Analysis and modelling of initial delay time and its impact on propagation delay of CMOS logic gates

Y.-H. Yang  
C.-Y. Wu

Indexing terms: Metal-Oxide-semiconductor structures, Modelling

Abstract: The initial delay times due to the capacitive feedthrough effects in CMOS inverters are characterised and investigated. Based on the MOSFET large-signal model, the initial delay is modelled for a chain of CMOS inverters under step and ramp inputs. Optimal design that results in the minimum initial delay is obtained. Correlation between the initial delay and the propagation delay is constructed in the case of characteristic waveforms. The initial delays are found to determine the propagation delay. Applying the model to evaluate the speed performance of a scaled-down CMOS, the delay improvements for various scaling laws are compared. It is found that the most effective law in reducing the initial delay for internal circuits is the constant voltage law, whereas that for the input stage is the constant electric field law. Comparisons to SPICE simulation results are also given and good agreement is achieved.

List of symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{S,D}$</td>
<td>area of source (drain) region</td>
</tr>
<tr>
<td>$C_f$</td>
<td>zero-bias bulk-junction bottom capacitance per unit area</td>
</tr>
<tr>
<td>$C_{JSW}$</td>
<td>zero-bias bulk-junction sidewall capacitance per unit length</td>
</tr>
<tr>
<td>$C_{BDS}$</td>
<td>bulk-drain (source) junction capacitance</td>
</tr>
<tr>
<td>$C_{SOE}$</td>
<td>gate-source (drain) overlap capacitance</td>
</tr>
<tr>
<td>$C_L$</td>
<td>loading capacitance of chain of CMOS inverters in each stage</td>
</tr>
<tr>
<td>$C_{Ox,PS}$</td>
<td>oxide capacitance of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$\delta$</td>
<td>parameter of narrow width effect for threshold voltage</td>
</tr>
<tr>
<td>$L_{eff}$</td>
<td>effective channel length of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$M_{JSW,p}$</td>
<td>bulk-junction sidewall grading coefficient of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$N_{DG}$</td>
<td>substrate doping concentration</td>
</tr>
<tr>
<td>$P_{S,D}$</td>
<td>perimeter of source (drain) region</td>
</tr>
<tr>
<td>$q$</td>
<td>magnitude of electronic charge</td>
</tr>
<tr>
<td>$T_{ox}$</td>
<td>gate oxide thickness</td>
</tr>
<tr>
<td>$V_{max}$</td>
<td>maximum drift velocity of carrier</td>
</tr>
<tr>
<td>$V_T$</td>
<td>threshold voltage under zero bias</td>
</tr>
<tr>
<td>$V_{T_{ox}}$</td>
<td>threshold voltage of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$V_{pp}$</td>
<td>voltage drop between bulk and source regions</td>
</tr>
<tr>
<td>$V_T$</td>
<td>thermal voltage</td>
</tr>
<tr>
<td>$W_{p}$</td>
<td>effective channel width of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$W_{ex}$</td>
<td>dimension of source/drain regions in channel length direction</td>
</tr>
<tr>
<td>$X_J$</td>
<td>metallurgical junction depth</td>
</tr>
<tr>
<td>$a_{st}$</td>
<td>geometrical factor of source (drain) junction with short-channel effect</td>
</tr>
<tr>
<td>$\beta_{n,p}$</td>
<td>transconductance parameter of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$\gamma_{n,p}$</td>
<td>channel-length modulation parameter of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$\gamma_{m,p}$</td>
<td>bulk threshold parameter of long channel n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$\epsilon_{Silox}$</td>
<td>permittivity of silicon (silicon dioxide)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>surface mobility</td>
</tr>
<tr>
<td>$\mu_{n,p}$</td>
<td>mobility of n-(p-) channel MOSFET</td>
</tr>
<tr>
<td>$\nu_{n,p}$</td>
<td>critical field coefficient for mobility degradation</td>
</tr>
<tr>
<td>$\nu_{n,p}$</td>
<td>critical field exponent for mobility degradation</td>
</tr>
<tr>
<td>$\phi_f$</td>
<td>Fermi potential</td>
</tr>
</tbody>
</table>

1 Introduction

The advantages of low-power consumption and high-noise immunity make CMOS the main technology in VLSI and ULSI [1]. Ideally, the signal of a CMOS logic gate is rail-to-rail between the power supply ($V_{DD}$) and ground (GND). However, transient simulations show that the actual transient voltage level at the output node of a CMOS logic gate is greater than $V_{DD}$ or smaller than GND during certain periods. When this voltage overshoot or undershoot occurs, a charging or discharging current, opposite in direction to the normal output device currents, is generated on the device capacitance and led to the output node of the gate to increase the voltage above $V_{DD}$ or decrease it below GND. Owing to such a capacitive feedthrough effect, extra delay time, called the initial delay time, is needed to remove those excess feedthrough charges at the output node and recover the signal to its normal level. It is found that this initial delay time usually dominates the total signal delay [2], especially in the case that the input waveform has finite rise or fall time. The speed performance of a CMOS gate is therefore strongly affected by the initial delay times. However, it has not yet been well characterised in recent timing papers [2-7].
To obtain a better understanding of initial delays, and since the behaviour of other CMOS inverting logic gates can be understood from that of the basic inverter, initial delays of CMOS inverters are investigated and analysed in this paper. Taking the physical nature of transient waveforms into consideration, and using the analytical current equations [8–11] which consider the small-geometry effects, the initial delays of CMOS inverters under step and ramp input excitations are analytically modelled. Based on the developed analytical models, the dependence of initial delay on the input slope, device parameters and the loading capacitance is investigated. Optimal device dimension resulting in minimum initial delay can also be derived.

