A Technique for Selecting CMOS Transistor Orders

Ting-Wei Chiang, C Y Roger Chen, and Wei-Yu Chen
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY 13244, USA

Abstract

Transistor reordering has been known to be effective in reducing delays of a circuit with nearly zero penalties. However, techniques to determine good transistor orders have not been proposed in literature. Previous work on this has to resort to running SPICE for all meaningful transistor orders and selecting a best one, which is extremely time-consuming. This paper proposes an efficient and accurate technique for determining best transistor orders without running SPICE simulations. Experimental results from SPICE3 show that the predictions are very accurate.

1. Introduction

Transistor sizing [1] has been a widely used technique for optimizing circuit performance. It is known for its associated side effects and penalties due to possible shifts of critical paths and changes of layout and loads. Transistor reordering was investigated [2][3] as a pre-step to sizing. It is reported that transistor reordering can improve circuit performance of a circuit by as much as 20% with nearly zero penalties or side effects. Conventional wisdom tells us that the transistor with a slowest arriving signal (i.e., critical input) should always be placed near the output [4], which has been shown to be incorrect [2][3]. Figs. 1(a) and 1(b) (from [2]) show two different transistor orders of a 4-input NAND gate, that is, two circuits which have an identical logical function, but different transistor orders. Fig. 1(a) is referred to as top-critical case due to that critical input I is connected to the top transistor, while Fig. 1(b) is referred to as bottom-critical case. With an input rise time of 5 ns, the output 1-to-0 propagation delay for the top-critical case is 23% longer than that of the bottom-critical case. When input rise time = 1 ns, the propagation delay for the top-critical case is 8% shorter than that of bottom-critical case. A major contribution of [2][3] is to point out the impact of transistor reordering. However, no techniques are proposed to determine a good transistor order. In [2][3], a good transistor order is obtained by running SPICE for all meaningful transistor orders and selecting the best one. Extensive SPICE simulations have to be performed on every gate of a circuit, which introduces very significant computational overhead.

2. Transistor Reordering Investigation

The input rise or input fall time (both are also generally referred to as input transition time) is denoted by the variable \( d \). SPICE simulations have shown that optimal transistor orders depend on \( d \), load capacitance, transistor sizes, and intrinsic capacitance. In this paper the focus of our discussions is to determine a good transistor order with respect to the value \( d \) for a gate whose load capacitance, transistor sizes, and intrinsic capacitance are given. Transistor reordering with respect to physical oxide thickness was discussed in [5]. However, impact due to input transition time \( d \) was not addressed, which is the primary focus of this paper. Consider a scenario where the output has a 1-to-0 transition as shown in Fig. 2(a). When the critical input \( I \) is low in the beginning, the capacitors \( C_1, C_2, C_3 \), and \( C_4 \) are charged until \( I \) switches to high. \( C_1, C_2, C_3 \), and \( C_4 \) are full initially. Fig. 2(b) shows that the four fully charged capacitors are discharging when \( I \) switches from low to high. In Fig. 3(a), only \( C_1 \) is charged when the input is high in the beginning. In Fig. 3(b), \( C_1 \) is charged when the input is low in the beginning. The capacitor \( C_1 \) is discharging through four resistors when the critical input \( I \) arrives as shown in Fig. 3(b). In Fig. 4(a), four capacitors are discharging to ground when the input \( I \) is high in the beginning. In Fig. 4(b), \( N_4 \) is turned off by the arriving critical input \( I \), and power source charges...
all empty capacitors at this time. In Fig. 5(a), all capacitors are discharging to ground in the beginning. When the critical signal $I$ arrives, only capacitor $C_1$ can be charged by the power source as shown in Fig. 5(b). SPICE3 simulation results of the cases of 1-to-0 and 0-to-1 transitions at output are obtained by varying $d$ while fixing all other parameters such as width and length of the transistor, and load capacitance. To more effectively observe the effect of $d$, the delay of the simulation is measured from 0% input waveform to 50% output waveform.\footnote{1} In Figs. 6(a) and 6(b), the two critical lines are plotted according to output delay data obtained for a wide range of $d$’s in both critical transistor orders. The cross point, $Q$, formed by crossing the two lines, determines a best transistor order of the gate. In this example, we can easily observe that the bottom critical should be chosen when

