A High-Level Compact Pattern-Dependent Delay Model for High-Speed Point-to-point Interconnects

Tudor Murgan, Massoud Momeni, Alberto García Ortiz, and Manfred Glesner
Institute of Microelectronic Systems
Darmstadt University of Technology
Karlstr. 15, D-64283 Darmstadt, Germany
{tam,mmomeni,agarcia}@mes.tu-darmstadt.de

ABSTRACT

This work introduces an extended linear pattern-dependent model for high-level signal delay estimation in high-speed very deep submicron point-to-point interconnects. The proposed model accurately predicts the delay in both inductively and capacitively coupled lines for the complete set of the switching patterns and not only for capacitively coupled lines or worst-case delay as in previous works. We also consider process variations in the formulation of the model and propose a moment-based approach for the inclusion of variations. The accuracy of the model has been assessed by means of extensive experiments. Moreover, we show how the model can be applied at high levels of abstraction in order to explore coding-based alternatives to improve throughput.

1. INTRODUCTION

The rapid evolution in process technology allows the integration of increasingly complex systems in VLSI chips operating at continuously rising frequencies. On the one hand, the increase in system complexity with shrinking device features implies longer global interconnection lengths, and thus larger delays. On the other hand, as the number of devices per area increases, a greater number of vertical wiring layers are required and the cross-sectional dimension of interconnections is reduced. This trends make interconnections tightly coupled and undesired crosstalk interference appears. Moreover, strict performance requirements demand global wires to carry high-frequency currents and exhibit low resistance. All the aforementioned factors rendered on-chip inductance as a main factor to be taken into account for timing and crosstalk analysis [10].

In order to address timing issues at higher levels of abstraction, accurate models capable to predict pattern-dependent signal delay are required. This is mandatory if delay-aware coding is to be employed in high-speed very deep sub-micron (VDSM) buses. Current techniques are restricted only to non-inductively coupled interconnects because of a lack of proper compact models [22, 23]. Some efforts have been done to identify worst-case switching patterns in inductively coupled lines [25]. However, they cannot predict the delay for a given input switching pattern. Further, the delay coding limits and delay improvement methods developed in [22] for capacitive coupling do not hold in the more general case of inductively-coupled lines and have to be revised.

The goal of this paper is to develop a high-level model which predicts delays in dedicated point-to-point interconnects for all switching patterns. The technique has to take into account the pattern-dependent behavior of the delay as well as the effect of interconnect-related process variations on delay. Moreover, we show how the model can be easily and rapidly applied at high levels of abstraction in order to explore coding-based alternatives for throughput improvement and to assess the efficiency of encoding schemes. In contrast with previous models, the developed model is able to predict that coding techniques for throughput improvement are less efficient when inductive coupling effects are not negligible.

ICCAD'06 November 5-9, 2006, San Jose, CA

This work is organized as follows: Sec. 2 describes the general approach for delay estimation and the interconnect model that we employ in this work as well as the methodology for the experimental set-up. The extended linear delay model is developed and validated in Sec. 3. Additionally, the effects of process variations are included into the model. Afterwards, Sec. 4 gives an interpretation of the results regarding the ELD model and Sec. 5 shows how the ELD model can be employed to analyze coding-based throughput improvement opportunities and assess the efficiency of encoding schemes. Finally, the paper ends with some concluding remarks.

2. EMPLOYED APPROACH AND INTERCONNECT MODELS

Classically, the analysis of the total delay induced by a buffer driving an interconnect network has been addressed by splitting the problem in two simpler ones as shown in Fig. 1: separate estimation of gate delay and intrinsic wire delay, also referred to as time-of-flight. In order to determine the equivalent delay of a gate, the complete network is abstracted as an equivalent load, typically an effective capacitance [2, 19]. The delay is then just a function of the input transition time and this equivalent load. After determining the transition time at the gate output, the waveform at the gate output is approximated with a saturated ramp. The interconnect delay is then calculated using this waveform as line input. Consequently, the key elements of this two-step methodology are the estimation of the equivalent (effective) capacitance and the delay of the wire. When inductive effects appear, the process becomes much more complex and this issue is only partly solved. Some efforts have been put in characterizing the intrinsic delay of a buffer when the load is dominated by inductive lines [10]. Further, in order to cope with this problem, in [2] the saturated ramp is replaced by a piecewise equivalent voltage source.

In this work, we focus on the intrinsic delay of the wire since we believe that this effort is the first step towards a solid and complete model for delay in general interconnects. We generalize the delay model proposed in [22] for non-inductive capacitively-coupled buses to include inductive coupling effects. For the purpose of characterizing the on-the-fly delay, we can use a simple driver model. By considering a trapezoidal signal together with a series impedance as driver model, i.e. a parameterized Thévenin model, we can get in the considered cases a precise approximation of the real waveform. The Thévenin voltage source is generally modeled via an equivalent driver resistance and a saturated ramp voltage characterized by a transition time and a delay [2].

The aforementioned approach provides an accurate and simple model for analyzing the intrinsic delay of the wire. The drawback is that the characterization of the driver should provide not only the intrinsic delay, but also the output transition time and the equivalent series impedance as a function of the equivalent capacitance and input transition time. For the purposes of characterizing the intrinsic delay of the wire, we assume that this parameters are known. It is important to mention that the delay approximation for the driving point is fairly insensitive to the value of the driver resistance [5].

Nevertheless, as crosstalk effects become very pronounced for high-speed VDSM interconnects, a single saturated ramp for the Thévenin model is not always sufficient to accurately model the waveform at the gate outputs. In such cases, a piecewise linear Thévenin voltage source model can be employed, as described in [2]. Other possibilities for modeling coupled interconnects include a modified C-effective calculation [8] and modeling of the victim driver gate with a so-called transient holding resistance [21].

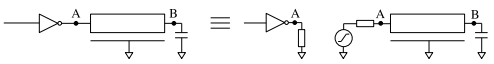

Figure 1: Total delay segregation: buffer and wire components.