Comparisons between the model calculations and SPICE [8] simulations for wide ranges of device dimensions, input slope and loading capacitance are given for step and ramp inputs, respectively. Good agreement is obtained. The correlation between the initial delay and the propagation delay of CMOS inverters is observed in the characteristic waveform case [12]. The model developed in Section 2.2 is modified to account for the dependence of input slope on device parameters, in the characteristic waveform case, and is applied to investigate the initial delay or the propagation delay of CMOS inverters scaled down according to various scaling laws [13, 14]. The experimental results of initial delay for various device dimensions and input ramp rates are also shown.

2 Models

2.1 Step input

2.1.1 Initial fall delay: Consider a chain of p-well CMOS inverters as shown in Fig. 1. When a step input voltage \( V(t) \) raises to \( V_{dd} \) instantly, the output voltage \( V(t) \) first overshoots above \( V_{dd} \) and then decreases back to \( V_{dd} \) at \( t = t_{df} \) (the initial fall delay). The input-voltage waveform and part of the output waveform are shown in Fig. 2a by solid lines in an enlarged scale; the waveforms in the undershoot case are also shown.

To determine the operating regions of MOSFETs, the drain-source saturation voltage \( V_{ds(app)} \), considering the velocity saturation effect, is calculated. The resultant \( V_{ds(app)} \) curve for the NMOS \( M_{1s} \) in the overshoot case is shown in Fig. 2a. It can be seen from Fig. 2a that \( M_{1s} \) is operating in the saturation region from \( t = 0 \) to \( t = t_{df} \). During this period, however, the output node becomes the source node of the PMOS \( M_{1p} \) because its voltage is greater than \( V_{dd} \), whereas the node connected to the power supply \( V_{dd} \) acts as the drain. This is opposite to the normal case and a positive substrate bias results, so that the magnitude of the PMOS threshold voltage is reduced [9]. \( M_{1p} \) will be turned on in the saturation region for a sufficiently large voltage overshoot. It is found, however, that the time interval when \( M_{1p} \) is

\[
V_{dd} = 5V, 
\frac{V_{dd}}{V_{ds(app)}} = V_{ds(app)}(t) = V_{dd}(t) - \mu C_{ox} V_{ds(app)}(t) + \frac{C_{ds}}{C_{ds} + C_{os}} V_{os} (t) + \frac{C_{os}}{C_{ds} + C_{os}} V_{ds} (t) - \frac{C_{ch}}{C_{ch} + C_{os}} V_{ch} (t) - \frac{C_{os}}{C_{ch} + C_{os}} V_{os} (t)
\]

Fig. 2a Schematic diagram of input and output voltage waveforms under step input

expressed as [15]
\[
\lambda_v = \left( X_{d} \left[ \frac{V_{DD}V_{max}/2}{\mu_{n}} \right]^{2} + V_{on} - V_{max} \right)^{2.5} 
- \frac{X_{d}^{2} \left[ V_{max}/2 \right]}{\left[ \mu_{n} \right]} \left[ L_{on} \right] \left[ V_{max} \right]
\]

All other parameters involved in the \( I_{d}(t) \) expression are further expressed in Table 1.

Table 1: Device parameters used in MOSFET current equation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_{n} )</td>
<td>( \frac{W_{n}}{L_{on}} \sqrt{\frac{2}{\mu_{n}}} )</td>
</tr>
<tr>
<td>( V_{on} )</td>
<td>( V_{on} - \frac{\mu_{n}}{2\sqrt{2}} + \left( (\delta_{n} - 1) (2\sqrt{2} - V_{max}) \right) )</td>
</tr>
<tr>
<td>( q_{n} )</td>
<td>( 1 + \frac{\mu_{n} \Delta T}{4 \Delta t} ) ((\frac{\Delta t}{V_{on}})^{2.5} \ Pc_{on} )</td>
</tr>
<tr>
<td>( \lambda_{n} )</td>
<td>( \frac{\mu_{n} \Delta T}{V_{on}} \left( \frac{V_{on} - V_{r}}{V_{on} - V_{on} - \mu_{n} \min (2\Delta t, V_{max})} \right) )</td>
</tr>
<tr>
<td>( \psi_{n} )</td>
<td>( \frac{\mu_{n} \Delta T}{V_{on}} \left( \frac{V_{on} - V_{r}}{V_{on} - V_{on} - \mu_{n} \min (2\Delta t, V_{max})} \right) )</td>
</tr>
<tr>
<td>( \lambda_{p} )</td>
<td>( \frac{X_{d}}{2L_{on}} \sqrt{\frac{2}{\mu_{p}}} )</td>
</tr>
<tr>
<td>( \psi_{p} )</td>
<td>( \frac{\mu_{p} \Delta T}{V_{on}} \left( \frac{V_{on} - V_{r}}{V_{on} - V_{on} - \mu_{p} \min (2\Delta t, V_{max})} \right) )</td>
</tr>
<tr>
<td>( C_{on} )</td>
<td>( C_{on} \left( \frac{1}{V_{on}} \right) )</td>
</tr>
<tr>
<td>( C_{p} )</td>
<td>( C_{on} \left( \frac{1}{V_{on}} \right) )</td>
</tr>
<tr>
<td>( C_{g} )</td>
<td>( C_{on} \left( \frac{1}{V_{on}} \right) )</td>
</tr>
<tr>
<td>( C_{d} )</td>
<td>( C_{on} \left( \frac{1}{V_{on}} \right) )</td>
</tr>
</tbody>
</table>

