IJCAT - International Journal of Computing and Technology, Volume 3, Issue 11, November 2016
ISSN: 2348 - 6090
www.IJCAT.org

Statistical Static Timing Analysis for VLSI Design of Complex Circuits

[1] R. S. Halke , [2] Bhushan Malkapurkar

[1,2] Dr. DYPSOET, Lohgaon, Pune, 412105

Abstract - As the CMOS technology is scaling down to the nanometer regime, process variations have been increased. In particular, the increase of delay variations has seriously affected the design periods and timing yields. To estimate more accurately these delay variations, statistical static timing analysis (SSTA), which considers delay variations statistically, has been proposed. The MOS transistor is the basic building block of integrated circuits. Scaling of the MOS transistor improves its size, cost and performance. Today's fabricated integrated circuits are many times faster and occupy much less area, like today's microprocessors that contain nearly one billion transistors on a single chip. In such designs, it is important to the timing yield at the design phase because, at this phase, we can consider the trade-offs between chip performance and yield.

Keywords **- SSTA, Scaling, variations etc.**

1. Introduction

The MOS transistor is the basic building block of integrated circuits. Scaling of the MOS transistor improves its size, cost and performance. Today's fabricated integrated circuits are many times faster and occupy much less area, like today's microprocessors that contain nearly one billion transistors on a single chip.[1] The role of supply voltage is vital for controlling the power consumption and hence reducing the power dissipation. It is reducing for each new technology generation.

With CMOS technology scaling down to the nanometer regime, process variations have been increased. In particular, the increase of delay variations has seriously affected the design periods and timing yields. To estimate more accurately these delay variations, statistical static timing analysis (SSTA), which considers delay variations statistically, has been proposed. SSTA is expected to shorten the design turnaround time (TAT) and predict the timing yields.[2] In application specific integrated circuit (ASIC) designs, sufficient margins for delay variations are required to achieve the target frequency at the expected yield. In process technologies above 90nm, the margins for delay variations are small enough and their impact on design can be eliminated. However, at 90nm and below, the increased delay variations enlarge the margins in circuit design. This results in overestimations of circuit delay and makes design work difficult.[3]In high performance microprocessor designs, excessive margins make it difficult to achieve the target performance. So, nominal values are used at the design phase. After a batch of chips has been fabricated, the frequency selection process sorts them into several ranks according to the measured maximum frequency. Then, the chips are priced. In such designs, it is important to the timing yield at the design phase because, at this phase, we can consider the trade-offs between chip performance and yield.[4] SSTA which analyzes circuit delays statistically by considering delay variations is attracting my interest as a solution to the above issues.[5]

2. Types of Variations

The delay variations are due to various types of process variations. Process variations result from perturbations in the fabrication process, due to which the nominal values of parameters such as the effective channel length (L_{eff}), the oxide thickness (t_{ox}), the dopant concentration (N_a), the transistor width (W), the interlayer dielectric (ILD) thickness (t_{ILD}), and the interconnect height and width (h_{int} and W_{int} respectively). Environmental variations arise due to changes in the operating environment of the circuit, such as the temperature or variations in the supply voltage (V_{dd} and ground) levels. Both of these types of variations can result in changes in the timing and power characteristics of a circuit.

3. Sources of Variations

3.1 Random Variations

The name itself implies random behavior that can be characterized in terms of a distribution. This distribution

may either be explicit, in terms of a large number of samples provided from fabrication line measurements, or implicit, in terms of a known probability density function (such as a Gaussian or a lognormal distribution) that has been fitted to the measurements. Random variations in some process or environmental parameters (such as those in the temperature, supply voltage, or L_{eff}) can often show some degree of local spatial correlation, whereby variations in one transistor in a chip are remarkably similar in nature to those in spatially neighboring transistors, but may differ significantly from those that are far away. Other process parameters (such as t_{ox} and *Na*) do not show much spatial correlation at all, so that for all practical purposes, variations in neighboring transistors are uncorrelated.

3.2 Systematic Variations

These variations show predictable variation trends across a chip, and are caused by known physical phenomena during manufacturing. Strictly speaking, environmental changes are entirely predictable, but practically, due to the fact that these may change under a large number (potentially exponential in the number of inputs and internal states) of operating modes of a circuit, it is easier to capture them in terms of random variations.

