Pre-Physical Design Analysis and Optimization of Repeaters Based on Technology Node, Materials, Devices, and Repeater Options

W. T. Lynch, Independent Consultant

4012 Harriat Dr., Apex, NC 27502, lynch@ntrnet.net

ABSTRACT

For the first time a *comprehensive methodology has been applied to the pre-physical design of hierarchical interconnect wiring* with consideration of the limitations of both the device and the wiring technologies (Fig. 1). The overall goal is an optimized distribution of power supplies and clocks, inclusive with a *histogram-based decision algorithm* for general wiring [2,3]. The specific goal is the use of a minimum number of wiring levels, with wiring sizes that achieve the performance requirements. *Decision criteria are established for the insertion of repeaters* by means of a normalized comparison of trombone-staged drivers, single-stage inverting repeaters, and two-stage non-inverting repeaters (Fig. 3), *as well as for Al vs. Cu options*. When the repeaters are optimized to minimize delay, *the MOSFET r_0 and c_0 device parameters are no less than equally important to the per unit wiring length values for R_L and C_L*. The fundamental delay inhibitors are the MOSFET I_{sat}/C_{gate} slew rate (V/s) and the nodal wire length L.

Keywords: Interconnect, Repeaters, Copper, ICs, Wiring

1 BACKGROUND FUNDAMENTALS

MOSFET device scaling tends to produce clock frequency enhancements that are directly inverse to the scaling of the device as long as total transistor counts for the "chip" remain constant. Frequency (performance) scaling is, therefore, achieved naturally only for "local" wiring. Since chip sizes are increasing, rather than decreasing, for each technology generation because of the greatly increased transistor counts [1], other ways must be employed to provide the "scaled" performance for global wires. These include: the use of hierarchical wiring sizes, with scaled wiring used only for local routing; material modifications (Cu metal, Low K dielectric) to reduce the per unit length values of wire resistance (R_L) and capacitance (C_L); reduced logic gate count (logic depth, LD) between clock edges; and repeaters. *In effect, the chip is no longer a scaled inner world, but takes on a "packaging" function when hierarchical wiring sizes are employed.* For the 50 nm technology node, it is anticipated that the ratio of global to local signal wire cross sections will be greater than 1000:1 (4000 nm x 2000 nm/50 nm x 150 nm). The need to expand dimensions and change materials for the uppermost levels, even as the lowest levels are shrinking, increases not only the number of wiring levels (and their net process volume), but also increases process complexity (and cost) and may dominate design [2, 3].

Figure 2 gives the intrinsic delay relation for a standard logic gate. It is empirically correct for a string of similarly loaded gates. L is the length of the line; R_L and C_L, as described earlier, are the resistance and capacitance per unit length of the line; VDD is the (scaled) power supply voltage; I_{sat} is the saturation current for the MOSFET driver stage (proportional to the width W of the driver device); R_T is the "effective" driver resistance, VDD/I_{sat}; and C_T is the end-of-line logic gate capacitance at the input to the next stage. C_L includes the uniform wiring capacitance $C_{L,w}$ plus any uniformly distributed "fanout" logic gate capacitance C_{FO}/L.

$$\text{Delay } \tau = R_T C_T + L(R_T C_L + R_L C_T) + 0.4 * R_L C_L L^2 \quad (1)$$

There are four terms. One is independent of length L ($R_T C_T$) and this pure term depends only on the device and the end-of-line capacitive loading. The two terms linear in L represent interactive terms between the device and the wiring ($R_T C_L L$, $C_T R_L L$). One is a pure wiring delay ($0.4 * R_L C_L L^2$). It is the L^2 delay term that causes the most concern in long wires. A sufficient condition that removes the dependence on L^2 is to have $R_L L << R_T$; however, increasing R_T to satisfy the relation will increase the delay. It is the introduction of repeaters (next section) that reduces the entire delay for a "global" wire to a linear dependence on L. Physically, a linear transmission delay is what is optimally expected.

