The Effects of Channel Width Tapering on the Power Dissipation of Serially Connected MOSFETs

Brian S. Cherkauer and Eby G. Friedman
Department of Electrical Engineering
University of Rochester
Rochester, New York 14627 USA

Abstract — **In this paper transistor channel width tapering in serial MOSFET chains is shown to 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. In this work 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 does not offer a speed advantage. In addition, dynamic CV^2f power dissipation of the serial chain is reduced. This behavior permits a designer to trade-off speed for a reduction in short-circuit and dynamic power dissipation, a trade-off not normally available with untapered chains.**

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 speed, area, and power dissipation, these factors must be considered together.

Many CMOS logic structures are composed of chains of MOSFETs serially connected between a power supply rail and the output of the subcircuit. These serially connected MOSFETs are a major source of delay and power dissipation [1]; therefore, optimal sizing of these transistors is important in reducing their delay and power dissipation. In this paper, the physical operation of serially connected MOSFETs is explained, and a method for sizing these transistors is presented. This method may improve circuit speed and simultaneously decrease both power dissipation and area.

II. Previous Research on Tapered Serial MOSFETs

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

In typical integrated circuits, the size of the transistors in the discharge chain is constrained to have the same channel dimensions such that the circuit satisfies its design criteria for speed and area. Shoji [2, 3] first pointed out 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 [2], 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). Exponential tapering assumes a fixed ratio, α (where $0 < \alpha \leq 1$), between the channel widths of adjacent transistors. A tapering factor of $\alpha = 1$ implies that each channel has equal width which represents an untapered chain. 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.

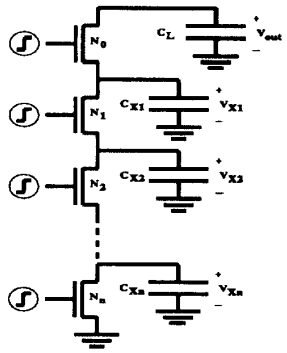

Fig. 1. Serially connected NMOS chain

Tapered widths have been successfully integrated into an automated layout system to increase circuit performance [4, 5]. In this physical synthesis tool, strings of serially connected MOSFETs are automatically sized by an approximate analytical formula to generate a tapered profile. Layouts are then synthesized based on this tapered profile specification.

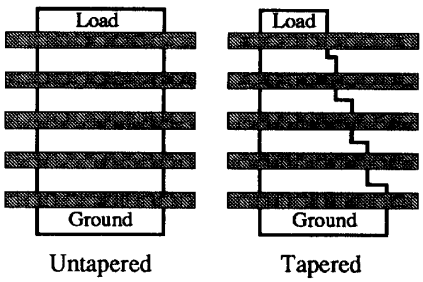

Fig. 2. Untapered and exponentially tapered MOSFET chains

This material is based upon work supported by the National Science Foundation under Grant No. MIP-9208165

III. Circuit Operation of Untapered and Tapered NMOS Chains

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

A. Region 1

Assuming worst case initial conditions, $V_{out}(t = 0^-) = V_{DD}$ and $V_{Xi}(t = 0^-) = 0$, and that simultaneous step inputs are applied at the gate of each transistor at time t = 0, transistor N_0 in Fig. 1 operates initially in the saturation region, then enters the linear region. The remaining transistors operate entirely in the linear region until the capacitances are fully discharged. The time during which the transistor closest to the output operates in the saturation region is referred to in this paper as Region 1.

Charging the drain/source parasitic capacitances stabilizes the node voltages such that the transistors are each 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" [6], 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, since the current sunk by the saturated device is relatively independent of its drain voltage.

With tapering, the channel width of N_0 is decreased, thereby lowering the transconductance of the device. Since the NMOS chain acts as a voltage divider, this has the effect of lowering the maximum voltage of the source terminal of N_0. The parasitic capacitance at the source node of N_0, C_{X1} (see Fig. 1), is also lowered by tapering. This allows the voltage at the source node, V_{X1}, to increase more quickly. In turn, this causes the saturation current to decrease more quickly due to the rising source potential of N_0.

In this paper it is shown experimentally that the effect of tapering during the time when N_0 is saturated tends to slow the output response. That is, the effects of decreasing W/L are not offset by the effects of increasing $V_{GS}-V_T$ and decreasing C_{X1}, and the discharge of the load is usually slower for a tapered chain during the saturated region of operation than it would be for an untapered chain. 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_0 operates in saturation.

B. Region 2

Once enough charge has been drained from the load capacitance to allow the saturated upper transistor, N_0, to pass into the linear state, its current becomes highly dependent upon its drain voltage, and the current begins to decrease as 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 capacitances are fully discharged.

