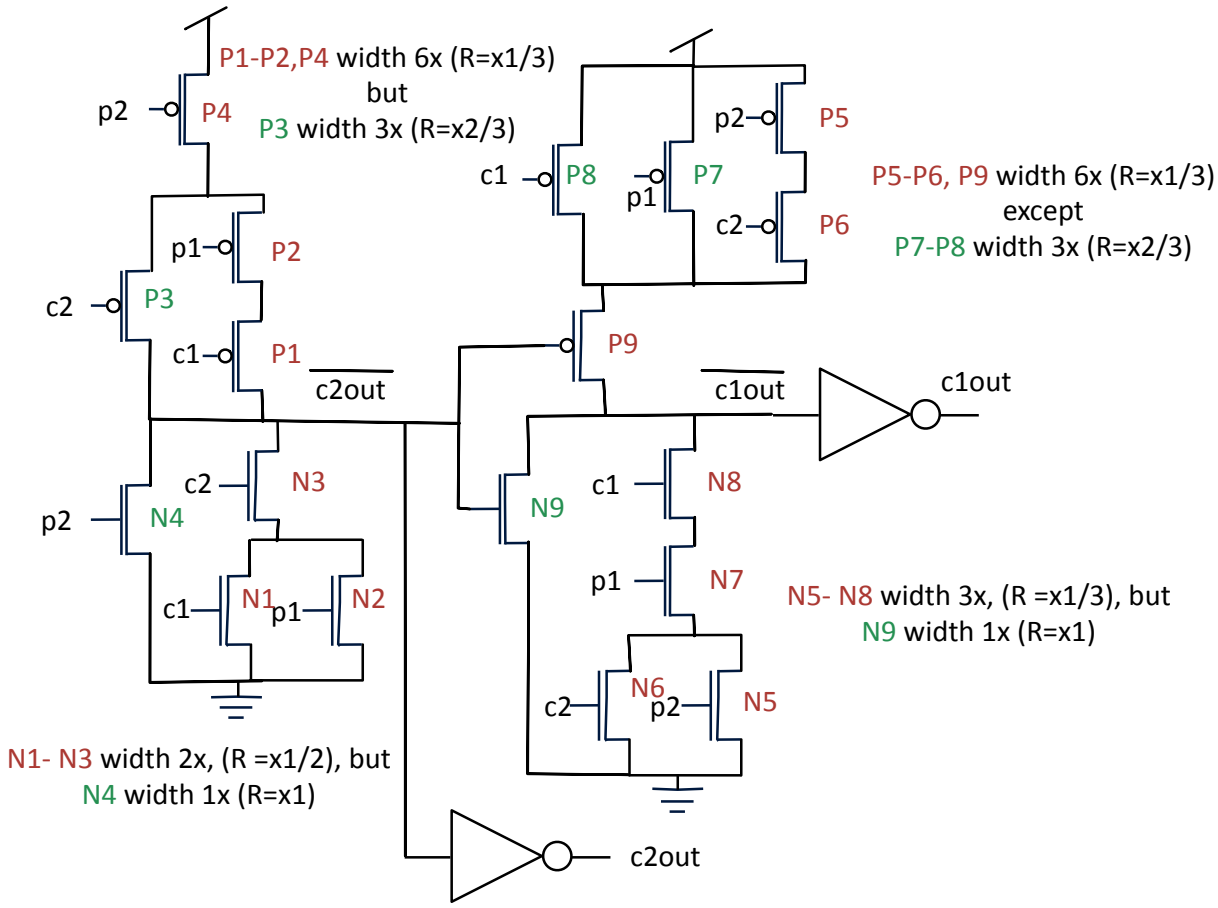
Solution to written exam in **Integrated Circuit Design MCC091**

Tuesday January 5, 2016, at 8.30-13.30 at Lecture halls, Hörsalsvägen

1) Logical effort, parasitic delay, layout

a) To compute the logical effort we first make the RC time constant the same for all paths. We then find the logical effort as the ratio of RC time constants from the complex gate and the reference inverter. If we make the worst-case **resistance** for all paths the same as that of the reference inverter, then we can take the ratio of only the input capacitances. In the layout supplied **all** transistors were made wider to make the resistance lower for transistors in series. Solution is given below with the **five** transistors that can be made narrower marked in green. The transistors are: P3 width 3x (rather than 6x), P8 and P9 width 3x (rather than 6x), N4 width 1x (rather than 2x), N9 width 1x (rather than 3x).



