Chapter 3: The CMOS inverter

This chapter is devoted to analyzing the static (DC) and dynamic (transient) behavior of the CMOS inverter. The main purpose of this analysis is to lay a theoretical ground for a dynamic switching model from which the propagation delay between the output and input signals can be calculated. In the next chapter, this model will be extended to any CMOS logic gate, and not be restricted to inverters only. In later sessions, these models will be used for estimating the propagation delay along critical timing paths in an integrated circuit.

A. Static properties

The overall aim of this lecture is to derive a model for the inverter voltage transfer characteristic (VTC), i.e. for the output voltage response to a slowly increasing (quasi-static) input voltage. As a result of the lecture, we will have derived a model based on the square-law MOSFET model for calculating the switching voltage, i.e. the input voltage for which the output voltage of an inverter flips from one state to the other (formal definition $V_{IN}=V_{OUT}$). The model can easily be extended for calculating the switching voltage of any logic gate by replacing the pull-up and pull-down networks by MOSFETs with an effective aspect ratio, W_{eff}/L.

Our approach for keeping track of whether the two MOSFETs of the inverter are OFF or ON, linear or saturated, is to start by overlaying the two diagrams showing the MOSFET regions of operation. The pMOSFET diagram is similar to that of the nMOSFET, but it has its origo in (V_{DD}, V_{DD}) since the source of the pMOSFET is connected to V_{DD} and not to V_{SS} as for the nMOSFET. By doing so, we obtain five regions of inverter operation as shown in Fig. 3.1. In regions A and E, when one of the MOSFETs are OFF, the output node is pulled to the rail by the ON MOSFET. The switching from high to low, or vice versa, occurs in the green region, C, when both MOSFETs are saturated.

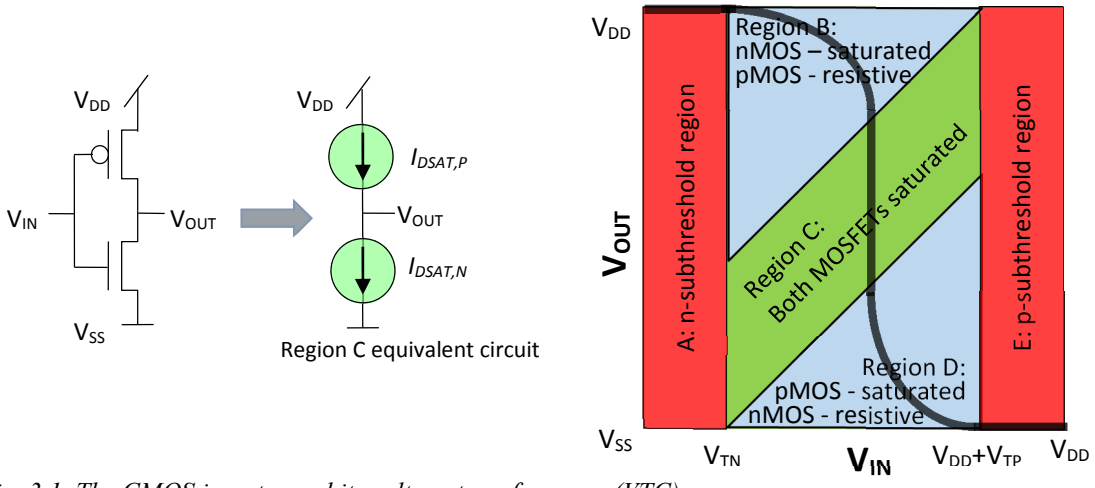


Fig. 3.1. The CMOS inverter and its voltage transfer curve (VTC).