For the purposes of this paper, a constrained definition of "*global" wires are those wires, with associated loading, for which the insertion of repeaters will reduce the total delay.*

Depending on the frequency, loading, and logic depth (LD), repeaters may be useful at any level in the hierarchical wiring. *The wiring hierarchy*, often described as local/semi-global/global, or as "skinny/standard/fat", *will be defined as ecto/meso/endo* in the spirit of ectomorph/mesomorph/endomorph (skinny/muscular/fat) human body *aspect ratios*.

2 REPEATER OPTIMIZATION

This analysis is an extension of work by Otten [4] and Lynch/Arledge [2,3]. Figure 3 gives the basis on which comparisons and optimizations are made. The three comparison cases are:

I: a single driver, with an optimized staging of up to 3 "trombone" inverters following the final logic gate;

II: an optimized staging of inverting repeaters that break up the line into k_{opt} segments;

III: an optimized staging of non-inverting (two stage) repeaters that has its own k_{opt} value.

In all cases, in order to have a proper comparison, the final logic stage has n-channel MOSFETs with W/L_{gate} values equal to 50, which is ~ 2-3x that for a nominal logic gate. Figure 4 gives all the fundamental definitions and units. r_0 and c_0 are the fundamental device parameters, and s represents the W/L_{gate} sizing of the repeater devices. r_0c_0 scales as L_{gate}. The outputs of the analysis are the k_{opt} and s values for each of the three comparisons. L_{equiv} is the wiring length that, with a single trombone multi-stage driver, has the same delay as k_{opt} repeaters for the wire length L and the specified loading parameters $\{\tau_{rep}(L) = \tau_{tromb}(L_{equiv})\}$.

Trombone design (I in Fig. 3) is known to minimize τ when

$$C_{out}/C_{in} = C_{rat} = (C_L L/3c_0s_0) = \exp(n+1) \qquad (2)$$

where n+1 is the number of trombone stages, including the last stage of the circuit logic. {For most realistic examples, the improvements are slight when n>3, and so a maximum n of 3 has been used for these analyses.} s_1, $s_2,\ldots,s_{n+1}$ are optimized to minimize the delay by means of the relations

$$s_j/s_{j-1} = (C_{rat})^{(1/(n+1))} \; ; \; s_n/s_0 = (C_{rat})^{(n/(n+1))} \qquad (3a, b)$$

C_L includes both the wiring itself and any laterally distributed fanout runners. $3c_0$ refers to a basic inverter loading, with the p-ch device width 2x that for the n-ch).

The delay summation over all k sections for the inv case (II in Fig. 3) is minimized with respect to s, producing

$$s_{opt} = \{[(r_0C_L)/(3c_0R_L)]/[(k-1)/k+r_0/(s_0R_LL)]\}^{0.5} \quad (4)$$

Since s_{opt} is a function of k, and k must be an integer, calculations of $\tau(s_{opt}(k))$ give the minimum τ and, therefore, the optimum k.

The delay sum for the non-inv case (case III) introduces the additional s_1 gate delay term. With s_1 fixed by the relation $s_1 = (s_0s_2)^{0.5}$, a minimization for fixed k produces a (solvable) cubic equation for s_2 (equations 5a, b, c below):

$$s_2^{0.5} = [-b/2+(b^2/4+a^3/27)^{0.5}]^{0.33}+[-b/2 -(\ldots)^{0.5}]^{0.33} ,$$

$$\text{with } a=-s_0r_0k/(r_0k+s_0R_LL) \; ; \; b=a(2C_LL/(k3c_0s_0^{0.5}))$$

Normalized delays $\tau_{rep,norm}$ are obtained by dividing by $3r_0c_0$. Ratios may also be taken for τ_{rep}/τ_{tromb} for cases II and III.

