On the Need for Statistical Timing Analysis

Farid N. Najm
ECE Dept., University of Toronto, Ontario, Canada
f.najm@utoronto.ca

ABSTRACT

Traditional corner analysis fails to guarantee a target yield for a given performance metric. However, recently proposed solutions, in the form of statistical timing analysis, which work by propagating delay distributions, do not conform to modern design methodology. Instead, new statistical techniques are needed to modify corner analysis in ways that overcome its weaknesses without violating usage models of timing tools in modern flows.

Categories and Subject Descriptors:
B.7 [**Integrated Circuits**]: Design Aids

General Terms: Design, Algorithms

Keywords: Variability, Statistical timing analysis

1. INTRODUCTION

Manufacturing *process variations* cause electrical *parameter variations* (such as transistor or wire parameters), leading to variations in certain chip performance metrics, such as the maximum operating frequency (FMAX) and the chip power dissipation. If these *performance variations* cause a particular chip to violate some *performance constraint*, then that chip is considered failed.

The underlying sources of the variations (in the process) are numerous and not always well characterized. Thus, there is not a universal agreement, for instance, that the variations can always be modeled as random variables. If they are modeled as random, there is not a universal agreement as to what their distribution may be. Nevertheless, there is a significant body of literature based on the assumption that the variations are random and that they can be modeled by normally distributed (Gaussian) random variables (RVs). This paper will also be based on this assumption.

For a circuit or chip designer, the challenge is to ensure that the unavoidable electrical parameter variations do not lead to excessive performance variations. It is useful to think of two *spaces*: the *parameter space*, which is typically multi-dimensional corresponding to the potentially large number of electrical parameters on modern chips, and the *performance space*, which also can be multi-dimensional (chip frequency, power dissipation, etc.). In the performance space, a number of performance constraints determine the *acceptability region*. A chip is considered failed if the performance variations put its performance metrics outside the *acceptability region*.

If the underlying process variations are modeled as RVs, then the electrical parameter variations are also RVs, and the performance variations are RVs as well. In the performance space, the *probability* that the performance metrics fall in the acceptability region gives the *yield*. Naturally, this yield is only the performance yield related to the set of variations under study, and there can be several other reasons for yield loss in an IC.

Let $X_1, X_2, \ldots, X_n$ be *independent* zero-mean RVs that represent the electrical parameter variations. The independence assumption is a simplification; in practice, if one is dealing with electrical parameters that are not independent, one can use techniques like *principal components analysis* to express them in terms of certain independent factors. Let Z be an RV that represents some performance metric, say chip frequency, so that:

$$Z = g(X_1, X_2, \ldots, X_n) \tag{1}$$

where $g(\cdot)$ is some function which in general can be non-linear. If $\mathcal{A}$ is the acceptability region in the performance space (the space of Z), then the yield is given by the probability: $\mathcal{Y} = \mathcal{P}\{Z \in \mathcal{A}\}$. When Z is single-dimensional, as in this case, then $\mathcal{A}$ is typically an interval on the Z axis, and the yield can be expressed as:

$$\mathcal{Y} = F_z(\mathcal{A}_{max}) - F_z(\mathcal{A}_{min}) \tag{2}$$

where $F_z(\cdot)$ is the cumulative distribution function (cdf) of Z and $\mathcal{A} = [\mathcal{A}_{min}, \mathcal{A}_{max}]$. When two performance metrics Z_1 and Z_2 are under study, then $\mathcal{A}$ is typically a rectangle in the (Z_1, Z_2) plane. For simplicity, and without loss of generality, we will focus on the single dimensional case.

2. CORNER ANALYSIS

In practice, one rarely has enough information to be able to construct the distribution of Z with any certainty during the design phase of an IC. Instead, and in order to ensure a good yield, a traditional approach is to apply *corner analysis*, as follows.

For a specific electrical parameter variation X_i, say transistor channel length, it is common to assume certain bounds on its magnitude, X_i^{min} and X_i^{max}. Transistors whose length falls within these bounds are considered normal and viable. Transistors outside these bounds are considered to be the result of serious deviations of the process; chips with such large variations would typically fail due to any number of reasons, not necessarily related to a performance metric. If X_i^* is a variable that can take a value of either X_i^{min} or X_i^{max}, then the vector $[X_1^*, X_2^*, \ldots, X_n^*]$ represents a *corner* in the parameter space. If, during chip design, one finds that the circuit meets the performance constraints for all such corners, *i.e.*, for all assignments $X_i = X_i^*$, then the design is deemed acceptable. This is corner analysis.

