Improved Delay Prediction for On-Chip Buses.

Real G. Pomerleau
North Carolina State University
CACC Box 7914
Raleigh, NC, 27695-7914
(919) 503-2007
rgpomerl@eos.ncsu.edu

Paul D. Franzon
North Carolina State University
CACC Box 7914
Raleigh, NC, 27695-7914
(919) 515-7351
paulf@eos.ncsu.edu

Griff L. Bilbro
North Carolina State University
CACC Box 7914
Raleigh, NC, 27695-7914
(919) 515-5101
glb@eos.ncsu.edu

ABSTRACT

In this paper, we introduce a simple procedure to predict wiring delay in bi-directional buses and a way of properly sizing the driver for each of its port. In addition, we propose a simple calibration procedure to improve its delay prediction over the Elmore delay of the RC tree. The technique is fast, accurate, and ideal for implementation in floorplanner during behavioral synthesis.

Keywords

RC wiring delay, High-Level Synthesis, Floorplanning, Buffer Optimization, Interconnect optimization.

1. INTRODUCTION

The problem of including interconnection delay during High Level Synthesis (HLS) has been addressed in a limited way by several researchers and its importance was recently highlighted by a panel of industry experts [9]. Prabhakaran and Banerjee [8] provide one of the latest publications on this subject. Their research describes a procedure to integrate floorplanning with the other sub-problems of behavioral synthesis. Their procedure to compute the delay was limited to the standard point-to-point formula with the Manhattan distance between modules. In VLSI however, depending on the strength of the driver and the wiring parameters, interconnect delay can be made arbitrary large or arbitrary small. In addition, buses are common structures used in design. The use of a simple point-to-point delay formulation without consideration of the drivers, wiring levels or topologies may lead to sub-optimal behavioral design. To produce better high-behavioral design especially in the area of bus partitioned design a simple, fast and accurate technique is needed.

1.1 Problem Statement

Our interest lies in integrating some aspects of physical design namely wiring delay and power density requirement directly into behavioral synthesis. Behavioral synthesis schedules, allocates and binds algorithmic operations to a given architecture. Operations (usually additions, multiplication, etc) represented by functional units are then interconnected together often in a bus structure to meet the requirements of the design.

DAC 99, New Orleans, Louisiana

Figure 1-1 shows a pictorial representation of a basic functional unit entities.

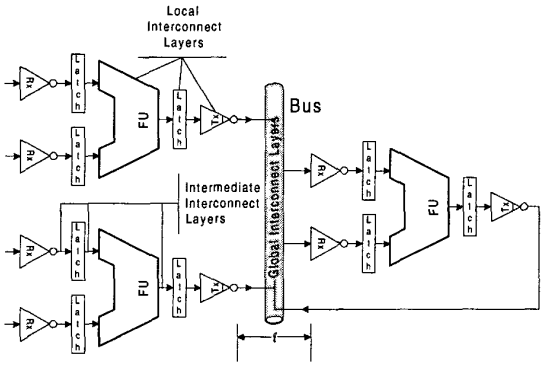

Figure 1-1 Basic Functional Unit Component

The literature refers to three different level of interconnect [4], [6]. We show a possible allocation of these levels in Figure 1-1. With each of these levels is associated a net length, wiring size and pitch. Of these three levels only the first level, the local interconnects, is defined in the SIA roadmap [12]. The other two levels are sized to accommodate increased net lengths and higher frequencies. Other then the proposed and potentially expensive *fat-wire technology* [10] little exists in terms of guidelines for these upper levels. Interconnect delay in these upper layers should be scaled or traded-off for functional units delay and vice-versa. What is needed is a fast, accurate and simple way to compute delay in buses that are driven by properly sized drivers.

2. DELAY PREDICTOR

In synchronous systems, delays are usually reported between latch boundaries. This makes the delay dependent on the output driver and wiring parameters of the interconnecting structure. Our interest lies in bus-oriented architectures with tapered drivers. The design of an optimal tapered driver has been widely investigated (see for example [1], and [2]). We assume that the wiring of the driver makes use of the deep-submicron layers as reported in the SIA roadmap and modified the optimization formulation of the drivers to include their wiring delay as a function of their growth factor. Details are presented in Appendix 5.1. Figure 2-1 shows a more detailed model of a bi-directional connection between the output of a transmitter latch and the input of a receiver latch.

