Scaling of Serially-Connected MOSFET Chains

Srinivasa R. Vemuru
Department of Electrical Engineering
City College of New York
New York, NY 10031

Abstract

Scaling of serially-connected MOSFETs results in reduced propagation delays and power dissipation. Delay formulae and optimum scaling factors are derived as functions of capacitive load, diffusion capacitance and number of MOSFETs in the serially-connected chain using exponential-scaling and linear-scaling methods.

1: Introduction

CMOS logic gates such as Domino gates, NAND gates, AOI gates, etc., are composed of serially-connected MOSFET chains between the output of the logic gate and one of the power supply rails. Sizing of transistors in these MOSFET chains result in reduced propagation delays and power dissipation. However, extensive simulations using techniques such as Monte-Carlo analysis have to be performed to obtain optimum delays [1, 2]. Scaling is a technique where the channel widths of transistors in the MOSFET chains are gradually increased from the transistor close to output node to the power supply node [1-5]. Scaling results in reduction of propagation delay in addition to reduction in overall circuit area and power dissipation in some circuit designs [5]. In exponential-scaling, the transistors are sized such that the ratio of widths of consecutive transistors in the MOSFET chain are the same [5]. In linear-scaling, the transistors are sized such that the difference in widths of consecutive transistors is the same within the MOSFET chain. In parabolic-scaling, the widths of transistors increase in a square law manner as one moves from the transistor connected to output node to the transistor connected to power supply. In this paper, analytical expressions are derived for the Elmore delay for exponentially- and linearly-scaled MOSFET chains. Also expressions are derived from which optimal scaling parameters for serially-connected MOSFET chain can be obtained. Results of parabolic-scaling are not presented due to space constraints. Comparisons are made in terms of propagation delays, areas, and power dissipation for exponentially-scaled, linearly-scaled and unscaled MOSFET chains. SPICE simulations are used to validate these results. The developed theoretical expressions can be used in optimal cell generation of gates such as NAND, NOR, AOI, OAI gates, Domino and dynamic CMOS logic gates. In addition, the delay formulae can be used in fast estimation of propagation delays for timing verification.

2: Analytical derivation of Elmore delay formula

Fig. 1(a) shows a typical N MOSFET chain seen in Domino CMOS logic gates and NAND gates. The equivalent RC model of the gate is shown in Fig. 1(b) [6]; Elmore propagation delay model is used to determine the delay of RC trees [7]. Elmore delay formulae for the discharge of this MOSFET chain is given as

\[ T_D = \sum_{i=0}^{n} C_i \sum_{j=0}^{i} R_j + C_L \sum_{i=0}^{n} R_i \]  

(1)

where \((n+1)\) is the number of serially-connected MOSFETs. \(C_L\), the load capacitance, includes the gate capacitance of next logic gate, drain capacitance of PMOS transistor(s) connected to the output node, and interconnect capacitance. \(C_0, C_1, \ldots, C_n\) are nodal capacitances composed of diffusion capacitances of the MOSFETs. Capacitances \(C_0, C_1, \ldots, C_{n-1}\) are dependent on the source/drain capacitances of two transistors connected to the node; \(C_n\) depends on the drain capacitance of NMOS transistor connected to the output node. The drain/source capacitances of the MOSFETs are proportional to the width of the MOSFETs and can be obtained from the pnp junction area capacitance and side-wall capacitances of the MOSFETs.

2.1: Exponential scaling

Using exponential-scaling, the width of the transistors in the MOSFET chain is given as

\[ W_i = W_0 a_i^e \]  

(2)

\(a_i\) is the scaling factor for the \(i\)-th transistor in the chain. \(W_0\) is the width of the transistor at the output node.
where \( W_0 \) is the width of the NMOSFET connected to the ground. If \( C_{d0} \) is the diffusion capacitance per unit width of the MOSFET, then the capacitances \( C_0, C_1, \ldots, C_{n-1} \) are given as

\[
C_i = \frac{C_0 \alpha^{i}}{1 + \alpha_e} \quad \text{for } i = 0, 1, \ldots, n-1
\]

\[
C_i = \frac{C_0 \alpha^{n}}{1 + \alpha_e} \quad \text{for } i = n
\]

\[
C_0 = C_{d0} W_0 (1 + \alpha_e).
\]

