

Importance of Modeling Non-Gaussianities in Static Timing Analysis in sub-16nm Technologies

Praveen Ghanta and Igor Keller,
Cadence Design Systems Inc., San Jose, USA.
{pghanta, ikeller}@cadence.com.

Abstract—Static Timing analysis (STA) of integrated circuits is becoming increasingly less predictable with technology scaling now in the sub 16nm-regime. Variability in the physical and electrical characteristics of ultra-small CMOS devices and metal wires due to manufacturing process variations has a very significant impact on the timing performance of today's VLSI circuits. Most of the commercial EDA timing tools model just the mean (μ) and the standard deviation (σ) of CMOS cell delays, slews and the arrival times. However, in the sub-16nm manufacturing technologies and even in the 28nm technology at ultra-low VDDs less than 0.6V, the CMOS cell delay and slew distributions can be strongly non-Gaussian due to the low Voltage headroom ($VDD - V_{th}$). Using $(\mu \pm k\sigma)$ to approximate the tail quantiles of the strongly non-Gaussian STA metrics like the cell delays, slews, path arrival and slack times can be significantly optimistic resulting in yield problems (especially for hold paths). In this paper, we propose to model the mean-shift ($\Delta\mu$), standard deviation (σ) and skewness (γ) of the individual cell delays and slews of all the library cells in the presence of process variations. We briefly discuss methods to propagate μ, σ and γ through the timing graph and paths and to compute the tail quantiles of the arrival times given their μ, σ and γ . In this paper, we focus on determining the statistical distributions of the arrival times, however the same exact techniques are applied to the slack times as well.

I. INTRODUCTION

Manufacturing process variations in the device and interconnect parameters such as gate oxide thickness (t_{ox}), interconnect width (W) and thickness (T), dopant density and threshold voltage (V_{th}) etc., cause significant variations in the cell and interconnect delays; and thus lead to significant uncertainties in the timing performance of integrated circuits [1]. To efficiently model the uncertainty in chip-level timing analysis, statistical timing analysis that generates timing path delays as a function of the (random) process variables has been proposed and widely developed over the past decade.

Statistical timing analysis has been a heavily researched topic. A wide-body of work has been published in the statistical timing area: (a) using linear (μ and σ) analyses in the process variables based on Clark's formula for the max of Gaussian random variables, deriving and propagating bounds for the max of Gaussian random variables, and tightness probabilities, etc. [2], [3], [8], [10], [16]; (b) using quadratic models in the Gaussian random variables [4]–[6], [11]; (c)

using linear and quadratic models in the Gaussian and non-Gaussian variables [7], [9]; and (d) employing Monte Carlo and Quasi-Monte Carlo techniques for statistical analysis [15]. A somewhat exhaustive summary of the approaches proposed so far in the statistical timing analysis area is available in the review papers [12], [13].

However in spite of all the rigorous body of research work and its implementation in commercial EDA tools, statistical timing analysis didn't gain acceptance in the vast majority of the semiconductor design community. Some of the important reasons were: (1) Designers were only comfortable with performing multi-corner analysis for inter-die (global) process variations rather than using statistical timing analysis, (2) Cost of characterization to generate a fully-statistical CCS/ECSM library was prohibitively expensive, (3) Run-time and memory cost of a statistical timing run with just the local process variables was very high anywhere from 3X-10X over a single corner (non-statistical) run, and (4) Gap between implementation tools and the static timing tools in employing statistical timing analysis.

In light of the above, static timing analysis tools still resort to corner analysis for handling global process variables and employ derating/statistical techniques only for handling local (mismatch) random variables. For handling local random variations, static timing analysis tools initially went back to using smarter derating techniques like depth-based derate generation that models statistical cancellation (depth-based OCV); and POCV [14] which is essentially a poor-man's (single-variable) version of statistical timing analysis with cell delay variation specified as normalized standard deviation relative to the corner delays. Also in POCV, variations specified do not vary with slews and loads.

With timing variations becoming much more critical in the sub-16nm technologies, POCV was no longer sufficiently accurate, which led to the development of the standardized liberty variation format (LVF) [17] libraries in the recent past. LVF libraries model the absolute standard deviation of the cell-arc and constraint-arc delays as a function of the input-slews and output loads in a (NLDM-like) table format. Further, due to significant speed-ups in the variational library characterization tools in the past 5 years, generating LVF libraries has become a computationally feasible task (matter of weeks compared to months before).