Straightforward application of corner analysis can be time consuming and expensive, because the number of corners is exponential in the number of electrical parameters, n. However, the situation is often simplified through some knowledge of how certain parameters affect circuit speed. Thus, if Z is circuit delay, it is known that Z_{min} may be obtained by setting transistor length to a minimum, rather than a maximum. Through considerations of this type, corner analysis has been applied for many years to design chips that are robust in the face of process variability.

When variations are modeled as RVs, it is common to think of the bounds X_i^{max} and X_i^{min} as being the $\pm 3\sigma$ limits of the normal distribution of X_i. In other words, if σ_i^2 is the variance of X_i then $X_i^{max} = +3\sigma_i$ and $X_i^{min} = -3\sigma_i$. For *any* normal distribution with mean μ and variance σ^2, the interval $\mu \pm 3\sigma$ covers 99.73% of the distribution:

$$\Phi_{\mu,\sigma}(\mu + 3\sigma) - \Phi_{\mu,\sigma}(\mu - 3\sigma) = \Phi(3) - \Phi(-3) = 0.99731 \tag{3}$$

where $\Phi_{\mu,\sigma}(\cdot)$ is the cdf of the normal distribution with mean μ and variance σ^2, and $\Phi(\cdot)$ is the cdf of the *standard normal distribution* (*i.e.*, the normal distribution with a mean of 0 and a variance of 1). Thus, for all practical purposes, almost all of the distribution lies within the $\pm 3\sigma$ limits. In many cases, engineers consider the 99.73% value (what one may refer to as the 3σ yield) to be a desirable yield that one should target.

DAC 2005, June 13–17, 2005, Anaheim, California, USA.
ACM 1-59593-058-2/05/0006.

Let Z_{max} and Z_{min} be the largest and smallest values of Z that are observed (say, by using a circuit simulator) upon checking all the corners. If $Z_{max} \leq \mathcal{A}_{max}$ and $Z_{min} \geq \mathcal{A}_{min}$, then the design is acceptable. Typically, design margins are very tight, so that one is often dealing with $Z_{max} \approx \mathcal{A}_{max}$ and $Z_{min} \approx \mathcal{A}_{min}$, and one may write:

$$\mathcal{Y} \approx F_z(Z_{max}) - F_z(Z_{min}) \tag{4}$$

In any case, the yield is at least as large as the value given by (4)[1]. A word of caution: knowing that the corners are defined by the 3σ limits of the X_i's is *not enough* to conclude that Z_{max} and Z_{min} are the 3σ limits of Z. There is more to it than that, as we will see below, so that (4) does *not* necessarily give a high yield value of 99.73%. Thus, when the limits on X_i are interpreted as the 3σ limits, a *failing* of corner analysis is that we do not know what performance yield is being guaranteed.

3. STATISTICAL ANALYSIS

Recently, with the increasing variability in the manufacturing process, traditional corner analysis has been viewed as inadequate, and statistical analysis has been proposed as a replacement approach. Specifically *statistical static timing analysis* (SSTA) has been proposed as an alternative to traditional static timing analysis (STA). However, SSTA represents a major overhaul of the design flow, and there is no universal agreement yet as to exactly what SSTA is, and whether it is actually needed.

Most SSTA proposals involve propagating distributions of delay through the logic network. Specifically, with the signal arrival time viewed as an RV, one propagates the pdf (probability density function) of that RV (typically a Gaussian) through the logic network. Path-based SSTA does this for a given path, while block-based SSTA generates the pdf for the maximum delay of the block. What is often ignored (or not spelled out) is exactly what the user should then do with these pdf's. Ideally, one would perhaps "chop off" that pdf at a point corresponding to the desired timing yield, and thereby determine the timing margin available for that path or block. However, if the pdf is for a single path (or for a single block), there is no way for the user to know what yield is desired for that path (or block), without knowledge of all the circuitry in the rest of the chip (the root cause for this fact is the presence of within-die variations). This effectively means that a block or a path cannot be "timed" in isolation. One needs to first design the whole chip, and run the SSTA on the whole chip, before knowing whether the chosen sizing for transistors on a given path are acceptable or not! This is contrary to the way static timing analysis is used today, where individual paths or blocks are "timed" in isolation (perhaps by different people), and iteratively improved until the stringent timing specs are met.

Designers expect that SSTA would somehow magically allow them to "deal with variability." However, tool developers seem to be working on tools that simply propagate delay pdf's. There is a disconnect between the two communities! A different approach to SSTA is required.

