Power Estimation in Global Interconnects and its Reduction Using a Novel Repeater Optimization Methodology

Pawan Kapur
Dept. of Electrical Engineering
CIS 112, Stanford, CA-94305
kapurp@stanford.edu

Gaurav Chandra
Dept. of Electrical Engineering
CIS, Stanford, CA-94305
gchandra@stanford.edu

Krishna C. Saraswat
Dept. of Electrical Engineering
CISX 326, Stanford, CA-94305
saraswat@cis.stanford.edu

ABSTRACT

The purpose of this work is two fold. First, to quantify and establish future trends for the dynamic power dissipation in global wires of high performance integrated circuits. Second, to develop a novel and efficient delay-power tradeoff formulation for minimizing power due to repeaters, which can otherwise constitute 50% of total global wire power dissipation. Using the closed form solutions from this formulation, power savings of 50% on repeaters are shown with minimal delay penalties of about 5% at the 50 nm technology node. These closed-form, analytical solutions provide a fast and powerful tool for designers to minimize power.

1. INTRODUCTION

The power dissipation in high performance integrated circuits (ICs) is quickly becoming a performance bottleneck. The scaling paradigm will exacerbate the power problem severely from many different angles. On one hand the device leakage power due to sub-threshold and gate leakage is likely to rise in the future. On the other hand, the dynamic power will also increase not only due to a larger number of devices and interconnects on the chip, but also owing to the burden of keeping the speeds of electrical interconnects compatible with increasing clock frequencies. Certain estimates, which merely include device power (both dynamic and static), have shown chip power densities to increase to about 200 W/cm^2 at 35 nm node [14]. Including the interconnect power would substantially increase these power density estimates. This may lead to a significant increase in the chip temperature, which would degrade both reliability, through greater susceptibility to electromigration failures, and performance, through increased interconnect resistance and poorer device characteristics. The situation presents an impending power crisis, which threatens to slow down the progress of the chip industry.

It is important to identify the major power consumption sources on a chip, quantify them, and focus on efficient technological, circuit and/or architectural solutions to minimize them. Toward this goal, we address a potentially large source of chip power: the power required for global signaling. We first deal with a realistic estimation of the power consumption in global wires including that due to repeaters, used to speed these wires. Having quantified this as a function of future technology nodes, we then introduce a novel, efficient, methodology which minimizes the power consumed by repeaters in global interconnects.

DAC 2002, June 10-14, 2002, New Orleans, Louisiana, USA.

2. POWER CONSUMPTION IN GLOBAL SIGNALING

2.1 Sources of Power Increase in the Future

The power consumed in signaling over long on-chip, global interconnects will rise in the future owing to many reasons. Firstly, the number and the length of global signaling wires is likely to grow because of an increase in the complexity and the area of the chip, and because of a shrinkage in module sizes. This, along with wire scaling would lead to an increase in the total capacitance of global wires, and hence dynamic power dissipation, despite the introduction of lower dielectric constant materials. Secondly, deterioration of interconnect performance (delay and bandwidth) with scaling will lead to solutions, such as repeater insertion, which attempt to fix this problem, but burn power. Finally, communication across chip will require increasing number of clock cycles in the future. This would require deeper wire pipelining, hence larger power consumption. The first two sources of power increase will be modeled in detail in subsequent sections. The extent to which the third source of power is addressed in this work is limited to modeling the deterioration in the global interconnect latency with respect to the clock period. Since this will dictate the pipelining depth, the extra power consumption may be deduced.

2.2 Global Wire and Repeater Power Calculations

In this section we present the global wire and repeater power dissipation calculations. First, we address the repeater power. The power dissipated due to global wires can, subsequently, be deduced from repeater power, as will be elucidated later. Before performing any calculations related to interconnects we need realistic, technologically constrained, modeling of the interconnect parameters, such as resistance. This accurate resistance modeling will be even more important in the estimation of interconnect latency, as presented in a later section.

2.2.1 Realistic Copper Resistivity

Ideal, constant, copper (Cu) resistivity (ρ) may lead to overly optimistic results for both power and latency. Scaling of interconnects increases the effective copper ρ due to the electron surface scattering effect and the copper barrier effect [10]. The surface scattering effect depends on barrier/copper interface quality. This is characterized by a parameter P, which could range from 0 to 1 with P=0.5 being a reasonable value for Cu [10]. The barrier effect depends on the barrier thickness and its deposition

technology. The possible barrier deposition technologies include Ionized Physical Vapor Deposition (IPVD), collimated PVD, PVD and Atomic Layer Deposition (ALD). ALD, because of its conformality, is most effective in lowering resistivity.

