

# A Compact Delay Model for Series-Connected MOSFETs

Kaveh Shakeri and James D. Meindl

Microelectronic Research Center, Georgia Institute of Technology

791, Atlantic Dr NW, Atlanta, GA, 30332, USA

Phone: (404) 894 9912

Email: kaveh@ece.gatech.edu

## ABSTRACT

A compact delay model for series connected MOSFETs has been derived. This model enables accurate prediction of worst-case delay of different logic families such as dynamic logic. It also provides insight into delay change as the device parameters change. Key results show that the relative delay of series connected MOSFETs is almost invariant for different generations of technology.

## 1. INTRODUCTION

Series connected MOSFETs (Fig. 1) are used in different logic families including dynamic logic families, static CMOS gates and many other logic families. A compact model for series connected MOSFETs when the drain/source capacitance of the MOSFET is small compared to the load capacitance has been derived by Sakurai [1]. Sakurai's model can be used for static logic families, in which the load capacitance is much larger than drain/source capacitances. However in logic families such as dynamic logic circuits, where the drain/source capacitance is not negligible compared to the load capacitance, there is no compact analytical model for the delay of series connected MOSFETs. Therefore, the purpose of this paper is to describe a model for series connected MOSFETs when the drain/source capacitance is not negligible compared to the load capacitance.

The alpha power law model has been used for the transistor model [2]. Although this model is very simple it represents accurately the velocity saturation effect of the transistor. Therefore, it is a useful model for sub-micrometer devices. The disadvantage of this model is that it is empirical and is not able to predict MOSFET behavior for future generations. The physical alpha power law model [3] for MOSFETs provides a physical interpretation of the device parameters; therefore it enables projections for future generations.

$$\left\{ \begin{array}{l} I_D = I_{DSAT} \left( 2 - \frac{V_{DS}}{V_{DSAT}} \right) \cdot \left( \frac{1 + \lambda \cdot V_{DS}}{1 + \lambda \cdot V_{DD}} \right) \cdot \frac{V_{DS}}{V_{DSAT}} \quad (V_{DS} < V_{DSAT}) \text{ Linear region} \\ I_D = I_{DSAT} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{T0}} \right)^\alpha \left( \frac{1 + \lambda \cdot V_{DS}}{1 + \lambda \cdot V_{DD}} \right) \quad (V_{DS} \geq V_{DSAT}) \text{ Saturation region} \end{array} \right. \quad (1)$$

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GLSVLSI'02, April 18-19, 2002, New York, New York, USA.

Copyright 2002 ACM 1-58113-462-2/02/0004...\$5.00.

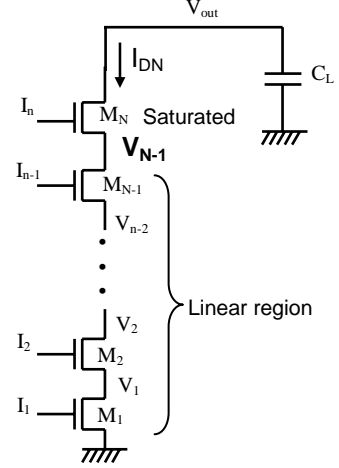


Fig. 1. Series connected transistors

In this model  $I_{DSAT}$  is the drain current when  $V_{GS}=V_{DS}=V_{DD}$ ,  $V_{DSAT}$  is the drain saturation voltage when  $V_{GS}=V_{DD}$ ,  $V_{T0}$  is the threshold voltage with no body bias,  $V_{TH}$  is the threshold voltage with body bias,  $V_{T0}$  is the threshold voltage with no body bias,  $\alpha$  is an empirical parameter and  $\lambda$  is the channel length modulation parameter. A linear approximation of the body effect is used for the device threshold voltage ( $V_{TH}$ ).

$$V_{TH} = V_{T0} - \gamma_1 V_{BS} \quad (2)$$

In this equation  $V_{BS}$  is the bulk-source voltage and  $\gamma_1$  is the body-bias factor.

