Analog Integrated Circuits and Signal Processing, 35, 97–105, 2003
© 2003 Kluwer Academic Publishers. Manufactured in The Netherlands.

An Interconnect Scaling Scheme with Constant On-Chip Inductive Effects

KAUSTAV BANERJEE[1] AND AMIT MEHROTRA[2]

[1]Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106
[2]Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana IL 61801
E-mail: kaustav@ece.ucsb.edu; amehrotr@uiuc.edu

Abstract. This paper introduces a new global-tier interconnect scaling scheme which ensures that inductance effects do not start dominating the overall interconnect performance. It is shown that for unscaled global lines, inductance effects increase as technology scales while for the scaling scheme proposed by ITRS [1], interconnects become extremely resistive and, while inductance effects diminish with scaling but the performance, specifically, delay per unit length, degrades with scaling. The effect of the proposed global interconnect scaling scheme on optimized driver size, interconnect length, delay per unit length and total buffer area is quantified and compared with the unscaled and the ITRS cases. It is shown that the proposed scaling scheme improves the delay per unit length without degrading inductive effects or increasing buffer area with scaling.

I. Introduction

For deep submicron technologies using copper, on-chip inductive effects are a concern for signal integrity and overall chip performance [2]. By inductive effects we refer to changes in the interconnect waveforms which cannot be predicted by considering the interconnect as a purely resistive-capacitive element. These include overshoots and undershoots in the waveforms due to improper termination, inductive coupling, crosstalk and oscillations. Inductance effects in global interconnects are more severe due to the lower resistance per unit length of these lines which results in the reactive component of the interconnect impedance to become comparable to the resistive component, and also due to significant mutual inductive coupling between interconnects resulting from longer current return paths. With the recent adoption of Copper as the interconnect metal [3,4], line resistances have decreased further and as a result, inductive effects have become more prominent. Additionally, inductive cross talk is being considered as the most serious concern in VLSI interconnects, specially for large busses. Hence inductance effects must be considered in the scaling schemes.

In the past a lot of effort has been devoted to the area of inductance computation and extraction. See [5] for a more complete review of existing literature in this area. However, accurate inductance modelling still remains a challenging problem. This is due to the fact that magnetic fields have much longer spatial range compared to that of electric fields and therefore, in practical high-performance ICs containing several layers of densely packed interconnects the wire inductances are sensitive to even distant variations in the interconnect topology and switching activity. Moreover, uncertainties in the termination of neighbouring wires can significantly affect the signal return path and also the return current distributions, and therefore the effective loop inductance and resistance. Moreover, since global wires are the farthest from the substrate, they are more susceptible to large variations in current return path and therefore large variations in the inductance.

Present interconnect scaling schemes do not consider inductance. As it will be shown later, inductance effects start dominating with technology scaling when global lines are not scaled, while they start diminishing with scaling for ITRS where global-tier interconnect dimensions are scaled too aggressively (Table 1). The drawback of the ITRS approach is that the interconnect delay increases since the line resistance per unit length dramatically increases with scaling.

Table 1. Technology parameters for top layer metal for different technology nodes as per ITRS.

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
$l_{\max}$ (nH/mm)	9.2	10.9	13.5	17.9	18.8
$h_{opt_{RC}}$ (mm)	3.33	2.5	2.22	1.32	1.06
$k_{opt_{RC}}$	174	151	110	82	53
$\tau_{opt_{RC}}$ (ns)	0.165	0.147	0.125	0.089	0.071
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

This paper introduces a new scaling scenario where the global tier metal lines are scaled such that inductive effects remain approximately constant. The motivation behind such an approach is that current design tools and methodologies are able to satisfactorily cope with inductive effects in current technologies (180 nm and 130 nm) but may not be able to do so if inductive effects increase with scaling. If on the other hand the line dimensions are scaled too aggressively as per ITRS, it has a deleterious effect on delay per unit length of global interconnects as shown in Section IV which directly results in severe performance penalty. Finally, performance, area increase due to optimal buffering and wireability are compared for the three scaling scenarios.