The rise in global wire effective Cu resistivity is shown in Fig. 1 [10]. The wire dimensions and other required parameters were taken from International technology roadmap for semiconductor (ITRS '99) [1]. With reasonable parameters defined as a chip temperature of 100 ^{0}C, a barrier thickness (BT) of 10 nm, P=0.5, and the ALD deposition technology, the copper effective resistivity rises to about 3.2μΩ-cm at the 35 nm node, much larger than the ideal value of 1.7μΩ-cm. The wire resistance was next calculated using above resistivity modeling for various technological conditions (Fig. 2). The effect of technology on resistance values at the 35 nm node is summarized in Table 1. It also indicates the error in resistance in parenthesis, if ideal ρ is used.

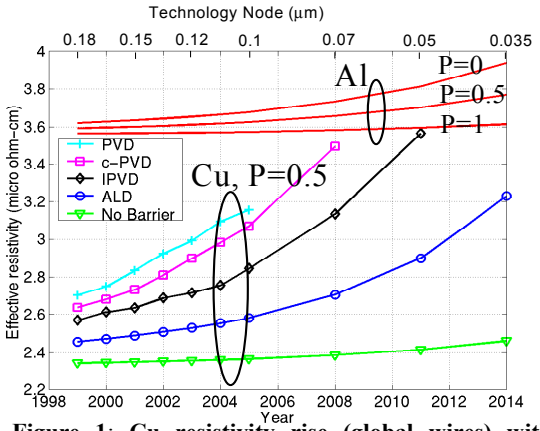

Figure 1: **Cu resistivity rise (global wires) with scaling. Barrier thick.=10nm Temp.=100 ^{0}C. Al resistivity showed for comparison**

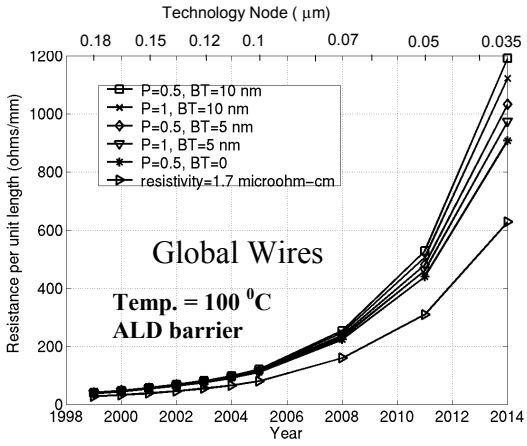

Figure 2: Global Cu wire resistance per unit length with ALD barrier with diff. P and barrier thickness

Table 1: **Summary of resistance per mm at 35 nm techn. node. T=100^{0}C, ALD Barrier dep. Techn**.

Practical Constraint	Global Resist. (Ω/mm)	Semi-global Resist. (Ω/mm)	Local Resist. (Ω/mm)
	Year 2014	Year 2014	Year 2014
None: ideal ρ=1.7μΩ-cm	628	1773	3275
P=0.5, BT=10 nm	1192 **(90%)**	4351 **(145%)**	9564 **(192%)**
P=1, BT=10 nm	1123 **(79%)**	3942 **(122%)**	8490 **(159%)**
P=0.5, BT=0	908 **(45%)**	2668 **(51%)**	5030 **(54%)**

2.2.2 Estimating the Number of Repeaters

The estimation of number of global wire repeaters requires 1) an accurate wire length distribution for various parts of the chip, 2) a demarcation method that decides the length boundaries between various wiring tiers (local, semi-global and global) as well as the length beyond which repeated wires are faster, and 3) the dimensions of the wires. In most previous approaches, which address the impact of scaling on wire length distribution, e.g. [7], the distribution is homogeneously calculated for the entire chip. However, a realistic wire length distribution necessitates a distinction between cache and logic areas and transistors. In our work, we partition the problem into logic and memory blocks and treat them separately. We get the area occupied by random logic by subtracting cache area from total chip area. A 6T SRAM cell occupies about 4.65 um^2 at 180 nm technology node and 3.45 um^2 at 150 nm technology node [2], giving a size of approximately 600λ². Here λ is the design parameter, such that, 2λ refers to the technology node in μm. Above memory size per bit along with the number of memory bits (taken from ITRS), is used to compute future cache area. Using the area occupied by random logic, and the number of logic transistors (from ITRS), we then obtain the wire length distribution based on the stochastic methodology proposed by Davis et. al. [5].

