Journal of Circuits, Systems, and Computers
Vol. 15, No. 3 (2006) 437–454
© World Scientific Publishing Company

SIZING CMOS INVERTERS WITH MILLER EFFECT AND THRESHOLD VOLTAGE VARIATIONS

BORIS ANDREEV*, EDWARD L. TITLEBAUM and EBY G. FRIEDMAN

Department of Electrical and Computer Engineering,
University of Rochester, Rochester,
New York 14627, USA

Revised 25 March 2006

The maximum speed of synchronous circuits is generally constrained by the worst case propagation delay, which limits the system clock frequency. Various techniques exist to manage the circuit delay, trading off speed for other system resources. One such approach is to equalize the rise and fall delay times. The primary design parameter for equalizing these delay times is the ratio between the width of the PMOS and NMOS transistors, which determines the relationship between the currents passed along the pull-up and pull-down paths. The variation of the pull-up to pull-down ratio for different circuit parameters is discussed in this paper under the constraint of equal rise and fall delay times. It is shown that the short-circuit current and the Miller capacitance affect the ideal linear relationship between the CMOS inverter delay times and the load capacitance, requiring the pull-up to pull-down ratio to be adjusted as circuit parameters are varied. These effects are more pronounced in deep submicrometer technologies with significant parasitic MOSFET capacitances and threshold voltage variations. Based on analytic and experimental observations, circuit design guidelines are proposed to minimize the Miller effect.

Keywords: CMOS inverter; Miller effect; threshold voltage variations.

1. Introduction

The CMOS inverter is one of the most ubiquitous electronic circuits and is often used as an example circuit for the analysis of the broad family of CMOS integrated circuits. Other CMOS circuits, such as a NOR and NAND gate, can be associated with an equivalent inverter where one pair of switching input and output nodes is considered. The basic performance properties of such an electronic circuit are the critical path delay, power dissipation, physical area, and noise immunity.

In this paper, the propagation delay of CMOS circuits is discussed. The CMOS inverter, as shown in Fig. 1, is treated as a representative example of a CMOS logic gate. The focus of this work is on the discrepancy between the propagation

*Boris Andreev is currently with QUALCOMM Incorporated, San Diego, California.

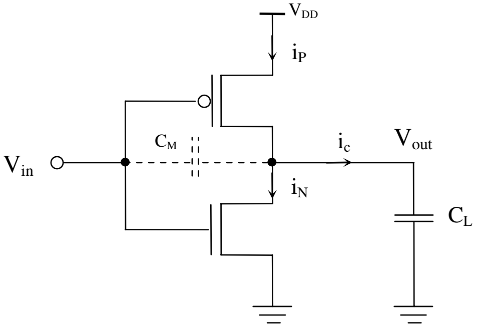


Fig. 1. CMOS inverter.

delay when the output voltage V_{out} is switching high-to-low (fall delay time, τ_{PHL}) versus switching low-to-high (rise delay time, τ_{PLH}). CMOS circuits are generally characterized by the worst case propagation delay. The rise and fall delay times are therefore typically designed to be equal to minimize the worst case delay. The pull-up to pull-down ratio W_p/W_n between the channel width of the PMOS and NMOS transistors at which the rise and fall delay times are equal ($\tau_{\text{PHL}} = \tau_{\text{PLH}}$) is referred to as the *optimal* W_p/W_n ratio (also denoted as β_{opt}) in the following sections.

This optimal W_p/W_n ratio is typically considered to be independent of design parameters such as the load capacitance C_L and input transition time τ. As discussed in this paper, however, due to parasitic impedances and imbalances between the pull-up and pull-down paths, the rise and fall delays differ so that the W_p/W_n ratio should be adjusted to restore the balance. The delay imbalance is demonstrated in Figs. 2 and 3 as the load capacitance C_L and the input transition time τ are varied. To compensate for these effects, the optimal W_p/W_n ratio is adjusted as shown in Figs. 4 and 5. The physical phenomena behind these effects are discussed in the following sections.

The noise immunity of a CMOS circuit is also affected by the W_p/W_n ratio. As this ratio is increased, the pull-up path becomes lower resistance and, correspondingly, the switching point of the logic gate rises. The motivation for strongly biasing the switching point is to reduce the voltage swing and propagation delay. The noise immunity, however, is reduced since it is easier for a voltage glitch at an input node to propagate to the output. The following discussion on the variation of the W_p/W_n ratio for different circuit parameters can also be applied into an integrated delay-noise optimization process.

CMOS buffer optimization has previously been analyzed by Hedenstierna and Jeppson in Ref. 1 and an expression is presented for the required W_p/W_n ratio to achieve the minimum average delay. The minimum average delay along a chain of inverters is achieved without the constraint of equal rise and fall delays and without considering the input-to-output capacitance (Miller capacitance C_M) and the second conducting transistor. The influence of the transistor gain ratio and coupling capacitance C_M on the CMOS inverter delay is modeled by Jeppson in Ref. 2. He

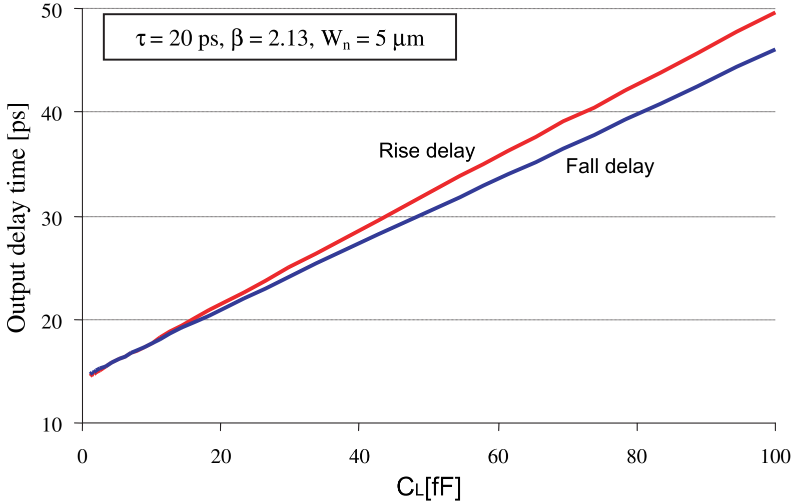


Fig. 2. Dependence of the output delay of a CMOS inverter on the load capacitance C_L.

