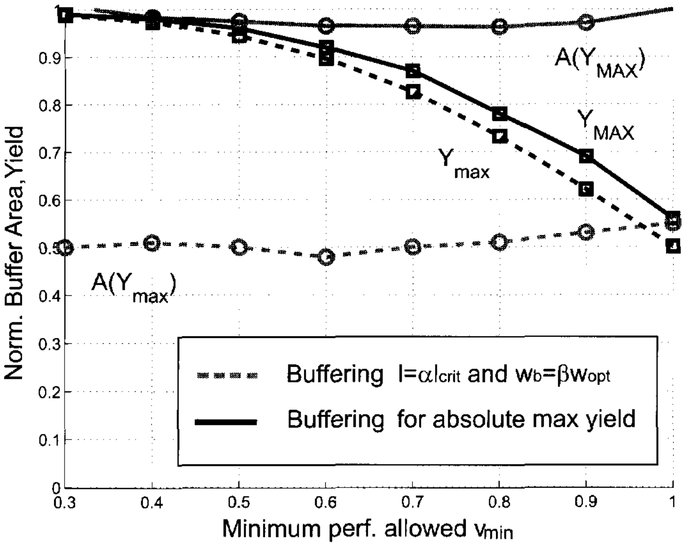


Figure 6.10. Comparison of the absolute optimal yield to the suboptimal yield obtained using the area constrained deterministic buffer planning

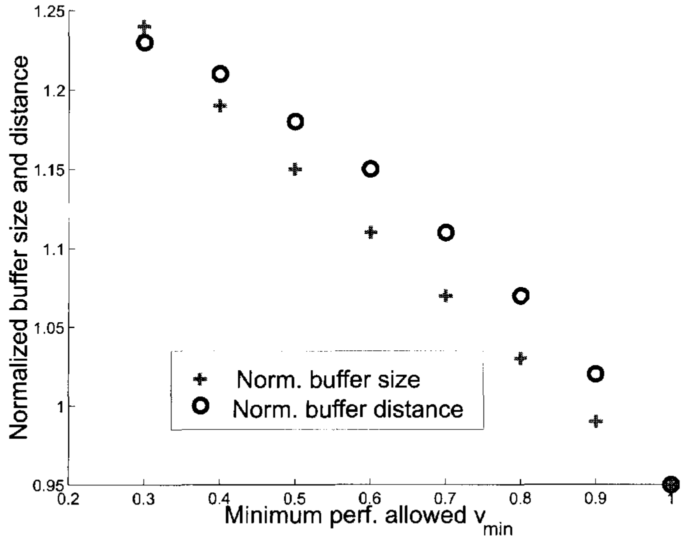


Figure 6.11. Buffer planning for absolute maximum yield

mains reasonably constant (i.e. 50% of the optimal unconstrained area) for every design specifications. This cost is higher for low design specifications.

Until now we have assumed to use a the specific buffer insertion, the one proposed in chapter 3 for the deterministic case with $w_b = \beta w_{opt}$ and $l = \alpha l_{crit}$ with $\gamma = \frac{\beta}{\alpha}$.

6.5 Conclusions

We have have presented a new approach for buffer insertion, which moves away from the traditional objective of delay minimization. Indeed, the process variations influence the delay prediction and consequently it makes uncertain the selection of the sizes and distances between buffers. In this chapter we propose a methodology to reduce this uncertainly. In practice we propose a robust buffer planning against process variations.

A very straight technique to buffer in presence of process variations is to collect all the solutions which minimizes the nominal delay and the solutions which minimize its variability. Among all the solutions found we select the most probable buffer planning. This means that we will have a buffering which will be optimal or close to it for most of the situations. We have presented also an heuristic which reduces the performance variations without degrading too much the nominal performance. The whole formulation remains analytical even if we were not able to achieve a close form expression for minimization of this heuristic.

We have introduces also the concept of buffering to enhance the parametric yield. We have demonstrated that exists a trade-off between yield and buffer area. Maximum yield costs a lot in terms of area when compared to the same performance without any process variations and this cost becomes higher for low design specifications (i.e. low min performance allowed). Moreover the relation between buffer size and distance found in chapter 3 can still be used with reasonable approximation also in the yield maximization problems. It gives values of the yield close the optimum without consuming too much buffer area. We can still improve yield but a better selection of the buffering, but we have shown that this is going relatively a lot in area.

Chapter 7

Application: Network on chip

Contents

IN order to cope with the increasing complexity of the systems and to boost design productivity it is important to be smart in using, reusing and/or adding new design parts. A wide part of emerging microelectronics research is dedicated to this goal. During the 1990s, the research aimed at integration of different large components and/or processing cores on a single die. The resulting systems became known under the name *System on Chip (SoC)*. As silicon technology has advanced, several problems have emerged, especially in the field of the communication part of the SoCs. Indeed a communication infrastructure has to be shared among a growing number different components and this makes performance of the communication very unpredictable. In addition of that there is the unpredictability introduced by using long wires in a deep submicron regimes. Then, the problem of finding suitable communication structures has to be addressed at all levels from physical to architectural to the operating systems and application level [9]. For the branch of research which focus simultaneously on integration of different design parts and communications structures the term *Network on Chip (NoC)* [29], [34] is used.

A common way to implement such NoC architecture is by using a regular mesh of switches and resources [37]. Resources can be processor cores, memories, custom hardware blocks or any generic IP. Switches are used to route and buffer messages and data between resources.

Switches and resources are connected by input and output channels. A channel consists of an unidirectional point-to point bus. Such NoC architectures have been proposed as solution in order to improve the scalability and the modularity such that the complexity and functional diversity can be handled systematically.

It seems obvious that this approach introduces challenging problems. The first problem is to identify the real nature of applications that are possible to implement optimally or near optimally on an NoC architecture. We are not attacking this problem because it has to be solved at a higher system level, while the work presented here remains more at the physical level. However, an important physical level problem is timing closure. Indeed, even small changes in adding gates to the netlist can result in a change of timing of the complete system with a consequent unexpected change of placement and routing. In this context, a systematic approach to predict and to select (by tuning the dimensions) the appropriate communication system, becomes essential. Then, the NoC platform consists not only of architecture but also of design methodology. The NoC design methodology has two tasks:

- map the architecture template to concrete physical hardware implementation. In practice decide about the number of resources, switches and network on the base of their shape. It is the architectural level that is integrated in the physical level.
- map the application onto this concrete architecture.

The design methodology which integrates architectural level and physical level can be accomplished only if appropriate performance (delay, cost, power etc.) can be estimated.

The throughput can be used in NoC as a figure of performance for unidirectional data streams. This choice can not be completely true when different type of data with variable data rate are travelling on that network. For example a mixture of control real time commands and high throughput video streams.

The network on chip architecture

We focus on the communication infrastructure between resources by considering the latter as stand-alone blocks. The main characteristics of this network are

- Hardware of each of the blocks can be developed then independently.
- The network build between blocks can be some how scalable and adaptable to different workload

It is out of the intend of this thesis to propose protocols to transmit data between resources and switches neither to propose solutions on how a packet of information can be passed through the network form an arbitrary sender to an arbitrary receiver. We focus on the physical layer in which the number and the length of the wires connecting resources and switches is determined.

For the type of on-chip buses that we consider, the wire parasitics are important. Coupling is becoming critical because the side wall capacitance is increasing with the wire aspect ratio. This has an effect on the signal propagation properties, and has to be accounted for during design. Thus, buffering can be required to increase the performance of the wires and to decrease the adverse affects of coupling.

In this chapter we will focus on unidirectional bus design connecting the system blocks. We model our bus by using a channel of uniform parallel wires that has fixed width W_{ch} and fixed length L as presented in figure 7.1. We propose a wire sizing approach combined with a repeater planning for the channel. We assume that wire width and spacing are uniform for every wire and also that the buffer insertion is uniform.

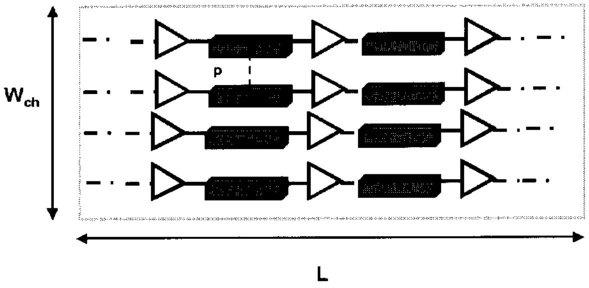


Figure 7.1. Bus Model.

We will now first discuss some preliminaries in section 7.1 and in particular some qualitative trends using a simplified capacitance model. Then, in section 7.2 we discuss the buffer and wire sizing solution, first for the unbuffered and the unconstrained case in section 7.3 and subsequently for the constrained case in section 7.4. We then conclude in section 7.6.

7.1 Preliminaries

Optimizing buses in fact entails selecting the optimal spacing and width of the wires, but also the proper size and distance of the buffers. For the buffering, we consider in the sequel of this chapter three different policies.

- *No buffering.* This is normally not a good policy, because of the quadratic dependence of the delay on the distance. However, we include it in our analysis because it helps in understanding the problem more fully.
- *Unconstrained buffering.* This policy refers to a dimensioning of the buffers (distance and size) for maximum performance. Given wire resistance and capacitance, Bakoglu [7], [25] has derived buffer size and distance that minimize the latency for a single line. However, the Bakoglu model only considers single lines.
- *Constrained buffering.* This policy refers to buffer size and distance such that the performance is maximal under total buffer area constraints. An analytic model for the area-latency trade-off for single lines was presented in [25].

Sometimes, we also use the adjective "optimal" in relation to buffering, but in fact it is redundant. For both constrained and unconstrained buffering we only consider the optimal case. When we would apply buffering methodologies introduced above for regular pipelined bus structures as the one described in figure 7.1, a measure of performance can not be exclusively the single wire latency. Instead, it has to be influenced also by the wire density. Indeed, design strategies (such as clocked and pipelined interconnect) might be able to tolerate latency, and the design objective could become throughput. Here, we define throughput as the amount of information that can be sent per unit time and per unit bus width (in microns, not in number of wires). In particular, it might be better to use many wires with small pitch than to use few wires with a large pitch. Wide wires have lower latency but more narrow wires fit in the same space. As a result, the throughput can be higher.

A throughput-centric design strategy has already been introduced in [54], but wire sizing is derived only for optimal unconstrained buffer insertion. In [36], a first attempt to maximize throughput under controlled buffer area usage is presented, but this work has the limitation that buffer area reduction is obtained by resizing only the buffer size without changing their distance and number. This is a suboptimal buffering scheme.

In [50], the throughput is maximized under constrained buffering, by using iterative algorithms based on closed form expressions that are more accurate than the Elmore delay. Our approach also optimizes throughput under constrained buffering. We use an analytic model as presented in [25] to arrive at optimal constrained buffer distance and size for a certain wire geometry, while [50] uses a global numeric optimization approach. It is in fact not clear how [50] finds the optimal number of buffers along a line, while in our approach this number unambiguously follows analytically. In addition, the increased insight gained by our analytic approach helps in understanding the exact nature of the bus optimization problem.

We will now first introduce our design resources. In particular we will present the relations between throughput and buffer area. Consider a bus of wires introduced above and in figure 7.1. L and W_{ch} denote the total physical length and width of the channel, respectively. Furthermore, N_W denotes the number of wires, which have width w and spacing s where the pitch p is given by $p = w + s$. Thus, we have $N_W = W_{ch}/(w + s)$. The wires are uniformly buffered, with l the distance between buffers and N_b the number of buffers. For simplicity we take $N_b = L/l$ (in reality the number of buffers is one plus the number of segments). Furthermore, we define τ as the delay of a single buffered segment and $v = l/\tau$ as the signal propagation velocity. Now we can express the throughput as

$$T = \frac{1}{N_b \cdot \tau} \times N_w = \frac{l}{L \cdot \tau} \times \frac{W_{ch}}{w+s} = \frac{v}{w+s} \times \frac{W_{ch}}{L} . \tag{7.1}$$

While shielding can improve the delay predictability, it usually does not improve the throughput as defined above. Although the signal line density is reduced, the reduced effective (Miller) capacitance does not sufficiently improve the performance of a single line. Thus, other techniques to reduce coupling effects, like signal alignment of adjacent wires are preferred to shielding. However, if necessary, our models can be extended to include shielding. For simplicity we will not do so in this chapter.

Motivated by the fact that T is proportional to W_{ch} and inversely proportional to L (if v is constant), we will in the rest of this chapter use the normalized or square throughput, $T_{\square}$, as the throughput for a square channel:

$$T_{\square} = \frac{v}{w+s} \tag{7.2}$$

By using the same bus structure and the same notation as introduced in defining the throughput, we can now define the total buffer area required for buffering in the channel, where w_b is the size of a single buffer:

$$A_{ch} = N_W N_b w_b = \frac{LW_{ch}}{l \cdot (w+s)} w_b \tag{7.3}$$

Note that w_b is actually only proportional to the buffer area. In this chapter we will ignore the constant multiplicative constant that relates the transistor width to this buffer area. This constant is only layout and technology dependent and does not change our model.

