Minimum Power and Area N-Tier Multilevel Interconnect Architectures Using Optimal Repeater Insertion

Raguraman Venkatesan, Jeffrey A. Davis, Keith A. Bowman and James D. Meindl

Georgia Institute of Technology
Atlanta, GA 30332-0269, USA
Phone: (404) 894-9910 Fax: (404) 894-0462 E-mail: vragu@ee.gatech.edu

ABSTRACT

Minimum power CMOS ASIC macrocells are designed by minimizing the macrocell area using a new methodology to optimally insert repeaters for n-tier multilevel interconnect architectures. The minimum macrocell area and power dissipation are projected for the 100, 70 and 50 *nm* technology generations and compared with a n-tier design without using repeaters. Repeater insertion and a novel interconnect geometry scaling technique decrease the power dissipation by 58-68% corresponding to a macrocell area reduction of 70-78% for the global clock frequency designs of these three technology generations.

1. INTRODUCTION

As CMOS semiconductor technology progresses towards gigascale integration, the power dissipated by a billion transistors is of paramount concern. Concurrently, the increasingly restrictive limits posed by interconnects make chip size wiring limited. Total chip capacitance, and therefore power dissipation, can be reduced by decreasing the chip size via optimal scaling of the wires using repeaters in a n-tier multilevel interconnect architecture [1]. A novel optimal repeater insertion methodology is demonstrated to minimize the power dissipation by minimizing the macrocell area for a n-tier multilevel interconnect architecture. The power dissipation and macrocell area are optimized for an ASIC macrocell case study using copper interconnects and a low-k dielectric (ε_r=2) for the 100, 70 and 50 *nm* technology generations projected by the International Technology Roadmap for Semiconductors (ITRS) [2].

2. N-TIER MULTILEVEL ARCHITECTURE

A tier is a collection of metal levels with the same cross sectional dimensions. The macrocell area is minimized by using a n-tier multilevel architecture which is designed by scaling the pitch on each tier so that the longest interconnect on a tier meets the timing delay constraint exactly [3]. A stochastic wiring distribution, validated previously with measured data [4], is used to estimate the longest interconnect on each tier. For the n^{th} tier, the range of interconnect lengths is calculated by equating the area available for wiring A_{av} to the area that is required for wiring A_{req}

$$A_{av} = n_l e_w A_m = \chi p_n \sqrt{\frac{A_m}{N}} \int_{L_{n-1}}^{L_n} l i(l) dl = A_{req} \qquad (1)$$

where n_l is the number of metal levels in the n^{th} tier, e_w (=0.4) is the wiring efficiency factor [5], A_m is the macrocell area, χ (=0.667) is a factor that converts the point-to-point interconnect length to wiring net length, p_n and L_n are the wire pitch and longest interconnect length of the n^{th} tier in microns and gate pitches, respectively, N is the number of logic gates and *i(l)* is the interconnect density function [4].

The wire pitch in (1) (without repeaters) for all non-local tiers (i.e. $p_n>2F$) is calculated using RC models for time delay as

$$p_n = 2\sqrt{\frac{1.1\rho\varepsilon_r\varepsilon_o 6.2 f_c}{\beta}}\sqrt{\frac{A_m}{N}}L_n \qquad (2)$$

where ρ is resistivity of metal, f_c is the clock frequency and β is the interconnect time delay expressed as a fraction of the clock period [5]. The maximum time delay is assumed to be 25% of the clock period (β=0.25) on the first tier and 90% (β = 0.9) on all other tiers.