b) The c1 input of the first gate has the original transistor sizes from the layout: 6x width for the pMOS transistor and 2x width for the nMOS transistor. So for the first gate $g_{c1}=\frac{6+2}{2+1}=\frac{8}{3}$. The parasitic delay, for that gate, p1, is then $\frac{3+6+1+2}{2+1}=\frac{12}{3}=4$. For the second gate the $\overline{c2out}$ input is connected to a 1x width nMOS transistor and a 6x width pMOS transistor; thus we have $g_{\overline{c2out}}=\frac{6+1}{2+1}=\frac{7}{3}$. For the parasitic delay, for the second gate, p2, we get $\frac{6+1+3}{2+1}=\frac{10}{3}$.

c) In the worst case, the resistances of the two gates are the same as that for the reference inverter, so we can proceed by taking the ratio of the capacitances only as we did in task b). Now, some transistors are wider in the layout than in the optimized schematic. On the other hand, in the layout we have used shared diffusion areas as much as possible. So we find that the output of the first gate is connected to

two shared diffusion areas only: one of width 2 for the n-net and one of width 6 for the p-net. Thus, we get $\mathbf{p1} = \frac{6+2}{2+1} = \frac{8}{3}$. For the second gate we have one shared area of width 3x for the n-net and one non-shared area of width 6 for the p-net, thus $\mathbf{p2} = \frac{6+3}{2+1} = \frac{9}{3} = \mathbf{3}$. So for both gates the parasitic delay of from the non-optimized layout is lower than the one calculated in b) due the use of shared diffusion areas.

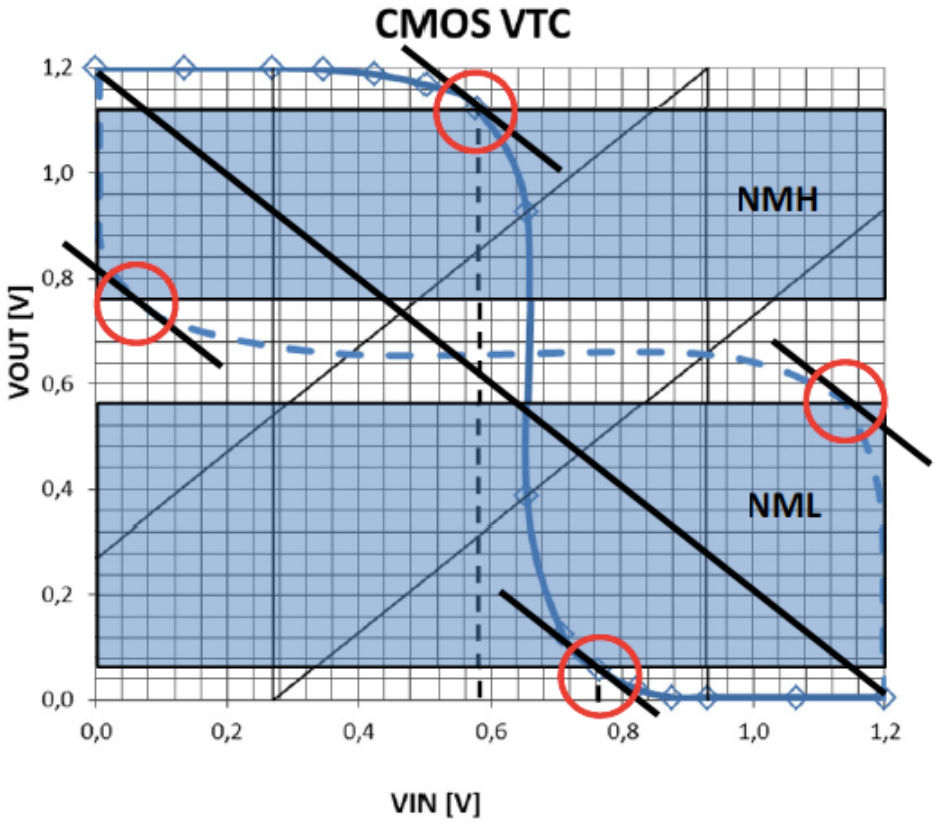
2) Noise margins

a) Noise margins indicate how large voltage deviations are tolerated on the input without the output voltage entering the "indeterminate" region. There are usually two margins: one for the low input voltage and one for a high input voltage. They are not necessarily the same.

b) High noise margins means that the circuit better withstands potentially corrupting noise. It is getting increasing important to do so, because there are lots of disturbances that can corrupt the digital signals in a modern chip and because the supply voltages are getting lower and lower as feature sizes shrink.