II. Preliminaries

Consider a uniform line with resistance, capacitance and inductance per unit length of r, c and l respectively, driven by a repeater of series resistance R_S and output parasitic capacitance C_P, and driving an identical repeater with load capacitance C_L (Fig. 1). For a given technology, let the output resistance, output parasitic capacitance and input capacitance of a minimum sized repeater be r_s, c_p and c_0 respectively. Therefore if the repeater size is k times the size of a minimum sized

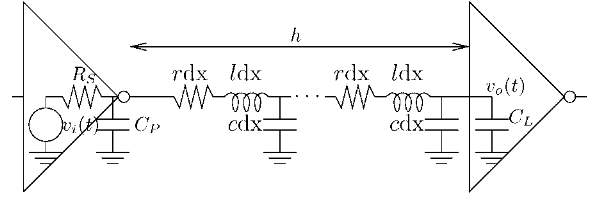


Fig. 1. Equivalent circuit of a driver-interconnect-load segment. The interconnect is uniform with resistance, capacitance and inductance per unit length of r, c and l respectively.

repeater, $R_S = r_s/k$, $C_P = c_p k$ and $C_L = c_0 k$. The transfer function derivation is outlined here from [5] for completeness. The ABCD parameter matrix for a uniform RLC transmission line of length h are given by [5]:

$$\begin{bmatrix} \cosh(\theta h) & Z_0 \sinh(\theta h) \\ \frac{1}{Z_0}\sinh(\theta h) & \cosh(\theta h) \end{bmatrix}$$

where

$$Z_0 = \sqrt{\frac{r+sl}{sc}}$$

s, is the complex frequency $\jmath\omega$ and

$$\theta = \sqrt{(r+sl)sc}$$

Therefore the ABCD parameter matrix of the configuration in Fig. 1 is given by

$$\begin{bmatrix} 1 & R_S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC_P & 1 \end{bmatrix} \begin{bmatrix} \cosh(\theta h) & Z_0 \sinh(\theta h) \\ \frac{1}{Z_0}\sinh(\theta h) & \cosh(\theta h) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC_L & 1 \end{bmatrix} = \begin{bmatrix} (1+sR_SC_P)[\cosh(\theta h) + sC_L Z_0 \sinh(\theta h)] + \frac{R_S}{Z_0}\sinh(\theta h) + sC_L R_S \cosh(\theta h) & (1+sR_SC_P) Z_0 \sinh(\theta h) + R_S \cosh(\theta h) \\ sC_P[\cosh(\theta h) + sC_L Z_0 \sinh(\theta h)] + \frac{1}{Z_0}\sinh(\theta h) + sC_L \cosh(\theta h) & sC_P Z_0 \sinh(\theta h) + \cosh(\theta h) \end{bmatrix}$$

and the input-output transfer function is given by

$$H(s) = \frac{V_o(s)}{V_i(s)}$$
$$= \frac{1}{[1+sR_S(C_P+C_L)]\cosh(\theta)+\left[\frac{R_S}{Z_0}+sC_LZ_0+s^2R_SC_PC_LZ_0\right]\sinh(\theta)}$$

The step-response of this system is given by $V_o(s) = \frac{1}{s}H(s)$ in the Laplace domain. However, computing the response in the time-domain is analytically intractable. The above transfer function is therefore approximated by a second order Padé approximation as

$$H(s) \approx \frac{1}{1+sb_1+s^2b_2} = \frac{1}{\left(1+\frac{s}{s_1}\right)\left(1+\frac{s}{s_2}\right)} \tag{1}$$

where

$$b_1 = R_S(C_P+C_L)+\frac{rch^2}{2!}+R_Sch+C_Lrh$$
$$b_2 = \frac{lch^2}{2!}+\frac{r^2c^2h^4}{4!}+R_S(C_P+C_L)\frac{rch^2}{2!}$$
$$+(R_Sch+C_Lrh)\frac{rch^2}{3!}+C_Llh+R_SC_PC_Lrh$$

The 50% delay τ is given by

$$0.5-\frac{s_2}{s_2-s_1}\exp(s_1\tau)+\frac{s_1}{s_2-s_1}\exp(s_2\tau)=0$$

This transfer function can be used to calculate the 50% delay [6]. Long VLSI interconnects are typically broken up into buffered segment of equal lengths and driven by identical repeaters. For minimum total delay in these long interconnects, the delay per unit length in the optimally buffered segment should be minimized. The driver size k and interconnect length h can be numerically optimized to give minimum delay per unit length [6,7].

