Buffer Planning for Global Wires Under Statistical Process Variations

Giuseppe S. Garcea, Nick P. van der Meijs, Ralph H. J. M. Otten *
Delft University of Technology, Faculty of EEMCS, The Netherlands
* Eindhoven University of Technology, Faculty of EE, The Netherlands
giuseppe@cas.et.tudelft.nl

Abstract—
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. This will give the possibility to develop a “variational aware” buffer insertion methodology for global wires. 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. In fact, our results show that the low performance designs that are sized for minimum area using an optimal deterministic approach, can suffer from low parametric yield.

I. INTRODUCTION

Due to the increasing complexity of the processes and due to the incredible number of parameters affecting the quality of their manufacturing, it is starting to become more and more difficult for technologists to guarantee that the design is “safely” manufacturable if the design rules are respected. Possible process tolerances, which technologists can not easily control or for which control could be not economical justified, then have to be considered at the design level and new analysis tools have to be provided and/or combined with the traditional one.

As semiconductor technology is aggressively scaled, the time constants of VLSI circuits are 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 [1], [2], [3].

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. Computational costs for this analysis by using simulations or Monte Carlo methods can be prohibitive especially in consideration of the fact that the design space is huge because of the number of statistical variables involved, their variances and the increasing number of nets interconnect delay dominated.

In this paper, we have chosen for an analytic approach. 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 consider the mean and (co-)variance of the relevant manufacturing parameters. We will assume Gaussian variability models, indeed enabling the analytic approach, which seems a natural choice if no more detailed information is available.

We will show that this approach can be useful for the problem of uniform buffer insertion for a point to point connection. Buffer insertion (also known as repeater insertion) is a common technique avoiding that the propagation delay becomes quadratic in the length of the line due to its RC properties [5], [6], [7]. The buffer insertion problem actually is an optimization problem of the buffer size \( w_b \) and distance between the buffers \( l \). There is a well defined tradeoff between performance and power consumption vs silicon area, the latter actually being proportional to \( w_b/l \).

In this paper we will in fact show that some part of the \( (w_b,l) \) design space that is considered optimal for small area in [7], actually suffers from large performance variability.

In this paper, we will only briefly summarize the models derived in [7] and [12]. Instead, we will connect the results from [7] and [12] to the parametric yield (defined as the fraction of systems that satisfy a designer-defined minimum performance) of point-to-point interconnect-buffer systems.

The rest of this paper is organized as follows. In Section II we introduce the statistical tools that we use. Then, in Section III we present the analytical model for optimal buffer insertion in point to point interconnections in a deterministic setting. Section IV introduces the variability sources and their impact on the performance. Subsequently we discuss in Section V the parametric yield which relates the interconnect and device variability to the vol-
ume of systems meeting the performance specifications.

II. MODELLING PERFORMANCE VARIABILITY

The approach proposed [10] uses the 2nd order Taylor expansion of a function \( \phi \), 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:

\[
\phi(P) = \phi(P^0) + \sum_i \frac{\partial \phi}{\partial p_i} (p_i - p_i^0) + \frac{1}{2!} \left( \sum_i \frac{\partial^2 \phi}{\partial p_i^2} (p_i - p_i^0)^2 + 2 \sum_{i,j} \frac{\partial^2 \phi}{\partial p_i \partial p_j} (p_i - p_i^0)(p_j - p_j^0) \right) + \ldots
\]  

(1)

where all the partial derivatives are evaluated at the point \( P^0 \). Thus, we have reduced our function to a 2nd order expansion of random variables which can be treated much easier.

The average (expected) value is given by \( E(\phi) = \phi(P^0) + \text{bias}(\phi(\phi)) \). A general first order expression for \( \text{bias}(\phi(\phi)) \) is from (1)

\[
\text{bias}(\phi(\phi)) = \frac{1}{2} \sum_i \frac{\partial \phi}{\partial p_i} \sigma(p_i)^2 + \frac{\partial^2 \phi}{\partial p_i \partial p_j} \text{cov}(p_i, p_j)
\]  

(2)

where \( \sigma(p_i) \) is the standard deviation of \( p_i \) and \( \text{cov}(p_i, p_j) \) is the covariance between \( p_i \) and \( p_j \). Thus, the bias is 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 \( \mu_x \) and \( \mu_y \) and variances \( \sigma_x^2 \) and \( \sigma_y^2 \), the total variance of \( z \) becomes \( \sigma_z^2 = a^2 \sigma_x^2 + b^2 \sigma_y^2 \) [10]. This result can be generalized for a generic number of random variables which together with (1) gives:

\[
\sigma^2(\phi(P)) = \sum_i \left( \frac{\partial \phi}{\partial p_i} \right)^2 \sigma(p_i)^2
\]  

(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.

