

Optimal n -tier Multilevel Interconnect Architectures for Gigascale Integration (GSI)

Raguraman Venkatesan, *Student Member, IEEE*, Jeffrey A. Davis, *Member, IEEE*, Keith A. Bowman, *Student Member, IEEE*, and James D. Meindl, *Life Fellow, IEEE*

Abstract—A multilevel interconnect architecture design methodology that optimizes the interconnect cross-sectional dimensions of each metal layer is introduced that reduces logic macrocell area, cycle time, power consumption or number of metal layers. The predictive capability of this methodology, which is based on a stochastic wiring distribution, provides insight into defining the process technology parameters for current and future generations of microprocessors and application-specific integrated circuits (ASICs). Using this methodology on an ASIC logic macrocell case study for the 100-nm technology generation, the optimized n -tier multilevel interconnect architecture reduces macrocell area by 32%, cycle time by 16% or number of wiring tracks required on the topmost tier by 62% compared to a conventional design where pitches are doubled for every successive pair of levels. A new repeater insertion methodology is also described that further enhances gigascale integration (GSI) system performance. By using repeaters, a further reduction of 70% in macrocell area, 18% in cycle time, 25% in number of metal levels or 44% in power dissipation is achieved, when compared to an n -tier design without repeaters. The key distinguishing feature of the methodology is its comprehensive framework that simultaneously solves two distinct problems—optimal wire sizing and wiring layer assignment—using independent constraints on maximum repeater area for efficient design space exploration to optimize the area, power, frequency, and metal levels of a GSI logic megacell.

Index Terms—Interconnections, modeling, multilevel systems, repeaters, system analysis and design, system-level interconnect prediction (SLIP), system optimization, wire-length distribution.

I. INTRODUCTION

AS CMOS semiconductor technology approaches gigascale integration (GSI), the increasingly restrictive limits posed by interconnects on processor performance make it imperative to optimize the wiring network for future technology generations [1], [2]. Reverse scaled multilevel wiring networks are extensively used in current VLSI systems to mitigate the impact of wiring on chip size and system performance [1]–[3]. A methodology to optimally design these reverse-scaled multilevel interconnect networks is presented that uses a stochastic wiring distribution [4] to estimate the interconnect lengths *a priori* and can be used to determine interconnect process parameters for future technology generations. Because repeater insertion reduces wire delay, repeaters can be an effective tool to reduce the stringent

interconnect limits on future GSI systems [5]. A novel repeater insertion methodology is presented in this work that optimally inserts repeaters in a multilevel wiring network to decrease chip size, cycle time, number of metal levels or power dissipation.

A similar layer-assignment algorithm has been described in [6], where wires are assigned to layers that are best fit depending on time delay constraints. The algorithm minimizes the total number of wiring layers for a fixed die area and for predefined upper and lower bounds on the wire pitches of each layer. In [6], the authors also consider the impact of vias on the area available for wiring, using models from [7]. However, they state that “a tight delay constraint forces an area-inefficient solution.” Also, since only optimal repeater width and count is used throughout the algorithm, “the solution that is produced by the algorithm . . . does not yet guarantee that it lies within the constraints of the repeater area” [6]. In contrast, the methodology presented in this paper determines the wire pitches for each metal layer to design an optimized wiring network that minimizes area, cycle time, power dissipation, or minimum number of levels [8]–[10] *within the constraints of available repeater area*. It is self-consistent and the solution obeys all the user-defined constraints on maximum permissible time delay for each metal layer and total repeater area (unless no solution exists for the set of constraints provided, in which case, the methodology returns no solution). This methodology is well suited for technology prediction and for defining interconnect process parameters for future technology generations.

Section II describes the models and the algorithm for the n -tier methodology. Section III analyzes an application-specific integrated circuit (ASIC) logic macrocell based on International Technology Roadmap for Semiconductors (ITRS) projections for the 100-nm technology generation and compares the advantages of an n -tier architecture to that of a conventional architecture where the wiring pitches are doubled for every pair of metal levels. Section IV describes the optimal repeater insertion methodology. The ASIC case study designs using the n -tier methodology with and without repeaters are compared in Section V. Finally, conclusions are provided in Section VI.

II. n -TIER MULTILEVEL ARCHITECTURE DESIGN METHODOLOGY

The key assumptions made in the n -tier design methodology are as follows.

- 1) The interconnects in a system obey the stochastic interconnect wiring distribution described in [4].

Manuscript received October 6, 2000; revised February 28, 2001. This work was supported by the Semiconductor Research Corporation under Contract SJ-374.002 and by the Defense Advanced Research Project Agency under Contract BAA 9415-A-007.

The authors are with the Microelectronics Research Center, School of ECE, Georgia Institute of Technology, Atlanta, GA 30332-0269 USA (e-mail: vragu@ee.gatech.edu).

Publisher Item Identifier S 1063-8210(01)07733-2.

- 2) Shortest wires are routed on the tier (collection of levels with the same wiring pitch) with the smallest pitch and successively longer wires go on tiers with progressively larger pitches.
- 3) The tier pitch is chosen based on given performance constraints (for the example case study, the maximum permissible time delay is 25% of the clock period for the lowest tier and 90% of the clock period for all other tiers).
- 4) For the example case study, the aspect ratio is chosen to be unity. However, the *n-tier* methodology is independent of the aspect ratio chosen; a nonunity aspect ratio requires a slight modification to the interconnect delay equation.
- 5) The wiring efficiency factor (c_w) is assumed to be constant for all the levels (for the example case study, it is assumed to be 40%).

The *n-tier* multilevel interconnect optimization is based on a stochastic wire-length distribution [4], which is used to obtain an *a priori* estimate of interconnect lengths in a logic block. The interconnect density function $i(l)$ that describes the macrocell wiring is given as [4]

$$\text{Region I: } 1 \leq l \leq \sqrt{N_g} \\ i(l) = \frac{\alpha k}{2} \Gamma \left(\frac{l^3}{3} - 2\sqrt{N_g}l^2 + 2N_g l \right) l^{2p-4} \quad (1)$$

$$\text{Region II: } \sqrt{N_g} \leq l \leq 2\sqrt{N_g} \\ i(l) = \frac{\alpha k}{6} \Gamma \left(2\sqrt{N_g} - l \right)^3 l^{2p-4} \quad (2)$$

where

- l interconnect length in gate pitches;
- N_g number of logic gates;
- p Rent's exponent;
- k Rent's coefficient;
- α fraction of sink terminals in the macrocell;
- Γ normalizing factor.

A gate pitch is defined as the average distance separating two gates and is equal to $\sqrt{A_m/N_g}$, where A_m is the macrocell area.

To reduce wiring layout problems, interconnects on adjacent *pairs* of metal levels are assumed to be routed orthogonally with the same wiring pitch. A collection of *pairs* having the same pitch is identified as a *tier*. In general, the range of interconnect lengths on the t^{th} tier is calculated by equating the area available for wiring A_{av} to the area that is required for wiring A_{req}

$$A_{\text{av}} = n_t c_w A_m = \chi p_t \sqrt{\frac{A_m}{N_g}} \int_{L_{t-1}}^{L_t} i(l) dl = A_{\text{req}} \quad (3)$$

The number of metal levels in the t^{th} tier is n_t , c_w is the wiring efficiency factor [3], [11], χ converts point-to-point interconnect length to wiring net length [4], p_t and L_t are the wire pitch and longest interconnect length on the t^{th} tier in microns and gate pitches, respectively.

The wiring pitch of the *local* tier is equal to twice the minimum feature size ($p_t = 2F$); for all nonlocal tiers (i.e., $p_t >$

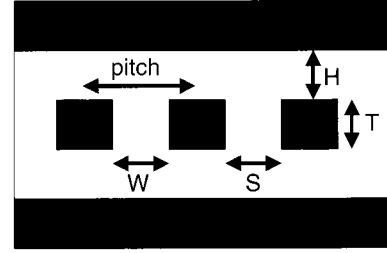


Fig. 1. Cross-sectional view of a multilevel interconnect architecture.

2 F), the pitch is obtained by equating the resistance capacitance (RC) time delay of the longest interconnect to an acceptable fraction of the cycle time [5], [12]

$$\tau = \frac{\beta}{f_c} \approx 4 \frac{1.1\rho\epsilon_r\epsilon_o 6.2 A_m L_t^2}{p_t^2 N_g} \quad (4)$$

Therefore

$$p_t = 2 \sqrt{\frac{1.1\rho\epsilon_r\epsilon_o 6.2 f_c}{\beta}} \sqrt{\frac{A_m}{N_g}} L_t \quad (5)$$

where

- τ interconnect time delay;
- β interconnect time delay expressed as a fraction of the cycle time ($= 1/f_c$);
- ρ resistivity of metal.