## 2. NEGLIGIBLE DRAIN/SOURCE CAPACITANCE

Sakurai describes a model for series connected transistors when the drain/source capacitance is small compared to the load capacitance,  $C_L$  [1]. In this case, the ratio of the delay of Series Connected MOSFETS (SCMS) to the delay of a single transistor, as a function of the number of transistors  $n$  is

$$\begin{aligned} F_D &= \frac{\text{delay(SCMS)}}{\text{delay(inverter)}} = \frac{I_{DSAT}}{I_{DN}} = 1 + \frac{1 - 1/\sqrt{2}}{1 - 1/\sqrt{2}} \frac{V_{DSAT}}{V_{DD} - V_{TH}} (1 + \gamma_1)(N-1) \\ &\approx 1 + \frac{1}{2} \alpha \frac{V_{DSAT}}{V_{DD} - V_{TH}} (1 + \gamma_1)(N-1), \end{aligned} \quad (3)$$

where  $F_D$  is the ratio of the delay of  $N$  transistors in series to the delay of a single transistor discharging the same load capacitance,  $I_{DN}$  is the equivalent current of the SCMS and  $N$  is the number of transistors in series. This model can be used only if the load capacitance,  $C_L$  is large compared to the drain/source capacitances of the MOSFETs. In normal static CMOS gates, the output capacitance is large compared to the drain/source capacitances so the results are in good agreement with the delay of normal static

CMOS gates. However, in dynamic circuits where the load capacitance is comparable to drain/source capacitance the model described by (3) doesn't have good agreement with Spice simulations.

### 3. ELMORE DELAY MODEL WHEN DRAIN/SOURCE CAPACITANCES ARE NOT NEGLIGIBLE

In this case, we cannot neglect the drain and source capacitances. In the Elmore delay model transistors are modeled as resistors and the delay  $T$  can be calculated using the Elmore delay rules [4].

$$T = 0.69 \times (R_1 C_1 + (R_1 + R_2) \times C_2 + (R_1 + R_2 + R_3) \times C_3 + \dots + (R_1 + R_2 + \dots + R_{N-1}) \times C_{N-1} + (R_1 + R_2 + \dots + R_N) \times C_L), \quad (4)$$

where  $C_L$  equal to the total capacitances at  $V_{out}$ ,  $R_1, R_2, \dots, R_N$  are the equivalent resistances of the MOSFETs  $M_1, M_2, \dots, M_N$  and  $C_1, C_2, \dots, C_{N-1}$  are their drain/source capacitances. If the transistors are equal we have

$$\begin{aligned} R_1 = R_2 = \dots = R_N = R \\ C_1 = C_2 = \dots = C_{N-1} = C. \end{aligned} \quad (5)$$

As a result, the delay expression can be simplified to

$$T = 0.69 \times \left( \frac{N^2 - N}{2} RC + N \cdot R \cdot C_L \right). \quad (6)$$

Figure 2 shows the  $F_D$  of a dynamic logic AND gate versus number of inputs. As shown in the figure these models do not have good agreement with Spice simulations. The Elmore model overestimates the delay and Sakurai's model underestimates it. Therefore in the next section a new model is introduced for the delay of series connected MOSFETs.

### 4. MODELING

Depending on which transistor switches, the delay of series connected MOSFETs changes. To find the worst-case delay for series connected MOSFETs different input combinations should be examined (Fig 1). Different input situations are:

- i) Transistors  $M_1$  to  $M_{N-1}$  are all on and transistor  $M_N$  is off. In this case, all of the drain/source capacitances are already discharged. Therefore when  $M_N$  is turned on, the switching is fast.
- ii) Transistors  $M_1$  to  $M_{k-1}$  and  $M_{k+1}$  to  $M_N$  are all on and transistor  $M_k$  is off. In this case, the drain/source

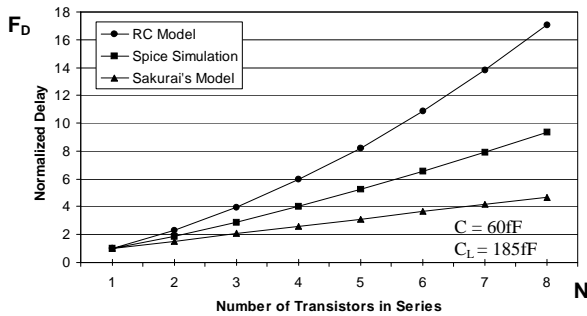


Fig. 2. Normalized delay versus number of transistors for Spice simulations and different models