The notation \( \sigma_n(\phi(P)) \) is the normalized standard deviation, denoted by the following ratio:

\[
\sigma_n(\phi(P)) = \sqrt{\frac{\text{var}(\phi(P))}{E(\phi(P))}}
\]  

(4)

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, which is defined as the signal velocity propagation. The standard deviation will then be estimated using (4). This variability estimation can drive a new approach to uniform buffer planning for point to point connections.

III. DETERMINISTIC BUFFER PLANNING

In this section, we will only summarize the results about optimal unconstrained buffer insertion, which are treated extensively in [5][6]. We start with the equivalent circuit model of a uniform RC line driven and loaded by an inverter, see Figure 1, of which the delay is denoted by \( \tau \).

![Fig. 1. Generic restoring buffer model](image)

Our delay model is a generalized Elmore delay model, determined by \( a \) and \( b \) (\( a = 0.37 \) and \( b = 0.69 \) for the 50% delay):

\[
\tau = bR_{tr}(C_L + C_p) + b(R_{tr}c + rC_L)l + arcl^2
\]  

(5)

Here, \( C_L \) is the buffer load capacitance, \( C_p \) the parasitic output capacitance, \( R_{tr} \) the buffer output resistance and \( c \) and \( r \) the interconnect capacitance and resistance per unit length, respectively. \( C_L \) is further defined as \( C_L = c_0 w_b \), where \( c_0 \) is the buffer load capacitance per unit buffer size, and \( w_b \) is the buffer size. Thus, \( l \) and \( w_b \) are the principal design parameters. Also, we have \( C_p = c_p w_b \), and \( R_{tr} = r_0/w_b \) where \( c_p \) is the normalized parasitic output capacitance and \( r_0 \) is the normalized buffer output resistance. Moreover, we define \( \tau_0 = r_0(c_0 + c_p) \) and \( \tau_l = rc \). These are technology related constants characterizing the devices and the interconnects, respectively.

If a global wire of total length \( L \) is regularly segmented into \( n \) cells such that \( n = L/l \), then the total delay along the wire is

\[
T = n\tau = L \left( \frac{b\tau_0}{l} + b\left( \frac{r_0}{w_b}c + r_c w_b \right) + a\tau_l l \right)
\]  

(6)

This delay can be minimized by separately optimizing \( l \) and \( w_b \) by setting the appropriate derivatives equal to zero. The result is

\[
l_{crit} = \sqrt{\frac{b\tau_0}{a\tau_l}} \quad w_{opt} = \sqrt{\frac{r_0 c}{c_0 r}}
\]  

(7)

This exact combination of \( l_{crit} \) and \( w_{opt} \) provides the absolutely best performance that can be obtained under the chosen deterministic delay model [5], [6].
In [7], optimal buffer planning is extended to the case in which only a limited total buffer area is available and/or the total buffer power is bounded. A precise and optimal tradeoff between \( \nu_b \) and \( l \) is presented, such that the total buffer size (modeled as \( \nu_b L/l = nw_p \) where \( n \) is the number of buffers) as a function of the performance and power is minimum. 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. Also, only 10% of the total, unconstrained optimum buffer area and 58% of the corresponding power is needed for 95% of the absolute maximum performance.

The study of variability introduced in the next sections will pose a limit to the minimum buffer sizes. We will show that too small buffers will degrade the predictability of the system.

**IV. SOURCES OF PERFORMANCE VARIABILITY**

**A. Interconnect variability**

Our interconnect model and the definition of some basic parameters are illustrated in Figure 2.

![Interconnect parameters](image)

Fig. 2. Interconnect parameters

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

\[
c = c_{\text{bottom}} + \Gamma c_{\text{top}} + SF (c_{\text{ib}} + c_{\text{hc}}) = \epsilon (1 + \Gamma) \frac{w}{t_{\text{ox}}} + 2SF \frac{h}{s}
\]  

(8)

where \( SF \) is the so-called switching factor \([11]\) and \( \Gamma = 0 \) if the wire is on the top interconnect layer and \( \Gamma = 1 \) if it is on an intermediate layer. In the rest of this paper, we set \( \Gamma = 0 \) (top layer) and we set \( SF = 1 \). However, it is possible to consider the switching factor as a statistical variable itself if so desired.