To obtain the length boundaries between local, semi-global, and global wires, we assume a modular design based approach proposed in [15]. In this approach, it is assumed that most modules will be on the order of 50,000 gates, within which the routing is done using local interconnects. The modules will be further grouped into isochronous regions connected by semi-global wires. The extent of these regions will be dictated by the distance traversed in 90% of the fast, local, clock cycle, over semi-global wires. Finally, fat global wires will connect various isochronous regions and communication between these regions will use a slower global clock. The idea of having isochronous regions operating on a local clock connected together by a network of global wires operating on a global clock parallels the present board level design scenario. Hence we think it is reasonable to believe that future on-chip designs would adapt to such design methodologies.

The third step for repeater number estimation is the knowledge of wire dimensions. We use the ITRS dictated future pitch and assume the width of wires to be half of that. There have been suggestions that the ITRS dictated pitch is only the minimum pitch, and in some cases, the global wires do not have to scale at all, staying at a constant pitch for future technology nodes [15]. However, this will lead to large number of metal levels. Concerns have already been made about the increasing signal wires leading to a rapid increase in the required number of metal layers [7]. We justify the use of ITRS dictated wire pitch by computing the number of metal layers required for just the signal wires with these pitches (Fig. 3). Two typical Rent's exponents of 0.55 and 0.6 are used for this calculation [6], [8], [13]. Additional layers would be needed to accommodate power/ground and clock wires. It is evident from the figure that even with the ITRS dictated aggressive wire scaling, the allocated metal levels by ITRS are not sufficient to accommodate the wires at far future nodes. For the present and near future technology nodes, the allocated metal layers appear to be in excess of the number required. However, owing to a large number of wires on the chip, slight increase in the pitch will lead to a rapid increase in metal levels. Thus, we don't expect a significant deviation in the average wire pitch from the ITRS dictated pitch even for near term technology nodes.

The only remaining thing to be established is the fraction of the global wires, which will have repeaters on them. We assume that repeaters on a wire will be inserted when the repeated wire delay is

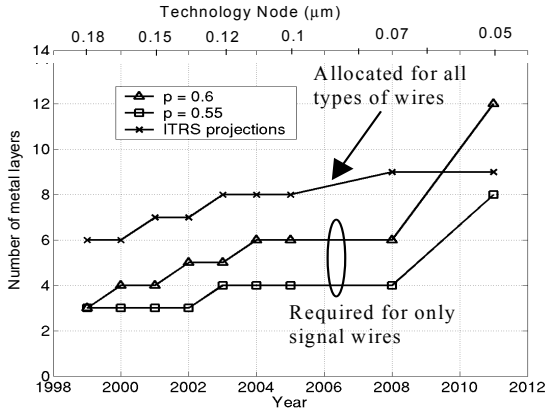

Figure 3: **No. of metal levels needed to accommodate only signal wires at future nodes. ITRS projection of total no. of metal levels also shown.**

smaller than the unrepeated delay. This occurs at the optimal distance between repeaters, (l_{opt}), originally formulated in [3] and expressed in the following form in [11]

$$l_{opt} = 3.24\sqrt{\frac{r_o C_{nmos}}{R_w C_w}} \quad (1)$$

Here, C_{nmos} and r_o are the capacitance and resistance of the minimum sized NMOS transistor, respectively. R_w and C_w are the resistance and capacitance per unit length of wires respectively. We find that l_{opt} for global wires is always less than the minimum global wire length. Hence, all global wires will have repeaters on them. We call the length, beyond which repeaters are inserted, as the crossover length. In our case, this length is the same as the minimum global wire length. Thus, for a wire of length *l,* the number of repeaters on that wire is:

$$n_{repeater}(l) = \begin{cases} 0, & \text{if } l < l_{crossover} \\ (round\left(\frac{l}{l_{opt}}\right)) - 1, & \text{otherwise} \end{cases} \quad (2)$$

Using the statistical wire length distribution, the minimum global wire length, and the number of repeaters at a given length from (2), we compute the total number of repeaters, $N_{repeater}$. The resulting number of repeaters, for two Rent's exponents of 0.55 and 0.6, are shown in Fig. 4, for realistic as well as ideal copper resistivity. The global signal wire repeaters are found to be as high as 5.5 million at the 50 nm technology node with reasonable copper resistivity and a Rent's exponent of 0.55. We compare our repeater number estimates with those obtained by other authors [4], [15] at the 70 nm technology node (Table 2). Our prediction of about 0.85 million repeaters, for a Rent's exponent of 0.55, lies between the two numbers predicted by references [15] and [4], where as, a Rent's exponent of 0.6 yields results which match well with [4]. The repeater estimate obtained in [15] is quite less because in this work the global wires are kept at a constant pitch at future nodes.