We present in the following subsection, by using qualitative trends, how to select the wire density depending on the resources just introduced. We will illustrate the cases of no buffering and of unconstrained buffering. This will help in understanding that the density has to be selected considering a trade-off between throughput and buffer area.

Simplified qualitative trends In this subsection exclusively, we assume use to use parallel plate capacitance model. Here this simplistic formulation gives clear qualitative insight in the factors involved in selecting the appropriate wire density. We postpone the quantitative results to the next section, using more accurate capacitance models. For simplicity, we will assume here that $w = s = p/2$ but in the rest of this chapter this assumption will be relaxed.

We use a simple first order parallel plate model for capacitance c per unit wire length, that includes the coupling to the layers above and below the wire as well as to the neighboring wires, as follows:

$$c = \epsilon w/t_{ox} + 2\epsilon SFh/s \tag{7.4}$$

Here, ϵ is the permittivity of the medium, w, s, h and t_{ox} are wire width, spacing, wire thickness and interlayer spacing, respectively and SF is the so-called switch factor accounting for the neighbor line activity [32].

Resistance per unit wire length will be given by

$$r = \rho/wh. \tag{7.5}$$

For long lines without buffers, the delay is of course proportional to rcL^2. Thus, the signal propagation velocity is proportional to $v = 1/rcL$. Using this and $w + s = p$, the normalized throughput (7.2) can be written as $T_\Box \propto 1/rcp$. Thus, maximizing throughput is now equivalent to minimizing rcp. Using (7.4) and (7.5) with $w = s = p/2$, this is equivalent to

$$minimize \quad f_1(p) = \tfrac{p}{ht_{ox}} + 8\tfrac{SF}{p} \,. \tag{7.6}$$

Then, an analytical value of the wire pitch which maximizes throughput is obtained by setting the first derivative with respect to p equal to zero. Thus, the wire pitch that maximizes the throughput for unbuffered buses is proportional to $p = 2\sqrt{2SFht_{ox}}$.

However, buffering can improve the single wire latency and then, implicitly, also the throughput. In this situation, we try to maximize the throughput for unconstrained total buffer area. It is possible to show [7, 25] that the reciprocal of the signal propagation velocity v^{-1} is then proportional to $\sqrt{rc}$. Thus, (7.2), can be rewritten as $T_\Box \propto 1/p\sqrt{rc}$. Note the square root when comparing to the equivalent expression above for the unbuffered case. Thus, maximizing the throughput is now equivalent to minimizing $p\sqrt{rc}$. By combining this with (7.4) and (7.5) and again using $w = s = p/2$, we find that maximizing throughput becomes equivalent to

$$minimize \quad f_2(p) = \sqrt{\tfrac{p^2}{ht_{ox}} + 8SF} \,. \tag{7.7}$$

Thus, in case of unconstrained buffering, the pitch should be chosen as small as possible.

In the first case, for unbuffered buses, the optimal pitch, actually balances between high density and small capacitance situations. In the second case, for unconstrained buffering, the optimal pitch is the smallest possible. Indeed, the buffers can compensate for the capacitive loading associated with the high density.

However, the increased throughput of unconstrained buffering is not always necessary and/or the cost (in terms of silicon area and power) of such buffering is too high. Then, it is natural to search for the best trade-off: either the minimum buffer area for a certain throughput or the maximum throughput achievable with a certain buffer area.

It is clear that throughput and buffer area can be traded by increasing the wire spacing. Informally, for constant latency of a line the increased spacing allows less buffering because the capacitance of a line is reduced. With constant latency but increased spacing, the density and hence the throughput is reduced. This reasoning, however, does not give a clue towards calculation of optimal trade-off, which might even be less clear when also the width of the line is allowed to change. However, in the next section we will investigate its solution.

7.2 Throughput driven buffer insertion

Until now we have considered a parallel plate capacitance model for the wire. However, it is known that fringing components of the capacitances become important for deep-submicron technologies where wires have a high aspect ratio. Therefore, we will use in the rest of this chapter the analytical model from [10], which can accurately model the capacitance including the fringing terms.

Given the switching factor SF, we can find for each technology and each channel (total width and length) in the maximum density configuration the Bakoglu solution [7] for the buffer area. We will denote this value as A^*_{ch}. It is the total unconstrained buffer area required by the bus for minimum latency in the maximum density wire configuration for that switching factor. For the specific case of $SF = 1$, we will denote the corresponding normalized throughput as $T^*_{\Box}$. If we exclude cases of $SF < 1$, based on the theoretical analysis of the previous section, this throughput is actually the maximum possible, which will indeed be confirmed below for the $0.18\mu m$ technology example.

A_{ch} will denote our design constraint for the buffer area available in the channel. The maximum useful buffer area is equal to A^*_{ch}, since this was determined for the most dense buses with the greatest capacitive loading, given the switching factor. The optimal buffer area for buses with a lower density is allays less then A^*_{ch}. Then, the cases of no buffering, unconstrained buffering and constrained buffering correspond to $A_{ch} = 0$, $A_{ch} = A^*_{ch}$ and $0 < A_{ch} < A^*_{ch}$, respectively.

We will use A^*_{ch} and $T^*_{\Box}$ as normalization values. In particular, we define the normalized total buffer area constraint as $\gamma_{ch} = A_{ch}/A^*_{ch}$ with $0 < \gamma_{ch} < 1$, and we use $T^*_{\Box}$ to normalize the throughput graphs below so that they can be compared.

Now, our optimization procedure works by sampling the space of allowable (w, s), and for each sample we will optimize the latency using the procedure from [25] as constrained by A_{ch}. For some cases with a low density, the A_{ch} actually exceeds the optimally needed buffer area. In that case, the procedure from [25] simplifies to the Bakoglu sizing [7]. After the sampling, we can just select the design point with the greatest throughput.

This sampling method is efficient, since the whole formulation is analytic. In most cases, it would not be necessary to apply a more efficient search method for the optimum but otherwise it could be implemented if that would be desirable.

Since our intent is only to explain a methodology to trade different resources in bus design, we will limit ourselves in the examples below to a particular technology. That is, in our examples we will use parameters typical for a $0.18\mu m$ technology. For the value of those parameters we refer to [25]. However, our methodology is completely generic and can also be applied to other processes.

For our sampling procedure we need to specify the minimum width and spacing, which are denoted by w_{min} and s_{min}, respectively. In this chapter, we take them representative for the top most metal layers in the $0.18\mu m$ example technology and assume a value of $0.6\mu m$.

Furthermore, since the delay of unbuffered lines is quadratic in the length and that of buffered lines is linear in the length, we need to specify the length of the bus when we intend to compare different scenarios. This is true even while we use normalized values of throughput, see (7.2). For our examples, we assume that the length $L = 2cm$. Also, we take $W_{ch} = 60\mu m$, allowing 50 minimum pitch wires.

7.3 Unbuffered and unconstrained bus design

Before we actually discuss constrained buffering, we will for comparison first consider the two limiting cases for bus design, namely that of no buffering ($A_{ch} = 0$) and of unconstrained buffering ($A_{ch} = A^*_{ch}$). The former case is trivial, and for the normalized throughput it follows that

$$T_\square \propto \frac{1}{Lrc(w+s)} . \tag{7.8}$$

Note that since the latency of unbuffered lines is quadratic in L, the normalized throughput $T_\square$ as in (7.2) actually depends on L.

For unconstrained buffering, we can apply the Bakoglu procedure. Thus, for each sample from our (w, s) space, we get values of l_{crit} and w_{opt}, being the optimal buffer distance and size, respectively. The resulting buffer area per line will be denoted by A^U, and is given by

$$A^U = \frac{L}{l_{crit}} w_{opt} . \tag{7.9}$$

since L/l_{crit} is the number of segments which approximates the number of buffers along a line if L is sufficiently large. Please note the difference between A^U as defined above, which is actually a function of w and s, and A^*_{ch}, which actually is the value of ion of w and s, and A^*_{ch}, which actually is the value of A^U when $w = w_{min}$ and $s = s_{min}$.

For the normalized throughput, $T_{\square}$, it then follows that

$$T_{\square} \propto \frac{1}{\sqrt{rc}(w+s)} \,. \tag{7.10}$$

Figures 7.2 and 7.3 illustrate the throughput of a bus for different wire densities for the two extreme cases considered in this subsection. Note that in these and the following throughput graphs, we actually show the ratio of $T_{\square}$ to $T^*_{\square}$. Thus, they can directly be compared to evaluate the benefit of buffering. (But this benefit of course depends on the particular value of L considered.)

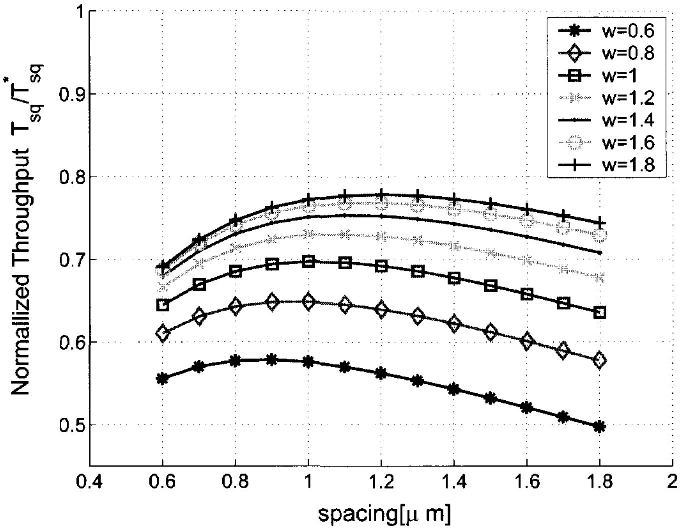


Figure 7.2. Throughput without buffering.

The results, as presented in figures 7.2 and 7.3, seem to confirm the trends illustrated in the previous section. If no buffering is considered, the wire density which maximizes the throughput has an intermediate value, balancing between wire latency and density. If we allow unconstrained buffering, the maximum wire density solution clearly is the optimal solution. We will show in the next subsection that the choice of the appropriate

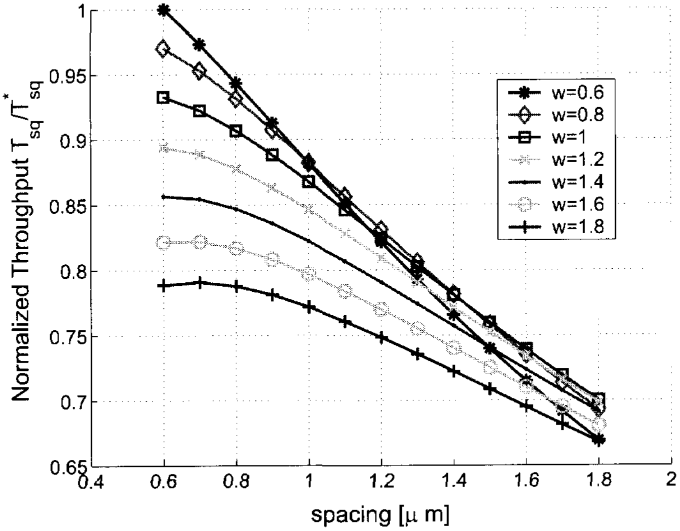


Figure 7.3. Throughput with optimal unconstrained buffer insertion.

wire density for the highest throughput will depend on the specific buffer area constraint.

7.4 Constrained bus design

Here we consider the case of $A_{ch} < A^*_{ch}$. For each sample in our (w, s) space, we will optimize the latency using the procedure from [25] as constrained by A_{ch}.

However, the procedure of [25] effectively works on single lines, with the interaction to other lines only modeled via the coupling capacitance. Therefore, we need to translate the overall A_{ch} into a specification per line. Moreover, the optimization procedure takes as input the buffer area constraint for a single line, say A, normalized to A^U. That is, the buffer area constraint is specified as

$$\gamma = \frac{A}{A^U}. \tag{7.11}$$

For bus design, we can relate this to A_{ch} by using $A = A_{ch}/N_W$, since each line will be buffered identically. Then,

$$\gamma = \frac{A_{ch}}{N_W A^U} \tag{7.12}$$

and because of (7.9) we get

$$\gamma = \frac{A_{ch} l_{crit}}{N_W L w_{opt}} \ . \tag{7.13}$$

Now, we can use the Bakoglu procedure to find l_{crit} and w_{opt} so that we can determine γ. Given γ and the relevant technology parameters, including wire resistance and capacitance, [25] calculates the buffer distance l and the size w_b for lowest latency under the area constraint specified by γ. We will denote this corresponding latency of a single segment by $\tau(\gamma)$. Then, the throughput (7.2) in the presence of a buffer area constraint γ can be written as follows:

$$T_{\Box} = \frac{l}{\tau(\gamma)} \frac{1}{w+s} \tag{7.14}$$

In the examples below, we will normalize l and w to the optimal unconstrained buffer distance l_{crit} and buffer size w_{opt} as $\alpha = l/l_{crit}(w,s)$ and $\beta = w_b/w_{opt}(w,s)$.