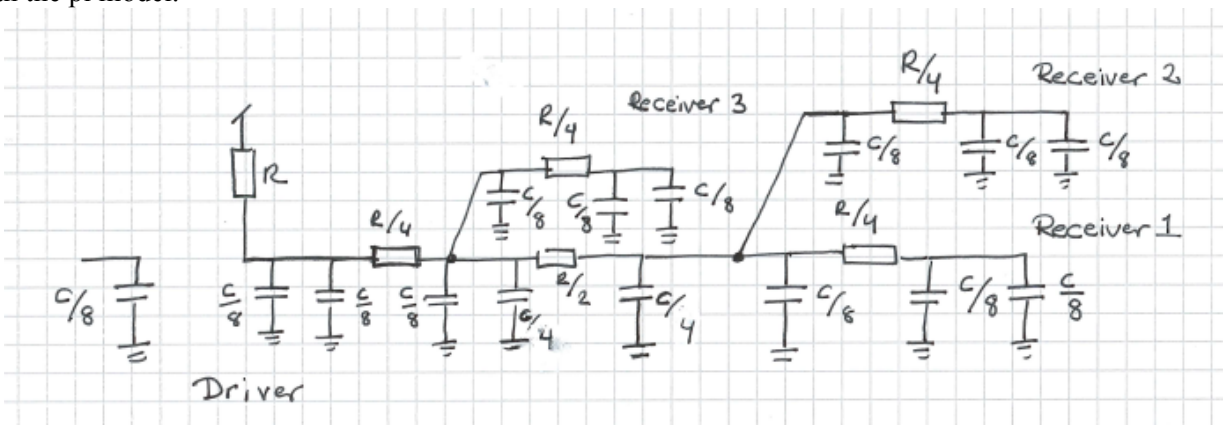
c) VOH,min=1.12 V, VIL,max=0.58 V; VOH,min =0.76 V; VOL,max =0.06 V.
NML= VIL,max- VOL,max = 0.58-0.06= 0.52 V; NMH = VOH,min- VOH,min=1.12-0.76=0.36V
Solid black lines indicate voltage gain $A_V = -1$.

d)



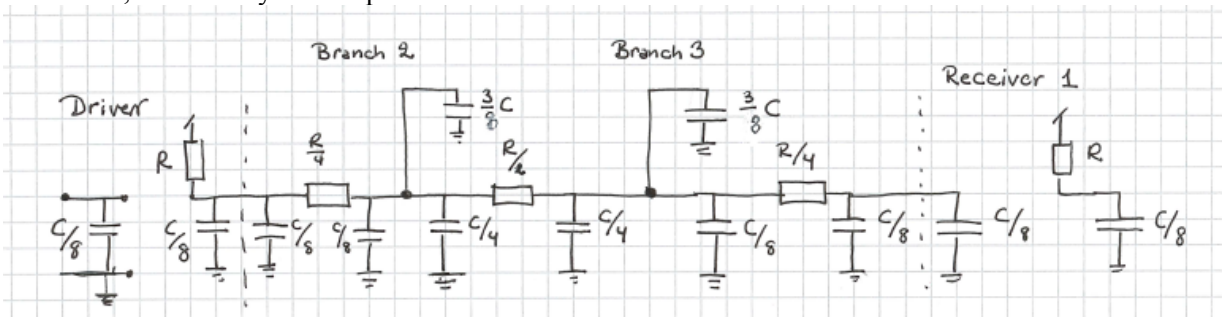
(6 p)

3) Wire and inverter delay The figure below shows the model for the entire circuit with all wires modeled with the pi model.

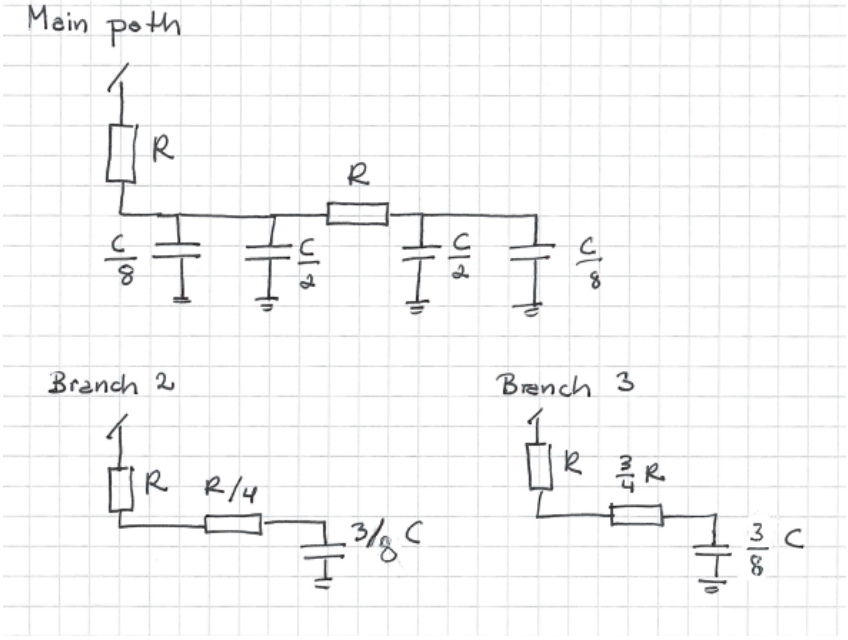


.

a) The figure is shown below. The Elmore branch model is used for the branches, where the resistances are zeroed, so that only their capacitances are included for the branches.