This formulation assumes that all the interconnect aspect ratios are unity, i.e., $W = T = S = H = p_t/2$, where W , T , S and H are the metal width, metal thickness, spacing between the interconnects and height of the inter-level dielectric, respectively, as illustrated in Fig. 1.

For a given value of A_m and f_c , the *n-tier* design methodology starts with the lowest tier and moves upwards, filling each tier with interconnects. To maximize wire density, it tries to employ minimum width interconnects on the local tier; solving (3) gives the longest interconnect length on the local tier. For nonlocal tiers, (3) and (5) are solved simultaneously, while scaling the pitch of each tier so that the longest interconnect on that tier satisfies the timing constraint (delay $\tau = \beta/f_c$). Thus, the wire density is maximized resulting in a minimum number of metal levels. Once the twice-die-edge-long interconnect has been accommodated on a tier, the algorithm stops, and counts the total number of levels. This procedure is repeated for different values of the macrocell area and clock frequency to determine the various optimizations described in the following section. A flowchart showing the complete *n-tier* design methodology is shown in Appendix A.

III. PERFORMANCE ENHANCEMENT USING OPTIMIZED *n-tier* ARCHITECTURES

The performance enhancement achieved by the optimized *n-tier* architecture is demonstrated for an ASIC logic macrocell using 100-nm technology projections from the ITRS [13]. In this case study, the macrocell has 11.3 M logic gates (assuming the use of three-input six-transistor NAND gates, this corresponds to approximately 68 M transistors), a low permittivity dielectric ($\epsilon_r = 2$) and copper interconnects (for

TABLE I
COMPARISON OF VARIOUS DESIGN POINTS USING CONVENTIONAL AND OPTIMIZED *n-tier* DESIGN METHODOLOGIES

	Conventional Designs	Optimized <i>n-tier</i> design		
		Min. Area optimization	Max. freq. optimization	Min. # levels optimization
$A_m =$	2-Tier (2F,2F,8F,8F) 0.82 cm ²	0.37 cm ²	0.82 cm ²	0.82 cm ²
$f_c =$	217 MHz	217 MHz	1.47 GHz	217 MHz
$n =$	8 levels	8 levels	8 levels	4.98≈6 levels
$A_m =$	3-Tier (2F,2F,4F,8F) 0.68 cm ²	0.47 cm ²	0.68 cm ²	0.68 cm ²
$f_c =$	672 MHz	672 MHz	1.16 GHz	672 MHz
$n =$	8 levels	8 levels	8 levels	6.57≈8 levels
$A_m =$	3-Tier (2F,4F,4F,8F) 1.01 cm ²	0.48 cm ²	1.01 cm ²	1.01 cm ²
$f_c =$	710 MHz	710 MHz	1.6 GHz	710 MHz
$n =$	8 levels	8 levels	8 levels	6.24≈8 levels
$A_m =$	4-Tier (2F,4F,8F,16F) 1.45 cm ²	0.98 cm ²	1.45 cm ²	1.45 cm ²
$f_c =$	1.56 GHz	1.56 GHz	1.86 GHz	1.56 GHz
$n =$	8 levels	8 levels	8 levels	6.77≈8 levels

the system to meet the ITRS projections of a 2-GHz clock frequency and a maximum of eight metal levels for the 100-nm technology generation, the maximum size of the macrocell is 11.3-M logic gates). The following values were assumed for the other parameters in (1)–(5): Rents exponent $p = 0.6$, Rent's coefficient $k = 4$ [14], $\alpha = 0.75$ ($\alpha = \text{fanout}/(\text{fanout} + 1)$); for three-input NAND gates, fanout = 3), $\chi = 0.667$ [4] and $e_w = 0.4$ [11]. Shorter interconnects (first tier) are more likely to constitute critical paths and, hence, are assigned a smaller β ($= 0.25$), so that the critical path gates can have a larger (remaining) time delay. A larger β ($= 0.9$) is assigned to longer interconnects because they would most likely be used for cross-chip communication only (requiring more delay). Three optimizations are defined in this section that minimize the number of metal levels, the macrocell area, or the cycle time. To demonstrate the advantages of an *n-tier* architecture, a comparison is made with conventional designs where the wiring pitch is arbitrarily scaled from one tier to the next.

A. Conventional Multilevel Network Designs

Initially, *two-tier* designs comprised of a thin *local* tier and a thicker *global* tier. Whenever the longest interconnect on the first tier cannot meet the timing constraint, the interconnects on that tier are moved to the global tier with a greater pitch, which consequently increases the number of metal levels. The left half of Table I shows a two-tier design having four local levels with pitch = 2 F and four global levels with pitch = 8 F. The operational frequency is determined by setting the maximum time delay among all the interconnects in the system equal to 90% of the clock period (25% for the interconnects on the local tier). In three-tier designs, there is an intermediate semi-global tier. Two examples of three-tier architectures are shown in Table I, one with four local, two semi-global and two global levels, and the other with two local, four semi-global and two global levels. In both cases, the semi-global pitch is equal to 4 F, while the local and global pitches remain unchanged. Sai-Halasz [11] recommends a multilevel network where the pitch is doubled for every

successive pair of metal levels. Such a four-tier network with wiring pitches equal to 2 F, 4 F, 8 F, and 16 F for each pair of metal levels is shown in Table I. The process described in [15] also uses a similar wire scaling pattern, and, therefore, this design would henceforth be called the *conventional baseline design*. Using this design, the macrocell in the case study has an area of 1.45 cm² (baseline area) with eight metal levels and can operate at 1.56 GHz (baseline clock frequency).

Such conventional designs are not optimal designs because the time delay of the longest interconnect on some tiers may be less than the maximum permissible delay (i.e., $\beta < 0.9$). This leaves room for decreasing the metal pitch or macrocell area, and consequently reducing the number of metal levels or increasing the operational clock frequency. This is the main motivation for the optimized *n-tier* design.

B. Optimized *n-tier* Design

The *n-tier* multilevel architecture is designed with the wire pitch of each orthogonal pair of levels dependent on the time-delay of the longest interconnect routed in that pair. This ensures that the time delay constraints are exactly satisfied and also the wiring density is maximum. Fig. 2 plots the number of metal levels versus macrocell area of the *n-tier* design for clock frequencies of 1, 1.56, 1.86, 2, and 3 GHz (the optimized *n-tier* designs have clock frequencies of 1.56, 1.86, and 2 GHz). The curves in Fig. 2 saturate when the minimum feature size width and spacing can no longer be used for any tier (i.e., $p_t > 2 F$ for all tiers). This is because, when (5) is substituted in (3), the longest interconnect length L_t of the lowest tier *also* becomes independent of macrocell area A_m . Therefore, the number of metal levels remains constant because the interconnect pitch and length increase in the same ratio as the macrocell area is increased. The improvement in system design using the *n-tier* methodology is quantified in the following three optimizations and the results are summarized in the right half of Table I. Detailed comparison between the conventional baseline design and the optimized *n-tier* design is shown in Table II.

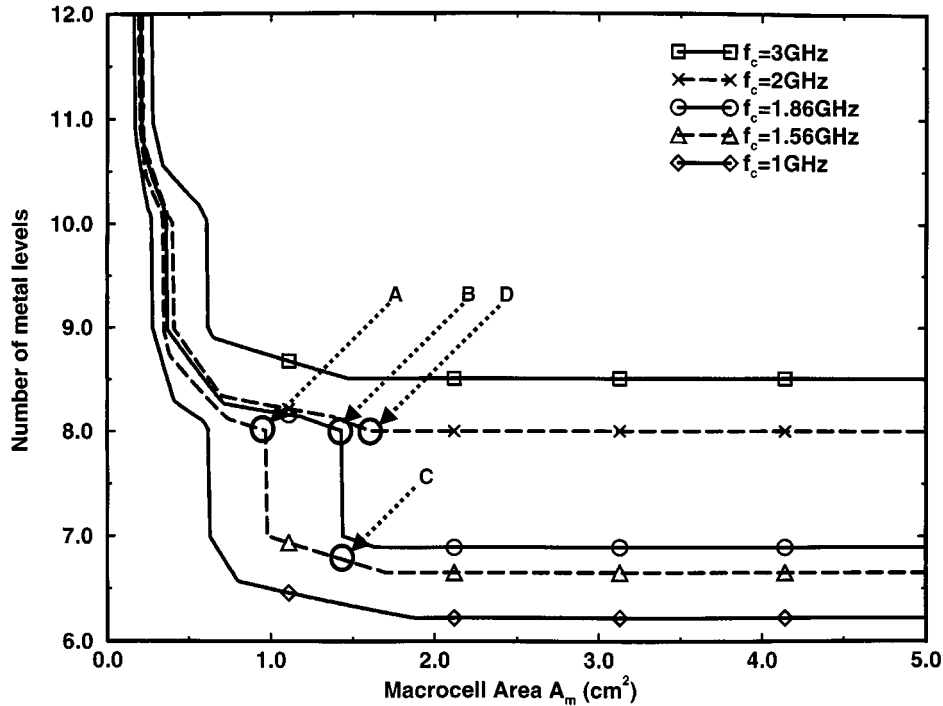


Fig. 2. Number of metal levels versus macrocell area for the n -tier design of a 68-M transistor macrocell.