The global interconnects section shown, in Figure 2-1 is represented as a point-to-point connection, but in fact, depicts an n-port bus. Reduction of an n-port bus to an equivalent point-to-point net is presented in Appendix 5.3.

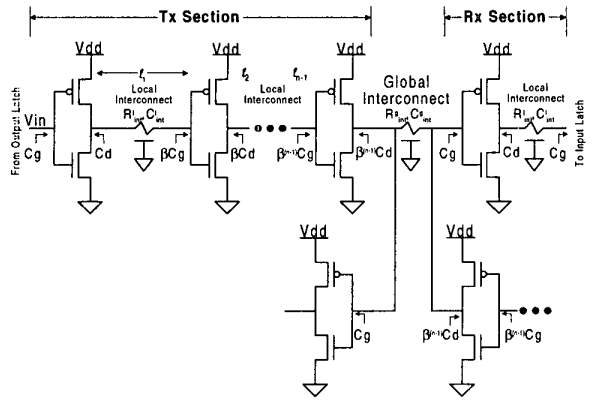

Figure 2-1 Tapered Bus Driver/Receiver Model

After considerable algebraic manipulation, we can write the delay for the tapered bus driver/receiver from a port *i* to a port *j* as

$$\tau_{ij} = \overline{\overline{K}}_2 \left(\frac{\left((\beta+1)(\beta^{2n}+4\beta^{n+1}) + \beta^2(4n(\beta-1)+7\beta-17) \right)}{16\beta^2(\beta-1)} \right) R_{\text{int}}^{\ell} C_{\text{int}}^{\ell} \ell_0^2 +$$

$$\overline{\overline{K}}_1 \left(\frac{\beta\left(\beta(n-2\beta^{-n}-1)+6\right)-n-3}{4(\beta-1)} R_{drv} C_{\text{int}}^{\ell} + \left(\frac{(\beta+\beta^n)^2}{4\beta(\beta-1)} - \frac{1}{\beta-1} \right) R_{\text{int}}^{\ell} C_g \right) \ell_0 +$$

$$\overline{\overline{K}}_0 \left(\beta(n-1)+n\xi+1\right) R_{drv} C_g +$$

$$\overline{K}_2 M W_{ij}^{l^2} R_{\text{int}}^{g} C_{\text{int}}^{g} \ell_g^2 +$$

$$\overline{K}_1 \left(\frac{R_{drv}}{\beta^{n-1}} C_{\text{int}}^{g} + W_{ij}^{l} \left(\beta^{n-1}\xi + \beta^0\right) R_{\text{int}}^{g} C_g \right) \ell_g +$$

$$\overline{K}_0 N_p \frac{\beta^{n-1}\xi + \beta^0}{\beta^{n-1}} R_{drv} C_g \qquad (2.1)$$

This equation is in the same form as Sakurai's [11] delay formulation. The equation is split into two components: the local interconnects component, which represents the tapered buffer delay, and the global component. The meaning of the variables are as follows: β is the growth factor of the buffer, n is the number of stages, R_{int}^{ℓ}, C_{int}^{ℓ}, R_{int}^{g} and C_{int}^{g} are the per-unit resistance and capacitance of their respective layers. ξ is the ratio of the drain/source capacitance to the gate capacitance as defined in [2] and [5]. ℓ_0 is the initial interconnection length of two minimal size device and ℓ_g is the total length of the bi-directional bus while N_p is the number of ports. The $\overline{K}_i$ are the line calibration coefficients and the Ws are the weighting factors used to appropriately scale the delay of each port of the bus relative to a two port bus. M is the Elmore scaling factor which must be set to 2. Finally, R_{drv} and C_g represent the driver resistance and gate capacitance of a minimum size devices. It can be shown that the part representing the local interconnect in equation (2.1) can be reduced to the expression derived in [1] for a tapered buffer.

The use of the $\overline{K}$ factors, which we refer as calibration coefficients, allows a trade-off between increased accuracy of the equation and range in its application. Since the local interconnects, are point-to-point connections we use Sakurai's derived coefficient values (IE $\overline{\overline{K}}_2 = 0.377$, $\overline{\overline{K}}_1 = \overline{\overline{K}}_0 = 0.693$). For the global interconnects, setting $\overline{K}_2 = \overline{K}_1 = \overline{K}_0 = 1$ and $M = 1/2$ would result in the computation of Elmore's delay [9] (an upper bound). However sometime greater accuracy is required than what can be obtained using Elmore's delay. Setting these coefficients to Sakurai's coefficient tends to underestimate the delay when the number of ports N_p is greater than two. For example, using line widths ranging from 0.6 to 4 μm and heights above a ground plane from .5 to 2 μm, Sakurai's delay equation underestimates the delay by almost 20% when compare to SPICE using a fully distributed RLC model. A more accurate estimate lies somewhere in between. To increase the accuracy of the equation we developed a simple calibration procedure to re-scale these coefficients. Details of the procedure can be found in Appendix 5.2.