A second order Padé expansion of the transfer function is sufficiently accurate if the system is not strongly underdamped. If that is the case, a higher order Padé expansion can be used [5] for higher accuracy. The advantage of using a second order expansion is that the concept of critical inductance can be conveniently defined and used to define inductive effects. Hence in this work we use only a second order Padé expansion. The second order transfer function given by (1) and discussed in [6,7] can be *critically damped*, *overdamped* and *underdamped* when $b_1^2-4b_2$ is equal to, greater than, or less than zero respectively. The response of an overdamped system is very similar to an RC line whereas for an underdamped system, the behaviour is significantly different from an RC line, i.e., inductive effects are significant. Since b_1 and b_2 are functions of h and k and b_2 is a function of l, it has been shown [6] that for optimum values of h and k where interconnect delay is minimum for a given line inductance, a value l_{crit} can be obtained for which the system will be critically damped [6]. If line inductance is less than l_{crit}, the system will be overdamped where as if line inductance is greater than l_{crit}, the system will be underdamped.

$$l_{crit} = \frac{\frac{b_1^2}{4}-\frac{r^2c^2h^4}{4!}-R_S(C_P+C_L)\frac{rch^2}{2!}-(R_Sch+C_Lrh)\frac{rch^2}{3!}-R_SC_PC_Lrh}{\frac{ch^2}{2!}+C_Lh}$$

As pointed out earlier, the self, mutual and loop inductance of an interconnect is not just a function of geometry but also depends on the current return path which is a function of input vectors. Therefore, in this work we consider a conservatively large range of line inductances which include variations in the self, loop and coupling inductance of the interconnect.

Figure 2 plots the critical inductance as a function of l for various technology nodes with unscaled global lines, while Fig. 3 plots the critical inductance as a function of l for various ITRS technology nodes with scaled global lines. Recall that the system is overdamped if $l < l_{crit}$. It can be observed from Fig. 2 that the fraction of line inductance l which is less than l_{crit} decreases if global tier metal lines are not scaled with technology. On the other hand, Fig. 3 shows that l_{crit} increases with technology scaling for ITRS, and $l < l_{crit}$ for an

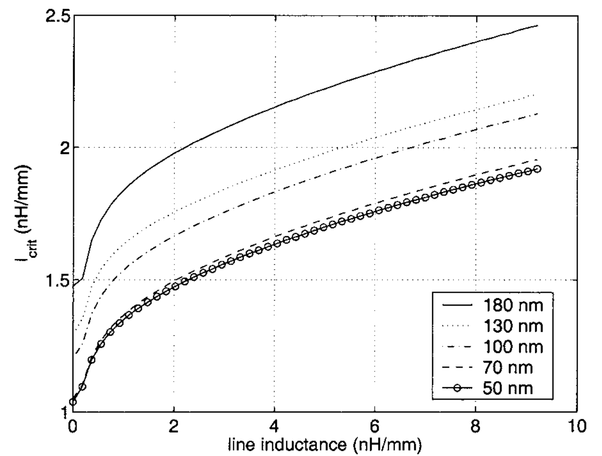


Fig. 2. Critical inductance as a function of line inductance for various technology nodes with unscaled global lines.

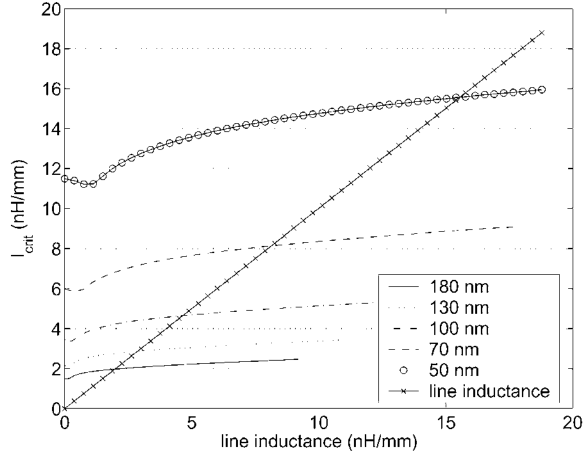


Fig. 3. Critical inductance as a function of line inductance for various ITRS technology nodes.

increasingly large range of l. This implies that inductive effects are diminishing with technology scaling. It can therefore be shown that [7] the range of line inductance (l) values which are less than the corresponding l_{crit} determine whether inductive effects are important for a given interconnect tier in a given technology.

