Channel Width Tapering of Serially Connected MOSFET's with Emphasis on Power Dissipation

Brian S. Cherkauer, *Student Member, IEEE,* and Eby G. Friedman, *Senior Member, IEEE*

Abstract— **Transistor channel width tapering in serial MOSFET chains is shown in this paper to simultaneously decrease propagation delay, power dissipation, and physical area of VLSI circuits. Tapering is the process of decreasing the size of each MOSFET transistor width along a serial chain such that the largest transistor is connected to the power supply and the smallest is connected to the output node. A detailed explanation of the effects of tapering on the output waveform is presented with specific emphasis on the power dissipation of tapered chains. It is demonstrated that in many cases tapering decreases delay and changes the shape of the output waveform such that the time during which a load inverter is conducting short-circuit current is reduced. This decrease in short-circuit current also occurs in many cases where tapering may not offer a speed advantage. In addition, dynamic CV^2f power dissipation of the serial chain is reduced. In those circuits where tapering does not decrease propagation delay, tapering permits a designer to tradeoff speed for a reduction in both short-circuit and dynamic power dissipation, a tradeoff not normally available with untapered chains. Thus the total power consumed by a serial chain of MOSFET's, as well as its propagation delay and area, can be reduced by channel width tapering. A design system for determining when tapering is appropriate, selecting the amount of tapering, and synthesizing the physical layout is presented. Physical layout issues unique to tapering are discussed, and fabricated test structures are described.**

I. Introduction

IN ORDER to increase the performance of a CMOS circuit, integrated circuit (IC) designers apply various specialized techniques to decrease the time, area, and power required for signals to propagate through combinatorial networks. One technique is the sizing of individual transistors for minimal delay and/or power dissipation. Since transistor sizing can greatly affect circuit speed, area, and power dissipation, these factors must be considered together when determining the channel width of a transistor.

Many CMOS logic structures are composed of chains of MOSFET's serially connected between a power supply rail and the output of the subcircuit. These serially connected MOSFET's are a major source of delay and power dissipation [1], therefore, optimal sizing of these transistors is important in reducing the delay and power dissipation of these circuit structures. In this paper, the physical operation of serially connected MOSFET's is explained, and the effects of tapering, in particular the power dissipation characteristics, are categorized. A circuit model for analyzing tapered chains is developed, design and layout issues of these tapered chains are discussed, and these results are verified by fabricated test circuits.

A brief background of previous research on tapering MOS chains is provided in Section II, followed by a detailed discussion of the circuit operation of an NMOS chain in Section III. In Section IV, channel width tapering is presented and qualitatively described. Resistive and capacitive models for analytically investigating tapering are presented in Section V. The propagation delay and power dissipation characteristics of tapered serial chains are presented in Section VI. An automated design system for synthesizing tapered serial chains to exploit the improved speed and power dissipation characteristics of these circuit structures is described, and related physical layout issues are reviewed in Section VII. Fabricated test structures are discussed in Section VIII. Some conclusions are made in Section IX. Details of the derivation of the analytic models are provided in the appendices.

II. Previous Research on Tapered Serial MOSFET's

For the purposes of this investigation, only NMOS transistor chains are discussed. The theoretical background as well as the design equations may be applied equally to PMOS transistor chains, with the polarities correspondingly reversed. Fig. 1 illustrates the transistor configuration of a serial chain of $n+1$ N-channel MOSFET's with the notation presented in this figure used throughout the paper.

In typical integrated circuits, the size of the transistors in the discharge chain is constrained to have equal channel dimensions while ensuring that the circuit satisfies its design criteria for speed and area. Shoji [2]–[5] in 1982 first pointed out qualitatively that under certain circumstances (specifically, the load capacitance must be of the same order of magnitude as the parasitic drain/source capacitances between the serial transistors), this constant width approach to transistor sizing may not be optimal. He proposed using either a linear tapering of transistor aspect ratios [4], or an exponential tapering of transistor aspect ratios [3], with the largest transistor closest to ground and the smallest closest to the load (see Fig. 2). Shoji further demonstrated that it was often possible, with the proper choice of tapering factor, to produce a circuit which would discharge a capacitive load more quickly than an untapered chain and therefore provide a faster transient response.

Tapered channel widths have been successfully integrated into an automated layout system to increase circuit performance [6],[7]. In this physical synthesis tool, strings of

Manuscript received December 11, 1992; revised August 16, 1993. This work was supported by the National Science Foundation by Grant MIP-9208165.

The authors are with the Department of Electrical Engineering, University of Rochester, Rochester, NY 14627.

IEEE Log Number 9214287.

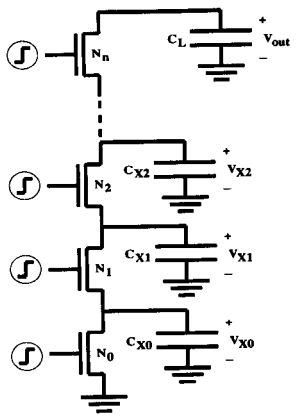

Fig. 1. Serially connected NMOS chain.

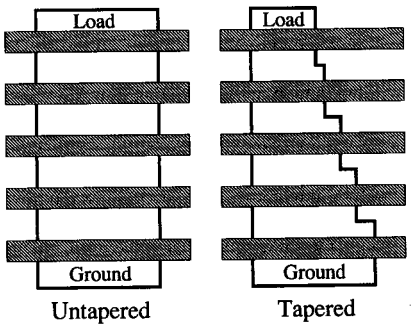

Fig. 2. Untapered and exponentially tapered MOSFET chains.

serially connected MOSFET's are automatically sized by an approximate analytical formula to generate a tapered profile. Layouts are then synthesized based on this tapered profile specification.

It is worthwhile noting that since the bottom transistor in a tapered serial chain retains its original width with tapering, the impact on the clock distribution network is minimal for those circuits which utilize Domino-based logic [8]. This permits existing Domino-based circuits to incorporate tapering, thereby minimizing power dissipation and possibly decreasing propagation delay without requiring a redesign of the system-wide clock distribution network.

III. Circuit Operation of NMOS Chain

Consider that physically the transistor chain consists not only of transistors, but also of parasitic junction capacitances at each transistor terminal. Decreasing the aspect ratio (W/L) of a transistor decreases its current drive, but also decreases these parasitic capacitances. Decreasing the current drive serves to slow the rate of discharge of the load, while decreasing the parasitic capacitance serves to increase this same rate of discharge. Thus, these two effects tend to conflict with each other.

In order to investigate the problem of transistor sizing of serially connected NMOS chains, it is first necessary to understand the behavior of the circuit. In particular, the regions of operation of the circuit must be determined.

A. Region 1

Consider the following initial conditions (see Fig. 1): $V_{\text{out}}(t = 0^-) = V_{\text{DD}}$ and $V_{Xi}(t = 0^-) = 0$, with the assumption that simultaneous step inputs are applied at the gate of each transistor at time $t = 0$. Note that these initial conditions represent the limiting worst case condition of the output transient response. While other initial conditions exhibit similar behavior with tapering [9], these initial conditions have been chosen to facilitate analysis. Under these initial conditions, the transistor closest to the output operates initially in the saturation region, then enters the linear region, while the remaining transistors operate entirely in the linear region until the capacitances are fully discharged.

This may be demonstrated as follows: at time $t = 0^+$, the gates of all of the transistors have just switched to V_{DD}, and the voltage across each internal capacitance remains at the initial value of 0 V. Therefore, every transistor has a gate-to-source voltage, V_{GS}, of V_{DD}. This assures that none of the transistors is cut off. The drain-to-source voltage, V_{DS}, of the transistor closest to the output, N_n, is equal to V_{DD}. Thus, $V_{\text{DS}} > V_{\text{GS}} - V_T$ for the topmost device, and N_n operates in the saturation region. For short-channel devices, the onset of saturation is due to velocity saturation rather than channel pinch-off, and this effect is difficult to predict analytically, relying on curve fitting to determine the precise drain saturation voltage [10]. The aforementioned discussion of saturation considers channel pinch-off induced saturation only. V_{DS} across the remaining transistors are all equal to zero, therefore no current will flow, and each of these transistors operates in the linear region.

None of the transistors which initially operate in the linear region ever enters the saturation region. This occurs since in order for a MOSFET operating in the linear region to transition into the saturation region, V_{DS} must be greater than $V_{\text{GS}} - V_T$. With the gate voltages all held constant at V_{DD}, this requires that the drain voltage of the linear transistor be raised higher than $V_{\text{DD}} - V_T$ to saturate the transistor, which, as described below, cannot occur.