The re-calibration of the $\overline{K}$ coefficients is performed by running a Spice simulation on the circuits shown in Figure 2-2 for different line lengths. The driver on-resistance, R_{drv}, and gate capacitance, C_g, values are set consistent with minimum size devices for the process at hand. Results for a .250μ process are summarized in Table 2-1.

The resulting effect is obvious. These coefficients are slightly larger than Sakurai's coefficients. We have observed that if these coefficients are within these ranges of values reasonable variations in the growth factor and line parameters produce consistent results with Spice simulation.

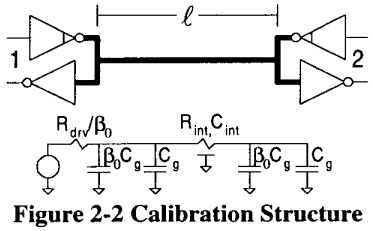

Figure 2-2 Calibration Structure

Cal Run	$\overline{K}_2$	$\overline{K}_1$	$\overline{K}_0$
A	0.3938	0.8280	1.3602
B	0.3885	0.8362	1.1528
C	0.3963	0.8246	1.6270

Table 2-1 Calibration of the K coefficients

3. Accuracy of Delay Estimates Compared to Spice Simulation.

We tested the procedure on the five port bus shown in Figure 3-1. Being bi-directional, it has three distinctive tree structures. We tested the procedure on length 15kμ and 8kμ. The tables display the resulting 50% delay from the driving port to the receiving ports in nano-seconds. The mapping of this five port bus to an equivalent two port is achieved by computing the appropriate scaling factors Ws as outlined in Appendix 5.3.

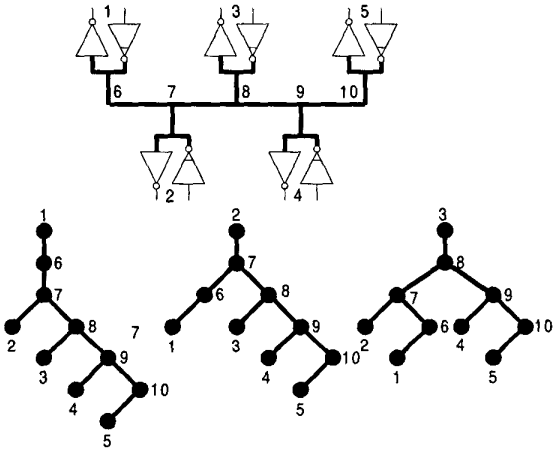

Figure 3-1 Example of a Five-Port Bus.

The rows labeled "Calibrated Coef" consist of setting the $\overline{K}$ coefficients to the value shown in Table 2-1. The rows labeled "Elmore Coef" consist of setting all the $\overline{K}s$ to one as discussed in the previous section. It can be observed that using the simple calibration procedure significantly enhances the accuracy of equation (2.1) when compared to Spice. Of course the actual CPU time to compute the port-to-port delay using equation (2.1) is practically negligible compared to Spice.

Driver on Port #1

ℓ_g (μm)	P#1	P#2	P#3	P#4	P#5
15k (Spice)	1.4397	2.4443	2.7415	2.9232	2.9997
Calibrated Coef		2.6368	2.8516	2.9884	3.0473
Elmore Coef		4.0012	4.5473	4.895	5.0442
8k (Spice)	0.9393	1.2071	1.2907	1.3427	1.3647
Calibrated Coef		1.2772	1.3395	1.3793	1.3964
Elmore Coef		1.7547	1.9115	2.0113	2.0543

Driver on Port #2

ℓ_g (μm)	P#1	P#2	P#3	P#4	P#5
15k (Spice)	2.2347	1.6391	2.4629	2.6453	2.7217
Calibrated Coef	2.4029		2.5588	2.6956	2.7545
Elmore Coef	3.4061		3.8029	4.1505	4.2998
8k (Spice)	1.1477	0.9835	1.2093	1.2271	1.2834
Calibrated Coef	1.2095		1.2546	1.2944	1.3115
Elmore Coef	1.5839		1.6978	1.7976	1.8406

