

Advanced Waveform Models for the Nano-Meter Regime

Sani R. Nassif and Emrah Acar
IBM Research - Austin
11501 Burnet Rd
Austin, TX 78746
(nassif,emrah)@us.ibm.com

ABSTRACT

In the early years of digital system design, it was sufficient to model the delay of elements, so digital simulation tools used step functions to represent signals. As switching speeds increased, it became increasingly necessary to take into account the transition time of the signal. The simplest way of modeling the transition time is to upgrade the signal model from a step to a ramp, and this has been the state of digital design since the early eighties [1, 2].

As speeds increase further, however, the limitation of this approximation have become increasingly apparent. One well known problem is that of threshold selection, which -when not done properly- can lead to to negative delays [3, 4]. Also, many of the deep-submicron phenomena (e.g. inductive interconnect, coupled noise, and power supply current) are difficult or impossible to model accurately with the ramp model.

In this paper, we develop the mathematical basis for a new approach to modeling the waveforms associated with digital circuits. The approach represents a logical extension to current waveform modeling methods and provides a straightforward method to quantify and increase the accuracy of waveform models.

1. PRIOR WORK

The simplicity of the ramp approximation of signal waveforms has several advantages: (1) it makes the task of building models for timing analysis easy, (2) it is conceptually easy to grasp and to translate from the model parameters to a pictorial representation or to a Spice input specification, (3) it is information dense in that two real numbers (delay, slope) and a boolean (rising/falling) completely encapsulate the behavior of the waveform in question.

Several researchers have recognized the need for more detailed waveform models, [3, 4, 5, 6]. Most, however, attempt to derive the models from a first-principles circuit behavior approach by deriving the waveform model from the analytical solution of the non-linear ordinary differential equation describing the behavior of a simple circuit, such as a CMOS inverter [7]. While such approaches are useful in order to improve our understanding of the behavior of such circuits, they have the following weaknesses:

- They are usually restricted to simple output stage topologies for which these analytical techniques can be applied.
- They are not related to or integrated into the current timing model characterization methodology.
- They do not allow for a tradeoff of complexity vs. accuracy, and do not gracefully degrade to the common ramp model.

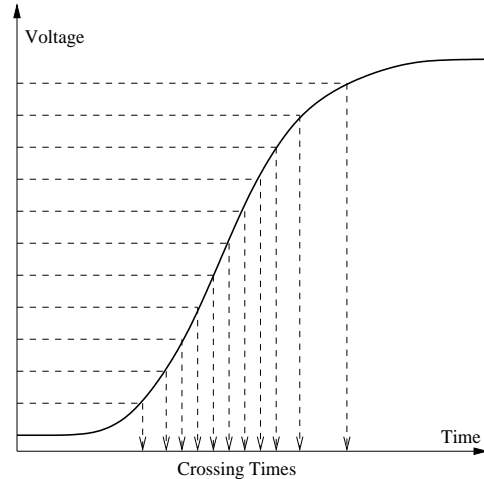


Figure 1: A typical digital waveform.

2. OUR APPROACH

Consider the waveform in Figure 1. Assuming that we are modeling legal waveforms that transition fully between the two voltage supply rails, we can choose to model the waveform in a piece-wise linear fashion by the vector of times at which the waveform reaches specified intermediate voltage levels between the low and high levels. Without loss of generality, we assume that the low and high levels are zero and V_{DD} , normalize these intermediate voltage levels by the V_{DD} , and denote the vector of such normalized voltages by V . We measure the crossing time for each of these voltages and denote the vector of resulting crossing times by T , and its length by N_T .

During the process of characterizing a cell in order to generate a timing model, the cell will be simulated under a variety of input, loading, temperature and power supply voltage conditions. Suppose that we perform M such simulations, and measure the resulting crossing times $T_i, i = 1..M$. If we view these M simulations collectively, we would assert that the resulting components of T are *correlated*. That is, we would expect that -for example- the 20% crossing point and the 30% crossing point, viewed as statistical variables, are *not* independent of each other, but have strong shared behavior. Figure 2 shows a plot of the 50% and 75% crossing times vs. the 5% crossing time from an experiment comprising 625 simulation of a simple CMOS inverter under various input, loading and power supply conditions. It is clear from the plot that the various crossing times are statistically correlated.