We show an example for $\gamma_{ch} = 0.3$. It means that the total area available for buffering in the channel is 30% of A^*_{ch}. We derive the normalized buffer area for a single wire, γ, from (7.13) and we evaluate the throughput using (7.14) in the space (w,s). That is, we sample this space and for each sample we find the optimal buffering. This is fast, because the whole formulation is analytic. The result is in figure 7.4.

As shown in figure 7.4, the maximum throughput possible for a given area constraint is achieved for a wire topology which differs from the minimum one. Higher densities require an aggressive buffering to achieve low enough latency, which is not possible due to the area constraint. Also lower density solutions can not achieve a low enough latency to compensate for the reduced parallelism. Note that the highest throughput for 30% of the maximum area is only 13% below the absolute maximum throughput. It is obtained by using $w = 0.8\mu m$ and $s = 0.8\mu m$, together with $w_b = 35\mu m$ and $l = 4.2mm$.

The optimal normalized buffer size β (in dashed line) and buffer distance α (in solid line) are presented as function of the wire topology in figure 7.5. The values on the y-axis equal to 1 represents the area unconstrained solution [7]. For large spacing (larger than $1.6\mu m$ in this example),

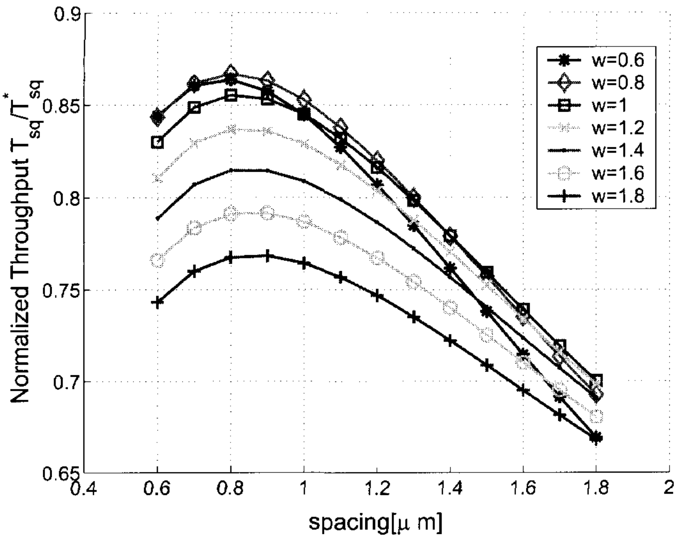


Figure 7.4. Wire sizing for optimal area-constrained buffer insertion $A_{ch}(w, s) = 0.3A^*_{ch}$.

the area constrained buffer insertion solution is more and more similar to the area unconstrained solution. This is because only few and relatively small buffers are necessary or in other terms the area constraints for those topologies become less stringent.

Figures 7.6 and 7.7 show contours of constant throughput as a function of the wire width and spacing for the case of $A_{ch} = A^*_{ch}$ (unconstrained buffering) and for the case of $A_{ch} = 0.3A^*_{ch}$ (constrained buffering), respectively. Note that for each w, s configuration shown, specific w_b and l have been computed that maximize the throughput for the given area constraint. Note that both graphs actually have a single optimal point from the perspective of throughput, as marked by the black square. However, these graphs illustrate what happens to the throughput if suboptimal w and s are chosen.

If we compare for example the curve at $0.85T^*_{\square}$ of figure 7.6 and 7.7, we notice that the same throughput requires higher density in the constrained case. This is to be expected considering the fact that higher density compensates, in the throughput budget, for the increased latency of a single wire. The increased latency of a single wire is a result of the limited available buffer area.

While the methodology in this chapter as developed thus far optimizes

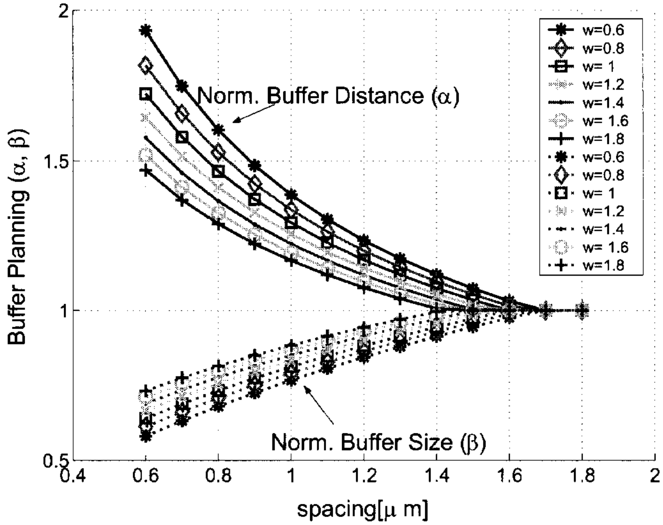


Figure 7.5. Normalized buffer size and distance for optimal area-constrained buffer insertion $A_{ch} = 0.3A^*_{ch}$.

throughput, these graphs actually also offer a perspective on latency. In fact, the latency is what is actually optimized for a single wire. Now, as the wire spacing is increased, the coupling capacitance decreases and a better latency can be achieved for the same buffer area.

Thus, given a certain throughput contour, the same throughput is achieved with the lowest possible latency for the configuration with the lowest density. This configuration corresponds to the point where the constant contour curve is tangential to the lines with constant density. In the graph, these constant density are the diagonal −45° lines, and the loci of optimal latency for a certain throughput are along the lines labeled ”optimal latency”. While not explicitly presented in this chapter, the actual latency in each design point can easily be obtained from the model. Thus, the model also offers a solution for latency constrained designs.

In figure 7.8, we plot the maximum throughput that is achievable using our model as a function of the normalized buffer area constraint γ_{ch}. The figure shows two sets of points, for different switching factors. Note that the achievable throughput strongly depends on the switching factor since it relates to the capacitive coupling effect. Also note that with $SF = 2$, the buffer area maximum throughput is actually 80% larger than for the

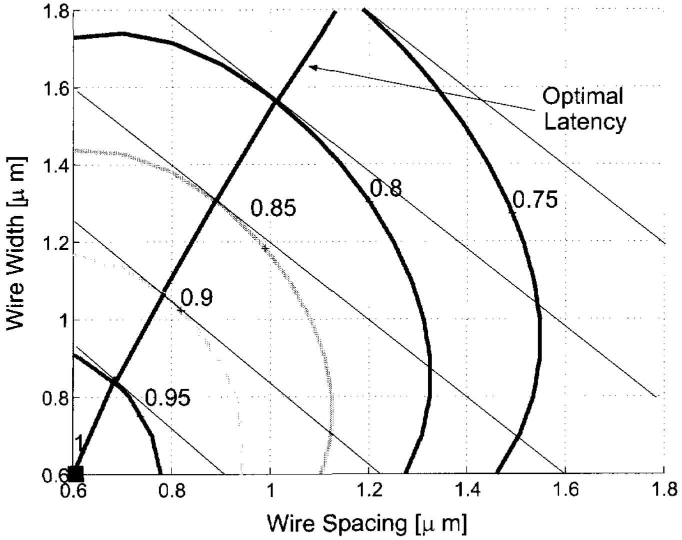


Figure 7.6. Constant throughput contours for unconstrained buffer area (A^*_{ch}).

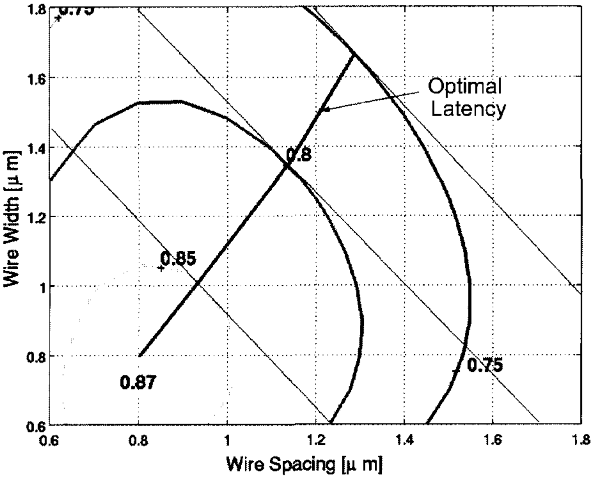


Figure 7.7. Constant throughput contours for constrained buffer area ($A_{ch} = 0.3A^*_{ch}$).

$SF = 1$ case. However, for either switching factor, the throughput is only degrading relatively little if the buffer area becomes small. For example, for 80% reduction of the buffer area compared to the buffer area for maximum throughput, the performance is only reduced by less than 25% for both extreme vales of the switching factor. Thus, the effectiveness of proper wire and repeater sizing is clearly demonstrated.

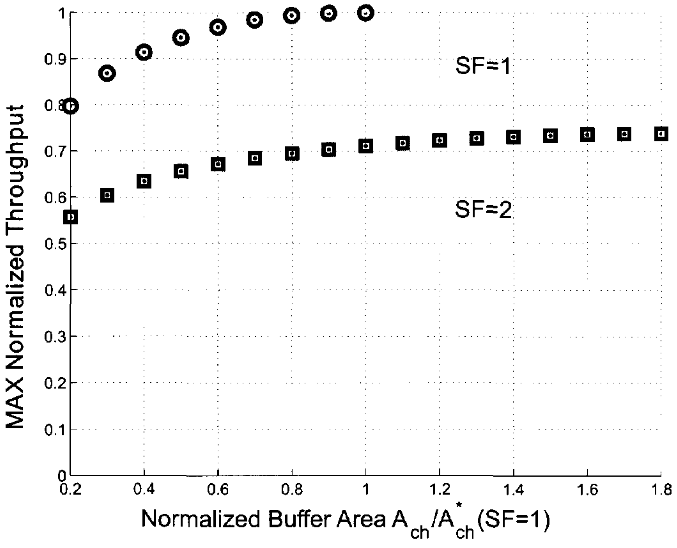


Figure 7.8. Maximum throughput as a function of the buffer area constraint γ_{ch}.

Finally, figure 7.9 is a companion figure to figure 7.8. It again presents 2 sets of points for the two extreme switching factors. On the y-axis, the wire spacing and width are given that are found to be the optimal ones for the specific buffer area and can realize the throughput in igure 7.8.

Thus, if such graphs are created once for a technology, possibly using unnormalized axis, they can be used for design. First, a suitable buffer area is selected that achieves the required throughput, by reading it off from the x-axis in figure 7.8. Subsequently, this value can be used to get the wire geometry from figure 7.9. Finally, the right buffering is determined from [25] or from corresponding graphs that could be prepared but are presented in figure 7.10-a and figure 7.10-b.

The buffer distance and size presented in those figures refers to $0.18\mu m$ technology. As expected a more aggressive buffering with larger and nar-

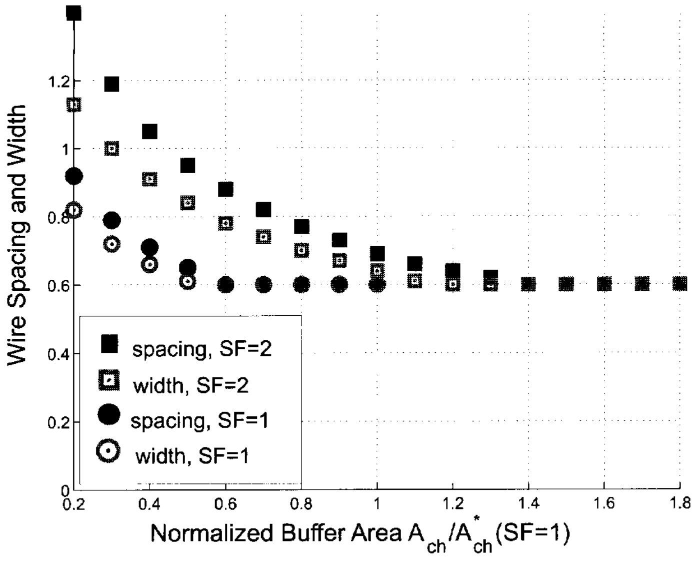
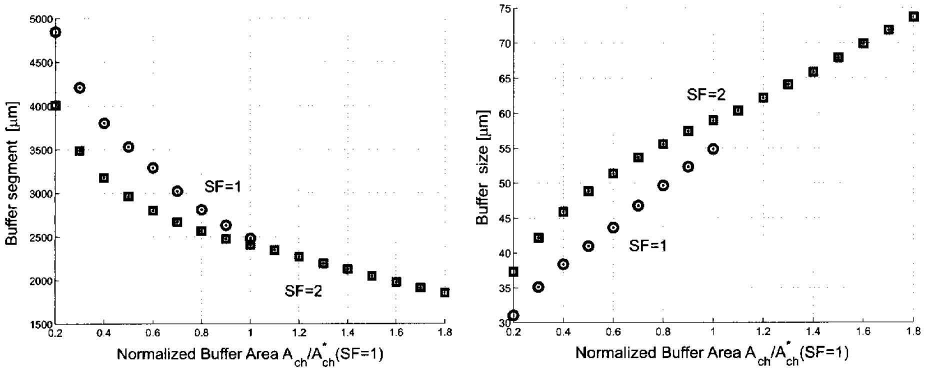


Figure 7.9. Relative wire density as function of the buffer area constraint of the channel.