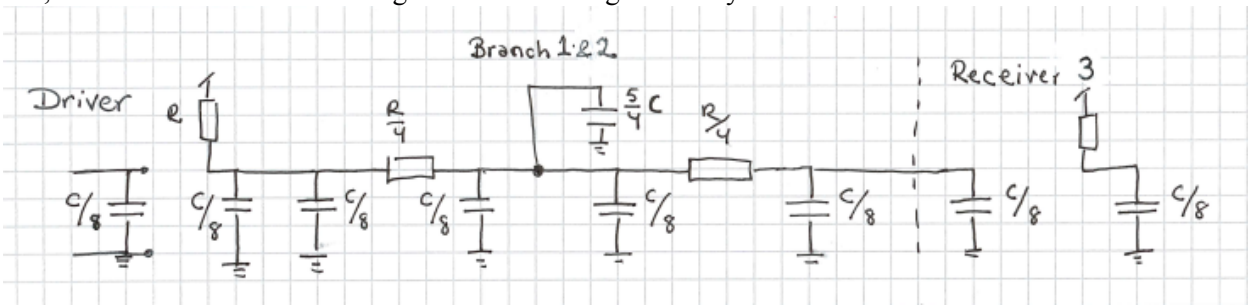
b) We calculate the delay for the main path and the side branches separately. See figure below:



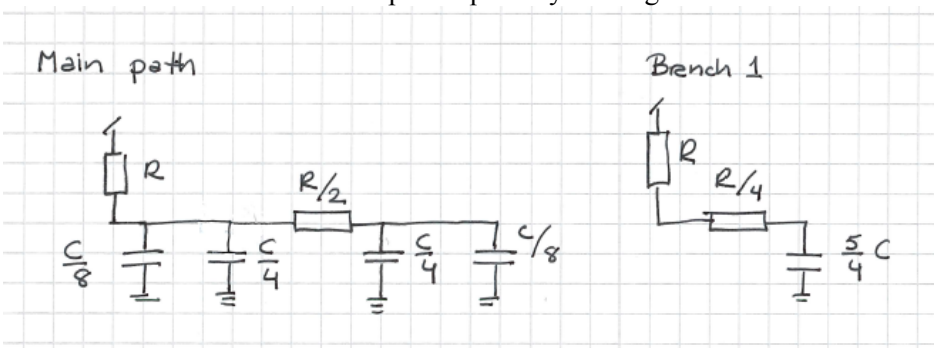
For the main branch we merge the four wire segments and then we find: $mainRC1 = R\left(\frac{C}{8} + C + \frac{C}{8}\right) + R(\frac{C}{2} + \frac{C}{8}) = \frac{15}{8}RC$. For the receiver 2 branch we find: $r2RC = \left(R + \frac{R}{4}\right)\frac{3}{8}C = \frac{15}{32}RC$. For the receiver 3

branch we similarly find *r3RC* = = $\left(R+\frac{3}{4}R\right)\frac{3}{8}C=\frac{21}{32}RC$. All in all we find the delay as *dr1* =0.7(*mainRC1*+*r2RC*+*r3RC*) = $0.7RC\left(\frac{15}{8}+\frac{15}{32}+\frac{21}{32}\right)=0.7RC\left(\frac{15}{8}+\frac{9}{8}\right)=\mathbf{0.7\times 3}\boldsymbol{RC}$.

c) The delay to the input of receiver 3, *dr3*, is the same as that for receiver 3. But the one for receiver 2, *dr2*, is not the same. Below is a figure for calculating that delay:



Also here we calculate the main path separately. See figure below:



The RC product for main path is mainrc2 = $R\left(\frac{C}{8}+\frac{C}{2}+\frac{C}{8}\right)+\frac{R}{2}\left(\frac{C}{4}+\frac{C}{8}\right)=R\frac{15}{16}C=\frac{15}{16}RC$. The branch adds the RC product, r3RC = $\frac{5}{4}R\frac{5}{4}C=\frac{25}{16}RC$. All in all we have dr3= 0.7(*mainRC2*+*r3RC*) = $\mathbf{0.7\times 2.5}\boldsymbol{RC}$. So the clock skew is **0** between the inputs of receivers 1 & 3 and for 1&2 and 3&2 is it $\mathbf{0.7\times 0.5}\boldsymbol{RC}$**.**

d) It will not change since changing the load at the output of an inverter does not change its input capacitance.

