Power Dissipation Issues in Interconnect Performance Optimization for Sub-180 nm Designs

Kaustav Banerjee
Center for Integrated Systems, Stanford University
Stanford CA 94305
kaustav@ee.stanford.edu

Amit Mehrotra
Coordinated Science Lab, University of Illinois at Urbana-Champaign
Urbana IL 61801
amehrotr@uiuc.edu

Abstract

This paper addresses the problem of power dissipation during the buffer insertion phase of interconnect performance optimization. It is shown that the interconnect delay is actually very shallow with respect to both the repeater size and separation close to the minimum point. A methodology is developed to calculate the repeater size and inter-buffer interconnect length which minimizes the total interconnect power dissipation for any given delay penalty. This methodology is used to calculate the power-optimal buffering schemes for various ITRS technology nodes for 5% delay penalty. Furthermore, this technique is also used to quantify the relative importance of the various components of the power dissipation for power-optimal solutions for various technology nodes.

1 Introduction

As VLSI circuits continue to be scaled aggressively past the 180 nm technology node, performance of these ICs is being increasingly dominated by the interconnects [1, 2]. With technology scaling, more and more functionality is being integrated on-chip which results in an increase in the die size in spite of the reduction in minimum feature size. As a result, the number of long global lines and the length of these global lines increases with technology scaling. Since the delay of a long unbuffered line is quadratic in its length, long interconnects are divided into a number of segments with repeaters or buffers. The delay of an optimally buffered line is linear in its length [3]. However, for large high-performance designs, the number of such repeaters can be prohibitively high [4] and can take up significant fraction of active Silicon and routing area [2]. Additionally, as the total chip capacitance (dominated by interconnect capacitance), operating frequency and leakage current increases with scaling, total chip power dissipation is increasing rapidly [1]. A significant fraction of the total chip power dissipation arises due to the loading caused by long global- and semi-global-tier interconnect networks, especially in high-performance designs. For example, it has been reported that around 40%-70% of the total power consumption could be due to the clock distribution network [5, 6].

In general, the repeaters are optimally sized and separated to minimize the interconnect delay. However, since these optimally sized repeaters are quite large (~450 times the minimum sized inverter available in the relevant technology for global-tier lines [7]) and also dissipate a significant amount of power, the total power dissipation by such repeaters in large high-performance designs can be prohibitively high. However, as shown in Figure 1, the interconnect delay is actually very shallow with respect to both the repeater size and separation close to the minimum point. Since, all global interconnects are not on the critical path, a small delay penalty can be tolerated on these noncritical interconnects and there exists a potential for large power savings by using smaller repeaters and larger inter-repeater interconnect lengths. In this work, we develop a methodology to estimate the repeater size and inter-repeater interconnect length which minimizes the total interconnect power dissipation for a given delay penalty. We use this methodology to find the power-optimal buffering schemes for various ITRS technology nodes for a given delay penalty. Furthermore, we use this methodology to show the relative importance of the various components of the power dissipation for various technology nodes. We show that for a given delay penalty, the relative power saving increases as the technology scales. This is shown to be due to the fact that leakage power dissipation becomes the dominating component of the total power dissipation, and therefore reducing the repeater and the number of repeaters results in large power savings.

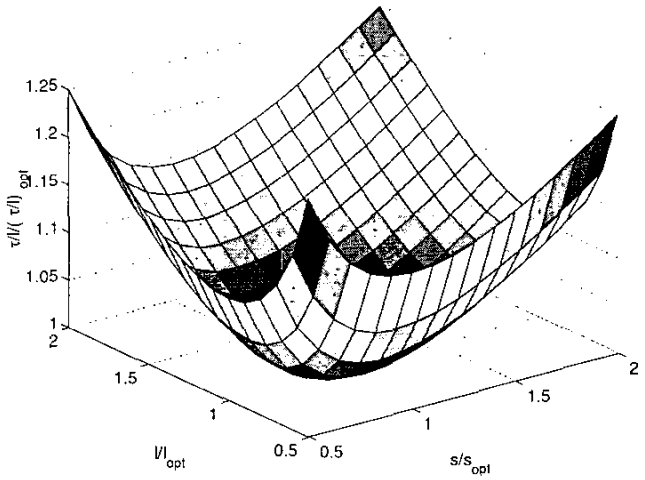


Figure 1: Normalized delay per unit length as a function of buffer size and interconnect length for 180 nm top layer metal.