\( R_0, R_1, \ldots, R_n \) represent the resistances of the NMOS transistors in RC tree. Since exponential-scaling is used

\[
R_i = R_0 \alpha_e^{i} \quad \text{for } i = 0, 1, 2, \ldots, n.
\]

Substituting for capacitance and resistance into (1)

\[
T_D = \sum_{i=0}^{n-1} C_0 \alpha^{i} \sum_{j=0}^{i} \frac{R_0}{\alpha^{j} + (C_0 \alpha^{n} + C_L) \sum_{j=0}^{i} \frac{R_0}{\alpha^{j}}}. \tag{6}
\]

Solving (6), the propagation delay is given as

\[
T_D = \frac{C_0 R_0 [(n + 1) (1 - \alpha_e^2) - 2 (\alpha_e - \alpha_e^{n+2})]}{(1 - \alpha_e^2) (1 + \alpha_e)} \frac{1}{C_L R_0 (1 - \alpha_e^{n+1})} \frac{1}{(1 - \alpha_e) \alpha_e^{n+2}}. \tag{7}
\]

The propagation delay is a complex function of load capacitance, diffusion capacitance and number of transistors in the MOSFET chain. Differentiating (7) with respect to \( \alpha_e \) and equating it to zero, the expression from which optimum scaling factor for the NMOS chain, \( \alpha_o, \) is given as:

\[
\frac{2 n (1 - \alpha_e^{n+2}) - 2 (n + 2) (1 - \alpha_e^{n+3})}{(n + 1) \alpha_e - n - \alpha_e^{n+1} \alpha_o} \frac{C_L}{C_0} = 0
\]

\[
\frac{2 n (1 - \alpha_e^{n+2}) - 2 (n + 2) (1 - \alpha_e^{n+3})}{(n + 1) \alpha_e - n - \alpha_e^{n+1} \alpha_o} \frac{C_L}{C_0} = 0
\]

\[
2 n (1 - \alpha_e^{n+2}) - 2 (n + 2) (1 - \alpha_e^{n+3}) \frac{1}{\alpha_e^{n+1} \alpha_o} = 0
\]

\[
(n + 1) \alpha_e - n - \alpha_e^{n+1} \alpha_o = 0
\]

2.2: Linear scaling

The width of the \( i \)-th transistor in a linearly-scaled MOSFET chain is given as

\[
W_i = W_0 (1 - i \alpha_t) \quad \text{for } i = 0, 1, 2, \ldots, n. \tag{9}
\]

where \( W_0 \) is the width of the NMOSFET connected to the ground. The nodal capacitances are given as

\[
C_i = \frac{C_{d0} (2 - (2i + 1) \alpha_t)}{C_{d0} (1 - n \alpha_t)} \quad \text{for } i = 0, 1, \ldots, n-1 \text{ and } \frac{C_{d0} (n + 1)}{C_0} \quad \text{for } i = n.
\]

\[
R_i = \frac{R_0}{1 - i \alpha_t} \quad \text{for } i = 0, 1, 2, \ldots, n
\]

Substituting for capacitance and resistance into (1)

\[
T_D = (C_L + C_n) \sum_{j=0}^{n} \frac{R_0}{1 - i \alpha_t} + \sum_{i=0}^{n} \frac{R_0}{1 - i \alpha_t} \sum_{j=i}^{n} C_{d0} [2 - (2j + 1) \alpha_t]. \tag{12}
\]

Simplifying (12), propagation delay of linearly-scaled chain is given as

\[
T_D = (C_L + C_n) \sum_{j=0}^{n} \frac{R_0}{1 - i \alpha_t} + R_0 C_{d0}.
\]