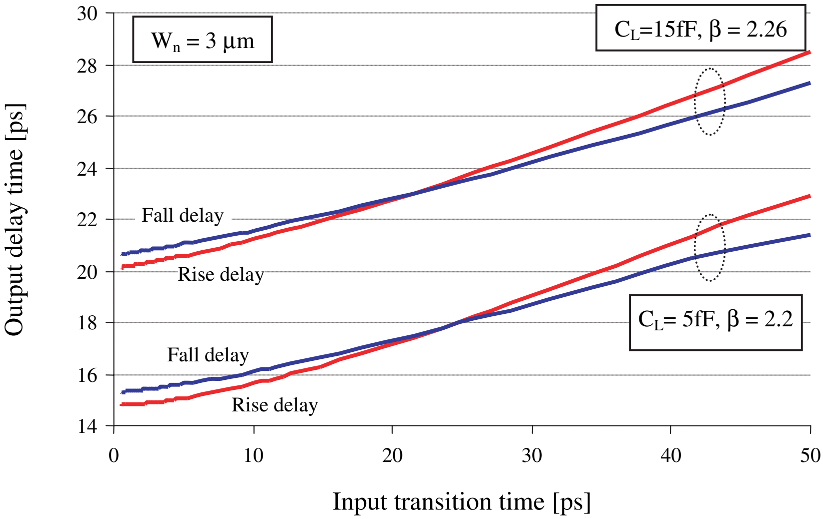


Fig. 3. Dependence of the output delay of a CMOS inverter on the input transition time.

shows that the short-circuit current and the Miller capacitance modify the ideal linear relationship between the inverter delay and the rise/fall time of the input ramp. Kung and Puri[3] propose an approach for selecting the W_p/W_n ratio in standard cell libraries so as to minimize the average delay. A theoretical framework is developed in Ref. 3 to determine the optimal W_p/W_n ratio for each logic gate utilizing empirical gate delay models that explicitly represent the dependency of delay on the W_p/W_n ratio and load capacitance.

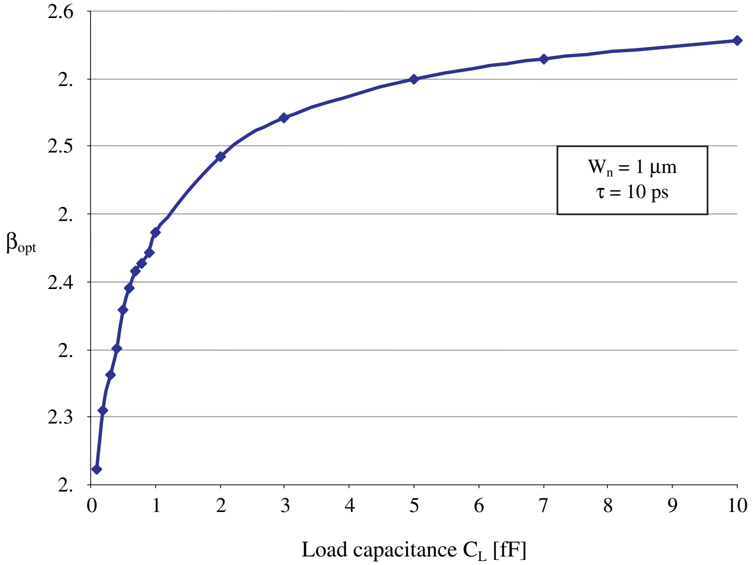


Fig. 4. Dependence of the $\beta_{\rm opt}$ ratio on the load capacitance C_L.

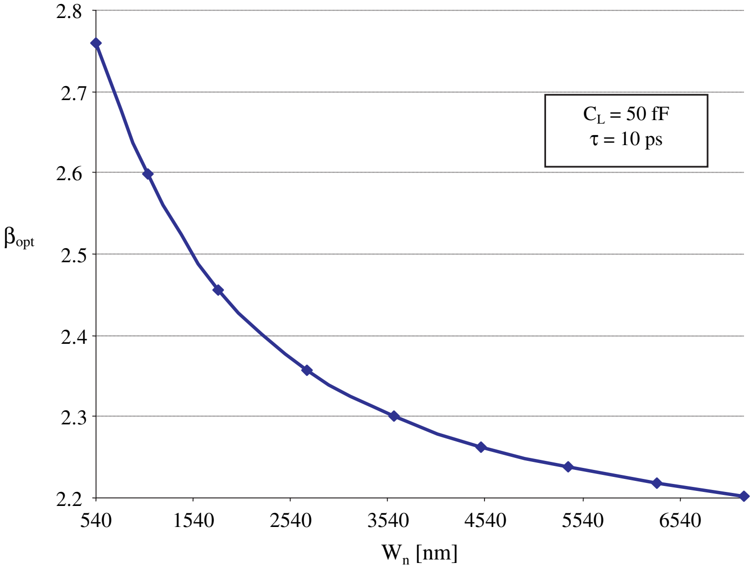


Fig. 5. Dependence of the $\beta_{\rm opt}$ ratio on the NMOS channel width W_n.

Previous publications have focused on the analysis of a chain of CMOS circuits, minimizing the overall average delay without requiring symmetric inverters with equal rise and fall delay times. Alternatively, the primary contribution of this paper is an analysis of the variation in the optimal W_p/W_n ratio (such that the rise and fall delay times are equal) for a CMOS inverter under the assumptions of a ramp input voltage and lumped load capacitance C_L. Physical phenomena affecting the delay imbalance, such as threshold voltage variations and the Miller capacitance, are described. A model is presented for the CMOS inverter delay that considers the Miller capacitance assuming a step input voltage, demonstrating the significance of the Miller effect on both the initial value of the output voltage and the higher output capacitance.

A theoretical background to the problem is described in Sec. 2 showing that a conventional delay model is inadequate to explain the delay imbalance, as shown in Figs. 2 and 3. The CMOS inverter delay is discussed in Sec. 3 considering the Miller capacitance, threshold voltage variations, and a ramp shaped input voltage. Additional simulation results are provided in Sec. 4 confirming the model described in Sec. 3. Based on these analytic and experimental observations, circuit design guidelines are proposed in Sec. 5 to minimize the Miller effect. Some final remarks are offered in Sec. 6.

2. Theoretical Background

The switching characteristics of a CMOS inverter are analyzed in this section with a focus on those conditions that produce a difference between the rise and fall delay times. These delay times are described under ideal conditions in Sec. 2.1. The difference between the threshold voltages has a significant effect on the propagation delays, as discussed in Sec. 2.2.