The capacitances \( C_{1}, C_{2}, \text{ and } C_{3} \) in Fig. 2b can be expressed as [8, 11]
\[
C_{1} = C_{on} + C_{on} + 2C_{on}^{3} \quad (4)
\]
\[
C_{2} = C_{on} + C_{on} + 2C_{on}^{3} \quad (5)
\]
\[
C_{3} = C_{on} + C_{on} + C_{on} + C_{on} + C_{on} + C_{on} + C_{on} + C_{on} + C_{on} + C_{on} + C_{on} + C_{on} \quad (6)
\]

and the voltage-dependent p-n junction capacitances \( C_{on} \) and \( C_{on} \) can be written as [8]
\[
C_{on} = C_{on} \left( \frac{1}{V_{on}} \right) \left( \frac{1}{V_{on}} \right)
\]

In eqns. 4-6, the loading capacitance \( C_{L} \) and the gate-oxide capacitances \( C_{on} \) and \( C_{on} \) have a large effect on \( C_{1} \) and \( C_{3} \) than other capacitances.

To analytically solve the output voltage \( V_{out}(t) \), necessary simplifications must be applied to linearise the current and capacitances. First, the channel-length modulation parameters \( \lambda_{n} \) and \( \lambda_{p} \) in the equation of \( V_{out}(t) \) on \( V_{on} \). Second, the load capacitance \( C_{on} \) is assumed to be constant in solving \( V_{out}(t) \). After the explicit expression of \( V_{out}(t) \) is obtained, the voltage dependencies of \( C_{on} \) and \( C_{on} \) are incorporated into the expression of \( V_{out}(t) \) to form a nonlinear equation. Iterations are then applied to get the final result. Based on the above assumptions, \( V_{out}(t) \) can be solved as
\[
V_{out}(t) = \left[ 1 + C_{on} \left( \frac{C_{max}/2}{\mu_{n}} \right)^{2} + V_{on} \right]
+ \frac{1}{C_{on}} \left( \frac{1}{V_{on}} - 1 \right) \left( \frac{1}{V_{on}} \right)
\]

where
\[
P_{f} = \beta_{n} \lambda_{n} \frac{V_{DD}^{2}}{V_{on}^{2}} \left( C_{2} + C_{3} + C_{on} + C_{on} \right)
\]

and
\[
V_{out}^{2} = \left( \frac{V_{DD}}{V_{on} - \lambda_{n} V_{out}/2} \right)^{-2} \left( \frac{2\lambda_{n}^{2}}{2\lambda_{n}^{3}} \right)
\]

From eqn. 8, the maximum output voltage \( V_{max} \) occurs at \( t = 0 \) and is equal to
\[
V_{max} = \frac{V_{DD} \lambda_{n} \lambda_{n} \lambda_{n} \lambda_{n}}{C_{on} + C_{on} + C_{on} + C_{on} + C_{on}}
\]

It can be seen from eqn. 10 that the voltage \( V_{max} \) is larger than \( V_{on} \) and the net voltage overshoot is just equal to the feedthrough voltage in the capacitance network of \( C_{2} \) in series with the output capacitances \( C_{on} \) and \( C_{on} \). This is the cause of the initial delay.

At \( t = 0 \), the voltage \( V_{on} \) and \( V_{on} \), in eqn. 7 can be written as
\[
V_{on} = V_{on} - V_{max} - V_{DD}
\]

Substituting eqns. 11 and 12 into eqn. 7 and then substituting eqn. 7 into eqn. 10, one can solve \( V_{on} \).

Define the initial delay \( t_{df} \), as the time interval from \( t = 0 \) to \( t_{df} \) at which \( V_{on} \) is lowered back to \( V_{on} \). \( t_{df} \) can be obtained from eqn. 8 as
\[
t_{df} = C_{on} V_{DD} \left( \frac{2\lambda_{n} \lambda_{n}}{V_{on} \lambda_{n} \lambda_{n} \lambda_{n} \lambda_{n}} \right)
\]

where \( V_{on} \) is given in eqn. 10. It is seen that the initial delay strongly depends on the device parameters as well as the power supply voltage. It should be noted that eqn. 13 is valid only for a finite \( \lambda_{n} \) which corresponds to the case of short-channel devices.