TABLE II
DETAILED INTERCONNECT PARAMETERS FOR THE CONVENTIONAL BASELINE DESIGN AND OPTIMIZED n -tier DESIGNS

Tier # (n)	Number of levels	L_{n-1}		L_n		P_n (μm)
		Gate pitches	(cms)	Gate pitches	(cms)	
Conventional baseline architecture : $A_m=1.45\text{cm}^2$, $f_c=1.56\text{GHz}$ and $n=8$ levels						
Tier 4	2	1884	0.67	6723	2.41	1.6
Tier 3	2	818	0.29	1884	0.67	0.8
Tier 2	2	174	0.06	818	0.29	0.4
Tier 1	2	1	0.0004	174	0.06	0.2
(n-tier architecture) Minimum area : $A_m=0.98\text{cm}^2$, $f_c=1.56\text{GHz}$ and $n=8$ levels						
Tier 4	2	1760	0.52	6723	1.98	1.17
Tier 3	2	735	0.22	1760	0.52	0.61
Tier 2	2	93	0.03	735	0.22	0.26
Tier 1	2	1	0.0003	93	0.03	0.2
(n-tier architecture) Maximum clock frequency : $A_m=1.45\text{cm}^2$, $f_c=1.86\text{GHz}$ and $n=8$ levels						
Tier 4	2	1847	0.66	6723	2.41	1.56
Tier 3	2	846	0.30	1847	0.66	0.85
Tier 2	2	174	0.06	846	0.30	0.39
Tier 1	2	1	0.0004	174	0.06	0.2
(n-tier architecture) Minimum number of levels : $A_m=1.45\text{cm}^2$, $f_c=1.56\text{GHz}$ and $n=6.77\approx 8$ levels						
Tier 4	$0.77\approx 2$	2000	0.72	6723	2.41	1.42
Tier 3	2	893	0.32	2000	0.72	0.85
Tier 2	2	174	0.06	893	0.32	0.38
Tier 1	2	1	0.0004	174	0.06	0.2

1) *Minimum Macrocell Area*: For the same number of metal levels and clock frequency, the optimized n -tier architecture reduces the macrocell area by decreasing the wiring pitch resulting in greater packing density for the interconnects. The minimum macrocell area optimization is shown by point A on the curve in Fig. 2, which plots the number of metal levels against macrocell area for the n -tier design at the baseline clock frequency of 1.56 GHz. For a maximum of $n = 8$ metal levels, the conventional baseline design requires a macrocell area of

1.45 cm^2 , whereas the n -tier design requires a macrocell area of only 0.98 cm^2 , which is a 32% reduction in cell size.

2) *Maximum Clock Frequency*: For the same number of metal levels and macrocell area, the optimized n -tier architecture improves the clock frequency by increasing the wiring pitch so that the area wasted by the conventional baseline design is utilized. The maximum clock frequency optimization using the n -tier design is shown by point B on the curve in Fig. 2, which plots the number of levels versus macrocell area

for the n -tier design at a clock frequency of 1.86 GHz. This corresponds to a 16% reduction in the cycle time over the baseline clock frequency of 1.56 GHz for the conventional baseline design. This is the maximum clock frequency achievable using the n -tier designs for $n = 8$ metal levels and $A_m = 1.45 \text{ cm}^2$. To achieve higher clock frequencies, the curves will shift upward and to the right, as seen in Fig. 2, necessitating a greater number of metal levels or larger area.

3) *Minimum Number of Metal Levels*: The optimized n -tier architecture reduces the number of metal levels, for the same area and clock frequency, by decreasing the wiring pitch resulting in greater packing density for the interconnects. The reduction in the number of metal levels using the n -tier design is shown by point C on the curve in Fig. 2, which plots the number of levels versus macrocell area for the n -tier design at the baseline clock frequency of 1.56 GHz. Since the levels are grouped in x - y orthogonal pairs, the number of metal levels should be rounded off to the next higher *even* integer. For the baseline macrocell area of $A_m = 1.45 \text{ cm}^2$, the n -tier design requires 6.77 metal levels (which when rounded off to eight metal levels, is the same as for the conventional baseline design). Thus, for this example, although there is no actual reduction in the number of metal levels, 62% of the wiring tracks on the topmost tier have been freed and can be used to accommodate additional power, ground and clock wiring resources.

Thus, the n -tier design methodology can be used to determine the interconnect pitches for different tiers so that optimal performance is extracted from the system for the available resources. Additional improvements in performance can be achieved through the insertion of repeaters. The following sections investigate an optimum repeater insertion methodology in the n -tier design process and quantify its impact.

IV. REPEATER INSERTION MODELS AND METHODOLOGY

Repeaters have been previously shown to improve the dependency of time delay on interconnect length from a square law to a linear relationship [3], [16]. As mentioned in Section I, repeater insertion provides a viable option to relieve the demanding interconnect restrictions placed on future GSI systems. Repeaters are increasingly being used by chip designers to improve the performance of microprocessors [17]. The potential of repeaters to improve the n -tier architecture design is demonstrated in this section.

A. Repeater Models

Bakoglu [2], [14] derived an expression for the time-delay, τ , of an interconnect when the number of equi-spaced repeaters is “optimal,” which minimizes the cumulative delay of repeaters and interconnect segments as

$$\tau = \frac{\beta}{f_c} = 2.46 \frac{2}{p_t} \sqrt{6.2\rho\epsilon_r\epsilon_o R_o C_o} \sqrt{\frac{A_m}{N_g}} L_t. \quad (6)$$

R_o and C_o are the output resistance and input capacitance of a minimum size inverter, respectively. However, if the number of

repeaters is a factor ζ times the “optimal” number of repeaters $0 < \zeta \leq 1$, then the time delay can be expressed as

$$\tau = \frac{\beta}{f_c} = \left(1.4 + 0.53\zeta + \frac{0.53}{\zeta}\right) \frac{2}{p_t} \sqrt{6.2\rho\epsilon_r\epsilon_o R_o C_o} \sqrt{\frac{A_m}{N_g}} L_t. \quad (7)$$

This “suboptimal” design in (7) utilizes a smaller number of repeaters resulting in a larger delay. The tradeoff between the performance and the number of repeaters is shown in Fig. 3, which plots the ratio of the suboptimal delay in (7) to the optimal delay in (6) versus ζ . A 50% reduction in the number of repeaters from the optimal number imposes a performance penalty of *only* 10%. The considerable savings in silicon area, wiring complexity and power dissipation encourages this worthwhile tradeoff. Assuming the area of the chip is wire limited, there is unutilized silicon area that is available for repeaters and is given by

$$A_{\text{rep}} = e_{\text{rep}} A_{\text{free}} = e_{\text{rep}} (A_m - A_{\text{logic_gates}}) \quad (8)$$

where

A_{rep}	area occupied by repeaters;
e_{rep}	repeater insertion efficiency;
A_{free}	free area;
$A_{\text{logic_gates}}$	area occupied by logic gates.

Only a fraction ($e_{\text{rep}} = 0.6$) of the free area is assumed available for repeater insertion to account for practical routing and placement constraints and additional silicon area needed for on-chip decoupling capacitors. The area occupied by logic gates and repeaters is estimated using the following gate area models.

