Design and Selection of Buffers for Minimum Power-Delay Product
Sandra Turgis, Nadine Azemard, Daniel Auvergne

To cite this version:

HAL Id: lirmm-00239400
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00239400
Submitted on 5 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Design and Selection of Buffers for Minimum Power-Delay Product

S.Turgis, N.Azemard, D.Auvergne
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
LIRMM UMR CNRS 9928 Un de Montpellier II 161 Rue ADA
34392 Montpellier FRANCE

Abstract

Using explicit modeling of delays we present and discuss real design conditions of CMOS buffers from the viewpoint of power dissipation. Efficiency of buffer implementation is first studied through the definition of limit for buffer insertion. Closed form alternatives to the design for minimum power-delay product are then proposed in terms of this limit. Validations are obtained through SPICE simulations on two stage inverter arrays. Applications are given to standard cell library in comparing implementations for different selection alternatives.

1. Introduction

Driving buffers have been extensively used to control delays on combinatorial paths. Values of tapering factors were determined depending on the performance modeling level and on the physical representation of the cells involved with a common objective: minimizing the delay of paths.

In an initial simple theory Lin and Linholm [1] introduced the fixed tapered buffer where the minimum propagation delay time is achieved when the output current drive to output capacitance ratio, in each buffer stage remains constant. Jeager [2] showed that the optimum tapering factor was a constant $e = 2.72$. This result has been then highlighted in the popular books of Mead and Conway [3] and Weste and Eshragian [4].

Improvements to this simple model were obtained by including inertial delay of gates [5,6,7] in the so called split capacitor model. Sakurai [8] also extended the model using standard cell characterization laws, to define optimal tapering factors.

First consideration on power minimization for buffer design have been given by Veendricks [9] in his widely recognized attempt to model the influence of input rise and fall time on short circuit power dissipation. Vemuru [10] introduced then a variable tapered approach to save area (and dynamic power). Using calibrated analytical equation of delays [11] Sha Ma [12] obtained 20-30% energy saving in designing multistage variable tapered buffers from multivariable Kuhn-Tucker optimization.

CAD tools have been proposed to size gates [13,14,15], moreover they use numerical programming techniques which are far to be intuitive for designers. An exception is offered in [16] where a local optimization has been used to define explicit sizing equations for equal rise and fall times on individual gates, resulting in a variable taper implementation with smaller area than that obtained with global methods. The last published work to date addresses the problem of buffer design with local interconnect capacitance [17] by conserving a constant load capacitance to current drive tapering ratio.

A complete treatment of buffers must use an accurate explicit modeling of delay and power, showing clearly up design and process parameters in order to useful in full custom design approach as well as with standard cell methodology... In the work presented here, we address this important problem of designing tapered buffers, with low power-delay product, using analytical means.

This paper is organized as follows. The delay modeling is described in section 2. In section 3 we specify the power component to be considered in the objective function. Using the results obtained in section 2 it is then possible to define a general sizing strategy for buffers, this is given in section 4.

Section 5 introduces the concept of load limit for buffer insertion. Design conditions for a minimum power delay product are examined in section 6 and applied to standard cell library in section 7. In section 8, we draw a conclusion and discuss a speed up strategy based on template selection and fanout limit evaluation.

The main contributions of this paper include: the general formulation of buffer sizing rule, the definition of buffer insertion limits and the application to the buffer design to satisfy minimum power delay product constraints.

2 - Delay Modeling

An accurate delay analysis must take into account real cell parameters such as technology, structural complexity, cell environment including total output load and input slope effects. To obtain a better accuracy than the RC model (18), we used an explicit formulation based on the physical modeling of the switching operations of individual transistors [11,19].

It allows a real delay evaluation of inverters as a linear combination of the step responses of the driving (i-1) and the controled (i) cell (Fig.1) as follows:

$$t_{HL, LH}^{i} = \frac{A \cdot t_{LHS, HLS}(i-1) + t_{LHS, HLS}(i)}{1 + \alpha . A \cdot t_{LHS, HLS}(i-1)} .$$ (1)

where $A$ specifies the voltage range to be considered in evaluating input slope effect

$$(A = 1 - 2 \cdot V_T \cdot V_{DD}) , \alpha$$ is a slow input ramp correcting factors [19] and $t_{LHS, HLS}$ are the fall and rise step responses of inverters, evaluated for a variation of the output voltage from the output static level to half supply voltage ($V_{DD} / 2$).
These responses can be directly obtained from the mean charge transfer evaluated across the node under consideration and produced by the imbalance current developed in the cell under evaluation, as follows:

$$t_{HLS}(i) = \frac{C_{load}}{2C_N(i)} \cdots t_{LHS}(i) = \frac{\mu_N}{\mu_P} \frac{C_{load}}{2C_P(i)}$$  \hspace{1cm} (2)

for an inverter made up of N (C_N) and P (C_P) transistors, loaded by a capacitance C_{load}, where:

$$\tau_T = \frac{2CoxW_NL_{min}}{\mu_NV_th(V_{th} - V_N)} \frac{8V_{th}}{V^2 + 4V^2} - \frac{12V_{th}V_N}{NC_{par}(i)}$$ \hspace{1cm} (3)

is the elementary fall time characteristic of the technology, which can easily be found to be the fall time of a sequence of ideal inverters (W_N=W_P and no parasitic capacitance) with minimum length transistors, loaded by an identical one. The ratios C_{load}/C_N, C_{load}/C_P can be understood as the ratio of load capacitance to the current drive of each switching structure. Note that $\tau_T$ and $\mu_N/\mu_P$ (eq. 2), represent speed characteristic parameters of the considered process, and are defined through average values of mobilities and transistor lengths across the full voltage excursion. Their values can be extracted directly from the characterization of specific oscillators or calibrated from simulation of inverters [20].

### 3 - Power Dissipation in CMOS Buffers

The main component of power to be considered arises during the input signal transition, and consists of two parts, a dynamic dissipation ($P_{DL} = C_{load}V_{DD}^2.f$) and short circuit one ($P_{SC} = I_{average}.V_{DD}$). The dynamic power is the power dissipated to modify the charge content of the capacitive load, it is proportional to the switching frequency and to the load capacitance, and thus depends on the area of the buffers. The short circuit power dissipation is produced by the simultaneous conduction of P and N transistors during the transition of the input signal. As shown in [9,21] the ratio of input to output transition times is a good indicator of short circuit power dissipation. For ratio values comparable or lower than unity, which is the objective of buffer design, the short circuit power dissipation can be neglected. Buffer design with fixed or variable tapering factors have been obtained [10] with short circuit power content lower than 10% of the total power dissipation. As a result, in the definition of our optimizing objectives, we will just consider the dynamic component:

$$P = f \cdot V_{DD}^2 \cdot C_{REF} \cdot \frac{C_{load}}{C_{REF}}$$ \hspace{1cm} (4)

where C_{tot} is the total value of the load capacitances involved in the design, it includes layout parasitic (diffusion and interconnect) and active loads (C_{IN}(i)= C_N(i) + C_P(i)) and C_{REF} is the minimum capacitance considered in the design methodology or in the library under consideration.

### 4 - Buffer Design

An interesting analysis of the power dissipation of buffers has been published recently [22]. A comparison of uniform and nonuniform tapered buffers for minimum switching energy, using simple RC model, is given. For an ideal array constituted of N stages, design rules can be summarized as follows:

$$T_N^* = \frac{C_{load}}{C_{IN}} \cdots N = \frac{\ln C_{load}}{\ln T}$$ \hspace{1cm} (5)

where C_{load} is the load capacitance, C_{IN} the minimum input capacitance and T the tapering factor between stages. The output ratio of PMOS to NMOS transistors is assumed to be the same for all stages, and no consideration is given on the symmetry of the responses. As expected for ideal array, the minimum delay time is obtained for T = 2.72 while the minimum power delay product is reached for higher values (T=4.25). Considering real structures with layout parasitic capacitances these factors are significantly increased depending on the relative parasitic content. A non uniform tapering factor can only be considered for minimizing the power delay product and results in only few per cent of improvement [22] with respect to uniform implementations. Minimization of the power-delay product can then be resumed (for regular structures, but this can be solved at the layout level) to the search for a minimum of the product active area-delay. Due to the tapering factor values involved in real buffer arrays it appears reasonable to limit the study to a maximum of N=3 stages. In order to take into account input controlling ramp [19], the minimum structure to be considered for design optimization must be constituted of two inverters and this guarantees convexity of delays. Let us now consider a real two stage buffer constituted of inverters with identical internal configuration ratios (k=C_P/C_N), the delay expressions become:

$$\begin{bmatrix}
\theta_{HL} = \frac{\mu_N}{\mu_P} \frac{1}{k} \frac{(1+k)C_{N2} + C_{par1}}{C_{N1}} \\
\frac{(1+k)C_{N3} + C_{par2}}{C_{N2}}
\end{bmatrix}$$ \hspace{1cm} (6)

