Optimal Transistor Tapering for High-Speed CMOS Circuits *

Li Ding and Pinaki Mazumder
Department of Electrical Engineering and Computer Science
The University of Michigan, Ann Arbor, MI 48109-2122
E-mail: {lding, mazum}@eecs.umich.edu

Abstract

Transistor tapering is a widely used technique applied to optimize the geometries of CMOS transistors in high-performance circuit design with a view to minimizing the delay of a FET network. Currently, in a long series-connected FET chain, the dimensions of the transistors are decreased from bottom transistor to the top transistor in a manner where the width of transistors is tapered linearly or exponentially. However, it has not been mathematically proved whether either of these tapering schemes yields optimal results in terms of minimization of switching delays of the network. In this paper, we rigorously analyze MOS circuits consisting of long FET chains under the widely used Elmore delay model and derive the optimality of transistor tapering by employing variational calculus. Specifically, we demonstrate that neither linear nor exponential tapering alone minimizes the discharge time of the FET chain. Instead, a composition of exponential and constant tapering actually optimizes the delay of the network. We have also corroborated our analytical results by performing extensive simulation of FET networks and showing that both analytical and simulation results are always consistent.

1. Introduction

Transistor sizing plays a pivotal role in improving speed and reducing power consumption of integrated circuits [1]. Transistor tapering is a special type of transistor sizing technique mostly used in dynamic CMOS logic circuits, first proposed by Shoji [2,3] in 1982. Instead of having FETs with uniform width, it is observed that a FET chain may discharge faster if the widths of the FETs decrease gradually from bottom to the top. There are two effects associated with reducing the channel width of the top transistor in a chain. First, smaller channel width means larger effective resistance. This will slow down the discharging of the load capacitance. Second, smaller transistors have smaller parasitic capacitances. This effect tends to reduce the total discharge time. When the parasitic capacitance is large in comparison with the load capacitance, the latter effect may dominate over the first one, hence resulting in a faster discharge process.

Ideally, one wants to size each of the transistors separately to fully explore the possibility of performance gain. A Monte Carlo optimization approach is proposed by Wurtz in [5]. The approach of sizing the transistors independently works well for FET chains where the number of transistors N is small [6]. As the size of N increases, the search space explodes because of increased number of variables. A commonly used way to reduce the problem complexity is to use a tapering shape. For example, under a linear tapering scheme, the widths of adjacent transistors are decreased by a constant, which is the only adjustable variable. This approach can be considered as searching for suboptimal solutions in a one-dimensional subspace of the multi-dimensional variable space. Besides linear tapering, exponential tapering is another widely used tapering shape. Under this scheme, the width ratio of any two adjacent FETs is a constant.

Now the question is whether either of the two commonly used tapering schemes is able to give optimal tapering solution. Shoji does not remark on this issue in his papers. One serious attempt in the quest for the optimal tapering scheme is presented by Bizzan *et al*. in [7], in which the discharge time of a FET chain is viewed as the sum of the delay terms through the effective resistance of each transistor. The authors proposed an analytical approach to the FET chain tapering problem based on an observation that the delay terms were equal in near optimally sized FET chains. This has been incorporated into an automated layout system [8]. However, they could not give any analytical proof for their observation.

In this paper, we propose a new tapering shape which is a composition of constant and exponential functions. We give analytical proof that the proposed tapering scheme is optimal in the long chain limit. Simulations confirm the su-

*This work was partially supported by a NSF grant.

periority of the scheme and show it outperforms linear and exponential tapering schemes for reasonable size of FET chains. Furthermore, we actually have a similar observation as the one by Bizzan and give an analytical proof.

2. Problem Formulation

2.1. RC Model of FET Chains

A series connected FET chain can be modeled as a resistor chain with parasitic capacitors [9]. Since long p-type FET chains are less common in high-speed CMOS circuits, our analysis will be based on NFET chains. However, the formulation and analysis presented in this paper apply to PFET chains the same way.