Steps of design process and their resulting timing uncertainties is depicted in fig:4. As the device size shrinks the device size the process parameters have predominant role in timing violations. The random behavior of process parameters makes the optimization a difficult task, thereby reducing the accuracy.

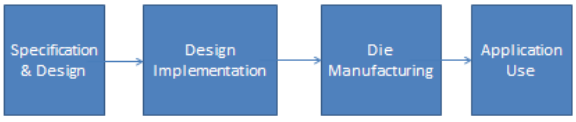

Figure 4: Steps of Design Process and Their Resulting Timing Uncertainties

The errors that occur while performing timing analysis can be classified in to three main categories.

1. Modeling and analysis errors
2. Manufacturing variations
3. Operating context variations

Once the design specifications are ready, next step is to model the design. After modeling, it undergoes several testing and verification stages like power consumption, delay, layout, floor planning etc. The outcome of this step may deviate from the expected ones. After making necessary corrections, it goes to fabrication level. The challenges faced at this level are variations due to any limitation in process, process equipment imperfections and imprecisions etc. After fabrication, the fabricated device ready to use. There it faces uncertainties from environment.

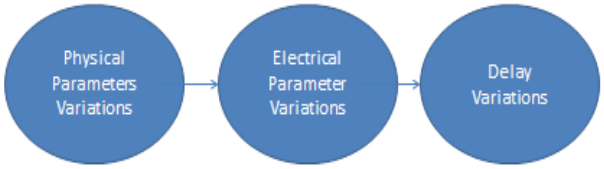

Figure 5: Sources of Variations in SSTA

Sources of Variations in SSTA are shown in fig: 1.5. Any variations in physical parameters like critical dimension, oxide thickness etc., will lead to electrical variations. Those electrical variations cause delay variations like gate

4. Basic of SSTA

In the conventional design flow, static timing analysis (STA) is used to estimate the circuit delay and maximum frequency. To assure sufficient yield, STA analyzes corner cases, in which all the factors of the delay variations are at the worst-case or best-case corner values. In actual chips, however, the probability of all factors being at the corner values is very low. Therefore, STA estimates a delay that rarely occurs in actual processes; that is, it analyzes using excessive margins for delay variations.

The main concept of SSTA is to statistically consider the random variations of WID in order to analyze circuit delay more accurately. The simplest method of statistical calculation is Monte Carlo simulation. However, the computation time of this method increases drastically according to the number of variation factors and the circuit scale. For this reason, Monte Carlo simulation is not practical for analyzing actual designs. Therefore, many researchers have studied the basic SSTA method, and many of their results have been reported, starting from about 2000. The basic SSTA method defines the random variations of the delay as random variables and calculates the probability density function (PDF) of circuit delay. The method saves computation time while producing results equivalent to those of Monte Carlo simulation.

4.1 SSTA Operations

In SSTA processing, a circuit is expressed by a graph that represents the gates and interconnects as nodes. Traversing

the graph, the PDF of the delay in each node is calculated using the statistical sum and max operations with the delay variations of the gates and interconnects as inputs.

The basic operations for two circuits with the interconnect delays ignored to simplify the calculations. Figure 6a shows a circuit with two gates connected in which two signals converge at the output pin of a gate. The delay of the series circuit in Figure 2a is calculated using a statistical sum operation. With the delay PDF of these gates denoted as f1 and f2, the delay PDF at the end point is the statistical sum of f1 and f2. However, when f1 and f2 have normal distributions, a simple formula can be used.

In Figure 2a, when f1 has a normal distribution with m1 (average) and s1 (3σ), and f2 has a normal distribution with m2 (average) and s2 (3σ), f has a normal distribution with m1 + m2 (average) and $\sqrt{S_1^2 + S_2^2}$ (3σ). The value equivalent to 3σ of this distribution is m1 + m2 + $\sqrt{S_1^2 + S_2^2}$. In this case, when this delay is calculated using the conventional STA method, the delay is the sum of the worst-case values. When the worst-case values are equivalent to 3σ, the delay is m1 + m2 + s1 + s2. Therefore, the delay of the series circuit calculated with SSTA is smaller than that calculated with STA. The output delay of the multiple-input gate in Figure 6b is calculated using a statistical max operation. It is generally difficult to calculate an accurate value for the statistical max. However, when the two random variables for f1 and f2 are independent of each other, an accurate solution can be obtained. Conversely, when f1 and f2 correlate with each other, it is difficult to obtain an accurate solution of the statistical ma operation and only an approximation is possible by using the upper bound or by lower bound of the PDF calculation or by using the moment matching technique. [6]When f1 and f2 are independent, the result of the statistical max operation is known to have an upper bound and the value equivalent to 3σ is greater than the delay calculated with STA.

