CMOS Gate Delays, Power, and Scaling

GATE DELAYS

In the last lecture (Lec. 15) we calculated the gate delay for a symmetrical CMOS inverter with

$$V_{Tn} = |V_{Tp}| \equiv V_T,\ C^*_{o\,xn} = C^*_{o\,xp} \equiv C^*_{o\,x},\ \text{and } K_n = K_p,$$

in which both the n- and p-channel devices were minimum gate length devices, i.e., $L_n = L_p = L_{min}$. The p-channel device was made twice as wide as the n-channel device to get the desired K equality, because we assumed $\mu_e = 2\,\mu_h$.

We found that the gate delay was given by:

$$\tau_{GD} \approx \frac{4\,C_L\,V_{DD}}{K_n(V_{DD} - V_T)^2}$$

Replacing C_L and K_n, to write this in terms of the device dimensions, we found after a bit of simple algebra:

$$\tau_{GD} \approx \frac{12\,n}{\mu_e}\,L^2_{m\,in}\,\frac{V_{DD}}{(V_{DD} - V_T)^2}$$

POWER

There is zero static power in CMOS so the only contribution is the dynamic power

$$P_{ave} = C_L\,V^2_{D\,D}\,f$$

where f is the operating frequency and C_L is the loading capacitance. This load will be the average fan-out, n, times the input capacitance of a similar CMOS gate, plus any parasitic interconnect capacitance:

$$C_L = n\,C^*_{o\,x}\,(L_{min}\,W_n + L_{min}\,W_p) + C_{parasitic}$$

$$= 3\,n\,C^*_{ox}\,L_{min}\,W_n + C_{parasitic}$$

Neglecting $C_{parasitic}$, we can write

$$P_{ave} = 3\,n\,C^*_{ox}\,L_{min}\,W_n\,V^2_{DD}\,f$$

MAXIMUM POWER

The maximum power dissipation will occur when the gate is operated at its maximum frequency (bit rate), which is in turn proportional to $1/\tau_{GD}$. Thus we can say

$$P_{ave\,max} \propto 3\,n\,C^*_{ox}\,L_{min}\,W_n\,V^2_{DD}\,\frac{1}{\tau_{GD}}$$

$$= \frac{1}{4}\,\frac{W_n}{L_{min}}\,\mu_e\,C^*_{ox}\,V_{DD}\,(V_{DD} - V_T)^2$$

$$= \frac{1}{4}\,K_n\,V_{DD}\,(V_{DD} - V_T)^2$$

The importance of keeping V_{DD} small is quite evident from this expression, but the situation is not black and white because making V_{DD} small makes τ_{GD} large; the same is true of making K_n small. The whole problem of what to reduce how while maintaining high performance and not frying the IC chips is a complex one and has led to the development of rules for scaling dimensions and voltages; we will discuss scaling rules after first looking at one more important parameter, the maximum average power dissipation per unit area.

POWER DISSIPATION PER UNIT AREA

In many situations the power dissipation per unit area is more important than the total power dissipation. To estimate how this factor varies with the device dimensions we make the assumption that the density of devices in an integrated circuit increases inversely with the gate area, $W_n L_{min}$. We have:

$$P_{density\ max} \propto \frac{P_{ave\ max}}{W_n L_{min}} \propto \frac{\mu_e C^*_{ox}}{4\, L^2_{min}} V_{DD} (V_{DD} - V_T)^2$$

SCALING RULES

We in general want to simultaneously reduce gate delays, decrease power dissipation, and increase packing density, while not exceeding a certain power density. The place we start is with a reduction of the gate length, but we quickly find we must do more than that or we get into trouble.

For example, as the gate length is reduced, the oxide thicknesses and the junction depths (of the sources and drains) must be reduced proportionally to obtain good transistor characteristics. One is essentially maintaining a long, thin geometry consistent with the gradual channel approximation, and this turns out to be just what is needed to get good saturation (flat curves; small g_o) of the device output (i_D vs v_{DS}) characteristics. Thus, if we reduce the minimum gate length, L_{min}, by a factor of s, we will also want to reduce the gate oxide, t_{ox}, by the same factor. To increase the packing density futher, we also reduce the gate width, W, by the same factor:

$$L_{min} \rightarrow L_{min}/s$$

$$W \rightarrow W/s$$

$$t_{ox} \rightarrow t_{ox}/s$$

With these changes we find that our gate delay, average power, device density, and power density change as follows:

$$\tau_{GD} \rightarrow \tau_{GD}/s^2$$