III. Proposed Scaling Scheme

In this work, a new scaling scheme for global tier lines for the ITRS technology nodes is proposed where inductive effects are kept constant with scaling [8]. For simplicity, the aspect ratio of global tier lines is same as that specified by the ITRS.[1] Interconnect pitch is assumed to be twice the interconnect width for all technologies. Interconnect capacitance was extracted using FASTCAP [9]. Repeater resistances and capacitances were extracted from SPICE simulations as follows: A five stage ring oscillator was simulated to compute the delay of each stage. Each identical stage consisted of an inverter followed by an RC interconnect of known resistance and capacitance per unit length. The inverter size and line length was varied to find $h_{opt_{RC}}$ and $k_{opt_{RC}}$ which minimized the stage delay per unit length. With these three simulated quantified ($h_{opt_{RC}}$, $k_{opt_{RC}}$ and $\tau_{opt_{RC}}$) inverter series resistance r_s and output and parasitic capacitances (c_0 and c_p) were back calculated. Using these parameters, optimum values of h and k which minimize the RLC delay per unit length for various line inductances l were numerically computed. This in turn yielded the critical inductance

Table 2. Global tier line widths and aspect ratios (line thickness/line width) for constant inductive effect scaling. Aspect ratios are the same as suggested by ITRS roadmap.

Tech. (nm)	Width (nm)	Aspect Ratio
180	525	2.2
130	460	2.5
100	425	2.7
70	400	2.8
50	380	2.9

for each value of l. For each technology, global tier line widths were varied until the l_{crit} vs l plot coincided with the 180 nm technology case. The proposed line widths and aspect ratios are shown in Table 2. It can be noted from Table 2 that interconnect widths reduce as the technology scales but not aggressively as the ITRS scaling. Additionally, since the interconnect dimensions are comparable to the mean free path of electrons in Cu, surface scattering starts having a non-negligible contribution to the resistivity compared to the contribution due to bulk scattering [10,11]. Therefore the resistivity of Cu interconnects needs to be appropriately increased. According to [10] the resistivity of a thin film of metal ρ can be expressed in terms of bulk resistivity ρ_0 as

$$\frac{\rho_0}{\rho} = 1 - \frac{3}{2k}(1-p)\int_1^{\infty}\left(\frac{1}{x^3} - \frac{1}{x^5}\right)\frac{1-e^{-kx}}{1-pe^{-kx}}\,dx$$

where $k = d/\lambda_{mfp}$, d is the smallest dimension of the film (in our case, the width), λ_{mfp} is the bulk mean free path of electrons and p is fraction of electrons which are elastically reflected at the surface. For Copper, $p = 0.47$ and $\lambda_{mfp} = 421$Å at 0°C [11]. Furthermore, since the thickness of the barrier material for Copper interconnects is not scaling with technology scaling, the effective area through which the current conduction takes place is reducing, therefore the effective resistivity of the interconnect is increasing [12]. The cumulative increase in resistivity of global tier lines due to these two effects for ITRS technology nodes is shown in Table 3. Similarly, the effective increase in resistivity for the metal line widths shown in Table 2 is also calculated. These modified resistivity values are used throughout the paper.

Figure 4 plots the critical inductance (l_{crit}) for global tier lines for various technology nodes with the proposed scaling. It is clear that critical inductance values for all the technology nodes are almost same for a given

Table 3. Resistivity ratios for the global tier metals for various ITRS technologies. All dimensions are in nm. Barrier thickness of 10 nm assumed for all technology nodes.