To raise the drain voltage of a transistor in the serial chain, the parasitic capacitance at each drain/source node must be charged by the transistor above it in the chain. However, an NMOS transistor is only capable of charging its source terminal to within a threshold voltage of its gate voltage, at which point the transistor enters the cutoff region. Since the gate voltage is fixed at V_{DD}, the maximum voltage the source node can attain is $V_{\text{DD}} - V_T$, which is insufficient to saturate the next transistor in the chain. Also, due to the body effect, the value of V_T would be larger for each transistor farther up the chain from ground, thus moving the transistor operating in the linear region farther from the saturation boundary condition. In general, other constraints, such as the available charge, will limit the maximum drain voltage reached by each of the linear

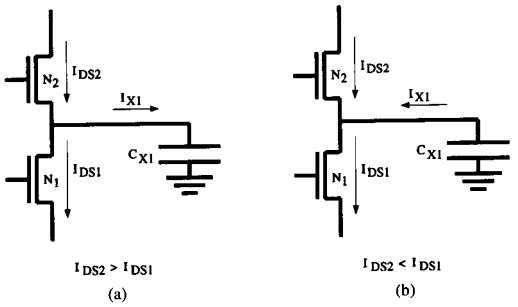

Fig. 3. Currents at a drain/source node.

transistors to a value lower than $V_{DD}-V_T$; these effects merely reinforce the inability of the lower transistors to saturate.

During the discharge cycle, the transistors within the serial chain play roles of differing importance which are governed by the magnitudes of the drain/source capacitances. Since the gate voltage is fixed at a voltage sufficiently large to allow each transistor to conduct current, the magnitude of the current is controlled by the individual transistor drain and source voltages, and hence by the charge stored on each capacitor.

Charging a capacitor occurs when the transistor above the capacitive node is biased such that it is sourcing more current than the transistor below the node is able to sink [see Fig. 3(a)]. As the voltage across the capacitor rises, it changes the V_{DS} bias on both the upper and lower transistors such that it simultaneously reduces the current sourced by the transistor above it while increasing the current sunk by the transistor below it.

Likewise, the capacitor begins to discharge when the transistor below the drain/source capacitance is biased such that it is sinking more current than the transistor above the capacitance is sourcing [see Fig. 3(b)]. As the voltage across the discharging capacitor decreases, it increases the current sourced by the transistor above it, while reducing the current sunk by the transistor below it.

Thus, these two effects act to stabilize the node voltage such that the transistors are all biased at a current equilibrium in which the transistor chain acts as a voltage divider. This is the state which Kang and Chen refer to as the "plateau voltage" [11], and the time required to reach this equilibrium state is referred to as the "initial charge distribution." The voltage across each capacitor remains at the plateau voltage as long as there is sufficient charge on the load capacitance to maintain the top transistor in saturation. This occurs since while N_n is saturated, the current sunk by the topmost device is relatively independent of its drain voltage.

B. Region 2

Once enough charge has been drained from the load capacitance to allow the saturated upper transistor N_n to pass into the linear state, the current through N_n becomes highly dependent upon the drain-to-source voltage across it, and the current begins to decrease once the drain voltage decreases. The time at which the transition between the first region of operation (saturation) and the second region (linear) occurs is defined in this paper as t_{12}. Once t_{12} is reached, the remaining capacitances begin to discharge in an attempt to maintain an equilibrium with respect to the drain currents, and this continues until all of the capacitances are fully discharged.

It is important to note that in those cases where the load capacitance C_L is of the same order of magnitude as the drain/source parasitic capacitance along the chain, the amount of charge on C_L may not be sufficient to maintain the upper transistor in saturation until the plateau voltage is reached. Since, as will be described later, this is the regime most suitable for tapering, the plateau voltages may not appear in those cases in which tapering is most applicable (i.e., where C_L is comparable to or less than the drain/source capacitance along the chain). Also, with nonsimultaneous step inputs, the initial conditions of the capacitances along the chain may not be zero, and this may prevent the top transistor from saturating. In this case, Region 1 is bypassed, and the discharge begins directly within Region 2.

During the initial phase of discharge (beginning at $t = 0^+$), the parasitic drain/source capacitors are charged by the current flowing through the saturated device. The saturated device defines the magnitude of the current for the entire chain, since the current flowing through the serial chain must all flow through N_n. As time progresses, the significance of the saturated transistor lessens to the point where all transistors become equally important. This occurs during the plateau region, or a time shortly after t_{12} if the plateau voltage is not reached. Then, the bottom transistor becomes the dominant device, as it forms the "bottleneck" through which all current must pass on its way to ground.

IV. Tapering of Serially Connected MOSFET's

The effects of tapering on the two regions of operation of a serial chain of MOSFET's are described in this section. The current in the top transistor N_n directly determines the discharge rate at the output of the serial chain through the KCL equation taken at the output node as shown below:

$$I_{DS} = -C_L \frac{d(V_{out})}{dt}. \tag{1}$$

A. Region 1

With tapering, the channel width of N_n is decreased, thereby lowering the transconductance of the device. With the saturated device having the smallest channel width of the chain, its effective on-resistance is the greatest. Since the NMOS chain acts as a voltage divider, this has the effect of lowering the maximum voltage of the source terminal of N_n. The parasitic capacitance at the source node of N_n, C_{Xn-1} (see Fig. 1), is also lowered by tapering. This allows the voltage at the source node V_{Xn-1} to increase more quickly. This, in turn, causes the saturation current to decrease more quickly, due to the rising source potential of N_n.

Assuming channel length modulation is negligible ($\lambda \approx 0$), the I–V equation for an NMOS device in saturation is given

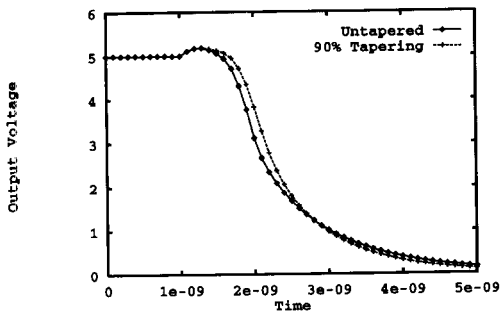

Fig. 4. Tapered versus untapered SPICE simulation.

in (2) using the alpha-power model developed by Sakurai and Newton [10],[12],

$$I_{DSat} = \frac{W}{L} P_C (V_{GS} - V_T)^m, \tag{2}$$

where W represents the channel width, L represents the effective channel length, and P_C and m are experimentally measured parameters. For long-channel devices, setting $m = 2$ and $P_C = K'/2$, with K' representing the MOS transconductance parameter, reduces (2) to the classical Shichman-Hodges saturation I–V equation [13].

It is shown experimentally herein that the effect of tapering during the time when N_n is saturated (Region 1) tends to slow the output response (see Fig. 4). That is, during Region 1, the effects of decreasing W/L are not offset by the effects of increasing $V_{GS} - V_T$ and decreasing C_{Xn-1}, and the discharge of the load is usually slower for a tapered chain during the saturation region of operation than it would be for an untapered chain. As described previously, the amount of time spent in the first region is directly proportional to the charge stored on the load capacitance. Therefore, a large load capacitance tends to lengthen the time that N_n spends in saturation, and tapering does little to improve the output response under these conditions. Tapering also shifts out the time t_{12} when N_n switches from saturation into the linear region. Due to the body effect, the voltage at the output at time t_{12} also increases.

B. Region 2

Once the chain enters Region 2, all of the transistors operate in the linear region. The drain-to-source voltage across each transistor below N_n is relatively small compared to the gate-to-source voltage ($V_{DS} \ll V_{GS}$—V_T).

Tapering has the effect of increasing the channel resistance and decreasing the junction capacitance of each successive transistor farther up the chain from ground. Note that the dominant transistor in Region 2, the transistor closest to ground, is unchanged with respect to tapering, so it maintains all its untapered ability to sink current. Also, as discussed previously, the voltages on these junction capacitances tend to decrease with tapering. Smaller capacitances with lower voltages across them imply that, with the exception of the output node, there is less charge stored on the parasitic capacitances which must be discharged, thereby speeding up the transient response within Region 2.

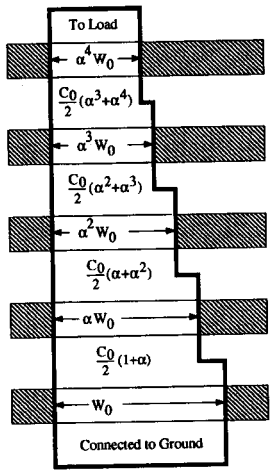

Fig. 5. Exponential tapering ($\alpha = 0.85$).

In summary, the effects of tapering are to increase the transistor on-resistance; decrease the junction capacitances; delay the transition time, t_{12}; and increase the output voltage at time t_{12}. Experimental evidence shows that for small loads, tapering tends to increase the discharge rate within Region 2, as shown in Fig. 4.

V. *RC* Models for Analyzing Tapered Serially Connected MOSFETs

Shoji states that of the two proposed tapering methods, exponential and linear tapering, exponential tapering yields higher speed circuits [3]. The primary advantage of exponential tapering over linear is that the ratio of adjacent transistor channel widths remains fixed. This is consistent with the results found in the optimal design of cascaded buffers [14],[15]. The results described in this paper have therefore concentrated on investigating channel width tapering with an exponential tapering factor. Furthermore, Bizzan *et al.* [7] recently showed that when applying an iterative, numerical optimization technique to the transistor sizing problem, the results very nearly match an exponential tapering factor.

Exponential tapering, when applied to each transistor moving up the chain from ground, decreases the transistor width by a fixed factor α with $0 < \alpha \leq 1$, and with values of α typically greater than 0.5. Therefore, with exponential tapering and assuming constant channel length, the width of the transistor closest to ground is W_0, the next transistor up the chain has a channel width of αW_0, the next transistor a width of $\alpha^2 W_0$, and so on, as shown in Fig. 5.