In order to accurately model the delay in VDSM buses, a first key point is a careful and precise model of the physical wires used to transmit the signals. With increasing rise times and interconnect length, inductive effects must be taken into consideration as lumped models become inadequate. The effects of off-chip and on-chip inductance on signal integrity, timing, repeater insertion, power consumption and other IC related performance aspects in interconnects have been often investigated in the literature during the past years [1, 7].

From the theoretical point of view, as self and mutual inductances are loop-dependent quantities, they can be determined only if the whole current loop, i.e. the return path, is known [10]. However, the return path is especially difficult to be determined and for this purpose, Ruehli proposed in [18] an alternative inductance extraction approach based on the partial element equivalent circuit (PEEC) method, which is well-suited for circuit simulation as it depends only on circuit geometry. Consequently, the return path is determined by simulation and no *a priori* knowledge is required.

The impact of inductance effects can be determined by analyzing the signal spectrum, i.e. the spectral magnitude at different frequencies. In this context, the concept of significant frequency, f_s, can be used to reduce the complexity of the required information. For a trapezoidal pulse with ramp time t_r, we have $f_s = 0.34/t_r$. Although more than 10% of the spectral components are at higher frequencies than t_s, they can be neglected as their overall magnitude is very small and the introduced error negligible [4]. At extreme high frequencies, the interconnect reactance becomes frequency dependent because of the skin effect and the proximity effect [4]. When dealing with fast transitions, the high frequency values must be taken into account, while the low frequency ones have to be considered for the slow transient parts of simulation. However, in circuit simulators like SPICE, the resistance and inductance can be included at only one frequency. There are two possibilities to cope with this problem: on one hand, simulators can be extended to allow frequency dependent data simulation in the time-domain [3]; on the other, only one representative frequency that gives minimal errors can be employed and it has been empirically shown in [4] that the significant frequency is a very good choice. The last mentioned method is the one we have employed.

In this work, we analyze typical scenarios consisting of signals with rise times, t_r, ranging from 25ps to 500ps transmitted through local, intermediate, and global lines in generic 130nm 1.8V, 90nm 1.5V, and 65nm 1.1V technology nodes. The significant frequency is thus in the range of 0.68GHz to 13.6GHz [6, 16]. The frequency spectrum of the transmitted signals and the dimensions of the bus set us in a range where lumped models are inaccurate and distributed models considering line inductance and inductive coupling have to be employed. Non-electrical full-wave models are not required. In order to develop the delay model and analyze its accuracy for inductive and non-inductive models, three distributed models are considered: *RC* with coupling capacitances, *RLC* considering self-inductance, and *RLMC*, which also takes into account mutual inductance. In order to determine the SPICE parameters from wire and bus geometries, state of the art extraction tools have been used. The total ground and coupling capacitances have been extracted with FastCap [15], while resistance and inductance values have been extracted using FastHenry [11]. The extracted interconnect parameters have been used to write a complete netlist for all three models, which is then simulated in SPICE. The signal delay was afterwards determined from these simulation results.

In this work, we model 5-bit and 8-bit wide bus structures. Width, w, thickness, t, and pitch, p, vary between 0.5 and 3µm, 0.5 and 1µm, 1.5 and 6µm, respectively. The distance to the lower and upper metal layer is considered to be 1µm. To obtain accurate interconnect parameter values, simulations with the abovementioned field solvers have been performed until insignificant errors were achieved. As a result, wires were split into segments of approximately 100µm length. The number of segments, s, has been increased along with the wire length to maintain sufficient accuracy.

As process technologies scale down, the variations in several process parameters are continuously increasing and affecting more and more the overall system performance. For simulations where the effect of process parameter variations is included via Gaussian distributions, these values have been used as the expected value with a standard deviation specified as 0.1µm.

3. EXTENDED LINEAR DELAY MODEL

With technology scaling, the coupling capacitance between neighboring wires steadily increases and mostly dominates the overall line capacitance. The capacitive coupling between two lines directly influences the dynamic power consumption and the time required for a transition to complete. The transition time is determined by the relative transitions of the lines. For example, a victim toggling from low to high has to charge twice the coupling capacitance to a neighbor switching from high to low because of the Miller effect. On the contrary, when the neighbor switches in the same direction, the coupling capacitance does not have to be charged. This way, the delay function of capacitively coupled lines can be easily constructed as shown in detail in [22].

In the sequel, we denote the transition in line i of an n-bit wide bus as $\Delta b_i = b_i^+ - b_i^-$, where b_i^- and b_i^+ represent the initial and final value on line i, respectively. We also define the transition vector, $\underline{\Delta b} = [b_1, b_2, \cdots, b_n]^t$. Basically, each line is characterized by the effect on the delay produced by four possible switching scenarios in each aggressor: $(b_i^-, b_i^+) = \{(0,0),(0,1),(1,0),(1,1)\}$. Generally speaking, the set of line delays is a function of the transition vector.

In capacitively-coupled interconnects, when inductances can be safely ignored, the delay in line k, δ_k, of a symmetric bus is given in [22] as:

$$\delta_k = \tau_0\big[(1+2\lambda)\Delta b_k^2 - \lambda\Delta b_k(\Delta b_{k-1} + \Delta b_{k+1})\big], \quad (1)$$

where τ_0 is the delay of the crosstalk-free line and λ is the ratio of the coupling capacitance to the ground capacitance. In a bus, due to the shielding effect of the first-order neighbors on the higher-order ones in terms of coupling capacitance, the effect of the aggressors of an order higher than two can be neglected without any loss of accuracy. Note that the term Δb_k^2 is used instead of $|\Delta b_k|$ for convenience. We can rewrite Eq. 1 in a more general fashion:

$$\delta_k = \alpha_k\Delta b_k^2 + (\alpha_{k-1}\Delta b_{k-1} + \alpha_{k+1}\Delta b_{k+1})\Delta b_k \quad (2)$$

$$= \sum_{i=k-1}^{k+1} \alpha_i \Delta b_i \cdot \Delta b_k, \quad (3)$$

where $\alpha_k = \tau_0(1+2\lambda)$, $\alpha_{k-1} = \alpha_{k+1} = -\tau_0\lambda$. It is worth mentioning here, that for non-inductive interconnects, α_k is positive, while α_{k-1} and α_{k+1} are negative, and that in the case of non-symmetric buses we generally have $\alpha_{k-1} \neq \alpha_{k+1}$.