3. REPEATER INSERTION MODELS AND METHODOLOGY

Repeaters improve the dependency of time delay on interconnect length from a square-law to a linear relationship [6]. Bakoglu [7] derives an expression for the pitch of an interconnect when an "optimal" number of equi-spaced repeaters are inserted so that the delay of each repeater equals the delay of the interconnect segment between repeaters. When the number of repeaters in an interconnect is less than this optimal number, it is considered a "sub-optimal" design [1]. If the number of repeaters is some factor ζ times the "optimal" number of repeaters, $0<\zeta<1$, then the pitch is given by

ISLPED '00, Rapallo, Italy.

$$p_n = \left(1.4 + 0.53\zeta + \frac{0.53}{\zeta}\right)\frac{2f_c}{\beta}\sqrt{6.2\rho\varepsilon_r\varepsilon_o R_o C_o}\sqrt{\frac{A_m}{N}}L_n \qquad (3)$$

where R_o and C_o are the output resistance and input capacitance of a minimum feature size inverter respectively.

Since the area of the chip is wire-limited, there is some unutilized silicon real estate which is used to accomodate the repeaters. It is assumed that only 60% of this free area is used for repeater insertion to account for practical routing and placement constraints.

The area of a gate (logic gate or repeater), A_g, is calculated as [8]

$$A_g = k_I\left(1 + \frac{4\sqrt{G_{ar}}(f_i - 1)}{\sqrt{k_I}}\right)\left(1 + \frac{(1+\beta_g)(w_k - 1)}{\sqrt{k_I G_{ar}}}\right) \qquad (4)$$

where k_I is the area of a minimum sized inverter, G_{ar} is the gate aspect ratio, f_i is the number of inputs, β_g is the ratio of PFET to NFET width and w_k is the NFET width to feature size ratio. The PFET width is constrained to satisfy equal worst case rise and fall times and w_k is calculated by equating the critical path delay to the clock period (= $1/f_c$),

$$\frac{1}{f_c} = \frac{n_{cp} T_{PDn} f_{ineff}}{b} \qquad (5)$$

where T_{PDn} is the NFET propagation delay including the transition time effect that is derived from the physical alpha-power law model [9], f_{ineff} is the effective fan-in factor for series connected MOSFETs [10], n_{cp} is number of gates in a critical path and b is the clock skew factor (=0.9).

Repeater insertion begins from the topmost tier and continues downward to lower tiers in a top-down design style. On each tier either the "optimal" number (ζ =1) or a "sub-optimal" number (ζ <1) of repeaters are inserted, depending on the number of repeaters available.

4. POWER DISSIPATION MODELS

The total power dissipation (P_{total}) in a macrocell is defined as

$$P_{total} = P_{logic} + P_{int} + P_{rep} \qquad (6)$$

where P_{logic}, P_{int} and P_{rep} are the power dissipations in the logic gates, interconnects and repeaters respectively. The power dissipation per gate (logic gate or repeater), P_g, is defined as

$$P_g = \frac{a}{2}C_g V_{DD}^2 f_c \qquad (7)$$

where a is the activity factor (=0.1), V_{DD} is the supply voltage and C_g is the total gate capacitance given as

$$C_g = w_k C_{go}. \qquad (8)$$

C_{go} is the gate overlap, junction and fan-out capacitance for a minimum feature sized gate. The product of the power per gate (7) and total number of logic gates results in the logic gate power dissipation as

$$P_{logic} = N\frac{a}{2}C_{g,logic}V_{DD}^2 f_c. \qquad (9)$$

The repeater power dissipation is calculated as

$$P_{rep} = n_{rep}\frac{a}{2}C_{g,rep}V_{DD}^2 f_c \qquad (10)$$

where n_{rep} is the number of repeaters. The power dissipation for all interconnects is calculated by

$$P_{int} = \frac{a}{2}c_o\left(L_{total}\sqrt{\frac{A_m}{N}}\right)V_{DD}^2 f_c \qquad (11)$$

where c_o is the distributed wiring capacitance per unit length, L_{total} is the total length of interconnects in gate pitches and $\sqrt{\frac{A_m}{N}}$ is the average gate pitch.