The parameters that we model as statistical variables are correlated. For example, the wire aspect ratio (defined as \( h/w \)) has technological limitations. However, we will assume that a strong correlation can only be identified between \( w \) and \( s \) such that \( w = p - s \). 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 simplifies the analytic treatment. However, correlation among these low level parameters can be included if necessary, by including the appropriate covariance term in (3) \([10]\).

We use typical parameters for a 0.18\( \mu \)m technology. Nominal values for interconnect parameters with the relative standard deviation used in this paper, are summarized in Table I.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Nominal value</th>
<th>( \sigma_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>2.2( \mu )\Omega\cdot cm</td>
<td>0.2</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>3.5*8.85( \text{pF/m} )</td>
<td>0.2</td>
</tr>
<tr>
<td>( w )</td>
<td>0.7( \mu )m</td>
<td>0.2</td>
</tr>
<tr>
<td>( h )</td>
<td>1.2( \mu )m</td>
<td>0.2</td>
</tr>
<tr>
<td>( t_{\text{ox}} )</td>
<td>1( \mu )m</td>
<td>0.2</td>
</tr>
<tr>
<td>( SF )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( p )</td>
<td>1.4( \mu )m</td>
<td>0</td>
</tr>
</tbody>
</table>

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.

**B. Device variability**

The delay model from Section III needs \( r_{0}, c_{0} \) and \( c_{p} \) as basic device parameters. Like \( r \) and \( c \) in Section IV-A, 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

\[
c_{0} = c_{\text{ox}} L_{\text{eff}}
\]  

(9)

where \( c_{\text{ox}} \) is the gate capacitance per unit area and \( L_{\text{eff}} \) is the effective channel length. Thus, we have \( c_{\text{ox}} \) and \( L_{\text{eff}} \) as two parameters for our model. We could further decompose \( c_{\text{ox}} \) 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
square MOS transistor is:

\[ R_{tp} = \frac{L_{eff}/W}{\mu c_{ox}(V_{DD} - V_{TH})} \]  

(10)

from which

\[ r_0 = kn_v R_{tp} * W = \frac{L_{eff}}{\mu c_{ox}(V_{DD} - V_{TH})} \]  

(11)

where \( kn_v \) is a calibration parameter which can be extracted using SPICE. 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 II.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Nominal value</th>
<th>( \sigma_{\mu} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_{ox} )</td>
<td>7.342e - 14 ( \mu F/\mu m^2 )</td>
<td>0.2</td>
</tr>
<tr>
<td>( V_{DD} )</td>
<td>1.8V</td>
<td>0.2</td>
</tr>
<tr>
<td>( V_{TH} )</td>
<td>0.35V</td>
<td>0.2</td>
</tr>
<tr>
<td>( L_{eff} )</td>
<td>0.115( \mu m )</td>
<td>0.2</td>
</tr>
</tbody>
</table>

V. PARAMETRIC YIELD

The statistical analysis method from Section II allows us to approximate the mean and variance of the performance of the point-to-point interconnections system designed using the method from Section III. We use as input variability the parameters from Section IV. Our performance metric is \( v \), the signal speed. 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 [8], [9].

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 in this paper work with the 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 [5], [6]. 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

\[ Y = \int_{v \geq v_{min}} pdf(v)dv \]  

(12)

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 (2), is zero) and by the standard deviation calculated according to (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 [10].

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 III [7]. Our analysis is based on the technology data with variability from Section IV.

In Figure 3, the solid line presents the optimal trade-off between performance and buffer area from [7] as summarized in Section III. 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 (12). 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%.

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 3 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 IV.

Figure 3 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 considerably 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.

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 4, which displays the yield as a function of the segment length \(l\) and the buffer size \(w_b\) for \(v_{min} = 0.8\).

In particular, Figure 4 suggests that small buffers are more strongly decreasing yield than long segments. We intend to investigate these trade-offs further in subsequent work.

VI. CONCLUSION

We have developed a general analytic approach for estimation of 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.

Using this model, we could estimate the parametric yield of uniformly buffered wires (defined as the fraction of chips that are faster than the design goal) as a function of buffer size and segment length. 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 2nd order Taylor expansion exists.

In subsequent work, we intend to investigate the trade-off mechanisms between buffer size, segment length, area, performance and yield more thoroughly. In this analysis, we can include the situation that the actual performance under statistical variations can be faster than the one predicted by deterministic model.

REFERENCES