with equivalent expression for the rising edge. In this equation the delay is expressed in reduced units ($\theta = 2t/(1+A)\tau_T$), and the subscripts (1,2,3) specify the rank of the buffer, from the input of the array to the output, respectively, $C_{par}(i)$ represents the parasitic output capacitance of each inverter. As deduced from the posynomial variation of delay equations, constraint on equality of delays between falling and rising edges results in the equality of loading factors (load to drive ratio):

$$\frac{(1+k)C_{N2} + C_{par1}}{C_{N1}} = \frac{(1+k)C_{N3} + C_{par2}}{C_{N2}}$$ \hspace{1cm} (7)

which gives:

$$C_{N2} = \frac{C_{par1}}{2(1+k)} - 1 + \frac{4(1+k)C_{N1} n(1+k)C_{N1} + C_{par2}}{C_{par1}^2}$$ \hspace{1cm} (8)

where it is assumed that C_{N3} is a tapered value of C_{N1}. Cancellation of the derivative of $\theta$ with respect to C_{N2} cannot be obtained in closed form. Replacing, in eq.6, C_{N2} by the value given in eq.8, we plot on figure 2 the variation of $\theta$. 

---

**Figure 1:** General configuration for a real delay evaluation.
For the total range of considered loads, sizing with k=1 results in delay degradation ranging from 36% for purely parasitic load to 5% for purely active one. However for buffer array where the parasitic content is not higher than the gate capacitance value of the corresponding transistors [24], the degradation of performance obtained, by selecting identical N and P type transistors to size the inverters, is not much higher than 20% of the minimum available value for a 75% reduction of the transistor area (dynamic power).

5 - Buffer Insertion Limit

Next point of interest is to characterize the performance of inverters in terms of load and design parameters, and to define the limit of loads at which any cell can be speed up by insertion of wider inverters. This corresponds to the buffer insertion limit definition for each cell. As previously discussed the general performance equation for inverters can be written as:

\[ \text{Delay} = t_0 + K \cdot \frac{C_L}{C_{IN}} \]  
(9)

where the signification of the different terms can be found from eq. 2. \(t_0\) corresponds to the delay introduced by the inverter layout parasitic capacitance (usually specified as inertial delay), \(K\) is characteristic of the gate strength and is defined (eq.8-14) by the sizing strategy of each inverter.

Figure 3 illustrates the load limit of an inverter, defined as the fanout limit, at which buffer insertion improves the speed of the cell. Using eq. 6 (fig.4) , we have to satisfy the condition: \(t(a) > t(b)\).

This results in the straight condition:

\[ t_01 + K_1 \cdot \frac{C_{\text{load}}}{C_{IN1}} > t_01 + K_1 \cdot \frac{C_{IN2}}{C_{IN1}} + t_02 + K_2 \cdot \frac{C_{\text{load}}}{C_{IN2}} \]  
(10)

In this inequality \(C_{\text{load}}\) and \(C_{IN1}\) are well identified, \(C_{IN2}\) is the input capacitance of the buffer to be selected. A direct determination of the optimal buffer \((C_{IN2})\) is obtained from the cancellation of the derivative, of the right part of equation 10, with respect to \(C_{IN2}\).

For real structures, evaluated on full custom cells, \(t_02/K_1\) can be easily evaluated from eq.14 as:

\[ \frac{t_02}{K_1} = \frac{C_{\text{par2}}}{C_{IN2}} = \frac{C_{\text{par2}}}{C_{IN1}} \sqrt{T_{\text{limit}}} \]  
(14)
showing clearly up design parameters. The buffer insertion condition is then obtained from:

\[ T_{\text{limit}} - 2V_{\text{t}}^{\text{IIIu}} - \frac{C_{\text{par}2}}{C_{\text{IN}1}} \sqrt{\frac{I}{V_{\text{t}}^{\text{IIIu}}}} \geq 0 \]  

(15)

Note that considering average value of delays, this result stands for individual inverters with internal configuration ratios balancing or not fall and rise delay times. From layout modeling considerations it can be easily shown that the ratio \( \frac{C_{\text{par}2}}{C_{\text{IN}1}} \) is nearly constant, for transistor widths greater than approximatively 10 times the minimum available length [23,24]. Table 1 summarizes the values of Tlimit deduced, from equ.14, for different parasitic content values.

<table>
<thead>
<tr>
<th>C_{\text{par}2}/C_{\text{IN}1}</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_{\text{limit}}</td>
<td>4.9</td>
<td>5.6</td>
<td>6.8</td>
<td>9.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Buffer insertion limit of single inverter for different output parasitic capacitive content.