2 Preliminaries

Consider a uniform interconnect of resistance r per unit length and capacitance c per unit length buffered by identical repeaters as shown in Figure 2. Assume that for a minimum sized repeater, the input capacitance is c_0, the output parasitic capacitance is c_p and output resistance is r_s. Therefore for a repeater of size s, the total output resistance $R_{tr} = \frac{r_s}{s}$, the total output parasitic capacitance $C_p = c_p s$ and the total input capacitance is $C_L = c_0 s$. If the line segment is of length l and the repeater size is s, then the delay of that segment is [3]

$$\tau = r_s(c_0 + c_p) + \frac{r_s}{s}cl + rlsc_0 + \frac{1}{2}rcl^2$$

and the delay per unit length is given by

$$\frac{\tau}{l} = \frac{1}{l}r_s(c_0 + c_p) + \frac{r_s}{s}c + rsc_0 + \frac{1}{2}rcl$$

This delay per unit length is optimal when [3]

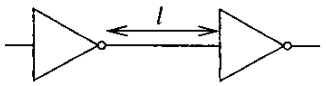
$$l_{opt} = \sqrt{\frac{2r_s(c_0 + c_p)}{rc}} \quad s_{opt} = \sqrt{\frac{r_s c}{rc_0}}$$

and

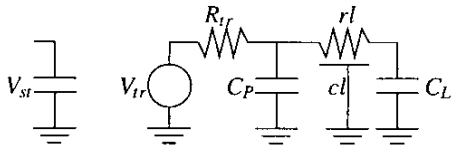
$$\left(\frac{\tau}{l}\right)_{opt} = 2\sqrt{r_s c_0 rc}\left(1 + \sqrt{\frac{1}{2}\left(1 + \frac{c_p}{c_0}\right)}\right)$$

It is widely believed that the total power dissipation due to optimum repeater insertion scheme can be excessive. As shown in Figure 1 the

 0-7803-7310-3/02/$17.00 ©2002 IEEE



(a) Schematic representation.



(b) Equivalent *RC* circuit.

Figure 2: Interconnect of length l between two identical inverters.

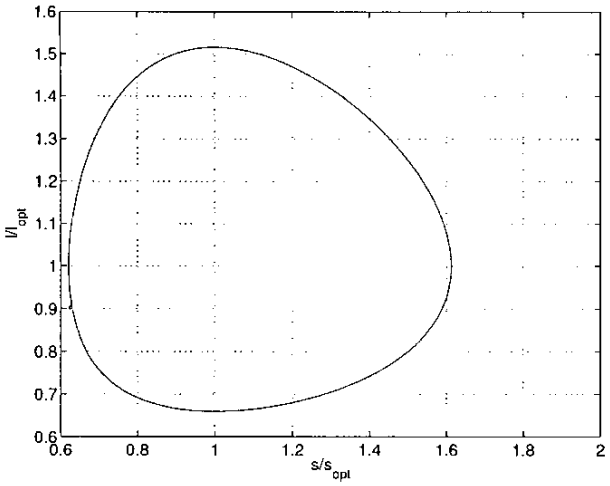


Figure 3: Set of s/s_{opt} and l/l_{opt} values for which $\frac{\tau}{l} = 1.05\left(\frac{\tau}{l}\right)_{opt}$.

minima of $\frac{\tau}{l}$ is very shallow both with respect to s and l. For this example, if the repeater size is $\frac{1}{2}s_{opt}$ and the interconnect length is $2l_{opt}$, the delay penalty is only 25%. Therefore in practice the repeater size is smaller than s_{opt} and the interconnect length is larger than l_{opt} in the hope that power dissipation of such a configuration will be small with minimal impact on delay.