5. OPTIMIZATION RESULTS

Repeaters are used to minimize the area and power dissipation of an ASIC logic macrocell consisting of N=12.4M NAND gates for the 100*nm* technology generation projected by the ITRS [2]. These optimizations are compared against the same designs without using repeaters. Then, the total power dissipation is further reduced by tweaking the physical geometry of the interconnects.

5.1. Optimization using repeaters

The area of a macrocell with 8 metal levels and f_c=2GHz is minimized using repeaters as shown in Fig. 1. The wire-limited, transistor-limited and heat removal limited areas are the minimum areas required for wiring (within 8 levels), for logic gates and repeaters and for the power dissipation density to stay within the specified upperbound, respectively. The macrocell area is the maximum of these three areas. The maximum heat removal capacity is assumed to be 50 W/cm^2.

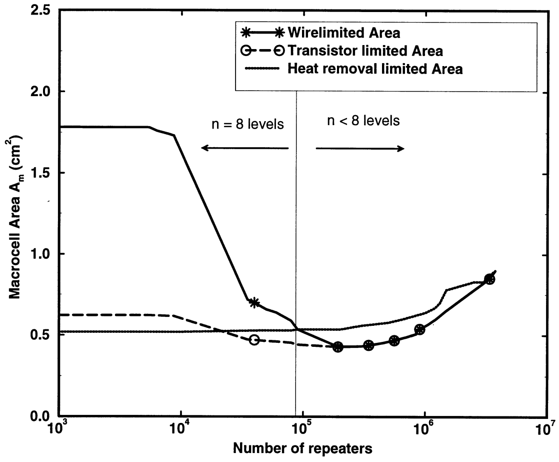


Fig.1 Minimum macrocell area optimization for a design with *N*=12.4M logic gates, f_c=2GHz and at most n=8 metal levels.

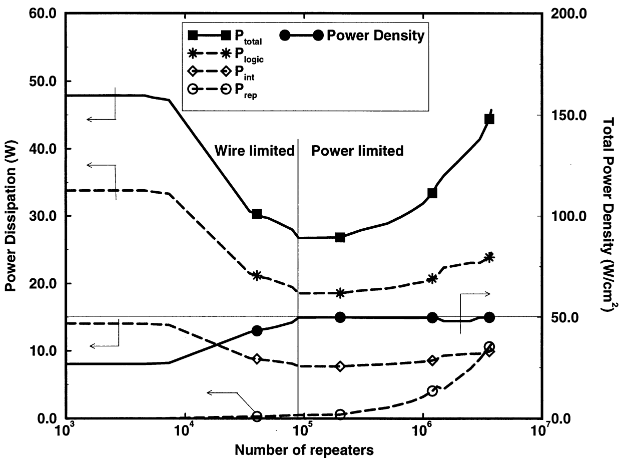


Fig. 2. Minimum power dissipation optimization for a design with *N*=12.4M logic gates, f_c=2GHz and at most n=8 levels.

Initially, the macrocell area is wire limited. By inserting repeaters, the interconnects become narrower which decreases the area required for wiring, reducing the average wiring capacitance, C_w. As C_w decreases, the logic gates can be made smaller thereby decreasing the transistor limited area. The increasing number of power-dissipating active devices (repeaters) increases the heat removal limited area. As the number of repeaters increases, the wire limited macrocell area decreases until it equals the heat removal limited area. If more repeaters are inserted, then the macrocell area becomes power limited and starts to increase. *Thus, for this example with a given number of metal levels, the macrocell area is minimized when the wire and heat removal limited areas become equal.* The wire limited area decreases till it equals the transistor limited area and then starts increasing due to an increase in the number of repeaters and the size of logic gates (required to achieve the desired clock frequency for a larger wiring capacitance). Repeater insertion can continue as long as the transistor limited area is less than the heat removal-limited macrocell area. As seen from Fig. 1, optimal repeater insertion decreases the macrocell area from 1.79cm^2 to 0.54cm^2, almost a 70% reduction in the cell size.