Using the Elmore delay formula, the discharge time of a FET chain can be written as the sum of *delay terms*:

$$t_D = \sum_{i=0}^{N-1} t_{D,i} = \sum_{i=0}^{N-1} \Big(r_i \cdot (\sum_{j=i}^{N-1} c_j + C_L) \Big), \quad (1)$$

where r_i is the effective resistance of the i-th FET, c_j is the parasitic capacitance between the i-th and the $i+1$-th FETs, C_L is the load capacitance and the delay term $t_{D,i}$ is the Elmore delay contribution of the i-th transistor. Equivalently, the discharge time can be written as the sum of the products of each of the parasitic and load capacitors and their respective resistances to the ground:

$$t_D = \sum_{i=0}^{N-1} \Big(c_i \cdot \sum_{j=0}^{i} r_j \Big) + C_L \cdot \sum_{j=0}^{N-1} r_j. \quad (2)$$

We further assume that the effective resistances and parasitic capacitances are inversely proportional and proportional to the width of the FETs, respectively [9]. Therefore the parasitic capacitance between the i-th and the $(i+1)$-th FET is: $c \cdot (w_i + w_{i+1})/2$, where c is the unit capacitance. Under this assumption, Eqn. (2) can be rewritten as:

$$t_D = \sum_{i=0}^{N-1} \Big(c \cdot \frac{w_i + w_{i+1}}{2} \cdot \sum_{j=0}^{i} \frac{r}{w_j} \Big) + C_L \cdot \sum_{j=0}^{N-1} \frac{r}{w_j}. \quad (3)$$

where r is the unit effective resistance and the undefined variable w_N is considered as zero.

As shown in Eqn. (3), t_D is a function of N variables: w_0, w_1, ..., w_{N-1}. The optimal tapering can be obtained by solving an array of N equations: $\partial t_D / \partial w_i = 0$, for $i = 0, 1, ..., N-1$. In practice, there is also a channel width upper limit for any transistor sizing problems. For the simplicity of presentation, we will thereafter assume the maximum transistor width is normalized to 1.

Before any further analysis of the FET chain tapering problem, we will first discuss a basic property of optimally tapered FET chains.

Lemma 1 *(Property of monotonicity) In an optimally tapered transistor chain, transistor widths decrease monotonically from bottom to top (that is, from ground to load).*

Proof of this lemma is omitted due to the stipulated length of the paper. Note that Lemma 1 does not state that transistor widths *restrictively* decrease monotonically. For example, when the load capacitance C_L is so large that parasitic capacitances of the FET chain are negligible, the optimal transistor sizing is constant.

2.2. Continuous Limit

The array of N equations for optimal transistor tapering can be solved numerically for small N values without much difficulty. The real problem arises when N is moderately large. Instead of trying to solve a large array of nonlinear equations, we study the limit when N is infinitely large. In this scenario, the sums in Eqn. (3) can be approximated by integrations. The new equation reads:

$$t_D = \int_0^1 c \cdot w(x) \cdot dx \cdot \int_0^x \frac{r}{w(\tau)} d\tau + C_L \cdot \int_0^1 \frac{r}{w(x)} dx, \quad (4)$$

where $w(x)$ is a normalized transistor width function. We will thereafter refer to the condition that the above approximation is valid as *continuous limit*. In this limit, a FET chain is actually modeled as an *RC sheet* with uniformly distributed resistance and capacitance.

Equation (4) involves dual integrations. To simplify the form, let us define the resistance to the ground as:

$$R(x) = \int_0^x \frac{r}{w(\tau)} d\tau. \quad (5)$$

Now Eqn. (4) can be transformed to a mathematically more manageable form with single integration:

$$t_D = \int_0^1 \Big(c \cdot r \frac{R(x)}{R'(x)} + C_L R'(x) \Big) dx. \quad (6)$$

It may be noted that Lemma 1 is applicable to the continuous case as well as the discrete case. Finally, the minimum FET chain discharge time problem can be stated as follows:

