

On Optimal Tapering of FET Chains in High-Speed CMOS Circuits

Li Ding, *Student Member, IEEE*, and Pinaki Mazumder, *Fellow, IEEE*

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 FET network where MOS devices are stacked over one another to form a series chain network, the dimensions of the transistors are decreased from the 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.

Index Terms—Circuit optimization, circuit theory and design, Elmore delay, field-effect transistors, transistor sizing, transistor tapering.

I. INTRODUCTION

TRANSISTOR sizing is frequently used in high-performance integrated circuit design in order to improve the speed of operation and concomitantly to reduce the power dissipation of a constituting circuit block [1]. Transistor tapering is a special class of transistor sizing technique where, instead of having a stack of uniformly wide FETs, the widths of FETs in the stack, monotonically increase from the top n-type transistor, connected to the output node, to the bottom FET, connected to ground. For a p-type stack, the top transistor is the widest and is connected to the supply voltage, while the FET with the smallest width will be the bottom p-type transistor, connected to the output node. In high-speed CMOS domino and dynamic logic families, only one type of switching network (pull-down in the case of NMOS network and pull-up in the case of PMOS network) is used to implement a Boolean function. In such a FET network, transistor tapering must be carefully selected in order to minimize delay and power dissipation of

the logic circuit as well as to improve its noise performance by reducing charge sharing between the output node and the FET network's internal nodes. Shoji [2], [3] first proposed the transistor tapering technique in 1982 and demonstrated that it can significantly improve performance of large FET networks.

It may be observed that an n-type FET chain may discharge faster if widths of the FETs decrease gradually from bottom FET to the top. There are two effects associated with the reduction of channel width of the top transistor in a chain. First, a smaller channel width means larger effective transistor-ON resistance, causing to discharge the load capacitance slowly. Second, by reducing the FET size, the input and output parasitic capacitances associated with the FET also reduce, which translates into reduction of the discharge time of the output node. 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. In the case when the load capacitance is not very small, transistor tapering may not produce significant improvement in speed. However, power consumption may still be significantly reduced [4], [5], mainly because of the smaller input capacitance.

Ideally, one wants to size each FET in the chain separately to fully explore the possibility of performance gain. By using a Monte Carlo optimization, Wurtz [6] developed an approach of independently sizing the transistors, which works well for FET chains having a small number of transistors N [7]. As the size of N increases, the search space quickly becomes too large, since the number of variables increases rapidly.

A commonly used way to reduce the problem complexity is to use a tapering scheme. For example, under a linear tapering scheme, the widths of adjacent transistors are decreased by a constant Δw . Therefore, Δw is the only adjustable variable. This approach can be considered as searching for suboptimal solutions in a one-dimensional subspace of the whole variable space. Besides linear tapering, exponential tapering is another widely used tapering scheme. Under this scheme, the width ratio of any two adjacent FETs is a constant. Example FET chains with linear and exponential tapering are shown in Fig. 1(b) and 1(c), respectively.

Now the question is whether either of the two commonly used tapering schemes is optimal, i.e., whether the global optimal solution lies within the one-dimensional subspace defined by either of them. Shoji does not remark on this issue in his papers; instead, he applied an exponential tapering scheme in [3]. One serious attempt to find the optimal tapering scheme is presented by Bizzan *et al.* in [8], 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

Manuscript received August 23, 2000; revised April 30, 2001. This work was supported in part by a grant from the National Science Foundation. This paper was recommended by Associate Editor X. Song.

The authors are with the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109-2122 USA (e-mail: ld Ding@eecs.umich.edu; mazum@eecs.umich.edu).

Publisher Item Identifier S 1057-7130(01)11371-6.

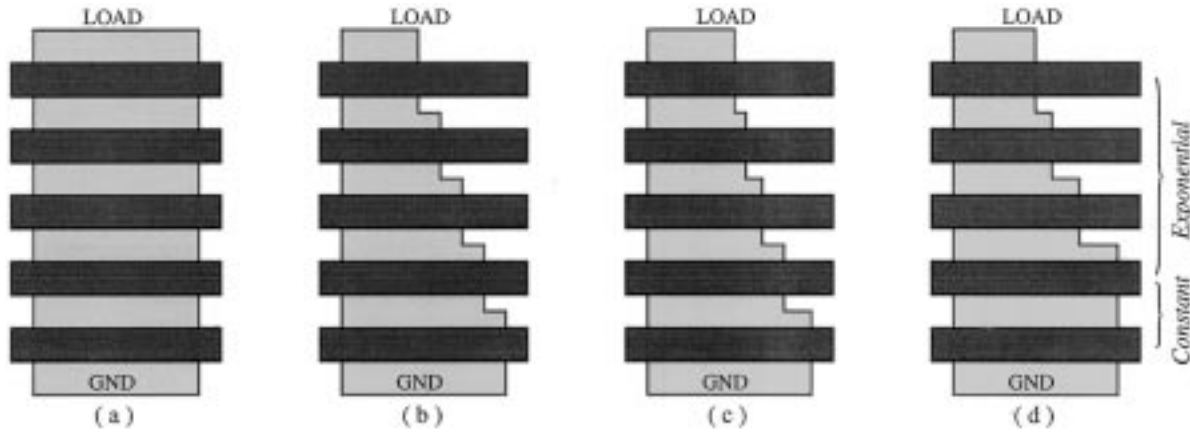


Fig. 1. FET chain. (a) Without tapering. (b) Linear tapering. (c) Exponential tapering. (d) Combination of constant and exponential (proposed).

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 [9]. However, they could not give an analytical proof for their observation.

In this paper, we propose a new tapering shape which is a composition of constant and exponential functions, as shown in Fig. 1(d). We give analytical proof that the proposed tapering scheme is optimal in the long chain limit (as N approaches infinite). Simulation confirms the superiority of the scheme and shows 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.

The rest of the paper is organized as follows. In Section II, the formulation for the transistor tapering problem under Elmore delay model is presented. This formulation is also extended to the continuous limit. In Section III, we propose a new tapering scheme and prove its optimality in the continuous limit. Section IV compares the three tapering schemes by simulation. The proposed scheme works consistently better than linear and exponential tapering schemes when N is moderately large. Finally, Section V concludes the paper.