2.1. *Idealized conditions*

Under the assumption of a step input voltage, the delay time is the time required for the output voltage to reach half of the supply voltage $V_{50\%}$ (from either zero or V_{DD}). The analysis is based on the classic expression for the current through a capacitor,

$$i_c(t) = C \cdot \frac{dV_c(t)}{dt} \,. \tag{1}$$

Applying (1) at the load capacitance of a CMOS inverter, the rise and fall propagation delay times are obtained as[4]

$$\tau_{\mathrm{PLH}} = \frac{C_L \cdot \Delta V_{LH}}{I_{\mathrm{avg}LH}} = \frac{C_L \cdot V_{50\%}}{I_{\mathrm{avg}LH}} \,, \tag{2}$$

$$\tau_{\mathrm{PHL}} = \frac{C_L \cdot \Delta V_{HL}}{I_{\mathrm{avg}HL}} = \frac{C_L \cdot (V_{DD} - V_{50\%})}{I_{\mathrm{avg}HL}} \,, \tag{3}$$

where $I_{\mathrm{avg}HL}$ and $I_{\mathrm{avg}LH}$ are the average currents, respectively, to discharge and charge a lumped load capacitance C_L.

The transient behavior of a CMOS inverter can be described by five regions of operation of the two MOSFET transistors.[4] Assuming idealized conditions, only one MOSFET transistor conducts during each transition while the other transistor is immediately turned off by the step input voltage. Under this simplified condition, delay expressions for the rising and falling transitions of the output voltage are[4]

$$\tau_{\mathrm{PLH}} = \frac{C_L}{k_p(V_{DD} - |V_{Tp}|)}\left[\frac{2 \cdot |V_{Tp}|}{V_{DD} - |V_{Tp}|} + \ln\left(\frac{4(V_{DD} - |V_{Tp}|)}{V_{DD}} - 1\right)\right], \tag{4}$$

$$\tau_{\mathrm{PHL}} = \frac{C_L}{k_n(V_{DD} - V_{Tn})}\left[\frac{2V_{Tn}}{V_{DD} - V_{Tn}} + \ln\left(\frac{4(V_{DD} - V_{Tn})}{V_{DD}} - 1\right)\right], \tag{5}$$

where k_n and k_p are the transconductance of the two transistors and V_{Tn} and V_{Tp} are the corresponding threshold voltages.

From Eqs. (4) and (5), if $V_{Tn} = |V_{Tp}|$, the condition for equal rise and fall delay times $\tau_{\mathrm{PHL}} = \tau_{\mathrm{PLH}}$ reduces to the equality of the NMOS and PMOS transconductances $k_n = k_p$. The MOSFET transconductance is a function of physical process parameters,

$$k = \frac{\mu \cdot C_{ox}}{2} \cdot \frac{W}{L}, \tag{6}$$

where μ is the mobility of the dominant charge carriers — μ_n for the electrons in the NMOS transistor and μ_p for the holes in the PMOS transistor; the unit gate oxide capacitance $C_{ox} = \varepsilon_{ox}/t_{ox}$ is a constant for a given technology, and depends on the dielectric constant ε_{ox} and the thickness t_{ox} of the gate oxide; and L is the effective channel length of the two MOSFET transistors.

The electron mobility μ_n is significantly higher as compared to the mobility of the holes μ_p. Neglecting secondary effects, the ratio μ_n/μ_p is constant and can be compensated by the design parameters W_p and W_n, such that the pull-up and pull-down MOSFET transconductances are equal.

As discussed in this section, under the assumptions of a lumped output load capacitance C_L, step input voltage waveform, single transistor charging/discharging the output capacitance (the other transistor is turned off at the edge of the input voltage), $V_{Tn} = |V_{Tp}|$, and constant ratio μ_n/μ_p, the rise and fall delay times τ_{PHL} and τ_{PLH} are equal if

$$\beta_{\mathrm{opt}} = \frac{W_p}{W_n} = \frac{\mu_n}{\mu_p}. \tag{7}$$

Note that according to Eq. (7), the optimal W_p/W_n ratio is not a function of the load capacitance C_L. This classic result postulates that the CMOS inverter can be designed for an arbitrary load capacitance C_L, maintaining the property of equal rise and fall delay times for varying C_L. This behavior, however, contradicts the empirical observations illustrated in Figs. 2 and 3. In the following sections, more realistic conditions are discussed along with the corresponding effects on the imbalance between the low-to-high and high-to-low delay times.

Table 1. Differences between V_{Tn} and V_{Tp}.

	TSMC CMOS technology	
	0.18 μm (see Ref. 8)	0.25 μm (see Ref. 9)
V_{Tn}	511 mV	457 mV
$\lvert V_{Tp}\rvert$	479 mV	653 mV
Difference (%)	6.8	42.8

2.2. *Unequal threshold voltages*

In CMOS technologies, the NMOS and PMOS devices are not manufactured with equal threshold voltages. The difference $V_{Tn} - |V_{Tp}|$ can be substantial in deep submicrometer technologies, as listed in Table 1. Manufacturers of CMOS integrated circuits provide different SPICE models for varying channel widths and lengths of MOSFET transistors.[6–9] While many parameters are assumed constant in MOSFET transistor models, the threshold voltage of both the PMOS and NMOS transistors exhibits significant variations.[8,9] The model parameters listed in Table 1 are for two different NMOS and PMOS technologies.

For equal transconductances $k_n = k_p$, the delay expressions (4) and (5), become linear functions of the load capacitance C_L with factors depending upon the threshold voltages, V_{Tn} and V_{Tp}. Differences in these factors can produce a significant imbalance between the rise and fall delay times. The condition for the W_p/W_n ratio to maintain balanced rise and fall delay times therefore changes to

$$\beta_{\text{opt}} = \frac{W_p}{W_n} = A(V_{Tn}, V_{Tp}) \cdot \frac{\mu_n}{\mu_p}\,, \tag{8}$$

where

$$A(V_{Tn}, V_{Tp}) = \frac{(V_{DD} - V_{Tn})}{(V_{DD} - |V_{Tp}|)} \frac{\left[\frac{2\cdot|V_{Tp}|}{V_{DD}-|V_{Tp}|} + \ln\left(\frac{4(V_{DD}-|V_{Tp}|)}{V_{DD}} - 1\right)\right]}{\left[\frac{2V_{Tn}}{V_{DD}-V_{Tn}} + \ln\left(\frac{4(V_{DD}-V_{Tn})}{V_{DD}} - 1\right)\right]}\,. \tag{9}$$

Rather than only being a function of the carrier mobility parameters, the optimal W_p/W_n ratio is also a function of the threshold voltages of the two transistors through the factor $A(V_{Tn}, V_{Tp})$. Note that the optimal ratio W_p/W_n, expressed in Eq. (9), is not a function of the load capacitance C_L or the channel width W_n.

3. MOSFET Device Parameter Variations

The effect of the Miller capacitance on the delay imbalance is analyzed in Sec. 3.1 assuming a step input waveform. Variations in the threshold voltage are described in Sec. 3.2. The effect of a finite slope ramp input waveform is discussed in Sec. 3.3.