(a) Buffer distance

(b) Buffer size

Figure 7.10. Buffering for maximum throughput as function of the buffer area constraint of the channel and for different switching factors

row buffers is required for $SF = 2$ when compared to the situation where $SF = 1$. Again buffering with normalized channel area larger than 1 for $SF = 1$ means that we are overdriving too much the segments and then larger area does not contribute anymore to increase the throughput. This explains the fact that there is no reason to investigate for $SF = 1$ the buffering $A_{ch} > A^*_{ch}(SF = 1)$.

7.5 Throughput under process variations

In the previous section we have presented a methodology to obtain the maximum throughput in presence of limited total buffer area. The results show that when large buffers are not allowed because of limited Si area, we have to increase the distance between the wires or, in other terms, decrease the wire density. However, if the density becomes to low, this will cause a degradation of the channel throughput. Thus, wire density can be traded with available buffer area to obtain maximum throughput.

In this section, we want to extend the study of buffer planning and wire density to the situation where process variations are considered. Essentially the questions we tackle are: what happen to the throughput if the process variations are taken into account? What will be then the trade-off between wire density and total Si area for buffering, which makes the throughput robust against process variations?

In order to answer to those questions we model the signal velocity propagation in a similar fashion to what we have already illustrated in [26] and more in detail in chapter 5. Our performance metric is the throughput of a channel of parallel wires uniformly buffered as already been presented in section 7.1.

We are interested in particularly on how the interconnect width and spacing affect the throughput in presence of process variations.

In first instance, we assume that the wire width and spacing variations are small enough such that is still possible to consider the wire pitch ($p = w + s$) constant. As a consequence of this assumption, the wire spacing will be inversely correlated to the wire width. In practice the number of the wires in the channel will remain the same. From (7.2) the throughput variation is just the signal propagation velocity variations as in [26] times the constant wire pitch.

Different is the situation if we assume that the width and spacing are varying independently from each other. In this case process variations can modify the wire density. For sake of simplicity we will limit our study to a fix constant spacing while the wire width will be varying randomly according to a Gaussian distribution with standard deviation 0.2. We simplify our problem assuming also that this standard deviation remains the same for different nominal value of the wire width. It is possible to extend our results also to the case where where random spacing variations are present. However, this extension results to be straightforward and can be done in a similar fashion to the modelling of only the wire width variation.

We model the effect of the process variations by using the second order Taylor's expansion of throughput and we evaluate the statistical proprieties of the throughput in terms of the expected value and of the throughput variability. This represents a simple extension to the throughput of the methodology already illustrated in [26] for the signal propagation velocity. The variability of the inverse throughput $T_{\square}^{-1} = v^{-1}.(w+s)$, derived from (7.2), can be estimated by:

$$\sigma_n(T_{\square}^{-1}) = \frac{\sqrt{\sum_i k_i^2.(w+s)\sigma_i^2 + (k_w(w+s)+v^{-1})^2\sigma_w^2}}{T_{\square}^{*-1}} \tag{7.15}$$

where the k_i represents the sensitivity of the inverse of velocity respect to the parameter i. Explicit relations for k_i have been already presented in 5. Notice that sensitivity w.r.t. the wire width is left explicit in (7.15) because the wire width is the only parameter which affects at the same time wire density and propagation velocity. In the equation (7.15) the throughput is normalized with respect to the maximum deterministic throughput for unconstrained buffer insertion and minimum wire density $T_{\square}^{-1}$.

We consider in our example only interconnect variations, since only the interconnect variations can affect the wire density and we assume that in the nominal situation spacing and width are the same $w = s = p/2$. We notice in figure 7.11 that for a given nominal wire density, really small buffer sizes increase the throughput variability. The throughput variability will be lower for high density and higher for low density and this difference is larger for small buffer sizes. In fact the width variability is defined as ratio of the nominal value, it means that smaller width (ie. higher wire density) have lower variability.

Summarizing, a general a guideline is to increase the wire density for small buffers. We notice also that the difference between high density and low density more pronounced for small buffer sizes. Moreover, for the example shown in figure where only width variability of 20% is considered, the throughput variability has a weak dependency from the wire density when compared to the dependency of the buffer size.

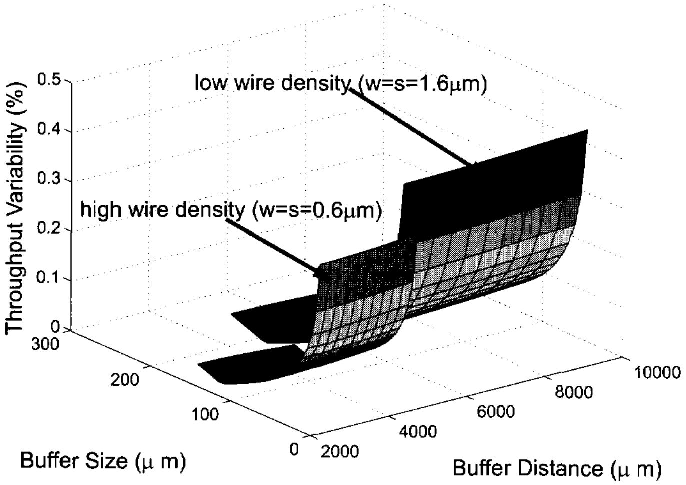


Figure 7.11. Throughput variability under 20% wire width variation.

We can extend the trade-off between buffer area, throughput and wire density to the parametric yield analysis. We fix analogously of what we have done in chapter 6 for the signal propagation velocity, the minimum acceptance value of the throughput. We know that the max throughput is given for the higher wire density and for the unconstrained buffer insertion $T_{\square}^{*}$ and the level of acceptance is defined w.r.t. to this value.

In figure 7.12 the acceptance level is fixed at $0.9T_{\square}^{*}$ and the throughput variation is induced by the wire width variation which has a Gaussian distribution with nominal width and a variance 20% of the nominal value. Three nominal cases are considered: $w = s = 0.6\mu m$ (max wire density), $w = s = 0.8\mu m$ and $w = s = 1\mu m$. The three curves represented in figure 7.12 are obtained each for the nominal width value selected. We decide to plot only those curves for $l = l_{crit}(w, s)$.

The parametric yield is decreasing for small buffer sizes. The value of the yield are expressed in percentage as function of $T_{\square}^{*}$, which explains the fact that for lower density the throughput is lower and its variability is higher. Those two effect contributed to the yield degradation. The figure 7.12 suggests then to use the minimum density configuration for high yield. High parametric yield for throughput requires then high density and consequently requires aggressive buffering, or in other terms Si area. For

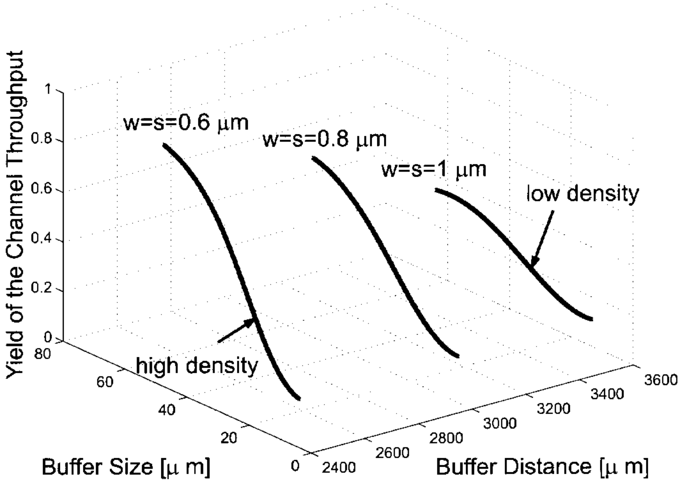


Figure 7.12. Buffer planning for the Parametric Yield of the throughput. We assume the acceptance lower $0.9 * T^*_{\square}$.

lower density the yield will degrade but less and smaller buffers will be necessary. How to trade yield for buffer area?

In figure 7.13 we present the parametric yield curves as function of the available are for buffering in the channel.

Depending on the area constraint the higher parametric yield is obtained by selecting the appropriate wire density. For example, if the area available is smaller than $0.6A^*_{ch}$ then it is necessary to increase the wire density. The increase of the width and spacing from $0.6\mu m$ to $0.8\mu m$ can increase the yield of 10% (from yield 30% to 40%) with a use of only 40% of the area.

7.6 Summary

At the physical layer of Network-on-Chip (NoC) implementations, blocks and switches are already assigned and inter block connections have to be planned. Most of the inter block connections will consist of a large number of long parallel wires. Such wires usually have significant resistance and (coupling as well as self) capacitance.

This chapter illustrates methodologies on how to perform buffering and wire sizing to obtain the maximum throughput. The results show that the

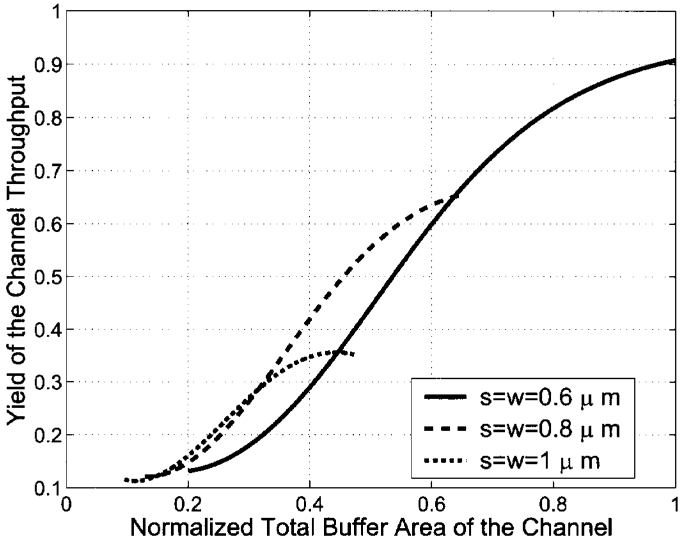


Figure 7.13. Normalized Throughput Parametric Yield as function of the total buffer area available in the channel $W_{ch}.L$. We assume the acceptance lower $0.9 * T^*_\square$.

appropriate wire density can be chosen depending on the total buffer area available in the channel. High wire density is not beneficial if the buffer area is small. Low wire density on the other hand can result in a low latency but not in a high throughput. Our methodology can explore this trade-off and produce optimal buffering and wire sizing for a given area constraint.

The effect of process variations on the throughput are quantified and the parametric yield of the throughput for a minimum acceptance limit shows again that buffering has to be conducted simultaneously with an appropriate wire sizing. Higher yield depends then on the total area available and in particular a small buffer area available requires a lower density. Yield for a given wire density is degrading drastically for small buffers, and this is a direct consequence of the increased throughput variability.

Chapter 8

Tools

Contents

IN this chapter we illustrate two tools that were developed during this project. In particular we present a tool which is able to extract accurate values of the wire capacitance per unit length.

In the second part of this chapter, a short description of a demo illustrating the methodologies of this thesis is given. It is explained its web based structure and how it is implemented.

8.1 CAIGE: CApacitance *from* Interconnect *Geometry* Extractor

It is well known that 3D geometry of interconnects and the intrinsic properties of the materials constituting the conductors and their interwire spacing, determine the electrical behavior of the wires. We try to characterize in this study the resistance and capacitance for unit length for a multilevel interconnect structure.

A preliminary study was done to assess an optimal 3D geometry for interconnect. Using the values of the Strawman's table presented in [58, 47], which contains a baseline prediction of interconnect geometries for future technologies, as nominal, the impact of RC delay of interconnect was evaluated by changing every single geometrical parameter.

In figure 8.1 a cross section of the structure studied is shown. A, B and C are three parallel conductors and above and below them there are two conductor planes connected to GND. The wire thickness is denoted by h, the inter-wire spacing by s, the wire width by w and the thickness of the oxide by tox. The wire pitch is defined as $P = w + s$.

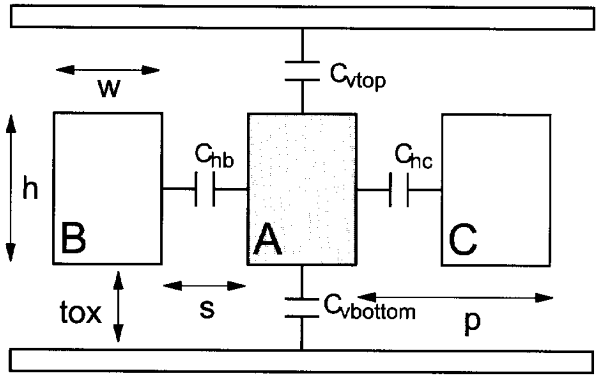


Figure 8.1. Cross-section diagram of parallel lines between two planes

The wire resistance is defined by the well known expression $R = \rho * l/wh$. It is trivial to show that a smaller cross section of the conductor translates in higher capacitance.

Wire capacitances are estimated using a tool called CAIGE (*Capacitance from Interconnect Geometry Extractor*) which derives accurate capacitance values per unit length of the wire.