Unless otherwise specified, a reference to an "Al" endo wire will refer to a 1 µm x 1 µm cross section with a 1 µm spacing of SiO_2 (K=3.9) to adjacent nearest-neighbor lines in either the vertical or horizontal direction. The Al wire acts as the unscaled endo control. Reference to "Cu" wire will refer to a 4 µm x 2 µm cross section with a 2 µm spacing of low K (K=1.95) on all sides. The Al wire will, therefore, have a horizontal pitch one third of the Cu wire, a vertical pitch one half of the Cu wire, an R_L 12x that of the Cu wire, and a $C_{L,w}$ 1.5x that of the Cu wire.

3 ANALYSIS RESULTS

The normalized (dimensionless) parameter

$$K_{tech} = L*[(0.4*R_LC_L)/(3*r_0c_0)]^{0.5} \qquad (1)$$

is the dominant parameter for determining τ (delay) minimization for the repeaters. Figure 5 gives the delay comparisons for Al and Cu at the 50 nm node.

It should be noted that, in Fig. 5 and in other figures for cases II and III, *the delay is linearly proportional to K_{tech} for a fixed technology node* (and, therefore, to L for uniform loading) and that the delay for the Al wire is not 18x longer than for Cu, but is only ~ 4x longer. The ~4x factor is reduced when the fanout C_{FO}/L is a significant portion of C_L. In fact, *for all optimized repeater cases,*

$$\text{Delay} \sim K_{tech}(3r_0c_0) \sim L(r_0c_0)^{0.5}(R_LC_L)^{0.5} \qquad (6)$$

Therefore, *the wire parameters are NEVER more important in determining delay than are the device parameters*. The worst case no-repeater delay dependency ~ $R_LC_LL^2$ has become ~ $(r_0c_0)^{0.5}(R_LC_L)^{0.5}$ L.

L_{equiv}/L calculations can be used to re-scale the length axis of a 3D histogram to include the effect of repeaters (Figs. 6 and 7 and [3]). The results are more dramatic for Al than for Cu.

The decision criteria for the insertion of repeaters to reduce global wire delay are:

$$L^2>L_{crit}^2 \sim= 3.7*(r_0/R_L)*(c_0/C_L)^{0.5} , \qquad (8)$$

with L_{crit} set by $\tau_{inv\ rep,k=2,s=sopt}(L_{crit}) = \tau_{tromb}(L_{crit})$;

$$L^2>L_{max}^2=(2.5f*T_{clock})/(R_LC_L),\ 1> f >/= (1/LD) \quad (9)$$

The L_{crit} relation (8) states that no improvement is achieved if a repeater is inserted for $L<L_{crit}$. The L_{crit} plot in Fig.7 includes an approximation that covers both Al and Cu, and technologies from 250 nm to 50 nm. *The L_{max} relation (9) says that there is no need for a repeater if $L<L_{max}$* , i.e., if the "L^2" delay term is less than the available (apportioned) fraction f of the clock period T_{clock}. f equals 1/LD for equally apportioned gate delays.

Fig. 9 gives the Al and Cu non-inv repeater delays, normalized to a single trombone driver, at the 50 nm node as a function of K_{tech}. Although the delay ratios are equivalent for the same K_{tech} , *the Al and Cu options cannot be compared*

as ever having the same K_{tech} except for the fact that additional lateral fanout loading may be assigned to the Cu trunk wire. Fig. 10 demonstrates that *ecto wires present no $R_LC_LL^2$ delay problems as long as wire lengths are less than 300-600 transistor spacings.*

4 SUMMARY

For values of L>Lcrit (8), the insertion of repeaters can reduce the net signal delay for K_{tech} >/= 5-8.
Non-inv (2-stage) repeaters provide slightly lower delays and also simplify the overall logic design.
2x4 μm^2 Cu/LoK wires require a considerably lower number of repeaters than do 1x1 μm^2 Al/SiO_2 wires, but the delay reduction at the 50 nm node is </= 4.2x. The ~4x reduction comes with the penalty of a 3x increase in endo layout area/pitch. Cu delay is only ~1.2x lower than 2x4 Al/LoK.
The Cu/LoK option can, however, service a lateral fanout area that is ~1.7x larger than for Al/SiO2 of the same sizing (e.g., meso, w/o employing lateral repeaters).
Since the repeater delay is ~$(r_0c_0)^{0.5}(R_LC_L)^{0.5}$, the parameters for the wire are never more important than for the device.
Repeaters can also be employed at the ecto level, but are not likely to be needed.
Signal frequency responses of 4-5 GHz for both local and global wiring should be achievable at the 50 nm node (clock distribution will require additional consideration).