As shown, T_{\text{limit}} value varies from 4 (which corresponds to the result previously obtained by Mead [3]), for an ideal structure or with negligible parasitic content, to 10 for a poorly designed inverter. \( \frac{C_{\text{par}2}}{C_{\text{IN}1}} = 4 \) is a typical value of 1\( \mu \)m standard cell library. This limit for buffer insertion, that we defined here, corresponds to the maximum acceptable load on an inverter implementing the lowest delay configuration. This structure appears then as the best candidate to be selected for a minimum power-delay product implementation.

6 - Design for Minimum Power Delay Product

In this part we define an explicit expression for the power-delay product of inverters and compare design solutions, for the minimum of this product, to the alternatives investigated in the preceding parts. We show that delays in CMOS structures depend on the ratio of the total load to the drive capacitance of each cell, and that this ratio could be expressed (\( \theta \)) as a fractional part of a reference capacitance, \( (C_{\text{REF}}) \) used as a measure unit for the load. In the same way, we expressed in eq. 4, the total dynamic power as the product of a reference power, \( P_{\text{REF}} = f.V_{\text{DD}}^2.C_{\text{REF}} \), by the total load capacitance normalized with respect to the reference capacitance \( C_{\text{REF}} \), as:

\[ P_{\text{Dyn}} = P_{\text{REF}} \frac{C_{\text{Tot}}}{C_{\text{REF}}} \]  

(16)

For an array of 2 inverters this results in:

\[ P_{\text{REF}} = \left[ \frac{C_{\text{load}} + C_{\text{par}2}}{C_{\text{REF}}} \cdot \frac{(1+k)C_{\text{IN}2} + C_{\text{par}1}}{C_{\text{REF}}} \cdot \frac{(1+k)C_{\text{N}2}}{C_{\text{REF}}} \right] \]  

(17)

The power-delay product can then be obtained as:

\[ \sum_{\theta} P_{\text{Dyn}} = \frac{P_{\text{REF}}} {1+k} \left\{ \frac{(1+k)C_{\text{IN}1}}{C_{\text{REF}}} \cdot \frac{C_{\text{load}} + C_{\text{par}2}} {C_{\text{REF}}} \cdot \frac{C_{\text{par}1}} {C_{\text{REF}}} \cdot \frac{C_{\text{load}} + C_{\text{par}2}} {C_{\text{REF}}} \right\} \]  

(18)

where \( C_{\text{IN}1} = (1+k)C_{\text{IN}1} \) defines the reference capacitance. Derivative of this expression with respect to \( C_{\text{IN}2} \) gives the optimum tapering factor minimizing the power-delay product. However for this equation no closed form solution is available.

<table>
<thead>
<tr>
<th>C_{\text{par}1}/C_{\text{par}2}</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_{\text{IN}2} min \theta</td>
<td>2.6</td>
<td>2.7</td>
<td>2.8</td>
<td>3</td>
<td>3.2</td>
</tr>
<tr>
<td>P</td>
<td>3.8</td>
<td>3.7</td>
<td>3.6</td>
<td>3.3</td>
<td>3.1</td>
</tr>
<tr>
<td>T_{\text{limit}}</td>
<td>0.45</td>
<td>0.54</td>
<td>0.63</td>
<td>0.82</td>
<td>0.99</td>
</tr>
<tr>
<td>P.\theta</td>
<td>1.2</td>
<td>1.5</td>
<td>2.3</td>
<td>3.6</td>
<td>5.2</td>
</tr>
<tr>
<td>C_{\text{IN}2} T_{\text{limit}}</td>
<td>2.5</td>
<td>2.1</td>
<td>1.8</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>T_{\text{limit}}</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>P</td>
<td>46</td>
<td>60</td>
<td>74</td>
<td>105</td>
<td>137</td>
</tr>
<tr>
<td>P.\theta</td>
<td>1.2</td>
<td>1.6</td>
<td>2.5</td>
<td>4.3</td>
<td>6.7</td>
</tr>
</tbody>
</table>

Table 2: Comparison of the power delay product minima obtained for a two stage array, to the values obtained by sizing the 2nd stage for \( \text{min} \) or at \( T_{\text{limit}} \).

For illustration we represent in Table 2 the variation, with \( C_{\text{IN}2} \), of the power-delay product (equ.18) for a configuration of 2 real inverters. The load has been imposed equal to 10 \( C_{\text{IN}1} \) and the configuration ratio \( k=1 \). As shown, depending on the parasitic content of the array implementation, tapering factors for minimum power delay implementations exhibit a small variation with respect to the parasitic load. Alternatives, with explicit solutions, allow interesting trade offs between speed and power with a few penalty on the power delay product. As shown, an explicit sizing at \( T_{\text{limit}} \) constitutes a good solution for low power implementations.