Once it has been defined a parameter range and nominal values, CAIGE generates the appropriate layouts. The capacitance values are extracted from the layouts using SPACE, a layout-to-circuit extractor developed in our group [61], embedded in this tool. The consistency of the extracted capacitance values is validated with respect to analytical models already available in literature [53, 14, 65, 7] and valid in the same range of geometrical parameters. Motivations to use CAIGE with respect to analytical models will be explained in the next section.

CAIGE's flow Different close expressions for capacitances have been proposed [53, 14, 65, 7]. They are well regarded for their simplicity, speed and accuracy, but they are rigidly constrained to simulate only few physical configurations, and often they are not extendible to other structures.

Figure 8.1 represents a typical configuration used to derive most of the analytical capacitance formulas available in literature. Moreover, the accuracy of these formulas is limited to a range of wire thickness, wire width, inter-wire spacing, thickness of the oxide and/or their ratios. To explore all possible geometrical configurations, it is necessary to guarantee accurate capacitance values also outside the parameter range fixed for each formula. CAIGE has been developed to extract accurate capacitance values for configurations which are not necessarily similar to that presented in Figure 8.1.The choice to consider structures similar to Figure 8.1 was done essentially for a validation purpose with respect to analytical formula.

CAIGE uses the layout to circuit extractor SPACE [61] as core engine to perform 3D capacitance extractions. CAIGE can give an idea on how much capacitances are sensitive to changing one or more geometrical parameters.

In Figure 8.2, a data-flow illustrates the operations performed by CAIGE. Nominal values, parameter range and number of samples constitutes the space-exploration of different geometrical configurations which we want to investigate. For each configuration, the layout generation, the update of the technology files for SPACE and the extraction are performed iteratively. Once all the space-exploration has been completed, the capacitance values per unit length of the conductor A as in Figure 8.1 are calculated.

Details about CAIGE's features will be illustrated below:

Parser The nominal geometrical configuration used is described using Strawman's technology table [58, 47]. Only configurations with $P = width_{nominal} + spacing_{nominal}$ have been extracted. This translates in considering only configurations in which the density of interconnects for each layer remains constant, in other words it retains the same wire-ability for each layer. Moreover, other technological constraints can be added by the user, for example on the maximum aspect ratio allowed.

Layout generation The layouts are described in LDM language and successively translated into GDS. Figure 8.3 presents a top view of the layout generated. The length of the parallel lines is fixed to $10 * pitch$.

Together with masks used to describe the conductors, symbolic masks have been used in our layout. The symbolic masks are masks which are not really present in the physical layout, but they are adopted to influence the 3D mesh opportunely.

Some technology files, which are characteristic of the process, are used

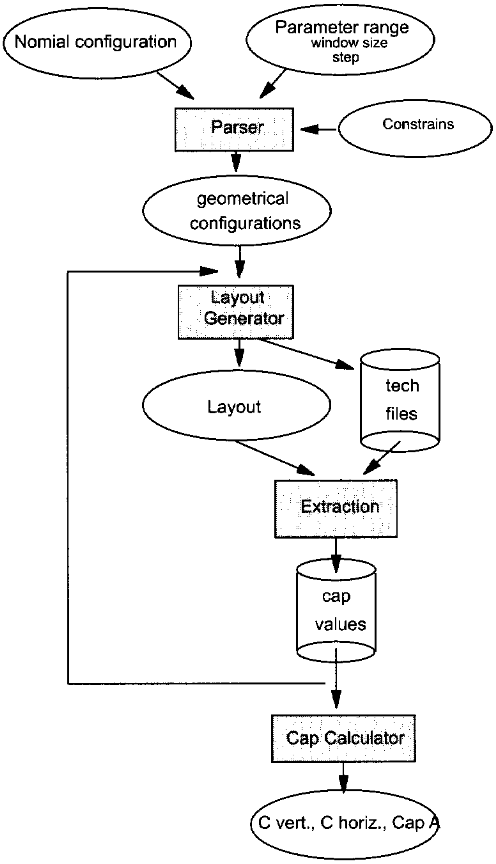


Figure 8.2. Caige's dataflow

by SPACE to perform the extractions. All the information about masks, elements to be recognized, values of parasitics, vertical dimensions, dielectric structure etc, are stored in those files. Since SPACE is used in this framework to extract 3D capacitance values, an update of the technology files is necessary for different thickness of $Si0_2$ or of the wire, or for different dielectric materials.

Extraction SPACE is used as core engine to perform 3D capacitance extractions. The capacitances will include edge effects as shown in Figure 8.4 with dashed lines; here only the capacitances related to conductor A are shown, the other conductors are such that they are connected to *gnd*. We indicate with C_{es} the edge side wall capacitance distributed along the length of the wire and with C_{ec} the contribution to the capacitance due to

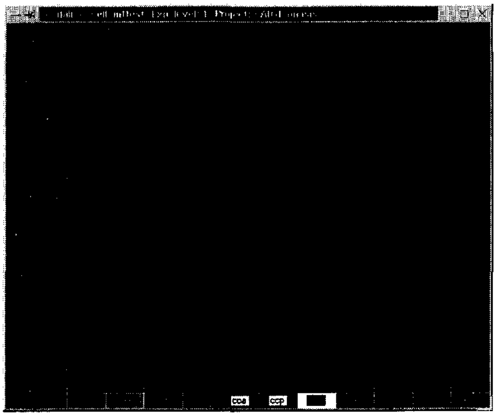


Figure 8.3. CAIGE's test layout

the front and back face of the conductor.

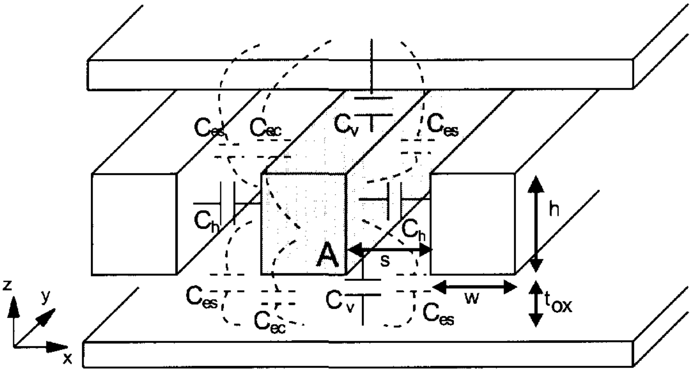


Figure 8.4. Edge effects

For each of the geometrical configurations of the wires, two extractions are necessary. The first extraction is an extraction of the layout generated as it is and the second is instead an extraction of the same layout but "stretched"in the y direction. C_{ec} in both the extractions will be the same. This consideration will be used to eliminate the contribution of the edge capacitances of the front and back face of the conductor from the capacitance values. The stretch of the layout is opportunely done such that the mesh doesn't change.

Cap Calculator The idea is to derive accurate values for the total horizontal (c_h) and vertical (c_v) capacitances per unit length as shown in Figure 8.1. These values have to include C_{es} but not the contribution of the edge capacitance of front and back faces of the Figure 8.4. Since the length of the wire is fixed to $10 * pitch$ and the stretch is 1% of this length, a simple approximation of c_h or c_v can be obtained using:

$$c_h = \frac{(C_{h_1} - C_h)}{(10.1 - 10)pitch} \tag{8.1}$$

and analogously for c_v. The index in the equation indicates values extracted from the stretched layout.

The capacitance c_v and c_h are derived for conductor A in the case in which all the other conductors and planes above and below are connected to gnd. In order to model the influence of the switching activity of conductors B and Con conductor A, we introduce the switching factor SF. The total capacitance of A is given by:

$$c_A = c_h + SF * c_v \tag{8.2}$$

where SF can have values in a range between 1 and 2.

8.2 WARP: *Wire And Repeater Planning*

It is iterative demo tool implemented on web (*http://space.tudelft.nl/warp*) which contains the results of our research on buffer insertion strategies for nanometer scale IC technologies.

Those features are currently available:

- Optimal tradeoff of repeater area and power versus performance of uniformly buffered single global interconnect lines.
- Statistical analysis methodology of the performance in view of variability of basic technological parameters.
- An online tool for plotting of 2D interconnect capacitance data as a function of the geometry and technological parameters. It is an on line version of CAIGE.

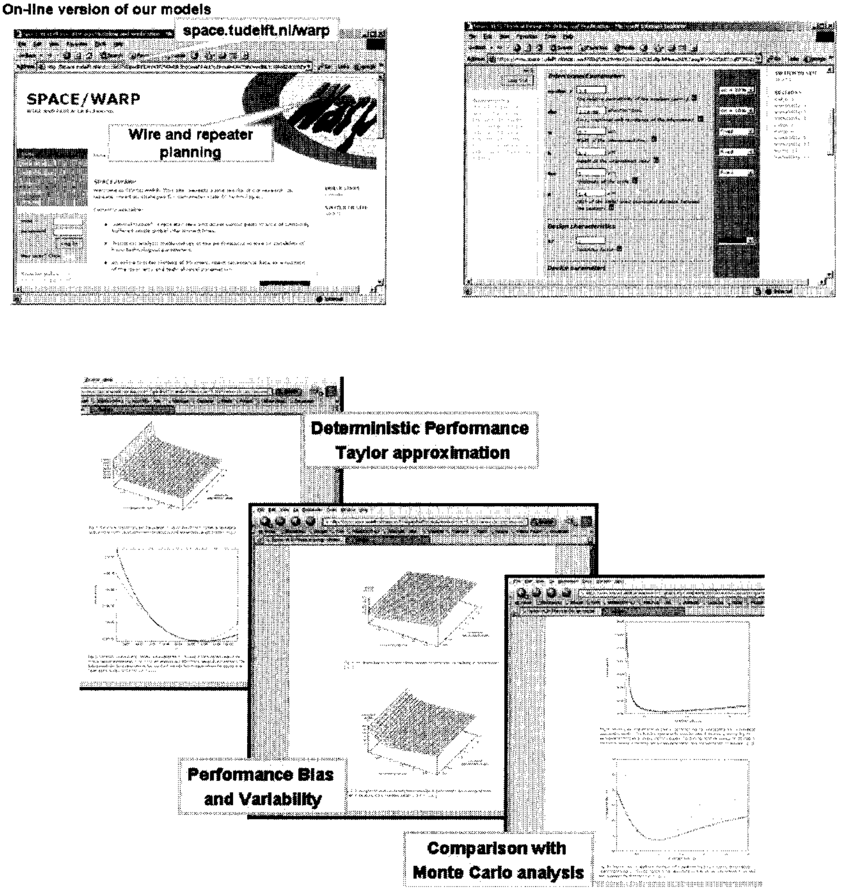

On-line version of our models
space.tudelft.nl/warp
SPACE/WARP
Wire and repeater planning
Deterministic Performance Taylor approximation
Performance Bias and Variability
Comparison with Monte Carlo analysis

Figure 8.5. Three WARP Screenshots

Implementation The structure of the tool developed, is such that the analytical models are solved symbolically by using Mathematica. A syntax converter was opportunely used to automate completely the conversion from the web to the syntax used by mathematica and to generate clean the Tcl code.

The actual value of the parameters are entered via the web. Those data denote as *query data* are mapped onto the parameters of the Tcl scripts generated by Mathematica. This mapping is done by using a Tcl Web Server which is able to interpret the http protocol server into Tcl. We use *Websh* (*http://tcl.apache.org/websh*) to build our web application. This has the function of web server and allows Http protocol to communicate with Tcl.

Each time a user enter the value on the web, a new session is created which actually store the state. Since the web is by nature stateless, if we want to create dynamic web pages we need to simulate somehow a state machine. The updating of the state can be handle by the web server itself but it has the problem that the web server can not be available at the same time for different users.

The alternative, that was preferred, was to implement a process which is responsible to handle the changing of the state by saving the state into file, by running the updating of the state via the Tcl scripts from Mathematica and by storing the new state. The web server then has only a sort of task manager function. In this way we keep the web server as free as possible giving the possibility to different users to run in parallel different sessions. The interface process run idle form the web server.

The web server can be invoked by the user in every moment to check if the computation is finished or is still running or even to stop the execution of the processes.

The data user data are then graphically visualized using Gnuplot and eventually the plots can be also downloaded.

The advantages of such implementation can be summarized in :

- actually the application is completely embedded leaving to the user only a friendly web-interface
- the processes run idle form the web server such tat different users can start sessions in parallel without waiting that the web server is available
- Mathematica during the simulation is not required and has only to be run again if the model is changed such that a new symbolic solution is needed.

Chapter 9

Conclusions

This thesis has investigated different issues in the field of buffer planning for global wires. We have opted for an analytical formulation which results to be very easy to incorporate in a complete design flow. Moreover, this approach is very efficient if implemented in iterative schemes and it can give better insight for designers to select the best trade-off in buffering. The idea is to provide fast and efficient screening tools and resources allocator tools for high level design.

The traditional buffering for maximum signal speed is very expensive in terms of buffer area and power and suboptimal solutions has to be pursued. Moreover, layout constraints together with limited routing resources, which reduce their available insertion locations, open new questions on how to efficiently benefit from placing those buffers along the wires. On top of that, process variability results in performance fluctuations, which makes buffer solutions that are valid in a determinist scenario not necessary valid when variations considered.