2.1.2 Initial rise delay: In the case of initial rise delay, the input is a falling step input. The input- and output-voltage waveform are also shown in Fig. 2a by dashed lines. According to the modelling method in Section 2.1.1, the voltage magnitude \( V_{max} \) at the minimum point can be similarly derived and expressed as
\[
V_{max} = \frac{C_{on} \left( V_{DD} \right)(C_{2} + C_{3} + C_{on} + C_{on}) \lambda_{n} \lambda_{n} \lambda_{n} \lambda_{n}}{C_{on} + C_{on} + C_{on} + C_{on} + C_{on}}
\]

The initial rise delay \( t_{rr} \), defined as the time interval from \( t = 0 \) to \( t_{rr} \), at which \( V_{on} \) is raised back to 0 V, can be obtained
\[
t_{rr} = C_{on} V_{DD} \left( \frac{2\lambda_{n} \lambda_{n}}{V_{on} \lambda_{n} \lambda_{n} \lambda_{n} \lambda_{n}} \right)
\]

where \( V_{on} \) is given in eqn. 10.

2.2 Ramp input

2.2.1 Initial fall delay: Fig. 3 shows the rising (falling) ramp input \( V(t) \) and the corresponding output waveform \( V(t) \) in solid (dashed) lines at the input and output nodes of the first stage in Fig. 1. In the case of rising ramp input, the displacement current flowing from the input node to output node is initially larger than the drain current \( I_{D} \) which has the output node as its source node. Therefore, a net charging current is presented at the output node so that \( V(t) \) increases above \( V_{on} \). As \( V(t) \) increases to turn on the NMOS \( M_{1} \), in the saturation region, the increase of total drain current \( I_{D}(t) \)
$I_d(t)$ flowing out of the output node tends to decrease the net current charging to the output node. This reduces the rate of increase of $V_o(t)$. When $V_o(t)$ reaches the maximum value $V_{omax}$ at $t = t_{max}$, the net current charging to the output node is reduced to zero and $V_o(t)$ begins to fall to ground after $t_{max}$.

Referring to the circuit in Fig. 2b, we have

$$C_z \frac{d(V_o(t) - V_b)}{dt} = I_d(t) - I_d(t) + (C_z + C_{sd} + C_{sh}) \frac{dV_o}{dt}$$

(16)

where $I_d(t)$ is given in eqn. 1 and $I_d(t)$ is the drain linear current of $M_{sd}$ and can be expressed as

$$I_d(t) = -\beta_n [(V_{ss} - V_{t} - \eta_s V_{ss}^2) V_{ds} - 2 \eta_{ds}^3]$$

$$\times [(2 \phi_{t} + V_{ds} - V_{th})^2 - (2 \phi_{t} + V_{ds})^2]$$

(17)

Note that the capacitance $C_z$ is equal to $C_{sd} + C_{sh}$ and the capacitance $C_{sh}(C_{sd})$ is evaluated at $V_{sd}(V_{sh}) = 0$.

At $t = t_{max}$, that $V_{omax}$ is equal to zero implies

$$\frac{dV_o}{dt} = 0$$

(18)

With the ramp rate $K$, the input voltage is written as

$$V_i(t) = Kt$$

(19)

Substituting eqns. 1, 17, 18 and 19 into eqn. 16, we have

$$C_z \frac{d(V_o(t) - V_b)}{dt} = I_d(t) - I_d(t) + (C_z + C_{sd} + C_{sh}) \frac{dV_o}{dt}$$

(16)

where

$$V_{ds} = (\eta_s - 1) V_{dd} - V_{t} - 2 \eta_{ds}^3 \cdot (2 \phi_{t})^{1/2}$$

(20a)

and

$$V_{sh} = V_{dd} - \theta_s V_{dd} + 2 \eta_{ds}^3 \cdot (2 \phi_{t})^{1/2} - (2 \phi_{t} + V_{dd} K_{sh})^{1/2}$$

(20b)

In eqn. 20, $V_{omax}$ and $t_{max}$ are functions of $V_{omax}$ and $t_{max}$ and can be calculated from eqns. 2 and 3, respectively.

Since the currents must be continuous, eqn. 16 can be differentiated as

$$C_z \frac{d^2(V_o(t) - V_b)}{dt^2} = \frac{dI_d(t)}{dt} - \frac{dI_d(t)}{dt} + (C_z + C_{sd} + C_{sh}) \frac{dV_o}{dt^2}$$

$$+ (C_z + C_{sd} + C_{sh}) \frac{d^2V_o}{dt^2}$$

(22)