A. Capacitance Model

The capacitance model used in this paper assumes that the width of the diffusion implant changes as an abrupt step midway between the gates of two adjacent transistors

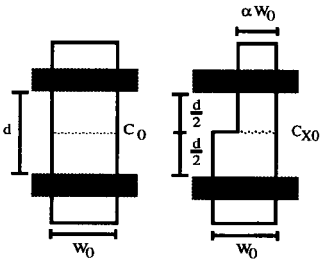

Fig. 6. Capacitance model of serially connected MOS transistors.

in the serial chain (see Fig. 6), and can be modeled as a single, voltage-independent capacitor. This approximation accurately models a transistor chain where the drain and source diffusions are adjacent, as is typical in NMOS serial chains. The capacitances of tapered and untapered drain/source nodes are given in (3) and (4), respectively, where C_{xi} is the capacitance at the drain node of transistor N_i (source node of transistor N_{i+1}), as defined in Fig. 1, where

$$C_0 = C_{j0} \cdot W_0 \cdot d + 2C_{jsw} \cdot W_0 + 2C_{jsw} \cdot d, \quad (3)$$

$$C_{Xi} = C_{j0} \cdot W_0 \cdot d\left[\frac{\alpha^{(i-1)}(1+\alpha)}{2}\right] + 2C_{jsw} \cdot W_0 \cdot \alpha^{(i-1)} + 2C_{jsw} \cdot d, \quad (4)$$

C_{j0} is the zero-bias bulk junction bottom capacitance, C_{jsw} is the zero-bias bulk junction sidewall capacitance, W_0 is the channel width of the bottom transistor, N_0, d is the length of the source/drain diffusion island between adjacent channels, and α, as was described previously, is the exponential tapering factor.

It is also assumed that the effects of tapering on the load capacitance, C_L, is negligible, as the output capacitance is dominated by the drain capacitances of any pull-up devices and by the wiring and input capacitances of the following stage. As shown in Fig. 7 and (5), in addition to all the capacitances associated with the drain of the topmost device in the chain (such as diffusion, sidewall, and overlap capacitance), C_{DN_n}, the load capacitance, C_L, is composed of all the capacitances associated with the drains of any pull-up devices, $C_{\mathrm{DP}_{\mathrm{pull-up}}}$; the interconnect (wiring) capacitance, C_{int}; and the gate capacitances of the following stage, $C_{G_{\mathrm{load}}}$. Of these, only C_{DN_n}, which is typically much smaller in magnitude than the other output load capacitances, is a function of tapering. Thus, it is a reasonable approximation to neglect the variation of the load capacitance with tapering.

$$C_L = C_{\mathrm{DN}_n} + C_{\mathrm{DP}_{\mathrm{pull-up}}} + C_{\mathrm{int}} + C_{G_{\mathrm{load}}} \quad (5)$$

B. Resistance Model

It is important to choose a model for the transistor chain which balances computational complexity with circuit accuracy. To do this, certain nonlinearities must be included in the circuit model. In both Regions 1 and 2, care must be taken to model the transistor closest to the output, N_n. In Region 1, this transistor is saturated. Therefore, the current

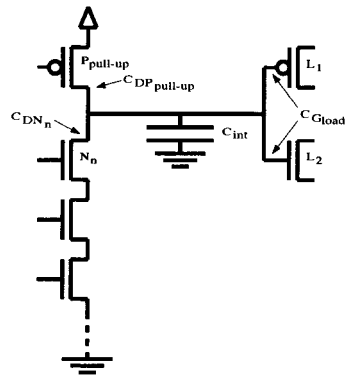

Fig. 7. Components of load capacitance.

through N_n is quadratically dependent upon the gate-to-source voltage and, assuming negligible channel length modulation (λ), independent of the drain-to-source voltage.

In Region 2, again assuming $\lambda \approx 0$, a large change in drain-to-source voltage as well as a large maximum drain-to-source voltage develops across N_n. The result is that the V_{DS}^2 term of the linear current equation, (6), is significant in this device. By comparison, the remaining transistors experience relatively small drain-to-source voltage, such that the dependence of these transistors on V_{DS} is nearly linear.

$$I_{\mathrm{DS}} = \frac{K'W}{2L}\left[2(V_{\mathrm{GS}} - V_T)V_{\mathrm{DS}} - V_{\mathrm{DS}}^2\right] \quad (6)$$

In order to predict the effects of tapering, it is best to begin by examining the node voltages, specifically, the source voltage of N_n, V_{Xn-1}. Investigation of a two transistor chain demonstrates that the computational complexity, when applying classical Shichman-Hodges [13] *I–V* equations for each of the two transistors, is extremely unwieldy, and extension of this approach to a greater number of transistors is analytically intractable. Therefore, a simpler strategy has been chosen to model the tapered serial MOS chain.

One approach is to model each transistor as a linear resistor and apply classical *RC* delay equations [4],[7]. However, this method has a significant drawback. It does not accurately model the significant nonlinear dependencies of transistor N_n. To overcome this, transistor N_n is modeled using Shichman-Hodges equations, while the remaining linear transistors are modeled as linear resistors, as shown in Fig. 8. This minimizes the intractability of the model by removing those nonlinearities which are of minor significance, thereby easing the computational complexity, while maintaining the nonlinear portion of greatest consequence.

The equation shown in (7) is used to estimate the effective resistance of each of the n linear devices. Simulation shows that for circuits where the load capacitance is of the same order of magnitude as C_0, such as in Domino circuitry [8], the resistance estimation of (7) produces 50% delay times within 5% of the values produced with the full large signal

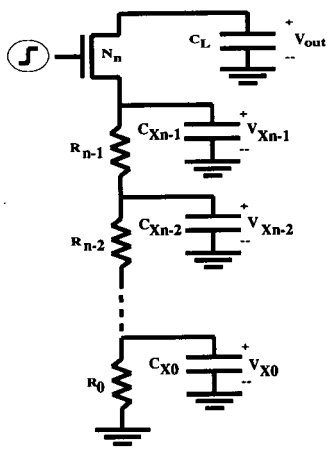

Fig. 8. Simplified model of NMOS transistor chain.

transistor model, but is analytically more tractable and requires significantly less computational time. The derivations for (7) and (8) are provided in Appendix A.

$$R_i = \frac{L}{\alpha^i K' W_0 \left[V_{DD} - V_{T0} - \frac{1}{2} V_{X(n-1)ss}\left(\frac{i+1}{n}\right)\right]} \quad (7)$$

where

$$V_{X(n-1)ss} = (V_{DD} - V_{T0})\left(1 - \frac{1}{\sqrt{n+1}}\right) \quad (8)$$

It is important to note that the delay of the serial chain is composed of two primary components, one which is dependent upon the current drive capabilities of the transistors and the parasitic drain/source capacitances, and one which represents the time required to move the transistors from the cutoff to the linear (or saturation in the case of N_n) region. The time required to turn on the devices is a function of the input waveform; however, this is not accounted for in the *RC* model described above. This delay shifts out the overall circuit delay by a fixed amount for a given input shape, and, in the authors' opinion, accounts for the difference between the *RC* delay and the SPICE delay shown in the paper by Bizzan *et al.* (Fig. 2 in [7]). Fig. 9 compares the *RC* model described in (7) with the Level-2 SPICE model for the discharge of a chain of seven serial transistors. Accuracies within 5% over the full range of operating voltage are shown.

An analytical description of this model may be found in Appendix B. Furthermore, a method for generating an approximate analytical solution to the system of equations describing the serially connected MOSFET structure is presented in Appendix C.

VI. Power Dissipation and Propagation Delay Characteristics of Tapering

Tapered circuits fall into one of three basic categories [16]. The first category of tapered circuits is comprised of those circuits where the serial chain is connected to a very small

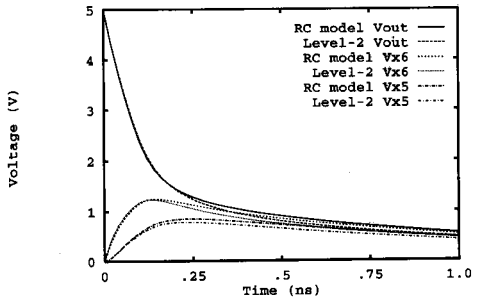

Fig. 9. Approximate *RC* model given by Fig. 8 versus Level-2 SPICE.

output load capacitance, as might be seen in Domino or Cascode Voltage Switch Logic (CVSL) [17] circuitry. The second category occurs in those circuits where the output load capacitance is somewhat larger, such as might be seen in those Domino or CVSL circuits which drive a large fan-out or in some smaller, physically close, static NAND/NOR logic gates. The third category of tapered circuits are those serial chains which are connected to a relatively large output load capacitance, as might be seen in large static NAND/NOR logic gates or circuits with long interconnect lines.

The first category encompasses the circuits which Shoji examined. This is the category of circuits for which tapering will actually reduce 50% propagation delay. In addition, in this paper it is shown that the output waveform has a shorter fall-time, translating into a reduction in short-circuit power dissipation [18] in the following stage. Dynamic power and area requirements are also both reduced as a direct result of reduced transistor geometric width. Many practical circuits may be constructed which do not fall into the first category, however, and the effects of tapering on the remaining two categories differ from that of the first.

