Probability-Based Delay Analysis and Tuning of VLSI Circuits Using a Variance-Covariance Method

Itiakorit J. Osele and Ramalingam Sridhar
Department of Electrical and Computer Engineering
The State University of New York at Buffalo
Buffalo, NY 14260

Abstract

A probability-based method is presented for the timing delay analysis and tuning of high performance VLSI circuits. The focus is on minimizing signal path skew, while using a set of wave-pipelined circuits as a vehicle of study. A recursive variance-covariance technique that assumes Gaussian delay distributions is used to i) balance, ii) quantify and iii) fine tune the critical path skew, and iv) calibrate performance. The results on the effectiveness of the technique to analyze and fine tune the wave-pipeline combinational circuits are reported.

1 Introduction

The main objective of this paper is to minimize the difference between the longest and the shortest path lengths. The goals are i) to obtain realistic and accurate differences between the longest and shortest paths, and ii) tune the path difference to improve circuit performance and reliability.

In wave-pipeline design methodology, the objective is overcoming delay variation because it is crucial that the number of waves of computation that may co-exist within the circuit do not interfere with each other, resulting in lose of data. Therefore, there is a need to keep the difference between the critical longest and shortest path lengths precise and minimal since the circuit throughput, in terms of the number of waves that can be pipelined, depends on having a minimal difference between the longest and shortest path lengths.

However, there are many variable factors that affect the delay variation, namely, i) circuit topology or architecture, ii) the data patterns of computations, iii) capacitive loading, iv) process, and v) temperature and voltage usually referred to as environmental variations. This approach, though, addresses delay changes influenced by process variations.

Until now this problem has been solved using deterministic modeling, unfortunately, when the delays, as done elsewhere [1] [2], are assumed i) fixed and ii) known in advance

subsequent delay variations are not accounted for, therefore use of probabilistic delay times. In a similar probabilistic approach to a related problem, an upper bound on the expected clock skew in synchronous systems is reported [3]. However, the results are based on the determination of the range of samples with identical and independent distributions, i.e., Gaussian random variables with equal means and variances. Therefore, the basic model is limited to instances where wire lengths are equalized in a multistage clock distribution system. But it does not, as in this instance, apply when wire lengths are disparate, capacitive loading is nonuniform, and delays are tuned. Here the delays are neither independent nor identical with equal means and variances.

The traditional Critical Path deterministic approach is not effective under high performance design, increased device and circuit variability, and submicron design, since the delays are probabilistic and the problem is complicated [4] [5], requiring a different solution approach. A probabilistic-based model approach in which the deterministic one is a special case of step function distributed delays is proposed.

Using a recursive variance-covariance technique, based on a unimodal Gaussian delay distribution assumption, the circuit is balanced using the deterministic approach developed by Wong *et al.* [1], while delays are fixed at mean values. Further, the longest and shortest path lengths are determined. These results are then used to tune the path delays by modifying delays of gates to whose delay changes the path length difference is sensitive, to improve performance. The path delays are tuned by modifying the delay distributions of the balancing gates, i.e., by resizing, and the performance is calibrated. Other results which connect the circuit delay and process or physical design have been established. For example, it is possible to tune the delays by optimizing technology parameters [6] or designing signal distribution networks that minimize signal spread [7]. In [6], it is found that in low voltage regimes the oxide thickness dominates the delay while gate length dominates in high voltage regimes. Using a 23-stage CMOS ringoscillator in the low voltage regime for a vdd of 1.2V, a 12.0 nm oxide thickness gives an optimal minimal delay distribution.

The problem is defined in section 2, while section 3 discusses the derivation of analytical formulations and the tuning procedures, and their implementation details. The results are presented in section 4 and, finally, concluding remarks are made in section 5.

2 Problem Definition

Given a combinational logic circuit network depicted in Figure 1, let a set of nodes $V = \{v_1, v_2, ..., v_n\}$ represent the gates or a group of gates. In addition, let the interconnections between the gates be represented by a set of edges $E = \{e_1, e_2, ..., e_m\}$.

Furthermore, let the delays at each node and edge take on values that are functions of architecture, computation data patterns, transistor and wirelength sizing, loading and loses, and process, environmental and manufacturing variations. The circuit is then modeled as an acyclic graph, whereby the inputs and outputs are represented as source and sink nodes. A set of paths are identifiable, in which each path is a sequence of nodes that connect the source and sink nodes.