From eqns. 1, 2, 7, 18 and 19, eqn. 22 can be further expressed at $t = t_{max}$ as

$$\beta_n [(K_{max} - \eta_s V_{omax}^2) V_{omax}$$

$$+ (K_{max} - V_{t} - \eta_s V_{omax}^2) V_{omax}$$

$$- \gamma_{sd} (2 \phi_{t} + V_{omax})^{1/2} V_{omax}]$$

$$+ \beta_p K (V_{dd} - V_{omax} + (C_z + C_{sd} + C_{sh}) V_o^*)$$

$$= 0$$

(23)

where

$$V_{omax} = K (V_{omax} - \eta_s V_{omax}^2) (V_{omax} - I_{max})$$

$$= \gamma_{sd} (2 \phi_{t} + V_{omax})^{1/2} V_{omax}$$

(24a)

and

$$V_o^* = \frac{d^2V_o}{dt^2}$$

(24b)

The detailed derivations of eqn. 24b are shown in the Appendix. From eqns. 20 and 23, $V_{omax}$ and $t_{max}$ can be obtained. After solving $V_{omax}$ and $t_{max}$, the initial fall delay $t_{df}$ can be calculated as follows: since the output voltage $V_o(t)$ is equal to $V_{dd}$ at $t = t_{df}$, the drain linear current $I_d(t)$ is equal to zero and eqn. 16 can be simply written as

$$K C_z = \beta_p V_{omax} (K_{df} - \eta_s V_{omax}^2)$$

$$\times (1 + \lambda_s V_{omax}) (C_z + C_{sd} + C_{sh}) V_o^*$$

(25)

where

$$V_{omax} = V_{t} + 2 \eta_{ds}^3 (3 V_{omax})$$

$$\times [(2 \phi_{t} + V_{omax})^{1/2} - (2 \phi_{t})^{1/2}]$$

(26a)

and

$$V_o^* = \frac{d^2V_o}{dt^2}$$

(26b)

$t_{df}$ can then be obtained from eqn. 25 with $V_{omax}$ and $t_{max}$ known from eqns. 20 and 23. Note that, in the above calculation, the parameters $\mu_n$ and $\lambda_s$ are evaluated at $V_{omax} - V_{dd}$.

Based upon eqns. 20, 23 and 25, the dependence of the initial fall delay on the input ramp rate, loading capacitance and device parameters can be calculated.

2.2.2 Initial rise delay: In the case of initial rise delay, the analysis proceeds in a manner similar to that in Section 2.2.1. The input voltage is a falling-ramp waveform

$$V_i(t) = V_{dd} - Kt$$

(27)

$t_{df}$ can be obtained after solving $V_{omax}$ and $t_{max}$ similar to the case of $t_{df}$. 
3 Comparison to SPICE simulation

3.1 Step input

Fig. 4 shows the calculated initial fall delay \( t_f \) together with SPICE simulation results. The effective channel length for both NMOS and PMOS varies from 3 \( \mu \)m to 1 \( \mu \)m. For simplicity, the device parameters for \( L_{eff} = 2 \mu m \) are assumed to be the same as those for \( L_{eff} = 3 \mu m \). The ratio of the aspect ratio (width/length) of PMOS to that of NMOS is equal to unity or the ratio of their carrier mobilities. To calculate the source/drain areas, the dimension \( W_n \) of the source/drain regions in the channel-length direction is assumed to be 8.5 \( \mu m \) for both PMOS and NMOS. \( W_n \) is proportionally scaled for devices with smaller effective channel lengths. In all these cases, a general agreement between model calculations and SPICE simulations is obtained.

![Graph showing initial fall delay for various channel lengths under step input](image)

Initial fall delay for various channel lengths under step input

- Simulation
- Calculation
- \( W_n = W_p = 32 \mu m \)
- \( W_n = (W_p/3) \mu m \)
- \( L_{eff} \) varies from 1 \( \mu m \) to 3 \( \mu m \)

When the device dimensions are scaled down, both \( C_x \) and \( C_z \) of NMOS and PMOS are decreased, while \( \lambda_0 \) is increased. The resultant \( t_f \) is decreased as predicted by eqn. 13 and is shown in Fig. 4. In general, a large \( t_f \) leads to large gate delay in CMOS logic gates.

Fig. 5 shows both calculated and SPICE-simulated initial rise delay \( t_r \). A satisfactory agreement is achieved. As shown in Fig. 5, \( t_r \) is larger for \( L_{eff} = 2 \mu m \) than for \( L_{eff} = 3 \mu m \). This is a special case due to the inherent transconductance degradation of PMOS in the adopted 2 \( \mu m \) CMOS process. Generally, \( t_r \) decreases in the channel length, as one compares the case of \( L_{eff} = 2 \mu m \) to that of \( L_{eff} = 1 \mu m \).