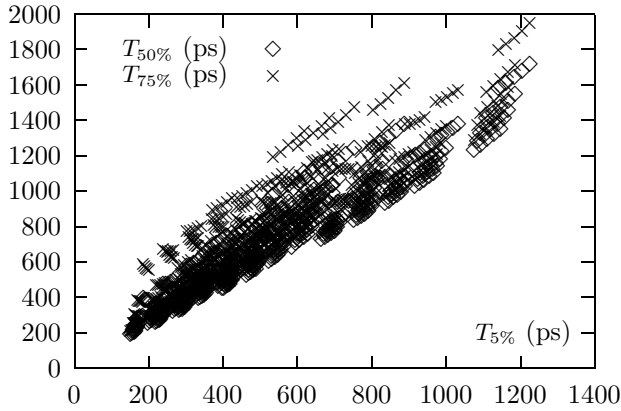


Figure 2: Plot of $T_{50\%}$ and $T_{75\%}$ vs. $T_{5\%}$. $T_{a\%}$ denotes the $a\%$ crossing point.

Principal component analysis (PCA) [8] is a statistical analysis technique that allows us to understand the joint behavior of a number of statistically correlated variables, and to map them into a reduced number of independent so-called *factors*. The PCA analysis works by performing a singular value decomposition (SVD) on the correlation matrix of the input variables, in our case, the N_T crossing times, and results in (a) a vector of N_T singular values, and (b) an N_T by N_T rotation matrix which maps the correlated crossing times to the uncorrelated factors.

We denote s_i as the i^{th} largest singular value resulting from the PCA analysis, and note that it can be directly interpreted as the amount of variability explained by the i^{th} factor. Furthermore, the quantity $d_K = \sum_{i=0}^K s_i / \sum_{i=0}^{N_T} s_i$ can be interpreted as the total variability explained by the first K factors altogether. Accordingly, the quantity of $1 - d_K$ would be the amount of variability left unexplained by the first K factors.

Figure 3 shows a plot of the singular values s_i , and the remaining variability (i.e. $1 - d_i$) resulting from the PCA analysis on the crossing times measured from the same 625 CMOS inverter simulations above. The plot clearly shows that the first 2 factors account for $\approx 99.9\%$ of the total variability in the resulting waveforms. This explains, to a certain degree, why the 2 parameter ramp model has been so successful at modeling the behavior of digital circuits. We will examine this in more detail in the next sections.

Thus far, we have shown: (a) that digital signal waveforms can be represented by the crossing times at N_T predefined voltage levels, (b) that these crossing times can be viewed as a set of correlated statistical variables, and (c) that the N_T crossing times can each be modelled as linear functions of a much smaller set of N_F independent variables. This means that we have defined a way in which the complete waveform can be expressed as a function of a small number of inputs: $T = \mathcal{F}^T V$, where \mathcal{F} is an $N_T \times N_F$ matrix, and the N_T vector V denotes the independent variables. The transform matrix \mathcal{F} can be derived from performing a PCA analysis of the waveforms that result during the normal process of timing characterization of a cell.

3. EXPLORING FACTORS FOR WAVEFORM MODELING

We now present the results of performing PCA analysis on the measured crossing times of several circuits under a variety of conditions. The purpose of these analyses is to show that the statis-

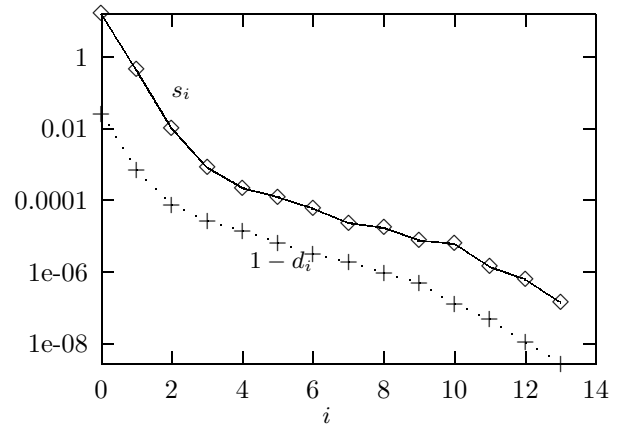


Figure 3: Plot of s_i and $1 - d_i$ vs. crossing time index i .