Approximating \( \frac{1}{1 - i \alpha_t} = 1 + i \alpha_t \), delay is given as

\[
T_D = C_L R_0 \left[ \frac{n + 1}{2} \frac{(n + 1)}{\alpha_t} + \right.
\]

\[
R_0 C_{d0} \left[ \frac{n + 1}{2} \frac{(n + 1)}{\alpha_t} + \frac{2 n^3 + 5 n^2 + 4 n}{6} \alpha_t^2 \right] +
\]

\[
n R_0 C_{d0} \frac{n^4 + 2 n^3 + 2 n^2}{4} \alpha_t^2
\]

Optimum linear scaling factor can be obtained from

\[
\left[ \frac{n^4 + 2 n^3 + 2 n^2}{4} \frac{1}{\alpha_o, l} + \frac{2 n^3 + 5 n^2 + 4 n}{6} \frac{1}{\alpha_2} \right.
\]

\[
C_L \frac{n(n + 1)}{2} = 0.
\]

Approximating \( \frac{1}{1 - i \alpha_t} = 1 + i \alpha_t + i^2 \alpha_t^2 \), delay is given as

\[
T_D = C_L R_0 \left[ \frac{n + 1}{2} \frac{(n + 1)}{\alpha_t} + \frac{2 n^3 + 5 n^2 + n}{6} \right.
\]

\[
R_0 C_{d0} \left[ \frac{n + 1}{3} \frac{(n + 1)}{\alpha_t} + \frac{n^3 + 2 n^2}{12} \frac{1}{\alpha_2} \right] -
\]

\[
n R_0 C_{d0} \left[ \frac{n^2 + 3 n + 2}{3} \frac{1}{\alpha_t} + \frac{n^3 + 2 n^2 + n - 2}{12} \frac{1}{\alpha_2} \right]
\]

Differentiating (16) with respect to \( \alpha \) and equating it to zero, the expression from which optimum scaling factor for the NMOS chain, \( \alpha_o, \) is given as:

\[
\frac{n^2 + 3 n + 2}{3} + \frac{n^3 + 2 n^2 - n - 2}{6} \alpha_o = 0
\]

\[
4 n^4 + 10 n^3 + 10 n^2 + 5 n + 1 \alpha_o^2 = 0
\]

\[
C_L \left[ \frac{(n + 1)}{2} + \frac{(n + 1)(2 n + 1)}{3} \frac{1}{\alpha_o, l} \right] = 0.
\]
3: Results and Discussion

In Fig. 2, propagation delay is plotted versus the number of transistors in MOSFET chain using exponential scaling for different values of taper factors $\alpha_e$ ranging from 1.00 to 0.76. Unscaled MOSFET chain ($\alpha = 1.0$) results in more propagation delay than scaled MOSFET chains ($\alpha = 0.96, 0.88, 0.82$). This is due to additional diffusion capacitances that have to be discharged by the unscaled MOSFET chain. However, excessive scaling will increase propagation delay as illustrated by the curve corresponding to exponential taper factor of 0.76. The taper factor of 0.88 yields minimum propagation delays for MOSFET chains of length greater than 4. For smaller chain lengths, there is not much difference in delays corresponding to taper factors of 0.88 and 0.96.

In Fig. 3, optimum exponential-scaling factor is plotted against number of serially-connected MOSFET stages for five different ratios of $C_L / C_o$. In Fig. 4, optimum linear-scaling factor is plotted against number of serially-connected MOSFET stages for the same values of $C_L / C_o$. The relative values of $C_L$ and $C_o$ affect optimal taper factors. For a given ratio of $C_L / C_o$, the optimal scaling factor normally reduces with increasing number of serially-connected MOSFETs. If the number of serially-connected transistors is small and for larger ratios of $C_L / C_o$, delay contribution towards discharge of $C_L$ dominates the overall falling propagation delay of output voltage. Increasing taper factor reduces the areas of MOSFETs and their source/drain diffusion capacitances and therefore discharge time. However this reduced discharge time do not compensate for increased delays in discharge of $C_L$. Therefore optimum taper factors corresponding to $C_L / C_o$ ratios of 1.5 and 2 reduces as number of serially-connected MOSFETs becomes smaller (2-5).