B. Gate Area Models

The area of a gate (either a logic gate or a repeater), A_g , is calculated as [18]

$$A_g = k_I \left(1 + \frac{4\sqrt{G_{\text{ar}}}(f_i - 1)}{\sqrt{k_I}}\right) \times \left(1 + \frac{(1 + \beta_g)(w_k - 1)}{\sqrt{k_I G_{\text{ar}}}}\right) F^2 \quad (9)$$

where

k_I	area of a minimum sized inverter with respect to F^2 ;
G_{ar}	gate aspect ratio;
f_i	number of inputs;
β_g	ratio of pFET to nFET width;
w_k	n-channel field effect transistor (nFET) width to feature size ratio.

The p-channel field effect transistor (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 cycle time ($= 1/f_c$) [19]

$$\frac{1}{f_c} = \frac{n_{cp} T_{\text{PDn}} f_{\text{ineff}}}{b} \quad (10)$$

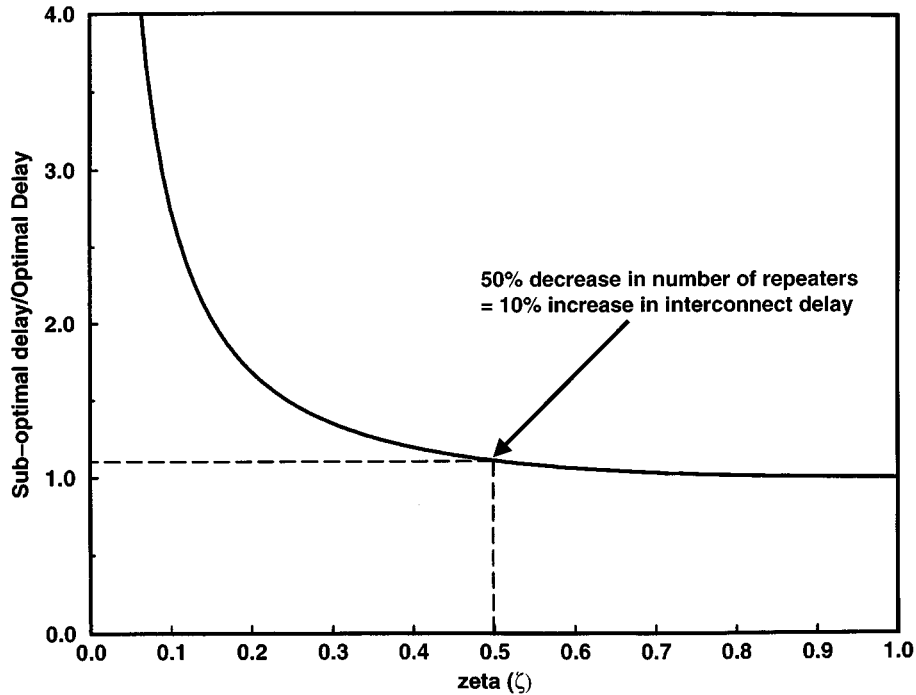


Fig. 3. Tradeoff between performance and number of repeaters.

where

T_{PDn} nFET propagation delay including the transition time effect that is derived from the physical alpha-power law model [19];

f_{ineff} effective fan-in factor for series connected MOSFETs [20];

n_{cp} number of gates in a critical path;

b clock skew factor ($= 0.9$).

Using (7)–(9), the total logic gate area is estimated, which defines the limit on the maximum amount of area available for repeaters.

C. Power Dissipation Models

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

$$P_{total} = P_{logic} + P_{int} + P_{rep} \quad (11)$$

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

$$P_g = \frac{a}{2} w_k C_{go} V_{DD}^2 f_c \quad (12)$$

where

a activity factor ($= 0.1$);

V_{DD} supply voltage;

C_{go} gate overlap, junction, and fan-out capacitance for a minimum sized gate.

The product of the number of gates and power dissipation per gate gives the total logic or repeater power dissipation. The power dissipated in the interconnects is calculated by

$$P_{int} = \frac{a}{2} C_w V_{DD}^2 f_c \quad (13)$$

where C_w is the total wiring capacitance given by

$$C_w = c_{int} \left(L_{total} \sqrt{\frac{A_m}{N_g}} \right) \quad (14)$$

where

c_{int} distributed wiring capacitance per unit length;
 L_{total} total length of interconnects in gate pitches;
 $\sqrt{A_m/N_g}$ average gate pitch.

D. Repeater Insertion Methodology

For a given macrocell area A_m and clock frequency f_c , the n -tier methodology described in Section II is first used to design its multilevel interconnect architecture. The models described in (7)–(14) are then used to determine the maximum number and size of repeaters that can be inserted in a multilevel architecture. Since thicker global interconnects benefit the most by using repeaters, a “top-down” repeater insertion methodology is adopted. Repeaters are first inserted in the uppermost tier. Repeater insertion then continues downward to the lower tiers depending on the amount of free silicon area available. This algorithm tries to insert 50% of the “optimal” number of repeaters ($\zeta = 0.5$) in all the wires on a given tier; if the area required for repeaters is not available, ζ is systematically decreased until either the repeater area constraint is satisfied or the number of repeaters reaches zero (see Fig. 3). If the interconnect width decreases to minimum width, then repeater insertion is discontinued. The resulting architecture is the new multi-

TABLE III
DETAILED INTERCONNECT PARAMETERS FOR THE *n-tier* BASELINE DESIGN (WITHOUT REPEATERS) AND OPTIMIZED *n-tier* WITH REPEATERS DESIGNS

Tier # (n)	No. of levels	L_n		p_n (μm)	Number of repeaters	zeta	NFET (W/L ratios)	Power Dissipation
		Gate pitches	(cm)					
Without repeaters : $A_m=1.62\text{cm}^2$, $f_c=2\text{GHz}$ and $n=8$ metal levels								
Tier 4	2	6723	2.55	1.71	0	0	$W/L_{logic}=17$	$P_{total}=43.0\text{W}$
Tier 3	2	1871	0.71	0.95	0	0		$P_{logic}=71\%$
Tier 2	2	885	0.34	0.45	0	0		$P_{int}=29\%$
Tier 1	2	208	0.08	0.20	0	0		$P_{rep}=0\%$
(With repeaters) Minimum area and power : $A_m=0.48\text{cm}^2$, $f_c=2\text{GHz}$ and $n=6.97\approx 8$ metal levels								
Tier 4	0.97 \approx 2	6723	1.39	0.73	49840	0.216	$W/L_{logic}=9$ $W/L_{rep}=114$	$P_{total}=24\text{W}$
Tier 3	2	1003	0.21	0.28	0	0		$P_{logic}=69\%$
Tier 2	2	283	0.06	0.20	0	0		$P_{int}=29\%$, $P_{rep}=2\%$
Tier 1	2	34	0.007	0.20	0	0		
(With repeaters) Maximum frequency : $A_m=1.62\text{cm}^2$, $f_c=2.44\text{GHz}$ and $n=6.49\approx 8$ metal levels								
Tier 4	0.49 \approx 2	6723	2.55	1.36	34855	0.308	$W/L_{logic}=63$ $W/L_{rep}=211$	$P_{total}=150\text{W}$
Tier 3	2	1684	0.64	0.94	0	0		$P_{logic}=89\%$
Tier 2	2	810	0.31	0.45	0	0		$P_{int}=10\%$, $P_{rep}=1\%$
Tier 1	2	193	0.07	0.21	0	0		
(With repeaters) Minimum number of levels : $A_m=1.62\text{cm}^2$, $f_c=2\text{GHz}$ and $n=4.99\approx 6$ metal levels								
Tier 3	0.99 \approx 2	6723	2.57	1.21	75390	0.26	$W/L_{logic}=17$ $W/L_{rep}=188$	$P_{total}=44.4\text{W}$
Tier 2	2	885	0.34	0.45	0	0		$P_{logic}=69\%$
Tier 1	2	208	0.08	0.20	0	0		$P_{int}=28\%$, $P_{rep}=3\%$

level interconnect architecture with repeaters (if any have been added). The complete repeater insertion methodology for the *n-tier* architecture is shown in Appendix B

V. PERFORMANCE ENHANCEMENT USING REPEATERS FOR *n-tier* ARCHITECTURES

The maximum clock frequency achievable by the optimized *n-tier* architecture (without repeaters), for the macrocell case study described in Section III, using eight metal levels (and area $A_m = 1.62 \text{ cm}^2$) is 2 GHz, as shown by point D in Fig. 2. This design is henceforth called the *n-tier baseline design*. In this section, repeaters are used to minimize the area, cycle time, number of metal levels or power dissipation of the macrocell case study. These optimizations are compared against the *n-tier* baseline design (without repeaters) and the results are summarized in Table III.