![Graph showing initial rise delay for various channel lengths under step input](image)

Initial rise delay for various channel lengths under step input

- Simulation
- Calculation
- \( W_n = W_p = 32 \mu m \)
- \( W_n = (W_p/3) \mu m \)
- \( L_{eff} \) varies from 1 \( \mu m \) to 3 \( \mu m \)

It can be seen from Figs. 4 and 5 that the initial delays vary with the channel-width ratio \( W_n/W_p \). Fig. 6 shows the variations of the initial fall delay, the initial rise delay and total initial delay \( (t_f + t_r) \) with channel-width ratio \( W_n/W_p \) for \( W_p = 2 \mu m \). It is seen from Fig. 6 that the minimum initial delay of a CMOS inverter can be achieved with a suitable value of \( W_n/W_p \). According to
eqn. 13 and 15, if the variations of $V_{\text{init}}$ and $V_{\text{max}}$ are negligible, the optimal width ratio $W_p/W_n$, which results in the minimum total initial delay, can be approximately obtained as

$$
(W_p/W_n)_{\text{opt}} = \frac{\mu_n T_{\text{delay}} I_p V_{\text{in}}} {\mu_n T_{\text{delay}} I_p V_{\text{in}}} \ln \left[ 1 + \frac{V_{\text{init}}}{V_{\text{delay}} + 1/\mu_n} \right]
$$

(28)

As can be seen from eqn. 28, the optimal ratio $W_p/W_n$ is proportional to the square root of the mobility ratio $\mu_n/\mu_n$. This is consistent with the results of previous work [3, 7].

### 3.2 Ramp input

Fig. 7a shows the calculated and SPICE-simulated dependence of $t_{\text{df}}$ and $t_{\text{dr}}$ on $W_p/W_n$ with the ramp rate $K$ as a parameter. Good agreement is obtained. As can be seen from Fig. 7a, $t_{\text{df}}$ increases, whereas $t_{\text{dr}}$ decreases, with the increase of $W_p/W_n$. The negative dependence of $t_{\text{dr}}$ on $W_p/W_n$ results from the increase of $W_p$ increasing the drain current of $M_1$, charging to the output node. Therefore, for a fixed value of $K$, there exists an optimal $W_p/W_n$ such that the total initial delay $(t_{\text{df}} + t_{\text{dr}})$ achieves a minimum value, similar to the case of step input. The resultant $(t_{\text{df}} + t_{\text{dr}})$ is shown in Fig. 7b for different values of $K$. It is seen from Fig. 7b that the optimal $W_p/W_n$, resulting in the minimum $(t_{\text{df}} + t_{\text{dr}})$, is equal to 1.75 for all $K$, which is close to the value 1.78 calculated from eqn. 28. This reveals that the optimal channel-width ratio is independent of input excitations and is determined only by the device parameters as approximately expressed in eqn. 28.

![Graph showing variations of initial fall and initial rise delays](image)

**Fig. 7A** Variations of initial fall and initial rise delays

$K = 0.25 \times 10^9 \text{ V/s}$

$K = 0.5 \times 10^9 \text{ V/s}$

$K = 1 \times 10^9 \text{ V/s}$

$K = 2 \times 10^9 \text{ V/s}$

Initial delay time, ns

Channelwidth ratio $W_p/W_n$

---

![Graph showing dependence of initial delays on loading capacitance for different ramp rate under ramp input](image)

**Fig. 8** Dependence of initial delays on loading capacitance for different ramp rate under ramp input

Simulation $\bigtriangleup$, $t_{\text{df}}$

Calculation $\triangle$, $t_{\text{dr}}$

L, = L, = 2 $\mu$m

$W_p = 2 \mu$m

$W_n = 4 \mu$m

---

The effect of loading capacitance $C_L$ on the initial delay is shown in Fig. 8, where SPICE-simulation results are also given for comparison. As can be seen from Fig. 8,
both $t_{lf}$ and $t_{lr}$ increase with the increase of loading capacitance. However, the rate of increase of $t_{lf}$ or $t_{lr}$ is not linearly proportional to the loading capacitance. As the loading capacitance increases, the net current charging to the output node is reduced, due to the increased transient current across $C_L$, so that the output voltage overshoot $V_{max}$ is decreased. However, the decrease of $V_{max}$ in turn decreases the device current $I_d(t)$ flowing out of the output node. Thus the resultant effects lead to a small increase of the initial delays when increasing the loading capacitance.

It can be seen from Figs. 7 and 8 that the dominant factor in affecting the initial delays is the input ramp rate $K$. Fig. 9 shows the calculated and simulated variations of $t_{lf}$ and $t_{lr}$ with the value $V_{dd}/K$ for different values of $K$. Also shown in Fig. 9 are the corresponding values of input voltage $V_g(V_i)$ calculated from eqn. 19 (eqn. 27) at $t = t_{lf}(t_{lr})$. The initial delays increase drastically and tend to be linearly proportional to $V_{dd}/K$ as $K$ decreases. However, the linear dependence of $t_{lf}$ and $t_{lr}$ on $V_{dd}/K$ occurs only for small values of $K$ (slowly ramped input), as shown in Fig. 9. When the ramp rate becomes large, the feedthrough currents from input node to the output node are so large that more current $I_d(t)$ or $I_f(t)$ is needed to draw the output node to its normal state. Thus, the input must take more time to increase to a sufficient value above the threshold voltage of $M_{np}$ or $M_{ip}$ to increase $I_d(t)$ or $I_f(t)$. The initial delays increase drastically and tend to be linearly proportional to $V_{dd}/K$ as $K$ decreases. However, the linear dependence of $t_{lf}$ and $t_{lr}$ on $V_{dd}/K$ occurs only for small values of $K$ (slowly ramped input), as shown in Fig. 9. When the ramp rate becomes large, the feedthrough currents from input node to the output node are so large that more current $I_d(t)$ or $I_f(t)$ is needed to draw the output node to its normal state. Thus, the input must take more time to increase to a sufficient value above the threshold voltage of $M_{np}$ or $M_{ip}$ to increase $I_d(t)$ or $I_f(t)$.