II. PROBLEM FORMULATION

A. RC Model of FET Chains

A series connected FET chain can be modeled as a resistor chain with parasitic capacitors [10] as shown in Fig. 2. 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 [12], the discharge time of the FET chain can be written as the sum of delay terms

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

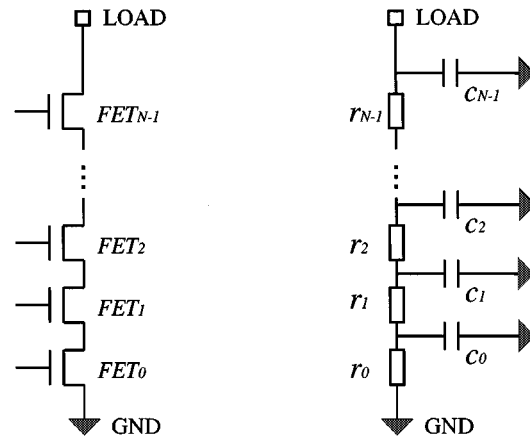


Fig. 2. RC model of FET chains.

where the delay term $t_{D,i}$ is the Elmore delay contribution of the i th transistor, r_i the effective resistance of the i th FET, c_j is the parasitic capacitance at node j and C_L is the load capacitance (not shown in Fig. 2). 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

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

We further assume that the effective resistances and parasitic capacitances are inversely proportional and proportional to the width of the FETs, respectively, [10], [11]. Therefore, the parasitic capacitance between the i th and the $(i + 1)$ th FET is $c \cdot (w_i + w_{i+1})/2$, as shown in Fig. 3, where c is the unit capacitance. Under this assumption, (2) can be rewritten as

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

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

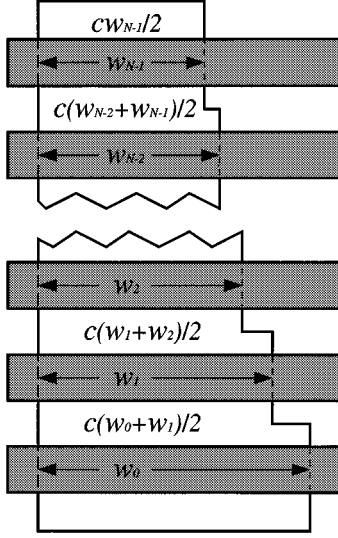


Fig. 3. Parasitic capacitances.

B. Optimal Transistor Tapering

As shown in (3), t_D is a function of N variables: w_0, w_1, \dots, w_{N-1} . The optimal tapering can be obtained by solving an array of N equations

$$\frac{\partial t_D}{\partial w_i} = 0, \quad i = 0, 1, \dots, N - 1.$$

Now substitute (3) to the above equation array. After simplification, the partial derivative equation with respect to variable w_i reads

$$\sum_{j=0}^i \frac{1}{w_j} - \sum_{j=i+1}^{N-1} \frac{w_j}{w_i^2} - C_L \frac{1}{w_i^2} = 0. \quad (4)$$

In practice, we have a channel width upper limit for a transistor sizing problem. For the simplicity of presentation, we will thereafter assume the maximum transistor width is normalized to 1. Obviously, increasing the channel width of the the lowest transistor will always bring down the discharge time. Hence the best size of this transistor is 1. And now we only need to solve the rest $N - 1$ equations under the constraint that w_i is less than or equal to 1 for any $i = 1, 2, \dots, N - 1$.

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 (ground) to top (load).

Proof: Assuming there exist two numbers i and j such that $i < j$ and $w_i < w_j$, let us swap the sizes of those two FETs.

Let us first consider the parasitic capacitance associated with those two FETs. The discharge time before swapping is

$$t_1 = \frac{cw_j}{2} \left(\frac{r}{w_j} + 2R_{II} + \frac{2r}{w_i} + 2R_I \right) + \frac{cw_i}{2} \left(\frac{r}{w_i} + 2R_I \right)$$

where R_I is the effective resistance below the i th FET and R_{II} is the resistance between the i th and the j th FETs. Similarly, the discharge time after swapping is

$$t_2 = \frac{cw_i}{2} \left(\frac{r}{w_j} + 2R_{II} + \frac{2r}{w_i} + 2R_I \right) + \frac{cw_j}{2} \left(\frac{r}{w_i} + 2R_I \right).$$

Therefore, the change in discharge time due to those capacitors reads

$$\Delta t_D = t_2 - t_1 = \frac{c}{2}(w_i - w_j) \cdot \left(\frac{r}{w_j} + 2R_{II} + \frac{r}{w_i} \right).$$

Since $w_i < w_j$, Δt_D is always negative.

The rest of the capacitors can be classified into three categories.

- 1) Apparently, the discharge time of the capacitors below the i th node does not change after the swapping.
- 2) The discharge time of the capacitors above the j th node (including the load capacitor C_L) also will not change because their respective resistance to the ground does not change.
- 3) The discharge time of the parasitic capacitors between the i th and the j th FETs will decrease because they have less resistance to ground after swapping.

In all, the FET chain discharges faster after swapping. This observation contradicts the claim that the original sizing is optimal. Therefore, the lemma is proved. ■

Note that Lemma 1 does not state that transistor widths *re-strictively* 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, i.e., $w_i = 1$ for all $i = 0, 1, \dots, N - 1$.

C. Continuous Limit

The array of $N - 1$ (4) 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 summations in (3) can be approximated by integrations. The new equation reads

$$t_D = \int_0^1 cw(x) \cdot dx \cdot \int_0^x \frac{r}{w(\tau)} d\tau + C_L \int_0^1 \frac{r}{w(x)} dx \quad (5)$$

where $w(x)$ is the normalized transistor width function.

We will thereafter refer to the condition that the above approximation is valid as the *continuous limit*. In this limit, a FET chain is actually modeled as an RC sheet with uniformly distributed resistance and capacitance.

It is known that as the value of N increases the problem becomes harder to solve, because the size of the equation matrix also increases. Continuous limit is the case when some sort of analysis is really demanded. It is also the case when the word optimal tapering *shape* makes sense. We will show later under Section IV what should be the value of N so that it can be considered as large enough. And it is derived by simulation.