A. Minimum Macrocell Area

The area of a macrocell with eight metal levels and $f_c = 2 \text{ GHz}$ is minimized using repeaters as shown in Fig. 4. The wire-limited and transistor-limited areas are the minimum areas required to have a maximum of eight metal wiring levels and for accommodating logic gates and repeaters, respectively. The power limited area is the minimum area required to keep the power dissipation density within the specified upper bound of 50 W/cm^2 , which is the assumed maximum heat removal capacity. This is calculated by dividing the total power dissipation of the macrocell by 50 W/cm^2 . The macrocell area is the maximum of these three areas.

Initially, the macrocell area is wire-limited as illustrated in Fig. 4. By inserting repeaters, the interconnects become narrower, which decreases the area required for wiring and reduces the wiring capacitance C_w (14). As C_w decreases, the logic

gates can be made smaller thereby decreasing the transistor limited area. Increasing the number of power-dissipating repeaters increases the power-limited area. As the number of repeaters increases, the wire-limited macrocell area decreases until it equals the power-limited area. If more repeaters are inserted, then the macrocell area becomes power-limited and begins to increase. Thus, for this example, the macrocell area is minimized when the wire and power-limited areas become equal. As seen from Fig. 4, optimal repeater insertion decreases the macrocell area from 1.62 cm^2 to 0.48 cm^2 , almost a 70% reduction in the cell size. If more advanced heat removal mechanisms are used such as liquid or two-phase cooling, then the power limited area curve would shift lower. Then the wire-limited area may decrease till it equals the transistor-limited area. From here the macrocell area would start to increase due to an increase in the number of repeaters and the size of logic gates required to achieve the desired clock frequency for an increased wiring capacitance.

B. Maximum Frequency Optimization

Fig. 5 plots the clock frequency versus number of repeaters to maximize the clock frequency through repeater insertion for a macrocell with $A_m = 1.62 \text{ cm}^2$ and eight metal levels. The wire-limited clock frequency is the maximum frequency for which all the transistors can be wired within eight metal levels for the specified macrocell area. The wire-limited clock frequency increases with an increase in the number of repeaters because repeaters decrease wire delay for a constant pitch. The transistor-limited clock frequency is the maximum frequency at which the logic critical path transistors can operate if they are enlarged to occupy all remaining macrocell area after accounting for repeaters. Transistor-limited clock frequency decreases with an increase in the number of repeaters due to a decrease in the area available for logic gates. Initially, when

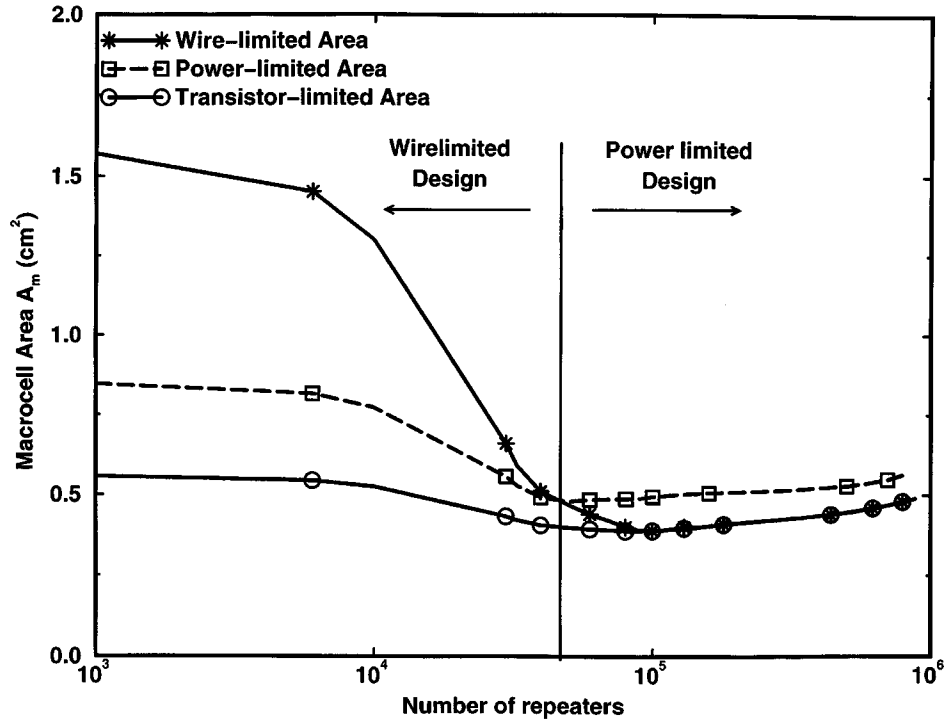


Fig. 4. Minimum area optimization using repeaters for $n \leq 8$ levels and $f_c = 2$ GHz.

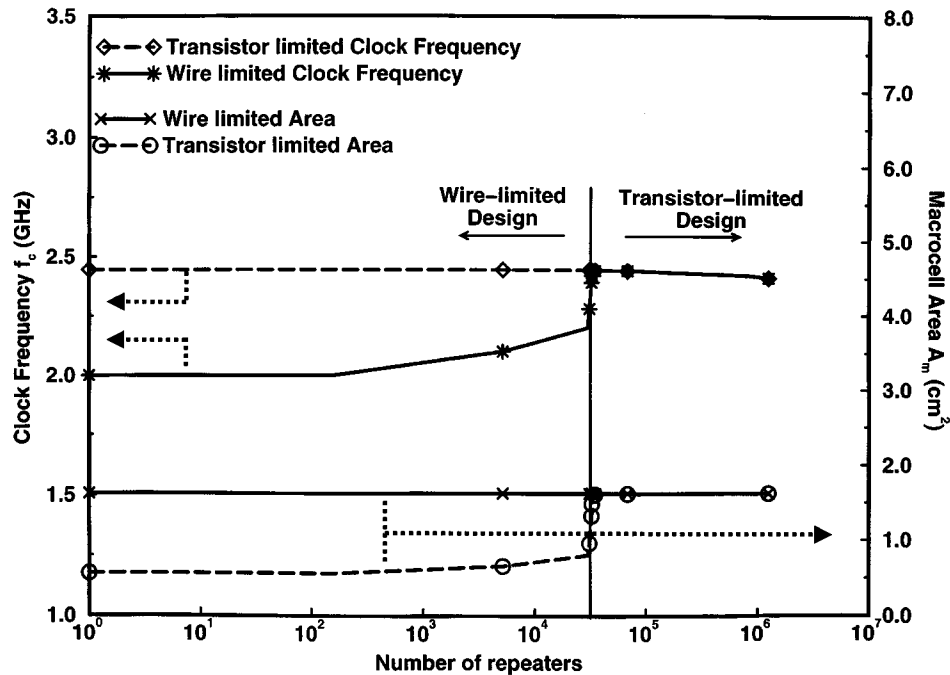


Fig. 5. Maximum clock frequency optimization using repeaters for $A_m = 1.62 \text{ cm}^2$ and $n \leq 8$ levels.

the macrocell area is wire-limited, the clock frequency is also wire-limited. As the number of repeaters increases, the wire and transistor limited-frequencies eventually converge when the macrocell area becomes transistor-limited. More repeaters can be added only by shrinking the logic transistors which reduces the transistor-limited frequency. Hence, the clock frequency peaks at the point where the wire and transistor limited frequency curves intersect. Therefore, repeater insertion

increases the maximum clock frequency from 2 to 2.44 GHz, which is a 22% improvement.

