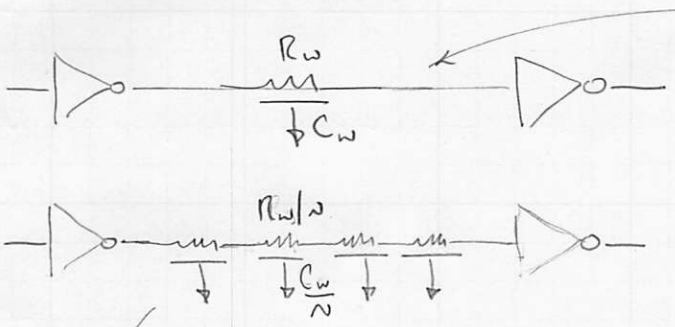
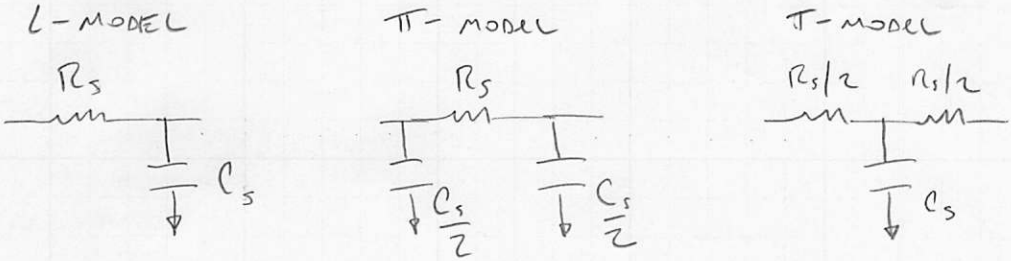
CMOS INTERCONNECT



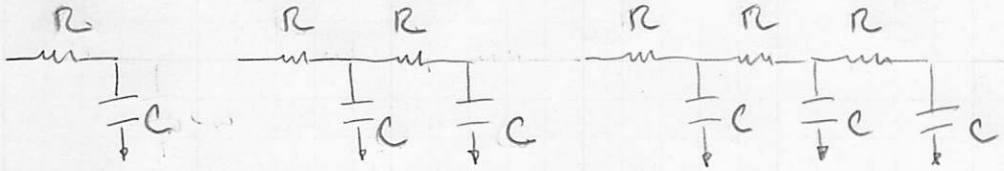
RESISTANCE ↑ WITH ℓ
CAPACITANCE ↑ WITH ℓ
RC TIME CONST ↑ WITH ℓ^2

THINK OF WIRE DIVIDED INTO MANY SMALL SEGMENTS EACH WITH AN EQUAL PORTION OF THE TOTAL WIRE RESISTANCE AND CAPACITANCE. THE LARGER R THE MORE ACCURATE THE RC MODEL

VARIOUS RC MODELS ARE POSSIBLE FOR EACH SEGMENT



ASSUME WE HAVE A WIRE OF LENGTH ℓ WHICH WE MODEL WITH A SINGLE SEGMENT USING THE L-MODEL. HOW DOES THE DELAY VARY AS WE LOOK AT LONGER WIRES WITH LENGTH 2ℓ, 3ℓ, $M \times \ell$



$t_{PD} = RC$ $\quad t_{PD} = RC + 2RC$ $\quad t_{PD} = RC + 2RC + 3RC$

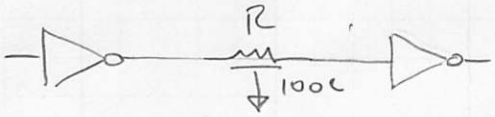
FOR WIRE OF LENGTH $M \times \ell$, $t_{PD} = RC + 2RC + 3RC + \ldots + MRC$

WHICH IS $O(M^2)$, IN OTHER WORDS THE PROPAGATION DELAY SCALES QUADRATICALLY WITH LENGTH WHEN WE DRIVE A LONG WIRE WITH A SINGLE RECEIVER