However in the sub-16nm technologies and even at 28nm at ultra-low VDDs (less than 0.6V), the cell delay and slew

In this paper, we cannot provide many details as these techniques have been implemented in a commercial EDA tool and are also currently part of a patent review. We focus more on the results in this paper.

distributions are strongly non-Gaussian and computing the tail quantiles of the path delays as $(\mu \pm k\sigma, k = 3 \text{ for Gaussian})$, either makes the timing analysis too optimistic or too pessimistic depending on the value of k . While optimism can lead serious yield issues, pessimism can make it very hard for the designers to sign-off on the design. So, the existing variational library characterization tools need to be enhanced to model non-Gaussian metrics like the mean-shift, skewness and kurtosis, etc. Static timing tools need to be enhanced to consume these non-Gaussian metrics and generate the tail quantiles with a reasonable accuracy when compared to Monte Carlo SPICE simulations.

II. A METHOD FOR HANDLING NON-GAUSSIANITIES IN STATIC TIMING ANALYSIS

In this paper, we propose to model the non-Gaussian metrics like the mean-shift, skewness, etc., as an extension to the existing LVF library format. We then briefly discuss ways to propagate the non-Gaussian metrics through the timing paths, and discuss ways to compute the tail quantiles of the arrival times from the non-Gaussian metrics. While we focus on arrival times in this paper, the same exact techniques are applied to the slack times as well in our tool.

A. Modeling non-Gaussian metrics of cells

The cell delay D (and output slew S) are either modeled as a non-linear Taylor series function of the underlying local (mismatch) random variable vector $\vec{x} = [x_1, \dots, x_N]^T$ or the delay and output slew Monte Carlo samples are generated from Monte Carlo SPICE sampling. From either the Taylor series or the Monte Carlo samples, the mean-shift ($\Delta\mu$), standard-deviation (σ) and skewness (γ) of the cell delay D can be obtained as follows. D_{corner} represents the corner delay, basically at $\vec{x} = \vec{0} = [0, \dots, 0]^T$.

$$\mu = E[D(x)] = \int_{-\infty}^{\infty} D(x)f(x)dx \quad (1)$$

$$\Delta\mu = \mu - D_{corner}(\vec{0}) \quad (2)$$

$$\sigma^2 = E[(D(x) - \mu)^2] = \int_{-\infty}^{\infty} D(x)^2 f(x) dx - \mu^2 \quad (3)$$

$$\gamma^3 = E[(D(x) - \mu)^3] / \sigma^3 \quad (4)$$

The values $\Delta\mu$, σ and γ are added as user-defined attributes in a LVF-format library.

B. Propagating non-Gaussian arrival moments through timing graph

A fundamental property that plays the critical role in the propagation of the non-Gaussian metrics through the timing graph, as well as computing the tail quantiles is the underlying probability distribution of the cell delays, slews and arrivals times. Given the first three moments, we need to first choose an underlying probability distribution for the delays, slews and arrival times. From our SPICE Monte Carlo experiments across

many designs and paths, we observed that some good choices of the probability distributions are: *skew-normal*, *log-normal*, *beta*, *Cauchy*, *chi-squared*, *gamma*, *student's t-distribution* etc.

The two statistical arithmetic operators that form the core of propagating the non-Gaussian metrics through the timing graph (and path) are addition and maximum. The arithmetic operators addition and maximum for non-Gaussian variables is a topic that has been dealt with extensively in the existing literature [4]–[7], [9], [11]. However, the difference here is that unlike most of the published work, it is very expensive to maintain the correlations between the individual arrival times for the statistical max operation. We therefore derive bounded non-Gaussian metrics for the maximum arrival time at the output of a logic gate in the graph-based analysis (GBA).

Let $A_j(\mu_1, \sigma_1, \gamma_1)$ and $A_k(\mu_2, \sigma_2, \gamma_2)$ represent two arrival times for which the max needs to be computed. Then,

$$\begin{aligned} & A_{max}(\mu_{max}, \sigma_{max}, \gamma_{max}) \\ &= \max(A_j(\mu_1, \sigma_1, \gamma_1), A_k(\mu_2, \sigma_2, \gamma_2)) \end{aligned} \quad (5)$$

Computation of the max operation requires that we assume an underlying probability distribution for A_i based on the first three moments. Once the underlying probability distribution is chosen, then the moments μ_{max} , σ_{max} and γ_{max} of the *estimated* maximum are obtained from moment-matching and numerical integration [4]–[7], [9], [11]. The correlation value between A_1 and A_2 is assumed so that the tail quantiles of A_{max} bound the tail quantiles of both A_j and A_k (for e.g., zero correlation).