3.1. *Miller effect*

The Miller capacitance C_M of a CMOS inverter is shown in Fig. 1. This parasitic capacitance is primarily formed by the gate-to-drain capacitances of the NMOS and PMOS devices. For each transition of the output voltage waveform, the Miller capacitance is discharged and then charged in the opposite direction. The effective parasitic capacitance, from an energy perspective, is therefore $2C_M$. From a delay perspective, however, significant nonlinearities are introduced by the Miller capacitance caused by changes in the operating regions of the two transistors.[4,5] Under the assumption of a step input waveform, the discharging current through the load capacitance C_L is

$$i_c(t) = C_L \cdot \frac{dV_{\text{out}}}{dt} = C_M\left(\frac{dV_{\text{in}}}{dt} - \frac{dV_{\text{out}}}{dt}\right) - i_n\,. \tag{10}$$

A delay model based on Eq. (10) is described in Appendix A, yielding

$$\tau_{\text{PHL}} = \frac{C_L + 3C_{M\,\text{fall}}}{2} \cdot \frac{V_{DD}}{I_{D0_n}}\,, \tag{11}$$

$$\tau_{\text{PLH}} = \frac{C_L + 3C_{M\,\text{rise}}}{2} \cdot \frac{V_{DD}}{I_{D0_n}}\,. \tag{12}$$

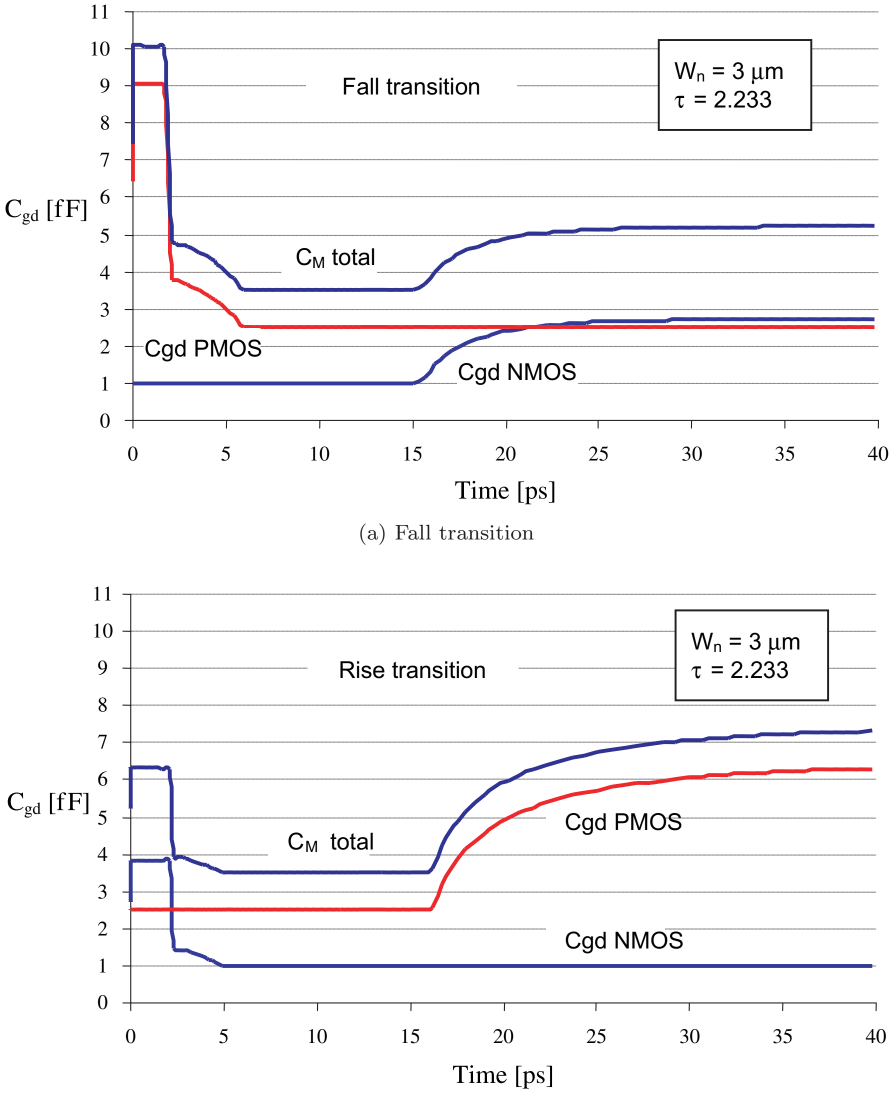
Expressions (11) and (12) reveal the relative significance of the Miller capacitance in the delay analysis of a CMOS inverter. For constant Miller capacitances and equal saturation currents, $I_{D0n} = I_{D0p}$, the output delays remain balanced as the load capacitance C_L is varied.

The MOSFET gate-to-drain capacitances C_{gd}, however, vary significantly with the applied terminal voltages.[4,5] The gate-to-drain capacitance can be expressed as

$$C_{gd} = C_{gd_{\text{overlap}}} + C_{gd_{\text{channel}}}\,, \tag{13}$$

where $C_{gd_{\text{overlap}}}$ is a static parallel-plate capacitor formed by the gate-to-drain overlap. $C_{gd_{\text{channel}}}$ is a dynamic capacitance which represents the distributed nature of the MOS transistor channel between the source and drain. When an inversion layer is formed, the gate oxide capacitance is shared between the source and drain.[4,5] In the saturation and cut-off regions, only the static overlap capacitance forms C_{gd}, while in the linear region, $C_{gd_{\text{channel}}}$ is added. Significant dynamic variations in the Miller capacitance during the fall and rise transitions of a CMOS inverter with a step input waveform are shown in Fig. 6. During each transition, the NMOS and PMOS devices switch into different operating regions. For the fall transition (see Fig. 6(a)), the PMOS transistor initially operates in the linear region (maximum C_{gd}) while the NMOS transistor operates in the saturation region (minimum C_{gd}). During the transition, the PMOS transistor gradually turns off and the NMOS transistor enters the linear region. The significant difference in the total Miller capacitance is due to the larger PMOS device.

The difference between the effective Miller capacitances produces a corresponding imbalance in the output delay, as expressed in (11) and (12). $C_{M\,\text{fall}}$ and $C_{M\,\text{rise}}$



(a) Fall transition

(b) Rise transition

Fig. 6. Dynamic variations of the C_{gd} capacitances in a CMOS inverter.

vary with any change in the primary circuit parameters C_L, W_n, and τ as the transient behavior of the CMOS inverter is modified. The longer the short-circuiting transistor operates in the linear region, the greater the difference in the effective Miller capacitance.