By applying Kirchhoff's current law (KCL) to the output node of the inverter, we can derive the inverter voltage transfer characteristic (VTC) on the form of five different functions, V_{OUT}=f(V_{IN}). The most simple to derive and the most "useful" of these five functions, except for those in regions A and E where,

$$V_{OUT} = \begin{cases} V_{DD} & V_{IN} \le V_{TN} \\ V_{SS} & V_{DD}+V_{TP} \le V_{IN} \le V_{DD} \end{cases}, \qquad (3.1)$$

are those for region C, where both MOSFETs are saturated. From KCL we obtain $I_{DSAT,N}=I_{DSAT,P}$, an expression from which the switching voltage, $V_{SW}=V_{IN}=V_{OUT}$, can be readily derived. The result is

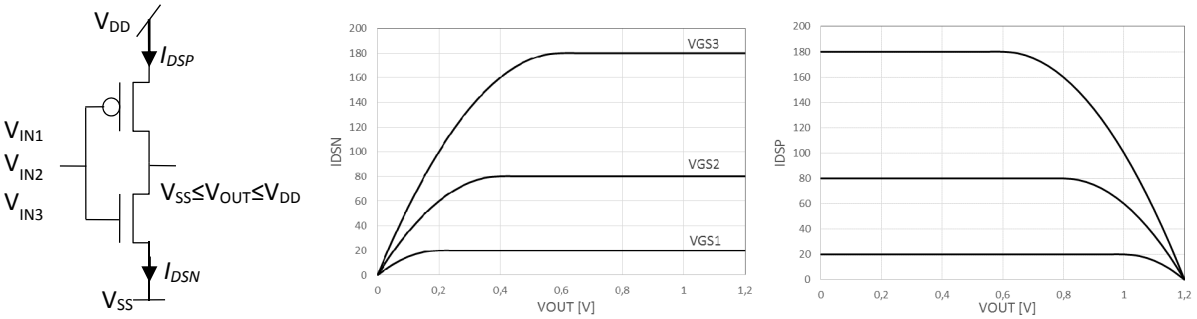
$$V_{sw}=\frac{V_{DD}+V_{TP}+\sqrt{x}\cdot V_{TN}}{1+\sqrt{x}}, \tag{3.2}$$

where $x=k_N/k_P$. The complete VTC is shown in Fig. 3.1. During the lecture and the hands-on laboratory session, we will also define and determine the high and low noise margins, NMH and NML. From the discussions of noise margins that we will have, we can conclude that CMOS is a robust technology.

Exercises

Exercise 3.1: Use the square-law MOSFET model and Kirchhoff´s current law to derive equation (3.2)! Often inverters are designed with x=1, but what is the effect on the VTC if x is increased to, say x=4?

Exercise 3.2: For three different input voltages, the output voltage of an inverter is swept from V_{SS} to V_{DD} while measuring the two MOSFET currents, I_{DSN} and I_{DSP}. The current-voltage characteristics thus obtained are shown below. Match the two MOSFET currents for each of the three inverter input voltages, and find the bias points where the two currents are equal! Mark each of these three bias points with B, C, or D, depending on to which region of operation in the V_{OUT} vs V_{IN} graph that they belong.



Exercise 3.3: In the inverter "regions of operation" diagram in Fig. 3.1, we can add a secondary axis for the plotting the "short-circuit" current, I_{SC}=min(I_{DSP}, I_{DSN}). Plot the "short-circuit" current after having identified the current-limiting MOSFET in the different regions of operation.

Exercise 3.4: *Derive the following VTC equations for the blue regions (regions B and D):

$$V_{out,high}=V_{DD}+V_{in}-b+\sqrt{(V_{in}-b)^2-x(V_{in}-a)^2} \quad \text{where } a=V_{TN} \text{ and } b=V_{DD}+V_{TP}. \tag{3.3}$$

$$V_{out,low}=V_{in}-a-\sqrt{(V_{in}-a)^2-\frac{1}{x}(V_{in}-b)^2}$$

Please, note, that with this nomenclature, the switching voltage can be written on the form

$$V_{sw}=\frac{b+\sqrt{x}\cdot a}{1+\sqrt{x}}, \tag{3.4}$$

Exercise 3.5: To account for voltage fluctuations, i.e. noise, the valid high and low output voltages are usually defined within certain ranges like $0\leq V_{OUT}\leq V_{OL,max}$, and $V_{OH,min}\leq V_{OUT}\leq V_{DD}$. Since CMOS is a robust technology, the input voltage can vary within ranges larger than those defined for valid output voltages without causing invalid output voltages, $0\leq V_{IN}\leq V_{IL,max}$, and $V_{IH,min}\leq V_{IN}\leq V_{DD}$. These regions