With the specified initial delays, the sum of the initial fall delay and the initial rise delay, defined as the gate initial delay $t_{delay}$, is just equal to the propagation delay $t_p$, as verified in Fig. 11, where the propagation delay against the gate initial delay is plotted from SPICE results for several chains of identical CMOS inverters. It is not surprising, as shown in Fig. 11, that the propagation delay is equal to the gate initial delay. When $V(t)$ in Fig. 10 increases from 0 V, an initial fall delay is taken before the output voltage $V_{out}(t)$ decreases below $V_{dd}$. Meanwhile, as $V_{out}(t)$ decreases from $V_{dd}$ a second time, the initial rise delay is spent before $V_{out}(t)$ increases above 0 V. Since $V(t)$ and $V_{out}(t)$ are duplicate waveforms with equal rise times, the propagation delay from $V(t)$ to $V_{out}(t)$ therefore results from the sum of the initial fall and the initial rise delays. The smaller the gate initial delay, the smaller the propagation delay will be.

Since the model developed in Section 2.2 considers the small-geometry effects in short-channel devices, it is used to calculate the delay performance of a chain of scaled CMOS inverters. In the calculations, a set of $3 \mu m$ device
parameters in Table 2, for both NMOS and PMOS, are scaled simultaneously in accordance with the three scaling laws: the constant-electric-field (CE), the constant-voltage (CV) and the quasiconstant-voltage (QCV) laws [13, 14] as shown in Table 3, where s is the scaling factor. The zero-biased threshold voltages for both NMOS and PMOS are scaled in the same way as that for the power supply voltage.

![Table 2: Device parameters to be scaled according to scaling laws in Table 3](Figure)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>NMOS</th>
<th>PMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$, μm</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$W$, μm</td>
<td>3.0</td>
<td>6.0</td>
</tr>
<tr>
<td>$V_{ds}$, V</td>
<td>0.725</td>
<td>0.725</td>
</tr>
<tr>
<td>$T_{on}$, m</td>
<td>7.01E-8</td>
<td>6.48E-8</td>
</tr>
<tr>
<td>$R_B$, Ω-1</td>
<td>0.775</td>
<td>0.85</td>
</tr>
<tr>
<td>$N_{a,s}$, cm$^{-3}$</td>
<td>2.0E16</td>
<td>1.75E15</td>
</tr>
<tr>
<td>$V_{ox}$, m/s</td>
<td>8.7E4</td>
<td>6.4E4</td>
</tr>
<tr>
<td>$X_{g}$, μm</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$\mu_{a,s}$, m²/V.s</td>
<td>820E-4</td>
<td>250E-4</td>
</tr>
<tr>
<td>$n_{s}$</td>
<td>0.113</td>
<td>0.295</td>
</tr>
<tr>
<td>$C_{ox}$, f/m</td>
<td>6.377E6</td>
<td>6.224E6</td>
</tr>
<tr>
<td>$r_{on}$, V/m</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>$f_{on}$</td>
<td>0.543</td>
<td>0.515</td>
</tr>
<tr>
<td>$M_{le}$</td>
<td>3.4</td>
<td>0.341</td>
</tr>
<tr>
<td>$C_{ox}$, f/m²</td>
<td>3.8E-4</td>
<td>1.2E-4</td>
</tr>
<tr>
<td>$C_{ox}$, f/m</td>
<td>3.6E-10</td>
<td>3.0E-10</td>
</tr>
<tr>
<td>$C_{ox}$, f/m²</td>
<td>1.23E-10</td>
<td>1.33E-10</td>
</tr>
<tr>
<td>$\phi_{ox}$, μm</td>
<td>8.5</td>
<td>8.5</td>
</tr>
</tbody>
</table>

$V_{dd}$ = 5 V

![Table 3: Three scaling laws for scaling of device parameters](Figure)

<table>
<thead>
<tr>
<th>Scaling law</th>
<th>Constant electric field</th>
<th>Constant voltage</th>
<th>Quasi-constant voltage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
</tr>
<tr>
<td>Voltage</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
</tr>
<tr>
<td>Oxide</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
</tr>
<tr>
<td>Dopant</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
<td>s^{-1}</td>
</tr>
</tbody>
</table>

Two cases, the constant ramp rate and the scaled ramp rate, are considered. In the case of constant ramp rate, the initial delay of the first stage in Fig. 1 is calculated under a fixed ramp-input excitation. In the case of scaled ramp rate, the input waveform in the characteristic waveforms case is simulated by a ramp waveform, with an appropriate ramp rate [7], to calculate the gate initial delay or propagation delay. In this case, the ramp rate of the input waveform of the present stage is also affected by the parameters of the previous stage. According to the expression of eqn. 9a, the ramp rate of the falling waveform can be approximately expressed as

$$K = (\beta_{i,c} V_{dd}^2 / A_s V_{dd}) (C_2 + C_3 + C_{dash} + C_{shut})$$  \(29a\)

whereas that for the rising waveform is

$$K = (\beta_{i,c} V_{dd}^2 / A_s V_{dd}) (C_2 + C_3 + C_{dash} + C_{shut})$$  \(29b\)

For the adopted parameters in Table 2, the calculated $t_{on}$ (Fig. 12a) by using eqn. 29a (eqn. 29b) is 0.838 ns (1.1345 ns), which agrees with the value 0.8027 ns (1.3187 ns) obtained from the characteristic waveform of SPICE simulations. This shows that the initial delays of characteristic waveforms can be calculated by simulating the input waveform with an appropriate ramp waveform.