Expression (A.5) can be reconfigured into

$$V_{\text{out}}(t)_{\text{fall}} = V_{DD} + \Delta V_1 - \frac{I_{D0_n}}{C_L + C_{M\,\text{fall}}} \cdot t\,, \tag{14}$$

where

$$\Delta V_1 = \frac{C_{M\,\mathrm{fall}}}{C_L + C_{M\,\mathrm{fall}}} \cdot V_{DD} \tag{15}$$

to distinguish the two effects of the Miller capacitance on the output transition time: (1) the *overshoot* voltage ΔV_1 — the initial value of the output voltage is offset from V_{DD} and (2) the slope of the output waveform is reduced. Both effects increase the output delay time. By analogy, in the rising transition, there is a corresponding *undershoot* offset from ground ΔV_2,

$$\Delta V_2 = \frac{C_{M\,\mathrm{rise}}}{C_L + C_{M\,\mathrm{rise}}} \cdot V_{DD}\,, \tag{16}$$

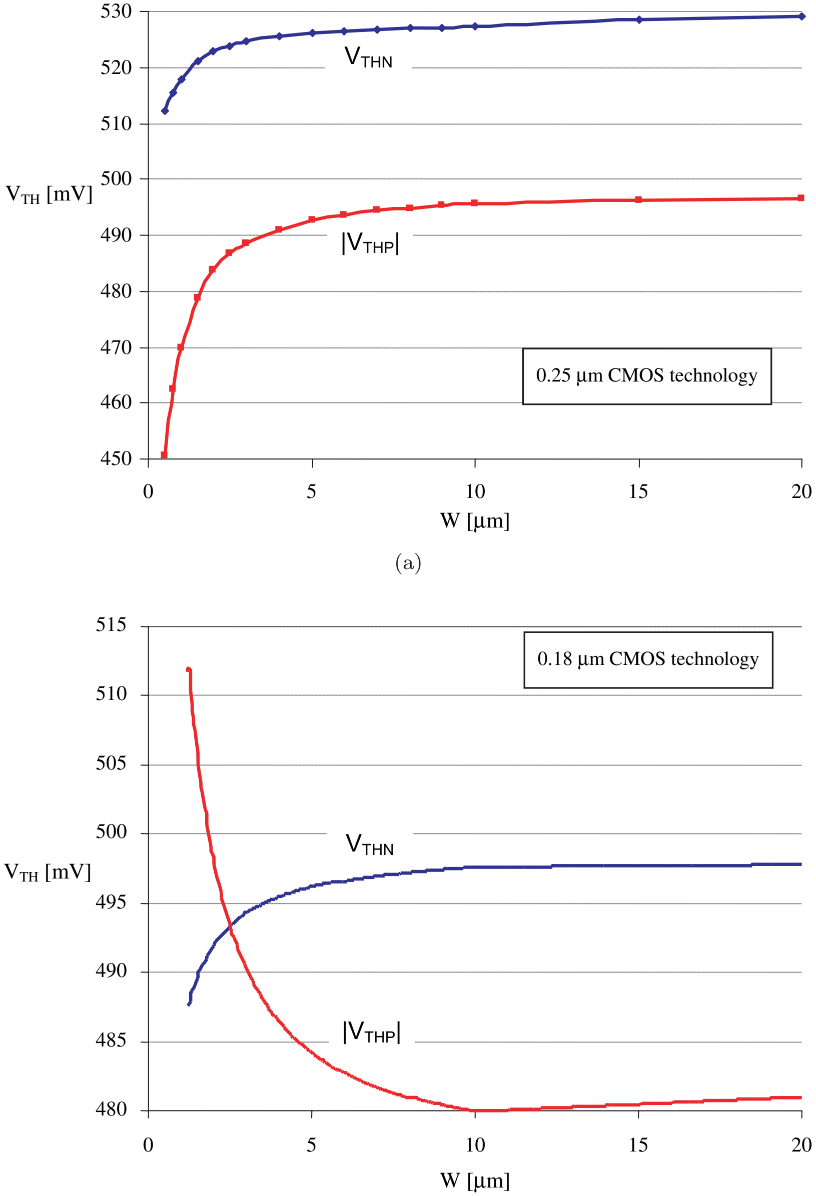
$$V_{\mathrm{out}}(t)_{\mathrm{rise}} = -\Delta V_2 + \frac{I_{D0_p}}{C_L + C_{M\,\mathrm{rise}}} \cdot t\,. \tag{17}$$

3.2. *Threshold voltage variations*

Short- and narrow-channel effects in deep submicrometer technologies produce significant variations in the MOSFET threshold voltages V_{TH}. This V_{TH} instability is manifested as a variation in the CMOS inverter delay. For a constant channel length and no body bias, V_{TH} variations are primarily due to the narrow-channel effect and drain-induced barrier lowering (DIBL).[5] Simulation data for CMOS inverters in 180 and 250 nm CMOS processes are shown in Figs. 7 and 8 and discussed later.

The narrow-channel effect in MOS devices is due to the increased significance of the gate-to-channel fringing field near the drain as the channel width is reduced. This additional electric field has an opposite effect on the MOSFET threshold voltage depending upon the field oxidation process technology, local oxidation of silicon (LOCOS) or shallow trench isolation (STI).[5] Modern CMOS processes typically employ STI to precisely form the channel region. In STI, the near-drain fringing field enhances the depletion of the channel and can be represented as an additional gate-to-channel capacitance. The effective threshold voltage is therefore reduced as shown in Fig. 7 as a lower gate voltage is sufficient to form the inversion layer. Increasing $|V_{\mathrm{THP}}|$ with decreasing transistor width W, as shown in Fig. 7(b) for the 180 nm CMOS process, is not typical for STI transistors[5] and may be due to foundry process overcompensation for the STI narrow width effect in order to reduce leakage reduction.

The MOSFET transistor threshold voltage is typically determined under the assumption of negligible drain-to-source voltage V_{DS}. In short-channel devices, however, the V_{DS}-modulated depletion region around the drain becomes significant. This depletion region reduces the amount of charge supplied from the gate required to invert the channel. The effective threshold voltage is thereby reduced as shown in Fig. 8. This phenomenon is what is known as DIBL. Because of process differences between the formation of the NMOS and PMOS devices, DIBL induced