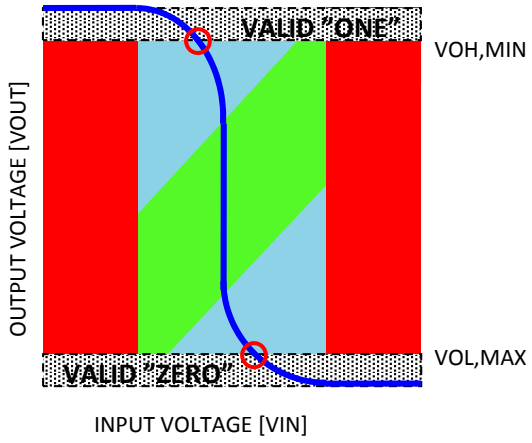
are usually defined from the two points, ($V_{OL,max}$, $V_{IH,min}$) and ($V_{OH,min}$,$V_{IL,max}$), on the VTC where the amplifications is equal to minus one, A_v=-1.

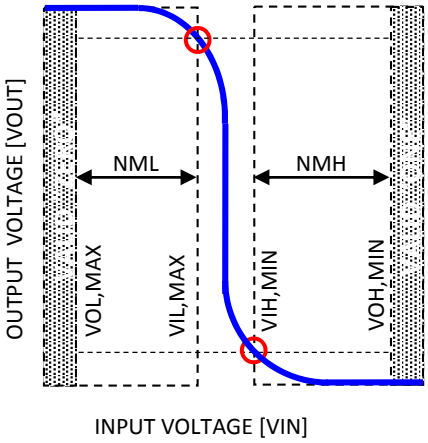
a) Derive expressions for the low and high noise margins, *NML* and *NMH*, as defined in the graph using the following expressions for ($V_{OL,max}$, $V_{IH,min}$) and ($V_{OH,min}$,$V_{IL,max}$)!

$$\left(\begin{array}{l} V_{OH,\min} = V_{DD} - \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \\ V_{IL,\max} = V_{sw} - \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \end{array} \right), \left(\begin{array}{l} V_{OL,\max} = \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \\ V_{IH,\min} = V_{sw} + \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \end{array} \right), \tag{3.5}$$

b) What are the explicit noise margin values in terms of fractions of V_{DD} if V_{TN}=-V_{TP}=V_{DD}/5?

c) *Derive the expressions for ($V_{OL,max}$, $V_{IH,min}$) and ($V_{OH,min}$,$V_{IL,max}$) given above for a CMOS inverter with k_n=k_p!





d) Derive expressions for the low and high noise margins, *NML* and *NMH*, as defined in the graph using the following expressions for ($V_{OL,max}$, $V_{IH,min}$) and ($V_{OH,min}$,$V_{IL,max}$)!

$$\left(\begin{array}{l} V_{OH,\min} = V_{DD} - \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \\ V_{IL,\max} = V_{sw} - \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \end{array} \right), \left(\begin{array}{l} V_{OL,\max} = \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \\ V_{IH,\min} = V_{sw} + \dfrac{V_{DD} + V_{TP} - V_{TN}}{8} \end{array} \right), \tag{3.6}$$

e) What are the explicit noise margin values in terms of fractions of V_{DD} if V_{TN}=-V_{TP}=V_{DD}/5?

f) *Derive the expressions for ($V_{OL,max}$, $V_{IH,min}$) and ($V_{OH,min}$,$V_{IL,max}$) given above for a CMOS inverter with k_n=k_p!

Suggested laboratory exercise: Use the .DC analysis tool of the Spice circuit simulator to derive the VTC for three different values of *x*! Use the slope marker tool to derive the noise margins from the points on the VTC where the gain is equal to -1! Finally, determine the voltage gain when V_{IN}=V_{OUT}!