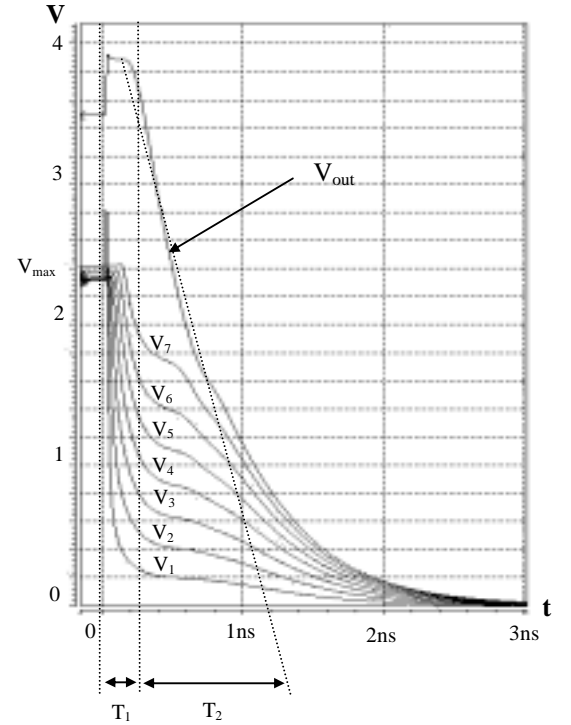


Fig. 3. Voltages of different nodes of eight series connected transistors

capacitances of transistors  $M_1$  through  $M_{k-1}$  are all already discharged and the drain/source capacitances of transistors  $M_{k+1}$  to  $M_N$  are all charged to  $V_{max}$ , which is the highest voltage to which they can be charged through an NFET transistor. When the transistor  $M_k$  is turned on the charged drain/source capacitances of transistors  $M_{k+1}$  to  $M_N$  are discharged through transistors  $M_1$  to  $M_{k-1}$ . Therefore the delay in this case is more than the previous case.

- iii) The worst-case delay is when transistors  $M_2$  to  $M_N$  are all on and  $M_1$  is off. In this case the drain/source capacitances are all charged to  $V_{max}$  before switching and when switched they are all discharged through transistors  $M_1$  to  $M_N$ . This case has the maximum delay and therefore has been modeled.

Fig. 3 shows the voltage of the nodes of eight transistors in series for the worst-case delay. In this case, the initial voltages of the drain/source nodes are  $V_{max}$ . When the lowest transistor,  $M_1$  is turned on, it starts discharging the drain/source capacitances until  $T_1$  without affecting the  $V_{out}$ . After  $T_1$  the series transistors start discharging the load capacitance. Therefore the output can be modeled as shown in Fig. 4. It is made of two parts  $T_1$  and  $T_2$ .  $T_2$  has been modeled by Sakurai (3) but the first part which is because of the discharge time of the drain/source capacitances has not been modeled. Therefore a model is needed for time  $T_1$ . During  $T_1$  transistors  $M_1$  to  $M_{N-1}$  are all non-saturated. Therefore, they can be represented as resistors. Using the alpha power law model the equivalent resistances of these transistors can be modeled as

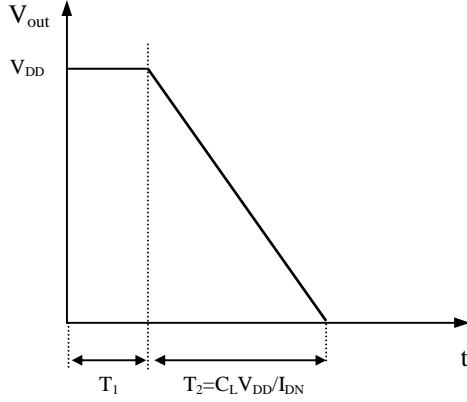


Fig. 4. Output voltage

$$R = \frac{(2 - \sqrt{2}) \cdot (1 + \lambda \cdot V_{DD}) \cdot V_{DSAT}}{\left(1 + \lambda \cdot V_{DSAT} \left(1 - \frac{\sqrt{2}}{2}\right)\right) \cdot I_{DSAT}} \quad (7)$$

Transistor  $M_N$  is saturated for  $t < T_1$  and its current is a function of its source voltage. Therefore it can be modeled as a resistor  $R_N$  connected to a power supply equal to  $V_{max}$ .

$$V_{max} = \frac{V_{DD} - V_{T0}}{1 + \gamma_1} \quad (8)$$

$$R_N = \frac{V_{max} \cdot (1 + \lambda V_{DD})}{2^{\left(\frac{1}{\alpha} - 1\right)} \cdot I_{DSAT} \cdot \left(1 + \lambda \left(V_{DD} - V_{max} \left(1 - 2^{-\frac{1}{\alpha}}\right)\right)\right)} \quad (9)$$

As a result, the series transistors during time  $T_1$  can be modeled as shown in Fig. 5. The current passing through resistor  $R_N$  discharges the load capacitance.  $V_x$ , which is