Tech.	w	$\frac{\rho}{\rho_0}$ (thin-film)	$\frac{\rho}{\rho_0}$ (barrier)	$[\frac{\rho}{\rho_0}]_{eff}$
180	525	1.0162	1.0487	1.0657
130	382.5	1.0224	1.0663	1.0902
100	280	1.0308	1.0914	1.1250
70	195	1.0448	1.1351	1.1859
50	137.5	1.0646	1.2003	1.2779

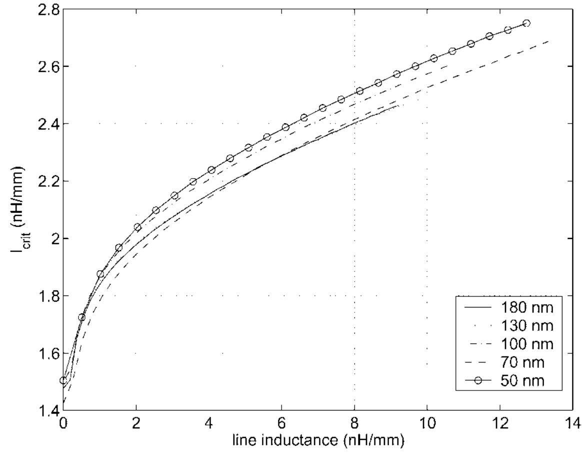


Fig. 4. Critical inductance as a function of line inductance for various ITRS technology nodes with the proposed global tier interconnect scaling scheme.

value of line inductance over a very large range of line inductances. This implies that inductive effects are also very similar for all the technology nodes with the proposed scaling.

IV. Performance Comparisons

We now compare optimized interconnect performance of the proposed scaling scheme for global wires with the ITRS and the unscaled cases for various technology nodes. The line inductance is fixed at 1 nH/mm. Figure 5 shows the optimum interconnect length between two inverters (h_{opt}) for minimum delay per unit length for the three cases for all technologies. It can be observed that h_{opt} is largest for the unscaled case and smallest for the ITRS case. Also note that h_{opt} decreases sharply with technology scaling for the ITRS case, whereas for the proposed scheme and the unscaled case, it does not change appreciably. Figure 6 shows the optimum driver size for the three cases for all technologies. Once again, a trend similar to the h_{opt} case is observed.

Figure 7 plots the optimized delay per unit length for the three cases for all technology nodes. The delay per unit length is the smallest for unscaled case and is the maximum for the ITRS case. It is instructive to note that for ITRS, the optimum interconnect delay per unit

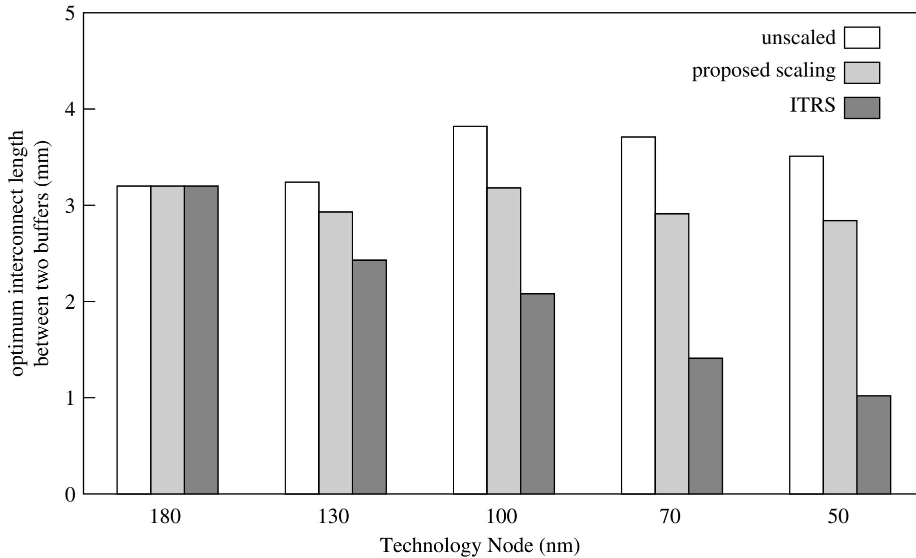


Fig. 5. Optimum interconnect length between two buffers (h_{opt}) for various technology nodes for global tier interconnects which are (a) unscaled, (b) scaled using the proposed scheme in Table 2 and (c) scaled according to ITRS specifications.