4) Path delay with wires, inverter sizing (See also example 5.11, page 195 in Weste and Harris.)

a) Since the logical efforts, *g*, are 1 for inverters, the stage effort is equal to the electrical effort, *h*, and we get *d*= *h1*+ *h3* + *h3*. The path delay expressed in terms of y and z is then:
$d=\frac{yC}{C}+\frac{(z+5)C}{yC}+\frac{10C}{zC}=\boldsymbol{y}+\frac{\boldsymbol{z+5}}{\boldsymbol{y}}+\frac{\mathbf{10}}{\boldsymbol{z}}$**.**
Note: The path delay comprises the stage efforts only, not the constant part of the delay (the p part).

b) To find the values for y and z we take the derivative of the expression for *d* from task a) wrt the two unknowns and set the derivatives to 0:

$$\frac{\partial D}{\partial y}=0=1-\frac{z+5}{y^2}\Rightarrow y^2=z+5$$
$$\frac{\partial D}{\partial z}=0=\frac{1}{y}-\frac{10}{z^2}\Rightarrow z^2=10y$$

The solutions to these equations are $\boldsymbol{y}\approx$ **3.3 and** $\boldsymbol{z}\approx$ **5.7** (The solution can rather easily be found by trial and error).

c) The total normalized delay is $d+\sum p$ = 3.3 + 10.7/3.3 + 10/5.7 + 3 = 11.3 if we assume p = 1 for the inverters. In this process 0.7RC is 5 ps, so the delay $\mathbf{t_d}$ **= 11.3*5ps = 56 ps.**

5) Scaling of delay and power

a) The FO4 delay is $5\times0.7RC$ when the inverter output capacitance is equal to its input capacitance. *C* is the transistor input capacitance, which is the transistors' gate capacitances: $C_G = W\times L\times Cox$. The equivalent resistance of the transistors is $R = \frac{V_{DD}}{I_{DSAT}}$. As hinted in Table 7.4 we have the maximum saturations current $I_{DSAT} = \mu Cox\frac{W}{L}(V_{DD} - Vt)^2$. So we find $RC = \frac{V_{DD}L^2}{\mu Cox(V_{DD}-Vt)^2}$. (As expected RC does not depend on the transistor width). KPn is the same as μCox. So with all the numbers from the problem inserted we get FO4 delay for 35 um process = **$5\times0.7RC$ = 43.7 ps**.

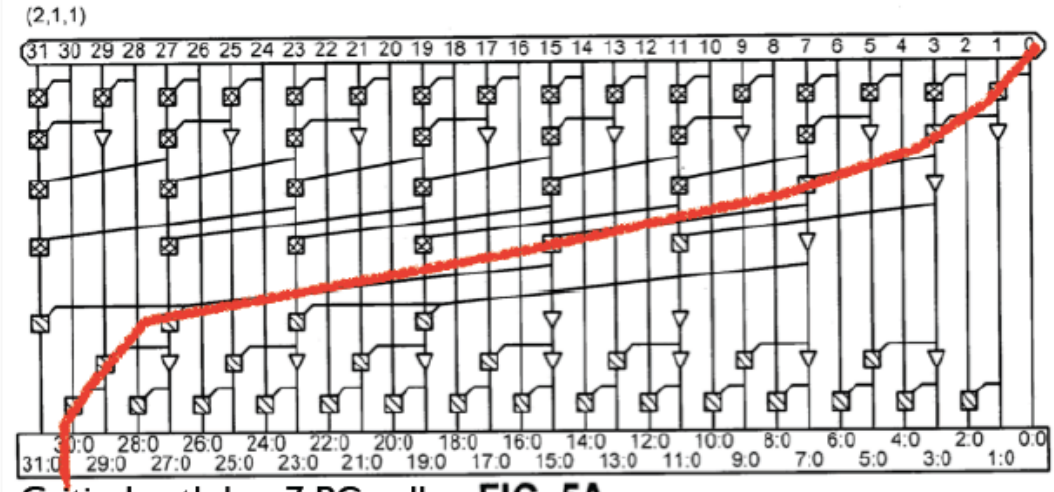
b) The scaling factor, S, between the two processes is 0.75/0.35 = 2.15 (which is a bit strange; usually the nominal scaling is $\sqrt[2]{2}$ between each process generation. Here it is a little more than that). Both processes use the same supply voltage, 3.3 V, so we have constant-voltage scaling, column 2 in Table 7.4. From this table we see that the *R* scales as 1/*S* and C also scales as 1/*S*. Thus, *RC* scales as $1/S^2$. Consequently, the FO4 delay in the 0.75 um process should be 2.15^2= 4.6 times that of the 0.35 um process, which is 201 ps. **Let's say, FO4 delay for 75 um: 200 ps**. A clock frequency of 200 MHz corresponds to a period time of 5 ns, that is 5000 ps. That corresponds to **25 FO4 delays** maximum delay in the Alpha processor logic. (In practice it would have to be a bit shorter than that due to setup times in the flip flops etc.)

