From Blind Certainty to Informed Uncertainty

Kurt Keutzer and Michael Orshansky
University of California, Berkeley

"All good things were at one time bad things; every original sin has developed into an original virtue."
Friedrich Nietzsche (1844–1900), The Genealogy of Morals, aph. 9 (1887).

ABSTRACT

The accuracy, computational efficiency, and reliability of static timing analysis have made it the workhorse for verifying the timing of synchronous digital integrated circuits for more than a decade. In this paper we charge that the traditional deterministic approach to analyzing the timing of circuits is significantly undermining its accuracy and may even challenge its reliability. We argue that computation of the static timing of a circuit requires a dramatic rethinking in order to continue serving its role as an enabler of high-performance designs. More fundamentally we believe that for circuits to be reliably designed the underlying probabilistic effects must be brought to the forefront of design and no longer hidden under conservative approximations. The reasons that justify such a radical transition are presented together with directions for solutions.

Categories and Subject Descriptors

J.6.1 **[Computer-Aided Engineerin**g]: Computer-aided design

General Terms: Algorithms

1. INTRODUCTION

During the design of an IC, and certainly before its manufacture, we wish to ensure that all sampled memory elements of a circuit have the proper logical value at the end of each clock cycle. Verifying this property first requires that under no circumstances does the computation of the correct logical value require longer than the clock cycle of a device. Secondly it requires that the latching of the correct logical value be not pre-empted by any rapid propagation of the results of the previous clock cycle. For more than a decade the industrial workhorse for verifying these two timing properties of synchronous digital integrated circuits has been static-timing analysis (STA). The accuracy, computational efficiency, and reliability of STA have made it the natural choice. A static approach eliminates the need of vector-set generation and more importantly provides *vector-independent* worst-case modeling. This approach obviates the concern that a key critical path sensitizing vector sequence may have been overlooked. Computing the longest and shortest paths without considering functionality is naturally linear time, and is therefore computationally efficient. After decades of research false-path eliminating function-dependent timing [6][22][23] has also been brought within affordable computational limits. Most importantly, STA has been reliable. With solid models of cells and wiring delay models fortified by back-annotated capacitances, STA has been able to provide reliable, i.e. consistently conservative, bounds on the delay of a circuit.

Each of the traditional strengths of STA also exposes a potential weakness. As Friedrich Nietzsche, shrewdly observes in *Beyond Good and Evil:* "One is punished most for one's virtues." Reliability has enforced conservatism, and the requirement of conservatism has led to worst-case delay estimation using an elaborate series of worst-case assumptions in delay modeling, delay calculation, and path traversal. Such a worst-case approach to timing is beginning to fail for several reasons. The major ones are that: (1) conservatism is sacrificing too much performance; (2) the amount of conservatism increases; and (3) under certain conditions "worst-case" timing may in fact be not conservative so that reliability is jeopardized.

The goals of accuracy and reliability are naturally at odds. Particularly in ASIC design, conservatism is aimed at ensuring that *no* manufactured circuit should exceed its clock period. The worst-case timing assumption that comes as a result requires that the entire delay variation due to processing get passed into the delay budget of each circuit. With the higher performance requirements that accompany the trend from ASIC to application-specific standard parts (ASSP) we believe that accuracy must not be entirely sacrificed in an attempt to achieve reliability through naïve conservatism.

This paper aims to do more than simply point out deficiencies of current timing approaches and propose a more accurate, statistical, approach to verifying timing. It may be natural to think of a circuit as a deterministic computing engine whose macro-behavior exhibits statistical variation due to processing effects [11]. This paper argues that there are so many and so diverse probabilistic effects now operating in circuit behavior that it is time to rethink our entire notion of the behavior of an integrated circuit and move from viewing an IC as a deterministic computing device to a non-deterministic one.

2. WHY WE NEED PROBABILISTIC STA NOW

In this section we explore the factors that drive the need in a shift from the deterministic treatment of circuit timing properties to an essentially probabilistic treatment.

2.1 Demise of ASICs and Higher Performance Demands on Chips

The standard ASIC design methodology is pessimistic in its approach to modeling timing. At every step in the design flow, the modeling and computational assumptions are made with the exclusive goal of guaranteeing the correctness of design. As a result, the conservative safety margin is accumulated. The traditional STA therefore fulfills its role as a cornerstone of such a design paradigm.