Perhaps the search for a better SSTA can be guided by the weaknesses of existing corner analysis. However, upon examination, these weaknesses don't seem insurmountable. There may be ways to develop statistical techniques, other than the proposed SSTA approaches, for solving the variability problem by extensions of classical corner analysis. To see this, consider that some commonly perceived weaknesses of corner analysis are as follows:

1. There are too many corners

Due to the increasing number of electrical parameters whose variability must be considered, the number of corners has gradually increased over the years. However, although the number of variables may be large, many practitioners are content to focus on only the key transistor parameters, such as channel length and threshold voltage, in order to study the impact on timing, which significantly reduces the number of corners that one should look at. In any case, there have been proposals made [1] for merging corners (corner clustering) and reducing the number of corners, and more work along these lines can be worthwhile.

2. Corner analysis cannot take care of within-die variations

This is true; corner analysis sets the value of a variable, at all its instances and across the whole die, to a specific extreme value. However, not everyone agrees that the impact of within-die variations is very big. Indeed, in [2], it is shown that the impact of within-die variations is small and manageable. In any case, there may be ways of factoring in the effect of within-die variations as part of a process by which new *virtual corners* are identified. This was illustrated in [3], where a technique is proposed by which a "factor of sigma" term, δ, is found (to replace the factor of "3" in "3σ"). The new corner value at $\delta\sigma$ is obtained by taking both die-to-die and within-die variations into account.

3. Corner analysis is overkill

This objection is usually raised because a corner becomes a very improbable occurrence in case of a large number of variables. While it is true that the probability of a corner decreases as the number of variables increases, this is not, however, a valid objection to corner analysis. Indeed, whether corner analysis is overkill or not depends, not on the corner probabilities, but on whether the implicit yield target, given by (4), is too large or not. It is reasonable perhaps to consider a good target value for this yield to be:

$$\mathcal{Y}_0 = \Phi(3) - \Phi(-3) = 99.73\% \tag{5}$$

If much larger than this (*i.e.*, if much closer to 1), then the yield may be called too large and the analysis an overkill. If lower than this, then the analysis is definitely not overkill. We will illustrate that this depends on the nature of the function $g(\cdot)$ in (1). For the case when $g(\cdot)$ is the simple sum function, $Z = \sum_{i=1}^{n} X_i$, then $Z_{max} = 3n\sigma$ and $Z_{min} = -3n\sigma$, and, because Z is normal with zero mean and variance $n\sigma^2$, it follows that:

$$\mathcal{Y} = \Phi\left(3\sqrt{n}\right) - \Phi\left(-3\sqrt{n}\right) \tag{6}$$

which indeed becomes increasingly larger than $\mathcal{Y}_0$ (closer to 1), for large n, so that the analysis is indeed overkill. However, consider the case when the function $g(\cdot)$ is the $\max(\cdot)$ function, so that:

$$Z = \max_{i=1,2,\ldots,n}(X_i). \tag{7}$$

In this case, the cdf of Z is given by $\Phi^n(z/\sigma)$, and, with $Z_{max} = 3\sigma$ and $Z_{min} = -3\sigma$, it follows that:

$$\mathcal{Y} = \Phi^n(3) - \Phi^n(-3) \tag{8}$$

which is gradually less than $\mathcal{Y}_0$ for larger n. Thus, for the max function, the analysis is definitely not overkill. Since circuit delay is a mix of both the sum and the max functions, it becomes impossible to make a blanket statement as to whether corner analysis is overkill or not. In this regard, the approach in [3] provides a way by which the corner location, δ, is chosen to give a desired yield, thus bypassing the "overkill" objection.

4. CONCLUSION

We find that it is unclear at this point whether an overhaul of the timing verification flow, as in many proposals for statistical timing analysis, is warranted or not. It should be possible to develop other kinds of statistical approaches to overcome the weaknesses and failings of traditional corner analysis.

5. REFERENCES

[1] M. Sengupta, et al. Application specific worst case corners using response surfaces and statistical models. In *IEEE International Symposium on Quality Electronic Design*, pages 351–356, San Jose, CA, March 22-24 2004.

[2] S. B. Samaan. The impact of device parameter variations on the frequency and performance of VLSI chips. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 343–346, San Jose, CA, November 7-11 2004.

[3] F. N. Najm and N. Menezes. Statistical timing analysis based on a timing yield model. In *ACM/IEEE 41st Design Automation Conference*, pages 460–465, San Diego, CA, June 7-11 2004.

[1] In case of multiple performance metrics $Z_1, Z_2, \ldots$, this approximation is not necessarily good and all we can say about (4), in that case, is that the yield is at least that much.