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Output 1-to-0 transition of bottom-critical case}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{Output 1-to-0 transition of top-critical case}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig4.png}
\caption{Output 0-to-1 transition of bottom-critical case}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig5.png}
\caption{Output 0-to-1 transition of top-critical case}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6.png}
\caption{(a) fall time by different input delay (b) rise time by different input delay}
\end{figure}

It is clear that 0-50% delay is equal to 50%-50% delay + $d/2$, and it does not matter which of the two is used for optimizing gate delays because $d/2$ is constant. The value of 50%-50% delay can be negative in some cases when there is smaller load capacitance with large input $d$. To normalize all delays to positive values, we have chosen 0-50% delay in our discussions.
is larger than point \( Q \), and the top critical should be chosen when \( d \) is smaller than \( Q \). However, in the output 0-to-1 transition, the delay of bottom critical is always longer and the two lines will not cross. By combining both situations, it becomes clear that the best way to determine the optimal transistor order of the gate is to find the point \( Q \), which is the key to the selection of optimal transistor orders.

3. Transistor Reordering Equivalent (TRE) Circuits

If we use Elmore delay model to represent the transistors of gates, as was done in a lot of previous work in transistor sizing (e.g., [6][7][8]), and apply it to each case in Figs. 2(b), 3(b), 4(b), and 5(b). We can easily calculate the delay for each case. However, the use of Elmore delay model will always conclude that the top critical case has better performance on the output 1-to-0 transition case. In other words, Elmore delay model is unable to explain effects of transistor reordering and its relation with input transition time \( d \). Thus, a new model needs to be developed.

Because output delay depends heavily on \( d \), an AC analysis method is used for the input source. An AC equivalent circuit is depicted in Fig. 7. The dependent source current is \( g_m \times V_{gs} \), where \( g_m \) is proportional to \( W/L \) of the MOS transistor. We also know that current going through the capacitor is \( C \frac{dV_{out}}{dt} \). Since our objective is to develop an equivalent switch-level circuit, through which we can determine a best transistor order, we can simplify our model such that when input begins to move up, gate output also begins to move down. Clearly, this might create a small discrepancy in calculating actual gate delay. However, we have found, through experiments, that as long as we apply them consistently in all cases, this assumption causes virtually no inaccuracy in predicting optimal transistor orders. The time between 0% input and 50% output (i.e., \( t \) in Fig. 7) can then be derived as follows:

\[
C \frac{dV_{out}}{dt} = g_m \times V_{gs} = -g_m \times V_m
\]

\[
-C \int_{v_{out}}^{V_{out}} dV_{out} = \int_0^t V_m \times dt
\]

\[
\frac{C}{g_m} \times \frac{V_{dd}}{2} = t \times \frac{t}{d} \times V_{dd} \times \frac{1}{2}
\]

\[
t^2 = \frac{C}{g_m} \times d
\]

The on-resistance \( R \) is given by \( \frac{V_{ds}}{i_d} \), where

\[
i_d = k_r \frac{w}{L} \left( V_{gs} - V_t \right) v_{ds} - \frac{1}{2} \frac{V_{ds}^2}{l_d}
\]