The second category of tapered circuits represents a transition between the first and the third categories both in terms of the magnitude of the output load capacitance and the utility of tapering. In this category, the load capacitance is large enough to maintain the circuit in Region 1 long enough to delay the 50% propagation time beyond what an untapered circuit would provide. However, in Region 2, tapering provides circuits with shorter 90%-to-10% fall times. The result is that the output waveform, though shifted out in time, is more rectangular. This behavior is illustrated in Fig. 4.

It is this characteristic of tapering which is most useful for reducing short-circuit power dissipation in those circuits which fall within the second category. The time during which the output voltage of the circuit operates between $V_{DD} + V_{Tp}$ and $V_{SS} + V_{Tn}$ is reduced, which translates into a reduction in short-circuit power dissipation in the following stage. This is the region of operation in which both NMOS and PMOS devices in the following stage conduct current, allowing short-circuit current to flow from V_{DD} to V_{SS}. Coupled with the reduction in dynamic power dissipation due to the decreased parasitic drain and source capacitances and input gate capacitance, tapering becomes advantageous for those circuits where power

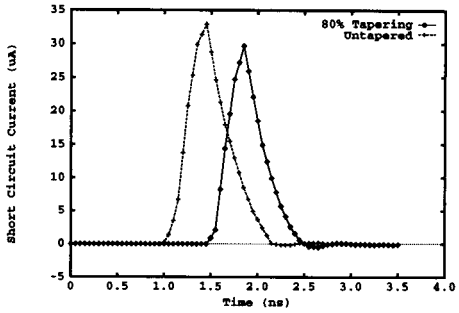

Fig. 10. Short-circuit current in an inverter following a Domino gate.

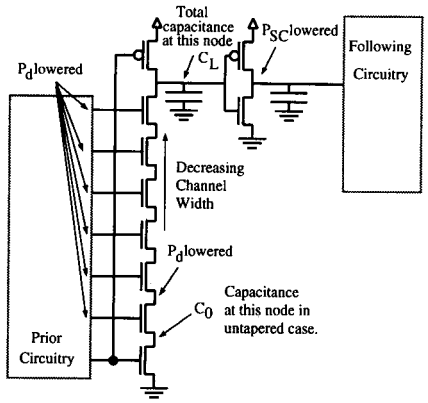

Fig. 11. Example system illustrating decreased power dissipation.

dissipation is of major concern. This reduction in dynamic and short-circuit power dissipation comes with only a minimal increase in 50% delay, a reduction in 90%-to-10% delay, and in certain cases, a reduction in overall system delay.

To illustrate this point, the short-circuit current of an example circuit belonging to the second category is shown in Fig. 10. This example circuit consists of a seven input Domino NAND gate with a minimum sized inverter at the output. Tapering the serial chain with $\alpha = 0.8$ results in a 22% reduction in energy expended through decreased short-circuit current in the load inverter. The maximum instantaneous short-circuit power dissipation in reduced by 10% as compared to an equivalent untapered serial chain.

In the third category, the load capacitance is large enough to swamp out both the delay and the short-circuit power dissipation advantages of tapering. The effects of tapering in this case are to increase both the delay and the signal fall time, which results in increased short-circuit power dissipation in the following stage. The area and dynamic power dissipation advantages remain; however, these advantages are outweighed by the disadvantages of decreased speed and significantly increased short-circuit power dissipation. Tapering circuits which belong to this third category is not recommended.

Fig. 11 depicts a typical CMOS Domino logic configuration which contains a serial chain of MOSFET's. The figure

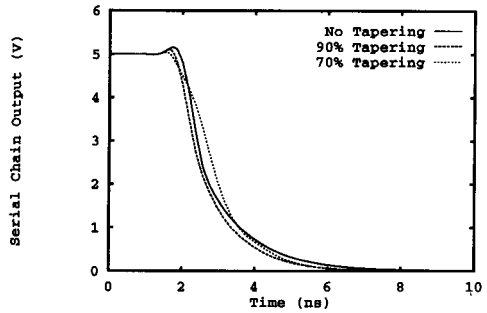

Fig. 12. Output waveform of Category 1. $C_L/C_0 = 0.87$.

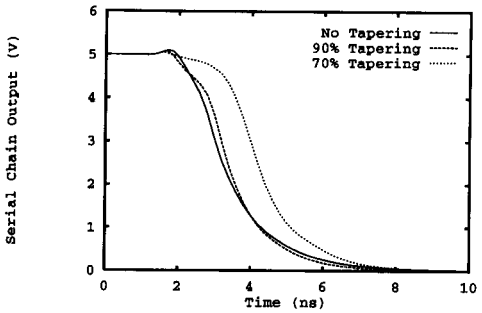

Fig. 13. Output waveform of Category 2. $C_L/C_0 = 1.2$.

illustrates those areas of the circuit in which power dissipation is decreased by tapering. Dynamic power dissipation, P_d, is smaller in the prior circuitry due to the decreased gate capacitance of the tapered serial chain. Likewise, dynamic power dissipation is reduced within the chain due to smaller source/drain parasitic capacitances in the tapered chain. Short circuit power, P_{SC}, is lowered in the load inverter for Category 1 and 2 circuits due to the change in shape of the output waveform of the serial chain, as discussed earlier in this section. The data presented in the remainder of this section quantify the effects of tapering on an example circuit, such as is shown in Fig. 11.

A. Propagation Delay

Figs. 12–14 depict the output waveforms for a seven input Domino NAND gate for each of the aforementioned three categories for three different values of α. Only the load capacitance at the output of the serial chain is varied so as to analyze the behavior of the circuit within each category.

As evidenced by Fig. 12, tapering can decrease delay in Category 1 circuits. It is important to note that there exists an optimal α beyond which tapering no longer decreases propagation delay, but actually increases it. This is shown both in Fig. 12 and in Table I, in which the 50% propagation delays from Figs. 12–14 are shown normalized to the delay of the untapered chains. Note that no improvement in delay through tapering is exhibited in Category 2 and 3 circuits.

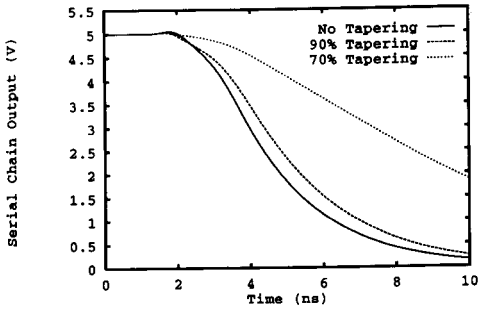

Fig. 14. Output waveform of Category 3. $C_L/C_0 = 5.0$.

TABLE I
PROPAGATION DELAY

Category	Capacitance Ratio	α = 1.0	α = 0.9	α = 0.7
1	C_L/C_0 = 0.87	100%	93%	116%
2	C_L/C_0 = 1.2	100%	107%	145%
3	C_L/C_0 = 5.0	100%	113%	221%

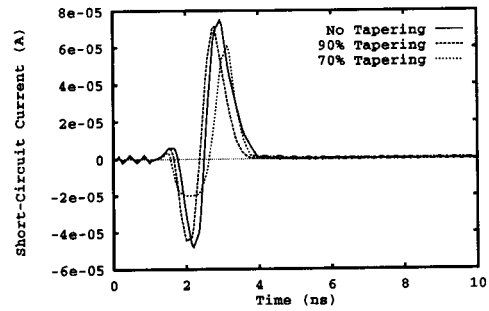

Fig. 15. Short-circuit current in stage following Category 1. $C_L/C_0 = 0.87$.

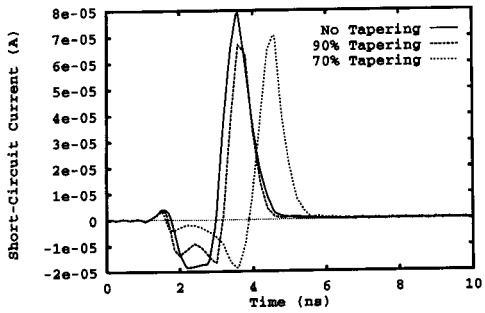

Fig. 16. Short-circuit current in stage following Category 2. $C_L/C_0 = 1.2$.

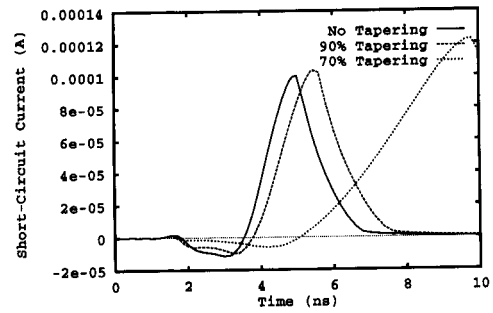

Fig. 17. Short-circuit current in stage following Category 3. $C_L/C_0 = 5.0$.

TABLE II
SHORT-CIRCUIT POWER DISSIPATION

Category	Capacitance Ratio	α = 1.0	α = 0.9	α = 0.7
1	C_L/C_0 = 0.87	100%	74%	59%
2	C_L/C_0 = 1.2	100%	85%	86%
3	C_L/C_0 = 5.0	100%	103%	122%

B. Short-Circuit Power Dissipation

Figs. 15–17 illustrate the short-circuit current of the circuit shown in Fig. 11 and, in particular, the current conducted in the stage following the serial MOSFET chain. Short-circuit current is determined for each of the cases illustrated in Figs. 12–14. Note that the negative short-circuit current is due to the Miller effect which is exacerbated by the small load capacitances of the Category 1 and 2 circuits.

