Problem Set 3 Solutions

February 10, 2015

1

(1 point) Draw a NOR2 gate and include the widths of its transistors in terms of W_{inv}, the width of an NMOS in an inverter. Assume $R_{0p} = 2R_{0n}$ and recall that we ignore C_D when sizing a gate.

Solution: In this gate the worst case pull down path is through a single NMOS transistor. That transistor will have the same width as one in an inverter: W_{inv}. This is denoted as a size of 1 in Figure 1, which depicts the sized NOR2 gate. The worst case pull-up path is through two PMOS devices, so the total pullup resistance would be $R_{pu} = 2R_{p0}/W_p = 2 \cdot 2 \cdot R_{0n}/W$. $W = 4W_{inv}$ will cancel out the factor of 4 in the R_{pu} expression.

2

(4 points) Calculate the pull-up and pull-down delays in a NAND4 gate sized, by our normal algorithm, to match delay with an inverter that has an NMOS width of W. Be sure to include C_D and C_G in these calculations.

Solution: The NAND4 gate has a pull-down network of four series NMOS devices, each needs to have a relative width 4. The pull up network is four parallel PMOS devices, each needs to have a relative width of 2. That means that pull-up is handled by a fairly simple delay calculation

$$t_{NAND4,pu} = \ln(2)(2\frac{R_{0n}}{2W})(12C_DW + C_l) = \ln(2)(12R_{0n}C_D + \frac{R_{0n}C_l}{W}) \tag{1}$$

The pull down expression is more complicated because of the many drain and gate caps. A full sketch of these capacitances indicating those which we can ignore (in red) appears in Figure 2. These capacitors can be ignored because they are attached to a DC voltage on both sides, so no charge flows into or out of them. We can say that the inputs are DC because we assume they are settled when we're doing this rise/fall time analysis. We can merge capacitors attached to V_{dd} and capacitors attached to ground as if they are in parallel because their opposite plates are DC values; if we were to look at the Thevenin resistance seen by that capacitor, V_{dd} would be shorted to ground.

We invoke Elmore delay to compute the delay.

$$t_{NAND4,pd} = \ln(2)(\frac{R_{0n}}{4W}(8C_DW + 4C_GW) + \frac{2R_{0n}}{4W}(8C_DW + 4C_GW)\ldots \tag{2}$$

$$\ldots + \frac{3R_{0n}}{4W}(8C_DW + 4C_GW) + \frac{4R_{0n}}{4W}(12C_DW + C_l)) \tag{3}$$

$$= \ln(2)R_{0n}(24C_D + 6C_G + C_l/W) \tag{4}$$

Using a complex gate like a NAND4 comes at a cost: we incur a delay penalty from the internal C_D caps and our input will have more total capacitance because the complexity of the gate means that we need to upsize our transistors. This doesn't mean that it is always bad to use complex gates; in the case the C_l is very small, you might pay less total capacitance for a single complex gates than a cascade of small ones. That said, for many logic applications it is better to use a series of small gates. VLSI covers rigorous techniques for this kind of logic design.