This thesis focusses on two main objectives:

- Trade-offs in buffering for deterministic constrained problems. In particular optimal area-power-performance trade-off solution is found for a single wire.
- Statistical modeling of performance in presence of random process variations. This open new possibilities to think to a buffering as a way to enhance robustness against those variations and also to derive a parametric yield-centric design.

Those results are presented for a single wire and are then extended also to the design of unidirectional multiwire structures typical used for NoC applications.

Summary of the results

For high design specifications, performance costs a lot in terms of buffer area under the assumption of uniform buffer insertion.

It is found the best trade-off curve performance versus buffer area and the relative buffer distance and size. Depending on the area constrained value, a combined use of smaller buffer placed and larger distance gives this optimum. An analytical expression for those quantities is derived. The results are presented in a normalized form respect to the unconstrained solution (which has indeed the maximum performance but uses the maximum buffer area). The normalized curve performance versus area is independent from the wire geometry (i.e layer assignment) and its weakly dependent from the technology (if we assume that c_p/c_0 remains almost constant with technology scaling). Thus, we know always how to trade optimally relative performance for relative area. For absolute values, we have only to know optimal unconstrained buffering, area and performance.

Power constrained problem is almost equal to area constrained problem. It is demonstrated that this difference is due to the switching power term. For high frequencies, the contribution of this term on the total power budget is almost negligible. Moreover, optimal unconstrained buffering or near optimal unconstrained buffering contributes to keep the switching power term small. This means that also the buffering for area constrained solves approximately also the power constrained problem. On the contrary for low frequencies, switching power becomes relatively important (if compared to the switching and leakage term) and then smaller and closer buffers (compared to the area-constrained problem) has to be used.

If non uniform buffering is considered, and only few of those buffers are not in their optimal position, the effect on the total wire delay is negligible. It more important to resize buffers opportunely depending on the transmitter and receiver size (driving and loading the unbuffered wire). The buffer area-constrained problem, applied to non uniform buffering, results in a further resizing of each of the buffers. This resizing is needed in the situation in which the area available for buffers is quite small. It is interesting to note how those problems of buffer resizing are solved by using iterative algorithms and the complexity of those problems grow linearly with the number of buffers on the wire.

A generic quality metric, being a non linear combination of the electrical parameter, can be modeled in presence of process variations by using a second-order Taylor approximation. The statistical behavior of this quality metric will be then characterized by its expected value and its standard deviation. This model can account also for parameter correlations when they are explicitly known. Otherwise we refer to uncorrelated primitives and propagate the variability. By using this method the sensitivity analysis is straightforward, the integration on new viability sources is very easy and it doesn't require perform all the simulation again (as in a monte carlo analysis). Performance variability uncertainly seems to become very high for small buffer sizes. If in a deterministic scenario going for small buffer size doesn't affect considerably the performance and for this reason is preferable, in a statistical scenario small buffer sizes introduce high performance uncertainly and then it becomes more difficult to have a reasonable performance prediction.

By introducing the concept of performance variability we can search for buffering methods to enhance performance variability tolerance. We introduce in this framework a buffer planning which performs an optimization for each of the sample in the variability space. We select from the buffer sizes and distances population, the most probable values. Another way to reduce the performance variability is to buffer for maximum parametric yield. A complete trade-off between area-yield-performance is proposed. Maximum yield for high design specifications requires less area than in deterministic case. For low performance specifications, high yield cost in terms of buffer area, on the other hand if we want to save buffer area, it will cost in terms of yield.

All the results proposed for a wire whose width and spacing is known are extended to the buffering for typical unidirectional multiwire structures. Assuming a channel of fix width containing parallel long wires, buffer area now is traded with wire density. The best configuration is the one which maximizes the throughput of the channel. We have shown for unconstrained buffer area that the maximum throughput is reached by using the highest wire density allowed by technology. However, if the buffer area available is limited, then buffers cannot compensate completely for very high coupling and resistance effects, and then is necessary to move the wires more far apart. We have also show in presence of process variations, the maximum parametric yield depends of the buffer area available. In particular for lower buffer area available should be associated also to lower density to maintain the maximum yield.

Since in this thesis we consider only point-to-point connections, a first natural question is on how to extend these methodologies to the buffer planing for tree structures. We believe that our techniques can be incorporated by visiting the tree bottom-up form the sinks to the driver. Each branch of the tree is then buffered in a point to point connection fashion.

However it should be investigate better how to handle split points.

It was not possible to access for us to real manufacturing data and process tolerances. Thus, a validation of our method of modelling variations in more realistic situations should be studied. This can open a further investigation on quality figures to compare process technologies.

Appendix A

Spice technology file

The $0.18\mu m$ spice technology file that we use to derive the device parameters r_0 and c_0 are taken from [10]

```
 .LIB CMOS_MODELS
* Predictive Technology Model Beta Version
* 0.18um NMOS SPICE Parameters (normal one)
*

.model CMOSN NMOS (
+Level = 49

+Lint = 4.e-08 Tox = 4.e-09
+Vth0 = 0.3999 Rdsw = 250

+lmin=1.8e-7 lmax=1.8e-7 wmin=1.8e-7 wmax=1.0e-4 Tref=27.0 version =3.1
+Xj= 6.0000000E-08         Nch= 5.9500000E+17
+lln= 1.0000000            lwn= 1.0000000            wln= 0.00
+wwn= 0.00                 ll= 0.00
+lw= 0.00                  lwl= 0.00                 wint= 0.00
+wl= 0.00                  ww= 0.00                  wwl= 0.00
+Mobmod=  1                binunit= 2                xl=  0
+xw=  0                    binflag=  0
+Dwg= 0.00                 Dwb= 0.00

+K1= 0.5613000               K2= 1.0000000E-02
+K3= 0.00                  Dvt0= 8.0000000           Dvt1= 0.7500000
+Dvt2= 8.0000000E-03       Dvt0w= 0.00               Dvt1w= 0.00
+Dvt2w= 0.00               Nlx= 1.6500000E-07        W0= 0.00
+K3b= 0.00                 Ngate= 5.0000000E+20

+Vsat= 1.3800000E+05       Ua= -7.0000000E-10        Ub= 3.5000000E-18
+Uc= -5.2500000E-11        Prwb= 0.00
+Prwg= 0.00                Wr= 1.0000000             U0= 3.5000000E-02
+A0= 1.1000000             Keta= 4.0000000E-02       A1= 0.00
+A2= 1.0000000             Ags= -1.0000000E-02       B0= 0.00
+B1= 0.00

+Voff= -0.12350000          NFactor= 0.9000000         Cit= 0.00
+Cdsc= 0.00                Cdscb= 0.00               Cdscd= 0.00
+Eta0= 0.2200000           Etab= 0.00                Dsub= 0.8000000
```

```
+Pclm= 5.0000000E-02       Pdiblc1= 1.2000000E-02      Pdiblc2= 7.5000000E-03
+Pdiblcb= -1.3500000E-02   Drout= 1.7999999E-02        Pscbe1= 8.6600000E+08
+Pscbe2= 1.0000000E-20     Pvag= -0.2800000            Delta= 1.0000000E-02
+Alpha0= 0.00              Beta0= 30.0000000

+kt1= -0.3700000           kt2= -4.0000000E-02         At= 5.5000000E+04
+Ute= -1.4800000           Ua1= 9.5829000E-10          Ub1= -3.3473000E-19
+Uc1= 0.00                 Kt1l= 4.0000000E-09         Prt= 0.00

+Cj= 0.00365               Mj= 0.54                    Pb= 0.982
+Cjsw= 7.9E-10             Mjsw= 0.31                  Php= 0.841
+Cta= 0                    Ctp= 0                      Pta= 0
+Ptp= 0                    JS=1.50E-08                 JSW=2.50E-13
+N=1.0                     Xti=3.0                     Cgdo=2.786E-10
+Cgso=2.786E-10            Cgbo=0.0E+00                Capmod= 2
+NQSMOD= 0                 Elm= 5                      Xpart= 1
+Cgsl= 1.6E-10             Cgdl= 1.6E-10               Ckappa= 2.886
+Cf= 1.069e-10             Clc= 0.0000001              Cle= 0.6
+Dlc= 4E-08                Dwc= 0                      Vfbcv= -1 )


*
* Predictive Technology Model Beta Version
* 0.18um PMOS SPICE Parametersv (normal one)
*

.model CMOSP PMOS (
+Level = 49

+Lint = 3.e-08 Tox = 4.2e-09
+Vth0 = -0.42 Rdsw = 450

+lmin=1.8e-7 lmax=1.8e-7 wmin=1.8e-7 wmax=1.0e-4 Tref=27.0 version =3.1
+Xj= 7.0000000E-08         Nch= 5.9200000E+17
+lln= 1.0000000            lwn= 1.0000000              wln= 0.00
+wwn= 0.00                 ll= 0.00
+lw= 0.00                  lwl= 0.00                   wint= 0.00
+wl= 0.00                  ww= 0.00                    wwl= 0.00
+Mobmod=  1                binunit= 2                  xl= 0.00
+xw= 0.00
+binflag=  0               Dwg= 0.00                   Dwb= 0.00

+ACM= 0                    ldif=0.00                   hdif=0.00
+rsh= 0                    rd= 0                       rs= 0
+rsc= 0                    rdc= 0

+K1= 0.5560000             K2= 0.00
+K3= 0.00                  Dvt0= 11.2000000            Dvt1= 0.7200000
+Dvt2= -1.0000000E-02      Dvt0w= 0.00                 Dvt1w= 0.00
+Dvt2w= 0.00               Nlx= 9.5000000E-08          W0= 0.00
+K3b= 0.00                 Ngate= 5.0000000E+20

+Vsat= 1.0500000E+05       Ua= -1.2000000E-10          Ub= 1.0000000E-18
+Uc= -2.9999999E-11        Prwb= 0.00
+Prwg= 0.00                Wr= 1.0000000               U0= 8.0000000E-03
+A0= 2.1199999             Keta= 2.9999999E-02         A1= 0.00
+A2= 0.4000000             Ags= -0.1000000             B0= 0.00
+B1= 0.00

+Voff= -6.40000000E-02      NFactor= 1.4000000          Cit= 0.00
```

```
+Cdsc= 0.00                Cdscb= 0.00                 Cdscd= 0.00
+Eta0= 8.5000000           Etab= 0.00                  Dsub= 2.8000000

+Pclm= 2.0000000           Pdiblc1= 0.1200000          Pdiblc2= 8.0000000E-05
+Pdiblcb= 0.1450000        Drout= 5.0000000E-02        Pscbe1= 1.0000000E-20
+Pscbe2= 1.0000000E-20     Pvag= -6.0000000E-02        Delta= 1.0000000E-02
+Alpha0= 0.00              Beta0= 30.0000000

+kt1= -0.3700000           kt2= -4.0000000E-02         At= 5.5000000E+04
+Ute= -1.4800000           Ua1= 9.5829000E-10          Ub1= -3.3473000E-19
+Uc1= 0.00                 Kt1l= 4.0000000E-09         Prt= 0.00

+Cj= 0.00138               Mj= 1.05                    Pb= 1.24
+Cjsw= 1.44E-09            Mjsw= 0.43                  Php= 0.841
+Cta= 0.00093              Ctp= 0                      Pta= 0.00153
+Ptp= 0                    JS=1.50E-08                 JSW=2.50E-13
+N=1.0                     Xti=3.0                     Cgdo=2.786E-10
+Cgso=2.786E-10            Cgbo=0.0E+00                Capmod= 2
+NQSMOD= 0                 Elm= 5                      Xpart= 1
+Cgsl= 1.6E-10             Cgdl= 1.6E-10               Ckappa= 2.886
+Cf= 1.058e-10             Clc= 0.0000001              Cle= 0.6
+Dlc= 3E-08                Dwc= 0                      Vfbcv= -1 )

.ENDL
```

Appendix B

Sensitivity parameters

This appendix gives details about the coefficients used in the performance sensitivities equations presented of chapter 5. We denote with k_x the performance sensitivity respect to the variation of a generic parameter x. We make a distinction between interconnect and device parameters. Interconnects sensitivities are presented in equation (5.6, 5.7, 5.8, 5.9, 5.10, 5.11) and while device sensitivities in equation (5.16, 5.17, 5.18, 5.19). Those equations are written explicitly as function of the buffer size and distance since we are interested in opportunities to reduce the sensitivities by tuning indeed size and distance of the buffers. However in chapter 5, there are omitted explicit expressions for the coefficients a_x, b_x and c_x. In this appendix we present the expressions for those coefficients. They are obtained by taking the first derivative of the performance (3.7) with respect to each of the varying parameters.