Table 2: Comparison of no. of repeaters of our approach with previous work. The numbers shown are for 70 nm technology

Number of repeaters estimated by [4]	Number of repeaters Estimated by [15]	Our approach, p=0.55	Our approach, p=0.6
1.6 million	0.2 million	0.85 million	1.61 million

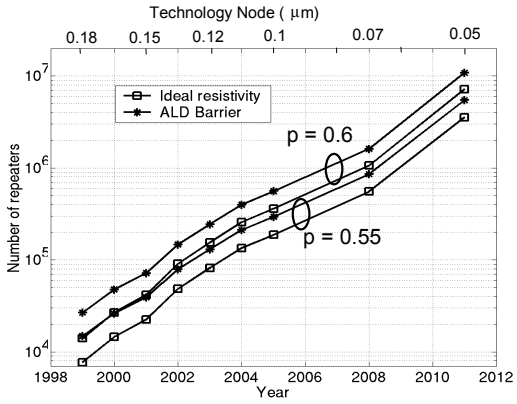

Figure 4: **Total no. of repeaters on global wires as a function of tech. node for different p (Rent's exponent)**

2.2.3 Power Due to Delay optimized Repeaters

The short circuit power of repeaters is neglected in our analysis. For estimating dynamic power, the capacitance due to all the repeaters on global wires, $C_{repeater}$, is given by

$$C_{repeater} = N_{repeater}(6 S_{opt} C_{nmos}) \quad (3)$$

Where,
$$S_{opt} = 0.58\sqrt{\frac{r_o C_w}{R_w C_{nmos}}} \quad (4)$$

and
$$C_{nmos} = C_g(2\lambda) \quad (5)$$

Here, S_{opt} is the optimal sizing of the NMOS in the repeater [3], [11]. C_g is the NMOS gate capacitance per micron, and is expected to stay constant at about 1.75 fF/µm for future technology nodes [9]. For a repeater, PMOS is assumed to be twice as large as NMOS. Also the diffusion capacitance is assumed to be the same as the gate capacitance. This leads to 6 times the NMOS gate capacitance in (3). The total dynamic power dissipation due to repeaters is

$$P_{repeater} = s_w C_{repeater} V^2 f_{clock} \quad (6)$$

Where, s_w is the switching activity factor, and V and f_{clock} are supply voltage and clock frequency, respectively. For a reasonable switching activity of 0.15 [16], the power dissipation due to global wire repeaters for future technology nodes is shown in Fig. 5. It is evident that the added power dissipation due to repeaters is a serious problem in the future. At 50 nm technology node, with a reasonable Rent's exponent of 0.55 [13] and using ideal copper resistivity, the repeater power dissipation is about 50 Watts, and with realistic copper resistivity it is about 60 Watts. The resistance plays a role in repeater power as it dictates the crossover length beyond which repeaters are inserted. The power numbers are much worse for a Rent's exponent of 0.6.

2.2.4 Power Due to Global Wires

The power dissipation due to global wires themselves can be simply obtained from the repeater power by realizing an interesting fact regarding the total capacitance of all the repeaters placed optimally on a single wire. This capacitance can be obtained by multiplying a single repeater capacitance ($6S_{opt}C_{nmos}$) by the number of repeaters on a wire (L/l_{opt}), where, l_{opt} and S_{opt} can be obtained using (1) and (4), respectively. The expression, thus obtained, is independent of the wire resistance, and is given by:

$$C_{repperline} = 1.07 C_{line} \quad (7)$$

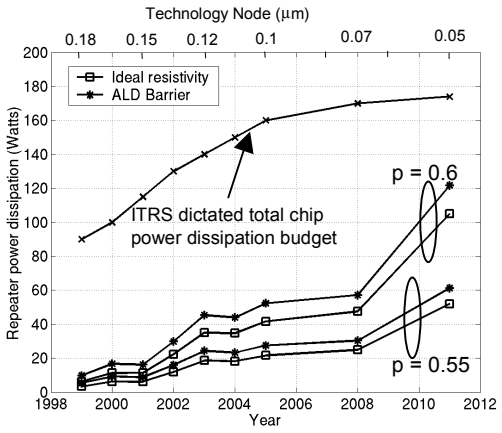

Figure 5: Repeater power dissipation as function of tech node. ITRS dictated total chip power budget also shown.