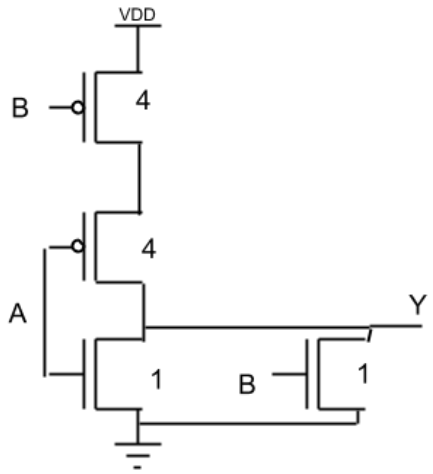

Figure 1: A NOR2 Gate

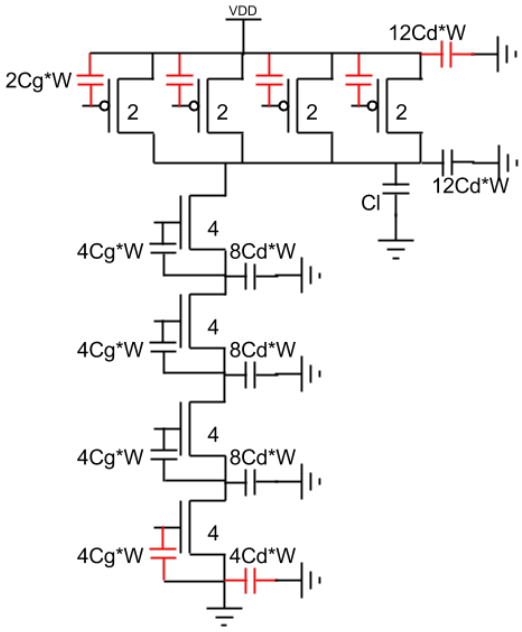

Figure 2: A NAND4 Gate with parasitic caps annotated. Capacitors with a fixed charge are highlighted in red. These capacitors don't affect circuit dynamics.

This problem also grants insight into some common digital practices. Often the bottom transistors in big stacks are made extra wide since they contribute to every Elmore term. In addition, early arriving signals are usually assigned to gates at the bottom of big transistor stacks so that intermediate nodes can be discharged.

One point for correctly sizing the gate, one point for $12C_D$ at the outuput, one point for C_D in the internal nodes in pull down, one point for C_G on the internal nodes on pull down.

3

3.1

(3 points) It is possible to express the delay of an inverter as $t_d = t_{inv}(f+\gamma)$ where $f = C_l/C_{in}$ and $\gamma = C_D/C_G$. What is t_{inv}? Note that $C_{in} = C_{GN,tot} + C_{GP,tot} = 3W_{inv}C_G$. Note that C_G and C_D are capacitance densities (capacitance per unit width).

Solution: If our gate is properly designed, then the delay for the pull up transition and the pull down transition will be the same.

$$t_d = \ln(2)\left(\frac{R_{0n}}{W}3C_DW + \frac{R_{0n}}{W}C_l\right) \tag{5}$$

$$= \ln(2)R_{0n}\left(3C_D + \frac{C_l}{W}\right) \tag{6}$$

$$= \ln(2)R_{0n}3C_G\left(\frac{C_D}{C_G} + \frac{C_l}{3WC_G}\right) \tag{7}$$

$$= 3\ln(2)R_{0n}C_G(\gamma + f) \tag{8}$$

$$\rightarrow t_{inv} = 3\ln(2)R_{0n}C_G \tag{9}$$

Note that t_{inv} is independent of the size of the gate. In fact, it only depends on technology parameters, so t_{inv} can be measured on test chips and given to designers. Designers are then able to design delays using only knowledge of the raio of on-chip capacitances.

Sometimes you will see a different version: $t'_{inv} = \ln(2)R_{on}C_G\gamma$, which means that the delay is reported as $t_d = t'_{inv}\left(1 + \frac{f}{\gamma}\right)$. This is more annoying to calculate with, but it means that t'_{inv} has a direct physical meaning: it is the delay of a totally unloaded (fanout=0) inverter.

2 points for getting delay setup right. 1 point for evaluation.

3.2

(3 points) Express the delay of three inverters in series driving a load of $C_L = 48C_GW_{inv}$ using this form of the delay. Assume the first inverter has a NMOS width of W_{inv}, the second has an NMOS width of $4W_{inv}$ and the last has a width of $16W_{inv}$. Also assume $\gamma = 1$. Could the delay of this series of inverters be reduced by changing the widht of the second or third inverter? Would it change the power consumption?

