EECS 151/251A Homework 7

Due Monday, March 20^{th}, 2024

Introduction

This homework is meant to test your understanding of the basic principles of transistor sizing, capacitive loads, and their impact on performance of digital circuit.

Problem 1: Transistor Sizing

Suppose you are given the layout information for some particular mobile integrated circuit (IC) developed with a planar process. A new technique you learned allows for you to change the size of all the transistors without making any other changes to the layout. Smaller transistors might mean less area consumed for your design, and therefore, less cost to manufacture the IC. However, the transistor sizing also impacts the performance of your circuit. Answer the following questions regardless the impact of transistor sizing on the performance:

1. Explain the impact of transistor sizing on the maximum clock frequency in general (i.e. How does the frequency change with larger transistors? How does the frequency change with smaller transistors?). Why does the clock frequency change due to transistor size?
2. Now assume you decide to "compact" the design to take advantage of the area savings from smaller transistors after decreasing the size of all the transistors. Explain now what would be the overall effect on frequency from the original design.
3. After shrinking the transistors, you realized one gate has a larger fanout (ex. an electrical fanout of 10, the output is connected to the input of 10 other gates). How could you compensate now that you have decreased the transistor size?

Solution:

1. For this analysis, let's consider to cases

   (a) A circuit composed of only gates scale so that the load of any gate is only the input capacitance of a gate(s). In this scenario, everything scales relative to width, therefore, the maximum frequency does not change (think inverters $t_{1/2} = ln(2)RC$); R and C have the opposite scaling relationship to width so that scaling factor cancels out).

   (b) A circuit with some constant capacitive load, such that the capacitance of the load is not dependent of transistor sizing (i.e. a capacitor or wire that does not decrease in length after transistor shrinking). A smaller transistor has lower drive strength and greater resistance, therefore the delay driving such a load is longer. Consequently, the maximum frequency would actually decrease*****. The opposite is true if we increase the width.

   .

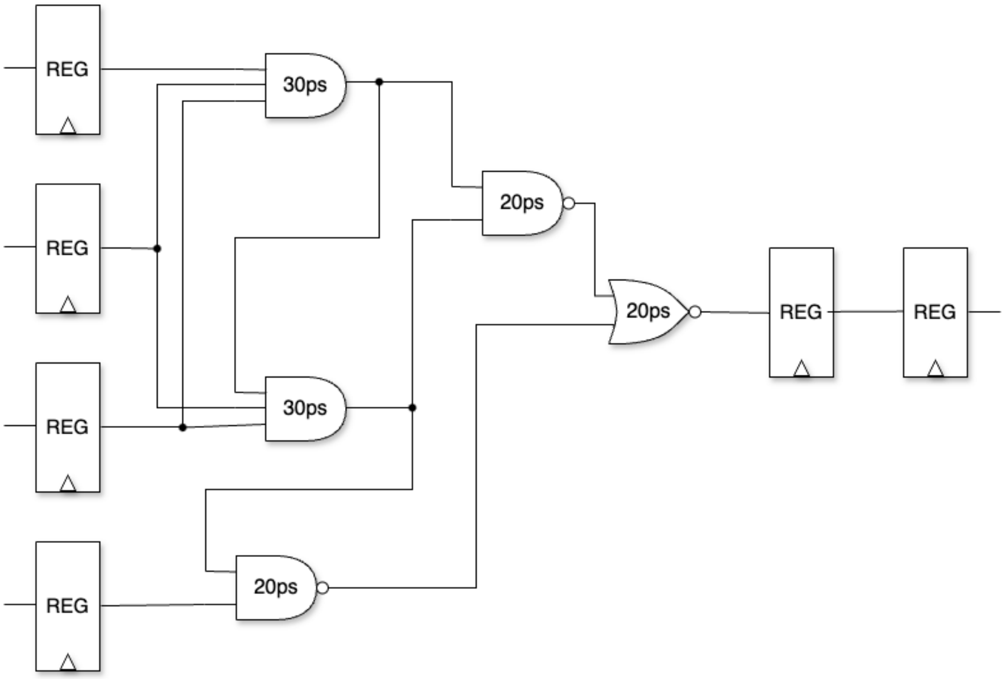
   ***Note**: This conclusion is actually a simplification. Due to device geometries, the drain capacitance does scale the same rate relative to width as resistance does with width. Therefore, when we scale up the capacitance actually does not increase as much as the resistance decreases. Therefore, larger transistors can switch faster due to reduced relative capacitance!

