Buffer Repeaters

David Harris

1.0 Introduction

Repeaters are widely used to combat the quadratic delay of long on-chip wires. A conventional repeater is a CMOS inverter placed periodically along the long wire. This design has two drawbacks. The wire must have an even number of repeaters to preserve signal polarity. This forces the designer to sometimes use a suboptimal number of repeaters. It is also awkward when a wire branches because repeaters must be placed in such a way that all branches have an even number of repeaters. The second drawback is that many repeaters are required. For floorplanning reasons, repeaters tend to be grouped into “gas stations” which may not be immediately under the best wire route. Thus, longer wires are required to reach the gas stations.

To deal with these problems, some designers have constructed repeaters from pairs of inverters. This automatically solves the polarity problem. Moreover, it may lead to fewer repeaters along the wire, since each repeater has lower input capacitance and higher output drive. This reduces the number of gas stations and extra routing required. Unfortunately, buffers are inherently slower than repeaters. This document explores the use of buffers as repeaters. It comes to a remarkably elegant conclusion showing that “optimal” buffer repeaters” are only slightly slower than “optimal” inverter repeaters.

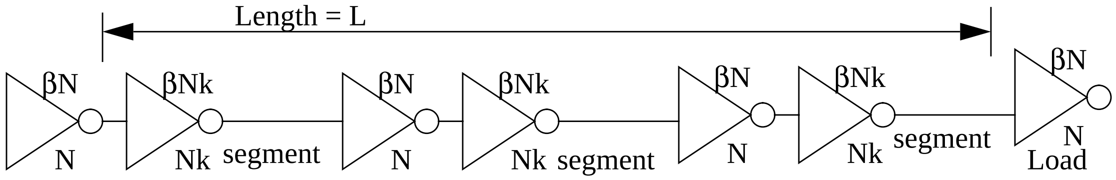
2.0 Model

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

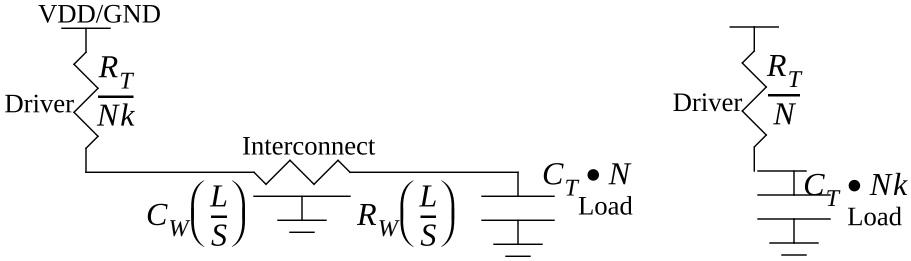
The interconnect of total length L is divided into S segments. Each inverter has a pulldown N microns wide and a pullup βN microns wide to achieve equal rise and fall times. The figure below illustrates a wire divided into three segments:

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



3.0 Delay

The delay of the interconnect is minimized by choosing the appropriate S, N, and k. Using the Elmore delay model, we find:

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

$$\text{simplifying, } t_{interconnect} = S\left(k + \frac{1}{k}\right)C_T R_T + L\left(C_T R_W N + \frac{C_W R_T}{Nk}\right) + L^2\left(\frac{C_W R_W}{2S}\right) \quad \text{(EQ 2)}$$

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

$$\frac{\partial t_{interconnect}}{\partial k} = SC_T R_T\left(1 - \frac{1}{k^2}\right) - L\frac{C_W R_T}{Nk^2} = 0 \Rightarrow k = \sqrt{1 + \frac{LC_W}{SNC_T}} \quad \text{(EQ 3)}$$

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

$$\frac{\partial t_{interconnect}}{\partial S} = \left(k + \frac{1}{k}\right)C_T R_T - \frac{C_W R_W L^2}{2S^2} = 0 \Rightarrow S = L\sqrt{\frac{C_W R_W}{2\left(k + \frac{1}{k}\right)C_T R_T}} \quad \text{(EQ 5)}$$

