Interconnect Delay Modeling

Delay of a single long wire in an integrated circuit is proportional to the square of the length because both the resistance and capacitance increase linearly with length. By breaking the wire into multiple stages and placing repeaters at the start of each stage, the delay can be made linear with length. This study explores the optimal sizing of buffers and number of repeaters to minimize the delay. It computes delay (ps/µm) as a function of process parameters at the optimal sizing.

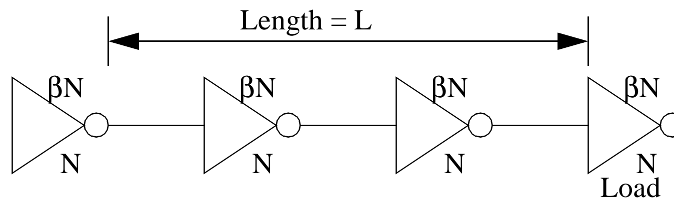
First order analysis

We begin with several simplifying assumptions to make the analysis cleaner.

- Source / drain diffusion capacitance is negligible
- Inverters are sized for equal rise and fall times (faster results may be possible with unequal times)
- Wire pitch and spacing are preset
- Propagation speeds are low enough that transmission line effects may be neglected
- Load capacitance equals repeater capacitance

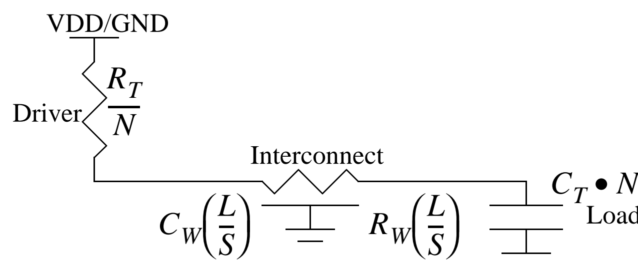
The interconnect of total length L is divided into S segments. Each inverter has a pulldown N microns wide and a pullup βN microns wide to achieve equal rise and fall times. The figure below illustrates a wire divided into three segments:

FIGURE 1 Interconnect divided into 3 segments



The path can be broken into S identical segments for analysis. A model of one segment is shown below. Resistances are in ohms / micron; capacitances are in pF / micron.

FIGURE 2 Model of single segment



Observing that a distributed RC line has the same delay as a lumped RC element with half the resistance, we can derive the following expression for the delay through a segment. The delay through the entire interconnect is S times the delay through each segment:

$$t_{interconnect} = S\left[C_T N\left(\frac{R_T}{N} + R_W\left(\frac{L}{S}\right)\right) + C_W\left(\frac{L}{S}\right)\left(\frac{R_T}{N} + \frac{R_W}{2}\left(\frac{L}{S}\right)\right)\right] \quad \text{(EQ 1)}$$

$$\text{simplifying, } t_{interconnect} = SC_T R_T + L\left(C_T R_W N + \frac{C_W R_T}{N}\right) + L^2\left(\frac{C_W R_W}{2S}\right) \quad \text{(EQ 2)}$$

Take partial derivatives with respect to N and S to minimize total delay:

$$\frac{\partial t_{interconnect}}{\partial N} = C_T R_W L - \frac{C_W R_T L}{N^2} = 0 \Rightarrow N = \sqrt{\frac{C_W R_T}{C_T R_W}} \quad \text{(EQ 3)}$$

$$\frac{\partial t_{interconnect}}{\partial S} = C_T R_T - \frac{C_W R_W L^2}{2S^2} = 0 \Rightarrow S = L\sqrt{\frac{C_W R_W}{2C_T R_T}} \quad \text{(EQ 4)}$$

Hence, the total delay is:

$$t_{interconnect} = L\sqrt{\frac{C_W R_W C_T R_T}{2}} + L(\sqrt{C_W R_W C_T R_T} + \sqrt{C_W R_W C_T R_T}) + L\sqrt{\frac{C_W R_W C_T R_T}{2}} \quad \text{(EQ 5)}$$

$$= L\sqrt{C_W R_W C_T R_T}(2 + \sqrt{2}) \quad \text{(EQ 6)}$$

and the optimal length of each segment is:

$$\sqrt{\frac{2C_T R_T}{C_W R_W}} \quad \text{(EQ 7)}$$

Observations

Observe that the interconnect delay is proportional to the number of stages. As expected, the number of stages is also linear in the number of stages. The buffer size is independent of wire length; it is only a function of the physical parameters of wire and transistors. The components of the delay caused by wire resistance and by driver resistance are equal.