We would therefore like to quantify the reduction in power dissipation when repeater sizes smaller than s_{opt} and interconnect lengths larger than l_{opt} are used for a fixed delay penalty. It is obvious from Figure 1 that for a given value of $\frac{\tau}{l} > \left(\frac{\tau}{l}\right)_{opt}$, there is a family of values of s and l which satisfy this equation which would be the closed curve formed by the intersection of the surface of solutions in Figure 1 with a plane parallel to the s-l axis. As an illustration, Figure 3 shows the set of solutions for which $\frac{\tau}{l} = 1.05\left(\frac{\tau}{l}\right)_{opt}$, i.e., a delay penalty of 5%. From this family of solutions we would like to select the one which gives the minimum total power dissipation for the line.

For a long interconnect of length L which is buffered several times the total power dissipation is

$$P_{line} = nP_{repeater} = \frac{L}{l}P_{repeater}$$

where $n = \frac{L}{l}$ is the number of repeaters for that line. For a fixed L, we therefore seek to minimize $\frac{P_{repeater}}{l}$ in order to minimize the total power dissipation.

Figure 4 shows the power dissipation per unit interconnect length for the curve shown in Figure 3. It is obvious from this figure that a optimum value of repeater size s and inter-repeater interconnect length l exists for which the delay penalty criteria is met and power dissipation is minimum.

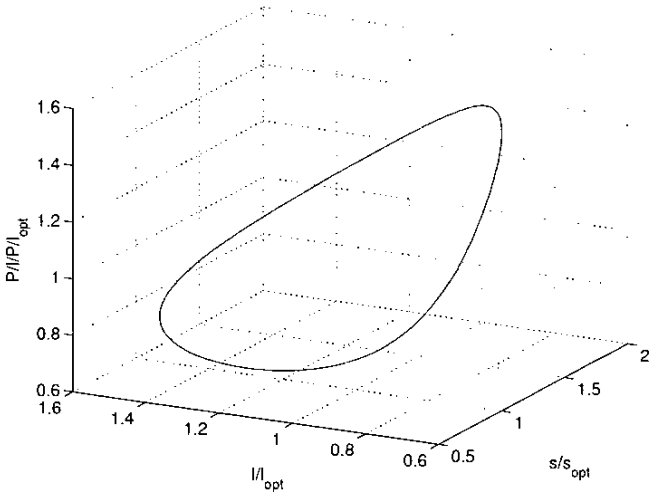


Figure 4: Normalized power dissipation per unit length for a 5% delay penalty as a function of s/s_{opt} and l/l_{opt}.

3 Methodology

The power dissipation of a repeater is given by [8]

$$P_{repeater} = P_{switching} + P_{short\ circuit} + P_{leakage}$$

The various components of the total power are expressed as follows:

Switching Power

The switching power is given by

$$P_{switching} = \frac{1}{2}\alpha(s(c_p + c_0) + lc)V_{DD}^2 f_{clk}$$

where V_{DD} is the power supply voltage, f_{clk} is the clock frequency and α is the switching factor which can be taken as 0.15 [8]. Note that as the repeater size is reduced and the inter-buffer interconnect length is increased, for a given line, the intrinsic repeater power dissipation reduces whereas the switching power due to line capacitance remains unchanged.

Leakage Power

The leakage power is given by

$$P_{leakage} = V_{DD}I_{leakage} = V_{DD}I_{off}W_n = V_{DD}I_{off}W_{n_{min}}s$$

where $I_{leakage}$ is the leakage current flowing through the repeater, I_{off} is the leakage current per unit transistor width, W_n is the width of the NMOS transistor and $W_{n_{min}}$ is the width of the NMOS transistor in minimum sized inverter. For long channel devices, I_{off} used to be negligible but for deep sub-micron technologies this can be significant. I_{off} values of NMOS transistors for all technologies is given in Table 1. Note that as the repeater size is reduced and the inter-buffer interconnect length is increased, the leakage power of one repeater decreases, as well as the total number of repeaters inserted along the line decreases. Therefore, this results is large savings in leakage power dissipation.

