1 Minimum-energy design

In this problem, we may either change the block device sizings or drop the supply voltage, so we will analyze both. First we will check the benefit from increasing device size. There are multiple ways to solve this problem, but here we will use the constrained optimization technique described in lecture 19.

\[ \Lambda(x) = E(x) + \lambda(D(x) - D_{\text{max}}) \]

Note that this is solving for the minimum energy under a maximum delay constraint, as stated in the problem.

\[ E = E_A + E_B, D = D_A + D_B \]

So we can write

\[ \Lambda(t_A, t_B) = \frac{6 \cdot 10^{-21} Js}{t_A} + \frac{10^{-21} Js}{t_B} + \lambda(t_A + t_B - 5\text{ns}) \]

To optimize, we can differentiate \( \Lambda \) with respect to \( t_A \) and \( t_B \), set equal to 0, and then solve for \( \lambda \). This yields two equations

\[ \lambda = \frac{6 \cdot 10^{-21} Js}{t_A^2}, \lambda = \frac{10^{-21} Js}{t_B^2} \]

Setting them equal to each other and then substituting either \( t_A \) or \( t_B \) into \( t_A + t_B = 5\text{ns} \) will give

\[ t_A = 3.55\text{ns}, t_B = 1.45\text{ns} \]

Finally, using the given energy-delay tradeoffs, we find that the minimum energy per block is

\[ E_A = 1.69pJ, E_B = 0.68pJ \]

which means \( E_{\text{tot,min}} = 2.37pJ \).

Now we analyze dropping the supply voltage. We know that increasing delay by 1% corresponds to dropping the energy by 2%. From the sizing energy delay tradeoffs and the initial delays, we know that the starting energy is at

\[ E_{\text{tot}} = E_A + E_B = 2pJ + 1pJ = 3pJ \]

We can calculate that we can drop the supply voltage by up to 20%, because that is the point at which the delays of the two blocks sum to 5ns. This corresponds to an energy decrease of 36%, which means that the total energy would be \( E_{\text{tot}} = 1.92pJ \).

Therefore, the minimum energy for this design is \( E_{\text{tot}} = 1.92pJ \), achieved by dropping the supply voltage.
2 Energy-delay optimization

a Minimize energy

For switching power,

\[ E_{sw} = \alpha(2C_g + 2C_d + C_w)V_{DD}^2 \]

Clearly, \( C_g \) and \( C_d \) increases linearly with inverter sizing, so we want to make them minimum-sized. Leakage, while ignored, also scales with sizing, thus supporting this conclusion.

b Minimize delay

The delay of this chain is (note the second inverter is unloaded):

\[ t_d = t_{unit}(1 + \frac{C_w + C_g}{C_g}) + t_{unit} \]

To minimize the delay, the fanout term on the first inverter should be minimized. Therefore, for minimum delay, we need to maximally-size the inverters.

c Minimize EDP

The EDP is a multiplication of the equations from a) and b), simplified since \( C_d = C_g \) and only for inverter 1:

\[ EDP = \alpha(2C_g + C_w)V_{DD}^2 \cdot t_{unit}(1 + \frac{C_w + C_g}{C_g}) \]

\[ = \alpha t_{unit}V_{DD}^2(4C_g + 4C_w + \frac{C_w^2}{C_g}) \]

Taking the derivative w.r.t. \( C_g \) and setting it equal to 0:

\[ \frac{\partial EDP}{\partial C_g} = 4 - \frac{C_w^2}{C_g^2} = 0 \]

\[ C_g = \frac{C_w}{4} \]

\[ C_g = 6 \text{ fF} \]

\[ W_p = W_n = 1.5 \mu m \]

d Sizing second inverter

The delay expression with a load capacitor is now:

\[ t_d = t_{unit}(1 + \frac{C_w + C_2}{C_1}) + t_{unit}(1 + \frac{C_L}{C_2}) \]

To minimize the delay, we set the derivative w.r.t. \( C_2 \) equal to 0:

\[ \frac{\partial t_d}{\partial C_2} = t_{unit}(\frac{1}{C_1} - \frac{C_L}{C_2}) = 0 \]

\[ \frac{C_L}{C_2} = \frac{1}{C_1} \]

\[ C_2 = \sqrt{C_1C_L} \]

\[ C_2 = 12.728 \text{ fF} \]

\[ W_{p,2} = W_{n,2} = 3.182 \mu m \]
3 Stack forcing

Since this inverter chain has no load, all 6 inverters are to be identically-sized for minimum delay. The overall delay as a function of $t_{\text{unit}}$ is:

$$t_{d,i} = \sum t_{\text{unit}}(1 + gh) = t_{\text{unit}}(5(1 + 1) + 1) = 11t_{\text{unit}}$$

The total energy consumed by this chain with 30% leakage is:

$$E_{\text{tot},i} = \frac{1}{0.7} \alpha(6C_g + 6C_d)V_{DD}^2 = 17.14\alpha C_g V_{DD}^2$$

By implementing stack forcing as seen in lecture (stack of 2 NMOS with W/2), the NMOS drive current is reduced by a factor of 3. **Note:** the PMOS is not stacked, so the current pulling up is unchanged while the current pulling down is decreased by a factor of 3. Furthermore, the drain capacitance from the NMOS is halved, so there is a factor of 0.75 on the parasitic delay term. Using the $t_d$ formula from above, for an input rising edge:

$$t_{p,LH} = t_{\text{unit}}(3(3 \cdot 0.75 + 3) + 2(0.75 + 1) + 0.75) = 20t_{\text{unit}}$$

while for an input falling edge:

$$t_{p,HL} = t_{\text{unit}}(3(0.75 + 1) + 2(3 \cdot 0.75 + 3) + 3 \cdot 0.75) = 18t_{\text{unit}}$$

The average delay is therefore $19t_{\text{unit}}$.

For the energy calculation, the drain capacitances have a 0.75 factor on them, but $\alpha$, and $V_{DD}$ are unchanged. The leakage is reduced by 10x, so the new energy is:

$$E_{\text{tot,stack}} = 0.03 \cdot 17.14\alpha C_g V_{DD}^2 + \alpha(6C_g + 0.75 \cdot 6C_g)V_{DD}^2 = 11.014\alpha C_g V_{DD}^2$$

With this, we can calculate the sensitivity:

$$S = \frac{\Delta E/E_{\text{tot},i}}{\Delta t_d/t_{d,i}}$$

$$= \frac{(11.014 - 17.14)/17.14}{(19 - 11)/11}$$

$$= -0.49$$

Intuitively, this makes sense. The energy decreases, but less so than the hit in the delay.