Equation (5) involves dual integrations. To simplify the form, let us define the resistance to the ground as

$$R(x) = \int_0^x \frac{r}{w(\tau)} d\tau. \quad (6)$$

Then transistor width function can be expressed in terms of $R(x)$

$$w(x) = \frac{r}{R'(x)} \quad (7)$$

where the prime refers to the derivative with respect to x . Hence, (5) can be transformed to a mathematically more manageable form with single integration

$$t_D = \int_0^1 \left(c \cdot r \frac{R(x)}{R'(x)} + C_L R'(x) \right) dx. \quad (8)$$

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 chains) given positive constants r , c and C_L , find the best functional form of $R(x)$ such that

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

is minimized and

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

is satisfied.

III. 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. Therefore, we will first introduce the fundamental equation in the variational calculus: the Euler–Lagrange differential equation.

If a cost function I is defined by an integral of the form

$$I = \int_{x_1}^{x_2} F(x, y, y') dx$$

where $y(x_1) = y_1$ and $y(x_2) = y_2$ are prescribed, then I has a stationary value if the Euler's equation

$$\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial y'} = 0 \quad (9)$$

is satisfied. Furthermore, in case $F(x, y, y')$ does not explicitly depend on x , the Euler's equation can be reduced to

$$F - y' \frac{\partial F}{\partial y'} = \text{const.} \quad (10)$$

The readers are referred to [13] and [14] for the proof of the above theorem. 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.

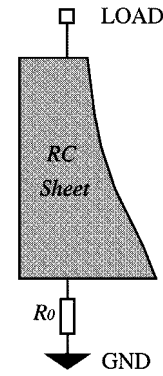


Fig. 4. RC sheet with fixed resistor.

A. RC Sheet With Fixed Resistor

According to Section II, our problem is to find the optimal width function of an RC sheet with a capacitive load of C_L . In this subsection, we will first obtain the optimal shape of an RC sheet connected with a fixed resistor, as shown in Fig. 4. This is stated in Lemma 2.

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

$$w(x) = \frac{r}{R_0 \alpha} e^{-\alpha x} \quad (11)$$

where the decay rate α is given by

$$\alpha = \ln \frac{R}{R_0} \quad (12)$$

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

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

Let $F(x) = c \cdot r R(x)/R'(x) + C_L R'(x)$ and $y(x) = R(x)$. Apparently, $F(x)$ does not explicitly depend on x . Therefore, we can use (10)

$$\text{const.} = \left(c \cdot r \frac{R(x)}{R'(x)} + C_L R'(x) \right) - R'(x) \frac{\partial}{\partial R'(x)} \left(c \cdot r \frac{R(x)}{R'(x)} + C_L R'(x) \right).$$

After simplification, we get

$$2c \cdot r \frac{R(x)}{R'(x)} = \text{const.}$$

from which the functional form for $R(x)$ can be obtained

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

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

$$A = R_0, \quad \alpha = \ln \frac{R}{R_0}. \quad (13)$$

Therefore, the width of the RC sheet as a function of x can be calculated as

$$w(x) = \frac{r}{R'(x)} = \frac{r}{R_0 \cdot \alpha} e^{-\alpha \cdot x}. \quad \blacksquare$$

B. RC Sheet Without Fixed Resistor

Now let us return to our original problem, i.e., an RC sheet without a resistor at the bottom. Letting $R_0 = 0$, we get $A = 0$ from (13). Therefore, $R(x) = 0$, which means that we cannot get a physical solution using Lemma 2. That is, exponential shape is no longer the optimal shape.

Similar to our earlier argument that the lower-most transistor should always be sized as large as possible, it is always beneficial to increase $w(x)$ as x is very small, which leads to a singular point at $x = 0$. This is the reason that Lemma 2 fails in this case. However, noticing 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 $0 \leq x \leq x_0$.

Proof: Euler's equation is not applicable here because of the singularity at $x = 0$. 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 $(x_0 - \Delta x/2, x_0 + \Delta x/2]$. 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.

$$\begin{aligned} \Delta t_{D,1} &= \left(C_L + \int_{x_0 + \Delta x/2}^1 c w(x) dx \right) \\ &\quad \times \left(\frac{r}{w(x_0 + \Delta w)} - \frac{r}{w(x_0)} \right) \cdot \Delta x \\ &\simeq - \left(C_L + \int_{x_0}^1 c w(x) dx \right) \frac{r}{w(x_0)^2} \cdot \Delta x \Delta w. \end{aligned}$$

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

$$\begin{aligned} \Delta t_{D,2} &= (c \Delta x \Delta w) \cdot \int_0^{x_0 - \Delta x/2} \frac{r}{w(x)} dx \\ &\simeq c \int_0^{x_0} \frac{r}{w(x)} dx \cdot \Delta x \Delta w. \end{aligned}$$

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 $0 \leq x \leq x_0$ is 1 for a very small x_0 . \blacksquare

C. 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 physical solution to the problem, the continuity condition at

position $x = a$ has to be satisfied. Note that since the starting point for the exponential decrease is now $x = a$ instead of $x = 0$, (11) is rewritten as

$$w(x) = \frac{r}{R_0 \cdot \alpha} e^{-\alpha \cdot (x-a)}.$$

The constant part of the RC sheet near the bottom has a resistance of $r \cdot a$. Therefore, we have

$$\frac{r}{(r \cdot a) \cdot \alpha} = 1$$

which, on simplification, yields

$$a = \frac{1}{\alpha}.$$

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 chains) given the load capacitance C_L and maximum width 1, the optimal tapering shape of a transistor chain in the continuous limit is a combination of constant and exponential functions given by

$$w(x) = \begin{cases} e^{1-\alpha x}, & \frac{1}{\alpha} < x \leq 1 \\ 1, & 0 \leq x \leq \frac{1}{\alpha} \end{cases} \quad (14)$$

where α is determined by solving the following equation:

$$\alpha \cdot e^{\alpha-1} = \frac{c}{C_L}. \quad (15)$$

Proof: First, the optimality of the proposed shape function can be verified in three regions.