Applying the model in Section 2.2, Figs. 12a and 12b show the calculated dependence of the initial delay ($t_{on}$) and gate initial delay ($t_{g}$) on the scaling factor $s$ for constant ramp rate and scaled ramp rate, respectively. As can be seen from Fig. 12, the initial delay decreases with an increasing scaling factor $s$ for all three scaling laws. Moreover, the linear dependence of the delays ($t_{on}$ + $t_{g}$) and $t_{on}$ occurs for the CE law, which agrees with the results predicted by the conventional first-order analysis.

When the input is driven by a waveform with a constant ramp rate, the scaling of devices with the CE law leads to the greatest improvement in the delay of the first stage than that with the other two laws, as can be seen.
from Fig. 12a. Furthermore, because the large driving capability of CMOS devices is offset by the effects of a large voltage swing in the CV law, the CV law leads to the least delay improvement for constant ramp rate.

To investigate the initial delay, a ramp input with rising or falling ramp rate equal to $V_{DD}/10$, $V_{DD}/20$ ns.

In the case of the characteristic waveform, however, the ramp rate of the input waveform is also affected, through eqns. 29a and 29b, by the scaling laws applied to the previous stage. Thus, in contrast to those in the constant ramp rate case, the reduction of gate initial delay for the QCV law is larger than that for the CE law but smaller than that for the CV law, as shown in Fig. 12b. This reveals that as one tries to improve the speed of internal circuits by the CV law, one will obtain less delay improvement in the input stage by the CV law than that by the other two laws. Therefore, the most effective manner in which to apply the scaling laws for delay improvement is to apply the CV law to the internal circuits, whereas the CE law is applied to the input stage of internal circuits.

5 Experiment

Three chains of CMOS inverters were fabricated with 3 μm p-well technology. The channel lengths of both NMOS and PMOS are 3 μm, 5 μm and 8 μm for the three chains. The channel width of the NMOS is 4 μm, 10 μm and 16 μm and that of the corresponding PMOS is 8 μm, 20 μm or 16 μm for the three chains. Each output node of the inverter chains is connected, through a two-stage buffer, to an output pad. The dimension of the first stage of the output buffer is the same as that of the corresponding inverter chain. With a 5 V power supply, Fig. 13 shows the measured waveform at the output node of the first stage. It is clearly seen that the voltage overshoot and undershoot indeed occur during the transitions of input waveform.

found that both the initial rise and initial fall delays in a CMOS inverter under step- and ramp-input excitations are characterised in detail. It is modelled for inverters with different dimensions. Optimal device dimension that results in minimum initial delay is obtained. The results are compared to those from SPICE simulations and good agreement between theoretical and experiment.

Satisfactory agreement is obtained between the theory and experiment.

6 Conclusion

The initial delays in a CMOS inverter under step- and ramp-input excitations are characterised in detail. It is found that both the initial rise and initial fall delays determine the propagation delay of CMOS logic gates. Thus, the initial delay actually determines the speed of a logic gate. From this point of view, the initial delays are important factors in determining the initial delay.

Based on the model, the delay time for scaled-down CMOS inverters is calculated. As the channel length is scaled, the delay is decreased as expected. Moreover, the calculations show that the constant-voltage-scaling law is the best choice of scaling laws in reducing the propagation delay of characteristic waveforms. However, for the delay response of the input stage, the constant-electric-field law is found to be the most effective law in reducing the initial delays.

Although the analysis concentrates on CMOS inverters only, other logic gates such as NAND, NOR, and transmission gate also have initial delays and similar modelling method can be applied accordingly.

7 References

8. NAGEL, L.W.: 'SPICE 2', a computer program to simulate semiconductor circuits. University of California, Berkeley, CA, USA, 1975

8 Appendix

The Taylor's expansion of $V_d(t)$ around $t = t_{max}$ can be written as

$$V_d(t) = V_{max} + (t - t_{max}) V_d'(t_{max})$$

where $H(t - t_{max})$ are those of higher order terms. The boundary condition of $V_d(t)$ is given by

$$V_d(0) = V_{dd}$$

Substituting eqns. 31 and 18 into eqn. 30, the second derivative of $V_d(t)$ at $t = t_{max}$ can be written as

$$V_d''(t_{max}) = 2[V_{dd} - V_{max} - H(t_{max})] / t_{max}$$

By neglecting the higher order terms in eqn. 32, eqn. 24b can be obtained.