CMOS Logic Design

Dinesh Sharma
EE Department, IIT Bombay

July 30, 2019

Contents

List of Figures

Chapter 1

Static CMOS Logic Design

Static logic circuits are those which can hold their output logic levels for indefinite periods as long as the inputs are unchanged. Circuits which depend on charge storage on capacitors are called dynamic circuits and will be discussed in a later chapter.

1.1 Static CMOS Design style

The most common design style in modern VLSI design is the Static CMOS logic style. In this, each logic stage contains pull up and pull down networks which are controlled by input signals. The pull up network contains p channel transistors, whereas the pull down network is made of n channel transistors. The networks are so designed that the pull up and pull down networks are never 'on' simultaneously. This ensures that there is no static power consumption.

1.2 CMOS Inverter

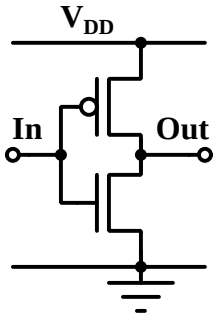


Figure 1.1: The basic CMOS inverter

The simplest of such logic structures is the CMOS inverter. In fact, for any CMOS logic design, the CMOS inverter is the basic gate which is first analyzed and designed in detail. Thumb rules are then used to convert this design to other more complex logic. The basic CMOS inverter is shown in fig. 1.1. We shall develop the characteristics of CMOS logic through the inverter structure, and later discuss ways of converting this basic structure more complex logic gates.

1.2.1 Static Characteristics

The range of input voltages can be divided into several regions. Let us compute the output voltage for a series of input voltages from 0V to V_{DD}.

nMOS 'off', pMOS 'on'

For $0 < V_i < V_{Tn}$ the n channel transistor is 'off', the p channel transistor is 'on' and the output voltage = V_{DD}. This is the normal digital operation range with input = '0' and output = '1'.

nMOS saturated, pMOS linear

In this regime, both transistors are 'on'. The input voltage V_i is $> V_{Tn}$, but is small enough so that the n channel transistor is in saturation, and the p channel transistor is in the linear regime. In static condition, the output voltage will adjust itself such that the currents through the n and p channel transistors are equal. The absolute value of gate-source voltage on the p channel transistor is V_{DD}- V_i, and therefore the "over voltage" on its gate is V_{DD}- V_i- V_{Tp}. The drain source voltage of the pMOS has an absolute value V_{DD}-V_o. Therefore,

$$I_d = K_p \left[(V_{DD} - V_i - V_{Tp})(V_{DD} - V_o) - \frac{1}{2}(V_{DD} - V_o)^2 \right] = \frac{K_n}{2}(V_i - V_{Tn})^2 \qquad (1.1)$$

Where symbols have their usual meanings.

We define $\beta \equiv \mathrm{K_n/K_p}$. We make the substitution $V_{dp} \equiv V_{DD} - V_o$, where $V_{dp}$is the absolute value of the drain-source voltage for the p channel transistor. Then,

$$(V_{DD} - V_i - V_{Tp})V_{dp} - \frac{1}{2}V_{dp}^2 = \frac{\beta}{2}(V_i - V_{Tn})^2 \qquad (1.2)$$

Which gives the quadratic

$$\frac{1}{2}V_{dp}^2 - V_{dp}(V_{DD} - V_i - V_{Tp}) + \frac{\beta}{2}(V_i - V_{Tn})^2 = 0 \qquad (1.3)$$

Solutions to the quadratic are:

$$V_{dp} = (V_{DD} - V_i - V_{Tp}) \pm \sqrt{(V_{DD} - V_i - V_{Tp})^2 - \beta(V_i - V_{Tn})^2} \qquad (1.4)$$