2. The maximum clock frequency (f_{max}) will either stay the same or decrease! Only the transistor width changed, the topology and circuit remained constant therefore the critical path is the same. In this case, the circuit has no constant capacitive load, the f_{max} remains constant. In the case whether there is a constant capacitive load, the f_{max} decreases due to the reduce drive strength of the transistors.

3. The natural inclination is to increase the transistor size as larger transistors have greater drive strengths. However, **all** transistors have been shrunk, therefore the only solution is to combine multiple transistors (i.e. put transistors in parallel and connect their gates together). The effect is similar to widen a single transistor which increases the combined drive strength. One could also add a buffer.

Problem 2: Retiming

Consider the circuit shown below. Assume all timing characteristics for the flip-flops are $10ps$ (i.e $t_{clk \to q}$, t_{setup}, and t_{hold} times). The delay for the gates are written within each gate symbol. *Ignore wire delay in this problem.*



1. What is the maximum clock frequency (f_{max}) for this circuit?
2. Without optimizing the logic, but only by retiming the circuit, now what is the new f_{max}? Draw your retimed circuit.
3. The simplest form of retiming does not allow the addition of extra registers (i.e the latency does not increase). With pipelining you add registers to long delay paths to increase clock frequency, but also can increase latency. Take the pipelining approach. What would be the best f_{max} you can achieve by adding an arbitrary number of new registers?

Solution:

1. This circuit has two paths will the same delay. Either path can serve as the critical path. One critical path exists from the top register on the left to the first register of the two registers on the right.

   Critical Path Delay: $(t_{clk\rightarrow q} + t_{AND} + t_{AND} + t_{NAND} + t_{OR} + t_{setup}) =$

   $10ps + 30ps + 30ps + 20ps + 20ps + 10ps = 120ps$

   Therefore, $f_{max} = \frac{1}{120ps} = 8.33GHz$

2. The two registers on the right provide no relief to the timing, therefore we can rebalance the registers by using this register to break the critical path. The optimal placement after the top AND gate. Consequently, in order to ensure functional the output of the second AND gate and the NAND gate also must be registered. Therefore we add two regsiters at there output. This is a retiming example where registers are added! The important thing is that the clock cycle latency remained constant!

3. Add pipeline registers at the output of each gate. The critical path would then be through a single AND gate: $10ps + 30ps + 10ps = 50ps \implies 20GHz$.