Driver on Port #3

ℓ_g (μm)	P#1	P#2	P#3	P#4	P#5
15k (Spice)	2.4744	2.3977	1.7080	2.3977	2.4744
Calibrated Coef	2.5397	2.4808			
Elmore Coef	3.7538	3.6045			
8k (Spice)	1.2120	1.1898	.9983	1.1898	1.2120
Calibrated Coef	1.2492	1.2321			
Elmore Coef	1.6838	1.6408			

4. Conclusion and Future Work

We reported on a simple procedure to compute delay in bi-directional buses driven with optimal tapered buffers. The equation is divided into two parts a local interconnect part and a global interconnect part. The local part of the equation accounts for the effects of the local interconnect by making the interconnecting line a function of the buffer growth factor. The global part of the equation handles bus type structures by computing weighting factors to properly scale the delay to each port. If the all segments of the bus have a predefined relationship to one another, these weighting factors can be computed in closed form. The accuracy of the global part of the equations can easily be increased by using calibration coefficients. These coefficients are obtained from a simple Spice simulation. However, as with all fitting procedures care must be taken to properly choose a valid range. We are currently, investigating this range. We are also in the process of enhancing this equation to allow individual driver optimization base on their locations on the bus. Because of its computational speed, flexibility, and enhanced accuracy we feel that this technique is particularly suited to behavioral synthesis with floorplanning.

5. Appendix

5.1 Accounting for the Delay in Local Interconnect Due to the Growth Factor β.

It is well known that in deep-submicron region interconnect delay dominates the gate delay. We attempt to account for this in the computation of the growth factor β by making the local interconnection length, ℓ_n, between each buffers stage, as shown

in Figure 2-1, a function of β. The length, ℓ_n, between any two stages can then be estimated using equation (5.1)

$$\ell_n = \frac{\ell_0}{4}\left(2+\beta^{n-1}(\beta+1)\right) \tag{5.1}$$

For example, Figure 5-1 demonstrates the growth in ℓ_0 for a growth factor $\beta = 2$ where the initial length ℓ_0 is set to length required to interconnect two minimum size devices.

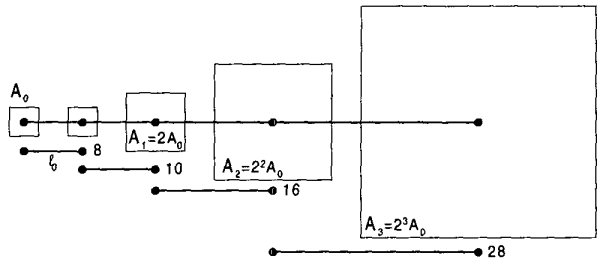

Figure 5-1 Local Interconnection Length dependency on the Growth Factor β.

5.2 Calibration procedure for the K coefficient

To be able to fit delay data to a second order polynomial in ℓ, where ℓ represents the length of the interconnect line as depicted by equation (5.2), would enhance the accuracy of our predictions.

$$\tau_d = \mathrm{K}_2\ell^2 + \mathrm{K}_1\ell + \mathrm{K}_0 \tag{5.2}$$

Since we are interesting in finding optimal driver size and predicting delay for bi-directional busses, we choose to calibrate our system to the structure shown in Figure 2-2.

The usual procedure for fitting data is to adjust one variable while all others remain constant. This means that the fitting coefficient will be computed for pre-determined driver strength, line parameters and loading conditions. Consequently, to allow for different value of growth factor β these fitting coefficients must be re-scaled accordingly.

To re-scale the fitting coefficients we re-write Sakurai's delay equation in the following form (5.3)

$$\tau_d = \overline{\mathrm{K}_2}R^0_{\mathrm{int}}C^0_{\mathrm{int}}\ell^2 + \overline{\mathrm{K}_1}\left(\begin{array}{l} R^0_{\mathrm{int}}(\beta_0\xi_0+1)C_g^{\min} \\ +\dfrac{R^{\min}_{drv}}{\beta^0}C^0_{\mathrm{int}} \end{array}\right)\ell \tag{5.3}$$
$$+\overline{\mathrm{K}_0}\frac{R^{\min}_{drv}}{\beta^0}(\beta_0\xi_0+1)C_g^{\min}$$

Equating the coefficient of ℓ between (5.2) and (5.3) and solving for the calibration coefficients $\overline{\mathrm{K}s}$ we get (5.4).