c) The dynamic power consumption is proportional to fCV_{DD}^2. Both processes have the same supply voltage, so any scaling of the dynamic power are due to the scaling of clock frequency, f, and the charged and discharged capacitances, C. As we saw in task b), if the clock frequency is scaled as the FO4 delay it scales as S^2, while the transistor capacitances are scaled down with 1/S. So, assuming that all capacitances are due to transistors, and none due to the interconnect, the result is that the dynamic power scales with S; so for the 35 um process we then have **P_{dyn} = 2.15 *30W = 64.5 W.**

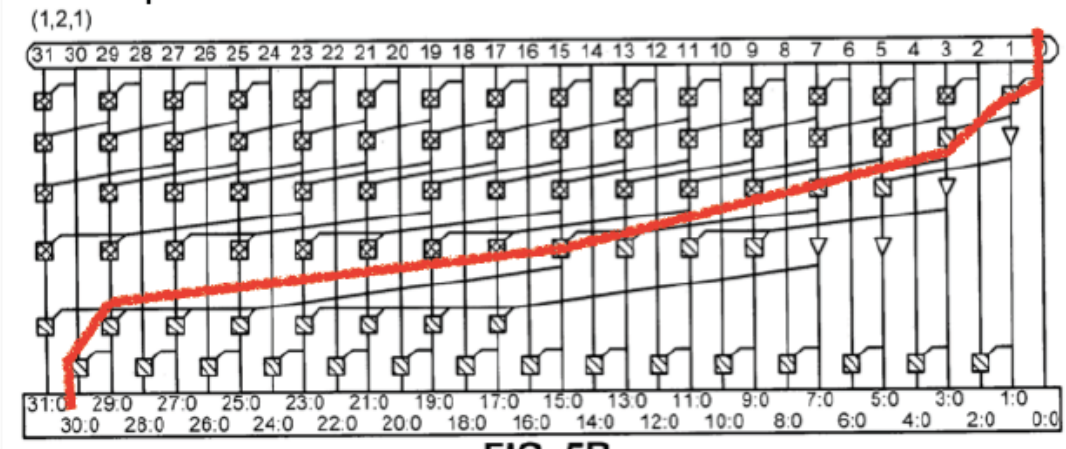
d) If V_{DD} were also scaled down with S, the dynamic power would instead also be scaled 1/S (with the same assumptions as in task c)). The result is then **P_{dyn} = 30W/2.15 = 14 W**. From this example, it is clear why constant voltage scaling was abandonded and it was necessary to start scaling down the supply voltages as the feature sizes were scaled down. (There were also other reasons of course).

6) Prefix adders

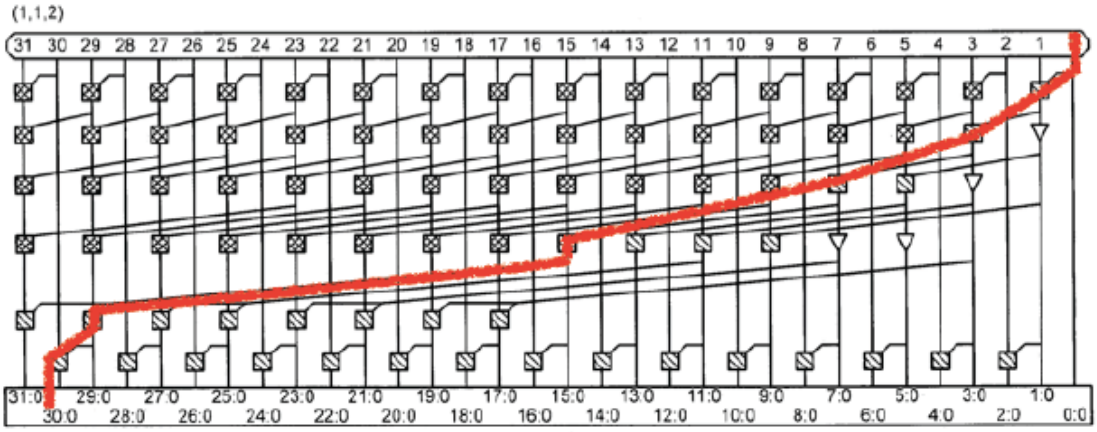
a) . Note. There are more paths that have the same lengths and fanout. See the answer to task d) for one example.