Problem 1 *(Optimal tapering of FET chain) Given positive constants r, c, and C_L, find the best functional form of $R(x)$ such that*

$$t_D = \int_0^1 \Big(c \cdot r \frac{R(x)}{R'(x)} + C_L R'(x) \Big) dx$$

is minimized and the following constraint is satisfied:

$$\frac{r}{R'(x)} \leq 1, \quad 0 \leq x \leq 1.$$

3. Analytical Approach to FET Chain Tapering

Our goal is to find the optimal shape of the transistor width function. This class of problems belongs to the domain of variational calculus in mathematics. The readers are referred to [10,11] for the background and proof of the fundamental equation in the variational calculus: the Euler-Lagrange differential equation. Similar to the maxima/minima problem in simple calculus, the basic principle of the variational calculus is that the stationary (optimal) function is one that is stable under any small variation upon the function.

3.1. RC Sheet with Fixed Resistor

In this subsection, we will first derive the optimal shape of an RC sheet connected to the ground through a fixed resistor R_0.

Lemma 2 *(Property of exponential shaping) Consider an RC sheet with a fixed resistor R_0 connected to the ground. The optimal shape of the RC sheet is an exponential function:*

$$w(x) = \frac{r}{R_0 \cdot \alpha} e^{-\alpha \cdot x}, \tag{7}$$

where the decay rate α is given by

$$\alpha = ln(R/R_0), \tag{8}$$

where R is the total resistance that includes the resistance of the RC sheet and R_0.

Proof: The discharge time in terms of $R(x)$ is

$$t_D = \int_0^1 \Big(c \cdot r \frac{R(x)}{R'(x)} + C_L R'(x) \Big) dx.$$

Let $F(x) = c \cdot r \cdot R(x)/R'(x) + C_L R'(x)$ and $y(x) = R(x)$. $F(x)$ does not explicitly depend on x. According to the Euler-Lagrange differential equation, we have

$$F - y' \frac{\partial F}{\partial y'} = const, \tag{9}$$

which yields:

$$R(x)/R'(x) = const.$$

Then the functional form for $R(x)$ can be obtained:

$$R(x) = A \cdot e^{\alpha \cdot x}.$$

where A and α are two constants that can be determined by the two boundary conditions, $R(0) = R_0$ and $R(1) = R$:

$$A = R_0, \quad \alpha = ln \frac{R}{R_0}. \tag{10}$$

The width of the RC sheet as a function of x can then be calculated as shown in Eqn. (7). ■

3.2. RC Sheet without Fixed Resistor

Now let us consider our original problem, i.e., an RC sheet without a resistor at the bottom. Letting $R_0 = 0$, we get $A = 0$ from Eqn. (10). Therefore $R(x) = 0$, which is not a physical solution. Mathematically, it is because the function $F(x)$ has a singular point at $x = 0$ in this case. Physically, it is always beneficial to increase $w(x)$ as x is very small. Noting the constraint that the maximum width is 1, we come to the following lemma.

Lemma 3 *(Property of fixed-width shaping) In an optimally tapered RC sheet, there exists a positive value x_0 such that $w(x) = 1$ for $0 \le x \le x_0$.*

Proof: Euler's equation is not applicable here because of the singularity. But we can still use the basic principle of variational calculus, i.e., an optimal shape is one that any perturbation upon it will only increase the cost function.

Consider a small increase in width, Δw, at location $(x_0 - \Delta x/2, x_0 + \Delta x/2]$. This has two effects on the discharge time. First, the portion of RC sheet above x_0 will be discharged faster since it has smaller effective resistance.

$$\Delta t_{D,1} \simeq -\Big(C_L + \int_{x_0}^1 c \cdot w(x) dx \Big) \frac{r}{w(x_0)^2} \Delta x \Delta w.$$

Second, larger width (therefore larger capacitance) at x_0 requires additional time to discharge through the lower portion of the RC sheet.

$$\Delta t_{D,2} \simeq c \cdot \int_0^{x_0} \frac{r}{w(x)} dx \cdot \Delta x \Delta w.$$