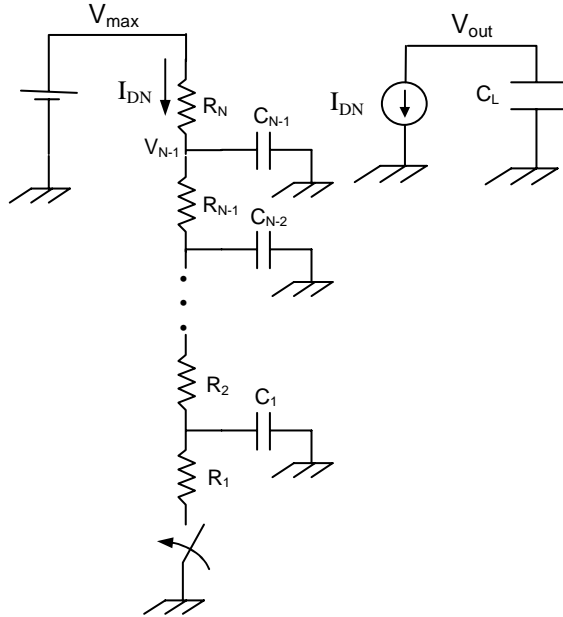


Fig. 5. Model for calculating  $T_1$

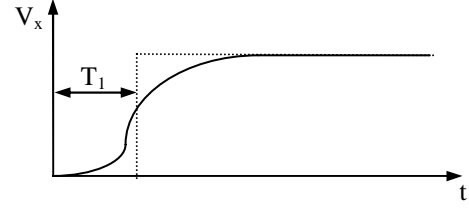


Fig. 6. Voltage of  $V_x$  as a function of time

$$V_x = V_{max} - V_{N-1}, \quad (10)$$

is shown in Fig. 6.  $V_x/R_N$  is the current discharging the load capacitance. Therefore, the charge removed from the output capacitance at time  $t$  is equal to the area under  $V_x/R_N$  curve at that time. This waveform can be approximated by a step function with the same area and delay  $T_1$  (Fig 6).

If  $V'_x$  is the impulse response of the circuit then  $T_1$  can be approximated by [4]

$$T_1 = \int_0^{\infty} t \times V'_x dt \quad (11)$$

The system is modeled as a linear system therefore the transfer function of the system can be written as

$$H(s) = \frac{1 + a_1 s + a_2 s^2 + \dots + a_n s^n}{1 + b_1 s + b_2 s^2 + \dots + b_m s^m}, \quad (12)$$

where  $a_i$  and  $b_i$  are real and  $m > n$ . By dividing the numerator by the denominator of the transfer function we have

$$H(s) = 1 - (b_1 - a_1)s + (b_1^2 - a_1 b_1 + a_2 - b_2)s^2 + \dots \quad (13)$$

From the definition of Laplace transform we have

$$H(s) = \int_0^{\infty} V'_x e^{-st} dt = 1 - s \int_0^{\infty} t \times V'_x dt + \frac{s^2}{2!} \int_0^{\infty} t^2 \times V'_x dt - \dots \quad (14)$$

$$= 1 - s \times T_1 + \dots$$

Therefore from equations (13) and (14),  $T_1$  can be calculated as

$$T_1 = b_1 - a_1. \quad (15)$$

For series connected MOSFETs and the worst-case initial condition  $a_1$  is zero and  $b_1$  is equal to

$$b_1 = \left( \frac{R_1 \times (R_2 + \dots + R_N)}{(R_1 + R_2 + \dots + R_N)} C_1 + \frac{(R_1 + R_2) \times (R_3 + \dots + R_N)}{(R_1 + R_2 + \dots + R_N)} C_2 + \dots \right) \quad (16)$$

$$\left( \frac{(R_1 + R_2 + \dots + R_{N-1}) \times R_N}{(R_1 + R_2 + \dots + R_N)} C_{N-1} \right)$$

Therefore,  $T_1$  is given by

$$T_1 = b_1 - a_1 = \left( \frac{R_1 \times (R_2 + \dots + R_N)}{(R_1 + R_2 + \dots + R_N)} C_1 + \frac{(R_1 + R_2) \times (R_3 + \dots + R_N)}{(R_1 + R_2 + \dots + R_N)} C_2 + \dots \right) \frac{(R_1 + R_2 + \dots + R_{N-1}) \times R_N C_{N-1}}{(R_1 + R_2 + \dots + R_N)} \quad (17)$$

When the transistors are of equal size, they can be modeled as equal resistors and capacitors indicated by

$$\begin{aligned} R_1 = R_2 = \dots = R_{N-1} = R \\ C_1 = C_2 = \dots = C_{N-1} = C. \end{aligned} \quad (18)$$