In ASIC design worst-case modeling is used in such a way that only function is presumed to require manufacture testing.

TAU'02, December 2–3, 2002, Monterey, California, USA.
Copyright 2002 ACM 1-58113-526-2/02/0012…$5.00.

In particular, expensive at-speed testing of every chip is not required. This is significant in the ASIC market where competition is largely based on cost-efficiency. It is clear that this approach has as it primary goal the functional correctness and cost-efficiency of testing rather than high performance of the final product, since a lot of performance is wasted being hidden under the worst-case assumptions. These days it is common to hear ASIC designers complain of weeks spent working to achieve timing closure only to discover that the manufactured part runs significantly faster than STA indicated [5].

We claim, however, that the era when such a paradigm could be sustained is coming to an end. The driving force behind the transformation is the demise of ASIC business, as we know it. Semiconductor market data [24] indicates that the number of manufactured ASICs (unique parts, not aggregate volume) has been steadily declining since 1995. ASSP designs have remained about constant over the same period. Thus in 2002 the nearly 7000 IC designs are approximately broken down between 70% ASSPs and 30% ASICs. This trend is expected to continue to a 85%/15% breakdown in five years. In short, ASSPs are steadily replacing ASICs in electronic systems. ASSPs must compete with other ASSPs for defined market niches. As a result, circuit speed and power constraints are higher for ASSPs than ASICs. Going forward, circuit designers, can no longer afford to lose performance due to inaccurate worst-case modeling assumptions.

2.2 Increased Parametric Variability and Randomization of Circuit Behavior

STA will need to become increasingly conservative due the growth of process variability in key device and interconnect parameters [15]. The growing extent of variation of these parameters can be attributed to several factors. The first is the difficulty of tightening process control margins as technology scales. The second is the rise of multiple systematic sources of parameter variability, e.g. optical proximity effects and the inter-layer dielectric thickness [16]. The most profound reason for the future increase in parametric variability is that the technology is approaching the regime of fundamental randomness in the behavior of silicon structures. The best example of that is the essentially random value of the transistor threshold voltages due to the processing phenomena at atomic scale. As a result, many other key circuit-level properties, such as delay and power, will exhibit random variation [1]. All these reasons lead to the greater relative variability of device parameters around the nominal value.

The patterns of variability are also changing: Specifically, the contribution to variability of the intra-chip (within chip) variation component is rapidly growing. By the time of 0.07um CMOS technology, the percentage of total variation of effective channel length L_{eff} that is accounted for by intra-chip variation will grow to 65% [9]. Similarly, intra-chip variation of the interconnect geometries is rising in relation to the total variation budget. The increase of intra-chip parameter variation is caused by the emergence of a number of variation-generating mechanisms located on the interface between the design and process. For example, variation of the effective channel length can be largely attributed to the optical proximity effect in which the length of the polysilicon line becomes dependent on the local layout surroundings of an individual transistor [16].

2.3 Non-Probabilistic STA Is Unacceptably Conservative

All the early methods of statistical circuit analysis made two key assumptions: (1) intra-chip parameter variability within the chip is negligible compared to inter-chip variability, and (2) all types of digital circuits, e.g. all cells and blocks, behave statistically similarly (i.e. in a correlated manner) in response to parameter variations.

For a long time these assumptions were valid. However, as we argued above, the intra-chip component of variation is steadily growing. The second assumption is also not holding up well, as many practicing engineers know. Specifically, different cells have different sensitivities to process variations [12]. This is especially true of cells with different aspect ratios, tapering, and cells designed within multi-threshold and multi-V_{DD} design methodologies. It is also clear that interconnect and gate delays are not sensitive to the same parameters, and with the increasing portion of delay being attributed to interconnect, this will have to be taken into account.

The breakdown of these two assumptions leads to significant intra-chip variation of gate and wire delays. In the presence of large intra-chip delay variation, the standard timing analysis is bound to result in an unreasonable level of conservatism. Indeed, the worst-case static timing analysis proceeds by setting each gate to its worst-case timing value, and performing a longest path computation to arrive at the worst-case critical path delay. The assumption that is implicit in this approach is that delay elements (gates and wires) are perfectly correlated with each other. The failure to consider the validity of the assumptions makes the probability of finding a chip, with characteristics assigned to it by the worst-case timing analysis, very small, leading to lost performance and expensive over-design.