Short-Circuit Power

The short circuit power is given by

$$P_{short\ circuit} = \alpha t_r V_{DD} I_{peak} f_{clk} = \alpha t_r V_{DD} W_{n_{min}} s I_{short\ circuit} f_{clk}$$

where I_{peak} is the peak short circuit current and t_r is the time for the input voltage to rise from V_{t_n} to $V_{DD} - V_{t_p}$ [8]. It has been empirically observed from SPICE simulations that $I_{peak} = W_{n_{min}} s I_{short\ circuit}$ where $I_{short\ circuit}$ is approximately 65 μA/μm across all technologies. Assuming that the input waveform is a single time-constant exponential and $V_{t_n} = V_{t_p} = \frac{1}{4}V_{DD}$,

$$t_r = \tau \log_e\left(\frac{V_{DD} - V_{t_p}}{V_{t_n}}\right) = \tau \log_e 3$$
$$= \left[r_s(c_0 + c_p) + \frac{r_s}{s}cl + rlsc_0 + \frac{1}{2}rcl^2\right]\log_e 3$$

Note that as the repeater size is reduced and the inter-buffer interconnect length is increased, the rise time t_r increases and therefore the short circuit power dissipation for one repeater may increase.

Therefore the total power can be written as:

$$P_{repeater} = k_1(s(c_p + c_0) + lc) + k_2 s + k_3 s\tau$$

where

$$k_1 = \frac{1}{2}\alpha V_{DD}^2 f_{clk}$$
$$k_2 = V_{DD} I_{off} W_{n_{\min}}$$
$$k_3 = \alpha V_{DD} W_{n_{\min}} I_{short\ circuit} f_{clk} \log_e 3$$

If the fractional delay penalty to be tolerated is f, then

$$\frac{\tau}{l} = (1+f)\left(\frac{\tau}{l}\right)_{opt} = \frac{1}{l} r_s(c_0 + c_p) + \frac{r_s}{s} c + r s c_0 + \frac{1}{2} r c l \qquad (1)$$

or

$$\tau = (1+f)\left(\frac{\tau}{l}\right)_{opt} l$$

Therefore

$$\begin{aligned} P_{repeater} &= k_1(s(c_p + c_0) + lc) + k_2 s + k_3(1+f)\left(\frac{\tau}{l}\right)_{opt} sl \\ &= k_1(s(c_p + c_0) + lc) + k_2 s + k_3' sl \end{aligned} \qquad (2)$$

where

$$k_3' = k_3(1+f)\left(\frac{\tau}{l}\right)_{opt}$$

and

$$\frac{P_{repeater}}{l} = k_1\left(\frac{s}{l}(c_p + c_0) + c\right) + k_2\frac{s}{l} + k_3' s \qquad (3)$$

Setting the derivative of this with respect to s to zero we have,

$$\frac{d\frac{P_{repeater}}{l}}{ds} = \frac{k_1(c_p + c_0)}{l} + \frac{k_2}{l} + k_3' - \left[\frac{k_1 s(c_p + c_0)}{l^2} + \frac{k_2 s}{l^2}\right]\frac{dl}{ds} = 0$$

$\frac{dl}{ds}$ can be calculated by differentiating (1). Therefore we have the following three nonlinear equations to solve:

$$\begin{aligned} &\frac{k_1(c_p + c_0)}{l} + \frac{k_2}{l} + k_3' - \left[\frac{k_1 s(c_p + c_0)}{l^2} + \frac{k_2 s}{l^2}\right]\frac{dl}{ds} = 0 \\ &\frac{1}{l} r_s(c_0 + c_p) + \frac{r_s}{s} c + r s c_0 + \frac{1}{2} r c l - (1+f)\left(\frac{\tau}{l}\right)_{opt} = 0 \\ &\left[\frac{1}{2} rc - \frac{r_s(c_0 + c_p)}{l^2}\right]\frac{dl}{ds} + r c_0 - \frac{r_s c}{s^2} = 0 \end{aligned} \qquad (4)$$

with three unknown l, s and $\frac{dl}{ds}$ out of which we are only interested in l and s. This can be solved numerically using Newton-Raphson.