Interconnects

We assume simple parallel plate capacitance model for the wires.

$$c = c_g + 2SFc_v = \epsilon(\frac{w}{t_{ox}} + 2SF\frac{h}{s}) \tag{B.1}$$

$$r = \frac{\rho}{wh} \tag{B.2}$$

This is done to derive easily close form expressions for the coefficients a_x, b_x and c_x.

$$a_\rho = a\epsilon(\frac{1}{ht_{ox}} + \frac{2SF}{wp - w^2}) \qquad b_\rho = b\tfrac{c_Q}{wh} \tag{B.3}$$

$$a_\epsilon = a\rho(\frac{1}{ht_{ox}} + \frac{2SF}{wp - w^2}) \qquad b_\epsilon = br_0(\frac{w}{t_{ox}} + \frac{2SFh}{p - w}) \tag{B.4}$$

$$a_w = a\rho\epsilon 2SF\frac{p - 2w}{w^2(p - w)^2} \qquad b_w = br_0\epsilon(\frac{1}{t_{ox}} + \frac{2SFh}{(p - w)^2}) \qquad c_w = \frac{bc_0\rho}{hw^2} \tag{B.5}$$

$$a_{t_{ox}} = \frac{a\rho\epsilon}{t_{ox}^2 h} \qquad b_{t_{ox}} = br_0\epsilon\frac{w}{t_{ox}^2} \tag{B.6}$$

$$a_h = \frac{a\rho\epsilon}{t_{ox}h^2} \qquad b_h = br_0\epsilon\frac{2SF}{p - w} \qquad c_h = \frac{bc_0\rho}{h^2 w} \tag{B.7}$$

Devices

$$a_{c_{ox}} = \frac{bL_{\text{eff}}c}{\mu(V_{DD} - V_{TH})c_{ox}^2} \qquad b_{c_{ox}} = brL_{\text{eff}} \tag{B.8}$$

$$a_{V_{DD}} = \frac{bL_{\text{eff}}^2(\alpha_p + 1)}{\mu(V_{DD} - V_{TH})^2} \quad b_{V_{DD}} = \frac{bL_{\text{eff}}c}{\mu c_{ox}(V_{DD} - V_{TH})^2} \tag{B.9}$$

$$a_{V_{TH}} = \frac{bL_{\text{eff}}^2(\alpha_p + 1)}{\mu(V_{DD} - V_{TH})^2} \quad b_{V_{TH}} = \frac{bL_{\text{eff}}c}{\mu c_{ox}(V_{DD} - V_{TH})^2} \tag{B.10}$$

$$a_{L_{\text{eff}}} = \frac{2bL_{\text{eff}}(\alpha_p + 1)}{\mu(V_{DD} - V_{TH})} \quad b_{L_{\text{eff}}} = \frac{bc}{\mu c_{ox}(V_{DD} - V_{TH})} \qquad c_{L_{\text{eff}}} = brc_{ox} \tag{B.11}$$

Bibliography

[1] Internetional technology roadmap, 2001.

[2] C. Alpert, C. Chu, G. Gandham, M. Hrkic, Jiang Hu, C. Kashyap, and S. Quay. Simultaneous driver sizing and buffer insertion using a delay penalty estimation technique. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Trans. on*, 23(1):136–141, Jan. 2004.

[3] C. Alpert and A. Devgan. Wire segmenting for improved buffer insertion. In *Proc. pf the 34th Design Automation Conference*, pages 588–593, Jun. 1997.

[4] C.J. Alpert, A. Devgan, J.P. Fishburn, and S.T. Quay. Interconnect synthesis without wire tapering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Trans. on*, 20(1):90–104, Jan. 2001.

[5] G. B. Arfken and H. J. Weber. *Mathematical Methods for Physicists*. Academic Press, 1985.

[6] Semiconductor Industry Association. National technology roadmap for semiconductors. San Jose, CA, 1997.

[7] H. B. Bakoglu. *Circuits, Interconnections and Packaging for VLSI*. Addison-Wesley, 1990.

[8] K. Banerjee and A. Mehrotra. Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling. In *Symp. on VLSI Circuits*, pages 195–198, 2001.

[9] L. Benini and G. De Micheli. Networks on chips: a new soc paradigm. *IEEE Computer magazine*, 35(1):70–78, Jan. 2002.

[10] BPTM. Spice technology files, Berkeley Predictive Technology Model (BPTM) *http://www-device.eecs.berkeley.edu/~ptm/introduction.html*.

[11] R.E. Bryant, Kwang-Ting Cheng, A.B. Kahng, K. Keutzer, W. Maly, R. Newton, L. Pileggi, J.M. Rabaey, and A. Sangiovanni-Vincentelli. Limitations and challenges of computer-aided design technology for cmos vlsi. In *Proc. of the IEEE*, volume 89, pages 341–365, Mar. 2001.

[12] Y. Cao, P. Gupta, A.B. Kahng, D. Sylvester, and J. Yang. Design sensitivities to variability: extrapolations and assessments in nanometer vlsi. In *Proc. of the 15th Annual IEEE Int. ASIC/SOC Conf.*, pages 411–415, 2002.

[13] Yu Cao, Chenming Hu, Xuejue Huang, A.B. Kahng, I.L. Markov, M. Oliver, D. Stroobandt, and D. Sylvester. Improved a priori interconnect predictions and technology extrapolation in the gtx system. *Very Large Scale Integration (VLSI) Systems, IEEE Trans. on*, 11(1):3–14, Feb. 2003.

[14] J.H. Chern. Multilevel metal capacitance models for cad design synthesis systems. In *IEEE Electron Device Letters*, volume 13, pages 32–34, Jan. 1992.

[15] C.C.N. Chu and D.F. Wong. Closed form solution to simultaneous buffer insertion /sizing and wire sizing. In *Proc. of the Int. Symp. Physical Design*, pages 192–197, 1997.

[16] J. Cong, Cheng-Kok Koh, and Kwok-Shing Leung. Simultaneous buffer and wire sizing for performance and power optimization. In *Proc. of Int. Symposium on Low Power Electronics and Design*, pages 271–276, Aug. 1996.

[17] J. Cong, Tianming Kong, and D.Z. Pan. Buffer block planning for interconnect-driven floorplanning. In *IEEE/ACM Int. Conference on Computer-Aided Design*, 1999.

[18] J. Cong and Zhigang Pan. Interconnect performance estimation models for design planning. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 20(6):739–752, Jun. 2001.

[19] Infineon. Infineon corporate research web site:
http://www.infineon.com.

[20] Z. Daoud and C.J. Spanos. Doric: design of optimal and robust integrated circuits. In *Proc. of the IEEE Custom Integrated Circuits Conf.*, pages 361–364, 1994.

[21] S. Dhar and M.A. Franklin. Optimum buffer circuits for driving long uniform lines. *IEEE Journal of Solid-State Circuits*, 26(1):32–40, Jan. 1991.

[22] W.C. Elmore. The transient responcse of damped linear network with particular regard to wideband amplifiers. *J. Applied Physics*, 19:55–63, 1948.

[23] A. Fan, A. Phman, and R. Reif. Copper wafer bonding. *Electrochem. Solid State Lett.*, 2(10):534–536, 1999.

[24] J.P Fishburn. Shaping a vlsi wire to minimize elmore delay. In *Proc. on European Design and Test Conference ED&TC 97*, pages 244–251, Mar. 1997.

[25] G.S. Garcea, N.P. van der Meijs, and R.H.J.M. Otten. Simultaneus analytic area and power optimization for repeater insertion. In *Proc. Int. Conf. on Computer Aided Design*, 2003.

[26] G.S. Garcea, N.P. van der Meijs, K.J. van der Kolk, and R.H.J.M. Otten. Statistically aware buffer planning. In *Proc of the Design, Automation and Test in Europe Conf. (DATE)*, pages 1402–1403, 2004.

[27] R. Ho, K.W. Mai, and M.A Horowitz. The future of wires. In *Proceeding of the IEEE*, volume 89, pages 490–504, April 2001.

[28] Y.I. Ismail and E.G. Friedman. Effects of inductance on the propagation delay and repeater insertion in vlsi circuits. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 8(2):195–206, 2000.

[29] A. Jantsch and H. Tenhunen, editors. *Networks on Chip*. Kluwer Accademic Publishers, 2003.

[30] J.A.G. Jess, K. Kalafala, S.R. Naidu, R.H.J.M. Otten, and C. Visweswariah. Statistical timing for parametric yield prediction of digital integrated circuits. In *Proc. of the 40th Design Automation Conf.*, pages 932–937, 2003.

[31] H. Johnson and M. Graham. *High-speed signal propagation: advanced black magic*. Prentice-Hall, 2003.

[32] A.B. Kahng, S. Muddu, and E. Sarto. On switch factor based analysis of coupled rc interconnects. In *Proc. of the 37th Design Automation Conf*, pages 79–84, 2000.

[33] S.P. Khatri, A. Mehrotra, R.K. Brayton, A. Sangiovanni-Vincentelli, and R.H.J.M Otten. A novel vlsi layout fabric for deep sub-micron applications. In *Proc. 36th Design Automation Conference,* 1999.

[34] S. Kumar, A. Jantsch, M. Soininen, J.-P.; Forsell, M. Millberg, J. Oberg, K. Tiensyrja, and A. Hemani. A network on chip architecture and design methodology. In *Proc. IEEE Computer Society Annual Symposium on VLSI,* pages 105–112, 2003.

[35] S.W. Lee and S.K. Joo. Low temperature poly-si thin film transistor fabrication by metal-iduced lateral crystallization. *IEEE Electron Device Lett.*, 14(4):160–162, 1999.

[36] Tao Lin and L.T. Pileggi. Throughput-driven ic communication fabric synthesis. In *Proc. of the Int. Conference on Computer Aided Design,* pages 274–279, 2002.

[37] J. Liu, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen. A global wire planning scheme for network-on-chip. In *Proc. of the 2003 Int. Symposium on Circuits and Systems, ISCAS '03,* volume 4, pages 892–895, May 2003.

[38] A.V. Metcalfe. *Statistics in Engineering, a practical approach.* Chapman and Hall, 1994.

[39] C. Michael and M. Ismail. *Statistical Modeling for Computer-Aided Design of MOS VLSI Circuits.* Kluwer, 1992.

[40] Mosis. Spice technology files. On MOSIS Site http://www.mosis.org.

[41] F. Mu and C. Svensson. Analysis and optimization of uniform long wire and driver. *IEEE Trans. on Circuits and Systems-I: fundamental theory and applications,* 46(9):1086–1100, Apr. 1999.

[42] S.R. Nassif. Modeling and forecasting of manufacturing variations. In *Proc. of the 5th Int. Workshop on the Statistical Metrology,* pages 2–10, 2000.

[43] S.R. Nassif. Modeling and forecasting of manufacturing variations. In *Proc. of the Asia and South Pacific Design Automation Conf. (ASP-DAC),* 2001.

[44] G.W. Neudeck, S. Pae, J.P. Denton, and T. Su. Multiple layers of silicon-on insulator for nanostructure devices. *Journal of Vacuum Science and Technology-B*, 17(3):994–998, 1999.

[45] I. O'Connor. Optical solutions for system-level interconnect. In *2004 Int. Workshop on System-Level Interconnect Prediction*, pages 79–88, Feb. 2004.

[46] R.H.J.M. Otten. Complexity and diversity in ic layout design. In *Proc. of Int. Conference on Circuits and Computers*, pages 764–767, Oct. 1980.

[47] R.H.J.M Otten and G.S. Garcea. Are wires plannable? In *Proc. of Int. Workshop on System-Level Interconnect Prediction*, April 2001.

[48] R.H.J.M. Otten and P. Stravers. Challenges in physical chip design. In *Proc. of Int. Conference on Computer Aided Design*, pages 84–91, Nov. 2000.

[49] R.J.M.H Otten and R.K. Brayton. Planning for performance. In *Proc. of the 35th Annual Design Automation Conf.*, pages 122–127, Jun. 1998.

[50] D. Pamunuwa, Li-Rong Zheng, and H. Tenhunen. Maximizing throughput over parallel wire structures in the deep submicrometer regime. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 11(2):224–243, Apr. 2003.

[51] R. Parsons. *Statistical Analysis: A Decision Making Approach*, chapter 3, pages 59–64. Harper and Row Publishers, 1974.

[52] A. Rahman and R. Reif. System-level performance evaluation of three-dimensional integrated circuits. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 8(6):671–678, Dec. 2000.

[53] T. Sakurai and K. Tamaru. Simple formulas for two and three-dimensional capacitances. In *IEEE Trans. on Electron Devices*, volume 30, pages 183–185, 1983.

[54] H. Shah, P. Shiu, B. Bell, M. Aldredge, N. Sopory, and J. Davis. Repeater insertion and wire sizing optimization for throughput-centric vlsi global interconnects. In *Proc. of IEEE/ACM Int. Conference on Computer Aided Design, ICCAD*, pages 280–284, Nov. 2002.

[55] S.J. Souri and K.C. Saraswat. Interconnect performance modeling for 3d integrated circuits with multiple si layers. In *Proc. of IEEE Int. Conference on Interconnect Technology*, pages 24–26, 1999.

[56] B. Stine, D. Boning, and J. Chung. Analysis and decomposition of spatial variation in integrated circuits processes and devices. *IEEE Trans. on Semi. Manuf.*, 10(1):24–41, Feb. 1997.