These equations are valid only when the pMOS is in its linear regime. This requires that

$$V_{dp} \equiv V_{DD} - V_o \leq V_{DD} - V_i - V_{Tp}$$

Therefore, we must choose the negative sign. Thus

$$V_{DD} - V_o = (V_{DD} - V_i - V_{Tp}) - \sqrt{V_{DD} - V_i - V_{Tp})^2 - \beta(V_i - V_{Tn})^2} \tag{1.5}$$

Therefore,

$$V_o = V_i + V_{Tp} + \sqrt{(V_{DD} - V_i - V_{Tp})^2 - \beta(V_i - V_{Tn})^2} \tag{1.6}$$

Since V_o must be $\geq V_i + V_{Tp}$, the limit of applicability of the above result is given by

$$(V_{DD} - V_i - V_{Tp})^2 = \beta(V_i - V_{Tn})^2$$

That is, the solution for V_o is valid for

$$V_i \leq \frac{V_{DD} + \sqrt{\beta}V_{Tn} - V_{Tp}}{1 + \sqrt{\beta}} \tag{1.7}$$

In the case where we size the n and p channel transistors such that

$$K_n = K_p; \text{ so } \beta = 1$$

we have

$$V_o = (V_i + V_{Tp}) + \sqrt{(V_{DD} - V_{Tn} - V_{Tp})(V_{DD} - 2V_i + V_{Tn} - V_{Tp})} \tag{1.8}$$

with

$$V_i \leq \frac{V_{DD} + V_{Tn} - V_{Tp}}{2}$$

nMOS saturated, pMOS saturated

At the limit of applicability of eq. 1.7, when the input voltage is exactly at

$$V_i = \frac{V_{DD} + \sqrt{\beta}V_{Tn} - V_{Tp}}{1 + \sqrt{\beta}} \tag{1.9}$$

both transistors are saturated. Since the currents of both transistors are independent of their drain voltages in this condition, we do not get a unique solution for V_o by equating drain currents. The currents will be equal for all values of V_o in the range

$$V_i - V_{Tn} \leq V_o \leq V_i + V_{Tp}$$

Thus the transfer curve of an inverter shows a drop of $V_{Tn}+ V_{Tp}$ at a voltage near $V_{DD}/2$. This is actually an artifact of the simple transistor model chosen for this analysis, which assumes perfect saturation of drain current. In a real case, the drain current does depend on the drain voltage (albeit weakly) in the saturation region. If the model incorporates an Early Voltage like effect, the drop near the middle of the characteristic is more gradual.

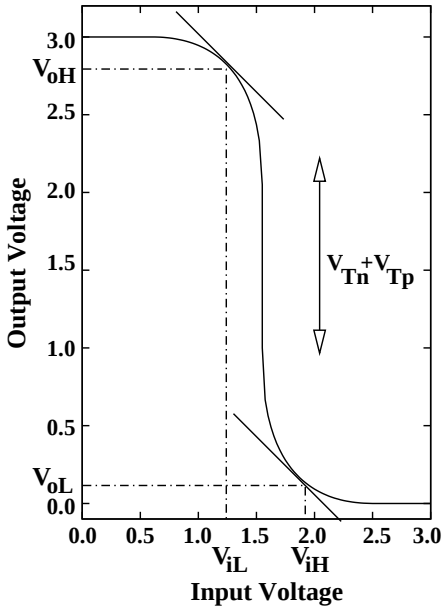


Figure 1.2: Transfer Curve of a CMOS inverter

nMOS linear, pMOS saturated

At the gate voltage given by eq. 1.9, both transistors are saturated. As we increase V_i beyond this value, such that

$$\frac{V_{DD} + \sqrt{\beta} V_{Tn} - V_{Tp}}{1 + \sqrt{\beta}} < V_i < V_{DD} - V_{Tp}$$

both transistors are still 'on', but nMOS enters the linear regime while pMOS gets saturated. Equating currents in this condition,

$$I_d = \frac{K_p}{2}(V_{DD} - V_i - V_{Tp})^2 = K_n \left[(V_i - V_{Tn}) V_o - \frac{1}{2} V_o^2 \right] \tag{1.10}$$