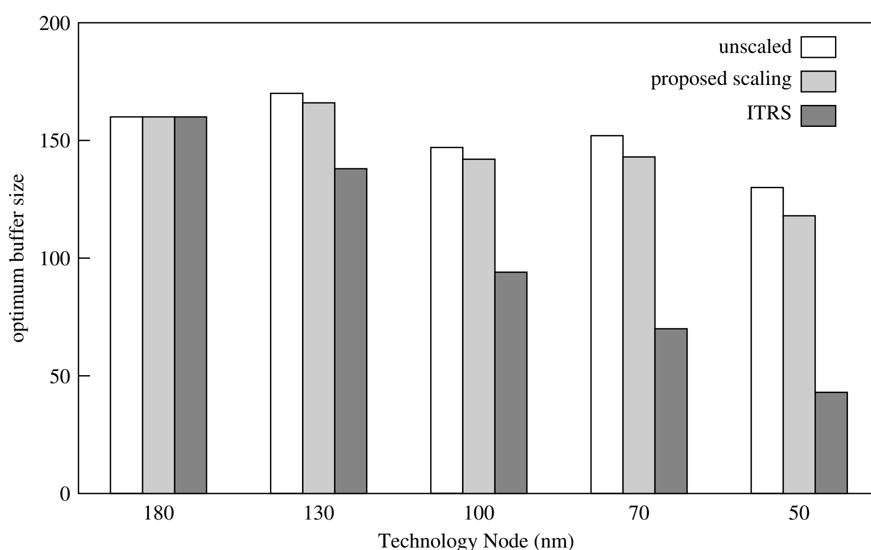


Fig. 6. Optimum buffer size for various technology nodes for global tier interconnects which are (a) unscaled, (b) scaled using the proposed scheme in Table 2 and (c) scaled according to ITRS specifications.

length *increases* with technology scaling. This is due to the fact the global tier interconnect dimensions are scaled very aggressively which dramatically increases line resistance. For the unscaled case and the proposed scaling of global wires, optimum interconnect delay per unit length decreases with technology scaling.

Note that while the optimum buffer size for ITRS reduces with scaling, h_{opt} also reduces which increases the number of buffered segments for an interconnect of a given length. Figure 8 plots the total repeater area (i.e., number of repeaters × area of one repeater) in an optimally buffered 4 cm long global tier interconnect

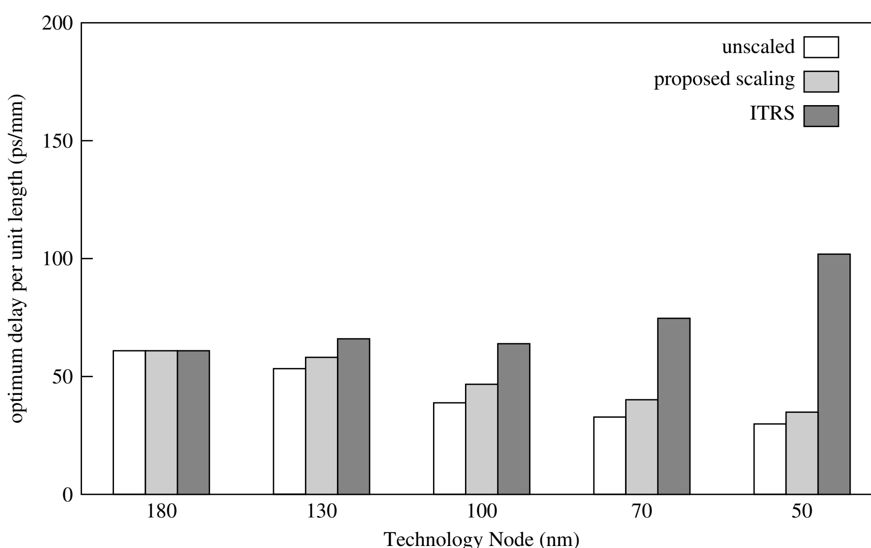


Fig. 7. Optimum delay per unit length for various technology nodes for global tier interconnects, assuming a line inductance of 1 nH/mm, which are (a) unscaled, (b) scaled using the proposed scheme in Table 2 and (c) scaled according to ITRS specifications.