Notice that $\Delta t_{D,2}$ has an integration from 0 to x_0. Therefore, when x_0 is very small, the value of $\Delta t_{D,2}$ will be very small. On the other hand, $\Delta t_{D,1}$ is always a finite negative value. This implies that the discharge time t_D will always decrease as one increases the width at position x_0. However, under our assumption, the maximum width is 1. Therefore, the optimal width as $0 \le x \le x_0$ is 1 for a very small x_0. ■

3.3. Optimal Transistor Chain Tapering

Based on Lemma 2 and 3, one can expect that the optimal width shape of the RC sheet is constant from the bottom till a specific point $x = a$, then it decreases exponentially. To make it a real solution to the problem, the continuity condition at position $x = a$ has to be satisfied. Using Eqn. (7), it is straightforward to derive the continuity condition as: $a = 1/\alpha$. The continuity constraint implies that the optimal shape is not a simple combination of arbitrary exponential and constant functions. Those two parts are related. For example, a small constant part suggests a large value of α; therefore, a steeper exponential decrease.

Theorem 1 *(Optimal tapering of FET chain) Given the load capacitance C_L and maximum width 1, the optimal tapering shape of a transistor chain is a combination of constant and exponential functions given by:*

$$w(x) = \begin{cases} e^{1-\alpha x} & \frac{1}{\alpha} < x \leq 1 \\ 1 & 0 \leq x \leq \frac{1}{\alpha} \end{cases}, \qquad (11)$$

where α is determined by solving the following equation:

$$\alpha \cdot e^{\alpha - 1} = \frac{c}{C_L}. \qquad (12)$$

Proof: First, the optimality of the proposed shape function can be verified in three regions.
1) When $\frac{1}{\alpha} < x \leq 1$, the RC sheet at $[0, \frac{1}{\alpha})$ is effectively a resistor of resistance $r \cdot \frac{1}{\alpha}$. Using Lemma 2, one obtains that the optimal shape of this portion of the sheet is an exponential function:

$$w(x) = \frac{r}{(r \cdot \frac{1}{\alpha}) \cdot \alpha} e^{-\alpha(x - \frac{1}{\alpha})} = e^{1-\alpha x}.$$

2) When $x = \frac{1}{\alpha}$, use the continuity property, we have

$$w(\frac{1}{\alpha}) = e^{1-\alpha \cdot \frac{1}{\alpha}} = 1.$$

3) When $0 \leq x < \frac{1}{\alpha}$, according to Lemma 1 (Property of monotonicity), the only possible solution is:

$$w(x) = 1, \quad 0 \leq x < a.$$

Next is to determine the value of parameter α. When $0 \leq x < \frac{1}{\alpha}$, $R(x) = r \cdot x$. When $\frac{1}{\alpha} \leq x \leq 1$, the resistance to ground can be calculated as:

$$R(x) = \frac{r}{\alpha} + \int_{\frac{1}{\alpha}}^{x} r \cdot e^{-1+\alpha x} = \frac{r}{\alpha} + r \cdot \frac{e^{-1+\alpha x} - 1}{\alpha}.$$

The total discharge time, after simplification, is

$$\frac{t_D}{r \cdot c} = \frac{1}{\alpha}\left(1 - \frac{1}{2\alpha}\right) + \frac{C_L}{c} \cdot \frac{1}{\alpha} \cdot e^{\alpha - 1}. \qquad (13)$$

Finally, the value of α, shown in Eqn. (12), can be obtained by solving the equation $dt_D/d\alpha = 0$. ■

3.4. Equal Delay Terms

Equal delay term observation [7] states that the RC delay of each transistor in an optimally tapered chain is equal. The RC delay of a FET is defined here as the product of the effective resistance of the FET and total capacitance that discharges through the transistor, that is, the total capacitance *above* the transistor. The following corollary is the continuous correspondence to the equal delay term observation.