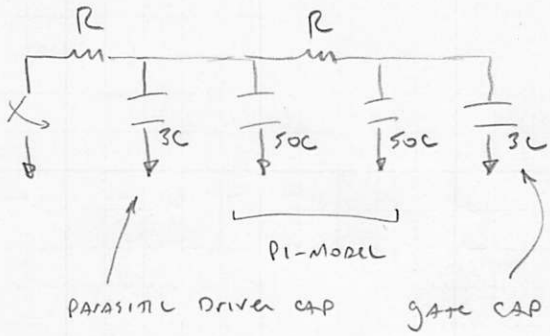
Let's look at three examples

1. Single small driver with no repeaters
2. With repeaters
3. Single larger driver with no repeaters

Single small driver



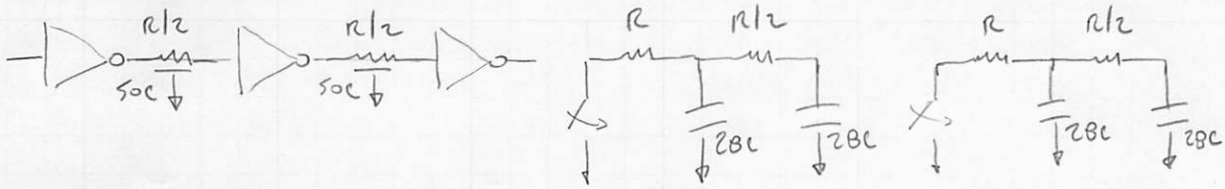
Assume total wire resistance for this length ≈ effective resistance of MW sized NMOS



$$t_{PD} = R \cdot 53C + 2R \cdot 53C = 159RC = \underline{53\uparrow}$$

With repeater

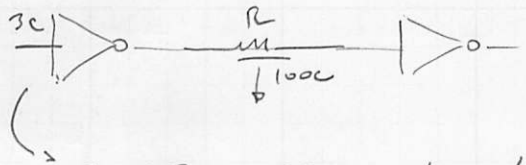
Break long wire into two parts by adding a second inverter in the middle of the wire.



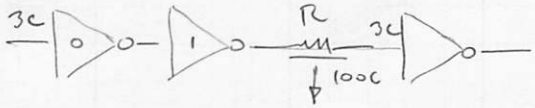
$$t_{PD} = 2\left(28RC + \left(R + \frac{R}{2}\right)28C\right) = 2\left(28RC + 42RC\right) = 140RC = \underline{46.7\uparrow}$$

So adding one repeater decreases the overall delay.

WITH LARGE DRIVER



$F = GBH = 33.3 \quad \log_4(33.3) \approx 2.5$ ← use two stages instead of 3 to save area / energy

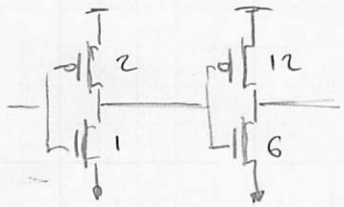


use logical effort to size second stage

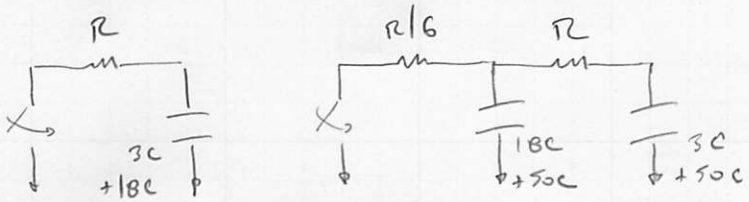
$$\hat{f} = \sqrt{F} = 5.77 \qquad C_{in} = \frac{g_i}{\hat{f}} C_{out}$$

$$C_{in,1} = \frac{1}{5.77} 103 = 17.4\,C$$

$$C_{in,0} = \frac{1}{5.77} 17.4 = 3.0\,C$$



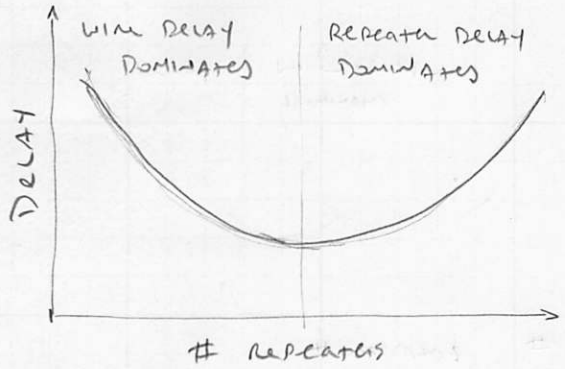
nearest even multiple of minimum sized inverter