[57] S. Strickland, E. Ergin, D.R. Kaeli, and P. Zavracki. Vlsi in the third dimension. *Integration, the VLSI Journal*, 25(1):1–15, Sept. 1998.

[58] D. Sylvester and K. Keutzer. Getting to the bottom of deep submicron. In *in Proc. of the IEEE/ACM Int. Conference on Computer-Aided Design*, pages 203–211, 1998.

[59] G. Taguchi. *Introduction to Quality Engineering*. Asian Productivity Organization, Tokyo, 1986.

[60] G. Taguchi and Yu-In Wu. *Introduction to Off-line Quality Control.* Central Japan Quality Control Association, Magaya, Japan (available from American Supplier Istitute, Inc., Dearborn, MI.), 1979.

[61] N.P. van der Meijs and A. van Genderen. Space : Accurate and efficient layout-to-circuit extraction for deep submicron technologies. web page: http://space.tudelft.nl.

[62] L.P.P.P. van Ginneken. Buffer placement in distributed rc-tree networks for minimal elmore delay. In *Int. Symp. Circuits Syst.*, pages 865–868, 1990.

[63] C. Visweswariah. Death, taxes and failing chips. In *Proc. of the 40th Design Automation Conf.*, pages 343–347, 2003.

[64] SPACE/WARP. Wire and Repeater Planning web site: *http://space.tudelft.nl/warp*.

[65] Shyh-Chyi Wong, Gwo-Yann Lee, and Dye-Jyun Ma. Modeling of interconnect capacitance, delay, and crosstalk in vlsi. *IEEE Trans. on Semiconductor Manufacturing*, 13(1):108–111, Feb. 2000.

[66] Yongtao You, B. Roetcisoender, A. Cheng, R. McGehee, and S. Sugiyama. Performance-driven layout through device sizing. In *Proc. of the IEEE Custom Integrated Circuits Conference*, pages 931–934, 1993.

[67] Qiang Zhang, J.J. Liou, J. McMacken, J. Thomson, and P. Layman. Development of robust interconnect model based on design of experiments and multiobjective optimization. *IEEE Trans. on Electron Devices*, 48(9):1885–1891, Sept. 2001.

[68] D. Zwillinger. *CRC Standard Mathematical Tables and Formulae, 31st Edition*. CRC Press, 2003.

Acknowledgements

This thesis comes at the conclusion of several wonderful years spent in Delft. This period has contributed greatly to my professional and personal growth. This university is not only a dynamic incubator of fresh ideas but it is also a free place to meet and to share experiences with a lot of people. In fact, I believe that my PhD is not only a personal achievement, but it is the result of their constant support and encouragement.

I want to start by acknowledging the people who have contributed to my scientific research, I would like to thank prof. Ralph Otten for his technical suggestions and his patience. He gave me the enthusiasm to accomplish the writing of this thesis. A special appreciation goes to my supervisor Nick van der Meijs for giving me the opportunity of being a Ph.D. student of the CAS group. I have received from him the freedom and the necessary calm to finish my research. I thank him for the many technical and sometimes more philosophical discussions.

I wish to thank prof. Patrick Dewilde, prof. Alle-Jan van der Veen and all the students and members of the Circuits and System group. It was a pleasure to work in this stimulating environment. Thanks to Kees-Jan van der Kolk who made a wonderful work by implementing a web based interface of the models presented in this thesis. I want to thank Eelko Schrik, my office roommate during the last four years. He has tolerated my noises everyday for all this time and he has found the time to help with the translation in dutch of the summary of this thesis. I don't want to forget also Relija Djapic and Antonio Trinidade really "vent valves" for my complains in some grey dutch days.

This project was founded by the interfaculty project DIOC 16: "Smart Product Systems". I would like to thank all the members of this project for giving me the opportunity to learn about very interesting aspects of environmental design, life cycle of the products and new trends in efficient car design.

I cannot forget the whole italian community, without them I could not have survived here for such long time. Many people have left and many others have arrived in those years. Among the friends that are now far away, I would like to mention Roberto, Paola, Marco D, Marco M, Roberta, Francesco, Davide, Marco P and Raysa. Our close friendship is deeply related to the many evenings spent together in Delft. I will never forget your presence, patience and support during a very critical period at the beginning of my PhD. Probably I would have given up with my PhD if I had not met you!

Since I was not able in those years to learn how to speak a decent dutch, I had at least the possibility to help dutch people willing to learn the italian language. Thanks to my fellows in this adventure Antonio, Erika and Barbara. I would like also to thank Alberto, Romina, Marcello, Cristina, Mark and Valentina for the countless dinners together and for listening patiently to my complains originating from my "dutch intoxication". I wish to thank also Marco S and Francesco de P. They have contributed in keeping me alive during the final stages of the writing. Last but not least I would like to mention also Marco B, a very good company during the long commuting to Eindhoven.

Finally I cannot forget my family. Thanks to my parents who have been certainly suffering a lot for living at great distance from their sons. I really thank them to have given me the freedom to choose always what was better for me. I want to mention also my two brothers Angelo and Alessandro because they were always present when I needed them. My list of acknowledgments cannot finish without thinking about my sweet Josine. I wish to thank her for the patience and understanding all the times when I was too busy with my "*x*+*y*" problems and I was not able to give her the love she deserved.

Giuseppe Garcea
May 2005, The Hague

Summary

Global interconnect lines in current and future technologies require insertion along their length of appropriate signal regenerating systems, typically simple inverters or inverter stages. This is done in order to prevent the wire delay from becoming quadratic in length. The sizes of the buffers and the distances between them have to be tuned to meet the performance specifications.

By using Elmore's delay metric, Bakoglu was able to estimate analytically for uniform buffered wires the distance and size which maximize their performance. However, the proposed solution has limited use in practice because these buffers occupy a very large area on the active layer, therefore it only contributes to fixing a theoretical upper performance limit. In fact, constrained buffer locations, limited total buffer area and power consumption should be considered in a realistic buffer planning for global wires. Uniform buffer insertion, while not theoretically optimal in presence of constraints, is still a useful theoretical model since in practice moderate deviations are only slightly non-optimal, and many practical buffer insertion techniques start from a uniform buffer insertion model.

In our work, we show that performance, in presence of design constraints, can be optimized considerably by resizing buffers and changing their distances. A method to achieve this buffer tuning is presented, which can be a useful tool for resource allocation while meeting performance specifications. In particular, we present an analytic formula for buffer insertion for global interconnects that simultaneously minimizes the silicon device area and power dissipation for a given performance. The analytic model requires only basic device and interconnect characterization data that are easily obtained. Thus, we have developed a practical pareto-optimal buffer insertion theory. This methodology shows for example that the use of small buffers in the deterministic case seems to have a relatively low impact on performance. Thus, in order to save silicon area, it is convenient then to use relatively small buffers. For example, 85% of the performance requires only 30% of the area and 67% of the power compared to absolute maximum performance, reached in the unconstrained (Bakoglu) case.

The approach to deep-submicron dimensions is pushing the fabrication technology to the physical limit. This means that tolerances on the geometrical dimensions are relatively increasing. Since the influence of interconnects on performance is predicted to increase with scaling, it is becoming relevant to identify the role of interconnect-caused variations on the total performance variability.

While we do not assume any given parameter distribution for a given process but only use nominal value and variance, we do propose a generic statistical approach. This approach is accurate for small to moderate process variations and is based on the second-order Taylor approximation of Elmore's delay model subject to variability. The impact of those variations on the performance can then be characterized by analytical expression for the the expected mean and variance of the performance. A buffer strategy having the objective of minimizing performance variability is then derived.

Compared to deterministic buffer planning, this study of performance variability indicates that by selecting small buffers there is a considerable increase of performance variability. Therefore the choice of small buffers becomes less convenient.

Buffer planning for single buffered wires is generalized for the design of wire buses typically used for network on chip (NoC) applications. A complete framework to optimally relate wire density, total buffer area, power and throughput (data rate) has been developed.

Samenvatting

Globale interconnect lijnen in huidige en toekomstige technologieën vereisen het invoegen van signaal regenererende systemen, typisch invertors of invertorschakels, langs hun lengte. Dit wordt gedaan om te voorkomen dat de delay (signaalvertraging) langs de draad kwadratisch wordt met de lengte van de draad. De grootte van de buffers en hun onderlinge afstand moeten dusdanig worden afgestemd dat de performance specificaties worden gehaald.

Met behulp van Elmore's delay-metriek is Bakoglu in staat geweest om de grootte en onderlinge afstand van de buffers in uniform gebufferde draden analytisch te schatten, zodanig dat de performance wordt gemaximaliseerd. De voorgestelde oplossing is in de praktijk echter beperkt bruikbaar, omdat de buffers een zeer grote oppervlakte beslaan op de actieve laag, en daarmee alleen bijdragen in het vaststellen van een theoretische bovenlimiet van de performance. In werkelijkheid kunnen de buffers niet geheel vrij worden geplaatst en zijn er beperkingen in de totale beschikbare oppervlakte voor de buffers en in het totale toegestane energieverbruik. Bij een realistische planning van buffers in globale bedrading moeten deze factoren in overweging worden genomen. Het uniform invoegen van buffers, hoewel theoretisch niet optimaal, is desalniettemin een bruikbaar theoretisch model omdat in de praktijk niet al te sterke afwijkingen maar beperkt niet-optimaal zijn en veel praktische bufferinvoegingstechnieken starten vanuit een uniform bufferinvoegingsmodel.

In ons werk laten we zien dat er aanzienlijke mogelijkheden voor performance optimalisatie bestaan door voortdurend de grootte van de buffers en hun onderlinge afstand te varieren. We presenteren een methode om op deze manier buffers op elkaar af te stemmen die een nuttig gereedschap kan zijn bij het zo goed mogelijk gebruiken van de beschikbare middelen (in termen van energie en oppervlakte) terwijl de performance specificaties worden behaald. In het bijzonder presenteren we een analytische formule voor bufferinvoeging op globale bedrading die, voor een gegeven performance, gelijktijdig oppervlakte en energieverbruik van de buffers minimaliseert. Het analytische model vereist slechts basisgegevens over de eigenschappen van de gebruikte devices en van de bedrading; deze gegevens kunnen gemakkelijk worden verkregen. Hiermee hebben we een praktisch bruikbare theorie ontwikkeld voor pareto-optimale bufferinvoeging. Deze methodologie laat bijvoorbeeld zien dat het gebruik van kleine buffers in deterministische situaties slechts een relatief kleine invloed heeft op de performance. Het blijkt daarmee gunstig om relatief kleine buffers te gebruiken. Bijvoorbeeld, vergeleken met de situatie berekend door de methode van Bakoglu, kan met 30% van de oppervlakte en 67energieverbruik nog altijd 85% van de maximale performance worden behaald.

De voortgang naar steeds diepere submicron-afmetingen drijft de fabricagetechnologie naar de fysieke limiet. Dit betekent dat de toleranties op de geometrische afmetingen relatief toenemen. Omdat er wordt voorspeld dat de rol van bedrading in de performance zal toenemen met de schaling van technologie, wordt het relevant om te identificeren wat de rol zal zijn van variaties in de bedrading op de variabiliteit van de totale performance.

Alhoewel we niet uitgaan van enige gegeven parameter distributie voor een gegeven proces, maar slechts nominale waarde en variantie gebruiken, stellen we een generieke, statistische aanpak voor. Deze aanpak is nauwkeurig voor kleine tot middelgrote procesvariaties en is gebaseerd op de tweede-orde Taylor benadering van Elmore's delay model onderhevig aan variabiliteit. Het effect van de variaties op de performance kan dan worden gekarakteriseerd door een analytische expressie voor het verwachte gemiddelde en de verwachte variantie van de performance. Een bufferstrategie die de minimalisatie van performance-variabiliteit tot doel heeft wordt dan afgeleid.

Vergeleken bij deterministische buffer planning, toont deze statistische studie aan dat kleine buffers de performance variabiliteit aanzienlijk doen toenemen, waarmee hun bruikbaarheid minder wordt.

Buffer planning vastgesteld voor enkelvoudige, gebufferde draden wordt gegeneraliseerd naar het ontwerp van bussen, zowel met als zonder afschermingsdraden.

Curriculum Vitae

Giuseppe Garcea was born on December 29, 1973 in Taranto, Italy. He attended the secondary school at Liceo Scientifico "G. Ferraris" in Taranto, Italy, where he obtained his diploma in 1992. In the same year, he started his study in Electrical Engineering at University of Florence, Italy. He completed his graduation project visiting, as Erasmus student, the Circuits & Systems (CAS) group at Delft University of Technology, The Netherlands. His M.Sc thesis was about the compilation methods for mapping a class of digital signal processing algorithms onto matching stream-based parallel dataflow architectures. After he graduated in 1998, he attended the post-graduate school DISC (Dutch Institute of Systems and Control) supported by the Control Laboratory group of Delft University of Technology.

From 1999, again as student of the Circuits & Systems group at TU Delft, he started to work towards his Ph.D. His project, which has resulted in this dissertation, was part of the DIOC (Delft Center for Interfacially Research) program *Smart Product Systems*.

Since October 2004 he is working for Magma Design Automation in the Netherlands.