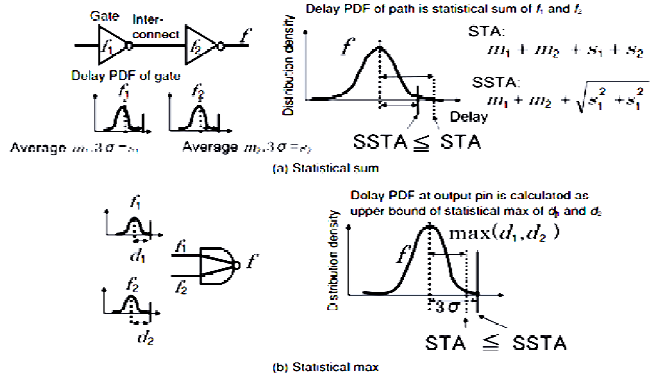

Figure 8: Basic operation of SSTA

4.2 Path-based and block- based SSTA

Path based technique: It finds the delay of each path. At the sink node it takes the maximum of those path delays. The advantage of this approach is it will not miss any critical path. The limitation of this approach is no slandered method to choose right path for analysis. Also the runtime increases exponentially.

The delay of an entire circuit can be analyzed by using the basic calculations shown above while traversing the graph. There are two types of graph analysis methods: Path-based SSTA and Block-based SSTA.

Figure shows the path-based SSTA. In this method, the delay PDF of each path is calculated individually, traversing from the source to the sink of the path. The advantage of this method is that it accurately calculates the delay PDF of each path because it does not use statistical max operations to analyze sequential paths. Also, it can consider the correlations between paths easily. However, its computation time drastically increases with the circuit scale because the number of paths increases exponentially with the circuit scale.

Block based analysis: In this approach it takes each interconnect and components as block. At each block it applies MAX operator to find the arrival time if those have multiple inputs. The runtime is linear and progressive computation is carried out. Due to non-linear behavior of MAX operator, there by reduction in accuracy. Block based analysis uses wide number of MAX operator.

Figure shows the block-based SSTA. In this method, all paths are analyzed simultaneously by traversing the graph, with the delay PDF of the entire circuit also being calculated at the end of the traversal. The advantage of this method is that it requires less computation time than the path-based method because more than one path can be analyzed simultaneously. However, the correlations between paths must be considered for the statistical max operation when multiple paths converge at a node. Therefore, there are trade-offs between accuracy and computation time.

4.3 Applying SSTA to Combinational and Sequential Logic

The advantage of applying SSTA to combinational and sequential logic is that, I can predict the timing yield with SSTA method. I can consider the trade-off between circuit performance and timing yield during the design phase. To accurately predict the timing yield, SSTA must analyze the entire circuit. Also, the analysis must be completed within a practical amount of time. For these reasons, the SSTA

for combinational and sequential design uses the block based SSTA method. The SSTA can statistically handle die to die variations as well as within die random variations, so the timing yield can be predicted more accurately. Figure 8 shows the SSTA flow to predict the timing yield. In this flow we can predict the within the die random variations of the gates and interconnects that are inputs of the SSTA tool. In timing sigh-off, all paths must satisfy their timing constraints. Therefore, we uses a path-based algorithm that can accurately calculate path delay.

We propose to use Dijkstra's algorithm to decide operating frequency of the circuit. The net list of logic gates will be the input to this algorithm and the longest path delay will the output of algorithm. The algorithm may use either path base or block base method to find probability distribution of delay.

5. Modeling Parameters

5.1 Components of Variations

In general, the intra-chip parametric variation δ can be decomposed into three parts: a deterministic global component δ_{global}, a deterministic local component $\delta_{local}$and a random component ϵ .

$$\delta = \delta_{global} + \delta_{local} + \epsilon \qquad (1)$$

The global component δ_{global}, is location-dependent. For example, across the die, it can be modeled by a slanted plane and expressed as a simple function of the die location:

$$\delta_{global}(x,y) = \delta_0 + \delta_x x + \delta_y y \qquad (2)$$

where (x,y) is its die location, δ_x and δ_y are the location-dependent gradients of parameter indicating the spatial variations of parameter along the x and y direction respectively.