- 1) When $1/\alpha < x \leq 1$, the RC sheet at $[0, 1/\alpha)$ is effectively a resistor of resistance $r \cdot 1/\alpha$. Using Lemma 2, one obtains that the optimal shape of this portion of the sheet is an exponential function

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

- 2) When $x = 1/\alpha$, using the continuity property we have

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

- 3) When $0 \leq x < 1/\alpha$, 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 $0 \leq x < 1/\alpha$, $R(x) = r \cdot x$. When $1/\alpha \leq x \leq 1$, the resistance to ground can be calculated as

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

The total discharge time is

$$\begin{aligned} \frac{t_D}{r \cdot c} &= \int_0^{1/\alpha} x \cdot dx + \int_{1/\alpha}^1 \left(\frac{1}{\alpha} + \frac{e^{-1+\alpha x} - 1}{\alpha} \right) \cdot e^{1-\alpha x} dx \\ &\quad + \frac{C_L}{c} \cdot \left(\frac{1}{\alpha} + \frac{e^{-1+\alpha} - 1}{\alpha} \right). \end{aligned}$$

After simplification, it reads

$$\frac{t_D}{r \cdot c} = \frac{1}{\alpha} \left(1 - \frac{1}{2\alpha}\right) + \frac{C_L}{c \cdot \alpha} e^{\alpha-1}. \quad (16)$$

Finally, the value of α can be determined by solving the equation $dt_D/d\alpha = 0$

$$\alpha \cdot e^{\alpha-1} = \frac{c}{C_L}.$$

Note that an α value less than 1 means that the whole RC sheet has a constant width 1. Therefore, the FET chain is actually not tapered. This is the case when C_L is large in comparison with the parasitic capacitances. From (15), one can easily obtain the condition that transistor tapering is not required

$$C_L \geq c \quad (17)$$

where c can also be viewed as the total capacitance of an RC sheet with constant width of 1, or, in the discrete case, total parasitic capacitances of a FET chain with all FETs having the maximum allowable width.

The left-hand side of (15) is a transcendental function. Hence one cannot write an analytical expression for α with respect to circuit parameters c and C_L . For the purpose of hand-calculation or CAD program, one can use the following iterative approach to get a numerical value of α

$$\alpha^{(n+1)} = 1 - \ln \left(\frac{C_L}{c} \cdot \alpha^{(n)} \right). \quad (18)$$

Since $\alpha \geq 1$, the term e^α dominates α in the left-hand side of (15). Therefore, a good starting point for the iterative equation can be chosen as

$$\alpha^{(0)} = \ln \left(\frac{e \cdot c}{C_L} \right). \quad (19)$$

Usually, it only takes a few iterations to get a good numerical solution.

D. Equal Delay Terms

Equal delay term observation [8] 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, i.e., the total capacitance above the transistor. The following corollary is the continuous correspondence to the equal delay term observation.

Corollary 1: (Equal delay terms) under the optimal tapering shape, differential delay term function

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

is a constant when $1/\alpha \leq x \leq 1$.

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

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

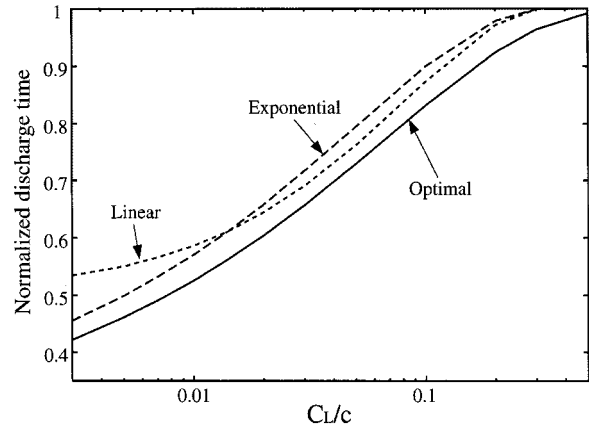


Fig. 5. Comparison of linear, exponential and proposed tapering schemes.

Therefore, the differential delay term can be calculated as

$$\begin{aligned} \Delta t_D(x) &= \left(\int_x^1 c e^{1-\alpha x} d\tau + \frac{c}{\alpha} e^{1-\alpha} \right) \cdot (r e^{-1+\alpha x}) \Delta x \\ &= \left(\frac{c}{\alpha} e^{1-\alpha x} \right) \cdot r e^{-1+\alpha x} \Delta x = \frac{r \cdot c}{\alpha} \Delta x. \end{aligned}$$

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 the corollary. It is also noted that neither linear nor exponential tapering has the equal delay term property.

IV. EXPERIMENTS

A. Continuous Limit

We will first compare linear, exponential and the proposed tapering schemes at the continuous limit of FET chains using the Elmore delay model. The discharge time of a FET chain under the proposed tapering scheme has been calculated in (16). Similarly, we can derive the discharge time of FET chains under linear and exponential tapering schemes.

The transistor width function for a linearly tapered transistor chain is $w(x) = 1 - \alpha x$. The effective resistance from a given position x to the ground is calculated as

$$R(x) = \int_0^x \frac{r}{1 - \alpha \tau} d\tau = -\frac{r}{\alpha} \ln(1 - \alpha x).$$

Therefore, the discharge time reads

$$t_{D,lin} = -\frac{rc}{\alpha} \int_0^1 (1 - \alpha x) \cdot \ln(1 - \alpha x) dx + C_L \left(-\frac{r}{\alpha} \ln(1 - \alpha) \right).$$

After simplification, we get

$$t_{D,lin} = \frac{rc}{2\alpha^2} \left((1 - \alpha)^2 \left(\ln(1 - \alpha) - \frac{1}{2} \right) + \frac{1}{2} \right) - \frac{rC_L}{\alpha} \ln(1 - \alpha). \quad (20)$$