Where, $C_{repperline}$ is the total capacitance of all delay optimized repeaters on a single wire and C_{line} is the capacitance of a single wire. Since we established before that all global wires use repeaters, the total power dissipated due to global wires is approximately same as that due to all global repeaters. Hence, the total power dissipation approximately doubles, yielding about 120 Watts of power at 50 nm node with a Rent's exponent of 0.55. This can be a substantial fraction of the total chip power.

3. REPEATER POWER MINIMIZATION METHODOLOGY

The exorbitant power consumption due to delay-optimized repeaters at future technology nodes can be of serious concern. A simple method to reduce repeater power is to decrease the repeater size and/or space them further apart. Both these solutions lead to a delay penalty. In this section, we develop a novel formulation which optimizes the separation and sizing of the repeaters such that the power savings is maximized for a given delay penalty.

The expression for delay due to repeaters which are spaced distance l apart and whose NMOS transistor is sized, S, (channel width to length ratio) can be simply obtained by applying Elmore delay model to a simplified RC network for a stage (one repeater to the next) and is given by

$$\tau_{rp} = \frac{L}{l} t_{stage} \tag{8}$$

$$= L\left(\frac{b(1+e)(1+f) r_0 C_{nmos}}{l} + a R_w C_w l + \frac{b r_o C_w}{S} + b(1+e) R_w C_{nmos} S \right)$$

Here, L is the length of the wire, a and b are switching model dependent parameters. If we assume that the output of the repeaters switches when the input reaches half of the voltage swing, a and b are found to be about 0.4 and 0.7, respectively [12]. Parameter e is the ratio of the PMOS to the NMOS size and f is the ratio of the diffusion capacitance to the gate capacitance of the transistors. Equation (8) can be optimized independently with respect to S and l to give minimum delay. This yields

$$\tau_{rpopt} = 2L\left(\sqrt{ab(1+e)(1+f)} + b\sqrt{1+e} \right) \sqrt{R_w C_w r_o C_{nmos}} \tag{9}$$

$$l_{opt} = \sqrt{\frac{b(1+e)(1+f)}{a}} \sqrt{\frac{r_o C_{nmos}}{R_w C_w}} \tag{10}$$

$$S_{opt} = \sqrt{\frac{1}{1+e}} \sqrt{\frac{r_0 C_w}{R_w C_{nmos}}} \tag{11}$$

For the typical value of e=2 (PMOS sized twice of NMOS), f=1 (diffusion capacitance is same as gate capacitance), and above stated a and b values, (10) and (11) reduce to (1) and (4), respectively. Now, in an attempt to reduce power, we decrease S and increase l, such that $S = x_s S_{opt}$ and $l = l_{opt}/x_l$. Here, x_s and x_l are less than one and denote the fractional change in sizing and spacing from delay optimal values. The total wire delay can then be written as

$$\tau_{rp} = L\left(\sqrt{ab(1+e)(1+f)} \left(x_l + \frac{1}{x_l} \right) + b\sqrt{1+e} \left(x_s + \frac{1}{x_s} \right) \right) \sqrt{R_w C_w r_o C_{nmos}} \tag{12}$$

For x_s and x_l equal to 1, (12) reduces to (9). The delay penalty, β, expressed as a ratio of delay with sub-optimal (x_s and x_l not equal to 1) repeaters to that with delay optimized repeaters (x_s and x_l equal to 1) can be written as

$$\beta = \frac{\tau_{rp}}{\tau_{rpopt}} = \frac{\left(x_s + \frac{1}{x_s} \right) + A\left(x_l + \frac{1}{x_l} \right)}{2(1+A)} \tag{13}$$

$$\text{where,} \quad A = \sqrt{\frac{a(1+f)}{b}} \tag{14}$$

Next, we examine the power consumption of a single repeated wire due to its capacitance and the capacitance of repeaters on it. This power for the delay sub-optimal case (general form) is given by

$$P = s_w \left[C_w L + (1+f)(1+e) C_{nmos} S_{opt} \frac{L}{l_{opt}} x_s x_l \right] V^2 f_{clock}$$

$$= s_w V^2 f_{clock} C_w L (1 + x_s x_l A) \tag{15}$$

The first and the second terms in the parenthesis correspond to the wire and the repeater contributions, respectively. For delay optimal case, where $x_s = x_l = 1$, the ratio of the capacitance of all the repeaters on a single wire to the wire capacitance turns out to be equal to A. For a reasonable value of f=1, A is 1.07 from (14), agreeing with (7).