Inductive coupling is a long-range effect in contrast to the short-range capacitive coupling. Therefore, the effect of the aggressors of order higher than two cannot be discarded for an accurate analysis. The delay model proposed in this work generalizes the abovementioned delay model developed for non-inductive capacitively-coupled buses. In this section, we show that this linear pattern-dependent delay model can be extended in order to include inductive coupling between neighbors of order higher than two. The signal delay can be approximated as a linear combination of the delay produced by the switching patterns on every line.

A simple and efficient approach to approximating the impact of capacitive coupling is to include its effect into the equivalent capacitance seen by the gate, either by adding a term to the ground capacitance or by multiplying it by a Miller factor [2, 19, 22]. The added term is pattern-dependent as the effectively seen capacitance depends on the relative togglings on the victim and the aggressors.

Conceptually, we can also formulate the problem by choosing a nominal pattern and constructing an equivalent pattern-dependent interconnect model instead of computing equivalent effective capacitances, inductances, or resistances. For this purpose, we first select a nominal pattern for each line, for instance the one when the victim line toggles from low to high and all aggressors are quiet. Afterwards, for a different switching pattern, a delay matching operation is performed, i.e. an equivalent interconnect model is constructed with different (pattern-dependent) PUL parameters for each line, such that the delay of the equivalent network under the nominal toggling pattern is equal to the delay of the real interconnect model with the actual switching pattern as input.

Consider the example with the influence of coupling capacitances on delay as a function of the switching pattern. As previously mentioned, the additional seen capacitance induced by the Miller effect can be added to the ground capacitance. This added extra capacitance is a linear function of the relative toggling patterns. There-

fore, we can write in the general case, that the equivalent PUL capacitance of line k, C_k, is a linear function of the pattern:

$$C_k(\underline{\Delta b}) = C_{k0} + \sum_{i=1}^{n} \Delta C_i \Delta b_i = C_{k0} + \Delta C_k(\underline{\Delta b}), \quad (4)$$

where C_{k0}, the PUL capacitance for the nominal pattern, is in general a non-linear function of many parameters like rise time, load impedance, and technological parameters, while ΔC_i is the extra capacitance due to the coupling to line i.

For simplicity, we can assume that for finding every equivalent PUL parameter, we have to add a linear term in Δb_i. However, in the case of inductances (or other parameters), this assumption is not exact and hence, we can theoretically expect higher errors with increasing inductive effects.

Let ψ_j be a PUL parameter or any other parameter one has to compute for delay matching and let m be the number of those parameters. We define the set of all parameters as $\underline{\Psi} = \{\psi_1, \dots \psi_m\}$. As a result of the abovementioned assumption, we can write for the delay matching:

$$\psi_j(\underline{\Delta b}) = \psi_{j0} + \sum_{i=1}^{n} \Delta\psi_{ji} \Delta b_i = \psi_{j0} + \Delta\psi_j(\underline{\Delta b}), \quad (5)$$

where ψ_{j0} represents a (non-linear) function of many factors, and $\Delta\psi_{ji}$ denotes the difference in ψ_j in line i. In particular, we have $\underline{\Psi} = \{R, L, M, C\}$.

The delay of a line can be expressed for a given bus and driver configuration as a continuous function of the PUL parameters. By neglecting the non-linear terms in Δb_i of the Taylor expansion around $\underline{\Psi_0} = \{\psi_{j0}\}_{j=\overline{1,m}}$, we obtain for a low-to-high transition, that is $\Delta b_k = 1$, the following:

$$\begin{aligned} \delta_k(\underline{\Delta b}) &= \delta_k\big(\underline{\Psi_0} + \underline{\Delta\Psi}(\underline{\Delta b})\big) \\ &\approx \delta_k(\underline{\Psi_0}) + \sum_{j=1}^{m} \frac{\partial \delta_k}{\partial \psi_j} \Delta\psi_j(\underline{\Delta b}) \\ &\approx \delta_k(\underline{\Psi_0}) + \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \frac{\partial \delta_k}{\partial \psi_j} \Delta\psi_{ji} \right) \Delta b_i. \end{aligned} \quad (6)$$

Hence, when the non-linear terms of the Taylor expansion are negligible, the delay in line k can be expressed as a linear function of the transition patterns in neighboring lines. This observation allows us to extend to inductively coupled lines the delay model described in [22] and [23].

When an aggressor line does not toggle, the effect on delay in the victim line is practically independent of the state. Thus, the contribution of all the patterns $(0,0)$ and $(1,1)$ can be modeled by a constant term and the only contributors which must be precisely modeled are the patterns $(0,1)$ and $(1,0)$. The currents generated by switchings with opposite transitions have opposite directions with opposite transitions. Therefore, the contributions of the patterns $(0,1)$ and $(1,0)$ are of opposite sign though equal as absolute values. The delay predicted in line k is thus:

$$\delta_k = \alpha_k \Delta b_k^2 + \sum_{i=1, i\neq k}^{n} \alpha_{ik} \Delta b_i \cdot \Delta b_k = \sum_{i=1}^{n} \alpha_{ik} \Delta b_i \cdot \Delta b_k, \quad (7)$$

where $\alpha_{kk} \stackrel{\text{def}}{=} \alpha_k$ represents the delay in line k with quiet aggressors, and α_{ik} for $i \neq k$ denotes the contribution to the delay of the aggressor line i on line k. We call this model the *Extended Linear Delay (ELD) Model* and the corresponding α_{ij}-s *model coefficients* or simply *coefficients*.

