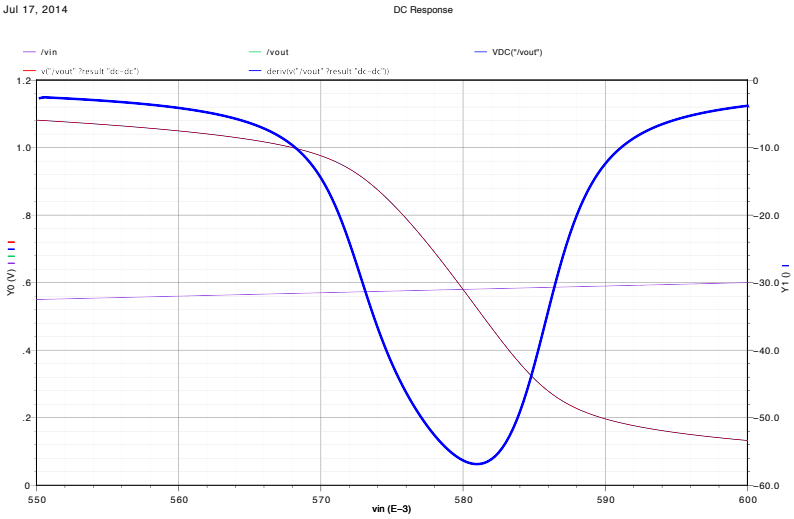
Solution **Integrated Circuit Design MCC091** Monday August 25, 2014

1. Inverter switching voltage and gain.

a) From the diagram we determine that voltage where $V_{IN} = V_{OUT}$ is $V_{SW} = 0.58$ V.

b) You had to determine the gain by measuring the slope of the curve in the diagram. Below we have let Cadence calculate the derivate of V_{OUT} w.r.t. V_{IN} along the curve. The small-signal gain is that derivate. At V_{SW} the gain is -58 times.



c-e) The small-signal gain is the derivative of V_{OUT} w.r.t. V_{IN}. For the inverter we have this expression for the small-signal gain

$$|A_v| = (g_{m_n} + g_{m_p}) r_{d_n} || r_{d_p} = \frac{g_{m_n} + g_{m_p}}{g_{d_n} + g_{d_p}}$$

where g_m is the transconductance and g_d is the output conductance. With the quadratic current equations we have these approximate expressions for the two small-signal parameters:

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \approx \frac{2I_D}{V_{GS} - V_T}$$

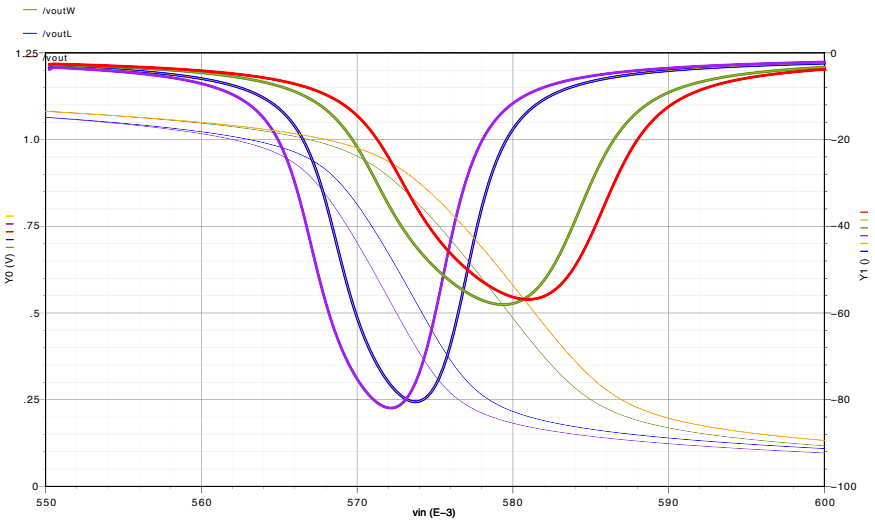
$$g_d = \frac{\partial I_D}{\partial V_{DS}} \approx \frac{I_D}{V_A}$$

(These are the n-transistor equations but the p-transistor ones are the same but with opposite voltages and absolute signs added on all currents and voltages). So the drain current will cancel and not be part of the expression for the gain:

$$|A_v| = \frac{g_{m_n} + g_{m_p}}{g_{d_n} + g_{d_p}} \approx \frac{\frac{2}{V_{GTn}} + \frac{2}{|V_{GTp}|}}{\frac{1}{V_{A_n}} + \frac{1}{|V_{A_p}|}}$$

Thus, the inverter gain depends only on the effective gate voltages and the Early voltages. The effective gate voltages will remain the same when the ratio between the two transistors is preserved. The Early voltages are higher for longer transistors, but do not depend on the transistor width. Thus, the answers are c) higher d) higher e) the same.

A simulation similar to the one above for all four cases shown below indicates that this is really the case. In our 65-nm process the magnitude of the gain increases from around 60 to around 80 when the length is doubled (from 1 µm to 2 µm).