Simplifying (17) results in

$$T_1 = \frac{RC \cdot N \cdot (N-1)}{R_N + (N-1)R} \left[ \frac{R_N}{2} + R \cdot \left( \frac{N-1}{6} - \frac{1}{3} \right) \right], \quad (19)$$

where  $R$  and  $R_N$  can be calculated by (7) and (9) respectively. During time  $T_2$ , the current discharging  $C_L$  can be calculated from (3), therefore

$$T_2 = \frac{C_L \cdot V_{DD}}{I_{DN}} = \frac{C_L \cdot V_{DD} \cdot F_D}{I_{DSAT}} \quad (20)$$

As a result the output voltage is

$$V_{out} = V_{DD} - \frac{I_{DSAT}}{C_L \cdot F_D} (t - T_1) \times u(t - T_1), \quad t < T_1 + T_2 \quad (21)$$

where  $u(t - T_1)$  is the step function delayed by  $T_1$ .

## 5. VALIDATION OF THE RESULTS

Fig. 7 shows the normalized delay of dynamic AND gates implemented with  $0.5\mu\text{m}$  transistors versus number of inputs. The results show good agreement between the new model and the Spice simulations.

## 6. RESULTS

Fig. 8 shows normalized delay as a function of the number of transistors in series for different sub-micrometer generations. The

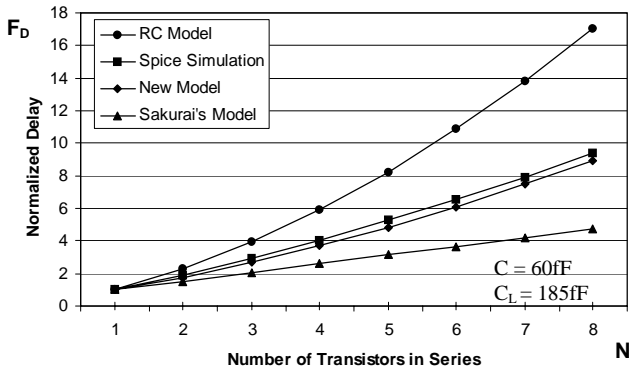


Fig. 7. Normalized delay of dynamic AND gates versus number of inputs for different models

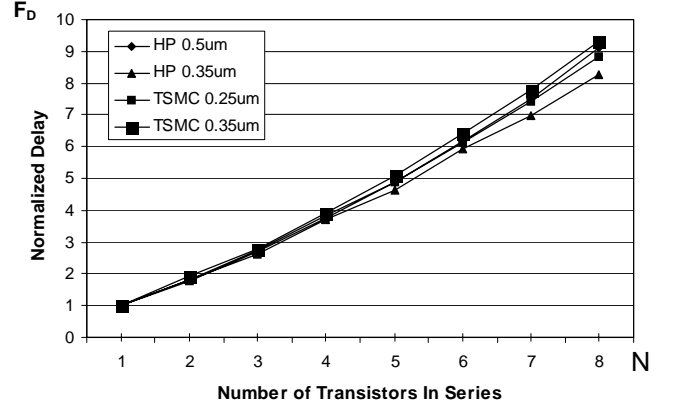


Fig. 8. Normalized delay versus number of transistors for different generations

model shows that the delay for series connected MOSFETs does not change for these sub-micrometer generations, because  $\alpha$  is almost constant. In other words these sub-micrometer devices are equally velocity saturated due to scaling both device dimensions and supply voltage.

## 7. CONCLUSION

A new model has been derived for the delay of series connected MOSFETs that can be used to calculate the delay of series connected MOSFETs used in any logic family. It also enables us to predict delay of different logic families for future generations and see how different parameters of the device affect the delay. Key results show that the relative delay of series connected MOSFETs is almost invariant for different generations of technology.

## 8. REFERENCES

- [1] T. Sakurai and A. R. Newton, "Delay Analysis of Series-Connected MOSFET Circuits," IEEE J. Solid-State Circuits, Vol. 26, NO. 2, Feb. 1991.
- [2] T. Sakurai and A. R. Newton, "Alpha-power model and its applications to CMOS inverter delay," IEEE J. Solid-State Circuits, Vol. 25, pp. 584-594, Apr 1990.
- [3] K. A. Bowman, B. Austin, J. Eble, X. Tang and J. D. Meindl, "A Physical Alpha-Power Law MOSFET Model," IEEE J. Solid-State Circuits, Vol. 34, No. 10, Oct 1999.
- [4] Elmore, W. C., "The transient response of damped linear networks with particular regard to wide-band amplifiers," J. Applied Physics, Vol. 19, 1948.