REFERENCES

[1] NTRS 1997 document, SIA, 181 Metro Drive, Suite 450, San Jose, CA 95110
[2] W. Lynch and L. Arledge, "Scaling and Performance Limitations…," Proc. for ICPDI 1998, Feb 2-3.
[3] W. Lynch and L. Arledge, "Power Supply Distribution and Other Wiring …," MRS vol 514, p 11.
[4] R. Otten, "Global Wires Harmful?", Proc. of ISPD I998

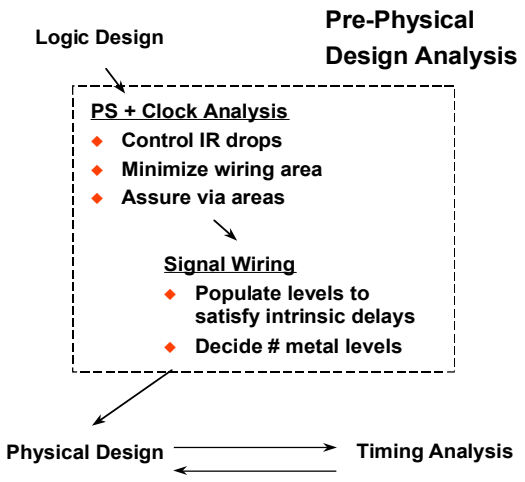
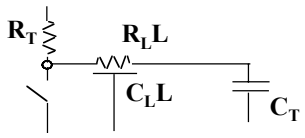


Fig.1. Focus on Pre-Physical Design Analysis

Circuit intrinsic delay
= [gate delay (including cap loading)] + [wiring delay]
$= [R_TC_T + R_TC_L\,L] + [C_TR_LL + 0.4*R_LC_L*L^2]$

$R_T = VDD/I_{sat}$
C_T = output gate capacitance (at the end of the line)



$C_L = C_{L,w} + \Sigma\,(C_{lat,w,i}*L_{lat,i}/L) + C_{FO}/L$

C_{FO} = fanout gate capacitance along the line

With $R_L*L << R_T$, the delay is represented by:

Delay = (VDD/Isat)*C, where $C = (C_L*L+C_T)$

Fig. 2. Intrinsic Delay (assumes that adjacent circuits are quiet)