However, in the case of the addition operation between the arrival time A_i and the cell-arc-delay $D_{\{i,i+1\}}$, the correlation between A_i and $D_{\{i,i+1\}}$ usually results from the input slew used to calculate $D_{\{i,i+1\}}$ and is readily available. So, the moments of the output arrival time A_{i+1} can be readily computed from moment matching and numerical integration [4]–[7], [9], [11].

$$\begin{aligned} & A_{i+1}(\mu_{i+1}, \sigma_{i+1}, \gamma_{i+1}) \\ &= A_i(\mu_i, \sigma_i, \gamma_i) + D_{\{i,i+1\}}(\mu_d, \sigma_d, \gamma_d) \end{aligned} \quad (6)$$

C. Obtaining the tail quantiles of arrival times

One of the key final steps after propagating the arrival times in the forward traversal (and a similar backward traversal of required times) is to compute the tail quantiles of the arrival and slack times. Given the first three moments, based on the underlying probability distribution being assumed (*skew-normal*, *log-normal*, *beta*, *Cauchy*, *chi-squared*, *gamma*, *student's t-distribution* etc.), the tail quantiles are computed analytically and whenever possible based on table lookup of the mean-shift, sigma and skewness values.

Since path-based analysis (PBA) employs only the addition operation, arrival and the slack quantiles can be obtained with relatively-good accuracy compared to Monte Carlo simulations. Since graph-based analysis (GBA) employs both the maximum and the addition operations, it's sufficient that the GBA arrival quantiles are accurate enough to bound the PBA arrival quantiles.

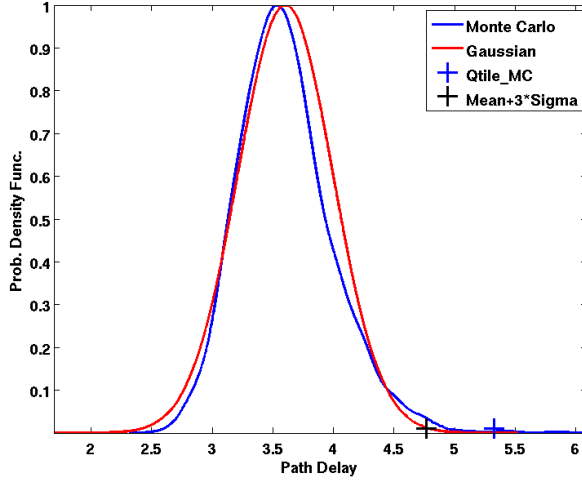


Fig. 1. Monte Carlo PDF vs. Gaussian Analysis in tool

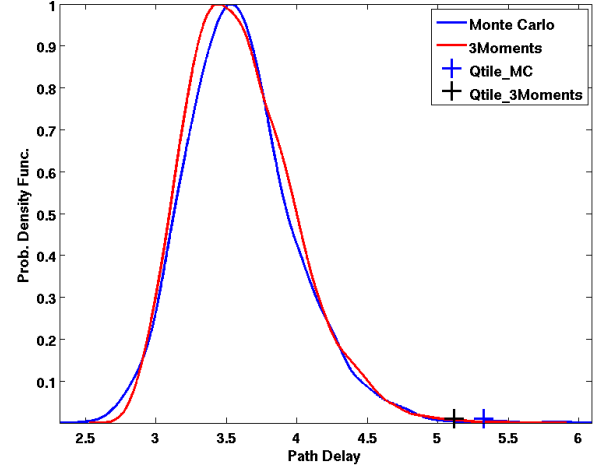


Fig. 2. Monte Carlo PDF vs. 3Moment Analysis in tool

III. RESULTS

As part of the experimental setup, we ran sub-16nm and 28nm ultra-low-VDD ($< 0.6V$) designs through our static timing tool while propagating the non-Gaussian arrival moments through the timing graph. To compare the accuracy of our approach, we selected a representative set of several hundred paths from these designs, and generated the spice decks for all the launch paths from inside our static timing tool. We then performed latin-hyper-cube (LHS) sampling-based Monte Carlo SPICE simulations for the arrival times of these hundreds of paths each with 5000 samples. In this section, for about 50 of those representative paths, we present the PBA (path-based-analysis) arrival-time tail-quantile results from our tool compared to those from the golden-standard Monte Carlo SPICE simulations. In this work, we consider the 0.99865 tail quantile which corresponds to the 3σ quantile for a Gaussian-distributed random variable.