If we perform a more accurate analysis of path delay distribution, we find that clock period (binning) distribution is modified compared to the standard timing approaches. A sizable reduction of conservatism is possible with statistical analysis. Being able to exactly predict the shape of the clock speed distribution would allow significant reduction of the timing conservatism of the standard timing methodology, improving circuit performance and reducing over-design.

Having an accurate prediction of the speed binning distribution is also critical for a different reason. Apart from reducing conservatism of the standard timing methodology, another way to improve circuit performance of ICs developed in a standard ``ASIC'' flow in comparison to custom chips is to allow them to trade parametric yield for performance. It has been noted that an ASIC chip produced in the foundry may run up to 40% faster than predicted by standard timing analysis [5]. Trading parametric yield for speed would require implementing a full speed testing procedure and enable testing ASICs at a specifiable speed, which would be chosen to satisfy some set of yield performance constraints. ASIC vendors will trade yield for performance if revenue from faster chips will justify the additional expense in lost yield and testing overhead.

2.4 Non-Probabilistic STA Is Unreliable

Static timing analysis cannot be divorced from the underlying modeling methodology. The modeling and characterization methodology that is typically used with gate-level STA tools, such as PrimeTime™ [19], cannot guarantee that the resulting timing estimates are accurate. In that methodology, delay tables are generated for every cell in the library. Delay tables are functions of such factors as output load, input slew-rate, and intrinsic gate-delay. In order to capture the impact of process, voltage and temperature variation the nominal simulation is repeated using a SPICE model that contains the corresponding points in the PVT space. The problem is that because of non-linear delay response and capacitive feedback coupling due to Miller capacitance, the worst-case behavior of a circuit path does not always correspond to the combination of worst-case points of individual circuit elements [21].

To address this difficulty by non-statistical means, one would have to find a set of process parameters that would always lead to the worst-case circuit behavior. The fundamental problem is that due to complex patterns of coupling between gates and interconnects finding such a point in the space of process parameters is extremely difficult. As a result, timing estimates produced by the worst-case methodology may very well not be conservative enough.

2.5 Non-Probabilistic STA Is Not Sufficient

Some other trends also make the standard STA approaches insufficient. One of them is the importance of delay faults, such as parametric delay faults due to random particles, which are very hard to capture by deterministic means. These delay faults arise from singular events that can only be described by statistical means.

Particles are pieces of dust that land onto the chip during its manufacturing in the "clean room". Traditionally, random particles were taken into account within the framework of yield analysis: if a particle did hit the chip, the chip was presumed to be damaged beyond the possibility of further operation, e.g. catastrophically. In such a model, the probability of a particle hitting a chip was equal to the probability of a chip being catastrophically damaged. This analysis was subsumed under the rubric of "random yield" estimation. Within the CAD community, such faults were treated as "stuck-at" faults. Because these faults were considered 'catastrophic' events, there was no need to perform circuit timing analysis in the presence of such faults. Circuit timing was carried out from the assumption that no fault is present. Their only conceptual role was to serve as a model to be used for testability.

In reality the "stuck-at" fault model fail to properly describe the behavior of realistic faults due to resistive contacts, unexpected coupling, or resistive bridges. Most often such "soft" faults do not irreparably damage the chip. Rather, because of them, the affected devices and interconnects become marginally slower, which leads to longer delays for some paths and the violation of longest path constraints. In effect, these faults behave as parametric variations of device and interconnect properties. Because the number of such faults is likely to be bigger in scaled technologies, and because the sensitivity of paths to such faults will increase, we need to explicitly take them into account in the timing analysis.

The standard STA tools are not able to model such faults, since it would lead to the assumption that the worst-case event did happen, i.e. that the random particle did hit the chip, and, that would have to be done for every possible fault. Clearly, this would result in an unhelpful explosion of conservatism. Alternatively, statistical STA could incorporate the modeling of such faults very naturally.