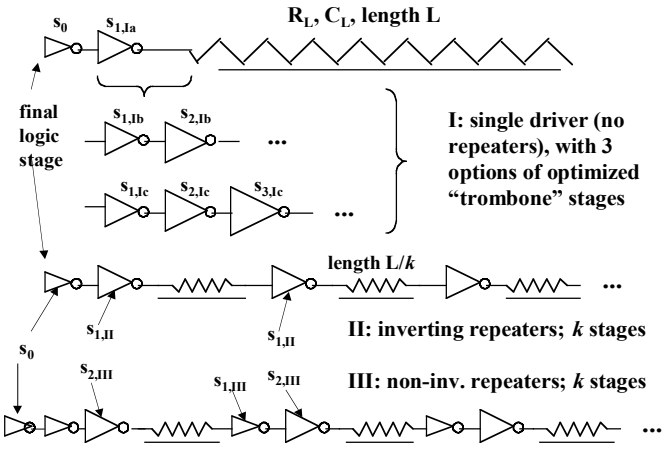


Fig. 3. Repeater Comparisons

- $r_0 = VDD/I_{sat}$, for $W=F=L_{gate}$ (Ω/unit square)
- $c_0 = C_{gate\ diel}$, for $W=F=L_{gate}$ (fF/unit square)
- $R_{L,w} = R_L$ = wiring res. per unit length (Ω/mm)
- $C_{L,w}$ = wiring cap. per unit length (fF/mm)
- $C_{L,FO} = C_{FO}/L$ = distributed fanout cap C_{FO} per unit length (fF/mm) - for cases when C_{FO} is distributed
- $K_{tech} = L*[(0.4*R_LC_L)/(3r_0c_0)]^{0.5}$, a "technology parameter"
- $\tau_{min,rep}$ = delay with an optimized number (k_{opt}) and sizing ($s_{1,opt}$, $s_{2,opt}$) of repeaters (ps)
- s_0 = sizing (n-ch W/L) of last logic gate before the first driver (chosen as 50 to standardize all comparisons)
- L_{equiv} = wiring length that, with a single multi-stage driver, has the same delay as k_{opt} repeaters for the length L

Fig. 4. Essential Definitions

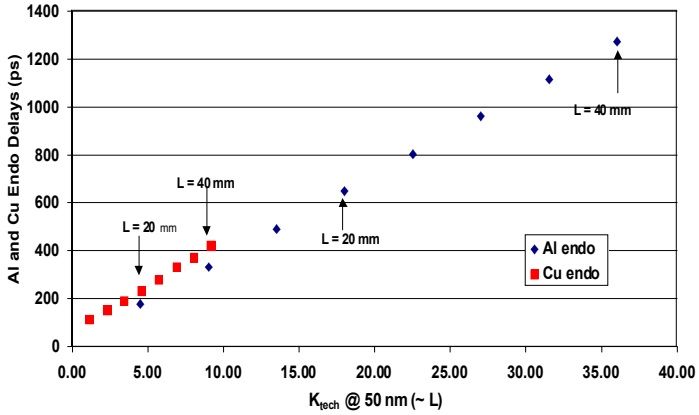


Fig. 5. Comparison of Delays at 50 nm for 1x1 Al/SiO2 and 2x4 Cu/LoK Endo Wires (L=5 to 40 mm; non-inv rep; $C_L=C_{L,w}$+133 fF/mm distributed fanout)

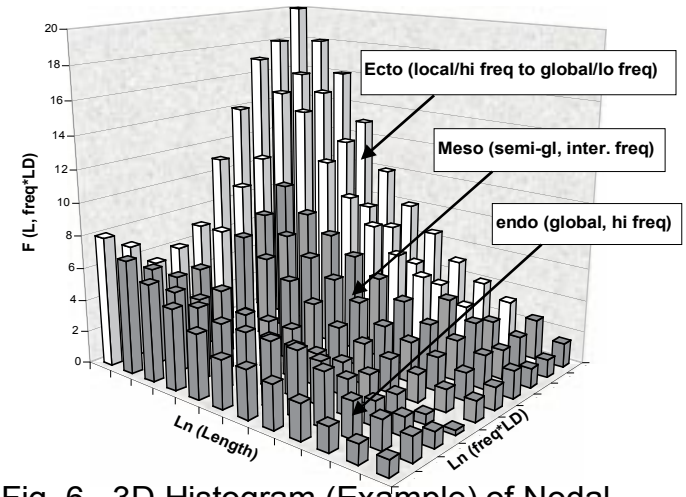


Fig. 6. 3D Histogram (Example) of Nodal Lengths and Inverse Gate Delays

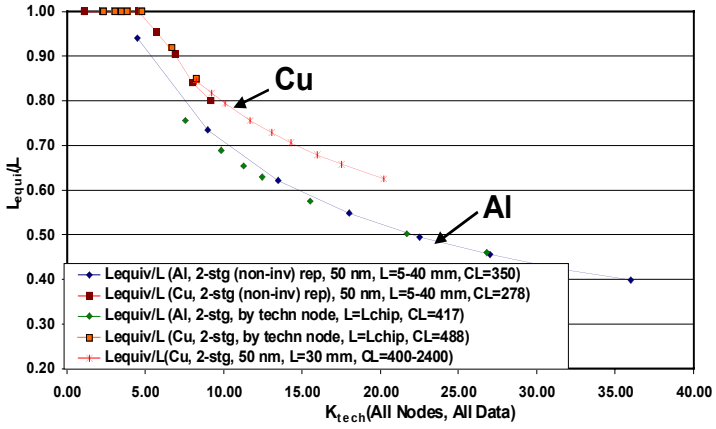


Fig. 7. L_{equiv}/L for "Al" and "Cu" Endo (Includes All Data for Nodes, Lengths, and Loading)