The local component, $\delta_{local,}$ is proximity-dependent and layout-specific. The random components, ϵ, stands for the random intra-chip variation and is modeled as a random variable with a multivariate normal distribution to account for the spatial correlation of the intra-chip variation**.**

5.2 Spatial Correlations

To model the intra-die spatial correlations of parameters, the die region may be partitioned into n rows x n columns = n grids. Since devices or wires close to each other are more likely to have similar characteristics than those placed far away, it is reasonable to assume perfect correlations among the devices [wires] in the same grid, high correlations among those in close grids and low or zero correlations in far-away grids. For example, in gates 'a' and 'b' (whose sizes are shown to be too large) are located in the same grid square, and it is assumed that their parameter variations (such as the variations of their gate length) , are always identical. Gates 'a' and 'c' lie in neighboring grids, and their parameter variations are not identical but are highly correlated due to their spatial proximity. For example, when gate 'a' has a larger than nominal gate length, it is highly probable that gate 'c' will have a larger than nominal gate length, and less probable that it will have a smaller than nominal gate length. On the other hand, gates 'a' and 'd' are far away from each other, their parameters are uncorrelated; for example, when gate 'a' has a larger than nominal gate length, the gate length for 'd' may be either larger or smaller than nominal.

Under this model, a parameter variation in a single grid at location *(x, y)* can be modeled using a single random variable *p(x,y).* For each type of parameter, n random variables are needed, each representing the value of a parameter in one of the n grids.

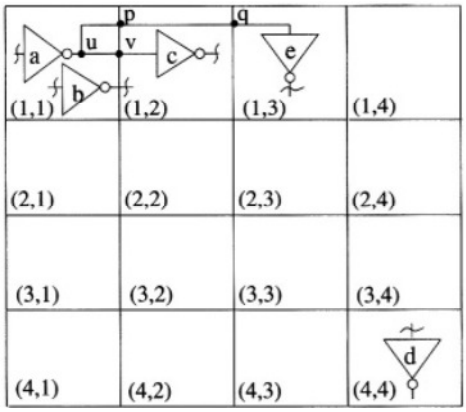

Figure 6. Grid model for spatial correlations

In addition, it is reasonable to assume that correlation exists only among the same type of parameters in different grids and there is no correlation between different types of parameters. For example, the L_g values for transistors in a grid are correlated with those in nearby grids, but are uncorrelated with other parameters such as T_{ox} or W_{int} in any grid. For each type of parameter, an correlation matrix, $\sum$, represents the spatial correlations of such a structure.

An alternative model for spatial correlations was proposed in [3][4]. The chip area is divided into several regions using multiple quad tree partitioning, where at *l* level the die area is partitioned into 2^l x 2^l squares; therefore, the uppermost level has just one region, while the lowermost level for a quad-tree of depth k has 4^k regions. A three-

level tree is illustrated in Figure 4. An independent random variable, $\Delta p_{i,r}$is associated with each region (i,r) to represent the variations in parameter p in the region at level r. The total variation at the lowest level is then taken to be the sum of the variations of all squares that cover a region.

For example, in Figure 4, in region (2,1), it represents the effective gate length due to intra-die variations, $\Delta L_{eff}(2,1)$, then

$$\Delta Leff(2,1) = \Delta L0,1 + \Delta L1,1 + \Delta\, L2,1 \qquad (3)$$

In general, for region (i,j),

$$\Delta p(i,j) = \sum_{[0<l<k,(l,r)\ \text{covers}\ (i,j)]} \Delta Plr \quad (4)$$

5.3 Structural Correlations

The structure of the circuit can also correlate in SSTA. Consider the circuit shown in Figure 5 the circuit has two paths, a-b-d and a-c-d. If, for example, we assume that each gate delay is a Gaussian random variable, then the Probability Density Function (PDF) of the delay of each path is easy to compute, since it is the sum of Gaussians,

Which admits a closed form? However, the circuit delay is the maximum of the delays of these two paths, and these are correlated since the delays of 'a' and 'd' contribute to both paths. It is important to take such structural correlations, which arise due to re-convergences in the circuit, into account while performing SSTA.