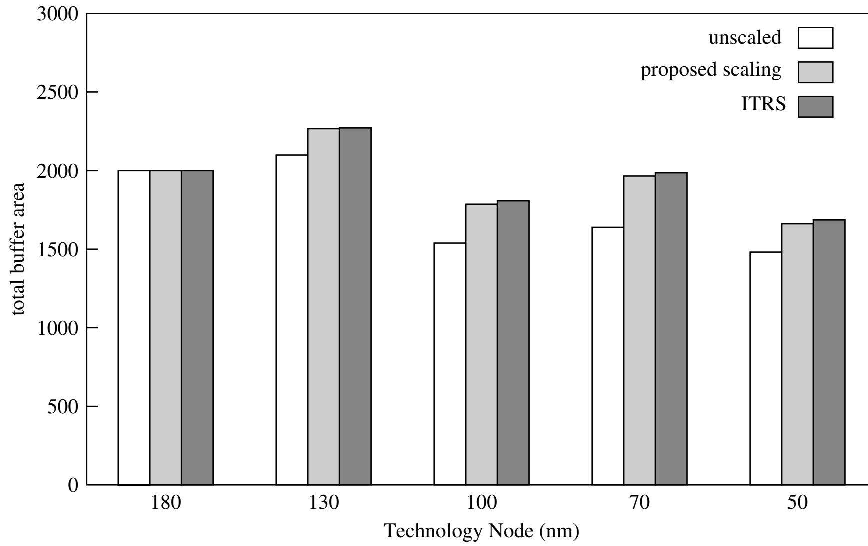


Fig. 8. Total repeater area for an optimally buffered 4 cm global tier interconnect for various technology nodes which is (a) unscaled, (b) scaled using the proposed scheme in Table 2 and (c) scaled according to ITRS specifications. Repeater area is in terms of the area of the minimum sized inverter in the corresponding technology.

for all technologies. It can be observed that total repeater area is smallest for unscaled lines and largest for the ITRS lines. Additionally, the total repeater area in terms of the area of the minimum sized inverter in the corresponding technology does not change appreciably across technologies.

Finally, since the minimum wire width (pitch) for the global lines increases as a result of this scaling methodology, there may be some concern regarding the wireability of the chip without increasing the number of metal layers or chip area. However a recent analysis of chip wireability optimization has shown that the semiglobal and global tiers may not be packed to the maximum [12]. Consequently, higher tiers are actually routed within a larger than required area. Since the wire delay per unit length will decrease with increasing wire width it is also possible to re-route some of the global wires on the semi-global tier without violating the maximum allowable length (or delay) for that tier. Hence, the increased wire width resulting from this scaling methodology can be accommodated both in the global and the semiglobal tiers. Furthermore, this scaling scheme can be applied only to the performance critical paths of a chip where the susceptibility of performance to delay variations due to inductance effects is the highest. For non-critical paths, the ITRS based scheme may be used in order to save routing resources.

V. Conclusions

In conclusion, a new global interconnect scaling scheme for deep sub-micron technology nodes is proposed which ensures that the inductive effects remain constant across technology nodes. It has been pointed out that for unscaled global lines, inductance effects increase as technology scales, while for the scaling scheme proposed by ITRS, interconnect becomes extremely resistive and inductance effects diminish with scaling but the performance, specifically delay per unit length, degrades with scaling. The effect of the proposed global interconnect scaling scheme on optimized driver size, interconnect length, delay per unit length and total buffer area is quantified and compared with unscaled and ITRS cases. It is shown that the proposed scaling scheme improves delay per unit length without changing inductive effects and buffer area consumption with scaling.

Note

1. This can be easily extended to the case where the interconnect thickness is kept the same at ITRS for each technology node and the width is varied.