Critical path has 7 PG cells **FIG. 5A**



Critical path has 6 PG cells **FIG. 5B**



Critical path has 6 PG cells **FIG. 5C**

b) 5A: Pro: Lower fanout for all PG cells & few wires on each level: Con: Highest number of PG cells in critical path.
5B: Pro: Few wires on each level and lower number of PG cells. Con: High fanout.

5C: Pro: Lower number of PG cells and lowest fanout. Con: Many wires have to cross on some levels.

c) There are 6 cells in the path. Disregarding the buffers drawn in the original diagram we find, as is shown in the figure below:

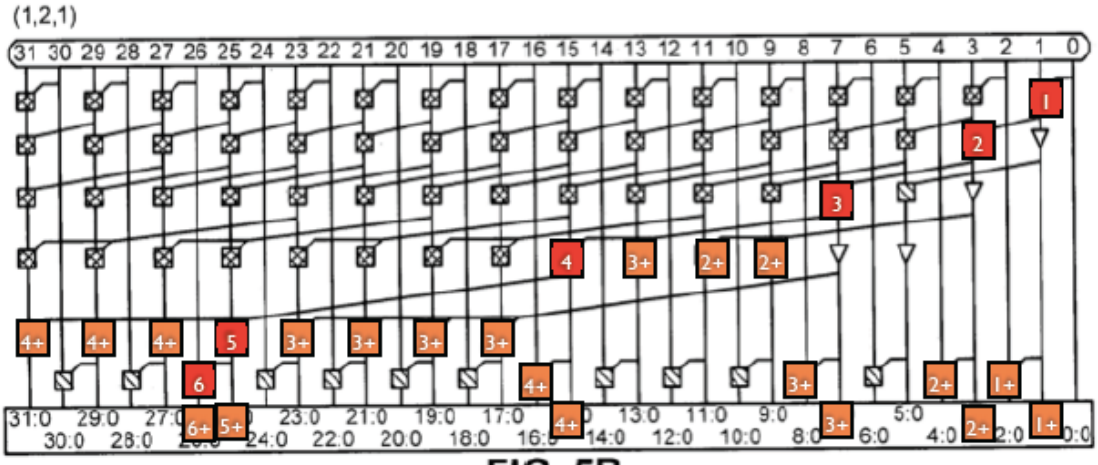


FIG. 5B

PG cell 1: fanout is 2 PG cells + 1 sum cell = 3 cells
PG cell 2: fanout is 4 PG cells + 1 sum cell = 5 cells
PG cell 3: fanout is 7 PG cells + 1 sum cell = 8 cells
PG cell 4: fanout is 4 PG cells + 1 sum cell = 5 cells
PG cell 5: fanout is 1 PG cells + 1 sum cell = 2 cells
PG cell 6: fanout is 1 sum cell = 1 cell.

To calculate the path delay we have g = 2 for all cells. So the path delay is then $d = g\sum_{i=1}^{6} h_i$.= 48. Note that the path delay does not include the constant part (p part) of the delay.

d) **BONUS QUESTION** A distinct path is one that differs in at least one PG cell from the others. All paths of length 7 end up at one of sum bit 30, 26 and 22. The 3 last PG cells are unique for each of these three sum cells. So it easiest to start from the end and mark the tree that leads to those cells. On close inspection we find 16 paths for sum bit 30 and 8 each for sum bit 26 and 22; so all in all 32 paths with7 PG cells. The three trees are drawn separately on the next page to make them clear to see.