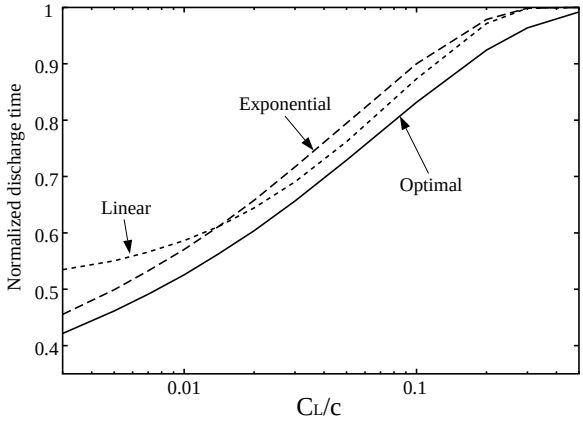

Figure 1. Comparison of tapering schemes.

Corollary 1 *(Equal delay terms) Under the optimal tapering shape, differential delay term function*

$$\Delta t_D(x) = \Big(\int_x^1 c(\tau) \cdot d\tau + C_L \Big) \cdot r(x) \cdot \Delta x.$$

is a constant when $\frac{1}{\alpha} \leq x \leq 1$.

Proof: We have the following equations under the optimal tapering:

$$c(x) = c \cdot e^{1-\alpha x}, \quad C_L = \frac{c}{\alpha} \cdot e^{1-\alpha}, \quad r(x) = r \cdot e^{-1+\alpha x}.$$

Therefore, the differential delay term can be calculated as:

$$\Delta t_D(x) = \Big(\int_x^1 c e^{1-\alpha x} d\tau + \frac{c}{\alpha} e^{1-\alpha} \Big) \cdot \big(r e^{-1+\alpha x} \big) \Delta x$$

$$= \Big(\frac{c}{\alpha} \cdot e^{1-\alpha x} \Big) \cdot r \cdot e^{-1+\alpha x} \Delta x = \frac{r \cdot c}{\alpha} \Delta x.$$

which is a constant independent of variable x. ■

Here the equal delay term argument is valid only in the exponential tapering region. However, since the transistors in the lower part have maximum channel width, this does not add complexity in using this corollary. It is also observed that neither linear nor exponential tapering has the equal delay term property.

4. Experiments

4.1. Continuous Limit

We first compare linear, exponential and the proposed tapering schemes at the continuous limit of FET chains using the Elmore delay model. Figure 1 shows the least discharge time of a FET chain with respect to the C_L/c value under each of the three tapering schemes. The discharge times are

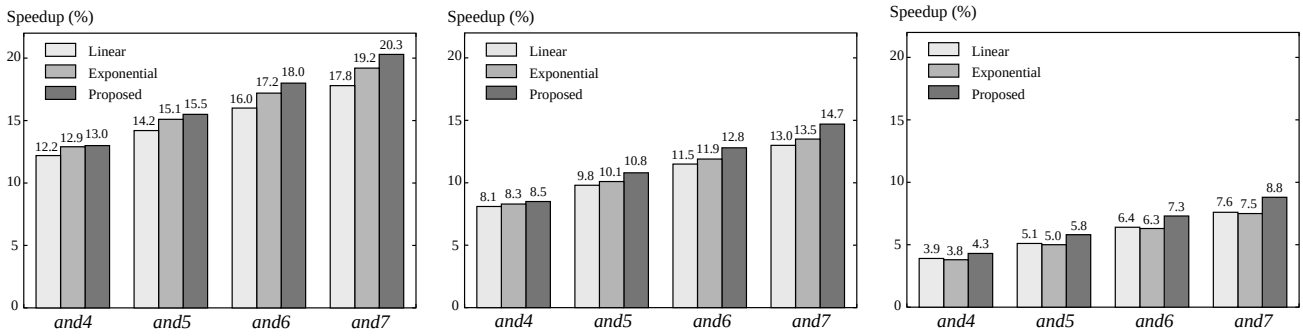

Figure 2. Comparison of the tapering schemes by simulation (a) 1-INV; (b) 2-INV; (c) 4-INV.

normalized to those obtained without performing transistor tapering. We show C_L/c in a logarithmic scale so that it covers a very wide range. The proposed tapering scheme consistently outperforms both linear and exponential tapering schemes. This is especially the case when C_L/c is relatively large, that is, when tapering starts to be beneficial. In this region, exponential and linear tapering schemes do not produce any noticeable performance improvement while the proposed scheme does. The simulation result, in another way, verifies the superiority of the proposed tapering scheme.