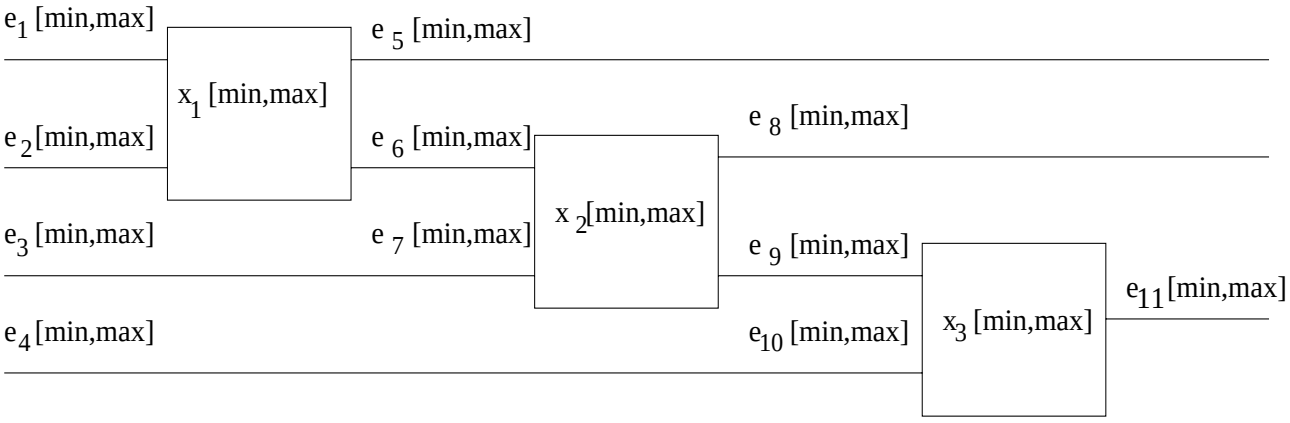

Figure 1: Combinational Network to be Wave-Pipelined

2.1 Probabilistic Modeling

Each node delay is modeled as a random variable, and the path lengths being sums of node values are also random variables. Therefore, no one path will always be critical, i.e., the longest or the shortest, thus only probabilistic formulations are meaningful. Since in wave-pipeline design a key objective is overcoming delay variation, this approach uses the distribution of the nodal delays to develop methods of minimizing the differences between the longest and shortest paths to improve performance.

2.2 Performance of a Wave-Pipelined Circuit

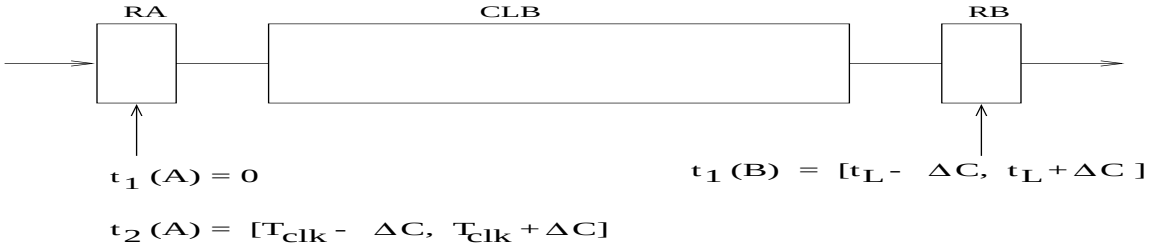

Figure 2: A Wave-Pipelined Combinational Circuit

Let a combinational logic circuit shown in Figure 2, be modeled as a network of a set of path lengths P, also referred to as a set of paths, each path being an interconnected sequence of nodes having delays X_i's, that connect the output of input register RA to the input of output register RB. Each path length P_j is then $P_j = X_1 + X_2 + ... + X_i + ... + X_l$, where l is the number of nodes in path P_j. The longest path length $L \in P$ is given by $\max P = \max[P_1, P_2, ..., P_k]$, where k is the number of paths in the network. Similarly the shortest path length S is given by $\min P = \min[P_1, P_2, ..., P_k]$.

Two key conditions determine the performance of a wave-pipeline circuit. One, the earliest clock to capture a wave launched at register RA at time $t_1(A) = 0$ should occur after the data wave propagating through the longest path length has arrived at register RB.

Two, the following earliest wave should arrive at register RB after the latest first wave has been clocked through register RB. Combining the two conditions gives the overall governing relationship as

$$U = \max P - \min P < T_{clk} - 2\Delta C - t_S - t_H \tag{1}$$

where T_{clk} is the clock width, ΔC is the clock skew, t_S and t_H are the setup and holdup times, at the registers RA and RB. The objective is to get a lowest bound on the clock period by minimizing the left hand side, the path length difference, U.