The amount of power saving obtained per wire can be expressed as the ratio of the total power per wire in the power saving repeaters to that in the delay optimized repeaters (δ). This is easily obtained using (15) and is given by

$$\delta = \frac{P}{P_{opt}} = \frac{(1 + x_s x_l A)}{(1+A)} \tag{16}$$

We propose that using the expressions for delay penalty, (13) and power savings, (16), one can find x_s and x_l, such that, for a required power saving, minimum delay penalty is incurred, or vice versa. This condition can be achieved by substituting x_l expressed in terms of δ and x_s from (16), into the expression for β, and minimizing β with respect to x_s. The minimum β and the corresponding x_{sopt} and x_{lopt} are obtained to be the following

$$\beta_{min} = \sqrt{\frac{A - 1 + \delta}{A + 1 - 1/\delta}} \tag{17}$$

$$x_{lopt} = \frac{\sqrt{(A + 1 - 1/\delta)(A - 1 + \delta)}}{A} \tag{18}$$

$$x_{sopt} = \delta \sqrt{\frac{A + 1 - 1/\delta}{A - 1 + \delta}} \tag{19}$$

The minimum delay penalty (β_{min}) can also be expressed in terms of the fractional decrease in power due only to the repeaters on a single wire (η) instead of δ. η is given by the product of x_s and x_l. Using (16) to relate δ and η, and substituting in the expression for β_{min} leads to the following compact expression after simplification:

$$\beta_{min} = \frac{\sqrt{(1+\eta A)(1+A/\eta)}}{(1+A)} \qquad (20)$$

Equations (17)-(20) can, thus, be used to minimize the delay penalty for a given power saving on repeaters, or vice versa.

To see the possibility of the most efficient delay-power trade off graphically, we first plot the delay penalty vs. desired power saving in Fig. 6 using (13). The power saving indicated by each curve represents saving only through size reduction at a fixed spacing. Different curves are for different spacings ranging from l_{opt} to $3l_{opt}$. The x-axis is the ratio of the repeater powers only. Different delay penalties for same desired power saving, with different combination of size and spacing, point to the possibility of optimization. In the past, a few attempts have been made to reduce repeater power dissipation [15]. In [15] the product of cube root of the width (proportional to S) and delay was taken to be the objective function for minimization. This minimization leads to an optimal spacing of l_{opt}, a size of $0.5S_{opt}$, a delay penalty, β of 12.5% and a 50% power saving on repeaters [15]. This particular data point is shown in Fig. 6. It is clear that for the same power saving there exists a different combination of S and l which gives a lower delay penalty of about 6%. Thus, the optimization we propose leads to a more efficient tradeoff, and provides a range in this tradeoff.

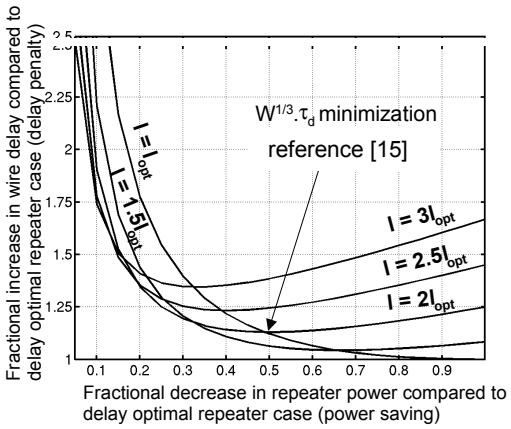

Figure 6: Plot showing the delay-power tradeoff. Each curve represents power saving only due to repeater size reduction

Fig. 7 shows the optimized curve for delay penalty-power saving tradeoff using (18)-(20). The figure also shows the corresponding fractional change in spacing and sizing required (compared to delay optimal repeaters) to achieve this optimization. We observe that with our proposed optimization it is possible to save large repeater power for moderate delay penalties.