From this, we get the quadratic equation

$$\frac{1}{2} V_o^2 - (V_i - V_{Tn}) V_o + \frac{(V_{DD} - V_i - V_{Tp})^2}{2\beta} = 0 \tag{1.11}$$

This has solutions

$$V_o = (V_i - V_{Tn}) \pm \sqrt{(V_i - V_{Tn})^2 - \frac{(V_{DD} - V_i - V_{Tp})^2}{\beta}} \tag{1.12}$$

Since the equations are valid only when the n channel transistor is in the linear regime ($V_o < V_i - V_{Tn}$), we choose the negative sign. This gives,

$$V_o = (V_i - V_{Tn}) - \sqrt{(V_i - V_{Tn})^2 - \frac{(V_{DD} - V_i - V_{Tp})^2}{\beta}} \tag{1.13}$$

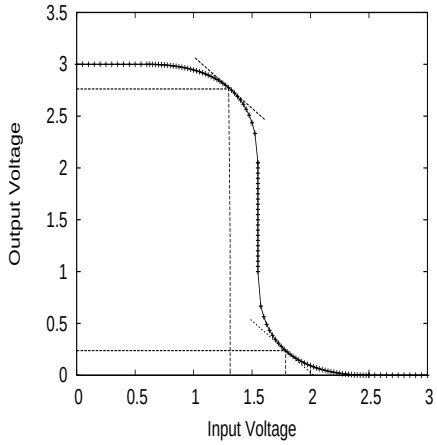
Again, in the special case where $\beta = 1$, we have

$$V_o = (V_i - V_{Tn}) - \sqrt{(V_{DD} - V_{Tn} - V_{Tp})(2V_i - V_{DD} - V_{Tn} + V_{Tp})} \tag{1.14}$$

nMOS 'on', pMOS 'off'

As we increase the input voltage beyond V_{DD}- V_{Tp}, the p channel transistor turns 'off', while the n channel conducts strongly. As a result, the output voltage falls to zero. This is the normal digital operation range with input = '1' and output = '0'.

The figure below shows the transfer curve of an inverter with V_{DD}= 3V, V_{Tn}= 0.6V and V_{Tp}= 0.5V, and $\beta = 1$.



The plot produced by SPICE for this circuit with realistic models is quite similar.

1.2.2 Noise margins

The requirement from a digital circuit is that it should distinguish logic levels, but be insensitive to the exact analog voltage at the input. This implies that the flat portions of the transfer curve (where $\frac{\partial V_o}{\partial V_i}$ is small) are suitable for digital logic. We select two points on the transfer curve where the slope ($\frac{\partial V_o}{\partial V_i}$) is -1.0. The coordinates of these two points define the values of (V_{iL},V_{oH}) and (V_{iH},V_{oL}). Robust digital design requires that the output high level be higher than what is acceptable as a high level at the input ($V_{oH} > V_{iH}$). The difference

between these two levels is the 'high' noise margin. This is the amount of noise that can ride on the worst case 'high' output and still be accepted as a 'high' at the input of the next gate. Similarly, we require $V_{oL} < V_{iL}$. The difference, $V_{iL} - V_{oL}$ is the 'low' noise margin. Obviously, it is of interest to evaluate the values of these noise margins. For the discussion which follows, we shall use the expressions derived earlier for $\beta = 1$ to keep the algebra simple.