B. Dynamic properties

The overall aim of this second half of the inverter chapter is to make a simple electrical model of the inverter for rough estimations of its dynamic switching behavior (in contrast to its static behavior

discussed in the previous half). In particular, we are interested in estimating the propagation delay between the input and output signals due to the capacitances of the loading gates. For simplicity, we will only consider inverters where the two MOSFETs are sized for equal driving capability. By doing so, we need not treat the delay of rising outputs differently from the delay of falling outputs. This is simple enough for back-of-the-envelope calculations prior to a more accurate computer-aided timing analysis performed by using state-of-the-art electronic design automation (EDA) tools.

Charging and discharging a load capacitor through a constant-current source, I_{ON}, yield linear relationships in time. Therefore, the step response propagation delay, i.e. the time it takes to reach the 50% level, $V_{DD}/2$, from either rail is given by

$$t_{pd} = \frac{C_L V_{DD} / 2}{I_{ON}} = 0.5 R_{eff} C_L . \tag{3.7}$$

However, in a real circuit the input signal, being the output from a previous gate, is better approximated by a linear ramp, see Fig. 3.2. Experience and many simulations have shown that the propagation delay for this situation is about 40% longer, i.e.

$$t_{pd} = 0.7 R_{eff} C_L . \tag{3.8}$$

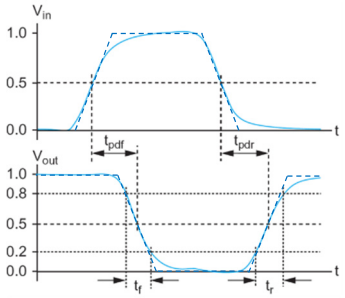


Fig. 3.2. CMOS inverter input signal and output response approximated by ramps.

This delay model yields the same delay as the delay when charging or discharging a capacitor, C_L, through a resistor, R_{eff}. To obtain a simple enough electrical model of the inverter for calculating its propagation delay given a certain load capacitance, C_L, we will simply replace the MOSFET current source by its effective resistance. The step-by-step derivation of the electrical circuit modeling the inverter output driving capability is illustrated in Fig. 3.3.

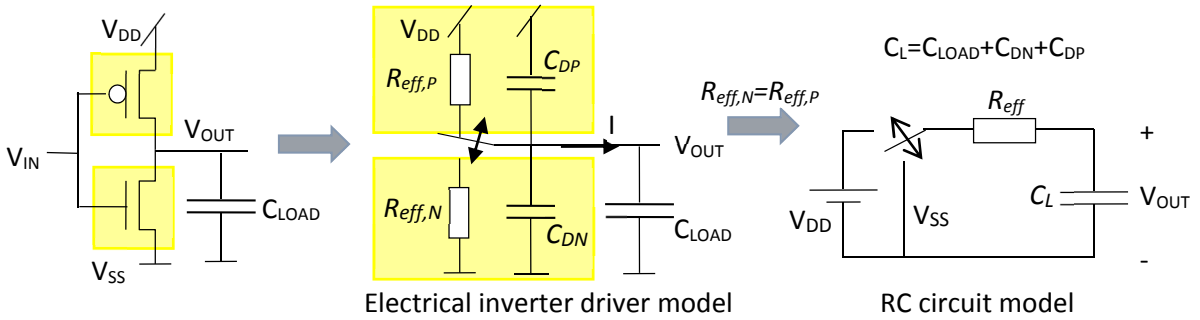


Fig. 3.3. The CMOS inverter and the derivation of its output equivalent RC circuit.

In a previous session, we have already learnt how to calculate the effective resistance of a MOSFET. In the case of an n-channel MOSFET delivering a maximum current I_{ON}=600 A/m at V_{DD}=1.2 V, we obtain

$$R_{eff} = \frac{V_{DD}}{I_{ON}} = \frac{2\ [\mathrm{k\Omega\times\mu m}]}{W_N\,[\mu\mathrm{m}]}, \tag{3.9}$$

where W_N is the channel width of the nMOSFET. In the following, we will always assume the pMOSFET being twice as wide as the nMOSFET), i.e. W_P=2W_N, to compensate for its inherent lower driving capability (due to a factor of two lower hole mobility). Since the C_{LOAD} most often is the input capacitance of another inverter, it might be appropriate to recall the expression for the inverter input capacitance

$$C_G = C_{GN} + C_{GP} = (W_N + W_P) L C_{ox} = 3 W_N L C_{ox} = 3.6 W_N \ \mathrm{fF}/\mu\mathrm{m}, \tag{3.10}$$

for an L=60 nm MOSFET with C_{ox}=20 fF/μm^2. The parasitic output capacitance is similarly assumed to scale with the inverter driving capability if we assume C_D=pC_G.