Figure 2 plots the power dissipation versus the number of repeaters for this design. Comparing Fig. 1 and Fig. 2, as the macrocell area decreases (with increase in the number of repeaters), the power dissipation in the logic gates and interconnects decreases. Reducing the macrocell area decreases the average interconnect length thereby reducing the average wiring capacitance. Hence, the size of the logic gates is smaller which decreases the gate capacitance in (8). From (9), the total logic gate power is reduced due to reduction in the gate capacitance. In Fig. 2, the repeater power dissipation monotonically increases because the effect of decreasing repeater size is overshadowed by the increase in the number of repeaters. From (11), the total interconnect power is also reduced due to a reduction in the macrocell area. Since the repeater power is small compared to the logic gate and interconnect power, the total power dissipation decreases and reaches a minimum when the macrocell area is at its minimum. Beyond this point, the increase in the macrocell area and number of repeaters causes the power dissipation in the logic gates, interconnects and repeaters to increase resulting in an increase in the total power dissipation. *Therefore, the power dissipation in the macrocell is minimzed when the macrocell area is minimized.* Table I provides all the interconnect parameters for the designs with and without repeaters.

For simplicity, (6) does not include the leakage power. However, it can be demonstrated that the minimum area design point corresponds to the minimum total leakage power for this example. The leakage power for a CMOS chip is given by

$$P_{off} = W_{tot} V_{DD} I_{off} \qquad (12)$$

where W_{tot} is the total macrocell turnedoff device width and I_{off} is the off-current per device width [11]. From Table I, the *W/L* ratio of the logic gates decreases 55% by inserting repeaters. Although the *W/L* ratio of the repeaters is comparatively larger, the number of repeaters is two orders of magnitude smaller than the number of logic gates. Therefore, when the macrocell area is minimized, the *W/L* ratio for logic gates is at a minimum and this minimizes the total leakage power. Hence, the inclusion of leakage power should not change the minimum area and power design point.

Table I. Interconnect parameters for the various design points

Tier # (n)	No. of levels	L_n		p_n (μm)	Number of repeaters	NFET (*W/L* ratios)	Power Dissipation
		Gate pitches	(*cm*)				
Without repeaters : A_m=1.79cm^2, f_c=2GHz and n=8 metal levels							
Tier 4	2	7042.7	2.6758	1.79	0	W/L_{logic}=22	P_{total}=47.93W
Tier 3	2	1845.2	0.7011	0.94	0		P_{logic}=71%
Tier 2	2	880.9	0.3347	0.45	0		P_{int}=29%
Tier 1	2	208.2	0.0791	0.20	0		P_{rep}=0%
(With repeaters) Minimum area and power : A_m=0.536cm^2, f_c=2GHz and n=8 metal levels							
Tier 4	2	7042.7	1.4642	0.38	9.00E4	W/L_{logic}=10	P_{total}=26.73W
Tier 3	2	1010.5	0.2101	0.28	0	W/L_{rep}=60	P_{logic}=69%
Tier 2	2	291.5	0.0606	0.20	0		P_{int}=29%, P_{rep}=2%
Tier 1	2	35.13	0.0073	0.20	0		

Fig. 2 also shows that when the macrocell area is wire limited, the power density is less than 50 *W/cm*^{2}; and it gets close to this value at and beyond the minimum power point proving that the macrocell design is power density-limited. From Fig. 2, repeater insertion decreases the total power dissipation from 47.9*W* to 26.7*W*, a reduction of 44%, that corresponds to the minimum area design point of A_m=0.54cm^2.