7 - Application to Standard Cells

From the previously shown results it appears that if the inverter is the most efficient gate in terms of load derating factor and fanout limit, the cost of each gate selection can be evaluated, on line, in terms of transistor size and fanout factor compared to its limit value (\( T_{\text{limit}} \)).

<table>
<thead>
<tr>
<th>RISE (ns)</th>
<th>FALL (ns)</th>
<th>T_{\text{limit}}</th>
<th>C_{\text{IN}1} (pF)</th>
<th>C_{\text{IN}2} (pF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV1</td>
<td>0.29+ 2.95</td>
<td>0.29+ 2.68</td>
<td>5.9</td>
<td>0.34</td>
</tr>
<tr>
<td>INV2</td>
<td>0.22+ 1.44</td>
<td>0.20+ 1.21</td>
<td>5.9</td>
<td>0.65</td>
</tr>
<tr>
<td>INV3</td>
<td>0.25+ 0.95</td>
<td>0.22+ 0.76</td>
<td>5.9</td>
<td>0.96</td>
</tr>
<tr>
<td>INV4</td>
<td>0.21+ 0.74</td>
<td>0.19+ 0.63</td>
<td>5.9</td>
<td>1.20</td>
</tr>
<tr>
<td>INV5</td>
<td>0.22+ 0.59</td>
<td>0.21+ 0.50</td>
<td>5.9</td>
<td>1.50</td>
</tr>
</tbody>
</table>

Table 3: Performance equations of a 1\( \mu \)m standard cell library.

We applied these results to the different inverter cells of an industrial 1\( \mu \)m library, which performance equations are given in Table 3 together with the values of T_{limit}, the maximum load allowed and the input capacitance of each template. These values constitute the first determination, at the cell level, of real buffer insertion limits allowing, by a direct inspection of loading factors on the output nodes, the definition of speed up or reconfiguration strategies as discussed later. In Table 4 we
compare delay, power and power-delay product values for different selection of inverters necessary to feed different loads. Different selection of design choices are investigated: implementation with the optimal number of stages (equ.5), minimum number of stages defined at the minimum delay with the condition T<T_{limit}, with regular tapering factor values, and selection of templates at T_{limit} resulting in irregular tapering factor values.

As shown in the Table the last condition results always in minimum power-delay product with very low penalty in delay. Sizing loaded stage at T_{limit} always results in non uniform tapering factors with a lower P\_q product. Note that due to the limited choice in the library (discrete variation of the drive capabilities of the cell), some alternatives are equivalent for particular loading conditions.

<table>
<thead>
<tr>
<th>Load (pF)</th>
<th>Design strat</th>
<th>Stage nbr</th>
<th>Cell selection</th>
<th>Delay r.unit</th>
<th>P r.unit</th>
<th>p _q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9</td>
<td>1 3</td>
<td>2INV5+INV2 INV3 INV1</td>
<td>2.1</td>
<td>2.74</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 2</td>
<td>2INV3 INV1</td>
<td>2.19</td>
<td>2.28</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 2</td>
<td>2INV3 INV1</td>
<td>2.19</td>
<td>2.28</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1 3</td>
<td>2INV5 INV3 INV1</td>
<td>2.03</td>
<td>2.22</td>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 2</td>
<td>INV5+INV1 INV1</td>
<td>2</td>
<td>1.87</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 2</td>
<td>INV5 INV1</td>
<td>2</td>
<td>1.51</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1 3</td>
<td>INV3+INV4 INV3 INV1</td>
<td>1.8</td>
<td>1.59</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 2</td>
<td>INV5 INV1</td>
<td>1.73</td>
<td>1.31</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 1</td>
<td>INV3 INV1</td>
<td>1.74</td>
<td>1.22</td>
<td>2.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Comparison of design alternatives for a low power-delay product implementation: 1 is the solution with optimal number of stages, 2 is the minimum delay implementation and 3 the selection at T_{limit}.

8 - Conclusion

Design and sizing of tapered buffers has been considered for minimizing the power-delay product. Rules for buffer insertion have been defined as a function of physical design parameters. They have been shown to constitute a good alternative for efficient use of inverter templates as well as for searching heavily loaded nodes to apply speed up alternatives.

From a general definition of the power-delay product it has been shown that sizing at the buffer insertion limit constitutes an efficient solution for buffer implementation with a minimum power-delay product. Moreover this allows closed form solutions of this problem, which can be completely determined from the performance characterization laws of library cells.

Application to a 1\mu m standard cell library has been given. Comparisons of implementations realized for an optimum buffer stage number, a minimum delay and a selection of templates at T_{limit} gives evidence of the benefit of this last solution for low power design.

References