$$\overline{\mathrm{K}_2} = \frac{\mathrm{K}_2}{R^0_{\mathrm{int}}C^0_{\mathrm{int}}},$$
$$\overline{\mathrm{K}_1} = \frac{\mathrm{K}_1}{\left(\dfrac{R^{\min}_{drv}}{\beta_0}C^0_{\mathrm{int}} + (\beta_0\xi_0+1)C_g^{\min}R^0_{\mathrm{int}}\right)}, \tag{5.4}$$
$$\overline{\mathrm{K}_0} = \frac{\mathrm{K}_0}{\dfrac{R^{\min}_{drv}}{\beta_0}(\beta_0\xi_0+1)C_g^{\min}}$$

Where $R^{\min}_{drv}$, and $C_g^{\min}$ are resistance and capacitance value of a minimum size device. β_0 sets the size of the driver used for calibration. When choosing β_0 it is important to remember the validity range of the RC delay model [4]. R^0_{int} and C^0_{int} are the line parameters. For calibration purposes ξ_0 may be set to one. The Ks are obtained by simultaneously solving the three sets of simulation data performed at different values of ℓ as shown in equation (5.5).

$$\begin{bmatrix} K_2 \\ K_1 \\ K_0 \end{bmatrix} = \begin{bmatrix} \ell_0^2 & \ell_0 & 1 \\ \ell_1^2 & \ell_1 & 1 \\ \ell_2^2 & \ell_2 & 1 \end{bmatrix}^{-1}\begin{bmatrix} \tau_0 \\ \tau_1 \\ \tau_2 \end{bmatrix} \tag{5.5}$$

5.3 Reduction of an n-port bus to an equivalent point-to-point representation.

In this section, we formulate a simple technique that allow scaling the delay of RC trees networks to an equivalent point-to-point topology. The idea is to re-formulate Elmore's equation to obtained a weighting function that scale the delay from any sources to any sinks in a bi-directional bus. For a given bus topology these weight sets can be formulated in a close form equation and then used in the standard point-to-point formula. This makes it very convenient in behavioral optimization.

The Elmore delay from a root node i to a leave node j is given by equation (5.6).

$$\tau_{i\to j} = R_i\sum_{k\in Desc(i)}C_k + \sum_{m\in \mathrm{P}_j}R_m\left(\frac{C_m}{2} + \sum_{n\in Desc(j)}C_n\right) \tag{5.6}$$

Where R_i is the resistance of the root node (driver resistance), R_m, C_m, C_n are the interconnect resistance and capacitance of the individual branches in the tree including any load capacitance. P_j is the path from node i to node j and $Desc(j)$ is the set of nodes that are descendants of node j. Equation (5.6) may easily be expressed in matrix notation as:

$$\tau_{ij} = R_i\left(\mathbf{C}^{\mathrm{T}}\mathbf{L} + \mathbf{e}^{\mathrm{T}}\mathbf{C}_{\mathrm{L}}\right) + \frac{1}{2}\mathbf{L}^{\mathrm{T}}\mathbf{C}_{\mathrm{b}}\mathbf{Q}_{\mathrm{ij}}\mathbf{R}_{\mathrm{b}}\mathbf{L} + \mathbf{C}_{\mathrm{ij}}^{\mathrm{T}}\mathbf{R}_{\mathrm{b}}\mathbf{L} \quad (5.7)$$

Where $\mathbf{C}^{\mathrm{T}}$ is the per unit capacitance value vector of each branch in the net, $\mathbf{L}$ is the branch length vector, $\mathbf{e}^{\mathrm{T}}$ is a unit vector with a one at each load port location and zeros elsewhere, and $\mathbf{C}_{\mathrm{L}}$ is the port load vector. $\mathbf{C}_{\mathrm{b}}$ and $\mathbf{R}_{\mathrm{b}}$ are the branch matrix with their respective per-unit value on the main diagonal and zeros elsewhere. The $\mathbf{Q}_{\mathrm{ij}}$ and $\mathbf{C}_{\mathrm{ij}}$ are the path and capacitance vector matrices.

The construction of these matrices is as follow: a value of one is placed on its main diagonal for each branch in the forward path. The descendent of the nodes located on the forward path are then accounted for by filling the remaining rows either above or below the diagonal with a value of 2. The capacitance vector, $\mathbf{C}_{\mathrm{ij}}$, represents the sum of the loads for each descendent node on the forward path.