4 Results

The methodology outlined above was used to optimize power for global tier interconnects for ITRS technology nodes for a 5% delay penalty as an illustrative example. The ITRS technology parameters are shown in Table 1. r_s, c_0, c_p and $I_{short\ circuit}$ were obtained by SPICE simulations. I_{off} at 100°C was taken to be 0.2 μA/μm for the 180 nm technology node [9] and was estimated for other technology nodes using a subthreshold swing of 100 mV/decade at that temperature [9].

The power optimization results are shown in Table 2. s/s_{opt} is the ratio of the new repeater size with respect to the delay optimal repeater size, l/l_{opt} is the ratio of the new interconnect length between successive repeaters with respect to the delay optimal interconnect length, P/P_{opt} is the ratio of the power dissipation of a *single* repeater with respect to the power dissipation of the delay optimal repeater and $\frac{P}{l}/\left(\frac{P}{l}\right)_{opt}$ is the ratio of the power dissipation per unit length with respect to the power dissipation per unit length of the delay optimal case. From the table, it is obvious that for optimal power dissipation for a given delay penalty, the repeater size needs to be reduced and the interconnect length between successive repeaters needs to be increased. Additionally, the total power savings increase as the technology scales. This is due to that fact that leakage current I_{off} increases substantially with scaling and therefore reducing the repeater size results in large savings in total power dissipation.

Tech. node (nm)	180	130	100	70	50
width (nm)	525	382.5	280	195	137.5
height (nm)	1155	1033	756	546	399
t_{ins} (nm)	7699	6664	6022	5571	4116
ε_r	3.75	3.1	1.9	1.5	1.25
r (kΩ/m)	36.3	60.1	103.9	206.6	401.3
c (pF/m)	269	240	154	125	106
l_{opt} (mm)	3.33	2.5	2.22	1.32	1.06
s_{opt}	174	151	110	82	53
$\left(\frac{\tau}{l}\right)_{opt}$ (ps/mm)	49.5	58.8	56.3	67.4	67.0
r_s (kΩ)	8	9.5	10	15.8	12.5
c_0 (fF)	1.9	1.7	1.5	1.3	1.2
c_p (fF)	4.8	3.5	2.5	1.5	0.75
V_{DD} (V)	1.8	1.5	1.2	0.9	0.6
V_t (V)	0.45	0.375	0.3	0.225	0.15
I_{off} (μA/μ)	0.2	1.13	6.33	35.6	200
f_{clk} (GHz)	1.2	1.6	2.0	2.5	3.0

Table 1: Technology and equivalent circuit model parameters for top layer metal for different technology nodes based on the ITRS. c was obtained using FASTCAP [10].

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{l}/\left(\frac{P}{l}\right)_{opt}$
180	0.6673	1.2512	0.9091	0.7266
130	0.6756	1.2634	0.9187	0.7272
100	0.6898	1.2915	0.9008	0.6974
70	0.7113	1.3337	0.8275	0.6204
50	0.7386	1.3877	0.7669	0.5526

Table 2: Power per unit length optimization results for 5% delay penalty for various ITRS technology nodes.

This fact is further illustrated in Figure 5 which plots the relative contributions of $P_{switching}$, $P_{short\ circuit}$ and $P_{leakage}$ as the technologies scale. It can be observed that leakage power starts dominating as the technology scales. Also note that the short circuit power is also nontrivial across all technology nodes. Therefore short circuit power needs to be considered in any power optimization.