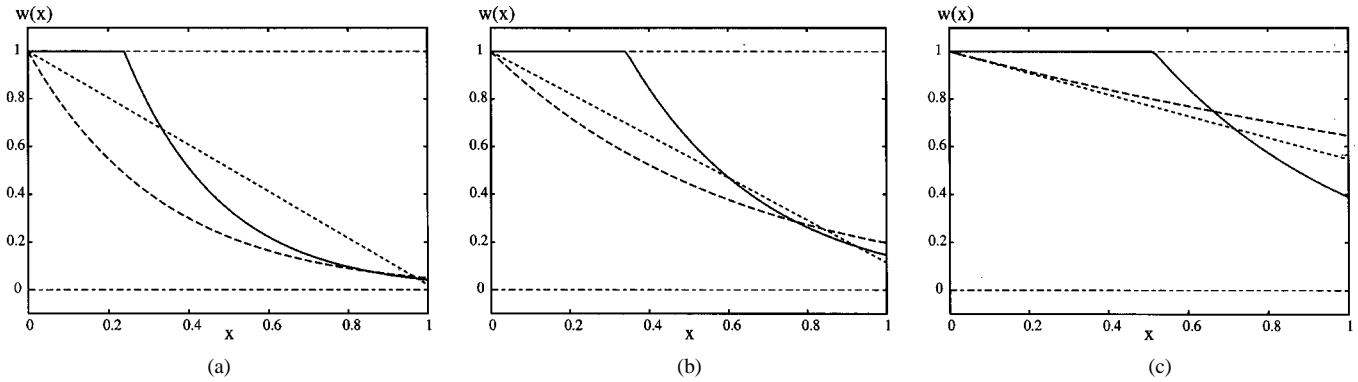


Fig. 6. FET chain width shape. (a) $C_L/c = 0.01$. (b) $C_L/c = 0.05$. (c) $C_L/c = 0.2$.

The transistor width function for an exponentially tapered transistor chain is $w(x) = e^{-\alpha x}$. The effective resistance from position x to the ground is calculated as

$$R(x) = \int_0^x r e^{\alpha\tau} d\tau = \frac{r}{\alpha} (e^{\alpha x} - 1).$$

Discharge time reads

$$t_{D,\text{exp}} = \frac{rc}{\alpha} \int_0^1 e^{-\alpha x} (e^{\alpha x} - 1) dx + C_L \frac{r}{\alpha} (e^\alpha - 1).$$

After simplification, we get

$$t_{D,\text{exp}} = \frac{rc}{\alpha} \left(1 - \frac{1 - e^{-\alpha}}{\alpha} \right) + \frac{rC_L}{\alpha} (e^\alpha - 1). \quad (21)$$

In each case, α is a parameter indicating the relative location in the one-dimensional subspace under a given tapering scheme. The minimal discharge time under each of the tapering schemes can be found by solving the equation: $\partial t_D / \partial \alpha = 0$.

Fig. 5 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 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 begins 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 actually 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.

The best FET width shapes under different tapering schemes are shown in Fig. 6 for three different C_L/c values: 0.01, 0.05, and 0.2, where the horizontal axis is x and the vertical axis is normalized FET width $w(x)$. Solid, dashed, and dotted lines

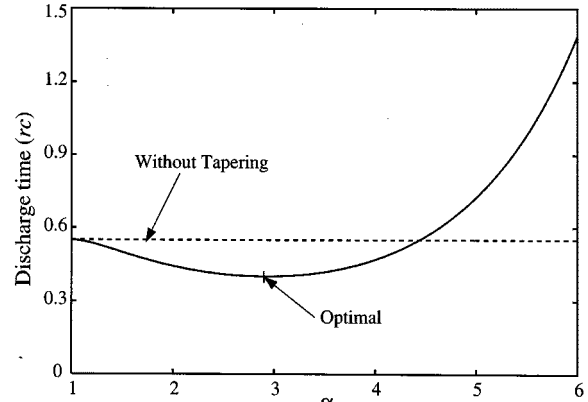


Fig. 7. Discharge time as a function of α ($C_L/c = 0.05$).

represent the width shapes under the proposed, exponential and linear tapering schemes, respectively.

Using the proposed tapering scheme, it is observed that the discharge time is not very sensitive to the parameter α . As shown in Fig. 7, there is a very flat valley region around the optimal α value. To put it more quantitatively, when the value of α lies between 2.29 and 3.53, the discharge time increase is within 5% of the optimal value, which corresponds to a α value of 2.92.

Fig. 8 depicts the range of FET width shape under the same 5%-increase assumption for three different C_L/c values. The solid lines represent optimal shapes while the dashed ones stand for the boundaries. It is also observed that the shape can have larger variation as C_L/c is large, as is the case in Fig. 8(c). When C_L/c is even larger, one can expect the dashed line marked by α_- will eventually hit the upper-right corner at a certain point. This is the case when tapering cannot generate more than 5% performance gain over nontapered FET chain and is sometimes a more practical criterion to determine whether tapering is needed. Numerically, we obtained the following equation:

$$\frac{C_L}{c} > 0.265 \quad (22)$$

which is approximately one fourth of what (17) gives.

B. Finite Number of FETs

The comparisons in the above subsection are valid only in the continuous limit. In a real circuit, the number of transistor N is a finite value. In this subsection, we discuss the effect associated with the discretization. Elmore delay model is still assumed.

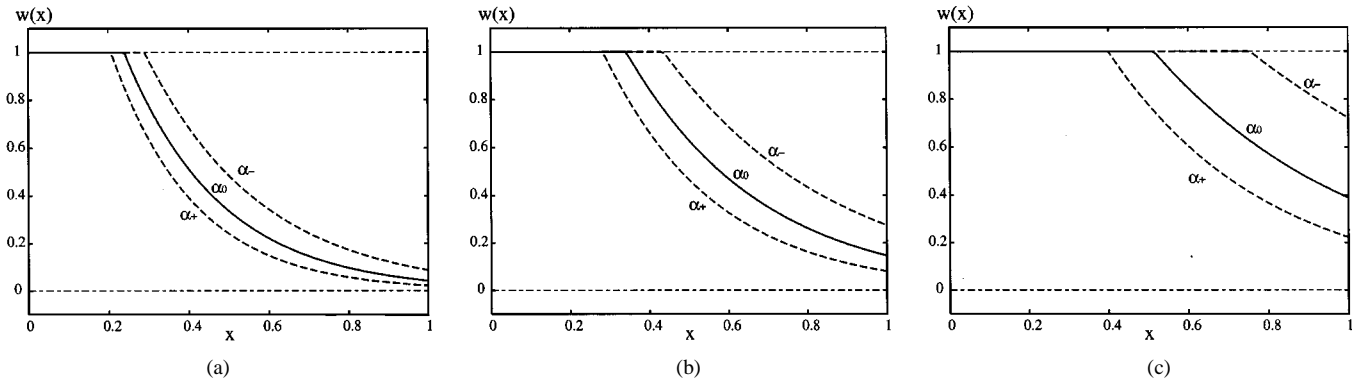


Fig. 8. Sensitivity of discharge time on the parameter α . (a) $C_L/c = 0.01$. (b) $C_L/c = 0.05$. (c) $C_L/c = 0.2$.