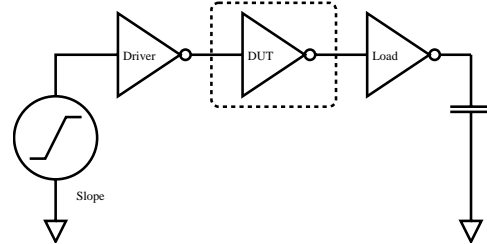


Figure 4: Circuit used to simulate a CMOS inverter.

tics governing the crossing times hold over a wide range of operating conditions, device sizes, model parameters and circuit topologies. This strengthens the case for using PCA as a general purpose method to model *all* the waveforms that one might encounter during the process of digital system analysis.

The next sections explore the validity and generality of the PCA analysis as it is applied to circuits under various conditions.

3.1 Operating Conditions and Device Sizes

For this first example, we simulated a 1.8V, 0.18 μ CMOS inverter using the circuit in figure 4. The variables in the simulation were:

1. The input slope to the driver was varied from 0.1ns to 1ns.
2. The size of the input driver was varied from 0.1X to 1X a standard size inverter.
3. The size of the output load was varied from 1X to 10X a standard size inverter.
4. The P/N ratio of the device widths was varied from 1.2 to 2.

A full-factorial experiment design with five levels (values) for each variable was performed, resulting in a total of 625 circuit simulations using 0.18um bulk technology. We used a voltage step size of 0.1V, and measured all 17 crossing times *between* 0.0 and VDD (1.8V). Figure 5 shows a plot of the correlations amongst the observed crossing times, from which we observe (a) that the minimum correlation between the observed variables is ≈ 0.86 , and (b) that the correlations is highest for neighbouring crossing times (along the diagonal) and decreases monotonically with separation between voltage levels.

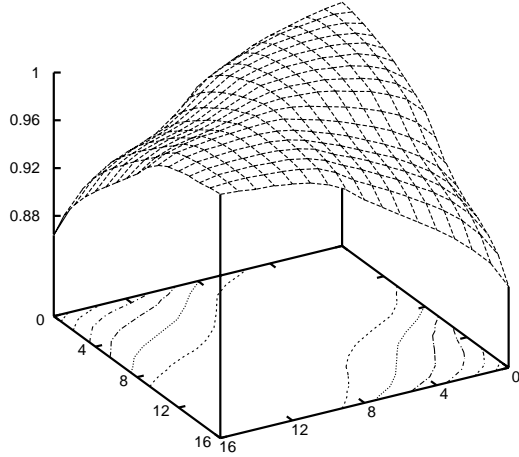


Figure 5: Correlation matrix amongst the crossing times.

Figure 3 presented the singular values obtained from the PCA analysis on this data. It is also interesting to consider the factors obtained from this analysis. In PCA analysis, each of the uncorrelated factors is expressed as a weighted sum of the original variables in question. Hence, we denote the factor F_i by an N_T long vector of the weights on the individual crossing times (t_i 's). Figure 6 shows a plot of the first three factors F_0 , F_1 , and F_2 in terms of the 17 t_i variables. From the figure we make the following important observations:

1. The first factor, F_0 , represents an approximately equal weighting of all the crossing times, t_0 through t_{16} , i.e. $F_0 = k_0 \sum_{i=0}^{17} t_i$, with k_0 being a circuit-specific constant. This is straightforwardly interpreted as the *average* of all the crossing points and in that sense represents the *delay* of the waveform.
2. The second factor, F_1 , represents an approximately linear weighting of the crossing times, i.e. $F_1 = \alpha \sum_{i=0}^N (\beta - i)t_i$ where α and β are circuit-specific constants. The expression for F_1 can be simplified to: $F_1 = k_{10}F_0 - k_{11} \sum_{i=0}^N it_i$ where k_{10} and k_{11} are algebraic functions of α and β . In this form, F_1 can be interpreted as the average of the quantity it_i which is proportional to the average slope of a linear ramp approximation to the original waveform.
3. The third factor, F_2 , represents an approximately second order (parabolic) weighting of the crossing times, i.e. $F_2 = \alpha \sum_{i=0}^N (\beta - i)^2 t_i$. Like F_1 , this can be re-written in terms of the average of the quantity $i^2 t_i$ that appears to be a dispersion metric.

It is important to note how the various factors appear to be closely related to the standard statistical moment formulae (e.g. mean, variance, kurtosis etc...). One might consider the direct application of such formulae to the crossing point statistics, but the weightings generated by the PCA analysis are not -in fact- constant (as we shall see in subsequent examples).

3.2 Model Parameters

In this second example, we simulated the same structure shown in figure 4 with the same simulation variables, but we took only 3 levels for each variable creating 81 different simulation conditions.

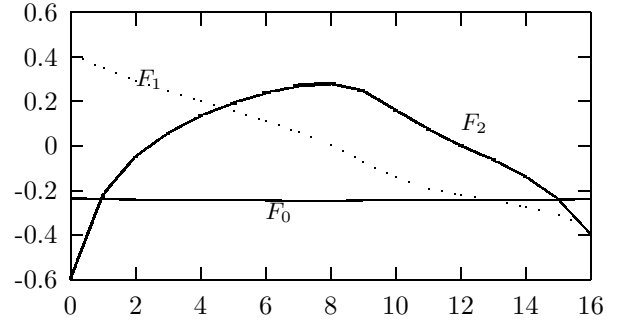


Figure 6: First three PCA factors vs. crossing time index.

N_M	s_0	$1 - d_0$	s_1	$1 - d_1$	s_2	$1 - d_2$
1	16.55	2.65%	0.438	0.068%	0.0104	0.008%
28	16.44	3.27%	0.546	0.056%	0.0079	0.009%

Table 1: Singular values with a single and multiple models.

We performed the simulation for every one of the 81 conditions for 28 distinct device model parameters obtained directly from a commercial foundry. These different device models enable us to consider the impact of the process parameter variability on the PCA analysis for crossing points. Table 1 compares the resulting factor rankings obtained for a single device model ($N_M = 1$) and the case of device model parameter variations ($N_M = 28$).

Figure 7 shows the differences between the first three factors generated from these simulations (using 28 distinct device model parameters, totaling $81 \times 28 = 2268$ simulations) and the same three factors generated for a single model in the previous section, i.e. $\Delta F_i = F_{i,28models} - F_{i,singlemodel}$.

It is clear that the differences are small and in the order of 1%. This means that the mapping from correlated crossing times to uncorrelated factors is substantially *independent* of the details of the device model within a typical technology.

3.3 Loading

In this third example, we performed a similar experiment to that in the previous section, but this time we performed the simulations for each structure with realistic resistive interconnect models of lengths $10\mu m$, $100\mu m$ and $1000\mu m$. It is well known that resistive interconnect results in waveforms that are qualitatively different from those generated with pure capacitive loading [5, 4]. Figure 8 shows a plot of the first three factors resulting from this experiment. We note that the first two factors are essentially the same as for the first experiment, while the third factor, while still second order in nature, is inverted in sign. This is an important indication that the simple moment-like formulae presented in Section 3.1 may not necessarily be as general as one would like, and that weights of these moments can range widely depending on -in this case- loading conditions. A possible reason of this outcome is the formation of tail shapes in signal waveforms mainly due to the presence of significant interconnect resistance.