Next, we apply the optimal trade off formulation to estimate the total power savings due to all the repeaters on a chip. Fig. 8 shows the total power dissipation due to repeaters as a function of technology node, for different tolerable delay penalties. Even for a small delay penalty of 5%, repeater power dissipation is reduced by approximately 50% (from 61 watts to about 30 watts at 50 nm node). The savings saturate rapidly for larger delay penalties and are 62% and 80% respectively for a 10% and 25% delay penalties. An important point to mention here is that the particular optimization for delay-power tradeoff is done using only a single

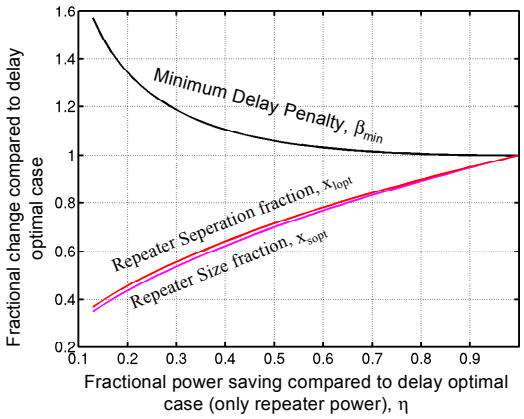

Figure 7: **Optimized delay-power tradeoff curve with our methodology. Also shown repeater size and spacing to achieve the respective power saving.**

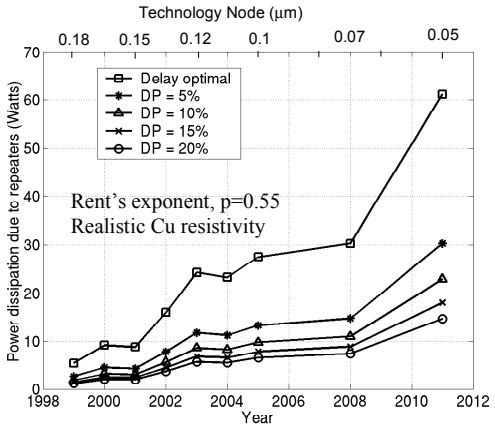

Figure 8: Plot showing the effectiveness of delay-power tradeoff. Power saving for various delay penalties (DP)

wire. However, there can be a third source of power saving which comes about through a decrease in the number of repeaters in a different way. This reduction occurs if the crossover length, beyond which repeaters are inserted, is increased. For this to happen the crossover length has to be dictated by the repeater spacing and not by the minimum wire length at the global tier. While the methodology we choose does not lead to this possibility, we consider this case for the sake of completeness. In such a case, the constraining equation for power has to be modified to the following instead of (16).

$$\delta_{global} = \frac{P_{global}}{P_{optglobal}} = \frac{(1+x_s x_l A)\int_{l_{opt}/x_l}^{L_{max}} xn(x)dx}{(1+A)\int_{l_{opt}}^{L_{max}} xn(x)dx} \qquad (21)$$

Here, n(x) is the number of wires at length x and is given by wire distribution function obtained using the Rent's rule. If repeater spacing is greater than minimum global wire length, (21) along with (13) should be used for optimization of power-delay tradeoff. The solution would involve using numerical methods.

4. INTERCONNECT LATENCY

It is important to analyze interconnect latency with respect to clock period because a large interconnect latency for communication across the chip would require deeper pipelining, hence larger power

dissipation. The second reason an analysis of wire latency is warranted, is to complete the discussion on repeater power-delay tradeoff by quantifying the latency penalty in absolute terms for power saving repeaters. Using the realistic resistance values, the latency calculations for global wires were performed and are shown in Fig. 9 [11]. Fig. 9 shows the chip-edge long global wire delay with and without the repeaters as a function of the clock period. It is seen from Fig. 9, that despite repeaters, the latency rises to about 7.5 times the clock period at 35 nm node, an increase of about 8X compared to the 180 nm node. The 8X rise in latency occurs due to three independent factors: 1) increase in clock frequency (~3X), 2) increase in chip edge (~ 1.45X), and 3) increase in the delay per unit length (1.85X). This latency rise will cause a further increase in power consumption due to global wires.

Fig. 9 also depicts the latency curve relevant to power saving repeaters with a delay penalty of 25% compared to delay optimal case. It should be recalled that this small delay penalty results in a large power saving of approximately 80% over delay optimal repeaters. Further, it is obvious from the figure that the 25% delay penalty curve is still much faster than the non-repeated wires.