Table I shows the 0.99865 arrival quantiles from Monte Carlo simulations, arrival quantiles from our tool using the first three moments (mean-shift, sigma, skewness) and arrival quantiles from our tool using the Gaussian moments (mean-shift, sigma). Table I also shows in braces *error_cmp2MC*, which is the % relative error in estimating the 0.99865 tail quantile compared to Monte Carlo simulations. *error_cmp2MC* is defined as $(Qtile_{tool} - Qtile_{MC})/Qtile_{MC}$.

From Table I, it can be observed that propagating just the Gaussian metrics of the arrival time distributions can lead to large optimistic errors of the order of -10% to -19% in the 0.99865 arrival quantiles. In view of the large outliers in the low-voltage-headroom technologies due to Gaussian analysis, it is imperative that non-Gaussian metrics like the mean-shift, skewness, etc., be modeled and propagated in the static timing tools.

To illustrate the inaccuracies involved in performing only a

Gaussian analysis in STA, Figure 1 compares the probability density function (PDF) and the 0.99865 tail quantile of a PBA path from our tool to the results from Monte Carlo SPICE simulations. For the same path, Figure 2 compares the PDF and tail quantile from our tool using the first 3 moments to the results from Monte Carlo SPICE simulations.

IV. CONCLUSIONS

In this paper, we described the importance of modeling the non-Gaussian metrics like cell delays, slews and arrival times in the low-voltage-headroom technologies like the sub-16nm and even the 28nm at ultra-low VDDs. We proposed to capture the non-Gaussian metrics like mean-shift, skewness, etc., through variational cell-characterization; and proposed enhancements to the existing LVF library format in the form of user-defined attributes to include these non-Gaussian metrics. We also briefly discussed the core operations involved in propagating these non-Gaussian metrics through the timing-graph and paths.

We demonstrated the significant optimism risk in modeling the arrival-time tail quantiles through Gaussian analysis in STA by comparing to the golden-standard Monte Carlo SPICE simulations. We also demonstrated the good accuracy in modeling the arrival-time tail quantiles through propagation of the non-Gaussian metrics by comparing to the Monte Carlo SPICE simulations. The results presented in this paper demonstrate that modeling the non-Gaussianities in static timing analysis is inevitable in the low-voltage-headroom technologies to avoid risking yield issues.

TABLE I

RESULTS COMPARING 0.99865 QUANTILE OF ARRIVAL TIMES FROM OUR TOOL VS 5000 MC SIMULATIONS

Path #	Qtile_MC	Qtile_3Moments (modeling $\Delta\mu$, σ , γ)	Qtile_Gaussian (modeling $\Delta\mu$, σ)
	nanosec	nanosec % error_cmp2MC	nanosec % error_cmp2MC
1	6.742	6.942 3.0	6.554 -2.8
2	8.209	7.551 -8.0	6.920 -15.7
3	5.742	5.773 0.5	5.344 -6.9
4	5.393	5.329 -1.2	4.976 -7.7
5	8.315	8.287 -0.3	7.706 -7.3
6	11.243	10.997 -2.2	9.691 -13.8
7	7.923	7.762 -2.0	7.202 -9.1
8	8.563	8.126 -5.1	7.436 -13.2
9	6.082	6.133 0.8	5.765 -5.2
10	8.733	8.582 -1.7	7.980 -8.6
11	6.434	6.197 -3.7	5.654 -12.1
12	6.626	6.571 -0.8	6.158 -7.1
13	7.252	7.180 -1.0	6.609 -8.9
14	6.074	6.148 1.2	5.761 -5.1
15	9.569	8.954 -6.4	8.188 -14.4
16	11.071	10.761 -2.8	9.603 -13.3
17	7.361	7.232 -1.8	6.562 -10.8
18	6.831	6.816 -0.2	6.318 -7.5
19	6.322	6.196 -2.0	5.837 -7.7
20	8.089	7.978 -1.4	7.320 -9.5
21	7.833	7.907 0.9	7.082 -9.6
22	6.296	6.087 -3.3	5.705 -9.4
23	12.829	12.170 -5.1	10.498 -18.2
24	10.083	9.715 -3.7	8.990 -10.8
25	8.505	8.292 -2.5	7.340 -13.7
26	6.911	6.906 -0.1	6.487 -6.1
27	9.385	8.569 -8.7	7.566 -19.4
28	9.551	9.198 -3.7	8.449 -11.5
29	4.383	4.332 -1.2	4.128 -5.8
30	4.674	4.529 -3.1	4.295 -8.1
31	3.369	3.399 0.9	3.255 -3.4
32	3.260	3.145 -3.5	3.031 -7.0
33	5.131	5.137 0.1	4.948 -3.6
34	6.733	6.608 -1.9	6.213 -7.7
35	4.818	4.852 0.7	4.605 -4.4
36	4.918	4.850 -1.4	4.649 -5.5
37	3.721	3.733 0.3	3.660 -1.6
38	5.349	5.335 -0.3	5.072 -5.2
39	3.719	3.753 0.9	3.568 -4.1
40	4.060	3.964 -2.4	3.802 -6.4
41	3.306	3.320 0.4	3.204 -3.1
42	4.302	4.168 -3.1	3.972 -7.7
43	6.234	6.246 0.2	5.883 -5.6
44	5.428	5.411 -0.3	5.209 -4.0
45	4.193	4.178 -0.4	4.033 -3.8
46	5.306	5.310 0.1	5.015 -5.5
47	4.693	4.522 -3.6	4.249 -9.5
48	3.816	3.724 -2.4	3.623 -5.1
49	4.072	3.933 -3.4	3.795 -6.8
50	5.462	5.497 0.6	5.245 -4.0