Solution: Using the above expression we can quickly sum the delays of the inverters:

$$t_d = t_{inv}(\gamma + f_1) + t_{inv}(\gamma + f_2) + t_{inv}(\gamma + f_3) = t_{inv}(3\gamma + f_1 + f_2 + f_3) \tag{10}$$

The value f in this expression refers to the fanout, which is the ratio of the load capacitance to the input capacitance of a gate. For our first two inverters, the fanout is easy to determine as the ratio of their widths. To convince you of that, I have calculated their input capacitances below:

$$C_{in,1} = 3C_G(W_inv) = 3C_GW_{inv} \tag{11}$$

$$C_{in,2} = C_{l,1} = 3C_G(4W_{inv}) = 12C_GW_{inv} \rightarrow f_1 = 4 \tag{12}$$

$$C_{in,3} = C_{l,2} = 3C_G(16W_{inv}) = 48C_GW_{inv} \rightarrow f_2 = 4 \tag{13}$$

$$C_l = 48C_GW_{inv} \rightarrow f_3 = 1 \tag{14}$$

Thus we can caluclate that $t_d = 12t_{inv}$.

If we reduce the size of the third gate so that it's NMOS is $8W_{inv}$, then $f_2 = 2$ and $f_3 = 2$. This reduces the sum of our delays to $t_d = 11t_{inv}$ and reduces the total transistor width in our chain. Reducing the total width reduces the total capacitance, which reduces the total power in turn. Sizing your gates matters a lot for improving power and delay performance. There are rigorous ways to size gates optimally. Some of them are covered in VLSI, and we'll talk more about the sense of power optimality in two weeks.

1 point for correctly calcualting fanouts, 1 point for any counterexample that reduces delay, 1 point for recognizing that reducing total transistor width will reduce power

4

(3 points) A capacitor is connected to a voltage source which has a variable value (i.e. a "step source"). The source steps from 0 to $V_{dd}/2$, remains at that value until the capacitor is fully charged, and then steps from $V_{dd}/2$ to V_{dd}. What is the total amount of energy pulled from the source?

Solution: For the first step, the charge pulled from the supply is $CV_{dd}/2$ and the supply value is $V_{dd}/2$. So the total energy is

$$E_{step1} = QV_{sup} = C \cdot V_{dd}/2 \cdot V_{dd}/2 = CV_{dd}^2/4. \tag{15}$$

For the second step the charge pulled from the supply is $Q = C(V_{dd} - V_{dd}/2) = CV_{dd}/2$ and the supply value is V_{dd}. The total energy is

$$E_{step2} = QV_{sup} = C \cdot V_{dd}/2 \cdot V_{dd} = CV_{dd}^2/2. \tag{16}$$

Adding the energy from the steps together we find

$$E_{tot} = 3CV_{dd}^2/4. \tag{17}$$

This is a bit weird. A single step to V_{dd} on this cap would pull CV_{dd}^2 from the supply. Slowly charging the capacitor has resulted in less total energy being used. This is called adiabatic charging and people have tried to build computer systems around the concept. They largely didn't work, leakage energy (a subject for next week) killed the power savings.

1 point first step, 1 point second step, 1 point total.

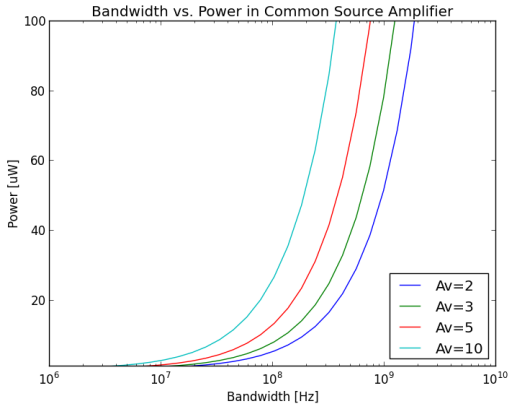

Figure 3: Bandwidth vs. Power in a common source amplifier

5

Assume $C_D = 1\text{fF}/\mu\text{m}$, $V_{th,p} = V_{thn} = 0.4\text{V}$ and $V_{dd} = 1\text{V}$. Make sure all plots capture interesting features of the curves they depict, vary the range of sweeps to capture this if necessary.