Now if we let $\mathbf{W}$ be a vector representing the weighted length of each segment of the nets such as $\ell_{ij} = w_{ij}\ell$, $\sum_{\forall i,j} w_{ij} = 1$ and $\sum_{i,j \in Desc(1)} \ell_{ij} = \ell$ we may then easily put (5.7) in the same form as given in [11]. Since equation, (5.7) use be used directly in providing that proper modification are made for β and ξ. However, our goal is to use this equation in a pre-layout stage were it must quickly be evaluated. The most obvious simplification is to assume a topology and make the per-unit resistance and capacitance equal for each segment.[1] Equation (5.8) may then be simply re-written as

$$\tau_{ij} = \tfrac{1}{2} w_{ij}^{l^2} R_{\mathrm{int}}^{g} C_{\mathrm{int}}^{g} \ell_g^2 + \left(R_i C_{\mathrm{int}}^{g} + w_{ij}^{l} R_{\mathrm{int}}^{g} C_L\right)\ell_g + R_i \mathrm{N}_p C_L \quad (5.8)$$

Where $w_{ij}^{\ell^2} = \mathbf{W}^{\mathrm{T}}\mathbf{Q}_{\mathrm{ij}}\mathbf{W}$ and $w_{ij}^{l} = \mathbf{W}\mathbf{C}_{\mathrm{ij}}$ are the weighting factors which can now be either evaluated in closed form or presented in table form. N_p is the number of ports and R_{int}^{g} and C_{int}^{g} are the per-unit resistance and capacitance of the global interconnect. For example, if we assume a linear Steiner tree topology where each segment of the net are of equal length we may easily derive closed form solution for the weighting factors $W_{1m}^{\ell^2}$ and W_{1m}^{ℓ} for any size buses.

$$W_{1m}^{l^2} = 1/2(m-1) + m(2\mathrm{N_P} - m)$$
$$W_{1m}^{\ell} = 1/2(2\mathrm{N_P} - m + 1)m$$

Where m is the port number.

[1] This assumption is certainly reasonable if each of the segments is laid-out on the same layer. The assumption is also very good for two layers stacked-up between two power planes (asymmetric stripline). For buried microstrip structures keeping the ratio of the line width to its height above the plane constant and adjusting their thickness to match their resistivity would also produce similar segment values. In the worst case the unit resistance and capacitance for each layer could simply be averaged.

REFERENCES

[1] Bakoglu H. B., "Circuits, Interconnections, and Packaging for VLSI", *Addison-Wesley Publishing Company,* 1990.

[2] Choi J-S, and Lee K. "Design of CMOS Buffer for Minimum Power-Delay Product", *IEEE Journal of Solid-State Circuits,* 29(9):1142-1145, 1994.

[3] Cong J., He L., Khoo K-Y, Koh C-K, Pan Z. "Interconnect Design for Deep Submicron ICs", *IEEE/ACM International Conference on Computer-Aided Design, 478-485, 1997*

[4] Deutsche A., et al, "When are Transmission-Line Effects Important for On-Chip Interconnections", *47th Electronic Components & Technology Conference,* 704-711, 1997.

[5] Li N., Haviland G., Tuszynski A., "CMOS Tapered Buffer", *IEEE Journal of Solid-State Circuits,* 25(4):1005-1008, 1990.

[6] Lynch W., "Power Supply Distribution and Other Wiring Issues for Deep-Submicron ICs", *NCSU VLSI Seminar,* 1997.

[7] Prabhakaran P., and Banerjee P, "Simultaneous Scheduling, Binding, and Floorplanning in High-Level Synthesis", *Proc of the 11th International Conference on VLSI Design*, 428-434, 1998.

[8] Pedran M., "Panel: Physical Design and Synthesis Merge or Die", *Proc. 32sd Design Automation Conference,* 238-239, 1995.

[9] Penfield P., and Rubinstein J., "Signal Delay in RC Tree Networks", *Proc. 18th Design Automation Conference,* 613-617, 1981.

[10] Sai-Halasz G., "Performance Trends in High-end Processors", *IEEE Proceeding,* 83(-):20-36, 1995.

[11] Sakurai T., "Closed-Form Expressions for Interconnection Delay, Coupling, and Crosstalk in VLSI's", *IEEE Transactions on Electronic Devices,* 40(1):118-124, 1993.

[12] Semiconductor Industry Association, "The National Technology Roadmap for Semiconductors", 1997 Edition.