3. COMPREHENSIVE PROBABILISTIC TREATMENT

In the previous section we discussed the reasons why it is necessary to treat static timing analysis as a probabilistic problem. While this proposal sounds conceptually innocent, it brings up a whole set of fundamental challenges to the way we traditionally think of circuit design and analysis. In this section we attempt to face these bigger issues and to develop a consistent probabilistic world-view as it applies to static timing analysis.

3.1 A taxonomy of uncertainties

It is natural to think about uncertainty as something to be quarantined and eliminated. However, since this is impossible we must face it directly. Uncertainty refers simply to something we cannot describe deterministically, precisely. This may be due to the fundamental randomness of the phenomena (e.g. the atomic-scale effects) or due to our inability or difficulty of modeling the phenomenon by deterministic means even if it can theoretically be so described (e.g. local capacitive coupling). We also implicitly assume that our only option in describing uncertain information is by probabilistic means.

The first step then is to consider what kinds of uncertain information we have to deal with in circuit timing analysis. First of all, we may distinguish *static or physical* uncertainty from *dynamic or environmental* uncertainty. *Static* uncertainty is uncertainty about the values of device and wire parameters as they are affected by process variation. After chip manufacturing has been finished, every individual chip may, in principle, be given a well-defined description, because its properties have been fixed by processing. This is different from *dynamic* uncertainty, exemplified by the uncertainty due to capacitive coupling (cross-talk), 2-D temperature and voltage distribution, and conservative delay modeling (e.g. due to unknown input arrival times). These distributions change dynamically, depending on the functional performance of the chip. Dynamic uncertainty stems from the lack of information about the exact patterns of switching events and signal interactions.

Second, we may distinguish *fundamental* uncertainty from *computational* uncertainty. Fundamental uncertainty refers to truly random behavior, which by definition cannot be described deterministically. For example, as devices reach atomic scale, transistor threshold voltages can only be described probabilistically because their values are defined by the placement of atoms in the channel, which is indeed a truly random process. Computational uncertainty is related to the practical limitations of our knowledge, to the things we could theoretically know but which cannot be practically made determinate due to the complexity of the modeling or computational structure of the problem. For example, it is theoretically possible to describe cross-talk behavior by an exhaustive analysis of all the switching transitions, but that is computationally impractical because requires an enormous computational effort.

It is worthwhile noting that probabilistic description is to be accepted only when deterministic description is impossible. This is not because probabilistic description is in any way epistemologically suspicious. The simple reason is that given a choice, deterministic description is to be preferred only because potentially it provides a more precise description of phenomena and allows exploiting them more fully. (Note that the worst-case model is, in a sense, deterministic but it is not more accurate than a probabilistic model.) While the idea of being able to describe chip's properties by probabilistic means only is unusual, a close look at the existing practice shows that fundamentally this treatment is de facto already happening. What's left is to explicitly acknowledge what is already happening. For example, chip reliability at the device level has been treated probabilistically for a long time. It is a known fact that the thin oxide film that forms the basis for transistor functionality will fail at some point in the future. We can only say that the expected (average) time until chip's failure is given by a certain quantity.

3.2 Integrating Different Sources of Uncertainty in One Conceptual Framework

A move to statistical timing analysis would mean that we treat many key chip properties as random values to be described by probabilistic distributions. A natural extension of this approach is to provide a unified treatment of the various sources of delay uncertainty in predicting chip's timing behavior. In the taxonomy of the previous section, such a treatment would combine the analysis of static and dynamic uncertainty. To a limited extent this has already been happening: for example, process variation is combined with temperature and voltage variation to generate the "worst-case" timing models. Our proposal goes further and also assumes the unified probabilistic treatment of dynamic uncertainty due to cross-talk, input-pattern dependent gate delay models.

Such a move, however, is quite consequential. If we design a chip using probabilistic treatment of static uncertainty only, then after manufacturing, timing properties of each chip are uniquely defined and we can find out these properties (e.g. the maximum clock cycle) by testing each manufactured chip. Once testing is performed, the "good" remaining chips are guaranteed to function properly. (The only additional requirement from point of view of the design flow is the requirement of doing complete at-speed testing.) On the contrary, treating dynamic uncertainty probabilistically means that even after manufacturing there is uncertainty about the timing behavior of each chip. All we can do is to provide a distribution of path delays, and say, for example, that this chip has a probability of 1 in a million to exceed the clock period. Then, the designer can decide which particular signal deserves more accurate analysis. While there is the above difference in the design implications of probabilistic treatment of static and dynamic uncertainty we may still find it beneficial to treat them within a unified conceptual framework. This is because dynamic uncertainty often stems from computational uncertainty, from our practical inability to describe, say cross-talk, by precise means.