The ELD model can be written in a compact way also for an n-bit wide bus. For this purpose, we consider the following notations: $\underline{\delta} = [\delta_1, \delta_2, \dots \delta_n]^t$, $\underline{\alpha} = [\alpha_1, \alpha_2, \dots \alpha_n]^t$, $\underline{e} = [1, 1, \dots 1]^t$, $\mathbf{A} = [\alpha_{ji}]_{n\times n}$, $\mathbf{A_i} = diag(\alpha_i)$, $\mathbf{B} = diag(\Delta b_i)$. It can be easily shown that $\underline{\alpha} = \mathbf{A_i}|\underline{\Delta b}|$. Thus, the two following forms can be used for the matrix formulation of the ELD model:

$$\underline{\delta} = \mathbf{B} \cdot \mathbf{A} \cdot \mathbf{B} \cdot \underline{e} = \underline{\alpha} + \mathbf{B} \cdot (\mathbf{A} - \mathbf{A_i}) \mathbf{B} \cdot \underline{e}. \quad (8)$$

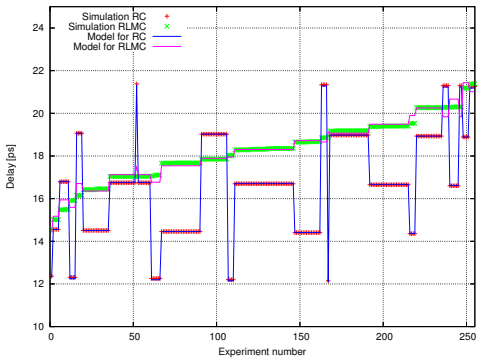

Figure 2: RC & RLMC buses: simulated and estimated delays.

For an n-bit wide bus, the ELD is able to characterize 4^n possible delays by storing only n^2 coefficients. These coefficients can be stored in the form of *a priori* computed and/or look-up tables and are functions of driver characteristics, wire geometry, effective load and input slew. For input values falling inside the range of the table index, they can be extended to be derived via interpolation.

As indicated in Sec. 2, extensive simulations have been carried out in order to assess the accuracy of the extended linear delay model. Both inductively and capacitively dominated crosstalk scenarios have been covered. The signal delay for a low-to-high transition has been measured as the delay from the time the near-end first reaches 50% of the final value to the time the far-end is stable above 50% of the final value.

As an initial approach, we have calculated the model coefficients on a given line, by performing for a switching on this line SPICE simulations for all possible patterns on the neighbors. The obtained delays are then used to compute the coefficients by the minimum square error approach. It is important to notice, that this coefficients can be pre-calculated and than used for fast delay estimation.

In the case of capacitively dominated coupling, a transition in an aggressor in the opposite direction increases the total capacitance that the victim has to charge and the transition is thus slowed down, i.e. $\alpha_i < 0$. On the contrary, due to Faraday's law of induction, in purely inductively coupling dominated lines, a transition of an aggressor in the same direction induces a current flowing in the opposite direction to the one in the victim line. Consequently, the effective current decreases and the delay increases, i.e. $\alpha_i > 0$. In buses exhibiting both capacitive and inductive coupling the coefficients for the first-order neighbor can be either negative or positive while the second-order coefficients are non-positive.

Tab. 1 shows the maximum absolute error, ε_{max}, and the root mean square error, ε_{rms}, of the proposed model for four specific cases of inductively-coupled lines: medium-high (two cases), low-medium, and medium inductive coupling. Moreover, for the purpose of comparing the model errors, we have also represented the errors when modeling the same interconnects as purely RC-coupled ones. In general, with regard to capacitive coupling, neighbors of at least second-order are almost completely shielded by the intermediate wires. The short-range nature of capacitive coupling explains the better approximation through the linear assumption. The very small errors appear f.i., because a second-order aggressor may slightly influence the victim through the two coupling capacitances which separates it from the victim. Nevertheless, this influence is extremely small. In the case of inductively-coupled interconnects, the assumption of linearity introduces slightly higher errors.

It is however important to notice, that although two of the presented cases correspond to highly inductively coupled interconnects, the error of the model is still small. Moreover, the highest maximum error appears when estimating the small delays. The maximum error when approximating higher delays has been in all simulated cases lower than 3.5%. This is especially interesting for delay reducing encoding schemes as shown in Sec. 5.

As a typical example, Fig. 2 shows the simulated and estimated delays for a 1000µm long line with $t_r = 100ps$. Both RC and

Table 1: Model Errors for Different Cases of Coupling.

	ε_{max}	ε_{rms}	ε_{max}	ε_{rms}	ε_{max}	ε_{rms}	ε_{max}	ε_{rms}
RC	2.06	0.78	1.68	0.65	2.27	0.94	1.92	0.76
RLMC	5.34	1.45	4.23	1.29	5.83	1.71	4.97	1.37

RLMC models are depicted for the 256 different switching patterns of the neighbors (the experiments are ordered after increasing values of the RLMC delay). The proposed model fits very well the experimental results. The error for the RC case is almost negligible. For the RLMC network, the error is slightly higher but less than 2%. Moreover, we can observe that the worst delay patterns for the RC and RLMC models are completely different and that the simplified model for *RC* buses is unable to predict the delays induced by the switching patterns in *RLMC* buses. The proposed ELD model is able to predict with high accuracy that behavior.

Process variations are fluctuations in the value of process parameters. The impact of process and environmental variations on performance and power consumption has been increasing with each semiconductor technology generation [24]. Variations have been classically divided into two categories: inter-die (die-to-die) and intra-die (within-die) variations. The inter-die variations are usually assumed to have a Gaussian distribution and when a number of process parameters are considered simultaneously it is important to take into account the correlation between these parameters. When device parameters vary within a single die as a function of their location, we talk about intra-die variations. Depending on the source of variations, within-die variations may be spatially correlated or uncorrelated and generally, modeling intra-die variations results in a huge complexity. Briefly, one can also say that variations are either spatially uncorrelated or correlated. Depending on their correlation distance, correlated variations are of inter-die or intra-die nature [24]. In order to model intra-die variations, a huge number of random variables is required. Some techniques have been developed in order to simplify analysis techniques when dealing concomitantly with correlated and independent sources of variations. For example, the Principal Component Analysis (PCA) (Karhunen-Loeve Transform or Hotelling Transform) is a statistical technique that maps a given set of correlated random variables to another set of uncorrelated random variables. The latter are called principal components, they are independent random variables, and the first few capture the most of the variability [24]. Actually, PCA represents the optimal linear transformation for choosing the subspace with the largest variance. PCA-based techniques are used to simplify the correlation structure of variations in process parameters across a chip.