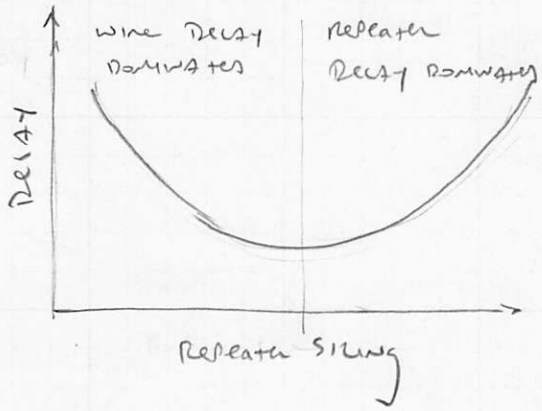


$$t_{pd} = 21RC + \frac{R}{6} 6BC + \left(\frac{R}{6} + R\right) 53C = 94RC = \underline{31.4/\tau}$$

so more optimally sizing the driver can also reduce overall delay

There will be an optimal number of repeaters and an optimal size of those repeaters which balances wire delay with repeater delay



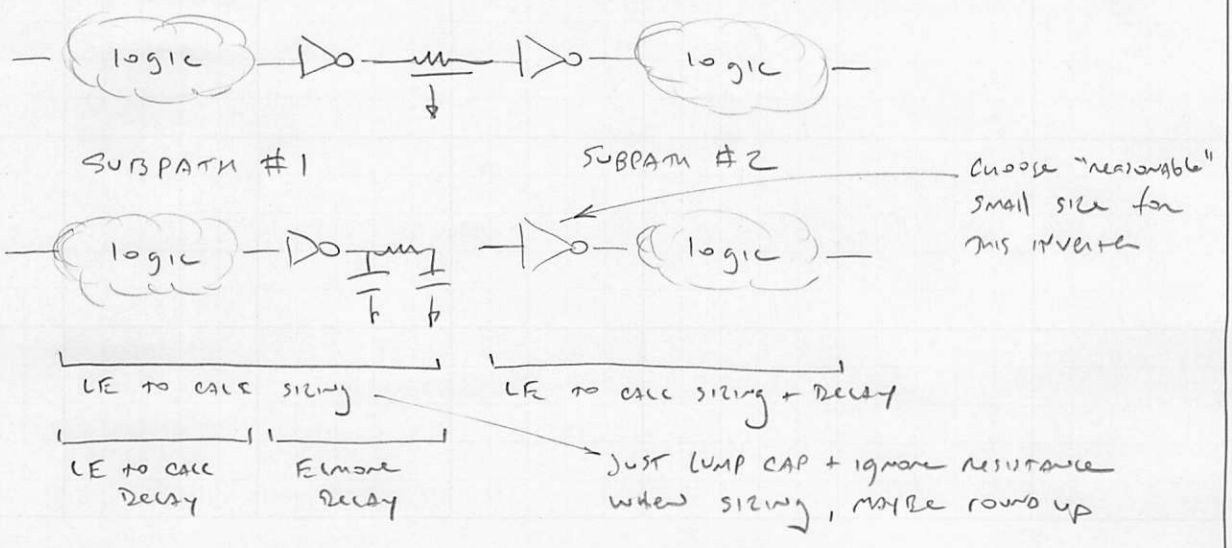


Of course the delay is a function of both the number and size of the repeaters so this is a multi dimensional optimization problem

Logical Effort and Interconnect

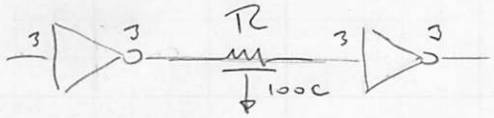
Logical effort does not work well with large fixed cap on intermediate nodes - which is exactly what interconnect can insert into our models.