3 Approach

The path delays are the sum of individual node delays which can be approximated to a Gaussian distribution for the following reasons. One, by the Central Limit Theorem, the sum of random variables regardless of their distribution tends to a normal distribution. Two, in the case of process delays, the individual process node delays, in general, tend to have normal distributions. Three, the path dependencies are easily resolved since the node delay means and variances are additive. The problem of finding the longest and shortest paths is then recast as that of finding the greatest and the smallest of the normally distributed random variables. A solution and an approximate solution of finding the greatest of a set of n random variables, for $n = 2$ and $n > 2$, was first reported by C. E. Clark [8]. In addition, Sculli *et al* [9] report recursive formulations extending the same results to determining an approximate solution to a PERT problem. By using an early signal model, this study, in turn, extends these results to finding the shortest paths, and subsequently the difference between the shortest and longest paths. The determination of the longest path is a direct application of the recursive approximate formulations. Finding the shortest path is recast as finding the greatest of negative random variables.

The estimates of the path lengths are less accurate if the means and the variances of the random variables are identical, a characteristic of a balanced wave-pipelined circuit. However, that is offset when the random variables are dependent, which is also a common feature of signal paths in any combinational circuit.

Considering k paths, the mean and variance, μ_i and σ_{ii} respectively of path P_i can be obtained by adding the respective means and variances of delays of the nodes that make up path P_i, assuming that path P_i is normally distributed. The covariance, σ_{ij} of any two paths P_i and P_j is determined as follows. Paths P_i and P_j will have a set of nodes that are common to both of them, otherwise the covariance will be zero, $\sigma_{ij} = 0$. Let the sum of delays of nodes common to paths P_i and P_j be v_{ij} and let w_i and w_j be the random variables representing the delay of nodes uncommon to paths P_i and P_j, then

$$P_i = v_{ij} + w_i \tag{2}$$

$$P_j = v_{ij} + w_j \tag{3}$$

The covariance, σ_{ij}, is given by

$$\sigma_{ij} = \mathrm{cov}\{P_i, P_j\} = \mathrm{cov}\{v_{ij} + w_i, v_{ij} + w_j\} = \mathrm{var}\{v_{ij}\} \tag{4}$$

Thus, the covariance of paths P_i and P_j is equal to the sum of variances of delays of the nodes common to both paths. Let the P^T be the transpose of $P = \{P_1, ..., P_k\}$, $\Lambda = [\sigma_{ij}]$, the multivariate normal distribution symmetric variance-covariance matrix and $M^T = [\mu_1, ..., \mu_k]$ is vector of means.

$$f(P, M, \Lambda, k) = \frac{\exp\{-\frac{1}{2}(P - M^T)^T \Lambda^{-1}(P - M)\}}{(2\pi)^{\frac{1}{2}}|\Lambda|^{\frac{1}{2}}} \tag{5}$$

The literature is replete with a derivation of the above distribution multivariate equation, from which the mean, variance and distribution of the shortest and longest paths of the set of paths $P = \{P_1, ..., P_k\}$ is determined using recursive formulations reported in [9], and for the sake of easy reference reproduced below in equations 6 - 20.

$$\begin{aligned}
E_1 &= \mu_1 && (6)\\
V_1 &= \sigma_1 && (7)\\
E_2 &= \mu_1\Psi(\beta_2) + \mu_2\Psi(-\beta_2) + b_2\psi(\beta_2) && (8)\\
V_2 &= (\mu_1^2 + \sigma_2^2)\Psi(\beta_2) + (\mu_2^2 + \sigma_2^2)\Psi(-\beta_2) + (\mu_1 + \mu_2)b_2\psi(\beta_2) - E_2^2 && (9)\\
E_k &= E_{k-1}\Psi(\beta) + \mu_k\Psi(-\beta_k) + b_k\psi(\beta_k) && (10)\\
V_k &= (E_{k-1}^2 + V_{k-1}^2)\Psi(\beta_k) + (\mu_k^2 + \sigma_k^2)\Psi(-\beta_k) + (E_{k-1} + \mu_k)b_2\psi(\beta_k) - E_k^2 && (11)\\
b_k^2 &= V_{k-1} + \sigma_k^2 - 2\sqrt{V_{k-1}}r[\max(P_1, ..P_{k-1}), P_k] && (12)\\
\beta_k &= \frac{E_{k-1} - \mu_k}{b_k} && (13)\\
r[\max(P_1, .., P_j), P_k] &= \frac{\sqrt{V_{j-1}}[\max(P_1, .., P_{j-1}), P_k]\Psi(\beta_j) + \sigma_j r(P_j, P_k)\Psi(-\beta_j)}{\sqrt{V_j}} && (14)\\
r[\max(P_1, .., P_{k-1})] &= r(\max[\max(P_1, .., P_{k-2}), P_{k-1}], P_k) && (15)\\
r(\max(P_1, P_2), P_3) &= \frac{\sigma_1 r(P_1, P_2)\Psi(\beta_2) + \sigma_2 r(P_2, P_3)\Psi(-\beta_2)}{\sqrt{V_2}} && (16)\\
b_2^2 &= \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2 r(P_1, P_2) && (17)\\
\beta_2 &= \frac{\mu_1 - \mu_2}{b_2} && (18)\\
\psi &= \frac{1}{\sqrt{2\pi}}\exp(-\beta^2/2) && (19)\\
\Psi &= \int_{-\infty}^{\beta}\frac{1}{\sqrt{2\pi}}\exp(-\beta^2/2) && (20)
\end{aligned}$$