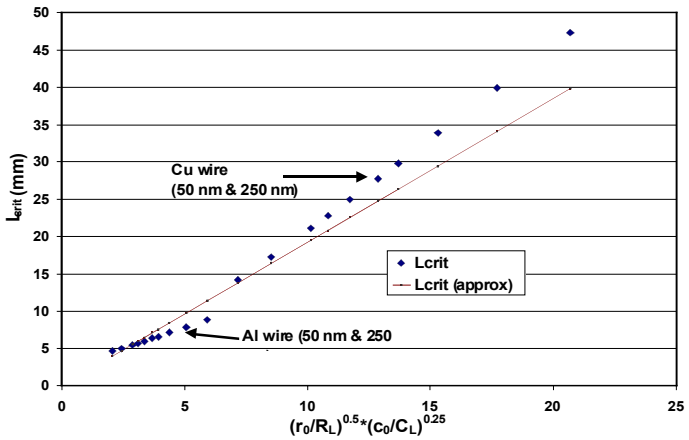


Fig. 8. Endo Lcrit (determined directly, and approximated) for 50 nm and 250 nm Device Technologies

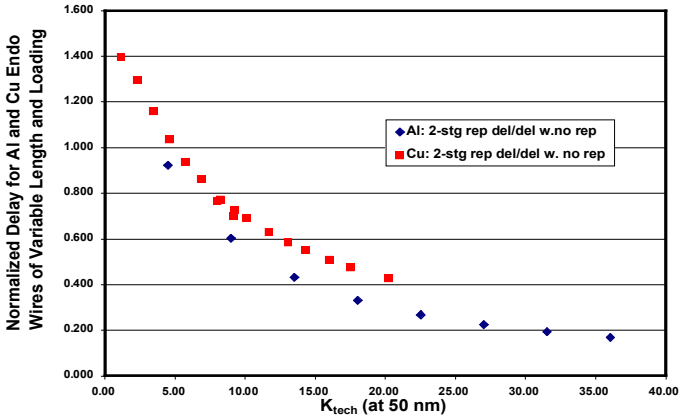


Fig. 9. Endo Delays for Non-Inv Repeaters Normalized To a Single Trombone Driver: 50 nm Technology

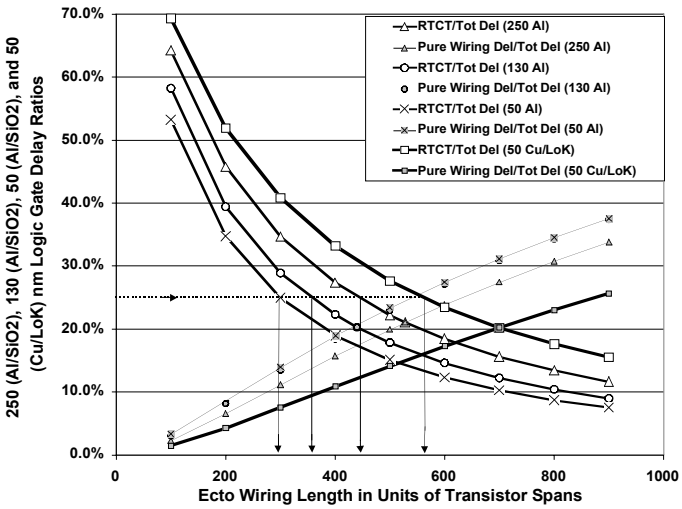


Fig. 10. Ratios of Pure R_TC_T and Pure $R_LC_LL^2$ Delays to Total Gate Delays, for Nominal Device Sizes, a FO of 3, and Ecto Wiring for Technology Options