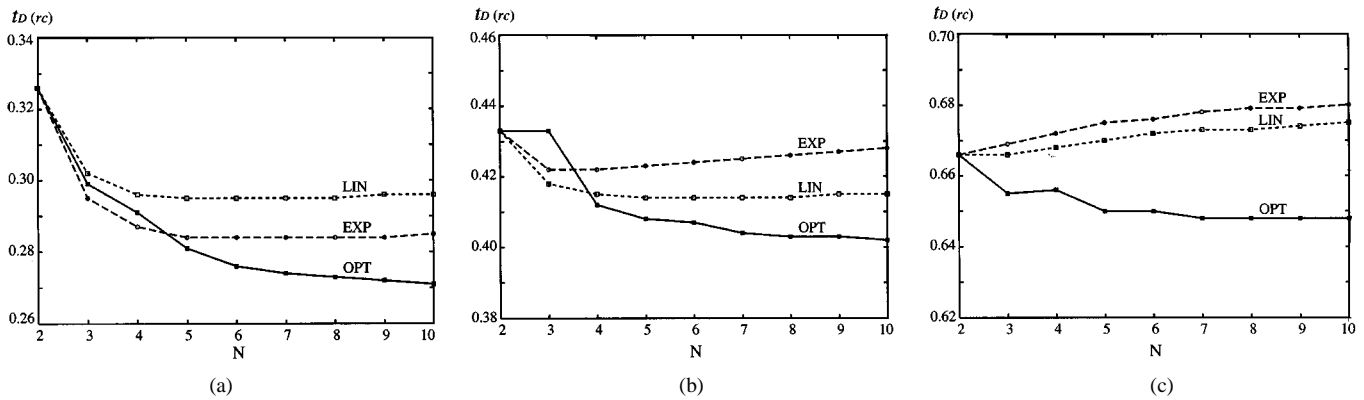


Fig. 9. Effect of discretization. (a) $C_L/c = 0.01$. (b) $C_L/c = 0.05$. (c) $C_L/c = 0.2$.

Note that the proposed tapering scheme is proved to be optimal only in the continuous limit. It is again a suboptimal solution when a FET chain has only finite number of transistors.

Fig. 9 compares the discharge times under three different load conditions: $C_L/c = 0.01, 0.05, 0.2$. Note that there is no difference among the three schemes when $N = 2$, since there is one controllable width anyway. The superiority of the proposed tapering scheme is apparent, however, when N is relatively large. From Fig. 9, one can observe that the proposed tapering scheme is worthwhile for N equals to or greater than five. An example circuit having this N value is a domino *and4* gate. For FET chains with smaller N , none of the three tapering schemes shows consistent advantage over the other two.

In the case that the maximum allowable width is very large, the value c becomes large in comparison with C_L . Therefore, the proposed tapering shape approaches traditional exponential tapering for moderate sized FET chain. This conforms to the empirical observation made by circuit designers which indicates that the exponential tapering performs well.

C. SPICE Simulation

Note that 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 sub-micron 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 if the result obtained from the simple model works adequately for real world circuits. Finite number of FETs is assumed and real FET models (HSPICE Level 49 [16]) are used. As real cases that FET chains are used in circuit design, we study a set of domino *AND* gates: *and4*, *and5*, *and6*, and *and7*. Simulation is carried out in a $0.18\text{-}\mu\text{m}$ technology at 1.6 V , 55°C and typical process.

Fig. 10(a) shows the *AND* gate implementation in domino CMOS style [15]. 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 $20\text{ }\mu\text{m}$. 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 $2\text{ }\mu\text{m}$ and NFET width of $1\text{ }\mu\text{m}$. The spacing between two adjacent transistors in the chain is set to 4λ according to MOSIS deep sub-micron rule.

To determine the parameter α , one needs to know both the load capacitance C_L and the 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 C is developed serving as a batch program which calls HSPICE for each different parameter and reads the measured delay data generated by HSPICE.

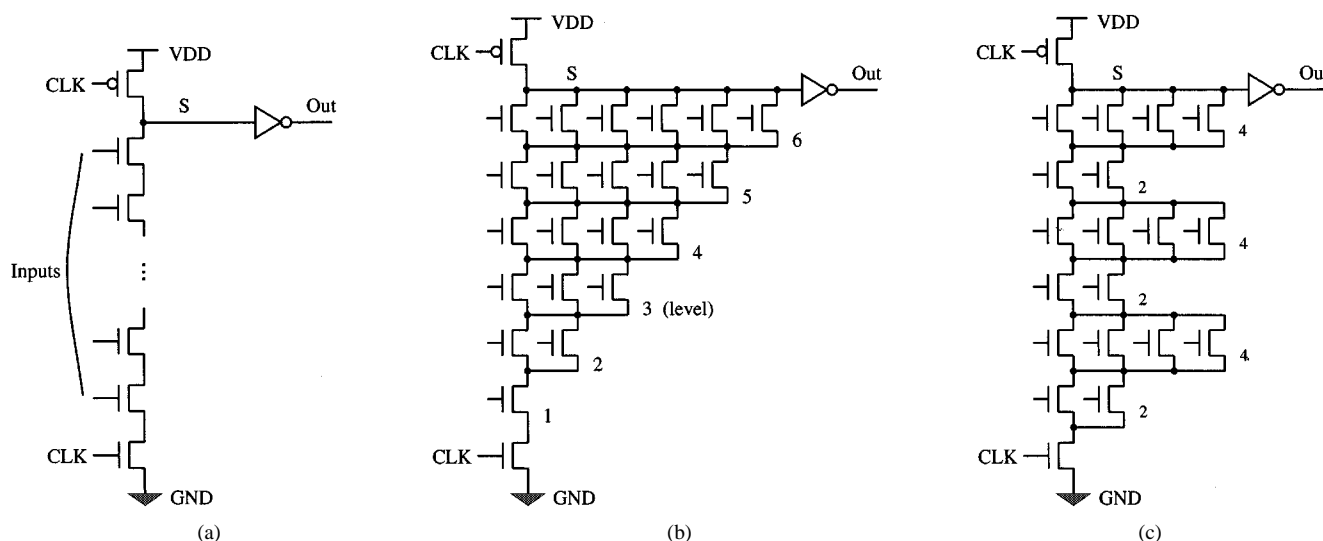


Fig. 10. Domino gates. (a) Multi-input AND gate. (b) OA654 321 gate. (c) OA424 242 gate.