Therefore, in order to show how the ELD model can be extended to take into consideration effects of process variations, we choose for simplicity to model variations as normal distributed random variables like proposed in [24]: $\phi = \phi_{nom} + \Delta\phi$, where ϕ_{nom} is the nominal value of the process parameter and $\Delta\phi$ is a zero-mean random variable that captures variations. We model variations in the width and thickness of the interconnect via Gaussian distributions with a standard deviation of 0.1 μm. This variations have an important impact on all PUL parameters. To be noticed that the PUL parameters become thus random variables that do not follow a Gaussian distribution.

In order to prove the suitability of the aforementioned technique, we have constructed a model for the simplified scheme for process variations presented in Sec. 2. Nonetheless, as previously explained, our approach is not limited just to that scheme. We have generated 1000 sets of Gaussian distributed values for pitch, width and thickness. For each of this 1000 sets, we have performed the extraction of the RLMC parameters. Afterwards, we have completed for each set SPICE simulations for calculating the delay tables and the corresponding coefficients, i.e. the $\mathbf{A}$ matrixes.

The goal of this work is to provide the designer with a high-level model which can be used to abstract completely the physical world. By letting the model coefficients to be random variables, it is possible to model in a compact way the effects of process variations on delay. Eq. 8 can be thus rewritten in order to include process variations:

$$\underline{\delta} + \underline{\Delta\delta} = \mathbf{B} \cdot (\mathbf{A} + \mathbf{\Delta A}) \cdot \underline{\Delta b}, \tag{9}$$

where $\underline{\Delta\delta}$ and $\mathbf{\Delta A}$ represent the variation in delay and model coefficients respectively. In our scenario, the random variables are modeled as independent processes. Thus, only the one-dimensional probability function of each coefficient must be determined.

In Fig. 3, we notice that the histogram of the coefficients is very close to a Gaussian. However, since all the PUL parameters but the resistance are skewed and non-linear function of the interconnect dimensions, the delay does not follow a Gaussian distribution. The probability density function (PDF) of a random variable can be accurately estimated by computing the first few moments [17]. Hence, we can employ this approach for an efficient yet accurate PDF estimation method. Thus, considering that we employ the first

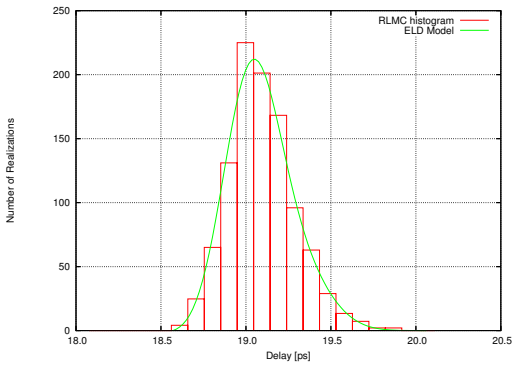

Figure 3: Process variations: simulated and estimated delay.

s moments, each α_{ij} of the ELD model is replaced by a set of s moments $\{\mu_k^{(\alpha_{ij})}\}_{k=\overline{1,s}}$, where $\mu_k^{(\alpha_{ij})}$ is the k-th moment of the random variable α_{ij}. The moments of the δ_k-s and their PDF can be easily calculated from the ELD (see Eq. 9).

An expansion of the probability function in terms of Hermite polynomials [20] is very suited for close-to-Gaussian distributions as the current scenario. By fitting just the first three moments of the random variable, we get the following approximation for the distribution $f_\alpha(x)$ of an α-coefficient:

$$f_\alpha(x) = \frac{(x^3 - 3x)\gamma_3 + 1}{\sqrt{2\pi}} \cdot \exp\left(-x^2/2\right), \tag{10}$$

where γ_3 is the skewness of the original α.

Consequently, in order to include the effects of process variations in our scenario, we have to characterize the model coefficients not only by their mean value, i.e. the first moment, but also by the second and third moments (standard deviation and skewness, respectively). In the case when the PDF differs significantly with respect to a Gaussian distribution, the approximation with Hermite polynomials becomes poor. In this case, high accuracy can be obtained by using more moments with expansions of the PDF in terms of other polynomials like Legendre or Laguerre.

4. INTERPRETATION OF RESULTS

As mentioned in Sec. 2, figures of merit have been proposed to characterize the importance of on-chip inductance [2, 4]. When the interconnect is treated as an uniform RLC transmission line, inductance needs to be considered if $t_r/2\sqrt{LC} < l < 2R\sqrt{L/C}$ where R, L, C are PUL resistance, self-inductance, and coupling capacitance, respectively; t_r is the rise time of the input signal. The first inequality ensures that the time-of-flight is at least two times larger than the rise-time, i.e. $t_f > 2t_r$. The second rule is equivalent to the condition that the RLC circuit is underdamped. Even though these inequalities are for coupled models rather loose, they serve in the sequel for qualitatively explaining the dependency of the model coefficients on t_r. We can mention here that the second condition is in our scenario always fulfilled (the lines are underdamped), while the second one becomes true only for small rise times.

In Fig. 4, we notice that the worst and best case delays and switching patterns are almost not changing in RC and RLC interconnects. However, this is not the case for RLMC networks as not only worst and best case delays rapidly change, but also the patterns which induce those delays vary. Moreover, as the rise time decreases, the coupling coefficients change from negative values to positive ones. Thus, the model accurately captures the tendency of the inductive coupling to dominate over the capacitive one with decreasing t_r.

In Fig. 5, we observe in the case of $RLMC$-interconnects an oscillation of the coefficients with varying rise time. For high rise times, the interconnect behaves capacitively ($\alpha < 0$ for the aggressors) and as the rise time decreases, the interconnect behavior becomes dominated by the inductive coupling components, the coefficients of the neighboring lines becoming thus positive.