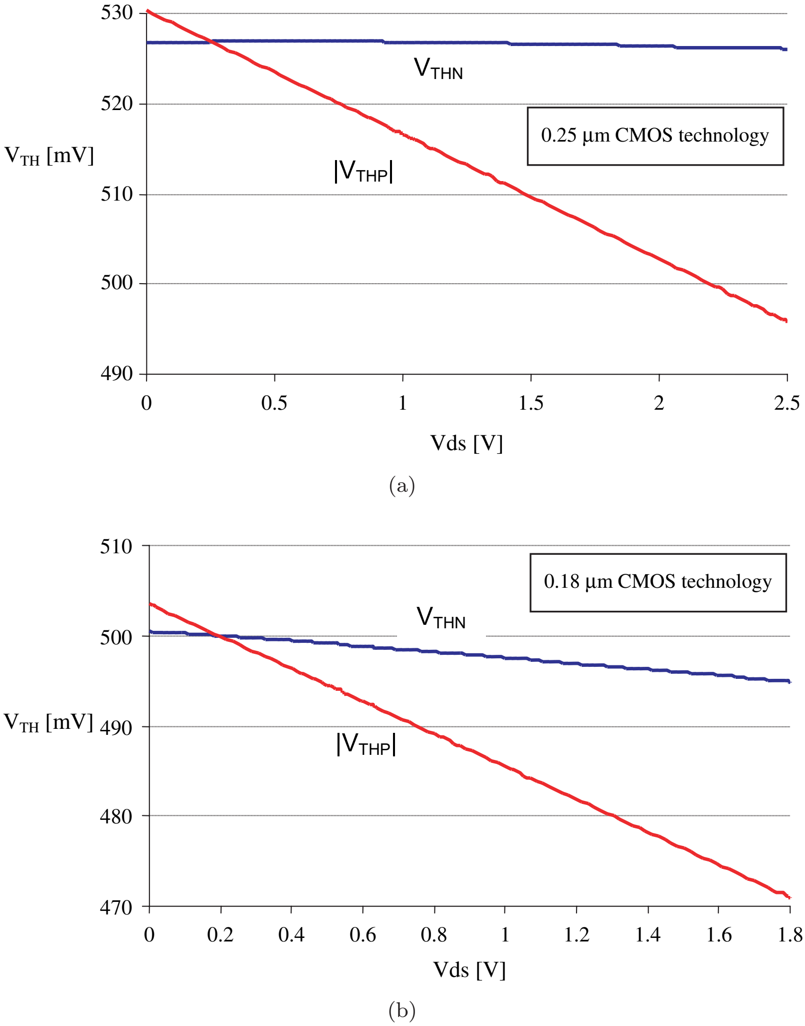
(a)

(b)

Fig. 7. Narrow-channel induced variations in threshold voltage for a (a) 0.25 μm CMOS technology.[8] (b) 0.18 μm CMOS technology.[9]

variations in $|V_{\mathrm{THP}}|$ are significantly larger as compared to DIBL induced variations in V_{THN} (see Fig. 8).

Threshold voltage variations directly affect the MOS transistor currents in all operating regions and correspondingly produce an imbalance in the rise and fall



(a)

(b)

Fig. 8. Variations in threshold voltage due to drain-induced barrier lowering (a) 0.25 μm CMOS technology.[8] (b) 0.18 μm CMOS technology.[9]

delay times. This imbalance is more pronounced in the 180 nm technology as shown in Fig. 7(b) where the threshold voltage difference grows significantly with W (since V_{THN} and $|V_{\mathrm{THP}}|$ change in opposite directions).

3.3. *Ramp input voltage and the Miller effect*

In practical integrated circuits, the input voltage waveform differs from a step function. The primary effects of a nonzero input transition time are (1) a gradual transition of the CMOS inverter from one region of operation to another region of operation and (2) the Miller effect is linearly reduced with linearly increasing input transition time.

Rather than only one transistor conducting current to/from C_L, both transistors actively participate in the current conduction process and switch through several different operating regions,

$$i_c(t) = C_L \cdot \frac{dV_{\text{out}}}{dt} = C_M \left(\frac{dV_{\text{in}}}{dt} - \frac{dV_{\text{out}}}{dt} \right) + i_p - i_n \,. \tag{18}$$

Equation (18) can be solved from multiple differential equations for different transition relations between V_{in} and V_{out}. Jeppson[2] presented an expression for the CMOS inverter fall delay time τ_{PHL} based on the model shown in Fig. 1 and expressed by Eq. (18),

$$\tau_{\text{PHL}} = \frac{1+2n}{6}\tau + \frac{\Delta Q}{I_n} + \underbrace{\frac{C_L \cdot V_{DD}}{I_n}\left(\frac{C_M}{C_M + C_L} + f(V_{\text{Dsat}}) \right)}_{\tau_{\text{Dstep}}}, \tag{19}$$

where V_{DD} is the power supply voltage, $n = V_{Tn}/V_{DD}$, I_n is the NMOS saturation current, $f(V_{\text{Dsat}})$ is a function of the NMOS drain-to-source saturation voltage, and τ_{Dstep} is the output delay time for a step input voltage given by Eq. (5). The effect of the PMOS transistor during the discharge transition is represented by the charge contributed by the PMOS transistor to the output node,

$$\Delta Q = (C_M + C_L) \cdot (\Delta V_1 - V_{DD}) \,, \tag{20}$$

where ΔV_1 is the overshoot voltage. This additional charge ΔQ contributes an output delay time of $\Delta Q/I_n$. Due to the different NMOS and PMOS device parameters, the three terms in Eq. (19) contribute to the delay imbalance when C_L, C_M, and τ are varied. The contribution of the second (short-circuiting) transistor depends on both capacitances $(C_M + C_L)$, as well as the transistor transconductance through the contributed charge ΔQ. In addition to the difference $V_{Tn} - |V_{Tp}|$, the transistor operating modes differ due to the change in the drain-to-source saturation voltages V_{Dsat}.

While a complete analytic expression for the optimal W_p/W_n ratio is not provided here, the delay model described by Eqs. (18) and (19) provides an intuitive perspective to the conflicting effects of input transition time on the output delay imbalance. As the rate of change of the input-to-output voltage is reduced (from the zero input transition modeled in Sec. 3.1), the Miller effect is reduced as described by Eqs. (10) and (18). The overshoot and undershoot voltages, (15) and (16), respectively, are correspondingly reduced. At the same time, the short-circuit power dissipation increases approximately linearly with the input transition time τ.[10] Increasing the input transition time therefore has conflicting effects on the output delay time: (1) reducing the output delay by reducing the Miller effect and (2) increasing the output delay due to larger short-circuit contention currents.

With fast input transitions, the imbalance in the output delay is primarily due to the difference in the effective Miller capacitances, as shown in Fig. 6. As the Miller effect is reduced for slower inputs, the contention current increases and any variations in the NMOS and PMOS threshold voltages, illustrated in Figs. 7 and 8, contribute significantly to the imbalance in the output delay.

The Miller effect is observed by the variations in the overshoot and undershoot voltages shown in Fig. 9. The significant difference between ΔV_1 and ΔV_2 is explained by the different initial values of $C_{M\,\mathrm{fall}}$ and $C_{M\,\mathrm{rise}}$, as shown in Fig. 6.