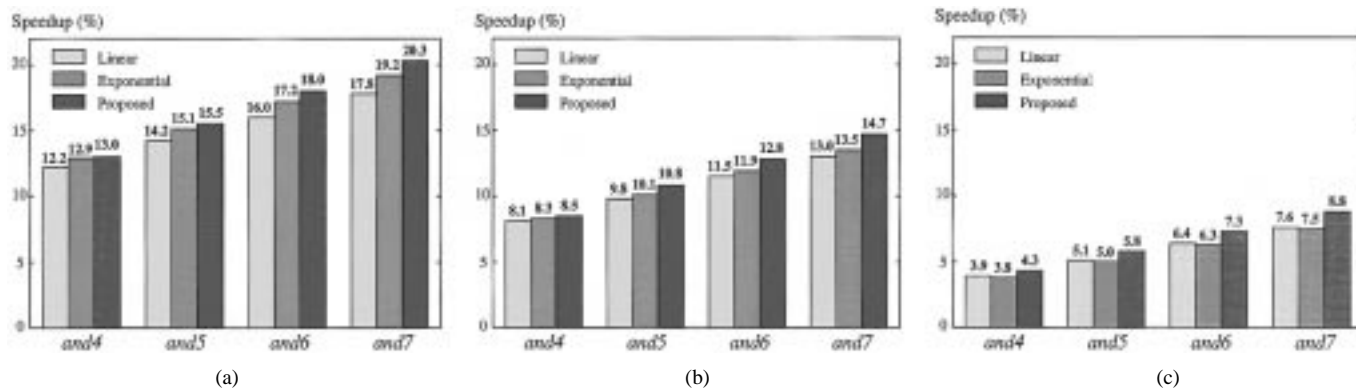


Fig. 11. Comparison of the tapering schemes by simulation. (a) 1-INV. (b) 2-INV. (c) 4-INV.

The shortest delay time and its corresponding α value are then reported.

Simulation results for the domino AND gates are shown in Fig. 11. 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. Finally, we show the optimal widths of the NFET chain in the *and7* gate under different inverter loads in Fig. 12.

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. In Fig. 10(b) and 10(c), we show two examples: domino OA654 321 and OA424 242 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 μm and we assume the size of the output inverters is 1 unit.

Simulation results on those OA gates are shown in Fig. 13. 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.

V. CONCLUSIONS

The problem of optimal tapering of a series connected FET chain is studied in this paper. In contrast to conventional approaches which try to first solve the problems for special cases when N is small, we begin with the limit when N approaches infinite. The results obtained from small N problems are not applicable 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.

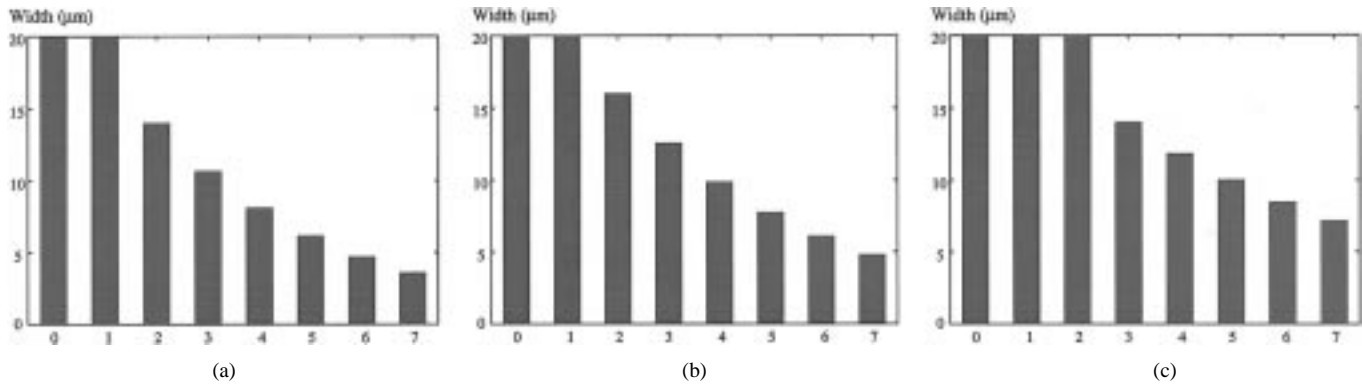


Fig. 12. Optimal transistor widths of *and7* gate under different loads. (a) 1-INV. (b) 2-INV. (c) 4-INV.

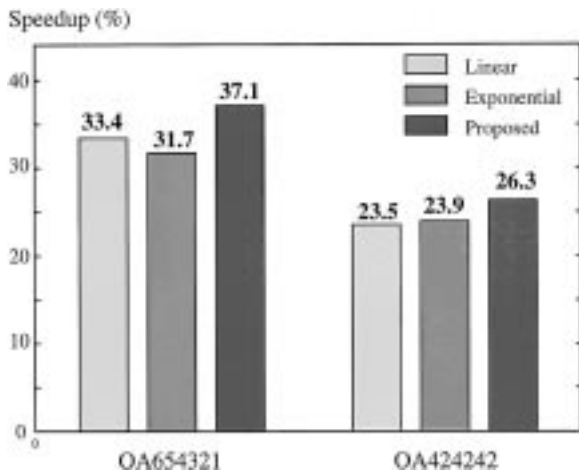


Fig. 13. Simulated results for two OR-AND gates.

Overall, the contributions of the 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; and 5) it is shown that the load capacitance has to be about a quarter of the total effective parasitic capacitance of a FET chain to make transistor tapering worthwhile for speed improvement.