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, 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.

IV. Experimental Results of Tapering

Tapered circuits fall into three basic categories. The first category of tapered circuits is comprised of those circuits where the serial chain is connected to a very small load capacitance, as might be seen in Domino logic. The second category occurs in those circuits where the load capacitance is somewhat larger, such as might be seen in Domino 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 and/or distant load capacitance, as might be seen in large static NAND/NOR logic gates.

The first category encompasses circuits for which tapering will actually reduce 50% propagation delay. In addition, the output waveform has a shorter fall-time, translating into a reduction in short-circuit power dissipation [7] in the following stage. Dynamic power and area requirements are also both reduced as a direct result of reduced transistor geometric width. This is discussed in further detail later in this section.

The second category of tapered circuits represents a transition between the first and the third both in terms of the magnitude of the 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%-10% fall times. The result is that the output waveform, though shifted out in time, is more rectangular.

It is this characteristic of tapering which is most useful in 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 $Vss + V_{Tn}$ is reduced, which translates into a reduction in short-circuit power dissipation in the following stage. Coupled with the reduction in dynamic power dissipation due to the decreased parasitic drain and source capacitances and input gate capacitance, tapering becomes advantageous in those circuit designs where power dissipation is often a major concern. This reduction in dynamic and short-circuit power dissipation comes with only a minimal increase in 50% delay, a reduction in 90%–10% delay, and in certain cases a reduction in overall system delay.

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 both increase the delay and the fall time, which results in increased short-circuit power dissipation in the following stage. The area and dynamic power advantages remain; however, these advantages are outweighed by the disadvantages of decreased speed and increased short-circuit power. Tapering circuits which belong to the third category is not recommended.

Fig. 3 depicts a typical CMOS Domino logic configuration which contains a serial chain of MOSFETs. The figure 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 because of 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 because of 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. 3.

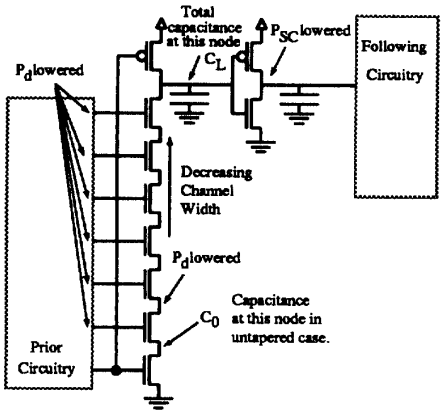

Fig. 3. Example system illustrating decreased power dissipation

A. Propagation Delay

As evidenced by Table I, 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 in Category 1 of Table I. The 50% propagation delays are shown normalized to the delay of the untapered chains. No improvement in delay through tapering is shown in Categories 2 and 3.

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%

B. Short-Circuit Power Dissipation

In Table II the short-circuit power dissipation of the example circuit is compared. Note that the data is normalized to the untapered case (α = 1.0). A reduction in short-circuit power dissipation is shown in both Categories 1 and 2. An increase in short-circuit power dissipation is apparent in Category 3 because the large load capacitance delays the transient response of the serial chain, which swamps out any positive effects of tapering.

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%

C. Dynamic Power

The ratio of dynamic power dissipation (CV^2f) originating from the input gate capacitance of an n-transistor tapered serial chain as compared to an untapered chain is shown in (1). A graph of dynamic power dissipation versus tapering factor, α, for different number of serially connected transistors, is shown in Fig. 4. As can be seen, as n increases and α decreases, the ratio of tapered to untapered dynamic power dissipation decreases significantly.

$$\frac{P_{d\ Tapered}}{P_{d\ Untapered}} = \frac{1}{n}\sum_{i=0}^{n-1}\alpha^i \qquad (1)$$

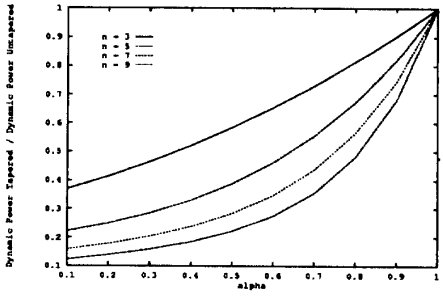

Fig. 4. Ratio of dynamic power in tapered versus untapered serial chains

D. Total Power Dissipation

The effects of tapering on the total power dissipation of the system depicted in Fig. 3 is shown by Figs. 5-7, which describe the total power dissipated in the system during a high-to-low transition. These results include both the dynamic and the short-circuit power dissipation of the serial chain, the input circuitry driving the serial chain, and the inverter loading the serial chain.