Calculation of V_{iL} and V_{oH}

from eq. (1.8)

$$V_o = (V_i + V_{Tp}) + \sqrt{(V_{DD} - V_{Tn} - V_{Tp})(V_{DD} + V_{Tn} - V_{Tp} - 2V_i)}$$

From this, we can evaluate $\frac{\partial V_o}{\partial V_i}$ and set it = -1.

$$\frac{\partial V_o}{\partial V_i} = -1 = 1 - \sqrt{\frac{V_{DD} - V_{Tn} - V_{Tp}}{V_{DD} + V_{Tn} - V_{Tp} - 2V_i}} \tag{1.15}$$

This gives

$$V_{iL} = \frac{3V_{DD} + 5V_{Tn} - 3V_{Tp}}{8} \tag{1.16}$$

Substituting this in eq.(1.8), we get

$$\begin{aligned} V_{oH} &= \frac{3V_{DD} + 5V_{Tn} - 3V_{Tp}}{8} + V_{Tp} + \\ &\quad \sqrt{(V_{DD} - V_{Tn} - V_{Tp})\left(V_{DD} + V_{Tn} - V_{Tp} - \frac{3V_{DD} + 5V_{Tn} - 3V_{Tp}}{4}\right)} \\ &= \frac{3V_{DD} + 5V_{Tn} + 5V_{Tp}}{8} + \sqrt{(V_{DD} - V_{Tn} - V_{Tp})\left(\frac{V_{DD} - V_{Tn} - V_{Tp}}{4}\right)} \\ &= \frac{3V_{DD} + 5V_{Tn} + 5V_{Tp}}{8} + \frac{V_{DD} - V_{Tn} - V_{Tp}}{2} \\ &= \frac{7V_{DD} + V_{Tn} + V_{Tp}}{8} \end{aligned}$$

So

$$V_{oH} = \frac{7V_{DD} + V_{Tn} + V_{Tp}}{8} = V_{DD} - \frac{V_{DD} - V_{Tn} - V_{Tp}}{8} \tag{1.17}$$

Calculation of V_{iH} and V_{oL}

When the input is 'high', we should use eq.(1.14).

$$V_o = (V_i - V_{Tn}) - \sqrt{(V_{DD} - V_{Tn} - V_{Tp})(2V_i - V_{DD} - V_{Tn} + V_{Tp})}$$

Differentiating with respect to V_i gives

$$\frac{\partial V_o}{\partial V_i} = -1 = 1 - \sqrt{\frac{V_{DD} - V_{Tn} - V_{Tp}}{2V_i - V_{DD} - V_{Tn} + V_{Tp}}} \tag{1.18}$$

$$\text{So} \qquad \frac{V_{DD} - V_{Tn} - V_{Tp}}{2V_{in} - V_{DD} - V_{Tn} + V_{Tp}} = 4$$

$$\text{Therefore} \qquad V_{DD} - V_{Tn} - V_{Tp} = 8V_{in} - 4V_{DD} - 4V_{Tn} + 4V_{Tp}$$

From where, we get

$$V_{iH} = \frac{5V_{DD} + 3V_{Tn} - 5V_{Tp}}{8} \tag{1.19}$$

Substituting the value of V_{iH} for V_{in} in eq.1.14), we get

$$\begin{aligned} V_{oL} &= \frac{5V_{DD} + 3V_{Tn} - 5V_{Tp}}{8} - V_{Tn} - \\ &\quad \sqrt{(V_{DD} - V_{Tn} - V_{Tp})\left(\frac{5V_{DD} + 3V_{Tn} - 5V_{Tp}}{4} - V_{DD} - V_{Tn} + V_{Tp}\right)} \\ &= \frac{5V_{DD} - 5V_{Tn} - 5V_{Tp}}{8} - \sqrt{(V_{DD} - V_{Tn} - V_{Tp})\left(\frac{V_{DD} - V_{Tn} - V_{Tp}}{4}\right)} \\ &= \frac{5V_{DD} - 5V_{Tn} - 5V_{Tp}}{8} - \frac{V_{DD} - V_{Tn} - V_{Tp}}{2} = \frac{V_{DD} - V_{Tn} - V_{Tp}}{8} \end{aligned}$$

So

$$V_{oL} = \frac{V_{DD} - V_{Tn} - V_{Tp}}{8} \tag{1.20}$$

Calculation of Noise Margins

The high noise margin is given by

$$V_{oH} - V_{iH} = \frac{V_{DD} - V_{Tn} + 3V_{Tp}}{4} \tag{1.21}$$

Similarly, the Low noise margin is

$$V_{iL} - V_{oL} = \frac{V_{DD} + 3V_{Tn} - V_{Tp}}{4} \tag{1.22}$$

The two noise margins can be made equal by choosing equal values for V_{Tn} and V_{Tp}.