In Table II, the short-circuit power dissipation depicted in Figs. 15–17 is compared. Note that the data in the table are normalized to the untapered case ($\alpha = 1.0$). A reduction in short-circuit power dissipation is shown for both Category 1 and 2 circuits. An increase in short-circuit power dissipation is apparent in the Category 3 circuit because the large load capacitance delays the transient response of the serial chain, which swamps out any positive effects of tapering.

C. Dynamic Power

Since dynamic power dissipation is directly proportional to the size of the load capacitance being charged and discharged, tapering, since it decreases this capacitance, lowers this power dissipation component. The ratio of dynamic CV^2f power dissipation originating from the input gate capacitance of an $n+1$ transistor tapered serial chain as compared to an untapered chain is shown in (9). A graph of dynamic power dissipation versus tapering factor, α, for different numbers of serially connected transistors is shown in Fig. 18. As can be seen, as n increases and α decreases, the ratio of tapered to untapered dynamic power dissipation decreases significantly.

$$\frac{P_{d\ \text{Tapered}}}{P_{d\ \text{Untapered}}} = \frac{1}{n+1}\sum_{i=0}^{n}\alpha^i \tag{9}$$

D. Total Power Dissipation

The effects of tapering on the total power dissipation are determined from the circuit of Fig. 11 and depicted in Figs. 19–21, which show the total power dissipated in the system during the high-to-low transition. These results include both the dynamic and the short-circuit power dissipation of the

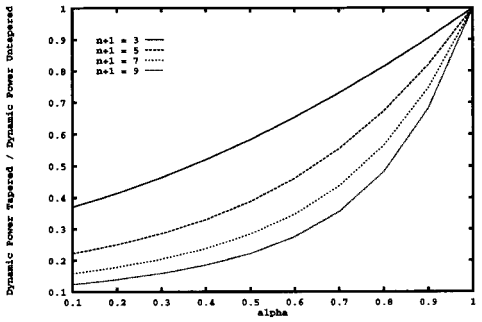

Fig. 18. Ratio of dynamic power in tapered versus untapered serial chains.

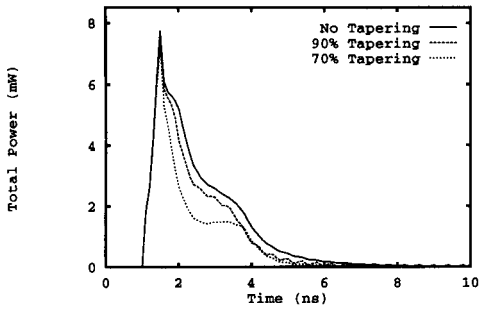

Fig. 19. Total power dissipation of high-to-low transition for Category 1. $C_L/C_0 = 0.87$.

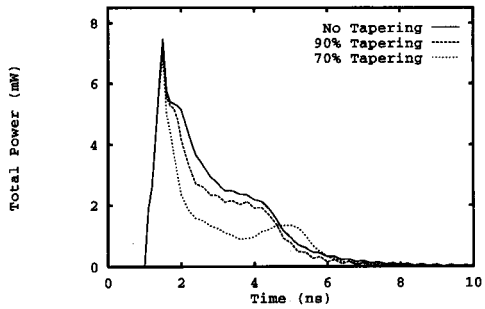

Fig. 20. Total power dissipation of high-to-low transition for Category 2. $C_L/C_0 = 1.2$

serial chain, the input circuitry driving the serial chain, and the inverter loading the serial chain.

In Table III, the average total power dissipation depicted in Figs. 19–21 is tabulated and normalized to the untapered cases. As can be seen from Table III and Figs. 19–21, channel width tapering decreases the total power dissipation even in those cases where short-circuit power dissipation increases. As is shown in the plots of total power dissipation (Figs. 19–21), when the inputs switch there is an initial spike of current due to the charging of all the gates in the MOSFET chain. The peak value of this current spike does not vary

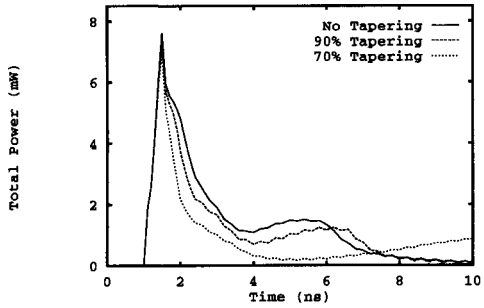

Fig. 21. Total power dissipation of high-to-low transition for Category 3. $C_L/C_0 = 5.0$

TABLE III
AVERAGE POWER DISSIPATION

Category	Capacitance Ratio	$\alpha = 1.0$	$\alpha = 0.9$	$\alpha = 0.7$
1	$C_L/C_0 = 0.87$	100%	84%	67%
2	$C_L/C_0 = 1.2$	100%	87%	65%
3	$C_L/C_0 = 5.0$	100%	88%	79%

significantly with tapering because it is mainly determined by the saturated current sourced by the inverters driving the MOSFET chain (the prior circuitry in Fig. 11). The magnitude of this saturated current only depends upon the output voltage, assuming non-negligible channel length modulation. Since the gate capacitances are smaller with tapering, the gates charge to their final voltage values more quickly, thereby decreasing the drain-to-source voltages of the transistors in the driving inverters. If channel length modulation is assumed, the saturation current is slightly reduced compared to the untapered case, which accounts for the small reduction in maximum power dissipation. In the example plotted in Figs. 12–17 and 19–21, a value of $\lambda = 0.035\ \mathrm{V}^{-1}$ has been assumed.

As the pull-up devices in the inverters enter the linear region, their currents become highly dependent upon the drain-to-source voltage across the P-channel devices. Due to this dependence, the difference in total power dissipation between the tapered and untapered chains becomes more pronounced as the currents in the inverters driving the tapered chains decrease. The difference in dynamic power dissipation due to decreased gate capacitance is most apparent in this region, which, for this example, occurs approximately when 2 ns $< t <$ 3 ns in Figs. 19–21.

As the voltage across the load capacitance is further discharged through the serial chain, a second hump is noted in the total power dissipation curves. This second hump is due to the power dissipated when the inverter loading the serial chain switches. As can be seen from the total power versus time plots, as the load capacitance is increased, the second hump is delayed, corresponding to increased delay of the serial chain. In these circuit configurations, the dynamic power dissipation component of the total power dominates the short-circuit power dissipation component.

VII. Automation of Channel Width Tapering with Layout Considerations

In order to exploit the speed and power dissipation characteristics of tapering, a method for determining the applicability and the amount of tapering must be devised. In this section, a simple automated design system for investigating these design tradeoffs is presented, and layout issues unique to tapered serial chains are discussed.

Using the C_L/C_0 guideline described in the previous section, the tapering category of a specific serial chain must initially be determined. If the circuit falls into a category such that tapering is beneficial, an appropriate value of α must be determined. If speed is the only criterion of concern, a linear resistive model of the transistors in the serial chain could be applied to determine a near-optimal tapering factor [6], [7], [19] through the use of *RC* delay approximations. However, an *RC* delay model provides no information about the shape of the discharge waveform. Thus, it is unable to predict variations in short-circuit power dissipation. For this reason, the interactive design system described in this section uses SPICE [20] in order to accurately estimate the effects of channel width tapering on power dissipation [21], [22] as well as to provide timing information. Alternatively, a power estimation tool such as described in [23] could be used. This would reduce simulation time while incurring only a slight loss of accuracy compared to SPICE simulation.

The selection of the tapering factor is performed by automatically sweeping the tapering factor over a small range of α (typically $0.7 \leq \alpha \leq 1.0$). SPICE circuit simulation files are generated and analyzed for an appropriate application-specific tapering factor. A block diagram of the design system is shown in Fig. 22. As this method sweeps the tapering factor while searching for the optimal α, the range and step size necessary for a time efficient, yet accurate search must be considered. Physical fabrication limitations constrain the minimum step size and range of the tapering factor, thereby ensuring that the design space is small. The two primary constraints are the minimum transistor dimensions and the minimum resolution of the optical reticles. Minimum transistor dimensions establish a lower limit on α beyond which design rules would be violated. Similarly, the minimum resolution of the reticles limits the minimum variation in α which is physically realizable. This leads to a lower limit on step size. Thus, the search space is kept small, allowing the search procedure to be sufficiently accurate without incurring a significant penalty in search time. The range and step size may either be specified by the designer or determined automatically based on minimum transistor dimensions and reticle resolution.

As the relative importance of speed and power dissipation greatly depends upon the application, selection of the proper tapering factor is done interactively by the designer. Once the proper tapering factor is chosen, a layout of the serial chain is automatically generated, as exemplified by Fig. 23. Currently, the physical layout is generated in Magic database format conforming to MOSIS scalable CMOS design rules.