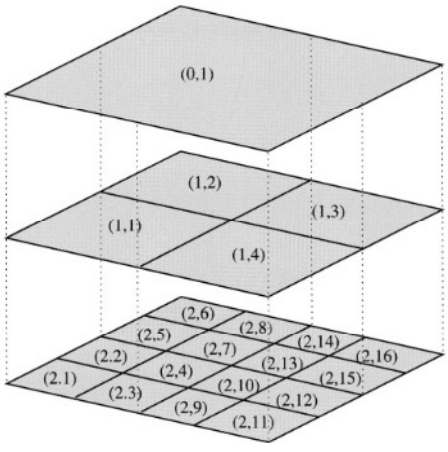

Figure 7: The quad tree model for spatially correlated variations

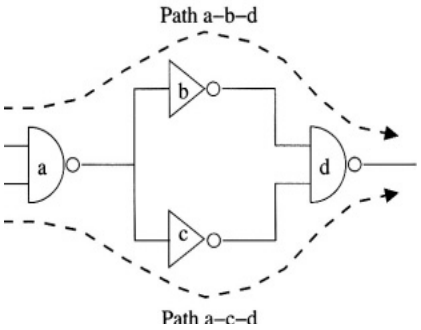

Figure 8: An example to illustrate structural correlations in a circuit.

5.4 Computing the PDF of Gate Delay

For a multiple-input gate, the pin-to-pin delay of the gate differs at different input pins. Let d_{gate}^{pini} be the delay of the gate from the i[th] input to the output. In general d_{gate}^{pini} can be written as a function of the process parameters ***P***of the gate, the loading capacitance of the driving interconnect tree C_w and the succeeding gates that it drives *Cg,* and the input signal transition time S_{in} at this input pin of the gate:

$$d_{gate}^{pini} = d(P, C_w, C_g, S_{in}) \quad (5)$$

The sensitivities of the gate delay to the process parameters can be found applying the chain rule for computing derivatives.

Since the gate delay d_{gate}^{pini}differs at the different input pins, in conventional static timing analysis, S_{out} is set to d_{gate}^{pini} if the path ending at the output of the gate traversing the i[th] input pin has the longest path delay. In statistical static timing analysis, each of the paths through different gate input pins has a certain probability to be the longest path. Therefore, S_{out} should be computed as a weighted sum of the distributions of the gate delays d_{gate}^{pini} where the weight equals the probability that the path through i[th] the pin is the longest among all others:

$$S_{out} = \sum_{\text{input pins i}} \{\text{Prob}[d_{pathi} > \max(d_{pathj})] \times d_{gate}^{pini}\} \quad (6)$$

where $d_{pathi}$is the distribution of path delay at the gate output through the i^{th} input pin. The calculation of d_{path} and max (d_{pathj}) can be achieved by the "sum" and the "max" operators. It is clearly to see that S_{out} is approximated as a normal distribution, since it is as a weighted sum of normal distributions d_{gate}^{pini}. Using the formulation above, the derivatives of S_{out} to the process parameters can also be

computed through the weighted sum of the derivatives of d_{gate}^{pini} to the process parameters.

6. Conclusion

Statistical Static Timing Analysis is better analysis than the traditional Static Timing Analysis. It is seen that the proposed method is easy to determine the propagation delay for the complex digital circuits.

References

[1] Izumi Nitta, Toshiyuki Shibuyu,Katsumi Homma. Statistical Static Timing Analysis Technology. FUJITSU Sci. Tech. J. 43,4,p.516-523 October 2007.

[2] Bhaghath P J, Ramesh S R. A Survey of SSTA Techniques with Focus on Accuracy and Speed. International Journal of Computer Applications (0975 – 8887) Volume 89 – No.7, March 2014.

[3] Masanor Hashito, Hidetoshi Onodera. A Performance Optimization Method by Gate Resizing Based On Statistical Static Timing Analysis. IEICE Trans. Fundamentals, VOL E83-A, No. 12, P2558-2568 December 2000.

[4] Takashi ENAMI, Shinyu NINOMIYA, et. Al: Statistical Timing Analysis Considering Clock Jitter and Skew due to Power Supply Noise and Process Variation. IEICE TRANS. FUNDAMENTALS, VOL. E93-A NO. 12 December 2010.

[5] Y. Liu, S. R. Nassif, L. T. Pileggi, and A. J. Strojwas. Impact of interconnect variations on the clock skew of a gigahertz microprocessor. In Proceedings of the ACM/IEEE Design Automation Conference, pages 168–171, June 2000.