The increasing difference $\Delta V_1 - \Delta V_2$ produces a corresponding difference between τ_{PHL} and τ_{PLH}. Accounting for the ramp input voltage, Miller effect, and threshold voltage variations, balancing the rise and fall delay times requires adjusting the optimal W_p/W_n ratio, as depicted in Figs. 3, 4, and 10.

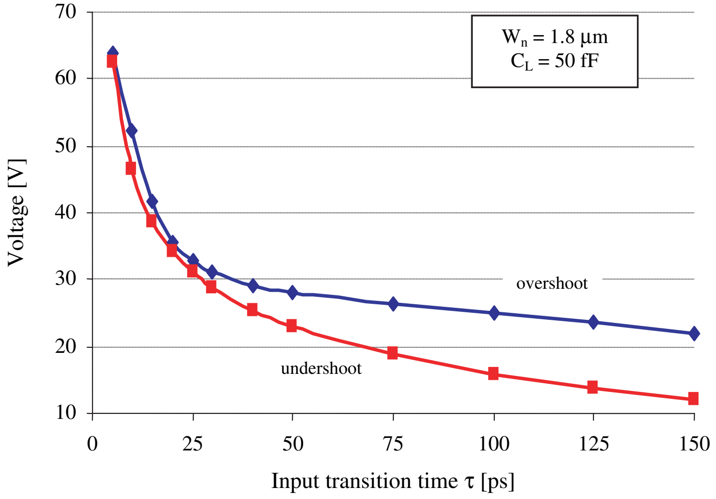


Fig. 9. Variations of CMOS inverter over/undershoot voltages with input transition time τ.

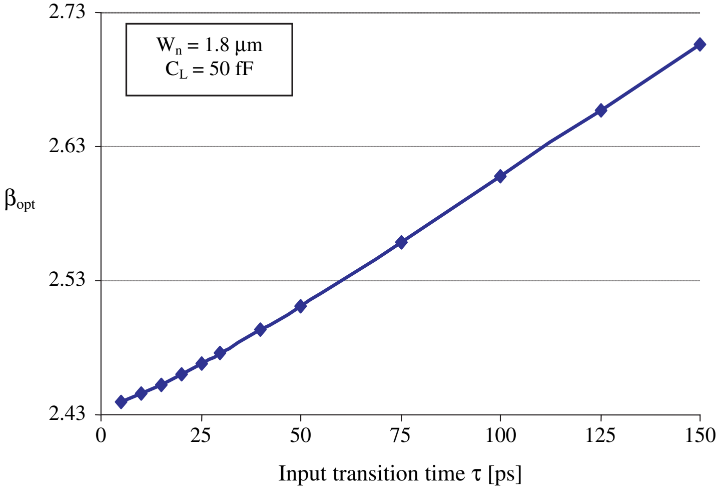


Fig. 10. Variations of β_{opt} with input transition time τ.

4. Simulation Data

A set of curves characterizing the variation of the optimal W_p/W_n ratio is shown in Figs. 11 and 12 for $W_n = 1.8\,\mu$m in a $0.18\,\mu$m CMOS technology.[8] The primary parameters are the load capacitance C_L, the input transition time τ, and the channel width of the NMOS transistor W_n (affecting the coupling capacitance C_M).

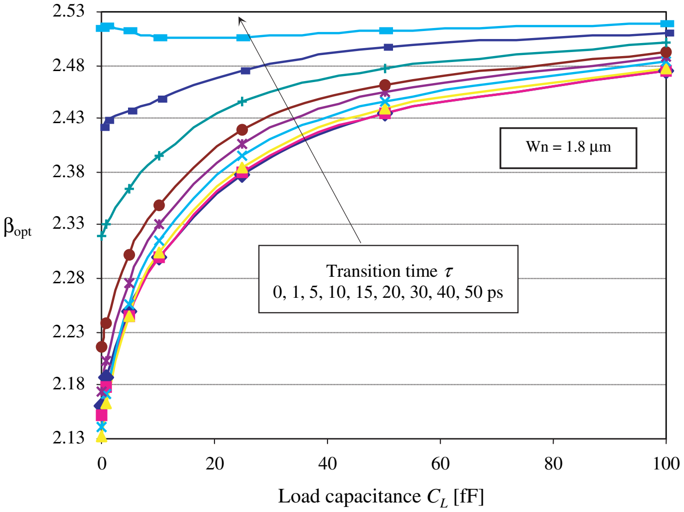


Fig. 11. Optimal W_p/W_n ratio as a function of the load capacitance C_L and input transition time τ.

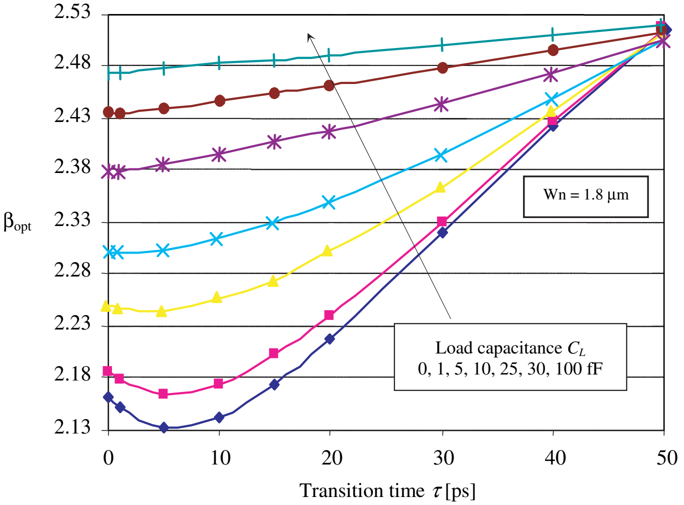


Fig. 12. Optimal W_p/W_n ratio as a function of the input transition time τ and load capacitance C_L.

The simulations shown in Figs. 11 and 12 agree with the theoretical discussion presented in Sec. 3. As the load capacitance C_L and the input transition time τ is increased, the influence of the Miller effect on the behavior of a CMOS inverter is reduced and the W_p/W_n ratio converges to a constant. The W_p/W_n inflection point shown in Fig. 11 illustrates the balance between the conflicting effects of the Miller capacitance C_M and the short-circuiting transistor. For relatively large C_M (with respect to C_L), as the transition time τ is increased, the Miller effect is reduced. The short-circuiting transistor remains active for a longer portion of the transition time, thereby contributing a larger ΔQ to the output node (increasing the contention current). The position and depth of the W_p/W_n minimum illustrated in Fig. 11 is a function of the C_M/C_L ratio and τ. As the Miller effect is reduced, the optimal W_p/W_n ratio behaves linearly with longer input transition time τ, primarily due to the short-circuit current expressed by the third term in Eq. (19).