Let \( g_m = \frac{1}{R} \). When \( v_{ds} \) is very small, the output delay \( t \) is then equal to \( \sqrt{RCd} \). Since we intend to represent transistors as switches, we will assume resistance of all transistors, when turned on, will be equal to \( R \). This assumption can also be explained conceptually. From the viewpoint of DC signals, the \( V_{ds} - i_d \) curve which represents the characteristic of the MOS transistor is fixed; that is, the slope which is approximately \( \frac{1}{R} \) in the linear mode is unchanged. From AC’s point of view, the fixed \( R \) implies that the capacitance associated with the critical input signal carries all effects when critical input goes through the transistor. Following such an approach, we will extract all relevant transistor behaviors and model them as part of the associated capacitors as long as the resulting equivalent switch-level representation achieves our objective (i.e., accurately determining a best transistor order). From the above analysis, the inverter is now modeled as a simple RC circuit as shown in Fig. 7, where the equivalent resistor \( R^* = R \) (i.e., the on-resistor) and equivalent capacitor \( C^* = \sqrt{\frac{C d}{R}} \). Thus, if we let \( R^* = R \), the delay obtained from the proposed model \( t = R^* C^* = R C^* \), where \( C^* \) is the equivalent capacitor. Next, we will show how to extend the concept of equivalent circuits to handle multi-input NAND gates. Consider the 5-input NAND gate shown in Fig. 8(a), where critical input \( I \) is connected to
transistor $N_3$. When signal $I$ arrives, each transistor is now modeled as a simple RC circuit, as shown in Fig. 8(b). Note that for each transistor above the critical input signal (i.e., input to $N_3$ in this case), the equivalent resistor and capacitor are simply the original on-resistor $R$ and intrinsic $C$. Note that for transistors below the critical input signal, the initial intrinsic capacitors are empty, and can be removed from the equivalent circuit. When $N_3$ is just turned on, it will be operating in saturation region, while $N_4$ and $N_5$ will be operating in linear region. Clearly the current flowing through $N_4$ and $N_5$ will be much smaller than that flowing through $N_3$. Thus, the voltage drops at $N_4$ and $N_5$ will be small and can be ignored, and $N_3$'s behavior is similar to the inverter case discussed earlier. Therefore, the equivalent $R^*$ and $C^*$ will still be assigned as $R^* = R$ and $C^* = \sqrt{\frac{C_d}{R}}$, respectively. This is shown in Fig 8(c). For NOR gates, the series-connected transistors will be on the PMOS side, and the above concept can be similarly applied. Simple extensions can be made similarly for complex gates. For the sake of description, each such equivalent circuit is referred to as a Transistor Reordering Equivalent (TRE) circuit, and the process of obtaining a TRE circuit is referred to as TRE transformation.

Fig. 8: Equivalent circuits for multi-input NAND gates.

4. Transistor Order Selection Technique

Throughout this section, a 4-input NAND gate will be used as an example for explaining the proposed technique. Firstly, for brevity, it is assumed that $C_1$ is equal to $C_z$; and $C_2$, $C_3$, and $C_4$ are all equal to $c$.

Case 1: Output 1-to-0 transition:

TRE circuits for the top-critical case and bottom-critical case for the output 1-to-0 transition are shown in Figs. 9(a) and 9(b), respectively. Following the concept of typical approach of delay calculation,

$$t_{top} = \sqrt{\frac{C_d}{R}} \times 4R = 4\sqrt{RC_d} .$$

(3)

$$I_{bottom} = C_z \times 4R + c \times 3R + c \times 2R + \sqrt{\frac{C_d}{R}} \times R$$

$$= 4RC_z + 5RC + \sqrt{4Rcd} .$$

(4)

For the second critical input and the third critical input, $t_{2nd}$ and $t_{3rd}$ can be obtained similarly.

$$t_{2nd} = C_z \times 4R + 3R + \sqrt{\frac{C_d}{R}} \times 2R = 4RC_z + 3RC + 2\sqrt{Rcd}$$

(5)

$$t_{3rd} = C_z \times 4R + c \times 3R + c \times 2R = 4RC_z + 3RC + 2\sqrt{Rcd}$$

(6)

From the calculations, it is clear that when $d$ is small,

$$t_{bottom} > t_{3rd} > t_{2nd} > t_{top} .$$

(7)

As soon as $d$ is large enough,

$$t_{top} > t_{2nd} > t_{3rd} > t_{bottom} .$$

(8)

That is, optimal transistor orders always occur in the top-critical or bottom-critical cases. The bottom critical case performs better when $d$ is large enough. Determining where the cross point $Q$ occurs becomes
the key to determining a best transistor order. The cross point $Q$ can be obtained by making $t_{top} = t_{bottom}$.

$$t_{top} = 4\sqrt{RC_c d} = 4RC_c + 5RC + \sqrt{Rcd} = t_{bottom} \quad (9)$$

$$4\sqrt{RC_c d} - \sqrt{Rcd} = 4RC_c + 5RC \quad (10)$$

Similarly, if the number of transistors connected in series is $n$, we have

$$d = \left(\frac{nRC_c + \left(\frac{n(n-1)}{2} - 1\right)RC}{n\sqrt{RC_c - RC}}\right)^2 = Q_{NAND} \quad (11)$$

Once the cross point $Q$ is obtained, we can easily determine the best transistor order of the gate when output switches from 1 to 0.

**Case 2: Output 0-to-1 transition**

No cross point $Q$ can be obtained in this case because the delay of the top critical $t_{top} = \sqrt{RC_c d}$ is always smaller than that of the bottom critical $t_{bottom} = 3RC + \sqrt{RC_L d}$ regardless of the value of $d$. It can be shown that $t_{bottom} > t_{3rd} > t_{2nd} > t_{top}$.