3.4 Multiple Input Switching

In this fourth example, we simulated a 2-input NOR gate under similar conditions to those in the previous experiments, but added the effect of multiple input switching. We applied input waveforms at both inputs of the NOR gate, but offset them apart in time by 6 values in the range from $0ns$ to $1ns$. The impact of multiple input

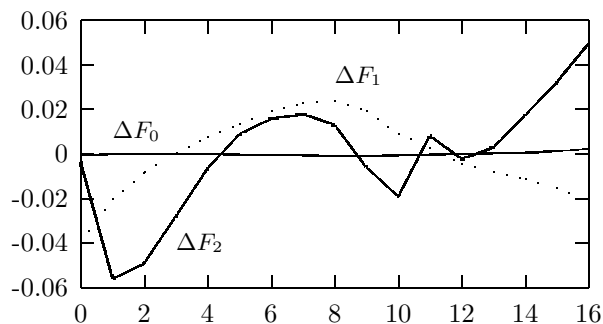


Figure 7: Difference between single and multiple model factors vs. crossing time index.

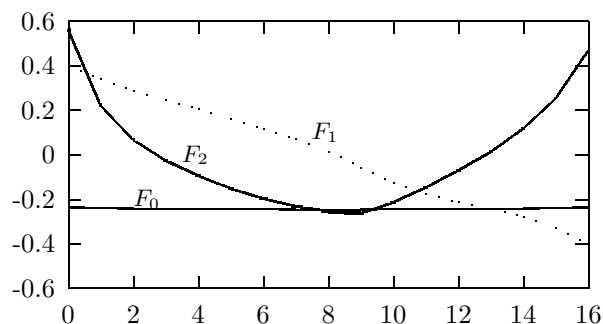


Figure 8: Factors with realistic interconnect loads vs. crossing time index.

switching is an important source of inaccuracy which has been recognized for some time by the CAD community, but relatively little has been done to address it, perhaps because it can lead to waveforms that are difficult to model with a single ramp. Figure 9 shows the resulting first three factors for this experiment.

We note that the three factors are essentially the same as for the first experiment with single input switching, which is an indication that the space of factors is large enough to represent waveforms even with the impact of multiple input switching.

3.5 Circuit Topology

In this final example, we performed the simulation experiments with a number of different circuit structures. These experiments included the same inverter circuit from section 3.1, along with 2 and 3-input NAND and NOR gates. We performed two PCA analyses: (a) for the data obtained from only the inverter, and (b) using the data from all five cells. We compare the singular values and sums for these two cases in table 2. It is apparent that when a smaller number of circuit topologies is included, more of the overall variability is explained with a fewer number of factors. Given the dominance of the first factor, these changes might appear small. This can be deceiving, however, since the previous sections have demonstrated that the subtleties of waveform shape appear to be contained in the higher (s_2 and above) factors.

4. ERROR ANALYSIS

Thus far, we have presented a new waveform modeling method and applied it to a variety of circuits to show that it is general enough to model waveforms of a broad class of circuits. We now

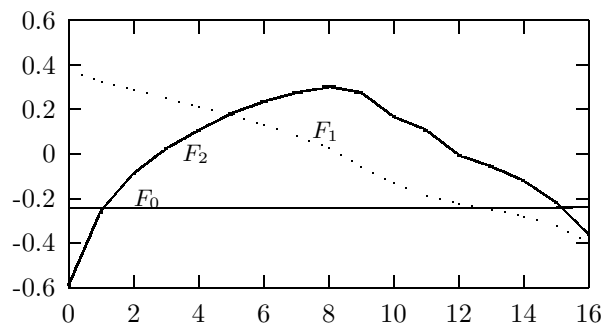


Figure 9: Factors for NOR2 gate with inputs offset in time vs. crossing time index.

Ckt	s_0	$1 - d_0$	s_1	$1 - d_1$	s_2	$1 - d_2$
INV	16.46	3.14%	0.533	0.011%	0.0128	0.0037%
ALL	16.40	3.51%	0.582	0.08%	0.0119	0.01%

Table 2: Singular values for 1 and 5 circuits.

turn our attention to a comparison of the accuracy of this proposed method as compared with current practice.