This discussion leads us to the important conclusion that the $R_{eff}C_G$ product of an inverter is constant, independent of W_N, or equivalently, independent of the inverter driving capability. Hence, the propagation delay of an ideal inverter without parasitics, loaded by an identical inverter is given by

$$\tau = 0.7 R_{eff} C_G = 5 \text{ ps}. \tag{3.11}$$

This is true independent of whether the inverter is of size X2 (with W_N =200 nm) or of size X10 (with W_N =1 μm). This fact will significantly simplify our lives as designers, since this is the only time throughout the course that we have to calculate the *RC* constant. All other delay calculations will be expressed as multiples or fractions of this fundamental *RC* constant. The value of 5 ps given above is for a 65 nm process, a value that in an ideal world scales as L^2/V_{DD} between process nodes, where again L is the channel length, or, equivalently, the minimum feature size.

The process of replacing the MOSFETs by their electrical models to obtain the complete inverter two-port model, and not only a model for its driving capability, is illustrated in Fig. 3.4.

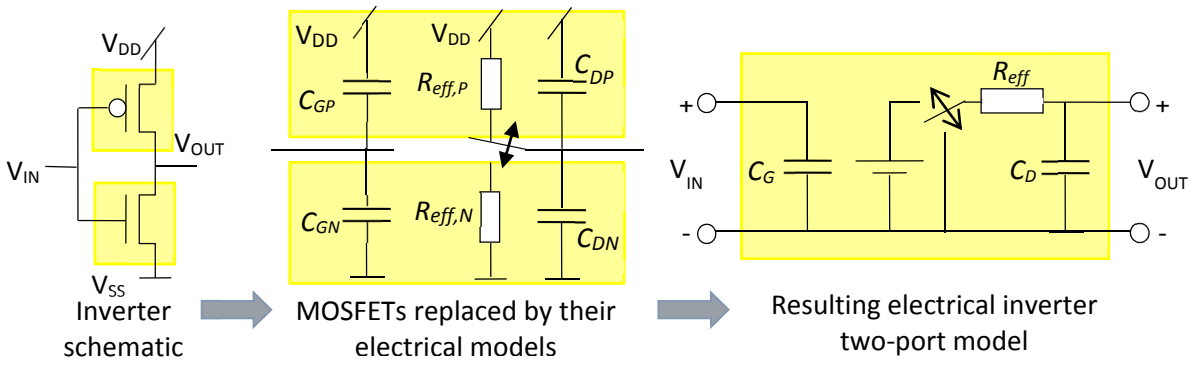


Fig. 3.4. The CMOS inverter and the derivation of its equivalent RC circuit.

Now, let us consider the case where an X2 inverter, for example, is loaded by four identical X2 inverters, or the equivalent case where an X2 inverter is loaded by an X8 inverter. In both cases, the fanout of the inverter is equal to four, since it is loaded by a capacitance four times its input capacitance. This delay is denoted the FO4 delay of the inverter. See Fig. 3.5. In both cases, the FO4 delay can be calculated as follows:

$$\text{FO4 delay} = \tau \times \left(\frac{C_D}{C_G} + \frac{C_L}{C_G} \right) \approx 5 \text{ ps} \times (p + f) = 25 \text{ ps}, \tag{3.12}$$

where the fanout $f=C_L/C_G=4$, and p is the relative delay due to the parasitic capacitance at the gate output, $p=C_D/C_G=1$. As we shall see during the next lecture, the ratio between the loading capacitance and the input capacitance is also called the electrical effort, h. But, for the case of an inverter, its fanout and its electrical effort are the same, $f=h$.