Moreover, the delay for all patterns do not increase monotonically with increasing rise times of the input signals as it was the case in simple RC networks. This effect is related with the multiple zeros and poles which appear in complex RLMC systems and it is

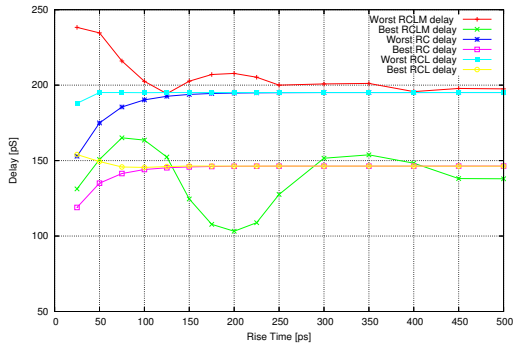

Figure 4: Variation of worst and best case with rise time for the third line of a 1000µm bus.

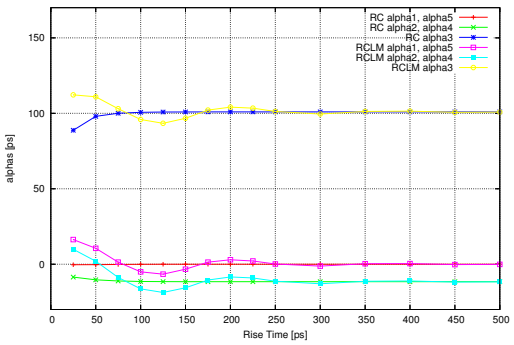

Figure 5: Model coefficients of line 3 for varying rise time for a 500µm bus.

modeled with an oscillating variation of the coefficients with t_r as we can see in both Fig. 4 and Fig. 5. Furthermore, in the case of the lines with the same number of neighbors, the coefficients are almost the same. Such properties can be exploited for improving even more the compactness of the model.

5. HIGH-LEVEL CODING-BASED THROUGHPUT IMPROVEMENT

In order to address timing issues at higher levels of abstraction, accurate models capable to predict pattern-dependent signal delay are required. This is mandatory if delay-aware coding is to be employed. Previous techniques in that direction like [9, 12, 22, 23, 26] are restricted only to non-inductively coupled interconnects because of a lack of proper models. Some efforts have been done to identify worst-case switching patterns in inductively coupled lines [21, 25]. However, those methods cannot predict the delay for a given input switching pattern. Further, the delay coding limits and delay elimination methods developed in [22] for capacitive coupling do not hold in the more general case of inductively-coupled lines and have to be revised. For this purpose, our model can be employed as shown in the following.

Let us consider first the case of capacitively coupled interconnects as done in [22, 23]. In the normal operation of data buses, the clock period T_{ck} is set so that all transitions can be completed, i.e:

$$T_{ck} \geq \tau_0(1 + 4\lambda). \tag{11}$$

As pointed out in [22], the abovementioned inequality suggests that we can speed up the bus by avoiding time-expensive transitions. For instance, by eliminating all transitions with delays equal to $\tau_0(1 + 4\lambda)$ and $\tau_0(1 + 3\lambda)$, the inequality becomes:

$$T_{ck} \geq \tau_0(1 + 2\lambda). \tag{12}$$

Thus, by prohibiting transitions that induce a large delay, the clock period can be significantly reduced. However, the number of bits that can be transmitted per transition is decreased. For an n-bit wide bus, we define the bit reduction factor, $\zeta_b(n, k)$, as the ratio between the maximum achievable information rate on the coded bus and the actual bus width. Further, the speed increasing factor, $\zeta_s(n, k)$, stands for the interconnect delay decreasing rate and is defined as:

$$\zeta_s(n, k) = \frac{1 + 4\lambda}{1 + k\lambda}, \tag{13}$$

where $k = \overline{0, 4}$ is a coefficient that indicates the highest allowed delay. Generally, for an efficient encoding we have k={2,3} [22]. Thus, we can define the total actual throughput increase rate as:

$$\zeta_t(n, k) = \frac{\zeta_s(n, k)}{\zeta_b(n, k)}. \tag{14}$$

In order for a code to be efficient, the achieved throughput increase rate must be higher than one, i.e. $\zeta_t(n, k) > 1$. It is to be mentioned that λ can also be regarded as a function of k and n.

In the following, we analyze the effect of increasing inductive coupling on maximum achievable throughput increase rates. Conceptually, the coefficients of the ELD model can be split in two parts: on the one hand, the coefficients standing for capacitive coupling ($\alpha_{ij,C}$) and on the other hand, the coefficients for the inductive coupling ($\alpha_{ij,L}$). Thus,

$$\alpha_{ij} = \alpha_{ij,C} + \alpha_{ij,L}, \ i \neq j, \tag{15}$$

where $\alpha_{ij,C} \leq 0$ and $\alpha_{ij,L} \geq 0$. Capacitive coupling is a short-range effect and thus, only first-order neighbors can be considered as in the models proposed in [22, 23]. Consequently, we can write the delay in line k as:

$$\begin{aligned}\delta_k = \alpha_k \Delta b_k^2 &+ (\alpha_{kk-1,C}\Delta b_{k-1} + \alpha_{kk+1,C}\Delta b_{k+1})\Delta b_k \\ &+ (\alpha_{kk-1,L}\Delta b_{k-1} + \alpha_{kk+1,L}\Delta b_{k+1})\Delta b_k \\ &+ \sum_{i\neq 0,1} (\alpha_{kk-i,L}\Delta b_{k-i} + \alpha_{kk+i,L}\Delta b_{k+i})\Delta b_k\end{aligned} \tag{16}$$

In a symmetric bus we have α_{kk-i}=$\alpha_{kk+i} \stackrel{\text{def}}{=} \alpha_k^{(i)}$, if the corresponding neighbor exists. We can define in a similar way $\alpha_{k,C}^{(i)}$ and $\alpha_{k,L}^{(i)}$. When the corresponding neighbors do not exist, we can define either the coefficients or the associated transitions as zero. For a symmetric bus we have then:

$$\begin{aligned}\delta_k \ = \ & \alpha_k \Delta b_k^2 + (\alpha_{k,C}^{(i)} + \alpha_{k,L}^{(i)})(\Delta b_{k-1} + \Delta b_{k+1})\Delta b_k \\ & + S_{ind}(k)\Delta b_k,\end{aligned} \tag{17}$$

where $S_{ind}(k)$ stands for the cumulative influence of the inductive aggressors of an order higher than two. It can be easily shown that $\tau_0 = \alpha_{kk} - 2|\alpha_{k,C}^{(1)}|$ and $|\alpha_{k,C}^{(1)}| = \lambda$.