These equations make physical sense. The number of stages S scales is proportional to length. It also depends on the ratio of wire delay to gate delay. The k+1/k term is the only difference from the inverter-based repeater solution, reflecting the extra delay of the buffer. In an inverter-based repeater, the term is unity because each inverter drives an identical inverter. In the buffer-based repeater, the term is k+1/k because the first inverter drives an inverter k times larger and the second inverter drives a receiver at the end of the line k times smaller. Thus, there are fewer repeaters because each introduces more gate delay. The transistor sizes are $\frac{N_0}{\sqrt{k}}$ and $N_0\sqrt{k}$ for the first and second inverters, respectively, where $N_0 = \sqrt{\frac{C_W R_T}{C_T R_W}}$ is the best transistor size if a single inverter were used as a repeater. In other words, the geometric mean of the transistor sizes remains constant, but one repeater is made larger for better drive and the other is made smaller to load the wire less. Finally, the step-up ratio k is chosen so that each inverter has equal fanout. The first inverter drives an inverter k times larger. The second inverter drives the wire capacitance LCW/S plus another inverter k times smaller. Solving this gives k = $\sqrt{\frac{Cgate + Cwire}{Cgate}}$, were Cgate is the total capacitance of the first inverter and Cwire is the total capacitance of the wire. This expression simplifies to the one shown in the equation.

For inverter-based repeaters, the design can be done by inserting repeaters such that the repeater delay equals the wire RC delay. For buffer-based repeaters, we have an additional unknown, so we need an additional constraint. We insert repeaters so that the total repeater delay through the two inverter stages equals the wire RC delay of each segment. We also choose the size of the second inverter so that the capacitive fanout of each inverter is equal.

Substituting the equations for N and S into the equation for k yields a 4th order equation $k^4 - 4k^2 - 1 = 0$ that can be solved for minimum delay with $k=\sqrt{2+\sqrt{5}}=2.06$. The other roots are non-physical. This is interesting because it shows the ratio of inverter sizes is independent of wire and transistor electrical characteristics. An explanation why is that Cgate and Cwire tend to scale in the same way: larger gates can driver longer wires and longer wires need larger gates for best drive. Since k is a ratio of these terms, k should be independent of physical parameters. Note, however, that it does assume no parasitic diffusion capacitance. Real diffusion capacitances lead to larger k.

The delay of the path is found by substituting N and S into the delay equation to yield:

$$t_{interconnect} = L\sqrt{C_W R_W C_T R_T}\left(2\sqrt{\frac{1}{k}} + \sqrt{2\left(k + \frac{1}{k}\right)}\right) \qquad \text{(EQ 6)}$$

Finally, substituting k=2.06, we obtain a delay of $3.64L\sqrt{C_W R_W C_T R_T}$. This compares with $3.41L\sqrt{C_W R_W C_T R_T}$ for repeaters built from single inverters. The buffer repeaters thus have a 7% delay penalty.

Repeater area is proportional to $\frac{N_0}{\sqrt{k}} + N_0\sqrt{k} = N_0\left(\sqrt{k} + \frac{1}{\sqrt{k}}\right)$. This a factor of 2.13 greater than a single inverter of size N_0 in an inverter-based repeater. However, there are only $\frac{1}{\sqrt{k + \frac{1}{k}}} = 0.62$ as many repeaters. Therefore, the area penalty is about 1.33 relative to inverter-based repeaters. This is approximate, because actual layout area is only indirectly related to transistor widths.

Power consumption is proportional to wire capacitance and gate capacitance. Gate capacitance increases by the same factor of 1.33 derived for total transistor gate area. Wire capacitance remains unchanged because the same total amount of wire is driven. If we know the relative gate and wire capacitances, we can predict how overall power scales. The capacitance of the inverters are NC_T and kNC_T. The load on the second inverter is k^2NC_T = Cwire + NC_T. Therefore, Cwire = $(k^2\text{-}1)NC_T$, and the total gate capacitance is $(k+1)NC_T$. The ratio of these capacitances is (k-1) = 1.06, indicating that wire and gate capacitance are approximately equal. Thus, power consumption scales by (1.33 + 1)/2 = 1.17.