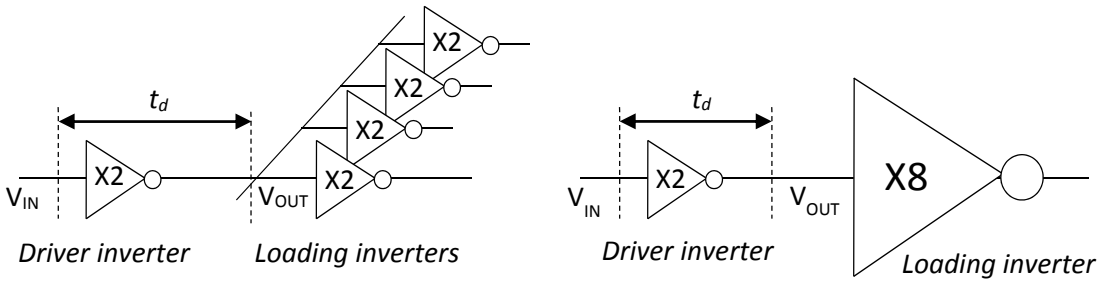


Fig. 3.5. Illustration of the fanout-of-4 (FO4) concept.

Now, what would be the propagation delay if the X2 inverter is loaded by three 2-input NAND gates, or by three 2-input NOR gates, both with a driving capability X2, i.e. the same as that of the inverter? To solve this problem, we have to go back to the schematics of the NAND and NOR gates, and size the gate MOSFETs so that all path resistances are equal to the effective resistance of the inverter. This is a two-stage process. First, as discussed before, the reference p-channel MOSFETs should, by default, be twice as wide as the reference n-channel MOSFETs. Second, when two MOSFETs appear in series, they must be sized for twice the reference width to reduce the effective path resistance to R_{eff}, the effective resistance of a reference inverter. The results of such sizing processes are shown for NAND2 and NOR2 gates in Fig. 3.6.

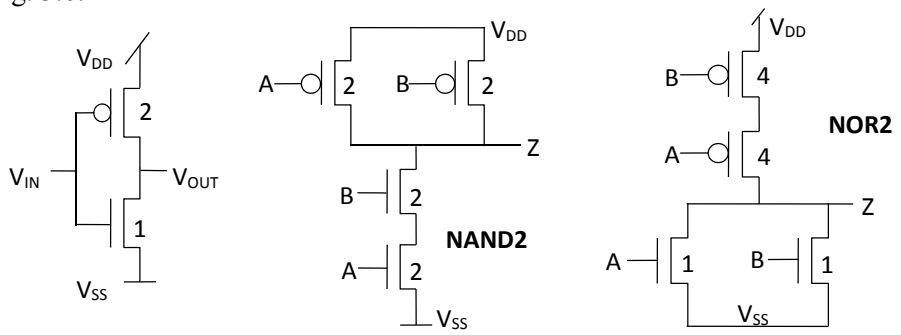


Fig. 3.6. Sizing MOSFET widths for the same effective resistances.

By simply counting the number of unit capacitances connected to the gate input in these MOSFET schematics, we can draw the conclusion that the gate input capacitance of the NAND gate is 4 unis, or 4/3 times that of the inverter, while the gate input capacitance of the NOR gate is five units, or 5/3 times the input capacitance of the inverter. From this discussion, we obtain the following inverter FO3 delays:

$$\text{FO3 delay}=\begin{cases}5\text{ ps}\times\left(1+3\times\dfrac{4}{3}\right)=25\text{ ps} & \text{NAND load}\\ 5\text{ ps}\times\left(1+3\times\dfrac{5}{3}\right)=30\text{ ps} & \text{NOR load}\end{cases}. \tag{3.13}$$

The conclusion of this discussion is that all CMOS logic gates, due to its inherent more complex topology, have an input gate capacitance larger than that of an inverter with the same driving capability. Of course, logic gates like the NAND and NOR gates in this example, can always be resized for the

same input capacitance as for the inverter, but then their driving capabilities will be only ¾ or 3/5 of the inverter driving capability, respectively. This is due to the fact that the RC products are constant also for logic gates. This will be discussed in more detail in the next chapter, where we will derive models for calculating propagation delays when the driver is a logic gate and not an inverter.