References

1. "International Technology Roadmap for Semiconductors (ITRS)," 1999.
2. Cheng, C.-K., Lillis, J., Lin, S. and Chang, N., *Interconnect Analysis and Synthesis.* John Wiley & Sons, 1999.
3. Edelstein, D., et al., "Full copper wiring in a sub-0.25 μm CMOS ULSI technology." *International Electron Devices Meeting Digest of Technical Papers,* pp. 773–776, 1997.
4. Rohrer, N., et al., "A 480 MHz RISC microprocessor in a 0.12 μm L_{eff} CMOS technology with copper interconnects." *International Solid-State Circuits Conference Digest of Technical Papers,* pp. 240–241, 1998.
5. Banerjee, K. and Mehrotra, A., "Analysis of on-chip inductance effects for distributed *RLC* interconnects." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 21, pp. 904–915, August 2002.
6. Banerjee, K. and Mehrotra, A., "Analysis of on-chip inductance effects using a novel performance optimization methodology for distributed *RLC* interconnects," in *Proceedings 2001 Design Automation Conference,* pp. 798–803, 2001.
7. Banerjee, K. and Mehrotra, A., "Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling," in *Proceedings 2001 IEEE Symposium on VLSI Circuits,* pp. 195–198, 2001.
8. Banerjee K. and Mehrotra, A., "Inductance aware interconnect scaling," in *Proceedings IEEE International Symposium on Quality Electronic Design,* pp. 43–47, 2002.
9. Nabors, K. and White, J. K., "FASTCAP: a multipole-accelerated 3-D capacitance extraction program." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 10, pp. 1447–1459, November 1991.
10. Anderson, J. C., ed., *The Use of Thin Films in Physical Investigations.* Academic Press, 1966.
11. Chen, F. and Gardner, D., "Influence of line dimensions on the resistance of Cu interconnections." *IEEE Electron Device Letters* 19, pp. 508–510, December 1998.
12. Banerjee, K., Souri, S. J., Kapur, P. and Saraswat, K. C., "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and system-on-chip integration," in *Proceedings of the IEEE* 89, pp. 602–633, May 2001.

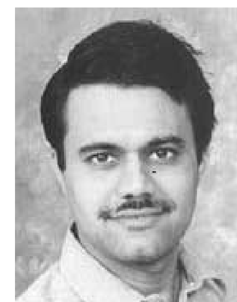
Kaustav Banerjee (S94–M99) received the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley in 1999.

He was with Stanford University, Stanford, CA, from 1999 to 2002 as a research associate at the Center for Integrated Systems. In July 2002, he joined the Faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, as an assistant professor.

His research interests include nanometer scale circuit effects and their implications for high-performance/low-power VLSI and mixed-signal designs and their design automation methods. He is also interested in some exploratory interconnect and circuit architectures such as 3-D ICs, integrated optoelectronics, and nanotechnologies such as single electron transistors. He co-advises several doctoral students at Stanford University, University of Southern California, Los Angeles, and the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

From February 2002 to August 2002 he was a visiting professor at the Circuit Research Labs of Intel in Hillsboro, Oregon. In the past, he has also held summer/visiting positions at Texas Instruments Inc., Dallas, Texas, and EPFL-Switzerland, and has consulted for several EDA companies in the Silicon Valley. He has authored or co-authored over 70 technical papers in archival journals and refereed international conferences and has presented numerous invited talks and tutorials.

Dr. Banerjee served as technical program chair of the 2002 IEEE International Symposium on Quality Electronic Design (ISQED 02), and is the conference vice-chair of ISQED 03. He has also served on the technical program committees of the ACM International Symposium on Physical Design, the EOS/ESD Symposium, and the IEEE International Reliability Physics Symposium. He is the recipient of a Best Paper Award at the 2001 Design Automation Conference.



Amit Mehrotra received B. Tech. degree in electrical engineering from Indian Institute of Technology, Kanpur in 1994 and Masters and Ph.D. from Department of Electrical Engineering and Computer Science,

University of California at Berkeley in 1996 and 1999 respectively.

In August 1999 he joined the University of Illinois at Urbana-Champaign where he is currently an assistant professor with the Department of Electrical and Computer Engineering and a research assistant professor with the Illinois Center for Integrated Micro-Systems group at the Coordinated Science Laboratory. He is a member of the IEEE.

His research interests include RF, analog and mixed signal circuit design, for mobile communication systems, simulation techniques for RF and mixed signal circuits and systems, interconnect performance and modelling issues in VLSI and novel circuits and physical design issues for high performance VLSI designs, model-order reduction of linear and nonlinear circuits. He has authored and coauthored over 30 technical papers in archival journals and refereed international conferences. He has served as the Techincal Program Committee member of International Symposium on Quality Electronic Design in 2002 and 2003. He received best paper awards at the 1997 International Conference on Computer Design and 2001 Design Automation Conference.