where E_i, V_i represent the mean and variance of the largest of i paths.

The recursive procedure works as follows: First, sort the path lengths in ascending means. Two, starting with a path length with the largest mean, μ_1, and variance, σ_1, the mean and variance of E_i, V_i, are determined recursively, each instance adding a path with the next largest mean until all paths are added.

3.1 Fine Tuning

After the circuit is rough tuned its nodes form an array. Fine tuning is done by direct optimization, i.e., varying the mean and the probability distributions of one *column* of nodes, at a time, starting from the output working towards the input, while keeping the means of the rest of the nodes fixed. The actual size of node depends on its capacitive loading.

The following is the summary of the tuning (both rough and fine) method:

1. balance (rough tune as in [1]) using mean gate delays
2. analyze using variance-covariance method
3. fine tune
4. repeat (2) and (3) until there is no decrease in the path length difference.

4 Results

Shown in Table 1, are the results of analysis and fine tuning (depicted in paranthesis) for the longest and shortest paths and their differences for the two combinational circuits; a generic circuit under design, CUD, Figure 3, and a four bit carry look ahead adder, CLA, Figure 4. Both are implemented using NAND/AND and delay elements of a family of wave-pipeline transmission gate logic [10], represented in the figures as boxes with solid and dotted lines. The distribution of the path difference is specified by the difference in their means and the sum of their variances, assuming the shortest and longest path lengths are independent. The CUD has six functional nodes, six balancing nodes, and six paths, while the CLA has equivalents of 15, 6, and 38. Gaussian delay distributions with typical mean delay values for the gates and delay elements of 0.8701 ns and 1.0459 ns were determined using HSPICE simulation of the CUD consisting of NAND/AND gates and balancing elements at nominal conditions of 25 degree Celcius, a 5V, Vdd, power source and a 2 μm MOSIS process.

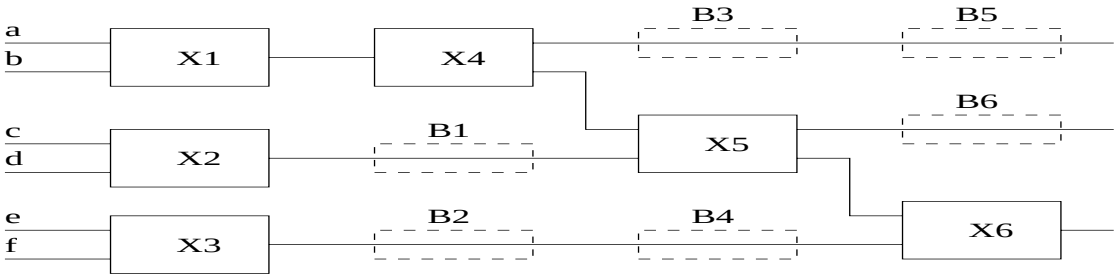

Figure 3: Circuit Under Design, CUD

5 Conclusion

Realistically accurate and computationally efficient estimations of the critical path lengths of wave-pipeline circuit networks are achieved using a recursive variance-covariance approach.