The factors predicted from a PCA analysis can be thought of as weighted linear sums of the various components of T . But we also have the freedom to choose any predefined linear transformation of T instead, and that predefined linear transformation need not use all N_T values of T , but may be as specific as a single component of T , or the sum/difference of any two components. The result is a less-than-optimal choice of factors (where optimality is defined as maximizing the explained variability). In fact, the ramp approximation can be thought to first order of as a selection of two factor, one (the delay) being the midpoint crossing time t_{50} , and the other (the slope or transition time) being the difference between -say- t_{70} and t_{30} .

To perform the comparison, we used the same 625 data samples generated in section 3.1. The analysis was also performed on all the other data sets with similar results. We performed four analyses:

1. We used linear least squares regression to build a first order (linear) polynomial model for each of the 17 switching times as a function of the first (most important) 3 factors generated from the PCA analysis. These results are represented by the curve labelled E_0 in figure 10. Note that there would normally not be a need to perform regression in this and the following case, but that we chose to do so in order to make the comparison between the four cases as consistent as possible.
2. We repeated the step above, but using only the first 2 factors generated from the PCA analysis. These results are labelled E_1 in figure 10.
3. We used linear least squares regression to build a first order (linear) polynomial model for each of the 17 switching times as a function of the midpoint crossing time t_{50} , and the transition time $t_{70} - t_{30}$. These results are labelled E_2 in figure 10.
4. We assumed a linear ramp defined from the midpoint crossing time t_{50} , and the transition time $t_{70} - t_{30}$. These results are labelled E_3 in figure 10.

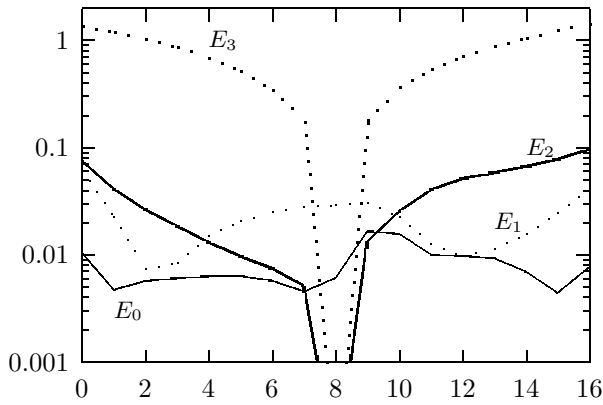


Figure 10: Error in estimating the 17 crossing times vs. crossing time index.

Figure 10 shows the standard deviation of the error (measured in ns) in estimating the 17 crossing times using the four methods above. It is clear from the figure that the error in estimating the crossing times for the PCA method is significantly better than the ramp method, widely used in practice (case E_3). We also note that using more factors (case E_0) improves the overall accuracy, which gives us confidence in extending the approach for future technologies.

Interestingly, we note that substituting the midpoint and transition time for the PCA factors is not significantly worse than using the first two factors of the PCA analysis. This gives current timing and modeling methodologies a possible easy method to improve overall accuracy while remaining compatible with legacy applications.

5. APPLICATIONS AND FUTURE WORK

As presented, the waveform models described in the paper would find application in any situation where having a more detailed description of digital switching waveforms is useful. Some of these applications are:

- Interfaces between analog and digital parts of a design. A problem that is growing in importance as mixed-signal SOC designs become more common.
- Interfaces between digital models and interconnect-dominated on-chip or off-chip buses. An example of the improvement in accuracy possible is shown in figure 11 which shows a comparison of the waveforms at the far end of a 2000μ line from Spice, the PCA approximation (dots) and the ramp approximation.
- Improved prediction of noise pulse generation and coupling.
- Modeling of power supply switching current, where the current ramp approximation predicts a simple pulse instead of the more typical triangle-like waveforms observed in practice. An example of results generated using this type of approximation is shown in figure 12.