C. Minimum Number of Metal Levels

The number of metal levels required for a macrocell with $A_m = 1.62 \text{ cm}^2$ and $f_c = 2$ GHz is minimized using repeaters. Fig. 6 plots the number of metal levels versus the number of

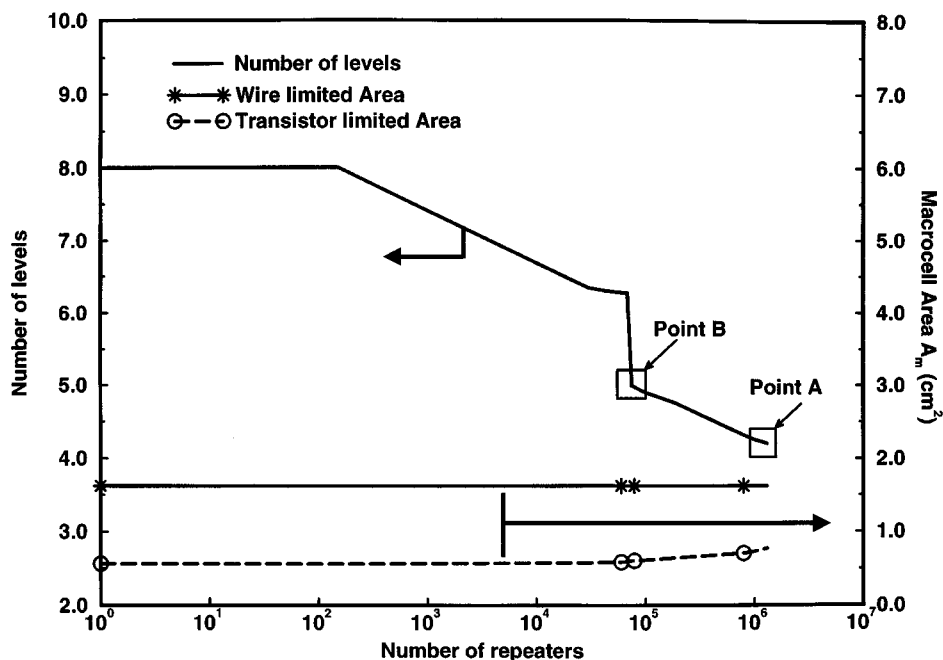


Fig. 6. Minimum number of levels optimization using repeaters for $A_m = 1.62 \text{ cm}^2$ and $f_c = 2 \text{ GHz}$.

repeaters for this design. Increasing the number of repeaters decreases the number of metal levels and increases the transistor-limited area. Once the wire-limited and transistor-limited areas become equal, continued repeater insertion will increase macrocell area. However, in Fig. 6, the minimum number of levels optimization occurs before the convergence of wire and transistor-limited areas when all tiers become saturated with repeaters (point A), such that either they have the quasi-optimum number of repeaters ($\zeta = 0.5$) or have reached minimum wire pitch. Since the levels are grouped in x - y orthogonal pairs, the number of levels should be rounded to the next higher *even* integral value. Therefore, in this example, optimal repeater insertion decreases the number of levels from eight to six, which is a reduction of two metal levels. From a designer's perspective, it would be prudent to choose point B in Fig. 6 as the design point since it has the same number of levels (six levels) but requires a far fewer number of repeaters than point A. The specifications for point B are tabulated in Table III.

D. Minimum Power Dissipation

The total power dissipation of the macrocell with eight metal levels and operating at $f_c = 2 \text{ GHz}$ is minimized using repeaters. Fig. 7 plots the power dissipation versus the number of repeaters for this design. Comparing Figs. 4 and 7, as the macrocell area decreases with an increase in the number of repeaters, the power dissipation in logic gates and interconnects also decreases. Reducing macrocell area decreases the average interconnect length and, hence, reduces the average wiring capacitance. Therefore, the size of the logic gates can be reduced, which decreases the gate capacitance in (12), reducing the total logic gate power. In Fig. 7, the repeater power dissipation increases monotonically because the effect of

decreasing repeater size is overshadowed by the increase in the number of repeaters. From (13), the total interconnect power is also decreased because of 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 minimized*. 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 and results in an increase in the total power dissipation. Moreover, the design becomes power-limited because the power dissipation density is at the maximum permissible limit of 50 W/cm^2 . From Fig. 7, repeater insertion decreases the total power dissipation from 43 to 24 W, a reduction of 44%, that corresponds to the minimum area design point of $A_m = 0.48 \text{ cm}^2$.

For simplicity, (12) 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

$$P_{\text{off}} = W_{\text{tot}} V_{\text{DD}} I_{\text{off}} \quad (15)$$

where W_{tot} is the total macrocell turned off device width and I_{off} is the off-current per device width [21]. From Table III, the W/L ratio of the logic gates decreases 46% 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.

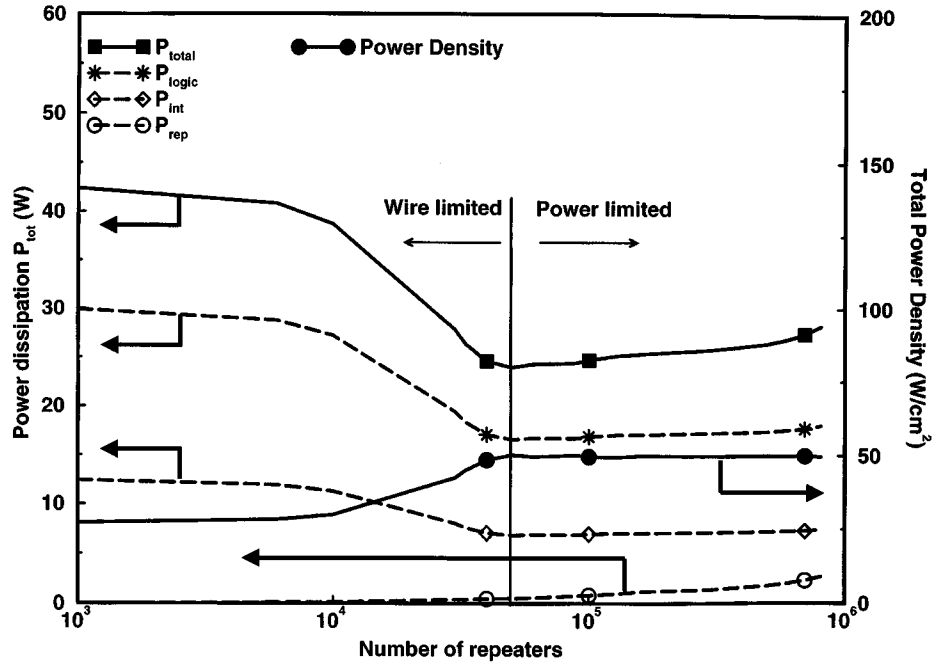


Fig. 7. Minimum power optimization using repeaters for $n \leq 8$ levels and $f_c = 2$ GHz.

The area occupied by repeaters in all the above optimized designs is less than 2% of the macrocell area. In [17], the authors report that the total area consumed by repeaters is about 6% of the total die area. This indicates that the results of the methodology demonstrated in this paper are well within industrially acceptable limits for repeater area.

VI. DISCUSSION OF ASSUMPTIONS

One of the key assumptions used in the n -tier methodology is that the wiring efficiency factor is constant for all metal layers. However, the wiring efficiency factor can be expressed as a product of three factors

$$e_w = e_{\text{rout}} \cdot e_{p/c} \cdot e_{\text{via}} \quad (16)$$

where

- e_{rout} router efficiency;
- $e_{p/c}$ power/ground/clock efficiency
(= 1 - power/ground/clock blockage);
- e_{via} via efficiency (= 1 - via blockage).

The router efficiency is assumed to be a constant for all metal layers. Since power/ground/clock lines are primarily routed on global layers, the blockage caused because of these lines is greater on the upper layers and lesser on the lower layers. Chen *et al.* [7] have described a new via blockage model that improves Sai-Halasz's empirical model in [11], according to [22]. Using an n -tier case study from [8], similar to the ones described in Section III, it is shown in [7] that the via blockage is significant for the lowest two layers and decreases rapidly for the upper metal layers. Because via blockage dominates at lower levels and power/ground/clock blockage dominates

at higher levels, assuming a constant wiring efficiency factor for all levels is a reasonable first order approximation. An enhancement to the n -tier methodology would entail the use of experimentally observed values for power/ground/clock blockage and the use of models in [7] to calculate the wiring efficiency factor for each metal level.

Another assumption is that the maximum permissible time delay is 90% of the clock period for interconnects in all the non-local tiers. However, for chips with large areas and high clock frequencies, this constraint might be impossible to achieve for full chip wires. One of the solutions would be to pipeline the interconnects by inserting flip flops as has been indicated in [17]. This would make the global interconnects much thinner but communication latency would be increased.

VII. CONCLUSION

A new n -tier multilevel interconnect optimization technique has been described in this paper. The key utility of this methodology is to find a set of wiring pitches for each metal level so that performance targets are achieved without wasting system resources. This methodology allows direct computation of the optimum solution without exhaustive iterative techniques saving considerable redesign time. The methodology has been demonstrated to reduce macrocell area by 32%, cycle time by 16%, or number of wiring tracks required on the topmost tier by 62% (although there is no reduction in the actual number of wiring levels), when compared to a conventional baseline design where wire pitches are doubled for every pair of levels.