1.2.3 Dynamic Considerations

In this section, we analyze the dynamic behaviour of the inverter. For the calculation of rise and fall times, we shall assume that only one of the two transistors in the inverter is 'on'. (Notice that this is more conservative than the input high and low conditions determined by slope considerations in eq.1.19 and 1.16). We shall continue to use the simple model described at the beginning of this booklet.

Rise time

When the input is low, the n channel transistor is 'off', while the p channel transistor is 'on'. The equivalent circuit in this condition is shown in fig. 1.3. From Kirchoff's current law at

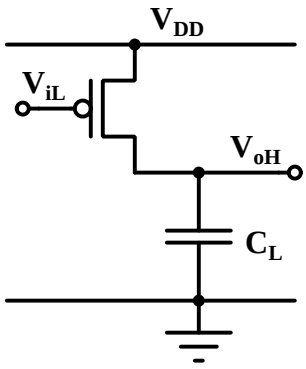


Figure 1.3: CMOS inverter with the nMOS 'off'

the output node,

$$I_{dp} = C\frac{dV_o}{dt}$$

so,

$$\frac{dt}{C} = \frac{dV_o}{I_{dp}}$$

This separates the variables, with the LHS independent of operating voltages and the RHS independent of time. Integrating both sides, we get

$$\frac{\tau_{rise}}{C} = \int_0^{V_{oH}} \frac{dV_o}{I_{dp}}$$

Till the output rises to $V_{iL}+ V_{Tp}$, the p channel transistor is in saturation. Since the current is constant, the integration is trivial. If $V_{oH} > V_{iL} + V_{Tp}$ (which is normally the case), the integration range can be broken into saturation and linear regimes. Thus

$$\begin{aligned}\frac{\tau_{rise}}{C} &= \int_0^{V_{iL}+V_{Tp}} \frac{dV_o}{\frac{K_p}{2}(V_{DD} - V_{iL} - V_{Tp})^2} \\ &\quad + \int_{V_{iL}+V_{Tp}}^{V_{oH}} \frac{dV_o}{K_p\left[(V_{DD} - V_{iL} - V_{Tp})(V_{DD} - V_o) - \frac{1}{2}(V_{DD} - V_o)^2\right]}\end{aligned}$$

We define $V_1 \equiv V_{DD} - V_o$ and $V_2 \equiv V_{DD} - V_{iL} - V_{Tp}$, so $dV_o = -dV_1$.
We get

$$\frac{K_p\tau_{rise}}{2C} = \frac{V_{iL} + V_{Tp}}{V_2^2} - \int_{V_2}^{V_{DD}-V_{oH}} \frac{dV_1}{2V_1V_2 - V_1^2}$$

The integral can be evaluated as

$$\begin{aligned} I &\equiv -\int_{V_2}^{V_{DD}-V_{oH}} \frac{dV_1}{2V_1V_2 - V_1^2} \\ &= \frac{1}{2V_2}\int_{V_{DD}-V_{oH}}^{V_2} \left(\frac{1}{V_1} + \frac{1}{2V_2 - V_1}\right) dV_1 \\ &= \frac{1}{2V_2}\left[\ln\frac{V_1}{2V_2 - V_1}\right]_{V_{DD}-V_{oH}}^{V_2} \\ &= \frac{1}{2V_2}\ln\frac{2V_2 - V_{DD} + V_{oH}}{V_{DD} - V_{oH}} \end{aligned}$$