5.2 Interconnect geometry based power reduction

The previous analysis was based on the assumption that all the interconnect aspect ratios are equal to unity i.e. $W=T=H=S$, where W, T, H and S are the width and thickness of the interconnect, height of the inter-level oxide and spacing between the interconnects, respectively, as shown in Fig. 3. This assumption ensures that the capacitance per unit length of interconnect, c_o, is constant for all the designs. However, from equation (11), the power dissipation in the interconnects can be reduced by decreasing c_o. If the thickness of the interconnect and the height of the oxide are scaled by a factor Ω such that

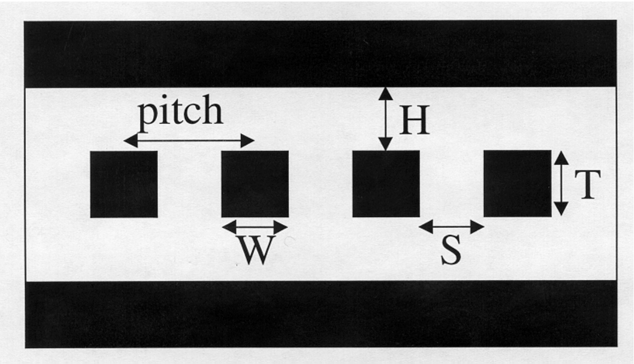


Fig. 3. Cross sectional view of a multilevel interconnect architecture.

$$T=\frac{W}{\Omega} \tag{13}$$

and

$$H=\Omega W \tag{14}$$

then c_o scales by a factor of $1/\Omega$ [12]. Using a parallel plate approximation for the ground and mutual capacitances, the RC time delay for an isolated line is approximated by [7]

$$\tau=RC\approx\frac{\rho\varepsilon_r\varepsilon_o}{TH}l^2, \tag{15}$$

and the crosstalk between two lines is approximately [5]

$$\frac{V_{peak}}{V_{DD}}\approx 0.5\frac{1}{1+\frac{C_{ground}}{C_{mutual}}}=0.5\frac{1}{1+\frac{WS}{TH}}. \tag{16}$$

Since (15) and (16) are independent of Ω , the time delay and crosstalk remain approximately unchanged. However, varying the value of Ω changes the power dissipation as shown in Fig. 4, which plots the total power dissipation versus Ω for the chosen case study.

For $\Omega > 1$, the ground and mutual capacitances decrease which reduces c_o. This decreases the average wiring capacitance C_w resulting in smaller logic gate sizes facilitating the use of more repeaters which help to decrease the macrocell area. The reduction in c_o and the macrocell area decreases the power dissipation. For $\Omega > 2$, the interconnect resistance increases, which forces the use of slightly wider interconnects resulting in a larger macrocell area as seen from Fig. 4. However, the effect of this small increase in the macrocell area is overshadowed by the continuous reduction in c_o leading to the monotonic decrease in total power dissipation with increase in Ω. For large Ω, the smaller ground capacitance results in higher crosstalk noise inspite of the reduction in the mutual capacitance due to the effect of fringing fields [12]. The upperbound on the value of Ω is determined by the constraints of the fabrication technology and the maximum permissible crosstalk noise. The ITRS predicts a maximum via aspect ratio ($H/W=\Omega$) between 2 and 3 for the future generations [2]. Choosing Ω=2.5, the total power dissipation is reduced from 26.73W to 15.35W, a reduction of 42.6%. From Fig. 4, as Ω decreases below 1, the power dissipation increases exponentially due to increases in the mutual and ground capacitances. This suggests that a low-power design requires thinner interconnects (Ω>1) whereas a high density design needs thicker wires (Ω < 1).