4.0 Parasitic Delay

Real repeaters have parasitic delay. This makes each repeater more costly, leading to fewer repeaters. Thus, longer wire lengths are required and a larger fanout k must be used on each stage. Repeater parasitic delay can be modeled by adding a term $2SpR_TC_T$ to the interconnect delay equation. p is the ratio of the parasitic delay to the intrinsic delay of an inverter driving its own gate (R_TC_T). There are S repeaters, and hence 2S parasitic delays of pR_TC_T each.

The best size N of each stage is unchanged; it is orthogonal to the parasitic delay. The number of stages S decreases to:

$$S = L\sqrt{\frac{C_W R_W}{2\left(2p + k + \frac{1}{k}\right)C_T R_T}} \qquad \text{(EQ 7)}$$

The fanout of each stage k does not directly change, but is inversely related to the number of stages and hence will increase as the number of stages decreases. Substituting the equations for S and N into the equation for k, we obtain a new 4th order equation that is slightly messier: $k^4 - 4k^2 - 4pk - 1 = 0$. This can be solved numerically to obtain a table of k for various values of parasitic delay. p=1 is typical in a CMOS process. p=0.5 is reasonable with folded transistors. Lengthy routes between the desired interconnect path and the actual repeater location increase parasitic capacitance and effectively increase p.

TABLE 1. Fanout k as a function of parasitic p

p	**k**
0	2.06
0.5	2.25
0.75	2.34
1	2.41
1.25	2.48
1.5	2.55
2	2.67
3	2.88

5.0 Scaling

For reasonable process parameters in a 0.4 micron process, the delay of interconnect with proper repeater placement is about 50 ps/mm. How does this scale with process? Consider a process shrink in which x and y dimensions, V_{DD}, and oxide thickness are all scaled by $\alpha<1$. C_T, the capacitance per micron, scales with L/t_{ox} and is unchanged. R_T scales as α. Wire parameters depend on the scaling of wire thickness. If wire thickness is also scaled, C_W is unchanged and R_W scales as $1/\alpha^2$. If wire thickness is not scaled, R_W scales as $1/\alpha$. C_W depends on the parallel-plate capacitance and sidewall capacitance. Parallel-plate capacitance scales as α, while sidewall capacitance scales as $1/\alpha$. Sidewall is becoming dominant, so realistic processes are expected to scale somewhere between 1 and $1/\alpha$.

Putting this all together, wire delay per unit length scales as $\frac{1}{\sqrt{\alpha}}$ when thickness is scaled. When thickness is not scaled, wire delay scales somewhere between constant and $\frac{1}{\sqrt{\alpha}}$, depending on how much sidewall capacitance dominates delay. This means that interconnect with proper repeaters should continue to operate at about 50 ps/mm, increasing only gradually with process shrinks. This is much better than the scaling of unrepeated wires, which gets worse with $R_W C_W = 1/\alpha^{1-2}$. Unfortunately transistor delay is scaling with α, so even if a chip remains constant in size and repeaters are used, the gap between gate delay and global wire delay is growing as $\alpha^{1-1.5}$.

6.0 Summary

In summary, we have found a remarkably elegant answer to the cost of buffer-based repeaters. The repeaters are inserted such that the repeater delay equals the wire RC delay. The stage ratio is chosen such that each stage has equal fanout. This best fanout is independent to first order of process parameters and is about 2.06. Realistic diffusion capacitance increases the best fanout to about 2.4. Wires driven with buffer-based repeaters are 7% slower, have 33% more gate area, and consume 17% more power than inverter-based repeaters. Only 62% as many repeaters must be placed. They may be more suitable for repeater insertion CAD tools because they automatically achieve the correct polarity and because they have a smaller impact on routing congestion and floorplanning.