To deal with this we can break a path into two subpaths with the first path ending at the large interconnect load and the second starting with the receiver. This is reasonable as long as the interconnect load is much larger than the receiver gate cap.



INTERCONNECT ENERGY

We can use a similar approach as before. Estimate total switches capacitance and activity factor



TOTAL SWITCHES CAP = 112c

$$E_{\substack{BIT \\ TRANSITION}} = \frac{1}{2} C V_{DD}^2 = \frac{1}{2} (112c)\left(\frac{0.5fF}{c}\right)(1V)^2 = 28\,fJ/bit$$

Note that this assumes input is always toggling. Usually when you see energy/bit in the literature they assume a random input signal.

0 → 0	0.25
0 → 1	0.25
1 → 0	0.25
1 → 1	0.25

} Bit transitions only occur 50% of time

So with random data $\alpha = 0.5$

$$E_{BIT} = \alpha \frac{1}{2} C V_{DD}^2 = 14\,fJ/bit$$

If frequency is 500 MHz

$$P_{BIT} = \alpha f \frac{1}{2} C V_{DD}^2 = (0.5 \times 10^9)(14 \times 10^{-15}) = 7\,\mu W$$

Sometimes energy is reported in units of mW/Gb/s

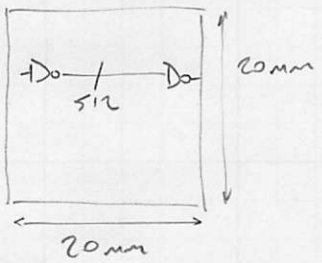
$$\frac{mW}{Gb/s} = \frac{10^{-3}\,W}{10^9\,b} \cdot s = 10^{-12}\,\frac{W}{b} \cdot s = 10^{-12}\,\frac{J/s}{b} \cdot s = 10^{-12}\,J/b$$

In other words 1 pJ/b = 1 mW/Gb/s

So in above example

$$E_{BIT} = 14\,fJ/bit = 0.014\,mW/Gb/s$$

Let's assume we have a wide 512b global bus crossing the entire length of a large 20mm x 20mm die in 65nm



Let's ignore the energy overhead of repeaters. Assume 0.2 pF/mm wire cap and random data

$$E_{BIT} = \alpha \frac{1}{2} C V_{DD}^2 = 0.5 \frac{1}{2} (20\text{mm} \times 0.2\text{pF/mm})(1\text{V}^2) = 1\text{pJ/bit}$$

w/ 512 wires $\boxed{E_{BUS} = 512 \times E_{BIT} = 512\text{pJ}}$

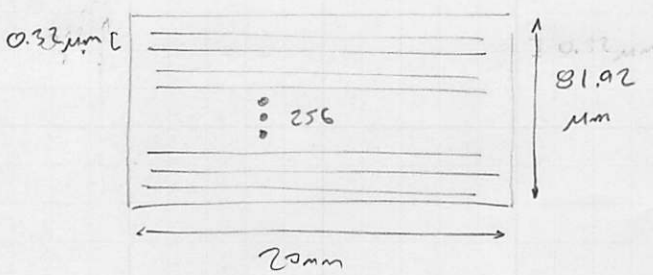
$$\boxed{P_{BUS} = f \cdot E_{BIT} = (0.5 \times 10^9)(512 \times 10^{-12}) = 256\text{mW}}$$

AREA

Unlike with CMOS logic/state where we often estimate area just by using the total million of gate width, often we can hide the logic "under" the wires in interconnect dominated units.

In these cases we can estimate the area just from the wire pitch and potentially information on the number of metal layers

For example if we assume the above global bus is wire dominate we can hide the repeaters under the wires. Assume the bus is mapped to two metal layers, and that the minimum wire pitch is 0.32 µm.



Total area is just

$256 \times 0.32\mu\text{m} \times 20\text{mm}$

$= 82\mu\text{m} \times 20\text{mm}$