The performance gain due to transistor tapering is more significant for smaller C_L/c values under all of the three schemes. Linear tapering is better than exponential tapering when C_L/c is larger than about 0.02. However, when C_L/c is very small, the performance of linear tapering deteriorates because of the choked neck effect near $x = 1$. On the other hand, the optimal tapering shape approaches the exponential shape at the limit that C_L/c goes to zero. Therefore, exponential tapering performs well as C_L/c is very small.

4.2. SPICE Simulation

The delay model we used is a simplified one. In a real FET chain, many assumptions we made like simple RC model of FET chains, Elmore delay formula, etc., may not exactly hold. One should also consider short channel effects, Miller and back-gate coupling effects, and so on, in a real deep submicron circuit. However, it is very difficult to include those second-order effects into a manageable formulation. It is even more difficult to solve the formula, once obtained, analytically.

Therefore, in this subsection, we study the FET chain based on HSPICE simulation to see whether the result obtained from the simple model works for real world circuits. We use HSPICE Level 49 MOSFET model for the transistors. As real cases that FET chains are used in circuit designs, we study a set of domino AND gates: *and4*, *and5*, *and6*, and *and7*. Simulation is carried out in a 0.18 μm technology at 1.6 V, 55 ^{o}C, and typical process. For simplicity, we do not use the optional keeper as that is not our main purpose. The maximum channel width of a transistor is set to 20 μm. We change the size of the output inverter to change the C_L value of the FET chain. Three different sizes of load inverters are used: $1\times$, $2\times$, and $4\times$ sized inverters, where a unit inverter has a PFET width of 2 μm and NFET width of 1 μm.

To determine the parameter α, one needs to know both load capacitance C_L and parasitic capacitance c. Since it is difficult to obtain the accurate value of effective parasitic capacitance of the transistors, we search the parameter space exhaustively to get the lowest delay value under each of those three different schemes. A program written in the C language is developed serving as a batch program which calls HSPICE for each different parameter and reads the measured delay data generated by HSPICE. The shortest delay time and its corresponding α value are then reported.

Simulation results for the domino AND gates are shown in Figure 2. Delay is measured from inputs to node *S*. It is observed that the performance gains of the three tapering schemes are quite comparable in each single cases. This is a common phenomenon when several suboptimal solutions, each of which is the optimal solution in their respective subspace, are compared. It also explains the reason that linear and exponential taperings have worked fine for the past two decades. Nevertheless, the proposed tapering scheme consistently outperforms linear and exponential schemes in all cases.

4.3. Beyond Simple FET Chain

The mixed constant and exponential tapering shape is proposed for a simple chain of FETs. However, the basic idea that we should keep the transistor width of the lower portion of a long chain constant may apply to a larger class of circuits like complex domino OR-AND (OA) gates. We will use two examples: domino OA654321 and OA424242

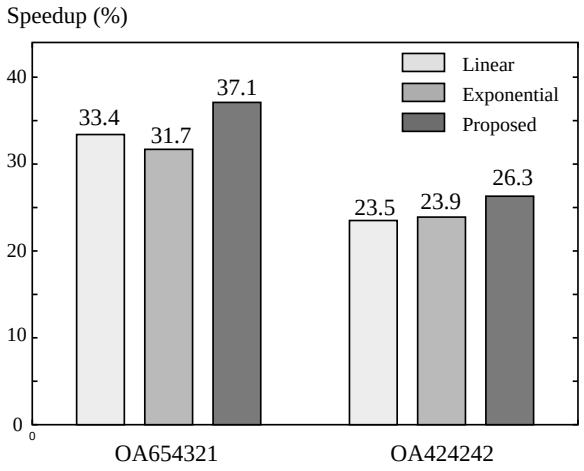

Figure 3. Simulated results of two OA gates.

gates. The former one is the example used in Shoji's pioneer paper on transistor chain tapering [3]. It is a special case when transistor tapering will generate large performance gain. On the contrary, our second example represents a more general scenario. Note that those circuits are not unreasonably large since domino CMOS circuits usually have very complex logic combination [3]. In practice, the internal nodes of large NFET trees are also precharged to V_{DD} in the precharge phase to solve the charge sharing problem. Therefore, we assume the internal nodes discharge from V_{DD} in our simulation. The maximum channel width for the transistors in the OA gates are set to $10\mu m$ and we assume the size of the output inverters is 1 unit.