Therefore,

$$\frac{K_p\tau_{rise}}{2C} = \frac{V_{iL} + V_{Tp}}{V_2^2} + \frac{1}{2V_2}\ln\frac{2V_2 - V_{DD} + V_{oH}}{V_{DD} - V_{oH}}$$

or

$$\frac{K_p\tau_{rise}}{2C} = \frac{V_{iL} + V_{Tp}}{(V_{DD} - V_{iL} - V_{Tp})^2} + \frac{1}{2(V_{DD} - V_{iL} - V_{Tp})}\ln\frac{2V_2 - V_{DD} + V_{oH}}{V_{DD} - V_{oH}}$$

Thus,

$$\begin{aligned} \tau_{rise} &= \frac{C(V_{iL} + V_{Tp})}{\frac{K_p}{2}(V_{DD} - V_{iL} - V_{Tp})^2} \\ &+ \frac{C}{K_p(V_{DD} - V_{iL} - V_{Tp})}\ln\frac{V_{DD} + V_{oH} - 2V_{iL} - 2V_{Tp}}{V_{DD} - V_{oH}} \end{aligned} \tag{1.23}$$

The first term is just the constant current charging of the load capacitor. The second term represents the charging by the pMOS in its linear range. This can be compared with resistive charging, which would have taken a charge time of

$$\tau = RC\ln\frac{V_{DD} - V_{iL} - V_{Tp}}{V_{DD} - V_{oH}}$$

to charge from $V_{iL}+ V_{Tp}$ to V_{oH}.

Fall time

When the input is high, the n channel transistor is 'on' and the p channel transistor is 'off'. If the output was initially 'high', it will be discharged to ground through the nMOS. To analysis the fall time, we apply Kirchoff's current law to the output node. This gives

$$I_{dn} = -C\frac{dV_o}{dt}$$

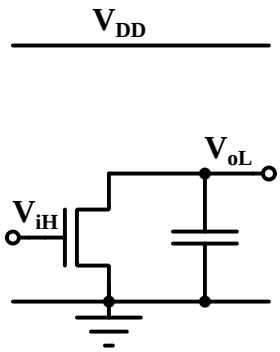


Figure 1.4: CMOS inverter with the pMOS 'off'

Again, separating variables and integrating from the initial voltage (= V_{DD}) to some terminal voltage V_{oL} gives

$$\frac{\tau_{fall}}{C} = -\int_{V_{DD}}^{voL} \frac{dV_o}{I_{dn}}$$

The n channel transistor will be in saturation till the output voltage falls to V_i- V_{Tn}. Below this voltage, the transistor will be in its linear regime. Thus, we can divide the integration range in two parts.

$$\begin{aligned}
\frac{\tau_{fall}}{C} &= -\int_{V_{DD}}^{V_i-V_{Tn}} \frac{dV_o}{I_{dn}} - \int_{V_i-V_{Tn}}^{V_{oL}} \frac{dV_o}{I_{dn}} \\
&= \int_{V_i-V_{Tn}}^{V_{DD}} \frac{dV_o}{\frac{K_n}{2}(V_i-V_{Tn})^2} \\
&\quad + \int_{V_{oL}}^{V_i-V_{Tn}} \frac{dV_o}{K_n[(V_i-V_{Tn})V_o - \frac{1}{2}V_o^2}
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{K_n\tau_{fall}}{2C} &= \frac{V_{DD}-V_i+V_{Tn}}{(V_i-V_{Tn})^2} + \int_{V_{oL}}^{V_i-V_{Tn}} \frac{dV_o}{2V_o(V_i-V_{Tn}) - V_o^2} \\
&= \frac{V_{DD}-V_i+V_{Tn}}{(V_i-V_{Tn})^2} + \frac{1}{2(V_i-V_{Tn})} \int_{V_{oL}}^{V_i-V_{Tn}} dV_o \left(\frac{1}{V_o} + \frac{1}{2(V_i-V_{Tn})-V_o}\right)
\end{aligned}$$