It should be noted that tapering may cause design rule checking programs to highlight a possible error based on the violation of a minimum spacing rule between a polysilicon and a diffusion layer. This warning is intended to prevent the unintentional creation of transistors through process misalignment in those places where polysilicon and diffusion are in close proximity. In the case of tapering, these misalignments can cause only slight variations in the drain/source capacitance and, in extreme cases, slight variations in the effective W/L ratio of the transistors. However, no violation is possible since the polysilicon gate remains completely overlapped across the entire diffusion island, thereby maintaining the original transistor operation. Since misalignment occurs globally on a layer, all transistors along the chain are affected proportionately, and therefore the tapered transistor behavior is preserved. Hence, this design rule violation does not apply in the case of tapering.

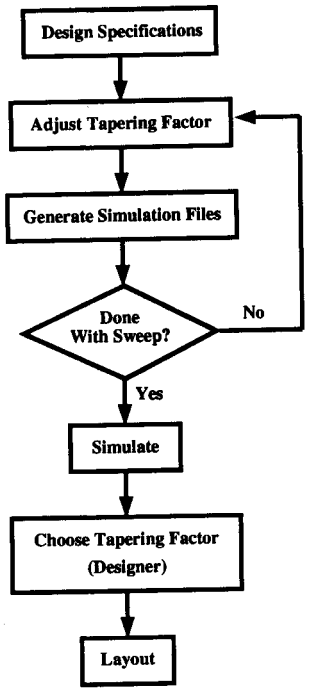

Fig. 22. Program flow of the automated tapering system.

The minimum polysilicon overhang should be increased based on the maximum misalignment which may occur in the specific process technology. In order to guarantee that this overhang is not violated by misalignment, the overhang should be determined from the edge of the widest portion of the drain/source implant of each transistor. These design rule issues are illustrated in Fig. 24.

VIII. Fabricated Test Structures

A simple test chip containing tapered serial chains fabricated using structures produced by a prototype of this tapered design system was developed to verify the waveform characteristics of the tapered chains. An example of the test circuits is the

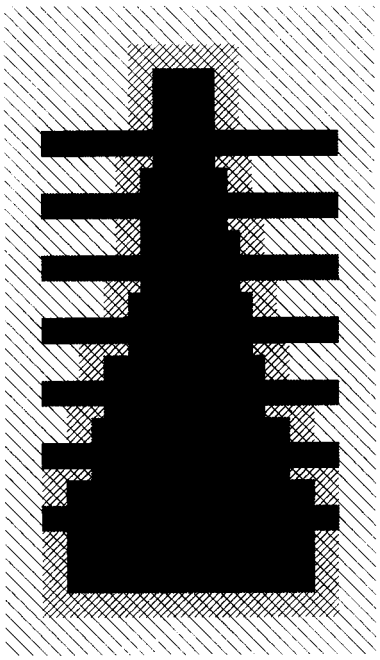

Fig. 23. Example of automated layout of tapered serial chain with $\alpha = 0.8$.

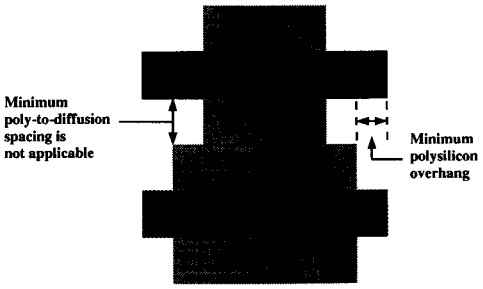

Fig. 24. Layout issues with tapering.

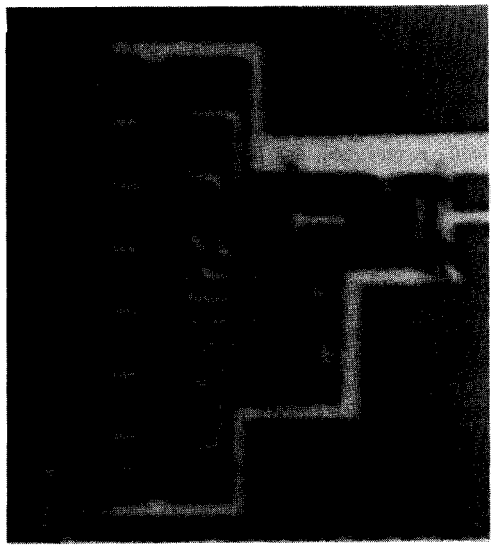

Fig. 25. Photomicrograph of fabricated test structures with $\alpha = 0.7$.

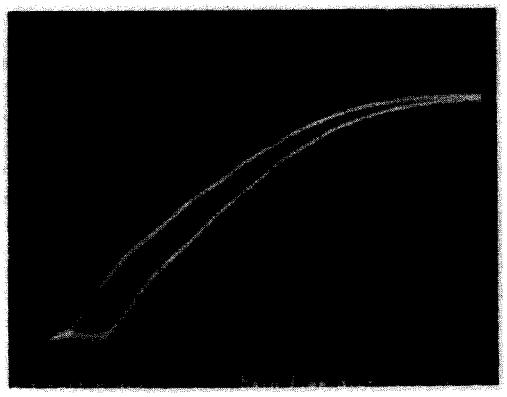

Fig. 26. Oscilloscope photograph of tapered (α = 0.9) versus untapered Category 2 circuits.

structure shown in Fig. 25, which has a tapering factor of $\alpha = 0.7$. The tapered circuit structures were fabricated using an Orbit Semiconductor 2 μm double level metal, double polysilicon P-well CMOS process and are fully functional.

Fig. 26 shows an oscilloscope photograph of the output traces from two otherwise identical test structures, one containing a tapered ($\alpha = 0.9$) serial chain and one containing an untapered serial chain ($\alpha = 1.0$). This photograph illustrates the phenomena typical of a Category 2 circuit. The delay is somewhat increased in the tapered structure over that of the untapered structure. However, the slope of the output is also increased over that of the untapered structure, leading to decreased short-circuit power dissipation in the following stage, as was previously described in Section VI of this paper. Note that the buffering of the output of the serial chain is responsible for the reversed polarity of the transition shown in Fig. 26.

IX. Conclusion

This paper provides a detailed explanation of the effects of channel width tapering on a chain of serially connected MOSFET's, a common structure in CMOS-based VLSI circuits. Tapering is shown to decrease area and dynamic power dissipation under all conditions and to decrease propagation delay and short-circuit power dissipation under certain conditions. The propagation delay of a tapered serial chain operating within Category 1 was shown to decrease by 7% with a tapering factor of α = 0.9 applied to the example Domino circuit shown in Fig. 11. Short-circuit power dissipation can

be decreased by tapering a Category 1 or Category 2 circuit, and reductions of 15% and 26% with $\alpha = 0.9$ for Category 1 and Category 2 circuit, respectively, were shown for example circuits. A reduction of over 10% in average total power dissipation is demonstrated for all three categories with $\alpha = 0.9$ and over 20% with $\alpha = 0.7$. Thus, channel width tapering is shown to be useful in those systems where load capacitance is of the same order of magnitude or less than the parasitic drain/source capacitance of the serially connected MOSFET's, such as in Domino logic, and in those circuits where power dissipation is of primary concern.

An analytic model has been developed which reduces the computational complexity of the serial chain while retaining the ability to accurately account for the topmost transistor being saturated. Physical layout issues have been addressed, and an automated design system for determining the appropriate tapering factor and synthesizing the physical layout of channel width tapered serial chains has been described. This capability permits in IC designer to more easily weigh the tradeoffs between power and speed in serial chains. Finally, a test chip containing tapered MOSFET circuits designed with this system is described, and oscilloscope data are presented which display the change in the output waveform shape of Category 2 circuits, illustrating how short-circuit current is decreased.

Appendix A: Resistance Derivation

The first step in determining the effective resistance of each of the serially connected linear transistors is to approximate the linear transistors below the topmost output device, which is initially in saturation, as a single transistor with

$$\beta = \frac{\beta_n}{n}, \tag{A.1}$$

where n represents the number of linear transistors [11], and where

$$\beta_i = K' \frac{W_i}{L} \tag{A.2}$$

represents the transconductance of the i^{th} transistor using the numbering convention defined in Fig. 1.

With this two-transistor approximation and assuming negligible body effect and channel length modulation, the maximum attainable voltage for the source of the output device, $V_{X(n-1)\text{SS}}$ (the "plateau voltage"), may be calculated by setting the currents in the two devices equal.

$$\frac{\beta_n}{2}(V_{\text{DD}} - V_{T0} - V_{X(n-1)\text{SS}})^2 = \frac{\beta_n}{n}\left[(V_{\text{DD}} - V_{T0})V_{X(n-1)\text{SS}} - \frac{V_{X(n-1)\text{SS}}^2}{2}\right] \tag{A.3}$$

From (A.3), the plateau voltage is

$$V_{X(n-1)\text{SS}} = (V_{\text{DD}} - V_{T0})\left(1 - \frac{1}{\sqrt{n+1}}\right). \tag{A.4}$$

The plateau voltages for each of the drain/source voltages in the complete chain may be approximated as equally spaced voltages below $V_{X(n-1)\text{SS}}$, and the average voltages of the drain/source nodes may be approximated as half the plateau voltage. Assuming the effects of the quadratic term in the linear current equation are minimal since the drain-to-source voltage is small compared to the gate-to-source voltage, the average resistance may be approximated as

$$R_i \approx \frac{1}{\beta_i\left[V_{\text{DD}} - V_{T0} - \frac{1}{2}V_{X(n-1)\text{SS}}\left(\frac{i+1}{n}\right)\right]}. \tag{A.5}$$

Substituting (A.2) and the tapered expression for channel width given by

$$W_i = \alpha^i W_0 \tag{A.6}$$

into (A.5), the tapered resistance model becomes

$$R_i = \frac{L}{\alpha^i K' W_0\left[V_{\text{DD}} - V_{T0} - \frac{1}{2}V_{X(n-1)\text{SS}}\left(\frac{i+1}{n}\right)\right]}. \tag{A.7}$$

Experimental evidence shows that although the model neglects the variation of $V_{X(n-1)\text{SS}}$ with α, sufficient accuracy is obtained without including this effect. As mentioned in Section V, close agreement (within 5%) over the range of interest is shown with these approximations.