With this basic framework, various power optimization alternatives can be compared. For instance, a naïve approach would be to minimize the power dissipation of individual repeaters instead of minimizing the repeater power per unit length. For this case, (2) needs to be used instead of (3) in the set of nonlinear equations (4). The results of this optimization are shown in Table 3. Comparing these results with Table 2 we observe that if power dissipation of one inverter is minimized, the *power-optimal* inter-repeater interconnect length l is *smaller* than the delay optimal length l_{opt}. Therefore, even though the power dissipation of one repeater is smaller than that in Table 2 (column 4), since the number of repeaters for a given line length is larger for this case, the total power dissipation (or equivalently power dissipation per unit length) (column 5) is higher than that in Table 2.

Similarly, the effect of ignoring short circuit power and leakage power on the optimization can be quantified. Table 4 shows the optimization results considering only the switching component of the power dissipation.

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{l}/\left(\frac{P}{l}\right)_{opt}$
180	0.6827	0.7766	0.6985	0.8994
130	0.6888	0.7723	0.7003	0.9069
100	0.6769	0.7948	0.6995	0.8802
70	0.6531	0.8721	0.6815	0.7814
50	0.6312	0.8923	0.6112	0.6914

Table 3: Power minimization results for 5% delay penalty for various ITRS technology nodes.

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{T}/\left(\frac{P}{T}\right)_{opt}$
180	0.6967	1.3160	0.9628	0.7316
130	0.7053	1.3287	0.9840	0.7406
100	0.7134	1.3415	1.0049	0.7491
70	0.7257	1.3625	1.0382	0.7620
50	0.7412	1.3926	1.0838	0.7782

Table 4: Switching power per unit length minimization results for 5% delay penalty for various ITRS technology nodes.

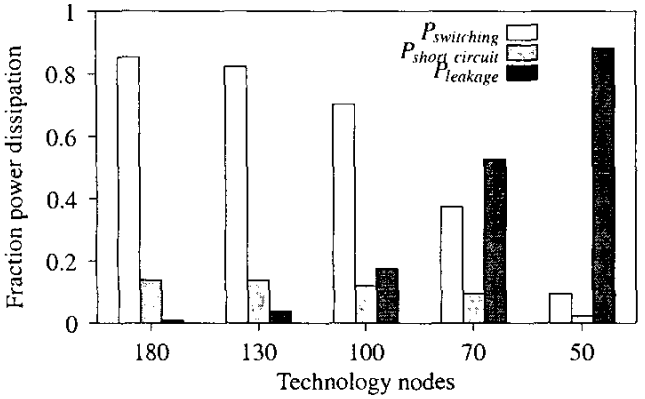


Figure 5: Relative contributions of the three components of overall power dissipation for 5% delay penalty for various technology nodes

Similarly, Table 5 shows the optimization results considering only the switching and short circuit component of the power dissipation and Table 6 shows the optimization considering only the switching and leakage component of the power dissipation. It can be observed from these tables that ignoring leakage power results in large errors in power optimization at future technology nodes. Similarly, ignoring short circuit power also results in errors throughout the optimization process.

Figure 6 shows the normalized power per unit length as a function of delay penalties for various technology nodes. As expected, $\frac{P}{T}/\left(\frac{P}{T}\right)_{opt}$ reduces as the delay penalty increases. Note that the incremental reduction in $\frac{P}{T}/\left(\frac{P}{T}\right)_{opt}$ is high for small values of delay penalty and starts decreasing as the delay penalty increases. Also note that the curves for 180 nm and 130 nm technology nodes are very similar. However, for a given delay penalty, $\frac{P}{T}/\left(\frac{P}{T}\right)_{opt}$ reduces as the technology is scaled beyond 130 nm. This is entirely due the leakage power. From Figure 5 it can be observed that for both 180 nm and 130 nm technology nodes, leakage power is a negligible portion of the overall power dissipation whereas for other technology nodes, it becomes progressively significant and is the dominant fraction of total power dissipation for the 50 nm technology node.