(2,1,1) 8 paths of length 7 PG cell end at sum cell 22

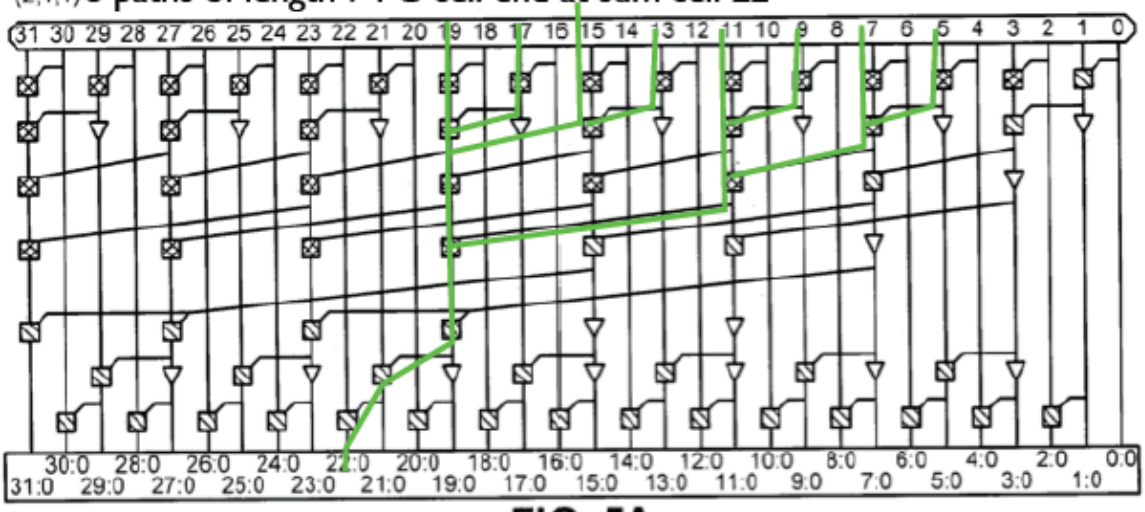


FIG. 5A

(2,1,1) 8 paths of length 7 PG cell end at sum cell 26

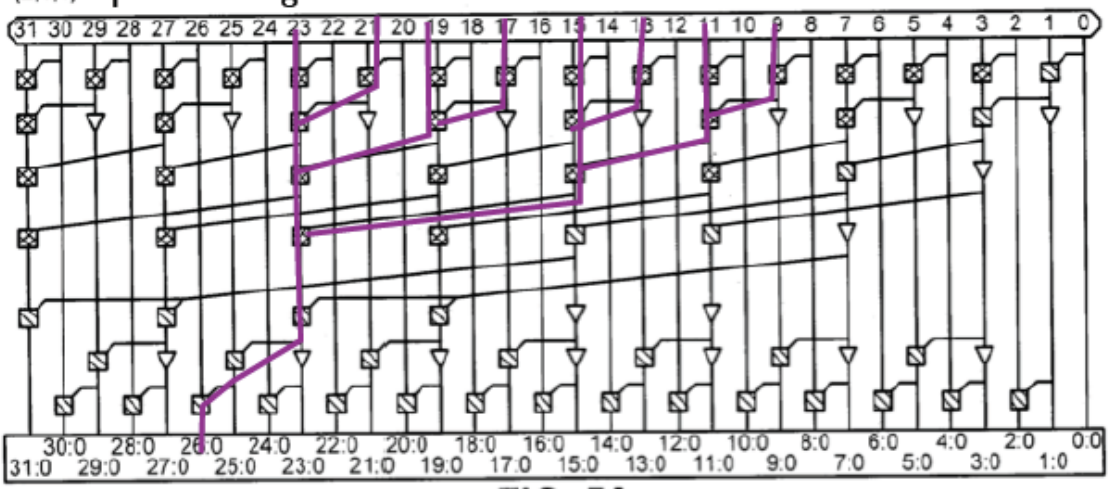


FIG. 5A

16 paths of length 7 PG cells end at sum cell 30
Note that the two sub trees overlap for inputs 13 and 15!

(2,1,1)

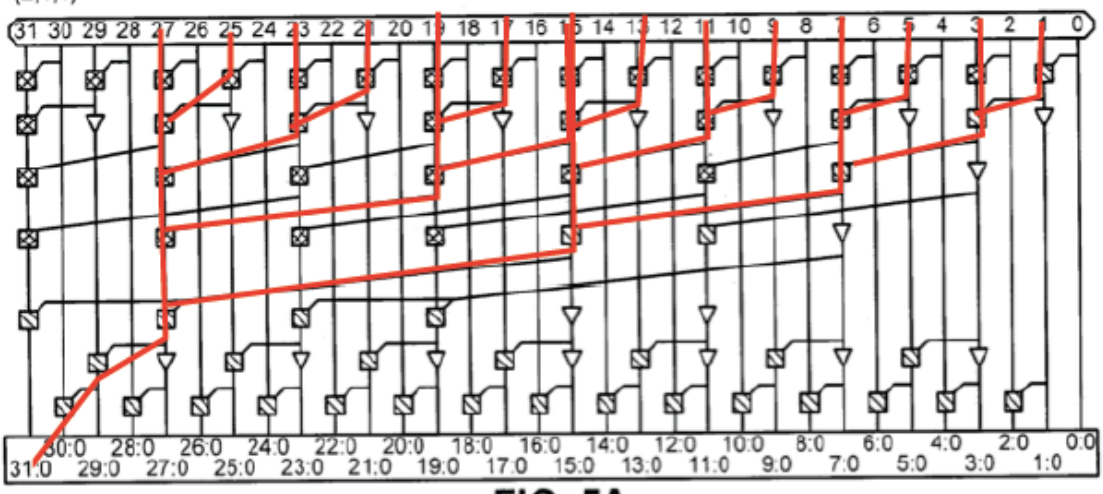


FIG. 5A