Appendix B: Analytical Model

In this appendix, the analytical model for the circuit structure shown in Fig. 8 is described. The model is generated with the Shichman-Hodges saturation *I–V* equation for Region 1 and linear *I–V* equation for Region 2 by applying the KCL equation at the output of the serial chain.

A. Matrix Equations

The body effect on the threshold voltage can be approximated by a two term Taylor series expansion around $V_{\text{SB}} = 1\text{V}$, as shown below [24],

$$V_T \approx V_{T0} + \gamma\sqrt{V_{\text{SB}}} \approx V_{T0} + \frac{\gamma}{2} + \frac{\gamma}{2}V_{\text{SB}}, \tag{B.1}$$

with γ representing the bulk threshold parameter. The approximation is necessary to allow the *I–V* equations to take polynomial form of order two. The expansion was done around one volt for two reasons. The first reason is that an expansion around $V_{\text{SB}} = 0$ V would have a zero radius of convergence, whereas an expansion at one volt converges. Second, it is demonstrated with circuit simulation that the source voltage (V_{Xn-1}) varies from zero to approximately 3 V in a 5 V system, so the error in expanding around 1 V is small.

This model has two specific regions of operation, as does the actual circuit under most initial conditions. If the initial conditions of the drain/source capacitance are such that the topmost transistor, N_n, never saturates, then the Region 1 equations become unnecessary, and analysis should begin directly with Region 2. In Regions 1 and 2, the node equations of an $n+1$ transistor serial chain form the systems described

in (B.2) and (B.3), respectively. The constants are provided below.

$$\begin{bmatrix} \dot{V}_{Xn-1} \\ \dot{V}_{Xn-2} \\ \dot{V}_{Xn-3} \\ \dot{V}_{Xn-4} \\ \cdot \\ \cdot \\ \cdot \\ \dot{V}_{X1} \\ \dot{V}_{X0} \end{bmatrix} = \begin{bmatrix} a_{n-1}V_{Xn-1}^2 \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix}$$

$$+ \begin{bmatrix} b_{n-1} & c_{n-1} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ a_{n-2} & b_{n-2} & c_{n-2} & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & a_{n-3} & b_{n-3} & c_{n-3} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & a_{n-4} & b_{n-4} & c_{n-4} & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & a_1 & b_1 & c_1 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & a_0 & b_0 \end{bmatrix}$$

$$\times \begin{bmatrix} V_{Xn-1} \\ V_{Xn-2} \\ V_{Xn-3} \\ V_{Xn-4} \\ \cdot \\ \cdot \\ \cdot \\ V_{X1} \\ V_{X0} \end{bmatrix} + \begin{bmatrix} K \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix} \tag{B.2}$$

$$\begin{bmatrix} \dot{V}_{\text{out}} \\ \dot{V}_{Xn-1} \\ \dot{V}_{Xn-2} \\ \dot{V}_{Xn-3} \\ \cdot \\ \cdot \\ \cdot \\ \dot{V}_{X1} \\ \dot{V}_{X0} \end{bmatrix} = \begin{bmatrix} D_n V_{\text{out}}^2 + E_n V_{Xn-1}^2 + F_n V_{\text{out}} V_{Xn-1} \\ D_{n-1} V_{\text{out}}^2 + E_{n-1} V_{Xn-1}^2 + F_{n-1} V_{\text{out}} V_{Xn-1} \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix}$$

$$+ \begin{bmatrix} y_n & z_n & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ x_{n-1} & y_{n-1} & z_{n-1} & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & x_{n-2} & y_{n-2} & z_{n-2} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & x_{n-3} & y_{n-3} & z_{n-3} & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & x_1 & y_1 & z_1 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & x_0 & y_0 \end{bmatrix}$$

$$\times \begin{bmatrix} V_{\text{out}} \\ V_{Xn-1} \\ V_{Xn-2} \\ V_{Xn-3} \\ \cdot \\ \cdot \\ \cdot \\ V_{X1} \\ V_{X0} \end{bmatrix} \tag{B.3}$$

Solving these systems of nonlinear differential equations to compute the propagation delay requires the endpoint voltages of Region 1 at t_{12} in order to establish the initial conditions for the solution of Region 2. In Region 2, a solution for V_{out} at the 50% voltage $\left(\frac{V_{\text{DD}}}{2}\right)$ is necessary, and this requires a solution for V_{Xn-1} as well.

An approximate analytical solution to the above system of equations is possible. One such solution may be generated using the perturbation method [25]. Details of this method are provided in Appendix C. However, this method does become rather cumbersome as the number of serial transistors increases.

B. Matrix Constants

Given that

$$\beta_0 = k' \frac{W_0}{L}, \tag{B.4}$$

$$\beta_i = k' \frac{W_i}{L} = \alpha^i \beta_0, \tag{B.5}$$

and C_{Xi} as defined by (4), (B.6)–(B.9) describe the constants in (B.2),

$$a_i = \begin{cases} \frac{\beta_0}{8C_{Xn-1}}(2+\gamma)^2 & i = n-1 \\ \frac{1}{C_{Xi}} & 0 \le i \le n-2 \end{cases} \tag{B.6}$$

$$b_i = \begin{cases} \frac{\beta_0(2+\gamma)[2(V_{\text{DD}}-V_{T0})-\gamma]}{4C_{Xn-1}} - \frac{1}{C_{Xn-1}R_{n-1}} & i = n-1 \\ \frac{-1}{C_{Xi}R_{i+1}} - \frac{1}{C_{Xi}R_i} & 0 \le i \le n-2 \end{cases} \tag{B.7}$$

$$c_i = \begin{cases} \frac{1}{C_{Xi}R_i} & 1 \le i \le (n-1) \\ 0 & i = 0 \end{cases} \tag{B.8}$$

$$K = \frac{\beta_0[2(V_{\text{DD}} - V_{T0}) - \gamma]^2}{8C_{Xn-1}} \tag{B.9}$$

while (B.10)–(B.16) describe the constants in (B.3),

$$D_0 = \frac{\beta_0}{2C_L} \tag{B.10}$$

$$D_1 = -\frac{\beta_0}{2C_{Xn-1}} \tag{B.11}$$

$$E_0 = D_0(\gamma - 1) \tag{B.12}$$

$$E_1 = D_1(\gamma - 1) \tag{B.13}$$

$$x_i = \begin{cases} \frac{\beta_0}{2C_{Xn-1}}(\gamma + 2(V_{\text{DD}} - V_{T0})) & i = n-1 \\ \frac{1}{R_{i+1}C_{Xi}} & 0 \le i \le n-2 \end{cases} \tag{B.14}$$

$$y_i = \begin{cases} -\frac{\beta_0}{2C_L}(\gamma + 2(V_{\text{DD}} - V_{T0})) & i = n \\ -\frac{\beta_0}{2C_{Xn-1}}(\gamma + 2(V_{\text{DD}} - V_{T0})) - \frac{1}{R_{n-1}C_{Xn-1}} & i = n-1 \\ -\frac{1}{R_{i+1}C_{Xi}} - \frac{1}{R_iC_{Xi}} & 0 \le i \le n-2 \end{cases} \tag{B.15}$$

$$z_i = \begin{cases} \frac{\beta_0}{2C_L}(\gamma + 2(V_{\text{DD}} - V_{T0})) & i = n \\ \frac{1}{R_iC_{Xi}} & 1 \le i \le (n-1) \end{cases} \tag{B.16}$$

Appendix C: Perturbation Solution

The first step in applying the iterative perturbation solution [25] to (B.2) is to write the right-hand portion of the equation as the sum of a linear function and a nonlinear function. A small dimensionless parameter μ is then introduced as a

coefficient of the nonlinear function. The resulting matrix is shown in (C.1).

$$\begin{bmatrix} \dot{V}_{Xn-1} \\ \dot{V}_{Xn-2} \\ \dot{V}_{Xn-3} \\ \dot{V}_{Xn-4} \\ \cdot \\ \cdot \\ \dot{V}_{X1} \\ \dot{V}_{X_0} \end{bmatrix} = \mu \begin{bmatrix} a_{n-1}V_{Xn-1}^2 \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix}$$

$$+ \begin{bmatrix} b_{n-1} & c_{n-1} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ a_{n-2} & b_{n-2} & c_{n-2} & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & a_{n-3} & b_{n-3} & c_{n-3} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & a_{n-4} & b_{n-4} & c_{n-4} & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & a_1 & b_1 & c_1 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & a_0 & b_0 \end{bmatrix}$$

$$\times \begin{bmatrix} V_{Xn-1} \\ V_{Xn-2} \\ V_{Xn-3} \\ V_{Xn-4} \\ \cdot \\ \cdot \\ V_{X1} \\ V_{X0} \end{bmatrix} + \begin{bmatrix} K \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix} \qquad \text{(C.1)}$$

The next step is to allow the unknown variables to be represented by an infinite series in the form of (C.2).

$$V_{Xi}(t) = V_{Xi0}(t) + \mu V_{Xi1}(t) + \mu^2 V_{Xi2}(t) + \cdots \qquad \text{(C.2)}$$

These series are then substituted into (C.1), keeping only those terms with powers of μ less than a specific cutoff point. Usually the first or second power of μ is sufficient. Initially, only the terms of power μ^0 are used, and the linear system is solved for V_{Xi0}. The next step is to use the terms power of μ^1, substituting the previous solution for V_{Xi0} where necessary, and to solve this linear system for V_{Xi1}. This process continues until the system comprised of the highest order terms of μ is solved. In this manner, a number of linear systems can be solved in order to build an approximate analytical solution to the original nonlinear system.