Certain parameters of the BSIM3 MOSFET device models[6–9] are varied to observe significant variations in the optimal W_p/W_n ratio produced by short channel length effects. The optimal W_p/W_n ratio is substantially affected by variations in the carrier velocity saturation, mobility reduction coefficients, channel length modulation coefficient, and drain-induced barrier lowering coefficients. The difference in these model parameters between the NMOS and PMOS devices affects the switching characteristics of the CMOS inverter, contributing to the imbalance between the rise and fall delay times.

5. Circuit Design Guidelines

A general circuit design approach that considers variations in the optimal W_p/W_n ratio is to size the CMOS inverter for a small C_M/C_L ratio, thereby reducing the Miller effect. This behavior is supported by Figs. 11 and 12. As shown in Fig. 12 and described by Eq. (18), the Miller effect is substantially reduced for longer input transition times, permitting slower driver circuits to be used without significantly degrading the output delay time. For fast CMOS buffers with large channel widths (and therefore large Miller capacitances), Eq. (18) suggests sizing the stages for equal input and output transition times or an equal current-to-capacitance ratio to minimize the Miller effect.

6. Concluding Remarks

The ratio between the channel width of the PMOS and NMOS transistors (W_p/W_n) is a primary design parameter in CMOS circuits which is used to equalize the rise and fall delay times. It is shown in this paper that due to parasitic effects, it is suboptimal in terms of delay to maintain a fixed W_p/W_n ratio while other circuit parameters are varied. The primary parameters that affect the optimal W_p/W_n ratio (at which the rise and fall delays are equal) are the load capacitance C_L, the transition time of the input voltage τ, and the channel width of the NMOS transistor W_n. The issue of how best to adjust the W_p/W_n ratio is discussed and confirmed by SPICE simulations. For fast input transitions, the delay imbalance is primarily due to differences in the effective Miller capacitance during the rise and fall transitions. As the Miller effect is reduced for slower inputs, the contention current increases and variations in the NMOS and PMOS threshold voltages due to small channel effects contribute significantly to the imbalance in the output delay. In order to manage these effects, circuit design guidelines are proposed based on analytic and experimental observations of the Miller effect. These effects on the output delay imbalance in CMOS inverters are expected to increase in future deep submicrometer technologies.

Acknowledgments

This research is supported in part by the Semiconductor Research Corporation under Contract No. 2003-TJ-1068 and 2004-TJ-1207, the National Science Foundation under Contract No. CCR-0304574, the Fulbright Program under Grant No. 87481764, a grant from the New York State Office of Science, Technology & Academic Research to the Center of Advanced Technology in Electronic Imaging Systems, and by grants from Xerox Corporation, IBM Corporation, Intel Corporation, Lucent Technologies Corporation, and Eastman Kodak Company.

Appendix A. CMOS Inverter Delay Model Considering the Miller Capacitance with Step Input Waveform

A CMOS inverter delay model is derived considering the effect of the Miller capacitance assuming a step input waveform. With a step input transition time, the short-circuiting transistor is cut off and only the primary transistor conducts current during each transition. The differential equation for the fall output transition is

$$i_c(t) = C_L \cdot \frac{dV_{\text{out}}}{dt} = C_M\left(\frac{dV_{\text{in}}}{dt} - \frac{dV_{\text{out}}}{dt}\right) - i_n\,. \tag{A.1}$$

Conventional Laplace domain analysis yields

$$C_L[sV_{\text{out}}(s) - V_{\text{out}}(0-)] + C_M[sV_{\text{out}}(s) - sV_{\text{in}}(s) - V_{C_M}(0-)] = -i_n(s)\,, \tag{A.2}$$

$$s(C_L + C_M)V_{\text{out}}(s) - (C_L + 2C_M)V_{DD} = -i_n(s)\,, \tag{A.3}$$

$$V_{\text{out}}(s) = \frac{C_L + 2C_M}{C_L + C_M} \cdot \frac{V_{DD}}{s} - \frac{i_n(s)}{C_L + C_M} \cdot \frac{1}{s}\,. \tag{A.4}$$

In deep sub-micrometer technologies, the drain-to-source saturation voltage V_{Dsat} is typically lower than half of the supply voltage V_{DD}. Therefore, before the output voltage reaches $V_{DD}/2$, the NMOS transistor only conducts current in the saturation region and i_n can be expressed by a constant average saturation current I_{D0n},

$$V_{\text{out}}(t) = \frac{C_L + 2C_M}{C_L + C_M} \cdot V_{DD} - \frac{I_{D0_n}}{C_L + C_M} \cdot t\,. \tag{A.5}$$

$V_{\text{out}}(t)$ reaches half of the supply voltage at time

$$\tau_{\text{PHL}} = \frac{C_L + 3C_{M\,\text{fall}}}{2} \cdot \frac{V_{DD}}{I_{D0_n}}\,. \tag{A.6}$$

By analogy, the low-to-high delay can be expressed as

$$\tau_{\text{PLH}} = \frac{C_L + 3C_{M\,\text{rise}}}{2} \cdot \frac{V_{DD}}{I_{D0_P}}\,. \tag{A.7}$$

The Miller capacitances C_M in Eqs. (A.6) and (A.7) are replaced by effective Miller capacitances $C_{M\,\text{fall}}$ and $C_{M\,\text{rise}}$. As described in Sec. 3, the difference between these capacitances produces a delay imbalance between the rising and falling output delay times.

References

1. N. Hedenstierna and K. Jeppson, CMOS circuit speed and buffer optimization, *IEEE Trans. Computer-Aided Design of Integrated Circuits Syst.* **CAD-6** (1987) 270–281.
2. K. Jeppson, Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay, *IEEE J. Solid-State Circuits* **29** (1994) 646–654.
3. D. Kung and R. Puri, Optimal P/N width ratio selection for standard cell libraries, *Proc. IEEE Int. Conf. Computer-Aided Design* (1999), pp. 178–184.
4. S.-M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design* (McGraw-Hill, 1999).
5. Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd edn. (WCB/McGraw-Hill, 1999).
6. B. Sheu, D. Scharfetter, P. Ko and M. C. Jeng, BSIM: Berkeley short-channel IGFET model for MOS transistors, *IEEE J. Solid-State Circuits* **SC-22** (1987) 458–466.
7. W. Liu *et al.*, *BSIM3v3.2 MOSFET Model — Users' Manual* (University of California, Berkeley, April 1999).
8. TSMC 0.18 μm 1P6M Salicide 1.8 V SPICE Models (July 1999).
9. TSMC 0.25 μm Logic Salicide (1 PM5, 2.5 V) SPICE Models (November 1998).
10. K. Nose and T. Sakurai, Analysis and future trend of short-circuit power, *IEEE Trans. Computer-Aided Design of Integrated Circuits Syst.* **19** (2000) 1023–1030.