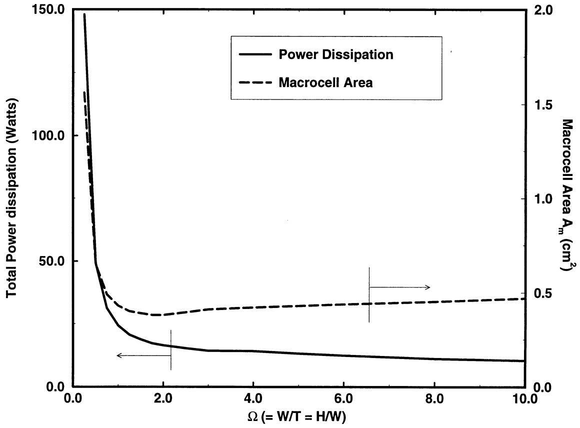
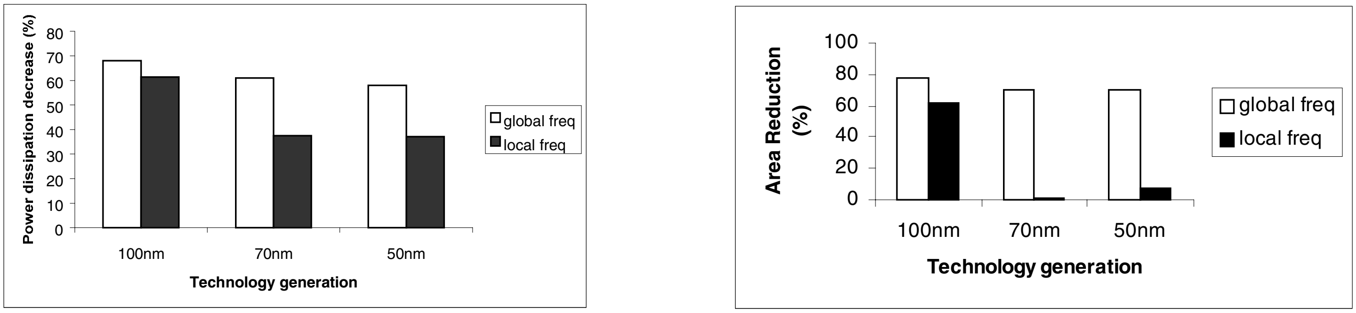


Fig. 4. Interconnect geometry based power reduction for 100*nm* ASIC macrocell with f_c=2GHz.

6. PROJECTIONS ACROSS THE ROADMAP

The methodology described in the previous section is used to predict the power reduction, obtained by employing repeaters, for future technology generations based on the ITRS projections. Table II shows the power dissipation predictions for three different technology generations. The number of gates for each design point is chosen so that the n-tier architecture without repeaters meets the specified clock frequency within the projected number of metal levels. The reduction in power dissipation (and macrocell area reduction) for the three technology generations is shown in Fig. 5.



(a) **(b)**

Fig. 5. (a) Power dissipation and (b) macrocell area reduction for the 100, 70 and 50nm technology generations for Ω=2.5. This plot is based on the data in Table II.

Table II. Technology predictions across the ITRS

ITRS 1999 projections							No repeaters		With repeaters			
							(Ω=1)		(Ω=1)		(Ω=2.5)	
f_c *(GHz)*	No. of levels	N ($\times 10^6$)	n_{cp}	t_{ox} *(nm)*	V_{DD} *(V)*	V_T *(V)*	A_m *(cm^2)*	P_{total} *(W)*	A_m *(cm^2)*	P_{total} *(W)*	A_m *(cm^2)*	P_{total} *(W)*
100*nm* technology generation												
2	8	12.4	10	1.8	1.0	0.125	1.79	47.93	0.536	26.73	0.40	15.35
3.5	8	7.6	7	1.8	1.2	0.12	0.94	82.91	0.75	74.23	0.36	32.06
70*nm* technology generation												
2.5	9	14.4	10	1.2	1.0	0.10	0.95	28.71	0.40	19.09	0.28	11.20
6	9	7	6	1.2	1.0	0.10	0.38	61.14	0.15	39.60	0.38	38.23
50*nm* technology generation												
3	10	15.5	10	0.8	0.8	0.10	0.50	15.55	0.22	10.76	0.15	6.48
10	10	5.7	5	0.8	0.8	0.10	0.14	52.97	0.11	48.76	0.13	33.29

Aggressive values for the number of gates in the critical path (n_{cp}), V_{DD} and threshold voltage (V_T) are needed to meet the higher local clock frequency compared to the global clock frequency. The maximum heat removal capacity is assumed to be 50 W/cm^2 for the lower global clock frequency for all the three generations; it would be far greater for the higher local clock frequency which is a high performance design.