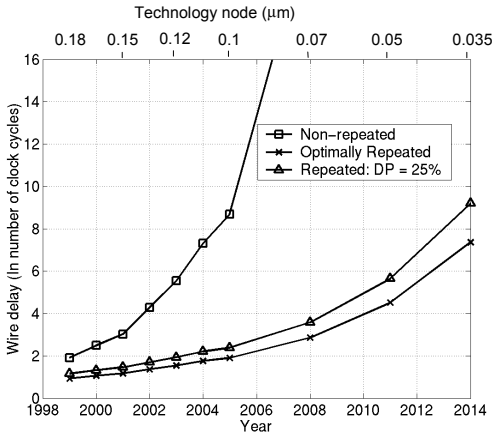

Figure 9: Chip -edge long global wire latency (in clock cycles) with and without repeaters. 25% delay penalty compared to delay-optimal repeaters also shown. ALD barrier assumed

5. CONCLUSION

In this work, we have identified and quantified the future power expenditure on long distance signaling in high performance chips, and suggested an efficient power-delay trade-off formulation to lower this power. In an attempt to accurately model various sources of power dissipation in global signaling, we use practical interconnect resistance, dictated by technological constraints. The power dissipation due to repeaters and wires, as well as that arising indirectly because of multi-clock cycle across chip communication latency, can be misleadingly optimistic, if ideal copper resistivity values are used. The power due to global interconnects and repeaters was quantified and was found to be as large as about 120 Watts at 50 nm technology node. This could be a substantial fraction of the total chip power. Approximately, half of this power was due to repeaters. With the motivation to ameliorate the large repeater power, closed form solutions for an efficient repeater delay-power tradeoff were developed. As much as 50% repeater power can be saved with merely a 5% delay penalty using this formulation. The closed form solutions give designers a powerful tool for a quick back of the envelope calculation for saving power. Finally, interconnect latency was discussed to both partially address the third power dissipation source in global signaling (arising from multi-clock cycle interconnect latency), and to quantify the delay penalty in power-delay tradeoff, for power saving repeaters. On one hand, it showed that the even 25% latency penalty over delay optimized repeaters, which saved 80% of repeater power, was very less compared to non-repeated wires. On the other hand, it pointed to about 8 times clock period latency even with delay optimized repeaters at 35 nm, leading to a possibility of further power dissipation due to deeper pipelining.

6. ACKNOWLEDGEMENTS

The authors would like to thank the MARCO Interconnect Focus Center for providing support for this research.

7. REFERENCES

[1] *The International Technology Roadmap for Semiconductors* (ITRS), 1999.

[2] [Online]. TSMC manufacturing specifications. Available: http://www.tsmc.com/technology/index.html.

[3] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*, Addison-Wesley, 1990

[4] J. Cong, "An interconnect-centric design flow for nanometer technologies," *Proceedings of the IEEE*, Vol. 89, No. 4, pp. 505-528, April 2001.

[5] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)- Part I: derivation and validation," *IEEE Transactions on Electron Devices,* Vol. 45, no.3, pp. 580-589, March 1998.

[6] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)- Part II: Applications to clock frequency, power dissipation, and chip size estimation," *IEEE Transactions on Electron Devices,* Vol. 45, no.3 March 1998.

[7] J. A. Davis, et. al., "Interconnect limits on gigascale integration (GSI) in the 21st century," *Proceedings of the IEEE,* Vol. 89, no.3, pp. 305-324. March 2001.

[8] P. Fisher and R. Nesbitt, "The test of time. Clock-cycle estimation and test challenges for future microprocessors," *IEEE Circuits and Devices Magazine*, pp.37-44, March 1998.

[9] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," *Proceedings of the IEEE*, Vol. 89, No. 4, pp. 490-504, April 2001.

[10] P. Kapur, J. P. McVittie, and K.C. Saraswat, "Technology and reliability constrained future copper interconnects-part I: Resistance modeling," *IEEE Transactions on Electron Devices*, Vol. 49, No. 4, pp. 590-597, April 2002

[11] P. Kapur, G. Chandra, J. P. McVittie and K. C. Saraswat, "Technology and reliability constrained future copper interconnects-part II: Performance implications," *IEEE Transactions on Electron Devices*, Vol. 49, no. 4, pp. 598-604, April 2002

[12] R. H. J. M. Otten and R. K. Brayton, "Planning for performance," *Proceedings of 35th Annual Design Automation Conference (DAC),* 1998, pp. 122-127.

[13] G. A. Sai-Halasz, "Performance trends in high-end processors," *Proceedings of the IEEE*, Vol.83, No. 1, pp. 20-36, January 1995.

[14] V. Swerdlov, Y. Naveh, and K. Likharev, "Nanoscale SOI Ballistic MOSFETs: An impending Power Crisis," *IEEE International SOI Conference,* pp. 151, 2001.

[15] D. Sylvester, and K. Keutzer, "Impact of small process geometries on microarchitectures in systems on a chip," *Proceedings of the IEEE*, Vol. 89, No. 4, pp. 467-489, April 2001.

[16] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron," *Proceedings of International Conference on Computer-Aided Design,* pp. 203-211, November 1998.