Which gives

$$\begin{aligned}
\frac{K_n\tau_{fall}}{2C} &= \frac{V_{DD}-V_i+V_{Tn}}{(V_i-V_{Tn})^2} + \frac{1}{2(V_i-V_{Tn})} \left[\ln \frac{V_o}{2(V_i-V_{Tn})-V_o}\right]_{V_{oL}}^{V_i-V_{Tn}} \\
&= \frac{V_{DD}-V_i+V_{Tn}}{(V_i-V_{Tn})^2} + \frac{1}{2(V_i-V_{Tn})} \ln \frac{2(V_i-V_{Tn})-V_{oL}}{V_{oL}}
\end{aligned}$$

and therefore

$$\tau_{fall} = \frac{C(V_{DD} - V_i + V_{Tn})}{\frac{K_n}{2}(V_i - V_{Tn})^2} + \frac{C}{K_n(V_i - V_{Tn})} \ln \frac{2(V_i - V_{Tn}) - V_{oL}}{V_{oL}} \tag{1.24}$$

Again, the first term represents the time taken to discharge at constant current in the saturation regime, whereas the second term is the quasi-resistive discharge in the linear regime.

1.2.4 Trade off between power, speed and robustness

As we scale technologies, we improve speed and power consumption. However, as we can see from the expression for noise margins, (eq 1.21 and eq 1.22) the noise margin becomes worse. We can improve noise margins by choosing relatively higher threshold voltages. However, this will reduce speeds. We could also increase V_{DD}- but that would increase power dissipation. Thus we have a trade off between power, speed and noise margins.

This choice is made much more complicated by process variations, because we have to design for the worst case.

1.2.5 CMOS Inverter Design Flow

The CMOS inverter forms the basis of most static CMOS logic design. More complex logic can be designed from it by simple thumb rules. A common (though not universal) design requirement is symmetric charge and discharge behaviour and equal noise margins for high and low logic values. This requires matched values of K_n and K_p and equal values of $V_{Tn}$and V_{Tp}. For a constant load capacitance, rise and fall times depend linearly on K_n and K_p. Thus it is a straightforward calculation to determine transistor geometries if speed requirements and technological parameters are given. However, as transistor geometries are made larger, self loading can become significant. We now have to model the load capacitance as

$$C_{Load} = C_{ext} + \alpha K_n$$

where we have assumed that $\beta = K_n/K_p$ is kept constant. α is a technological constant. We use the expressions for $K\tau/C$ which depend only on voltages. Once these values are calculated, the geometry can be determined.

In the extreme case, when self capacitance dominates the load capacitance, K/C becomes constant and τ becomes geometry independent. There is no advantage in using wider transistors in this regime to increase the speed. It is better to use multi-stage logic with tapered buffers in this regime. This will be discussed in the module on Logical Effort.

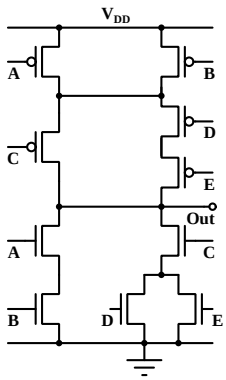


Figure 1.5: CMOS implementation of $\overline{A.B + C.(D + E)}$

1.2.6 Conversion of CMOS Inverters to other logic

Once the basic CMOS inverter is designed, other logic gates can be derived from it. The logic has to be put in a canonical form which is a sum of products with a bar (inversion) on top. For every '.' in the expression, we put the corresponding n channel transistors in series and the corresponding p channel transistors in parallel. for every '+', we put the n channel transistors in parallel and the p channel transistors in series. We scale the transistor widths up by the number of devices (n or p) put in series. The geometries are left untouched for devices put in parallel. Fig.1.5 shows the implementation of $\overline{A.B + C.(D + E)}$ in CMOS logic design style.