On Buffering Schemes for Long Multi-Layer Nets

Vani Prasad        Madhav P. Desai
Department of Electrical Engineering,
Indian Institute of Technology, Bombay,
Mumbai, India.
vani,madhav@ee.iitb.ac.in

Abstract

We consider the problem of minimizing the delay in signal transmission over point-to-point connections across multiple metal layers in a VLSI circuit. We present an exact solution for the two layer case. This exact solution, however, is irregular and dependent on the length of the net. We look for approximate solutions which are not only regular but are independent of the length of the net. We show that two of these approximate solutions yield delays that are within a constant of the optimal solution. We have seen that our results hold true even for the three-layer case. We conjecture that our models can be inductively extended for multi-layer nets as well.

1. Introduction

As we move into the deep sub-micron era, interconnects tend to limit system performance. This highlights the need to tackle the problem of interconnects early in the design process. Of the issues to be considered while planning the wire, we take up delay minimization, as this would give us the basic model for a wire before moving on to other problems. We therefore need to find a solution to this critical problem of interconnect delay minimization while controlling the cost/area/energy.

Multi-level routing schemes are gaining importance as technology scales, specially with the advent of the system-on-chip (SoC) era. Large logic modules (IPs) in SoCs need to be connected together using long global wires. Researchers are therefore trying to come up with good multi-level architectures. Venkatesan et al used a stochastic wiring distribution to find the wire-lengths to be run on a given metal layer [1]. Cong et al have proposed a multi-level advanced system [2] for routing global and local nets hierarchically. They essentially adopt resource reservation by setting aside some area for the nets and no explicit wire model has been developed. But none of the previous works have considered the delay optimization problem for wires which span multiple metal layers.

In this paper we consider the two-layer net problem and find the best delay solution. We then attempt to arrive at regular structures for point-to-point connections on a two layer net using a spliced solution. We show that these structures closely approximate the best solution in terms of delay as well as energy overhead and provide simple routing models for the wires. We will show in the subsequent sections, how the predictable spliced solution to the delay minimization problem for two layers, can be extended inductively for multiple layers as well. These simple models would aid us in an easier development of CAD tools for back-end design.

We describe the problem definition and some preliminaries in Section 2. We present the experiments leading us to the various buffering strategies for wire delay minimization in Section 3. Finally, we give the results and conclusions in Section 4 and 5, respectively.

2. Problem Formulation

Given the information regarding the input capacitance of the buffered wire, the lengths of the two nets and the sink capacitance, our aim is the minimization of the signal propagation time between the source and the sink. The problem is depicted in Figure 1(a), which shows the existence of a discontinuity (via) as we move from one metal layer to the other. Given two points A and B on two different metal layers, our aim is to achieve signal transmission between them in the shortest possible time.

The problem can be modeled as shown in Figure 1(b), wherein \( C_{\text{in}} \) is the input capacitance as seen looking into the wire and \( C_{\text{sink}} \) is the load capacitance. The two nets run for distances \( L_1 \) and \( L_2 \) over different metal layers. For this problem, we will first find the best solution using buffer insertion and uniform wire sizing. The problem can be defined as follows:

given the input capacitance, a sink load and distances \( L_1 \)
3. Experiments

We start off by considering two nets over the same metal layer as well as on different layers. All our initial experiments therefore correspond to cases where the wire parameters (resistance and capacitance per unit length) vary by a factor of 5 for the second (descendant) net. We do this in order to get an idea about the worst case performance. We consider the wire parameter variations across the two layers as \( \frac{R_2}{R_1} = 5, \frac{C_2}{C_1} = 5 \) and so on. The wire-lengths that we chose for the experiments are greater than the critical length and range from 1mm-20mm. We have not considered the effect of via resistance and capacitance on the delay function. All the optimizations were performed in MATLAB.

3.1. Unconstrained Optimization

In order to get the optimal delay for a signal traversing through a single-via wire, we performed unconstrained delay optimization with \( (3(k + s) - 1) \) variables, where \( k \) is the total number of sections in the first net and \( s \) is that in the descendant net. Here we kept the section lengths, buffer widths and the wire widths variable. Figure 2 shows the unconstrained optimization scheme.

Equation 1 gives the function to be optimized \((unc_{2_{net}})\) and corresponds to the delay of a buffered wire of length \( L = L_x + L_y \), driven by a minimum sized driver and with a sink capacitance \( C_s1 \).

\[
unc_{2_{net}} = D_1 + D_2 + D_3
\]  

\[
D_1 = a_0x_1w_{inf}x_1 + a_1w_{x_1} + a_2x_1^2 + \frac{a_3x_1w_{x_1}}{w_{inf}x_1}
\]
Figure 4, has three stages: the first and the last comprising

$D_2 = a_0 \frac{l_{x,k} w_{in1xk}}{w_{x,k-1}} + a_2 l_{x,k}^2 + a_1 \frac{w_{y1}}{w_{y1-1}} + a_4 \frac{l_{y1} w_{in1yk}}{w_{in1yk}}$