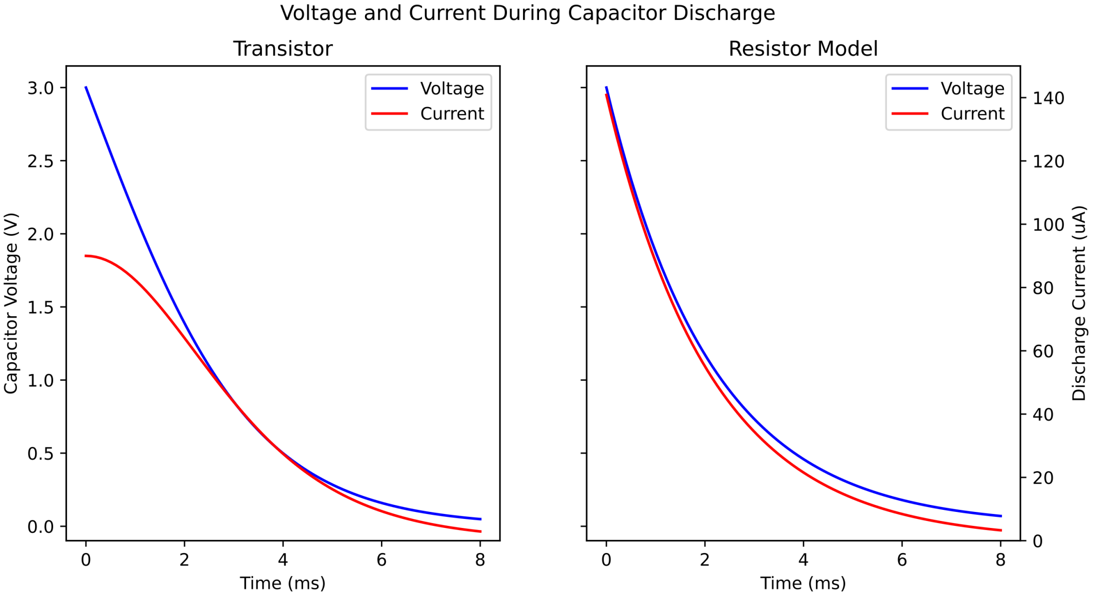
Problem 3: CMOS Circuits Driving a Load

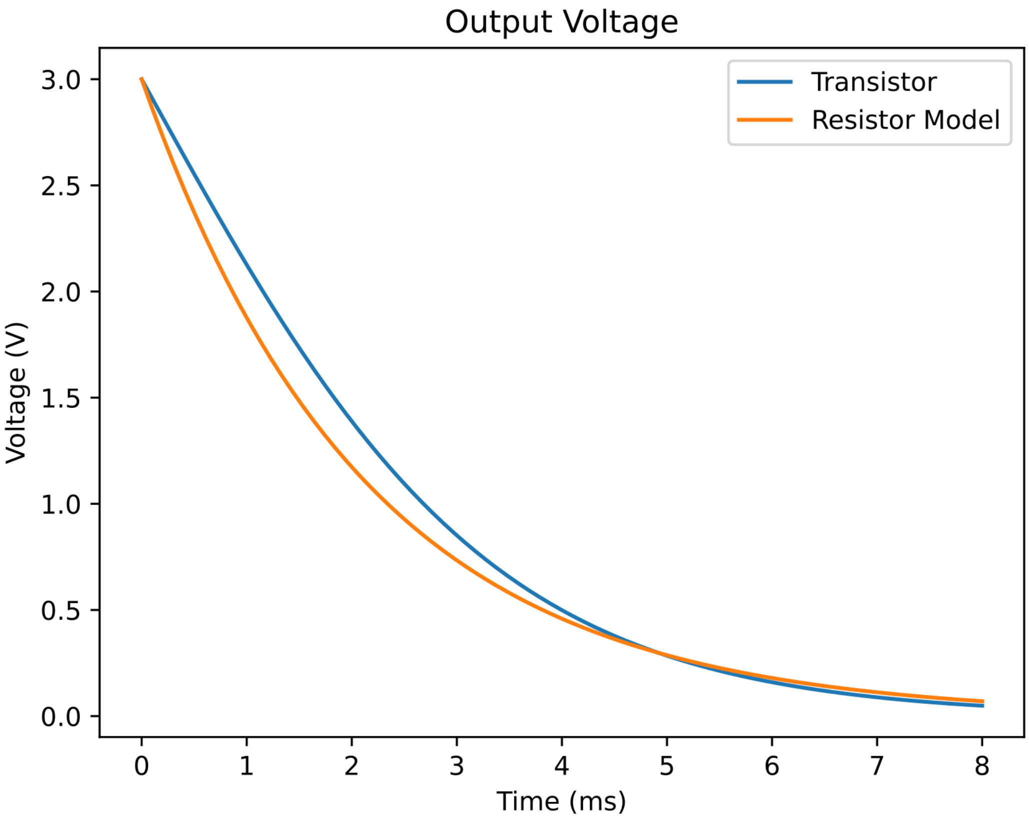
Consider a CMOS circuit in which a single NFET is used to pull the output capacitance from V_{DD} to 0 volts (as in an inverter). This circuit is driving a capacitive load. Sketch, on the same axis, the voltage on the output capacitence as a function of time assuming two different models for the transistor discussed in lecture: (1) transistor model as a simple resistor and (2) the non-linear model shown in the lecture notes.