5 Conclusions

In conclusion, we have developed a methodology for choosing the repeater size and inter-repeater interconnect length for a given line length which satisfies a given delay penalty criteria and minimizes the total power dissipation. Using this methodology, we have computed the power-optimal buffering schemes for various technology nodes for a 5% delay penalty. Furthermore, we have shown that short-circuit and leakage power are important components of the total power dissipation and

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{T}/\left(\frac{P}{T}\right)_{opt}$
180	0.6668	1.2499	0.9111	0.7290
130	0.6733	1.2573	0.9276	0.7377
100	0.6795	1.2656	0.9442	0.7460
70	0.6797	1.2525	0.9468	0.7560
50	0.6952	1.2836	0.9923	0.7731

Table 5: Switching and short circuit power per unit length minimization results for 5% delay penalty for various ITRS technology nodes.

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{T}/\left(\frac{P}{T}\right)_{opt}$
180	0.6967	1.3160	0.9591	0.7287
130	0.7053	1.3287	0.9671	0.7278
100	0.7134	1.3415	0.9289	0.6924
70	0.7257	1.3625	0.8294	0.6088
50	0.7412	1.3926	0.7648	0.5492

Table 6: Switching and leakage power per unit length minimization results for 5% delay penalty for various ITRS technology nodes.

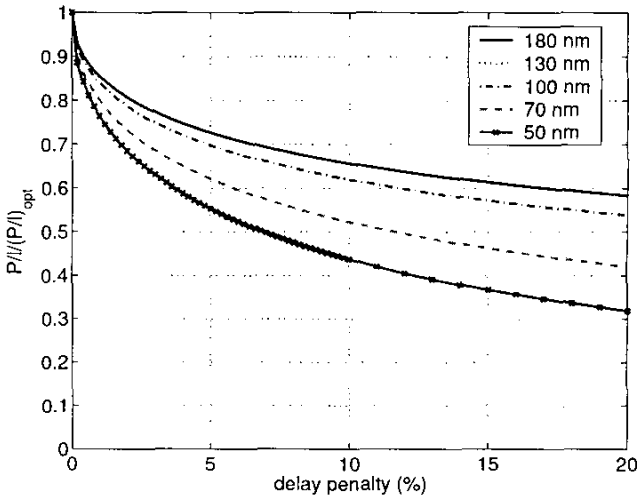


Figure 6: Power per unit length as a function of delay penalty for various technology nodes.

ignoring them in the power optimization process can lead to large errors. Short-circuit power becomes important as the allowed delay penalty increases since rise time of the signal increases. Similarly, leakage power increases exponentially with device scaling and is the dominant component of power dissipation for 50 nm technology node. We have also shown that for 180 nm and 130 nm technology nodes where leakage power is not significant, the relative power saving is almost the same for a given delay penalty. However, beyond 130 nm node, leakage power becomes significant and therefore the relative power savings increase with technology scaling for a given delay penalty.

References

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

[2] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proceedings of the IEEE*, vol. 89, pp. 602–633, May 2001.

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

[4] J. Cong and L. He, "An efficient technique for device and interconnect optimization in deep submicron designs," in *Proceedings International Symposium on Physical Design*, pp. 45–51, 1998.

[5] H. Kawaguchi and T. Sakurai, "A reduced clock swing flip-flop (RCFF) for 63% power reduction," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 807–811, 1998.

[6] T. Sakurai, "Design challenges for 0.1 μm and beyond," in *Proceedings ASP DAC*, pp. 553–558, 2000.

[7] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," in *Proceedings 1999 Design Automation Conference*, pp. 885–891, 1999.

[8] A. P. Chandrakasan and R. W. Brodersen, *Low power digital CMOS design*, ch. Sources of Power Consumption. Boston : Kluwer Academic Publishers, 1995.

[9] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proceedings 1999 International Symposium on Low Power Electronics and Design*, pp. 163–168, 1999.

[10] K. Nabors and J. K. White, "FASTCAP: a multipole-accelerated 3-D capacitance extraction program," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 10, pp. 1447–1459, Nov. 1991.