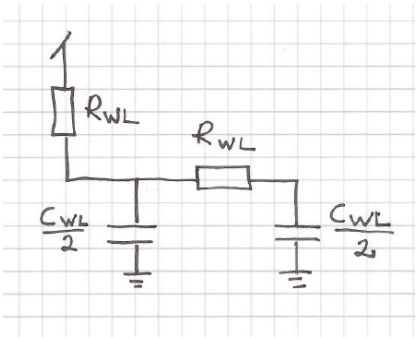
f) In the traditional model of channel-length modulation the Early voltage, V_A, is proportional to the transistor length so in c) and d) the magnitude of the gain doubles when the transistor length is doubled.

2.

a) The number of squares for one WL wire is its length/width = 128 / 0.1 = 1280. Thus, the resistance of the WL is R_{WL} = 1280 * 0.1 = 128 Ω.

b) The capacitance of one WL is C_{WL} = length * (C_{GND} + 2 * $C_{INTERWIRE}$) + #cells * 2 * C_G fF. In this case we have C_{WL} = 128 (0.1 + 2 * 0.02 + 2 * 0.1) = 128 * (0.34) = 43.5 fF.

c) Here is the model:



The delay can be computed as t_{dWL} = 2/3*$R_{WL}C_{WL}$ = 3.7 ps.

d) The energy is computer as E_{WL} =$C_{WL}V_{DD}^2$ = 43.5 pJ since only one WL is charged for each reading of the memory.

e) Resistance: The length of the wire is halved, but its width is not changed since the wire already has the minimum width. Thus, we have half the number of squares which gives R_{WL2} = R_{WL}/2 = 64 Ω. Capacitance: The parallel M2 wires are now approximately at half the distance they were before. If we assume plate capacitances, the capacitance doubles so we have $C_{INTERWIRE2}$ = 2* $C_{INTERWIRE.}$ C_{GND} and C_G, and the number of cells remain the same. C_{WL2} = length* (C_{GND} + 2 * $C_{INTERWIRE2}$) + #cells * 2 * C_G fF. So in this case we get: C_{WL2} = 64* (0.1 + 2 * 0.04 + 4 * 0.1) = 64 * (0.58) = 37.1 fF. The shorter WL wires improved the resistance much more than the capacitance.

The new delay is t_{dWL2} = 2/3*2.37 = 1.6 ps. The new energy E_{WL2} is 37 pJ. So making the memory smaller more than halved the delay, but the energy required is almost the same.

3. The relative delay is $g_n*h_n + p$ for all four stages. Here we just have inverters so g = 1 in all cases. Thus, the relative delay is $h_n + p$ for all stages. The factor h is the fanout factor, which is the ratio between the driven capacitance at the output and the gate capacitance. So we have $h_1 = f_1$, $h_2 = f_2$, $h_3 = f_3$, and $h_4 = x/f_1f_2f_3$.

The total relative delay is thus:

$$d = 4p + f_1 + f_2 + f_3 + \frac{x}{f_1 f_2 f_3}$$

The total relative delay is symmetrical in f_1, f_2, and f_3 so if we find a solution for one of the tapering factors it must hold for the others as well. Thus, we could rewrite the delay with just one tapering factor called f and find the optimal solution for that one:

$$d = 4p + 3f + \frac{x}{f^3}$$

The derivate of d with respect to f is:

$$\frac{dd}{df} = 3 - \frac{3x}{f^4}$$

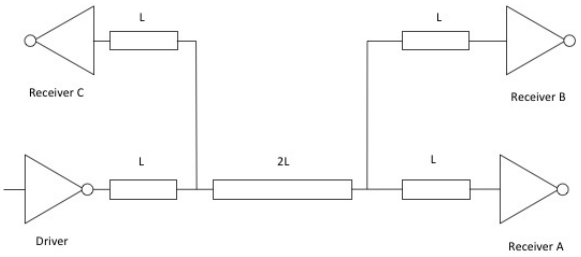
If we set this derivate equal to zero we find the solution

$$f = \sqrt[4]{x}$$

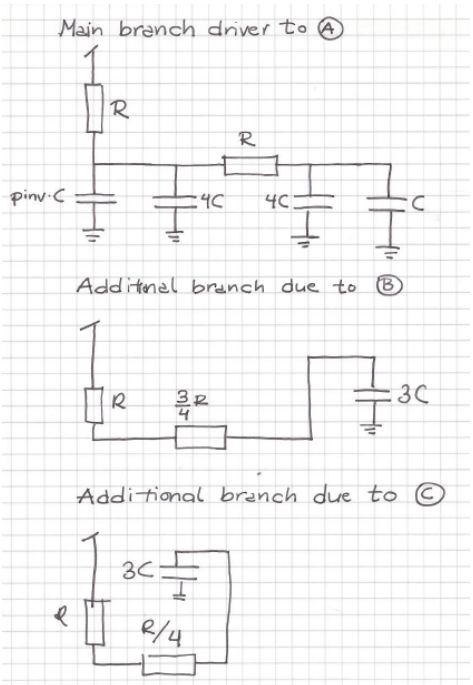
With that value of f we find that the total relative delay with optimal tapering factors is:

$$d = 4p + 4\sqrt[4]{x}$$

4. The setup is repeated below:



a)