In this paper, we study the original tapering problem [3] considering discharge time as the only optimization goal. In real circuit design practice, one may consider other design metrics like minimization of power consumption, charge sharing, and so on. For example, one may formulate the problem such that the power-delay product is minimized. Also, one may need to add certain penalty for FETs with very small channel width to guarantee adequate noise margin. Searching for optimal tapering shapes for multi-objectives under multi-constraints poses a challenge for future research.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their introspective comments and suggestions, which significantly improve the quality of this paper.

REFERENCES

- [1] C. Visweswariah, "Optimization techniques for high-performance digital circuits," in *Proc. IEEE Int. Conf. Computer-Aided Design*, 1997, pp. 198–205.
- [2] M. Shoji, "Electrical design of BELLMAC-32A microprocessor," in *Proc. Int. Conf. Circuits and Computers*, 1982, pp. 112–115.
- [3] —, "FET scaling in domino CMOS gates," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 1067–1071, May 1985.
- [4] B. S. Cherkauer and E. G. Friedman, "The effects of channel width tapering on the power dissipation of serially connected MOSFET's," in *Proc. IEEE Int. Symp. Circuits and Systems*, 1993, pp. 2110–2113.
- [5] —, "Channel width tapering of serially connected MOSFET's with emphasis on power dissipation," *IEEE Trans. VLSI Syst.*, vol. 2, pp. 100–114, Feb. 1994.
- [6] L. T. Wurtz, "An efficient scaling procedure for domino CMOS logic," *IEEE J. Solid-State Circuits*, vol. 28, pp. 979–982, Sept. 1993.
- [7] J. Yuan and C. Svensson, "Principle of CMOS circuit power-delay optimization with transistor sizing," in *Proc. IEEE Int. Symp. Circuits and Systems*, 1996, pp. 637–640.
- [8] S. S. Bizzan, G. A. Jullien, and W. C. Miller, "Analytical approach to sizing nFET chains," *Electron. Lett.*, vol. 28, no. 14, pp. 1334–1335, 1992.
- [9] G. A. Jullien, W. C. Miller, R. Grondin, Z. Wang, L. D. Pup, and B. Bizzan, "Woodchuck: A low-level synthesizer for dynamic pipelined DSP arithmetic logic blocks," in *Proc. IEEE Int. Symp. Circuits and Systems*, 1992, pp. 176–179.
- [10] T. Lin and C. A. Mead, "Signal Delay in General RC Networks," *IEEE Trans. Computer-Aided Design*, vol. 3, pp. 331–349, Apr. 1984.
- [11] J. Rubinstein, P. Penfield, and M. A. Horowitz, "Signal delay in RC networks," *IEEE Trans. Computer-Aided Design*, vol. CAD-2, pp. 202–211, Mar. 1983.
- [12] W. C. Elmore, "The transient analysis of damped linear networks with particular regard to wideband amplifiers," *J. Appl. Phys.*, vol. 19, no. 1, pp. 55–63, 1948.
- [13] W. S. Kimball, *Calculus of Variations*. London, U.K.: Butterworth, 1952.
- [14] M. J. Forray, *Variational Calculus in Science and Engineering*. New York: McGraw-Hill, 1968.
- [15] R. H. Krambeck, C. M. Lee, and H.-F. S. Law, "High speed compact circuits with CMOS," *IEEE J. Solid-State Circuits*, vol. SC-17, pp. 614–619, Mar. 1982.
- [16] *Star-HSPICE Manual*, Avant! Corp., Fremont, CA, 1998.



Li Ding (S'99) received the B.S. degree in physics from Peking University, Beijing, China, and the M.S. degree in electrical engineering and computer science from The University of Michigan, Ann Arbor, in 1997 and 2000, respectively. He is currently pursuing the Ph.D. degree in computer engineering at The University of Michigan.

He is a Graduate Student Research Assistant in the Electrical Engineering and Computer Science Department at The University of Michigan. His research interests include VLSI circuits and computer-aided design. He is currently working on signal integrity, reliability and other deep submicron issues.



Pinaki Mazumder (S'84-M'88-SM'94-F'99) received the B.S.E.E. degree from the Indian Institute of Science, Bangalore, India, the M.Sc. degree in computer science from The University of Alberta, Canada, and the Ph.D. degree in electrical and computer engineering from The University of Illinois at Urbana-Champaign, in 1976, 1985, and 1987, respectively.

During the summer of 1985 and 1986, he was a Member of the Technical Staff in the Indian Hill branch of AT&T Bell Laboratories. He was with Bharat Electronics Ltd. (a collaborator of RCA), India, for over six years, where he developed several types of analog and digital integrated circuits for consumer electronics products. For two years he was a Research Assistant with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. From 1996 to 1997, he spent his sabbatical leave as a visiting faculty at Stanford University, The University of California at Berkeley, and with Nippon Telephone and Telegraph, Japan. Currently, he is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, where he is leading ultrafast circuit design activities for nano and quantum electronic devices. He has successfully developed CAD tools for high-performance VLSI circuit simulation (NDR-SPICE) and numerous circuit topologies for quantum MOS and other quantum-well devices. Several U.S. and Japanese semiconductor companies including Texas Instruments, Hughes Research Laboratory, Lockheed Martin, NTT, and NEC have been collaborating with him on this research work. He is the coauthor of three books "*Testing and Testable Design of High-Density Random-Access Memories*" (Norwell, MA: Kluwer, 1996), "*Semiconductor Random-Access Memories: Testing and Reliability*" (Norwell, MA: Kluwer, 1996), and "*Genetic Algorithms for VLSI Design, Layout and Test Automation*, (Englewood Cliffs, NJ: Prentice-Hall, 1998). He has published over 160 papers and has two international patents pending.

Dr. Mazumder was a Guest Editor of the *IEEE Design and Test Magazine* special issue on multimegabit memory testing, March 1993, and the *Journal of Electronic Testing-Theory and Applications* special issue on memory testing and reliability, June 1994. He was the guest editor of two special issues on emerging nanoelectronic technologies and their applications in *IEEE TRANSACTIONS ON VLSI SYSTEMS* (December 1997) and the *PROCEEDINGS OF THE IEEE* (1998). He is on the Editorial Board of *PROCEEDINGS OF THE IEEE* and is also an Associate Editor of the *IEEE TRANSACTIONS ON VLSI SYSTEMS*. He is a member of Sigma Xi, Phi Kappa Phi, and the Association for Computing Machinery Special Interest Group on Design Automation.