5.1

(3 points) Plot the bandwidth vs. power curve for a common source amplifier. Use the minimum width amplifier for any given bandwidth. Finding this minimum width will require an iterative solution. Include curves for gains of 2, 3, 5, and 10.

This question requires us to formalize the relationship between g_m and I_d, it is

$$g_m = \frac{2I_D}{V_{gs,dc} - V_{th}} \text{ when } I_d/W < J_{crit} \tag{18}$$

where J_{crit} is some critical current density. We'll use $J_{crit} = 100\mu\text{A}/\mu\text{m}$.

In a common source amplifier, the $V_{gs,dc}$ would probably be $\approx 0.8\text{V}$. This limits the maximum input swing the amplifier can accept, but figuring out how to fix that is a subject for another class. Let $C_L = 20\text{fF}$.

Solution: The curve appears in figure 3. I've chosen to highlight the power wall that amplifiers hit by using a log-linear plot. Power increases linearly with bandwidth and then blows up due to self-loading. There's only so much bandwidth that you can squeeze out of an amplifier before the power makes it infeasible.

There are some tricks to bypassing this bandwidth limit – shunt peaking and delay line amplification for instance – but really wide band amplifiers are ultimately limited by their own load capacitance.

Note that adding a buffer wouldn't help here, eventually something has to drive the load capacitance, and consequently the buffer would mostly be driving itself too.

1 point if plot has right general trends, 1 point for shape matching, 1 point for numbers matching

5.2

(3 points) Plot the delay vs. energy curve for an inverter driving the load capacitance. Include curves for $C_L = 12\text{fF}, 40\text{fF}$ and 200fF. Assume $R_{0n} = 1\text{k}\Omega \cdot \mu\text{m}$.

Solution: The energy delay curve appears in Figure 4. Like the amplifier, at low delay (high frequency) the energy explodes as more and more energy is diverted to driving the inverter's own load capacitance. This can actually be compared very directly to the time constant of an amplifier, and I've included a plot of the time constant from the first question vs. power in Figure 5

1 point if plot has right general trends, 1 point for shape matching, 1 point for numbers matching

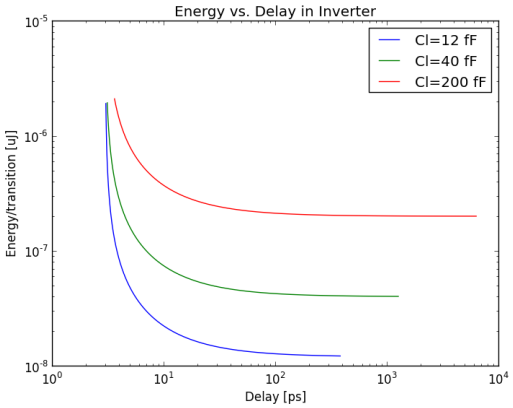

Figure 4: Energy vs. Delay in an inverter

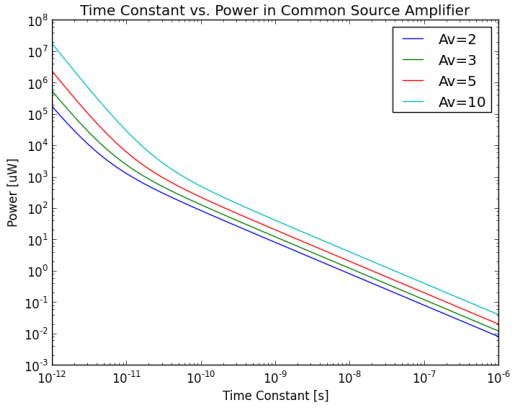

Figure 5: Time constant vs. power in a common source amplifier for direct comparison to energy-delay of inverter.