The main branch from Driver to A has the following expression:

$$t_{dA} = R \cdot (p_{inv}C + 4C) + 2R \cdot 5C = (p_{inv} + 14)RC$$

The branch to B adds this to the delay:

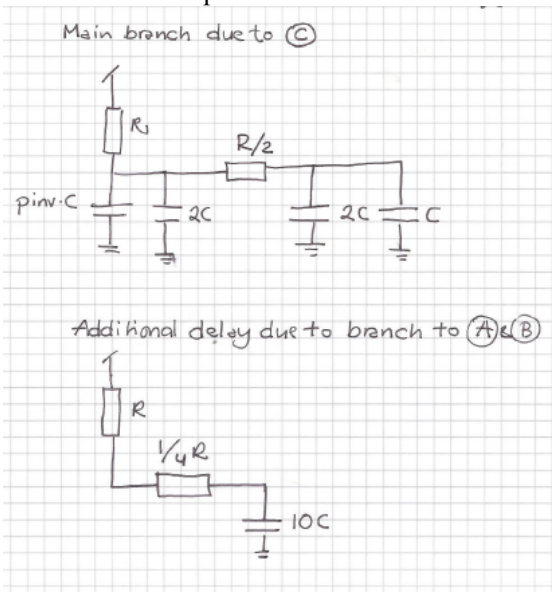
$$t_{dB} = (R + \frac{3}{4}R) \cdot 3C = \frac{21}{4}RC$$

The branch to C adds this to the delay:

$$t_{dC} = (R + \frac{1}{4}R) \cdot 3C = \frac{15}{4}RC$$

The total delay from the driver to receiver A is then **(23 + pinv) RC.**

b) The delay from Driver to Receiver B is the same as the one for A. The delay from Driver to C can be divided into two parts:



The main branch gives this delay:

$$t_{dC} = R \cdot (p_{inv}C + 2C) + (R + \frac{R}{2}) \cdot 3C = (p_{inv} + \frac{13}{2})RC$$

The additional delay due to the AB branch is:

$$t_{dAB} = (R + \frac{R}{4})10C = \frac{25}{2}RC$$

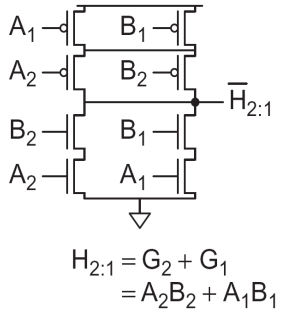
The total delay from the driver to C is then **(19 + pinv) RC**.

The clock skew is the difference in delay from the driver to B and C. It is **4RC.**

c) We can attach both the B and C branches at the middle of the 2L wire segment to achieve zero clock skew between B and C while maintaining the same delay from the driver to receiver A as before.

5.

a) Se figure below from Weste and Harris.



$H_{2:1} = G_2 + G_1$
$= A_2B_2 + A_1B_1$

b) Each possible branch has two transistors in series, so all pMOS transistors have to scale to width 4 and all nMOS transistors have to scale to width 2. Each of the four inputs is connected to one pMOS and one nMOS transistor. Thus, we have the same logical effort for all inputs. It is g = 6/3 = 2. To the output we have a width of 4+4+2+2=12 connected. The reference inverter has a width of 3 connected to the output. So we have p =12/3 pinv = 4 pinv.

c) Circuit diagram is repeated below. All pMOS transistors have to scaled to width = 4. When there are three nMOS transistors in series they all have width = 3, the other two nMOS transistors are scaled to width = 2.

Inputs A1 and B1 have the same logical effort. It is (4+3)/3 = 7/3. Inputs A2 and B2 have the same logical effort. It is (4+4+3+2)/3 = 13/3.

To the output we have a total width of 4+4+4+2+3=17 connected. Thus, we have p = 17/3 pinv.

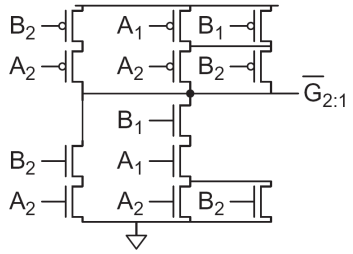


Figure 1: The 2-input generate gate from Weste & Harris figure 11.36 (a).

d) From our calculations in b) and c) it seems like it should always be beneficial to use a Ling adder rather than a regular prefix adder. The delay at the start of critical path should be slightly shorter both due to less parasitics (lower p) and due to simpler circuitry (lower g). However, one would have to investigate in more detail how much parasitics the additional XOR gate adds in the summation network even though that XOR gate is not in the critical path.

e) **(Bonus question)** Input A3 is the input for both the two parallel pMOS transistors connected to the output while B3 is missing in that part.

6. Layout of H2:1 cell with the diagram as the one shown in the solution for 5a, except that in the n-net the A2 transistor is above the B2 one.