REFERENCES

- [1] D. Boning and S. Nassif, Models of Process Variations in Device and Interconnect, in *Design of High-Performance Microprocessor Circuits, 1st ed.*, A. Chandrakasan, W. J. Bowhill, and F. Cox, Eds. IEEE Press, 2001.
- [2] H. Chang and S. S. Sapatnekar, Statistical Timing Analysis Considering Spatial Correlations using a Single Pert-Like Traversal, *IEEE/ACM international conference on Computer-aided design, November 2003*.
- [3] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, S. Narayan, First-order incremental block-based statistical timing analysis, *Proceedings of the 41st annual Design Automation Conference, June 2004*.
- [4] L. Zhang, W. Chen, Y. Hu, J. A. Gubner, C.C. Chen, Correlation-preserved non-gaussian statistical timing analysis with quadratic timing model, *Proceedings of the 42nd annual Design Automation Conference, June 2005*.
- [5] H. Chang, V. Zolotov, S. Narayan, C. Visweswariah, Parameterized block-based statistical timing analysis with non-gaussian parameters, nonlinear delay functions, *Proceedings of the 42nd annual Design Automation Conference, June 2005*.
- [6] Y. Zhan, A. J. Strojwas, X. Li, L. T. Pileggi, D. Newmark, M. Sharma, Correlation-aware statistical timing analysis with non-gaussian delay distributions, *Proceedings of the 42nd annual Design Automation Conference, June 2005*.
- [7] J. Singh, S. Sapatnekar, Statistical timing analysis with correlated non-gaussian parameters using independent component analysis, *Proceedings of the 43rd annual Design Automation Conference, July 2006*.
- [8] A. Agarwal, V. Zolotov, D. T. Blaauw, Statistical timing analysis using bounds and selective enumeration, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, November 2006*.
- [9] S. Bhardwaj, P. Ghanta, S. Vrudhula, A framework for statistical timing analysis using non-linear delay and slew models, *Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design, November 2006*.
- [10] D. Sinha, H. Zhou, Statistical Timing Analysis With Coupling, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, December 2006*.
- [11] L. Cheng, J. Xiong, L. He, Non-linear statistical static timing analysis for non-Gaussian variation sources, *Proceedings of the 44th annual Design Automation Conference, June 2007*.
- [12] D. Blaauw, K. Chopra, A. Srivastava, L. Scheffer, Statistical Timing Analysis: From Basic Principles to State of the Art, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, April 2008*.
- [13] C. Forzan and D. Pandini, Statistical static timing analysis: A survey. *Journal of VLSI Integration June 2009*.
- [14] A. Mutlu, J. Le, R. Molina, M. Celik, A parametric approach for handling local variation effects in timing analysis. *Proceedings of the 47th Design Automation Conference, July 2009*.
- [15] V. Veetil, Y. Chang, D. Sylvester, D. Blaauw, Efficient smart Monte Carlo based SSTA on graphics processing units with improved resource utilization, *Proceedings of the 47th Design Automation Conference, June 2010*.
- [16] D. Sinha, C. Visweswariah, N. Venkateswaran, J. Xiong, V. Zolotov, Reversible statistical max/min operation: concept and applications to timing, *Proceedings of the 49th Annual Design Automation Conference, June 2012*.
- [17] <http://www.opensourceliberty.org>