In the capacitive case we have only five possible delay classes for a switching line: $1 + k\lambda$, with $k = \overline{0, 4}$. Moreover, the delay depends only on the first order neighbors. However, with increasing inductive coupling effects, the effects of the second-order neighbors cannot be neglected anymore. Additionally, in the case of inductively coupled lines, the worst- and best-case delays have been reported to vary as a function of the relationship between the magnitude of capacitive and inductive coupling [1, 25]. Cao et al. [1] concluded that when taking into account inductive coupling, worst case delay and noise are dominated more by the switching pattern $\uparrow\downarrow\uparrow\downarrow\uparrow$ than by the $\downarrow\downarrow\uparrow\downarrow\downarrow$ one. Furthermore, Tu et al. [25] showed that the former switching pattern becomes the worst-case scenario with increasing wire capacitance. However, for smaller coupling capacitance the worst-case pattern was reported to change to $\uparrow\uparrow\uparrow\uparrow\uparrow$. We have denoted low-to-high and high-to-low transitions with $\uparrow$ and $\downarrow$ respectively. All these cases are covered by the ELD model and are easy to identify, as seen in the sequel.

Being a long-range effect, inductive coupling allows neighbors of order higher than two to become inductive aggressors. In the case of neighbors of order one, one cannot know *a priori* whether they are inductive or capacitive aggressors. This is decided by wires geometry, propagation time, and rise times.

Let us define $\eta \stackrel{\text{def}}{=} \max\{S_{ind}(k)\} = 2\max\{\sum_{i\geq 2} \alpha_{k,L}^{(i)}\} \geq 0$. Several major cases with regard to the relationship between induc-

Table 2: Comparison between delay classes in the case of capacitive and inductive coupling.

Capacitive	*Capacitive and Inductive*
$1+4\lambda$	$1+4\lambda-2\alpha_{k,L}^{(1)}\pm\eta$
$1+3\lambda$	$1+3\lambda-\alpha_{k,L}^{(1)}\pm\eta$
$1+2\lambda$	$1+2\lambda\pm\eta$
$1+\lambda$	$1+\lambda+\alpha_{k,L}^{(1)}\pm\eta$
1	$1+2\alpha_{k,L}^{(1)}\pm\eta$

tive and capacitive coupling can be identified: (a) $\lambda \gg \alpha_{k,L}^{(1)} + 2\eta$: in this case, the capacitive coupling completely dominates the inductive one which can be neglected without any loss of accuracy; (b) $\lambda \gtrsim \alpha_{k,L}^{(1)} + 2\eta$: the inductive coupling cannot be neglected and this case corresponds to a low-medium inductive coupling and the delay classes are disjoint (see Tab. 2); (c) $\lambda \lesssim \alpha_{k,L}^{(1)} + 2\eta$: both inductive and capacitive couplings cannot be neglected, the inductive coupling is getting more important, and the delay classes are not disjoint; (d) $\lambda \ll \alpha_{k,L}^{(1)} + 2\eta$: the inductive coupling outweighs the capacitive one and the delay classes are totally mixed; this case is highly unrealistic as the corresponding crosstalk noise is usually at unacceptable levels. The ELD model takes into account such effects in a very simple way and it allows to analyze and optimize early in the design flow crosstalk-induced delay in point-to-point interconnects. With increasing inductive effects, the five previously mentioned delay categories start to dissolute and cover a wider range of values as shown in Tab. 2. When prohibiting delays from the $1+4\lambda$ and $1+3\lambda$ classes, the delay classes are disjoint only if $\lambda > \alpha_{k,L}^{(1)} + 2\eta$.

In the sequel, we compare the influence of inductive coupling on the total throughput increase rate. In the case of inductive coupling, $\zeta_b(n,k)$ is equal to the capacitive case. However, we have:

$$\zeta_s(n,k) = \frac{1+4\lambda-2\alpha_{k,L}^{(1)}+\eta}{1+k\lambda-(k-2)\lambda+\eta}. \quad (18)$$

For k=2, the equality becomes:

$$\zeta_s(n,2) = \frac{1+4\lambda-2\alpha_{k,L}^{(1)}+\eta}{1+2\lambda+\eta}. \quad (19)$$

It can be easily shown that coding for performance would be more efficient with inductive coupling only if $\lambda < 0$, which is not possible. Thus, even in a case of a low or medium inductive coupling, the possibilities to increase throughput deteriorate in comparison with the non-inductive case. Further, if we consider simple encoding schemes, like the so-called Fibonacci code [13] (which is similar to the code indicated in [22]), $\zeta_b(8,k)$ degrades. This clearly indicates that in the case of the modeled global line even though for the capacitive coupling $\zeta_t(8,2) > 1$, the code may become inapplicable when inductive effects appear.

Consequently, the effectiveness of coding schemes for throughput improvement must be assessed at high levels of abstraction especially in the case of inductive coupling and the information required for this purpose is intrinsically comprised in the ELD model.

6. CONCLUDING REMARKS

This work introduced an extended linear model for high-level signal delay estimation in both inductively and capacitively coupled on-chip buses. The developed model approximates the signal delay as a linear combination of the contributions induced by each aggressor line for the complete set of switching patterns and not only for capacitively coupled point-to-point interconnects or the worst case patterns, as in previous works. Moreover, we have shown that the model can be extended to include the effects of process variations. For a simplified scheme, we proved that by considering the coefficients of the model as random variables and employing their first three moments, we can get an accurate description of the delay variation. The accuracy of the model has been assessed by means of extensive experiments employing state-of-the-art 3D capacitance and inductance extraction tools and SPICE simulations.