The relationships also make physical sense. As wire capacitance increases or driver resistance increases, a stronger driver is needed. As driver capacitance increases or wire resistance increases, a relatively smaller driver should is preferred. As the wire gets more resistive or capacitive, it should be broken into more stages; as the driver resistance and capacitance get worse, more interconnect between stages is better.

The total delay scales with the R_TC_T product. This product decreases as transistor geometries shrink; therefore, the delay to drive a wire of fixed length and fixed capacitance and resistance / unit length should decrease as processes shrink. In practice, however, we observe wire performance improving more slowly than transistor performance. This is at least partially due to increase in sidewall and cross-capacitance as a fraction of the entire capacitance as feature size shrinks.

Relation to physical parameters

Let us write the expressions for capacitances and resistances in terms of physical parameters. Assume that wire width and spacing are large enough that sidewall capacitance is a negligible fraction of the whole (this is not generally realistic for deep submicron processes). Therefore:

$$R_W = \frac{R_{square}}{W_{wire}} \quad \text{(EQ 8)}$$

$$C_W = \frac{\varepsilon_{ox}}{t_{fieldox}} \bullet W_{wire} \quad \text{(EQ 9)}$$

$$R_W C_W = \frac{R_{square}\varepsilon_{ox}}{t_{fieldox}} \quad \text{(EQ 10)}$$

Hence, the wire RC product can be reduced by reducing the square resistance of the interconnect (using thicker wires or different materials) and by increasing the thickness of the field oxide. A process could conceivably use a different dielectric than field oxide to reduce ε, but in practice it is unlikely.

Next, consider the device parameters. Use a P device scaled to have drive strength equal to the N. Approximate the transistor with a linear I-V relation. The capacitance includes both the N device and the P device.

$$R_T = \frac{t_{ox}L}{\mu_n \varepsilon_{ox}(V_{GS} - V_t)} \quad \text{(EQ 11)}$$

$$C_T = \frac{\varepsilon_{ox}L\left(1 + \frac{\mu_n}{\mu_p}\right)}{t_{ox}} \quad \text{(EQ 12)}$$

$$R_T C_T = \frac{L^2\left(\frac{1}{\mu_n} + \frac{1}{\mu_p}\right)}{(V_{GS} - V_t)} \quad \text{(EQ 13)}$$

Unfortunately, submicron devices are generally operating in velocity saturation. In such a case, shrinking channel length or increasing gate drive no longer improves performance. Instead, we can use the following model based on the E-field at saturation:

$$R_T = \frac{t_{ox}}{\mu_n \varepsilon_{ox} E_{sat}} \quad \text{(EQ 14)}$$

$$R_T C_T = \frac{L\left(\frac{1}{\mu_n} + \frac{1}{\mu_p}\right)}{E_{sat}} \quad \text{(EQ 15)}$$

Reasonable parameters for a current 0.4 µm process are as follows:

TABLE 1 Typical process parameters

Parameter	Value
R_{square}	$5 \bullet 10^{-2}$ Ω
ε_{ox}	$1.04 \bullet 10^{-4}$ pF/µm
$t_{fieldox}$	0.5 µm
W_{wire}	1 µm
R_W	$5 \bullet 10^{-2}$ Ω/µm
C_W	$2 \bullet 10^{-4}$ pF/µm
$R_W C_W$	10^{-5} ps/µm^2
t_{ox}	10^{-3} µm
C_T	$1.5 \bullet 10^{-2}$ pF/µm
R_T	2000 Ω-µm
$R_T C_T$	30 ps
$t_{interconnect}$	$5.43 \bullet 10^{-2}$ ps/µm
seg length	2500 µm
N	23 µm

Conclusion

The interconnect time of 50 ps / mm is somewhat optimistic and the segment lengths between repeaters are shorter than typically seen in current processor designs. This is largely due to neglecting diffusion capacitance of the driver; when the parasitics of the driver are included, the optimal segment length and total delay will both increase. As device performance continues to improve and wire characteristics remains constant, segment length and total interconnect delay will both shorten. Hence, methodologies must be developed to support easy insertion of repeaters at appropriate points along interconnect.