Also, a top-down repeater insertion methodology has been developed that uses the free silicon area for inserting repeaters in the interconnects. Starting with the upper most tier, this methodology inserts repeaters until all free silicon area is consumed.

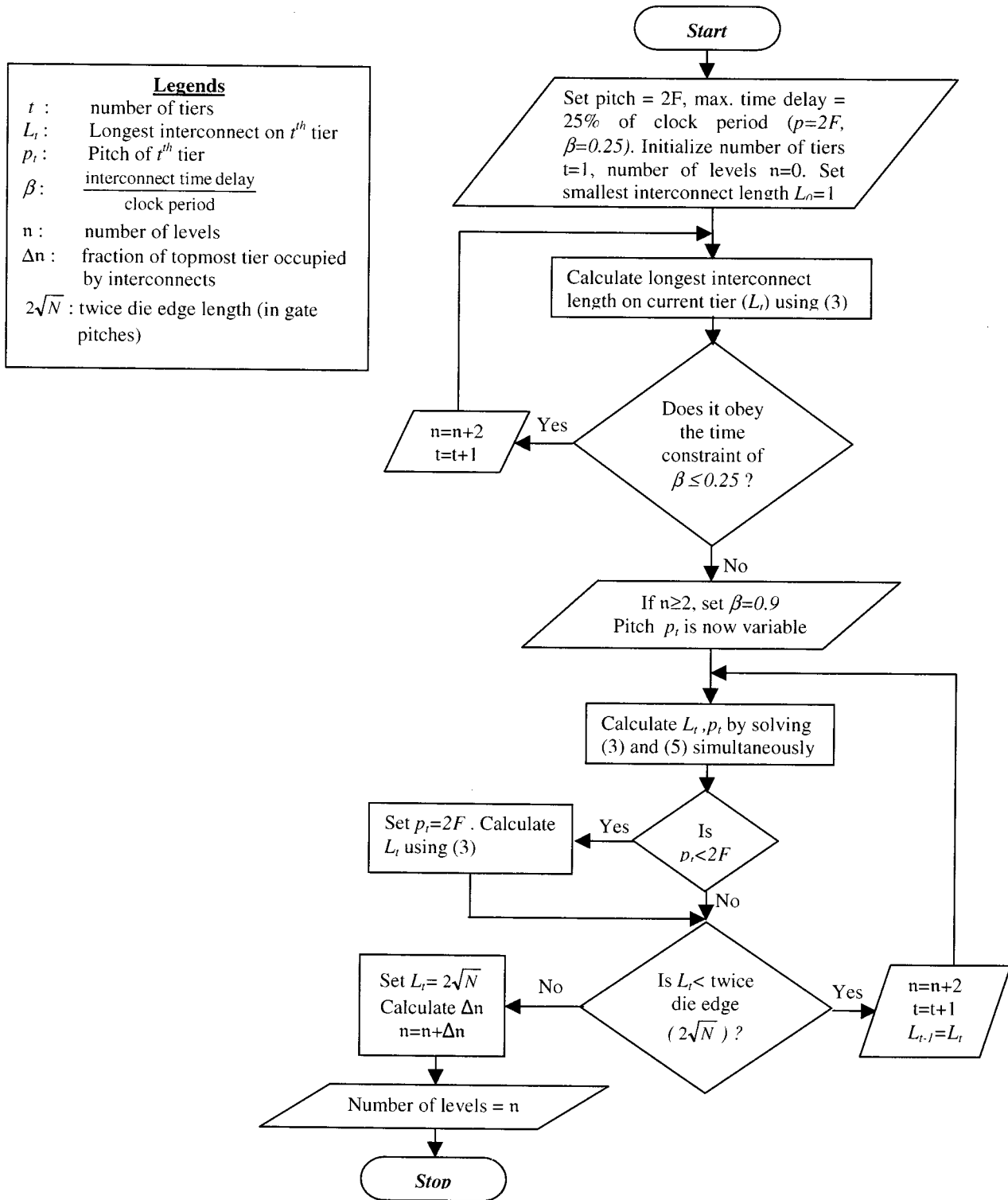


Fig. 8. Flowchart for *n-tier* design methodology.

This methodology has been demonstrated to reduce the macrocell area by 70%, cycle time by 18%, number of metal levels by 25%, or power dissipation by 44%, when compared to an *n-tier* baseline design without repeaters. It has been shown that power dissipation is minimized *simultaneously* by minimizing the macrocell area. These results illustrate the main advantages of extensively utilizing repeaters in a *n-tier* multilevel intercon-

nect architecture to alleviate the restrictive wiring demands of future GSI systems.

In summary, the methodologies presented in this paper demonstrate an *a priori* (i.e., before physical design and layout) optimization of a multilevel interconnect architecture. A general conclusion of this paper is that detailed system level optimization, instead of conventional design heuristics, is

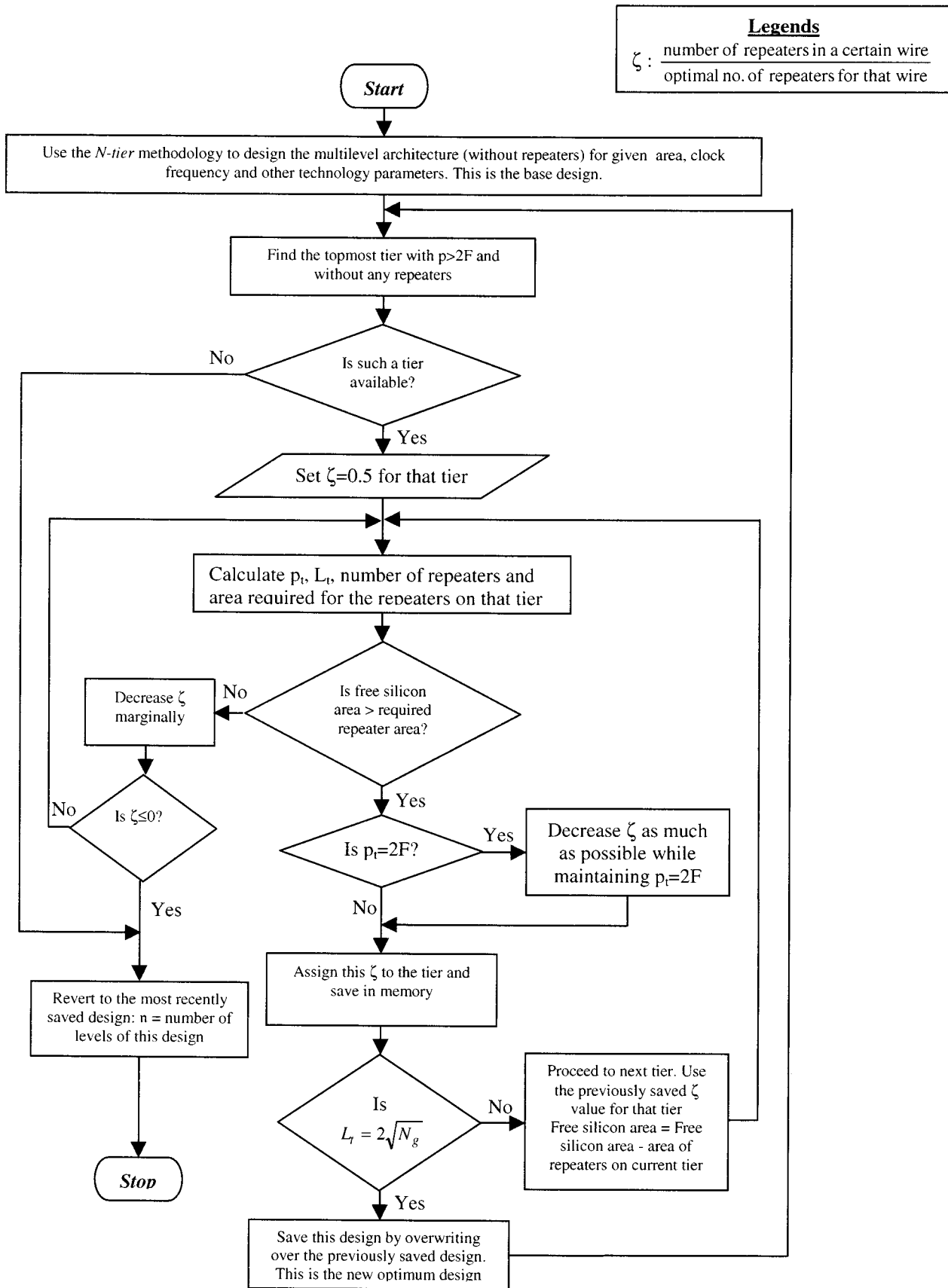


Fig. 9. Flowchart for *n-tier* with repeaters design methodology.

necessary to extract maximum system performance for future GSI designs.