Bizzan et al observed that minimum Elmore delay is obtained in RC trees by setting delay contributions of each resistor term to be equal [4]. This will be referred to as equal-delay scaling. In Table I, widths of transistors in a MOSFET chain made up of seven serially-connected transistors are compared for exponential, linear and equal-delay scaling approaches. Optimal linear- and exponential-scaling result in nearly twenty percent reduction in active device area as compared to equal-delay scaling. Since overall FET area is smaller, linear and exponential scaling result in reduced power dissipation. Scaling of transistors was done such that all implementations have similar Elmore delays. However, SPICE simulations show that equal-delay scaling results in smaller propagation delays. The reason for this difference is an increase in plateau time in the discharge process of exponential- and linear-scaling methods [8, 9]. The plateau time is the time during which the charge is initially shared between the load capacitance, $C_L$, and other capacitances [8, 9]; this depends on the size of transistor connected to output node, MNO. MNO operates in saturation mode during the initial discharge process where its effective resistance will be much larger than the value used in derivation of (7) and (16). Therefore increasing the width of transistor MNO in linear and exponential scaling methods will result in reduced propagation delays. In modified exponential-scaling and linear-scaling, the width of the transistor MNO in the chain is increased by a factor which is proportional to the ratio of $C_L / C_o$. Though the propagation delay is reduced by increasing the width of MNO, the output voltage fall time of modified scaling methods of MOSFET chain increases as compared to corresponding scaling method as seen in Table I; this is due to increased source and drain area capacitances. The output fall time of exponential and linear scaling methods is slightly smaller than equal-delay scaling implementation.

REFERENCES

Fig. 1(a) Serially-connected FET Chain. (b) RC model used in analysis.

Fig. 2. Delay of MOSFET chains for different exponent scaling factors.

Fig. 3. Optimal exponential scaling factor vs number of serially-connected MOSFETs.

Fig. 4. Optimal linear scaling factor vs number of serially-connected MOSFETs.

**TABLE 1.**
Comparison of Different MOSFET Scaling Schemes

<table>
<thead>
<tr>
<th></th>
<th>Equal Delay Scaling [4]</th>
<th>Exponential Scaling</th>
<th>Linear Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Normal</td>
<td>Modified MN0</td>
</tr>
<tr>
<td>Width of MN6</td>
<td>23 µm</td>
<td>20 µm</td>
<td>20 µm</td>
</tr>
<tr>
<td>Width of MN5</td>
<td>20 µm</td>
<td>17 µm</td>
<td>17 µm</td>
</tr>
<tr>
<td>Width of MN4</td>
<td>17 µm</td>
<td>14 µm</td>
<td>14 µm</td>
</tr>
<tr>
<td>Width of MN3</td>
<td>15 µm</td>
<td>12 µm</td>
<td>12 µm</td>
</tr>
<tr>
<td>Width of MN2</td>
<td>13 µm</td>
<td>11 µm</td>
<td>11 µm</td>
</tr>
<tr>
<td>Width of MN1</td>
<td>11 µm</td>
<td>9 µm</td>
<td>9 µm</td>
</tr>
<tr>
<td>Width of MN0</td>
<td>14 µm</td>
<td>7 µm</td>
<td>12 µm</td>
</tr>
<tr>
<td>Area</td>
<td>113</td>
<td>90</td>
<td>95</td>
</tr>
<tr>
<td>Elmore delay</td>
<td>0.82 ns</td>
<td>0.82 ns</td>
<td>0.83 ns</td>
</tr>
<tr>
<td>SPICE delay</td>
<td>0.84 ns</td>
<td>0.91 ns</td>
<td>0.89 ns</td>
</tr>
<tr>
<td>Fall time</td>
<td>2.05 ns</td>
<td>2.04 ns</td>
<td>2.11 ns</td>
</tr>
</tbody>
</table>