$+ b_0 \frac{l_{y1} w_{inty1}}{w_{y1-1}} + b_2 l_{y1}^2 + b_1 \frac{w_{y1}}{w_{inty1}}$

where $a_i$ and $b_i$ are buffer and wire parameters for the first and the second net respectively. The discontinuity point between the two nets was chosen as the origin. We term the small stubs from the individual nets across the discontinuity as a splice. In the above equations, $D_1$ and $D_3$ denote the two section delays and the splice delay is given by $D_2$. The delay equation, formulated using the Elmore delay model, is a posynomial in the buffer and wire widths. It is known that a convex optimization done on a posynomial would ensure that we arrive at the global minimum. This optimization procedure was also used in [5] to obtain an exact solution to the transistor sizing problem. By using this scheme, we found the optimal size of the buffers and the wire size such that the delay was minimized. We then used the trends that we obtained for the buffer and wire widths of the two nets (Figures 3(a), 3(b) respectively), to perform local optimization on the individual nets across the splice. The trends correspond to the case where the nets run from Metal1 to Metal2.

Based on these trends, we explored various buffering schemes suitable for different combinations of the wire parameters. For the unconstrained case the best solution is hard to model. Hence we performed local optimizations on the individual nets to arrive at regular structures which could approximate the reference solution and are easier to model.

3.2. The 2-p-m-p Architecture

Based on our results for the unconstrained optimization, we performed local optimization using the predictable pre-mid-post(p-m-p) buffer strategy [4] for a point-to-point connection. This architecture for a buffered wire, as shown in Figure 4, has three stages: the first and the last comprising of cascaded buffers and an intermediate section of equally spaced buffers with a fixed wire width.

The source is a minimum sized buffer which is boosted up by a factor of 4 to drive the buffer at the middle stage. The post-buffer stage again begins with a minimum sized buffer. The idea, here, is to decouple the sink capacitance from the wire. Hence the wire delay can be expressed as a function of the wire-length and the load capacitance, making the wire performance predictable.

Perhaps the simplest wire architecture possible using the p-m-p solution, is as shown in Figure 5. In this case, both the nets can be independently optimized for delay, as single p-m-p buffered wires. The prebuffer stage would depend on the first net. The post-buffer stage of this net, would in turn be dependent on the size of the intermediate buffer of the second net. Hence there is a common section of cascaded buffers between the two nets, at the discontinuity point. Finally, at the output end there is an independent section of post-buffers driving the total sink capacitance, which could be the contribution from the various nets branching out from this point.

The number of pre-stage buffers can be found by the following expression:
3.3. The Spliced Architectures

As mentioned earlier, we consider various combinations of the wire parameters across both the nets, which affect the buffer and wire sizes that optimize the total delay. We investigated the effect of making the stubs at the splice as variables and came up with the following architectures for the two-layer nets, which approximate the unconstrained optimization results.

1. Scheme 1

Here we consider the case where the contribution from the first net to the splice would be nil, as shown in Figure 6. Hence the intermediate buffer from the first net would drive the first section of the next net directly.

2. Scheme 2

In this case the intermediate buffer of the first net drives its own wire section which is connected to the first mid-buffer of the second net. The architecture is shown in Figure 7.

3. Scheme 3

Here the intermediate buffer from the first net directly powers up to drive the mid-buffer of the next net, through a cascaded configuration of buffers, which we term as translational buffers, as shown in Figure 8. The number of such translational buffers would depend on the buffer widths in the first as well as the second net. Such a case could arise when the buffer size of the descendant net is much larger than that of the first net. This would be evident for the cases where we move from a lower to a higher capacitance region.

4. Results and Discussion

In this section, we compare the optimized delay for each of the wire architectures described above, for all possible combinations of the wire parameters. Our objective is to decide the best topology for these cases among the 2-p-m-p and the spliced structures.

4.1. Results for 0.18μm technology

We compared the 2-p-m-p and the spliced solutions with the unconstrained reference solution, to finalize the wire ar-

---

**_normalized text**

3.3. The Spliced Architectures

As mentioned earlier, we consider various combinations of the wire parameters across both the nets, which affect the buffer and wire sizes that optimize the total delay. We investigated the effect of making the stubs at the splice as variables and came up with the following architectures for the two-layer nets, which approximate the unconstrained optimization results.

1. Scheme 1

Here we consider the case where the contribution from the first net to the splice would be nil, as shown in Figure 6. Hence the intermediate buffer from the first net would drive the first section of the next net directly.

2. Scheme 2

In this case the intermediate buffer of the first net drives its own wire section which is connected to the first mid-buffer of the second net. The architecture is shown in Figure 7.

3. Scheme 3

Here the intermediate buffer from the first net directly powers up to drive the mid-buffer of the next net, through a cascaded configuration of buffers, which we term as translational buffers, as shown in Figure 8. The number of such translational buffers would depend on the buffer widths in the first as well as the second net. Such a case could arise when the buffer size of the descendant net is much larger than that of the first net. This would be evident for the cases where we move from a lower to a higher capacitance region.