Once the process is complete, the approximate solution to the system of equations is given by the sum of the component solutions, as shown below.

$$V_{Xi}(t) \approx V_{Xi0}(t) + V_{Xi1}(t) + V_{Xi2}(t) + \cdots \qquad \text{(C.3)}$$

It is important to note two properties unique to the system being solved. The first is that since only a solution of V_{Xn-1} is necessary, and the only nonlinearities involved in the system are specific to V_{Xn-1},it is unnecessary to solve any of the intermediate steps for any value other than $V_{X(n-1)j}(t)$, where j is the index of the component terms. Second, again since the only nonlinearities in the system are specific to V_{Xn-1}, with the exception of the top row of (C.1), the system remains the same for each successive power of μ. Both of these properties make the solution of the intermediate systems by Laplace transform an attractive method. The Laplace transform of the system, other than the top row, is computed only once and placed in a form for substitution into the transform of the top row. In this manner, each successive step requires only the substitution and solution of a single linear differential equation.

A drawback of this approach is that it is not completely analytic for chains of more than four transistors. This occurs because the solution of the system of differential equations, even by perturbation, requires that the eigenvalues of an $n \times n$ coefficient matrix be found. These eigenvalues may be generated numerically by a number of well known methods [26].

REFERENCES

[1] T. Sakurai and A. R. Newton, "Delay analysis of series-connected MOSFET circuits," *IEEE J. Solid-State Circuits*, vol. 26, pp. 122–131, Feb. 1991.

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

[3] M. Shoji, "Apparatus for increasing the speed of a circuit having a string of IGFETs." U.S. Patent 4 430 583, Feb. 7, 1984.

[4] M. Shoji, "FET scaling in domino CMOS gates," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 1067–1071, Oct. 1985.

[5] M. Shoji, *CMOS digital circuit technology*. Englewood Cliffs, NJ:Prentice Hall, 1988, pp. 243–253.

[6] G. A. Jullien, W. C. Miller, R. Grondin, Z. Wang, L. Del Pup, and S. Bizzan, "Woodchuck: A low-level synthesizer for dynamic pipelined DSP arithmetic logic blocks," in *Proc. IEEE Int. Symposium on Circuits and Systems*, May 1992, pp. 176–179.

[7] 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, July 1992.

[8] 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, June 1982.

[9] M. Shoji, "Theory of CMOS digital circuits and circuit failures. Princeton, NJ: Princeton Univ. Press, 1992, pp. 137–142.

[10] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584–594, Apr. 1990.

[11] S. M. Kang and H. Y. Chen, "A global delay model for domino CMOS circuits with application to transistor sizing," *Int. J. Circuit Theory and Applicat.*, vol. 18, pp. 289–306, May/June 1990.

[12] T. Sakurai and A. R. Newton, "A simple short-channel MOSFET model and its application to delay analysis of inverters and series-connected MOSFETs," in *Proc. IEEE Int. Symp. on Circuits and Systems*, pp. 105–108, May 1990.

[13] H. Shichman and D. A. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuits," *IEEE J. Solid-State Circuits*, vol. SC-3, pp. 285–289, Sept. 1968.

[14] H. C. Lin and L. W. Linholm, "An optimized output stage for MOS integrated circuits," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 106–109, April 1975.

[15] R. C. Jaeger, "Comments on 'An optimized output stage for MOS integrated circuits'," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 185–186, June 1975.

[16] B. S. Cherkauer and E. G. Friedman, "The effects of channel width tapering on the power dissipation of serially connected MOSFETs," in *Proc. IEEE Int. Symp. on Circuits and Systems*, May 1993, pp. 2110–2113.

[17] L. G. Heller, W. R. Griffin, J. W. Davis, and N. G. Thoma, "Cascode voltage switch logic: A differential CMOS logic family," *Proc. Int. Solid-State Circuits Conf.*, pp. 16–17, Feb. 1984.

[18] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 468–473, Aug. 1984.

[19] L.T. Wurtz, "An efficient scaling procedure for Domino CMOS logic," *IEEE J. Solid-State Circuits*, vol. 28, pp. 979–982, Sept. 1993.

[20] A. Vladimirescu and S. Liu, "The simulation of MOS integrated circuits using SPICE2," ERL Memo M80/7, Univ. of California, Berkeley, Oct. 1980.
[21] S. M. Kang, "Accurate simulation of power dissipation in VLSI circuits," *IEEE J. Solid-State Circuits*, vol. SC-21, pp. 889–891, Oct. 1986.
[22] G. J. Fisher, "An enhanced power meter for SPICE2 circuit simulation," *IEEE Trans. Computer-Aided Design*, vol. 7, pp. 641–643, May 1988.
[23] F. Rouatbi, B. Haroun, and A. J. Al-Khalili, "Power estimation tool for sub-micron CMOS VLSI circuits," in *Proc. Int. Conf. on Computer-Aided Design*, Nov. 1992, pp. 204–209.
[24] J. P. Uyemura, *Circuit design for CMOS VLSI*. Boston, MA: Kluwer Acad., 1992, p. 33.
[25] W. J. Cunningham, *Introduction to Nonlinear Analysis*. New York: McGraw-Hill, 1958, pp. 123–133.
[26] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C, The Art of Scientific Computing*. Cambridge, UK: Cambridge Univ. Press, 1988.

Brian S. Cherkauer (S'93) was born in Buffalo, NY, in 1968. He received the B.S. degree in electrical engineering summa cum laude from the State University of New York at Buffalo in 1990 and the M.S. degree in electrical engineering from the University of Rochester in 1991.

He was previously employed as a software engineering technician at Mennen Medical, Inc, Clarence, NY, from 1988 to 1990 where he developed electrocardiogram simulation, digitization, and playback software. He received the Sproull Fellowship at the University of Rochester in 1990 and has been a teaching and research assistant there since 1992. He is currently working towards the Ph.D. degree at the University of Rochester. His research interests include high performance digital and analog integrated circuit design techniques; CMOS and BiCMOS integrated circuit design; speed, area, and power tradeoffs; and design for reliability.

Eby G. Friedman (S'78–M'79–SM'90) was born in Jersey City, NJ, in 1957. He received the B.S. degree in electrical engineering from Lafayette College, Easton, PA, in 1979 and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Irvine, in 1981 and 1989, respectively.

He was previously employed by Philips Gloeilampen Fabrieken, Eindhoven, The Netherlands, in 1978 where he worked on the design of bipolar differential amplifiers. From 1979 to 1983, he was employed by Hughes Aircraft Company, Newport Beach, CA, working in the areas of custom IC design, software compatible gate array design, one- and two-dimensional device modeling, circuit modeling, and double-level metal process development. From 1983 to 1991, he was with Hughes Aircraft Company, Carlsbad, CA, rising to the position of Manager of the Signal Processing Design and Test Department, responsible for the design and test of high performance VLSI/VHSIC CMOS and BiMOS digital and analog IC's, the development of supporting design and test methodologies and CAD tools, functional and parametric test, and the development of high performance and high resolution DSP and oversampled systems. He has been with the Department of Electrical Engineering, University of Rochester, since 1991, where he is an Associate Professor and Director of the High Performance VLSI/IC Design and Analysis Laboratory. His current research and teaching interests are in the areas of high performance VLSI/IC design and analysis with an emphasis on niche technologies, and their system applications. He is the author of a book chapter and many papers and presentations in the fields of VLSI design, high-speed and low-power CMOS design techniques and CAD tools, pipelining and retiming, and the theory and application of synchronous clock distribution networks to high performance VLSI-based digital systems.

Dr. Friedman is a member of the editorial board of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING, Secretary of the VLSI Systems and Applications CAS Technical Committee, a member of the technical program committee of several conferences (APAW, GLS VLSI, MWSCAS, ASIC) an officer of the Electron Devices Chapter of the IEEE Rochester Section, has chaired sessions at various IEEE Conferences, and is a recipient of the Howard Hughes Masters and Doctoral Fellowships, the NSF Research Initiation Award, and the DoD Augmentation Award for Science and Engineering Research Training.