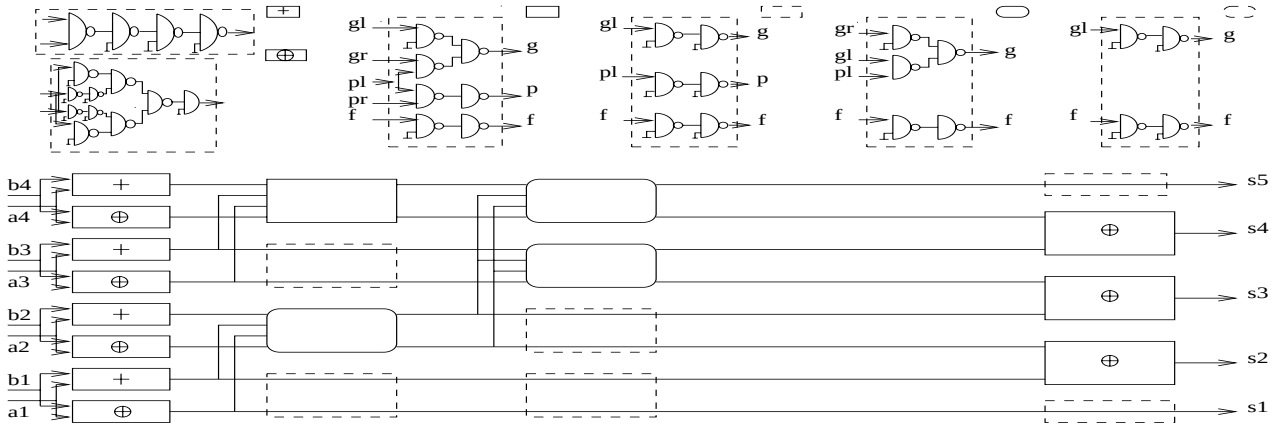

Figure 4: Carry look ahead, CLA, adder using NAND/AND gates

Table 1: Comparison of distributions of the longest, L, shortest, S, paths and their difference, U, of a circuit under design, CUD, and carry look ahead, CLA, adder for variance-covariance method, results after fine tuning are shown in paranthesis.

Circuit	Longest, shortest and difference paths	Path Length Estimates	
		Mean, μ, sec.	Variance, σ^2, sec^2
CUD	L	3.591060e-09 (3.490453e-09)	1.962982e-21
	S	3.484895e-09 (3.015779e-09)	8.558011e-21
	U	1.061649e-10 (0.474674e-10)	9.520993e-21
CLA	L	1.184675e-08 (1.094707e-08)	1.956117e-20
	S	1.020351e-08 (0.987145e-08)	1.909795e-20
	U	1.643242e-09 (1.075618e-09)	3.865912e-20

Both the longest and shortest path problems are solved using the same technique. Tuning or minimization of the difference length between the longest and shortest path lengths are shown. The possibility of pipelining more than one wave [11] is supported.

References

[1] D. C. Wong, G. D. Micheli, and M. J. Flynn, "Designing high-performance digital circuits using wave pipelining: algorithms and practical experiences," *IEEE Trans. on Computer-Aided Design*, pp. 25–45, 1993.

[2] C. T. Gray, W. Liu, and R. K. Cavin, *Wave pipelining: theory and CMOS implementation.* Boston, MA: Kluwer Academic Publishers, 1994.

[3] S. D. Kugelmass and K. Steiglitz, "An Upper Bound on Expected Clock Skew in Snychronous Systems," *IEEE Trans. on Computers*, vol. 39, no. 12, pp. 1475–1477, 1990.

[4] D. M. Nicol, R. Simha, and D. Towsley, "Static Assignment of Stochastic Tasks Using Majorization," *IEEE Transc. on Computers*, vol. 45, pp. 272–296, 1996.

[5] D. Wessels and J. Muzio, "Analysing and improving delay defect tolerance in pipeline combinational circuits," in *IEEE Int'nal. Workshop on Defect and Fault Tolerance in VLSI Systems*, pp. 181–188, 1995.

[6] H. Hoenigschmid, M. Miura-Mattausch, O. Prigge, A. Rahm, and D. Savignac, "Optimization of Advanced MOS Technologies for Narrow Distribution of Circuit Performance," *IEEE Trans. on CAD of ICAS*, vol. 16, no. 2, pp. 199–204, 1997.

[7] T. Gneiting and I. P. Jalowiecki, "Influence of Process Parameter Variations on the Signal Distribution Behavior of WSI Devices," *IEEE Trans. on CPMT-PART B.*, vol. 18, no. 3, pp. 424–430, 1995.

[8] C. Clark, "The greatest of a finite set of random variables," *Operations Research*, vol. 10, pp. 145–162, 1961.

[9] D. Sculli and Y. W. Shum, "An approximate solution to the PERT problem," *Computers Math. Applications*, vol. 21, pp. 1–7, 1991.

[10] X. Zhang and R. Sridhar, "CMOS wave pipelining using transmission gate logic," in *Proc. IEEE Intl. ASIC Conf. and Exhibit*, (Rochester, NY), pp. 92–95, 1994.

[11] K. J. Nowka and M. J. Flynn, *Environmental limits on the performance of CMOS wave-pipelined circuits.* Stanford University, 1994.