4. Results and Discussion

In this section, we compare the optimized delay for each of the wire architectures described above, for all possible combinations of the wire parameters. Our objective is to decide the best topology for these cases among the 2-p-m-p and the spliced structures.

4.1. Results for 0.18μm technology

We compared the 2-p-m-p and the spliced solutions with the unconstrained reference solution, to finalize the wire ar-
Architecture for the various metal layers in 0.18 \mu m technology. We suggest that among the different buffering techniques explored, Scheme 2 be used for all the long nets running from Metal4, Metal5 and Metal6 to Metal1. This kind of metal-run sees a transition from a lower to a higher resistance and capacitance region. For the rest of the combinations, Scheme 1 gives a near optimal solution.

In Table 1 and Table 2, we have listed the total delay from our simplified p-m-p spliced architecture (Scheme1-delay) as well as that from an unconstrained optimization for M1 to M2 and M4 to M1 metal transition. The buffer sizes are expressed relative to the minimum sized driver and the wire sizes relative to the minimum wire width corresponding to a given metal layer. An even simplified architecture would have a fixed buffer and wire width for a given set of lengths on a particular metal layer. The first column of the table gives the individual wire-lengths on the two layers. Our experiments are on a load capacitance of 100FF. We find that our solution for this two-layer delay minimization problem is within 5% of the optimal one for the M1-M2 case and within 10% for M4-M1. As expected, for longer lengths our solution approximates the reference solution better.

For the data in Table 1, the buffer and wire widths for the two nets are \( w_1 = 71; w_{int1} = 1; w_2 = 60; w_{int2} = 1 \)

Table 1. Delay comparison for M1-M2

<table>
<thead>
<tr>
<th>( L_{1,2} ) (mm)</th>
<th>no. of buffers</th>
<th>Scheme1-delay (ns)</th>
<th>unconstrained (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>14</td>
<td>0.686</td>
<td>0.591</td>
</tr>
<tr>
<td>2.4</td>
<td>22</td>
<td>1.0275</td>
<td>0.996</td>
</tr>
<tr>
<td>4.4</td>
<td>29</td>
<td>1.321</td>
<td>1.318</td>
</tr>
<tr>
<td>4.6</td>
<td>34</td>
<td>1.5247</td>
<td>1.513</td>
</tr>
<tr>
<td>6.8</td>
<td>45</td>
<td>1.988</td>
<td>1.961</td>
</tr>
</tbody>
</table>

For the data in Table 2, \( w_1 = 40; w_{int1} = 6.5; w_2 = 40; w_{int2} = 1.3 \)

We have arrived at a simple delay model for a buffered wire, the delay being a function of the wire-lengths and the sink capacitance. If the wire-lengths are in \( \mu m \), the interconnect delay for M1-M2 in (ns) can be modeled as shown in Equation 7.

\[
D_{\text{m1m2}} = 0.14 \times 10^{-3} L_1 + 0.095 \times 10^{-3} L_2 + 0.081 \log(C_s)
\]

(7)

We find from Figure 9 that the delay from the approximate solution is within a constant (1 gate delay) of the optimal solution. This is the penalty incurred across the splice. We expect that for each of the approximate schemes, \( D_{\text{approx}}(L_1, L_2) = D_{\text{optimal}}(L_1, L_2) + \Delta_{1,2} \)

where \( \Delta_{1,2} \) is the splice penalty. We observe empirically, that this indeed is the case. Hence the gap between the optimal solution for M1-M2 in (ns) can be modeled as shown in Equation 7.
The interconnect delay for M4-M1 in \((n/\Delta_2/\Delta_7)\) can be modeled as shown in Equation 8.

\[
D_{\text{approx}}(L_1, \ldots, L_n) = D_{\text{optimal}}(L_1, \ldots, L_n) + \sum_{i=1}^{n-1} \Delta_{i,i+1}
\]

where \(i\) is any metal layer and \(n\) is the total number of layers. \(\Delta_{i,i+1}\) denotes the splice penalty. We have found that \(\Delta_{i,i+1} \leq 40\ \text{ps}\), which is of the order of an inverter delay in 0.18\(\mu\text{m}\) technology. The next step would be to find the sub-optimality of the spliced solution for nets spanning larger (> 3) metal layers.

5. Conclusions

In this paper we have studied the delay minimization problem for a two-layer, 2-pin net. We studied the best solution for this and also explored regular structures for the two-layer nets. We have shown that despite the single discontinuity that is encountered while a net transits from one metal layer to another, our solution is close to the optimal one.

We conjecture that we can model the multi-layer nets by inductively extending the solution for two-layer nets. Hence we can come up with a predictable model for the interconnects routed on various metal layers. This simple interconnect model could then be used in routers in physical design tools and thus aid us during the design process.

References