Solution:

To understand this problem, you have to think about current. Current through a resister is linear with the voltage across the resistor. For a transistor, the current has three regimes based upon the I-V curve.

For a resistor model, the transistor is completely replace with a single resistor. Consequently, the circuit becomes a simple RC circuit so the voltage across the resistor follows an exponential decay. In the non-linear model, the transistor has three regimes. The saturation regime is important in this problem. The V_{DS} is greater than $V_{GS} - VTH$ and therefore I_{DS} is constant, therefore the voltage across the transistor is linear. This voltage decreases linearly until the transistor is no longer saturation as the capacitor drains and $V_{DS} < V_{GS} - VTH$. Once the transistor is no longer saturated, the current is proportional to V_{DS} and the voltage exponentially decays like an RC circuit.




Output Voltage
Transistor
Resistor Model
Voltage (V)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7
8
Time (ms)

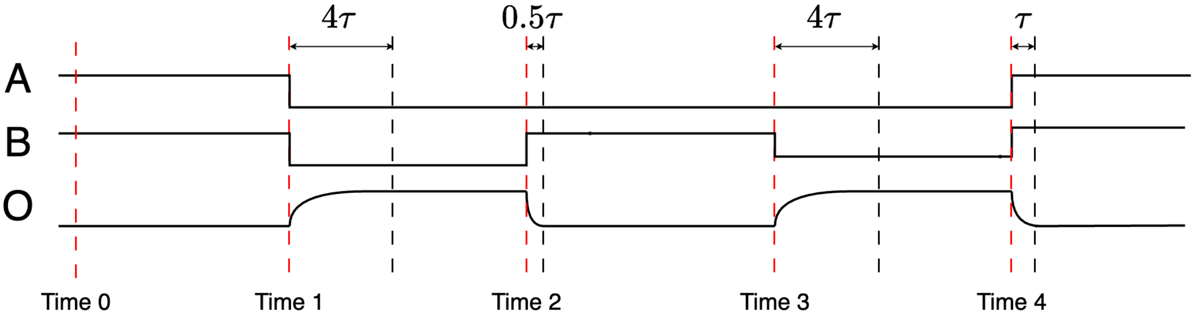
Problem 4: Rise and Fall

Consider a NOR gate, NOR(a,b), designed for a planar CMOS process. For this problem, we will assume that all transistors in the gate (all PFETs and all NFETs) are designed to the same physical width. On a single set of axis, and with modeling the transistors as resistors, draw rough waveforms for output transitions based on the following input values. The table below shows the input values over time:

Time	a	b
0	1	1
1	0	0
2	0	1
3	0	0
4	1	1

Solution:

For this problem, consider the resistance only (the capacitance of the transistors are the same; this is not *fully* true, but we will assume true for this class ;-)). PMOS transistors or PFETs have twice the resistance of an NFET (R). The PFETs in a 2-input NOR are in series, therefore the total resistance of the PUN is $4R$. The NFETs are in parallel, therefore the total resistance of the PDN is $R/2$ if both transistors are on, and R is only one is on. Consequently, switching on the PDN is 4x faster than the PUN.