Finally, I would like to point out that the ramp response of an inverter is not identical to the exponential curve form of the RC circuit. However, the two curve forms yield approximately the same delay at the 50% level as illustrated in Fig. 3.7.

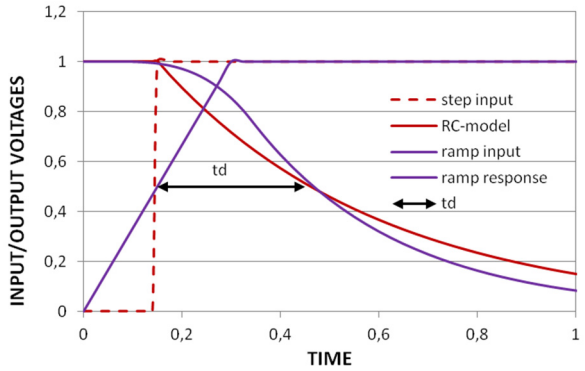


Fig. 3.7. Inverter ramp response as compared to the exponential decay of an RC circuit.

Exercises

Exercise 3.6: Analyze the RC circuit and show that the time needed for the exponential decay of the voltage across the capacitor to 50% of the initial voltage, V_{DD}, is given by $t_d=RC\ln 2$!

Exercise 3.7: What do we mean with an ideal inverter concerning its parasitic output capacitance?

Exercise 3.8: Calculate the propagation delay of an ideal inverter driving an identical inverter! Assume the following MOSFET data: n-channel MOSFETs can sink 500 µA/µm channel width at V_{DD}=1V, and their input capacitances are 1.3 fF/µm. The p-channel MOSFET is made twice as wide as the n-channel device to obtain the same driving capability.

Exercise 3.9: Assume that we, for simplicity, introduce a modified effective resistance $R'=R\ln 2$, how large would this resistance be [in Ω·µm] for the MOSFET in Exercise 3.3? How does the use of R' modify our delay model?

Exercise 3.10: The FO4 delay of the AMS 0.35 µm CMOS process running at 3.3 V is 125 ps. **What** would be the FO4 delay of a 0.13 µm CMOS process running at V_{DD}=1.8V and of a 65 nm CMOS process running at 1.2 V?

Exercise 3.11: Four is sort of a magic number, if the number of loading inverters becomes much larger than four, it is often more efficient to insert an extra inverter with a better driving capability as a buffer between the original inverter and the capacitive load.

a) What driving capability should the inserted buffer inverter have to minimize the delay?
b) For what number of loading inverters does the inserted buffer shorten the propagation delay?
c) How does the parasitic output capacitance influence these critical numbers?

Exercise 3.12: For how big a capacitive load would the insertion of a non-inverting, two-inverter buffer give the shortest propagation delay?

Exercise 3.13: Determine the number of buffer inverters needed to minimize the delay if the load capacitance is 1000 times larger than the inverter input capacitance? What would be the optimum tapering factor?

Exercise 3.14: As a preparation for the next chapter, determine the parasitic output capacitances of the two logic gates in the figure above simply by counting the number of unit drain capacitances connected to the output. How large are these parasitic capacitances with respect to the 3C input capacitance of the reference inverter?

Suggested laboratory exercise: Use the .tran[1] analysis simulation tool of the Spice/Spectre simulator to derive the rise and fall FO4 delays of an inverter that you have designed. Define your input signal so that its rise and fall times are about equal to those expected for the output. Did simulations give the output rise and fall times that you expected? Did you find them equal or different? Why or why not? Compare your results from the FO4 delay simulations with your pre-lab estimations using our simple *RC* model. If you have the time, run the same simulations also for the FF and SS process corners. What are the deviations from the typical case? A final task would be to check how, and if, the FO4 delay varies with the inverter driving capability by changing the channel widths. If the FO4 delay is different for inverters of different driving capabilities, how could that be explained?

[1] tran stands for transient, large-signal analysis suitable for digital circuits