Note that for NOR gates, the conducting series-connected transistors will be on the PMOS side, and the directions of current flow will be opposite. All the above delays can be similarly calculated.

$$d = \left(\frac{nRC_c + \left(\frac{n(n-1)}{2} - 1\right)RC}{n\sqrt{RC_c - RC}}\right)^2 = Q_{NAND} \quad (12)$$

The above analyses lead to the following theorems. Proofs are omitted due to limited space.

**Theorem 1:** Under TRE transformations, an optimal transistor order is either a top-critical transistor order or a bottom-critical transistor order.

**Theorem 2:** Under TRE transformations, (1) for output 0-to-1 transition of NAND gates, top-critical transistor order is optimal; (2) for output 1-to-0 transition of NAND gates, bottom-critical transistor order is optimal; (3) for output 0-to-1 transition of NOR gates, when $d < Q_{NAND}$, bottom-critical transistor order is optimal, and when $d > Q_{NAND}$, top-critical transistor order is optimal.

**Theorem 3:** Under TRE transformations, (1) for output 1-to-0 transition of NOR gates, bottom-critical transistor order is optimal; (2) for output 0-to-1 transition of NOR gates, when $d < Q_{NOR}$, bottom-critical transistor order is optimal, and when $d > Q_{NOR}$, top-critical transistor order is optimal.

Theorems 2 and 3 lead directly to the proposed technique. For complex gates, when conducting series-connected transistors occur on NMOS (PMOS) side, we can model it as NAND (NOR). Extensive SPICE simulations have been performed, and the results very much verify Theorems 1-3.

### 5. Experimental Results

$Q$ and $Q_{spice}$ are the cross points obtained by our selection technique and that observed from the results of the SPICE 3 simulations, respectively. Table 1 provides snapshots of $Q$ and $Q_{spice}$ for a 4-input NAND gate under a wide range of load capacitance. $d$ ranges from 3.58e-10 sec (which is obtained from the case when a small inverter is driven by another small inverter) to 2.15e-9 sec (which is obtained from the case when an inverter drives 10 other inverters). The experimental result of output 1-to-0 transition show that success rate of the proposed predication technique is higher than 98.02%. The success rate of the proposed selection technique of the output 0-to-1 transition is 100%. The results of output 1-to-0 transition for another 4-input NAND gates are shown in Table 2. A “√” stands for a correct selection, and an “X” stands for a wrong one. From these 11 different input transition time cases, the accuracy of our prediction for output 1-to-0 transition is 95.12 %. For the case of output 0-to-1 transition, the selection accuracy is 100%. This average improvement of propagation delay is 18.28% as shown in Table 3.

<table>
<thead>
<tr>
<th>$C_{LOAD}$</th>
<th>0.02pF</th>
<th>0.1pF</th>
<th>0.2pF</th>
<th>1pF</th>
<th>2pF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02pF</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1pF</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2pF</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1pF</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2pF</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$d$</th>
<th>$Q_{spice}$</th>
<th>$Q_{spice}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.58e-10</td>
<td>1.69e-10</td>
<td>3.88e-10</td>
</tr>
<tr>
<td>5e-10</td>
<td>1.22e-10</td>
<td>2.38e-10</td>
</tr>
<tr>
<td>7.5e-10</td>
<td>1.76e-10</td>
<td>3.86e-10</td>
</tr>
<tr>
<td>1.0e-9</td>
<td>2.68e-10</td>
<td>6.32e-10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$d$</th>
<th>$Q_{spice}$</th>
<th>$Q_{spice}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.58e-10</td>
<td>1.69e-10</td>
<td>3.88e-10</td>
</tr>
<tr>
<td>5e-10</td>
<td>1.22e-10</td>
<td>2.38e-10</td>
</tr>
<tr>
<td>7.5e-10</td>
<td>1.76e-10</td>
<td>3.86e-10</td>
</tr>
<tr>
<td>1.0e-9</td>
<td>2.68e-10</td>
<td>6.32e-10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$d$</th>
<th>$Q_{spice}$</th>
<th>$Q_{spice}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.58e-10</td>
<td>1.69e-10</td>
<td>3.88e-10</td>
</tr>
<tr>
<td>5e-10</td>
<td>1.22e-10</td>
<td>2.38e-10</td>
</tr>
<tr>
<td>7.5e-10</td>
<td>1.76e-10</td>
<td>3.86e-10</td>
</tr>
<tr>
<td>1.0e-9</td>
<td>2.68e-10</td>
<td>6.32e-10</td>
</tr>
</tbody>
</table>
Table 3: The percentage of the improvement for propagation delay