Problem 5: Just in Time

Assume the critical path in our design is through the next state logic of a FSM. The registers we used have $t_{clk\to q} = t_{setup} = 280ps$, and $t_{hold} = 556ps$. The delay through the combinational logic $t_{CL} = 4ns$. We set the clock period $T = 5ns$. Answer the questions below:

1. Will this circuit function correctly? If not, what can we change to correct it without a lose of performance? What the maximum frequency you can run the circuit at? Make sure to show your work.
2. A colleague points out another path between two registers (of the same type as in the critical path) except $t_{CL} = 200ps$. Why is this bad? What can you do to fix this issue?

Solution:

1. Check both setup and hold timing

   $$Setup : 5000ps - 280ps - 4000ps - 280ps = 440ps > 0 \text{ (}440ps\text{ of slack)}$$
   $$Hold : 280ps + 4000ps = 4280ps > 560ps$$

   The maximum frequency you can run at is $\frac{1}{5ns - .44ns} \approx 219MHz$
2. Hold time is violated on this path. You need to delay the path somehow. One example, is to add a chain of inverters. Another more complex solution is to redesign your circuit.

Problem 6: Electrical Fanout

Suppose we have a unit sized inverter, with input capacitance of $10fF$ and an intrinsic delay of $2ps$, that needs to drive a large capacitance of $6pF$. Assume for this process, $\gamma = 1$.

1. Calculate the delay of this single inverter driving the large capacitance.
2. Using staged buffers, calculate the optimal number of stages, and the total delay. Show your work.

Solution:

1. First, calculate the resistance of the inverter. $2ps = ln(2)R(10fF) \Rightarrow R = 288.539\Omega$. Therefore, the total total is: $2ps + ln(2)R(6pF) = 1202ps$.

   A much more efficient calculation is to use fanout. The delay is: $2ps(1 + \frac{6pF}{(1)10fF})$.
2. Assume optimal delay if each stage has same delay. Minimize the function $N + N\sqrt[N]{\frac{6pF}{10fF}} \approx$ 5 stages. The total delay is: $2ps(5 + 5\sqrt[5]{\frac{6pF}{10fF}}) \approx 46ps$

Problem 7: Delay through Chain of Inverters

Suppose we have three inverters daisy chained together. The wire between the first and second inverters is 100μm and the wire between the second and third inverters is 5μm. All inverters are of the same size and have input capacitance of 10 fF, internal capacitance of 10 fF, and effective pullup and pulldown resistance of 1kΩ. Both wires have a characteristic capacitance of 1 $fF/\mu m$ length and a characteristic resistance of 0.1 $\Omega/\mu m$ length.

1. Calculate the delay from the first inverter turning on completely to when the input to the third inverter reaches $V_{DD}/2$.
2. Without resizing transistors, what can you do to reduce this delay?
3. Approximately, what improvement could you achieve?

Solution:

1. Model the resistance and capacitance of the wire with equivalent resistor and capacitor ($0.38 * characteristic * L$). The load capacitance of any inverter is the wire capacitance and the gate capacitance of the next inverter.

$$R_{INV} = 1k\Omega$$
$$C_{int} = 10fF$$
$$C_{gate} = 10fF$$
$$R_{wire1} = 100\mu m(0.1\Omega/\mu m)$$
$$C_{wire1} = 100\mu m(1fF/\mu m)$$
$$R_{wire2} = 5\mu m(0.1\Omega/\mu m)$$
$$C_{wire2} = 5\mu m(1fF/\mu m)$$

Total delay: $0.69(R_{INV}(C_{int} + C_{wire1} + C_{gate}) + R_{wire1}(\frac{0.38}{0.69}C_{wire1} + C_{gate}) + R_{INV}(C_{int} + C_{wire2} + C_{gate}) + R_{wire2}(\frac{0.38}{0.69}C_{wire2} + C_{gate}) + R_{INV}C_{int}) \approx 103.95ps$

2. Move the second transistor closer to the first to decrease the length of the first wire
3. This expression is minimized if the total wire length in partitioned in half (this is left as an exercise; the total delay because linear with total wire length). Each wire would now be $\frac{105}{2}\mu m$. The new delay is $103.78ps$, which is a 0.16% decrease