Simulation results on those OA gates are shown in Figure 3. Linear tapering is better than exponential tapering in one case and worse in the other case. The proposed scheme wins out with a decent margin in both cases.

5. Conclusions

The problem of optimal tapering of a series connected FET chain has been studied in this paper. In contrast to conventional approaches which try to first solve the problems for special cases when *N* is small, we start from the case when *N* approaches infinite. The results obtained from small *N* problems cannot be used to problems of larger sizes; therefore, empirical tapering schemes like linear and exponential shapes are used. We show that the optimal tapering shape obtained in the continuous limit works well for FET chains with medium number of *N* and it actually outperforms linear and exponential tapering schemes in our simulations.

Overall, the contributions of this paper can be concluded as follows. 1) FET chain tapering problem is first formulated and analyzed in the continuous limit. 2) A new mixed constant and exponential tapering scheme is proposed and proved to be optimal in the long chain limit. 3) The proposed tapering scheme is demonstrated better than linear and exponential tapering schemes in most cases. 4) For the first time, an analytical framework is provided to the the equal delay terms argument/observation.

In this paper, we study the original tapering problem [3] considering speed as the main optimization goal. In real circuit design practice, one may consider other design matrices like power consumption, signal integrity, etc. For example, one might formulate the problem such that the power-delay product is minimized. Also, he may need to add certain constraints to avoid using FETs with very small channel width to ensure reliability. Pursuing optimal tapering shapes for multi-objectives under multi-constraints poses a challenge for future research.

References

[1] C. Visweswariah, "Optimization Techniques for High-Performance Digital Circuits," *Proc. Int. Conf. on Computer-Aided Design*, pp. 198-205, 1997.

[2] M. Shoji, "Electrical Design of BELLMAC-32A Microprocessor," *Proc. Int. Conf. on Circuits and Computers*, pp. 112-115, 1982.

[3] M. Shoji, "FET Scaling in Domino CMOS Gates," *IEEE Journal of Solid-State Circuits*, vol. SC-20, no. 5, pp. 1067-1071, 1985.

[4] B. S. Cherkauer, and E. G. Friedman, "Channel Width Tapering of Serially Connected MOSFET's with Emphasis on Power Dissipation," *IEEE Transaction on VLSI Systems*, vol. 2, no. 1, pp. 100-114, 1994.

[5] L. T. Wurtz, "An Efficient Scaling Procedure for Domino CMOS Logic," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 9, pp. 979-982, 1993.

[6] J. Yuan and C. Svensson, "Principle of CMOS Circuit Power-Delay Optimization with Transistor Sizing," *Proc. IEEE Int. Symp. on Circuits and Systems*, pp. 637-640, 1996.

[7] S. S. Bizzan, G. A. Jullien, and W. C. Miller, "Analytical Approach to Sizing nFET Chains," *Electronics Letters*, vol. 28, no. 14, pp. 1334-1335, 1992.

[8] G. A. Jullien, W. C. Miller, R. Grondin, Z. Wang, L. Del Pup, and B. Bizzan, "Woodchuck: A Low-Level Synthesizer for Dynamic Pipelined DSP Arithmetic Logic Blocks," *Proc. IEEE Int. Symp. on Circuits and Systems*, pp. 176-179, 1992.

[9] T. Lin and C. A. Mead, "Signal Delay in General RC Networks," *IEEE Transaction on CAD*, vol. 3, no. 4, pp. 331-349, 1984.

[10] W. S. Kimball, *Calculus of Variations*, Butterworths Scientific Publications, London, 1952.

[11] M. J. Forray, *Variational Calculus in Science and Engineering*, McGraw-Hill, 1968.