<table>
<thead>
<tr>
<th>(d_T - d_L)</th>
<th>0.02pF</th>
<th>0.1pF</th>
<th>0.2pF</th>
<th>1pF</th>
<th>2pF</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5e-10</td>
<td>2.2%</td>
<td>4.2%</td>
<td>1.4%</td>
<td>1.4%</td>
<td>1.2%</td>
</tr>
<tr>
<td>5.0e-10</td>
<td>2.5%</td>
<td>2%</td>
<td>2.5%</td>
<td>0.7%</td>
<td>0.4%</td>
</tr>
<tr>
<td>7.5e-10</td>
<td>18%</td>
<td>12.4%</td>
<td>7.1%</td>
<td>3.1%</td>
<td>1.4%</td>
</tr>
<tr>
<td>1.0e-9</td>
<td>35%</td>
<td>23%</td>
<td>11.3%</td>
<td>3.5%</td>
<td>1.2%</td>
</tr>
<tr>
<td>1.25 e-9</td>
<td>24.9%</td>
<td>36.1%</td>
<td>19.3%</td>
<td>5%</td>
<td>2.8%</td>
</tr>
<tr>
<td>1.5 e-9</td>
<td>35.9%</td>
<td>38.6%</td>
<td>25.2%</td>
<td>5.8%</td>
<td>4.4%</td>
</tr>
<tr>
<td>1.75 e-9</td>
<td>47.3%</td>
<td>37.9%</td>
<td>28.8%</td>
<td>9.3%</td>
<td>5.5%</td>
</tr>
<tr>
<td>2.0 e-9</td>
<td>62%</td>
<td>38.4%</td>
<td>31.2%</td>
<td>9%</td>
<td>5.8%</td>
</tr>
<tr>
<td>2.15 e-9</td>
<td>48.8%</td>
<td>55%</td>
<td>35.1%</td>
<td>13%</td>
<td>6.5%</td>
</tr>
<tr>
<td>2.25 e-9</td>
<td>72.6%</td>
<td>42.1%</td>
<td>35.3%</td>
<td>11.8%</td>
<td>7.6%</td>
</tr>
<tr>
<td>2.5 e-9</td>
<td>88.4%</td>
<td>41.4%</td>
<td>43.1%</td>
<td>14.7%</td>
<td>10%</td>
</tr>
<tr>
<td>average</td>
<td>32.6%</td>
<td>25.7%</td>
<td>21.8%</td>
<td>7%</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

Prediction of a 4-input NOR gate gives the accurate rate 92% when the same input transition region and parameters are applied. Complex gates such as AB+C and (A+B) C also give the accuracy rate of 98.4% and 100%, respectively. Another circuit shown in Fig. 10 is also tested. Fig. 10(a) shows the connections between gates in such a way that critical inputs always connect near the output of a gate. By using the proposed technique the circuit in Fig. 10(b) is obtained. From simulation results of SPICE3, it indicates that the arrival time at node (6) (i.e., the overall path delay from the primary input to the primary output in the circuit) in Fig. 10(a) is 4.38×10^{-9} s, and the arrival time at node (6) in Fig. 10(b) is 3.97×10^{-9} s. The total delay for this case is improved by 10.3%. Experiments were also performed for 0.12 µ technologies. The cross point Q is generated by measuring output high to low delay for both top- and bottom-critical cases. Accuracy of a 4-input NAND gate is around 95%. Note that for all cases when our proposed technique fail to select the optimal transistor orders, the difference between top- and bottom-critical delays are indeed very small. Top- and bottom-critical are virtually equally good. Thus, even though the proposed technique does not achieve 100% theoretical correctness, it is very accurate for practical purposes.

6. Conclusions

Electrical transistor behaviors over a wide range of load capacitance, input transition times, and transistor characteristics are extracted and incorporated into a switch-level model for equivalent circuits. Simple and efficient techniques for determining optimal transistor orders are then proposed. Properties of various transistor orders with respect to gate delays are described. Extensive experimental data show that the proposed technique for predicting transistor orders is very accurate.

References