APPENDIX A

FLOWCHART FOR *n-tier* DESIGN METHODOLOGY

See Fig. 8.

APPENDIX B

FLOWCHART FOR *n-tier* WITH REPEATERS DESIGN METHODOLOGY

See Fig. 9.

REFERENCES

- [1] J. D. Meindl, "Low-power microelectronics: Retrospect and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.
- [2] J. A. Davis, R. Venkatesan, A. Kaloyeros, M. Bylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, A. Reif, and J. D. Meindl, "Interconnect limits on gigascale integration (GSI) in the 21st century," *Proc. IEEE*, vol. 89, pp. 305–324, Mar. 2001.
- [3] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.
- [4] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)—Parts I and II," *IEEE Trans. Electron Dev.*, vol. 45, pp. 580–597, Mar. 1998.
- [5] J. A. Davis, "A hierarchy of interconnect limits for gigascale integration," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1999.
- [6] A. B. Kahng and D. Stroobandt, "Wiring layer assignments with consistent stage delays," in *Proc. Int. Workshop System Level Interconnect Prediction*, San Diego, CA, Apr. 2000, pp. 115–122.
- [7] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. D. Meindl, "A compact physical via-blockage model," *IEEE Trans. VLSI Syst.*, vol. 8, pp. 689–692, Dec. 2000.
- [8] R. Venkatesan, J. A. Davis, and J. D. Meindl, "Performance enhancement through optimal *n-tier* multilevel interconnect architectures," in *Proc. 12th Int. IEEE ASIC/SOC Conf.*, Washington, DC, Dec. 1999, pp. 19–23.
- [9] R. Venkatesan, J. A. Davis, K. A. Bowman, and J. D. Meindl, "Optimal repeater insertion for *n-tier* multilevel interconnect architectures," in *Proc. 3rd Int. Interconnect Technology Conf.*, San Francisco, CA, June 2000, pp. 132–134.
- [10] —, "Minimum power and area *n-tier* multilevel interconnect architectures using optimal repeater insertion," in *Proc. Int. Symp. Low Power Electronics and Design*, Rapallo/Portofino Coast, Italy, July 2000, pp. 167–172.
- [11] G. A. Sai-Halasz, "Performance trends in high-end processors," *Proc. IEEE*, vol. 83, pp. 18–34, Jan. 1995.
- [12] T. Sakurai, "Closed form expressions for interconnection delay, coupling and crosstalk in VLSI's," *IEEE Trans. Electron Dev.*, vol. 40, pp. 118–124, Jan. 1993.
- [13] *International Technology Roadmap for Semiconductors*, Semiconductor Industry Assoc., 1999.
- [14] W. E. Donath, "Wire length distribution for placement of computer logic," *IBM J. Res. Development*, vol. 2, no. 3, pp. 152–155, May 1981.
- [15] S. Tyagi *et al.*, "A 130 nm generation logic technology featuring 70 nm transistors, dual Vt transistors and 6 layers of Cu interconnects," *Proc. IEDM*, pp. 567–570, Dec. 2000.
- [16] H. B. Bakoglu and J. D. Meindl, "Optimal interconnection circuits for VLSI," *IEEE Trans. Electron Dev.*, vol. ED-32, pp. 903–909, May 1985.
- [17] R. McInerney, M. Page, K. Leeper, T. Hillie, H. Chan, and B. Basaran, "Methodology for repeater insertion management in the RTL, layout, floorplan and fullchip timing databases of the Itanium™ microprocessor," in *Proc. Int. Symp. Physical Design*, San Diego, CA, Apr. 2000, pp. 99–104.
- [18] J. C. Eble, "A generic system simulator with novel on-chip cache and throughput models for gigascale integration," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1998.
- [19] 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, pp. 410–414, Oct. 1999.
- [20] T. Sakurai and A. R. Newton, "Delay analysis for series-connected MOSFET circuits," *IEEE J. Solid-State Circuits*, vol. 26, pp. 122–131, Feb. 1991.
- [21] Y. Taur *et al.*, "CMOS scaling into the nanometer regime," *Proc. IEEE*, vol. 85, pp. 486–504, Apr. 1997.
- [22] A. B. Kahng, S. Mantik, and D. Stroobandt, "Requirements for models of achievable routing," in *Proc. Int. Symp. Physical Design*, San Diego, CA, Apr. 2000, pp. 4–11.



Raguraman Venkatesan (S'99) was born in Sindri, India, in 1976. He received the B.Tech. degree from the Indian Institute of Technology, Bombay, and the M.S. degree from the Georgia Institute of Technology, Atlanta, in 1998 and 2000, respectively, both in electrical engineering. He is currently working toward the Ph.D. degree in the Gigascale Integration Group, Georgia Institute of Technology.

In summer 2000, he worked on determining optimum interconnect system dimensions for the next generation microprocessors while interning with the Logic Technology Development Group, Intel Corporation, Hillsboro, OR. His research interests include designing optimal multilevel wiring networks and modeling inductive effects in high speed interconnects.

Mr. Venkatesan was awarded the Intel Graduate Fellowship for the academic year 2001–2002.



Jeffrey A. Davis (S'94–M'00) received the B.E.E., M.S.E.E., and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1993, 1997, and 1999, respectively.

He joined the faculty at the Georgia Institute of Technology as an Assistant Professor in 1999. His current research interests are in the areas of interconnect modeling, high-speed area efficient interconnect circuits, interconnect-centric design methodologies, and optimal multilevel interconnect network design for future GSI processors.

Dr. Davis is currently the general chair of the 2002 System Level Interconnect Prediction (SLIP) Workshop (www.sliponline.org).



Keith A. Bowman (S'97) received the B.S. degree in electrical engineering from North Carolina State University, Raleigh, in 1994 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1995 and 2001, respectively.

His doctoral research focused on the impact of power consumption and parameter fluctuations on future circuit performance to enable opportunities for further advancement of gigascale integration. He is currently a Senior Computer-Aided Design (CAD)

Engineer in the Technology CAD Division at Intel Corporation, Hillsboro, OR. In the summer of 2000, while interning at the Technology CAD Division, Intel Corporation, Santa Clara, CA, he performed research in modeling the impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for a 0.25- μm microprocessor.



James Meindl (M'56–SM'66–F'68–LF'97) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Carnegie Institute of Technology (Carnegie Mellon University), Pittsburgh, PA, in 1955, 1956, and 1958, respectively.

He is Director of the Interconnect Focus Center, a multi-university research effort, managed jointly by the Microelectronics Advanced Research Corporation and the Defense Advanced Research Projects Agency for DoD. He was Senior Vice President of Academic Affairs and Provost of Rensselaer

Polytechnic Institute, Troy, NY, from 1986 to 1993. He was with Stanford University, Stanford, CA, from 1967 to 1986 as the John M. Fluke Professor of electrical engineering, Associate Dean for Research, School of Engineering, Founding Director of the Center for Integrated Systems, Director of the Electronics Laboratories, and Founding Director of the Integrated Circuits Laboratory. He is also the Director of the Joseph M. Pettit Microelectronics Research Center and the Joseph M. Pettit Chair Professor of Microelectronics at the Georgia Institute of Technology. His current research interests focus on physical limits on gigascale integration.

Dr. Meindl is a Fellow of the American Association for the Advancement of Science and a member of the American Academy of Arts and Sciences and the National Academy of Engineering and its Academic Advisory Board. He received the Hamerschlag Distinguished Alumnus Award from Carnegie Mellon University in 1996, the Benjamin Garver Lamme Medal from ASEE in 1991, the IEEE Education Medal in 1990, and the IEEE Solid-State Circuits Medal in 1989. He has also been awarded the IEEE Electron Devices Society's J. J. Ebers Award, the 1997 Hamerschlag Distinguished Alumnus Award from Carnegie Mellon University, as well as five outstanding paper awards from the IEEE ISSCC. He also received the 1999 SIA University Research Award, the IEEE Third Millennium Medal, and, most recently, the Georgia Institute of Technology 2001 Distinguished Professor Award.