Repeaters and interconnect geometry scaling techniques are used to reduce the total power dissipation and the macrocell area. From Table II, the total power has been reduced by 58-68% while the macrocell area has been decreased by 70-78% for the global clock frequency designs. For the local clock frequencies, the greater power dissipated per unit area indicates that these are high performance designs and hence can be realized only by trading the power dissipation for performance. This calls for better heat distribution and removal techniques.

7. CONCLUSION

It has been demonstrated that the power dissipation can be minimized by minimizing the macrocell area. Optimal repeater insertion in a n-tier design of a CMOS ASIC macrocell coupled with modified interconnect geometry techniques have been shown to decrease the macrocell area by 70-78% and reduce the power dissipation by 58-68%, for the global clock frequency designs, as compared to a n-tier network design without using repeaters. The results indicate that repeater design should be incorporated at an early stage of the design cycle for future technology generations.

8. ACKNOWLEDGMENTS

The support of the Semiconductor Research Corporation (Contract: SJ-374.002) and the Defense Advanced Research Project Agency (Contract: BAA 9415-A-007) is gratefully acknowledged.

9. REFERENCES

[1] R.Venkatesan, J.A.Davis, K.A.Bowman and J.D.Meindl, "Optimal repeater insertion for n-tier multilevel interconnect architectures," *Proc. of 3rd International Interconnnect Technology Conference*, San Francisco, June 2000.

[2] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors," 1999.

[3] R.Venkatesan, J.A.Davis and J.D.Meindl, "Performance enhancement through optimal n-tier multilevel interconnect architectures," *Proc. 12th IEEE ASIC/SOC Conference*, Washington D.C., pp.19-23, Sept 1999.

[4] J.A.Davis, V.K.De and J.D.Meindl, "A stochastic wire-length distribution for gigascale integration (GSI) - Parts I and II," *Trans. Electron Devices*, vol. 45, No. 3, pp.580-597, Mar 1998.

[5] T.Sakurai, "Closed form expressions for interconnection delay, coupling and crosstalk in VLSI's," *IEEE Trans. Electron Devices*, vol. 40, pp. 118-124, Jan 1993.

[6] H.B.Bakoglu and J.D.Meindl, "Optimal Interconnection Circuits for VLSI," *IEEE Trans. Elect. Devices*, Vol. ED-32, No. 5, May 1985.

[7] H.B.Bakoglu, *Circuits, interconnections and packaging for VLSI*, Reading, MA: Addison-Wesley, 1990.

[8] J.C.Eble, "A generic system simulator with novel on-chip cache and throughput models for gigascale integration," *Doctoral Thesis*, Georgia Institute of Technology, Atlanta, Nov. 1998.

[9] K.A.Bowman, B.L.Austin, X.Tang, J.C.Eble and J.D.Meindl, "A physical alpha-power law MOSFET model," *IEEE J. Solid-State Circuits*, Vol. 34, No. 10, pp. 410-414, Oct. 1999.

[10] T.Sakurai and A.R.Newton, "Delay analysis for series-connected MOSFET circuits," *IEEE J. Solid State Circuits*, vol.26, No. 2, pp. 122-131, Feb. 1991.

[11] Y.Taur, et al., "CMOS scaling into the nanometer regime," Proc. of the IEEE, Vol. 85, No. 4, pp. 486-504, April 1997.

[12] J.A.Davis, "A hierarchy of interconnect limits for gigascale integration," *Doctoral Thesis*, Georgia Institute of Technology, Atlanta, July 1999.