The computational cost associated with using these waveform models includes two components. First is the cost of generation, which involves finding crossing times, performing the PCA analysis, and post-processing the results during the normal flow of cell

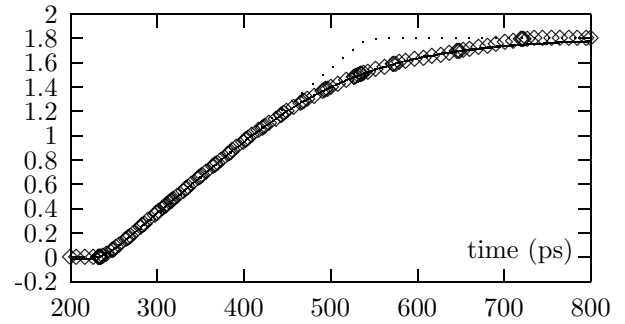


Figure 11: Waveforms at the far end of RC line with Spice, PCA and Ramp waveform approximations.

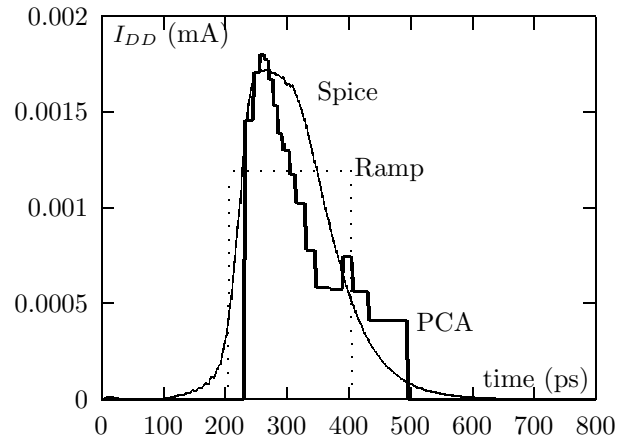


Figure 12: Power supply currents from Spice, PCA and Ramp waveform approximations.

characterization. In practice we find that all of these steps can be done in far less time than would be required to perform the circuit simulations required. Second is the storage cost associated with representing the waveforms. Assuming that the voltage axis is divided into N_V levels, and that a linear model is built for each of the crossing times as a function of -say- delay and transition time, then a mere $2N_V$ real numbers need to be stored with the cell model in order to allow it to generate full waveforms.

In the future, the PCA-based waveform modeling methodology proposed in this work will be applied to a larger and far more important and challenging task: enhancing the accuracy of current timing simulation tools. This accuracy is needed to: (a) capture difficult to model analog phenomena such as noise propagation, (b) model multiple input switching, (c) model complex loads difficult to represent by a single effective capacitance, (d) model dependence on environmental (e.g. power supply and temperature) and physical (e.g. ΔL and V_{th}) variations.

6. REFERENCES

- [1] N. Jouppi, "Timing analysis and performance improvement of mos vlsi designs," *IEEE Trans. CAD*, Jul 1987.
- [2] J. Ousterhout, "A switch-level timing verifier for digital mos vlsi," *IEEE Trans. CAD*, Jul 1985.
- [3] F. Dartu and L. Pileggi, "Modeling signal waveshapes for empirical cmos gate delay models," in *Proc. Intl. Workshop on*

Power and Timing Modeling, Optimization and Simulation,
1996.

- [4] F. Najm and J. Abraham, "Accounting for very deep sub-micron effects in silicon models," *EEdesign Magazine*, Jan 2001.
- [5] R. Arunachalam, F. Dartu, and L. Pileggi, "Cmos gate-delay models for general rlc loading," in *Proc. Intl. Conf. on Computer Design*, 1997.
- [6] L. McMurchie and C. Sechen, "Wta - waveform-based timing analysis for deep submicron circuits," in *Proc. Intl. Conf. on Computer-Aided Design*, 2002.
- [7] T. Sakurai and A. R. Newton, "Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas," *IEEE J. Solid State Ckts.*, Feb 1990.
- [8] G. A. F. Seber, *Multivariate Observations*. Wiley, 1984.