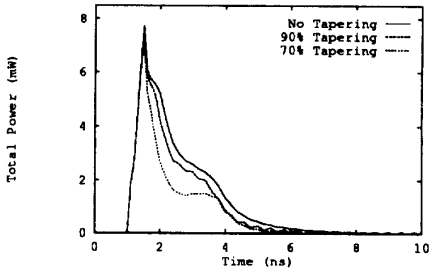

Fig. 5. Total power dissipation of high-to-low transition for Category 1. C_L / C_0 = 0.87

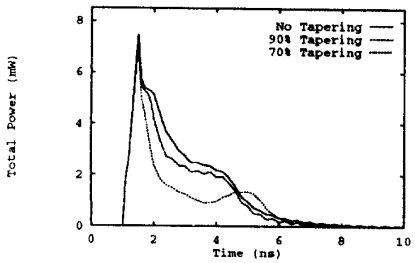

Fig. 6. Total power dissipation of high-to-low transition for Category 2. C_L / C_0 = 1.2

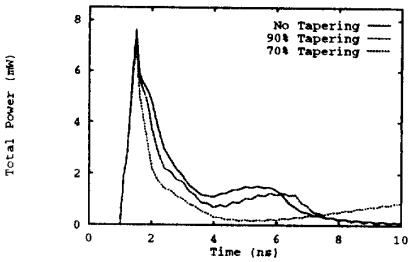

Fig. 7. Total power dissipation of high-to-low transition for Category 3. C_L / C_0 = 5.0

In Table III, the average total power dissipation depicted in Figs. 5-7 is tabulated and normalized to the untapered cases. As can be seen from Table III and Figs. 5-7, 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. 5-7), when the inputs switch, there is a 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 significantly with tapering because it is mainly determined by the saturated current developed from the inverters driving the MOSFET chain (the prior circuitry in Fig. 3). The magnitude of this saturated current only depends upon the input 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 slight reduction in maximum power dissipation. In this example, a value of $\lambda = 0.035\ V^{-1}$ is assumed.

As the pull-up devices in the inverters enter the linear region, their currents become highly dependent upon the drain-to-source voltages across the P-channel devices. Because of 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 occurs approximately when 2 ns < t < 3 ns in Figs. 5-7.

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 hump is due to the power dissipated when the inverter loading the serial chain switches.

Table III
Average power dissipation

Category	Capacitance Ratio	α = 1.0	α = 0.9	α = 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%

V. Conclusions

This paper provides a detailed explanation of the effects of channel width tapering on a chain of serially connected MOSFETs, a common structure in CMOS-based VLSI circuits. Tapering is shown to decrease area and dynamic power dissipation under all conditions and to decrease delay and short-circuit power dissipation under certain conditions. The propagation delay of a tapered serial chain was shown to decrease when the load capacitance is approximately equal to or less than the parasitic drain/source capacitance of the nodes along the serial chain, which is called Category 1 in this paper. As shown in Table I, propagation delay decreases by 7% with a tapering factor of α = 0.9 applied to the example Domino circuit shown in Fig. 3. Short-circuit power dissipation can be decreased by tapering a Category 1 or Category 2 circuit. In Table II, short-circuit power dissipation decreases by 15% or more with α = 0.9 for a standard Domino circuit for both Category 1 and Category 2 circuits. In Table III, a reduction of over 10% in average total power dissipation is demonstrated for all three categories with α = 0.9 and over 20% with α = 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 MOSFETS, such as in Domino logic, and in those circuits where power dissipation is of primary concern.

References

[1] T. Sakurai and A. R. Newton, "Delay Analysis of Series-Connected MOSFET Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-26, pp. 122–131, Feb. 1991.

[2] M. Shoji, "FET Scaling in Domino CMOS Gates," *IEEE Journal of Solid-State Circuits*, vol. SC-20, pp. 1067–1071, Oct. 1985.

[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] G. Jullien, W. Miller, R. Grondin, Z. Wang, L. Del Pup, and S. Bizzan, "Woodchuck: A Low-Level Synthesizer for Dynamic Pipelined DSP Arithmetic Logic Blocks," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 176–179, May 1992.

[5] S. S. Bizzan, G. A. Jullien, and W. Miller, "Analytical Approach to Sizing nFET Chains," *Electronics Letters*, vol. 28, pp. 1334–1335, July 1992.

[6] S. M. Kang and H. Y. Chen, "A Global Delay Model for Domino CMOS Circuits with Application to Transistor Sizing," *International Journal of Circuit Theory and Applications*, vol. 18, pp. 289–306, 1990.

[7] H. J. M. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and Its Impact on the Design of Buffer Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-19, pp. 468–473, Aug. 1984.