Root mean square errors less than 2% have been reported. Therefore, the ELD model is suitable for fast yet efficient high-level analysis of bus encoding schemes focused on delay minimization in inductively coupled lines. We have also shown how the model can be employed at high levels of abstraction in order to explore coding-based alternatives. In contrast with previous models, the developed model is able to predict that coding techniques for throughput improvement are less efficient when besides the capacitive coupling effects also inductive ones appear.

7. REFERENCES

[1] Y. Cao, X. Huang, N. H. Chang, S. Lin, O. S. Nakagawa, W. Xie, D. Sylvester, and C. Hu. Effective On-Chip Inductance Modeling for Multiple Signal Lines and Application to Repeater Insertion. *IEEE Trans. on VLSI Systems*, 10(6):799–805, Dec. 2002.

[2] M. Celik, L. Pileggi, and A. Odabasioglu. *IC Interconnect Analysis*. Kluwer, 1996.

[3] N. Chang, L. Barford, and B. Troyanovsky. Fast Time-Domain Simulation in SPICE with Frequency-Domain Data. In *47th Electronic Comp. & Tech. Conf.*, pages 689–695, San Jose, CA, May 1997.

[4] C.-K. Cheng, J. Lillis, S. Lin, and N. Chang. *Interconnect Analysis and Synthesis*. Wiley, 2000.

[5] F. Dartu, N. Menezes, and L. T. Pileggi. Performance Computation for Precharacterized CMOS Gates with RC Loads. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 15(5):544–553, May 1996.

[6] Device Research Group, UC Berkeley. BSIM 4.5.0 Release. http://www-device.eecs. berkeley.edu/bsim3/, Dec. 2005.

[7] K. Gala, D. Blaauw, J. Wang, V. Zolotov, and M. Zhao. Inductance 101: Analysis and Design Issues. In *DAC*, pages 329–334, Las Vegas, Nevada, June 2001.

[8] P. D. Gross, R. Arunachalam, K. Rajagopal, and L. T. Pileggi. Determination of Worst-Case Aggressor Alignment for Delay Calculation. In *ICCAD*, pages 212–219, San Jose, CA, Nov. 1998.

[9] K. Hirose and H. Yasuura. A Bus Delay Reduction Technique Considering Crosstalk. In *DATE*, pages 441–445, Paris, France, Mar. 2000.

[10] Y. I. Ismail and E. G. Friedman. *On-Chip Inductance in High Speed Integrated Circuits*. Kluwer, 2001.

[11] M. Kamon, M. Tsuk, and J. White. FastHenry: A Multipole-Accelerated 3D Inductance Extraction Program. *IEEE Trans. on Microwave Theory and Techniques*, 42(9):1750–1758, Sept. 1994.

[12] Z. Khan, A. T.Erdogan, and T. Arslan. Dual Low-Power and Crosstalk Immune Encoding Scheme for On-chip Data Buses. *Electronics Letters*, 39(20):1436–1437, Oct. 2003.

[13] T. Lindkvist, J. Löfvenberg, H. Ohlsson, K. Johansson, and L. Wanhammar. A Power-Efficient, Low-Complexity, Memoryless Coding Scheme for Buses With Dominating Inter-Wire Capacitances. In *IWSOC*, pages 257–262, Banff, Alberta, Canada, July 2004.

[14] B. Lu, D.-Z. Du, and S. S. Sapatnekar, editors. *Layout Optimization in VLSI Design*. Kluwer, 2001.

[15] K. Nabors and J. White. FastCap: A Multipole-Accelerated 3D Capacitance Extraction Program. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 10(11):1447–1459, Nov. 1991.

[16] NIMO Group, Arizona State Univ. Predictive Technology Model. http://www.eas.asu.edu/ptm/, Dec. 2005.

[17] L. T. Pillage and R. A. Rohrer. Asymptotic Waveform Evaluation for Timing Analysis. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 9(4):352–366, Apr. 1990.

[18] A. E. Ruehli. Inductance Calculation in a Complex Integrated Circuit Environment. *IBM J. on Res. and Dev.*, 16:470–481, Sept. 1972.

[19] S. Sapatnekar. *Timing*. Kluwer, 2004.

[20] S. Shanmugan and A. Breipohl. *Random Signals: Detection, Estimation and Data Analysis*. Wiley, 1988.

[21] S. Sirichotiyakul, D. Blauuw, C. Oh, R. Levy, and V. Zolotov. Driver Modeling and Alignment for Worst-Case Delay Noise. In *DAC*, pages 720–725, Las Vegas, Nevada, June 2001.

[22] P. P. Sotiriadis. *Interconnect Modeling and Optimization in Deep Sub-Micron Technologies*. PhD thesis, MIT, May 2002.

[23] S. R. Sridhara, A. Ahmed, and N. R. Shanbhag. Area and Energy-Efficient Crosstalk Avoidance Codes for On-Chip Buses. In *ICCD*, pages 12–17, San Jose, CA, Oct. 2004.

[24] A. Srivastava, D. Sylvester, and D. Blaauw. *Statistical Analysis and Optimization for VLSI: Timing and Power*. Springer, 2005.

[25] S.-W. Tu, J.-Y. Jou, and Y.-W. Chang. RLC Effects on Worst-Case Switching Pattern for On-Chip Buses. In *ISCAS*, vol. 2, pages 945–948, Vancouver, Canada, May 2004.

[26] B. Victor and K. Keutzer. Bus Encoding to Prevent Crosstalk Delay. In *ICCAD*, pages 57–69, San Jose, CA, Nov. 2001.

[27] S.-C. Wong, G.-Y. Lee, and D.-J. Ma. Modeling of Interconnect Capacitance, Delay, and Crosstalk in VLSI. *IEEE Trans. on Semiconductor Manufacturing*, 13:108–111, Feb. 2000.