4. SOME IMPLICATIONS OF PROBABILISTIC COMPUTATION

In this section we will consider in more detail several implementation issues related to probabilistic timing computation. Traditionally, statistical timing analysis has been concerned with delay variation due to process, voltage, and temperature (PVT). As we argued in section 4.2, it is appealing to incorporate the analysis of other sources of uncertainty into the single probabilistic framework. While many sources of modeling conservatism can potentially be treated probabilistically, here we discuss two specific mechanisms: capacitive coupling and conservative gate delay modeling.

In very deep sub-micron technologies, capacitive coupling, or cross-talk, strongly affects delay and integrity of the signals sent though long interconnect lines. As the neighboring lines switch they couple to each other leading to timing deterioration. Such active coupling may result in both increased and decreased delay, depending on the direction of the switching signal. It is thus key to include the modeling of such coupling effects into STA so that the longest or shortest delay computation remains conservative. While it is possible to come up with the worst-case estimate for the active coupling capacitance, such estimates severely limit the design space because of their conservatism. The use of switching windows helps to reduce the conservatism by identifying the time windows in which the coupling may happen [4]. However, such analysis is computationally prohibitive, since to be complete it would require an exhaustive analysis of logical signal dependencies and physical proximities of the adjacent wires.

Faced with the computational intractability of the deterministic cross-talk analysis, it appears natural to propose to treat cross-talk probabilistically. The first, to our knowledge, probabilistic account of cross-talk was given by Vrudhula [20] who proposed a model to estimate the likelihood of the capacitive coupling event. His model is based on a discrete probabilistic model of the noise event on a particular net happening. An alternative probabilistic model could be based on a continuous model of the distribution of an aggressor noise level. However, both models can lead to the same key estimation procedure: finding the probability density function (*pdf)* of the path delay.

Another significant source of modeling conservatism that has always formed the basis of STA is the assumption that only one input signal is switching at the same time [7]. This model once again assumes the worst-case scenario for gate switching time. However, if several inputs are switching simultaneously then this model becomes extremely conservative. In many estimates the conservatism is as high as 50% [5]. The probabilistic analysis of gate delay would assign a certain probability to the event in which multiple signals are switching simultaneously, and would find the distribution of gate propagation delays.

5. CONCLUSIONS

The displacement of an increasing percentage of ASIC designs by ASSPs is causing IC designers as a whole to seek higher performance. At the same time, the migration into smaller process geometries is bringing forth new significant variational effects and is changing the relative impact of existing variational effects. Attempts to sweep the impact of all these effects under a single rug of worst-case timing modeling will lead to unacceptable levels of timing conservativism at precisely the time that designers are seeking higher performance.

This paper argues that the time has come to consider a bold alternative: to bring the variational effects to the forefront and consider them directly. In particular we argue for a rethinking of static-timing analysis and argue for a statistical approach to determine the timing of digital synchronous circuits.

6. REFERENCES

[1] Burnett, D. et al., "Implications of Fundamental Threshold Voltage Variations for High -Density SRAM and Logic circuits," *Proc. of Symposium on VLSI,* pp. 15-16, 1994.

[2] Chandrakasan, A., Sheng, S., and Brodersen, R., "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuit*s, Vol. 27, No. 4, pp. 473-484, April 1992.

[3] Chang, E. et al., "Using a Statistical Metrology Framework to Identify Systematic and Random Sources of Die- and Wafer-level ILD Thickness Variation in CMP Processes," *Proc. of IEDM,* pp. 499-502, 1995.

[4] Chen, P., D. Kirkpatrick, K. Keutzer, "Switching Window Computation for Static Timing Analysis in Presence of Crosstalk Noise," *Proc. of ICCAD,* pp. 331-337, 2000.

[5] Chinnery, D., and Keutzer, K., Closing the Gap Between ASICs and Custom, Kluwer Academic Publishers, 2002.

[6] Devadas, S., Keutzer K., Malik S., “Delay Computation in Combinational Logic Circuits: Theory and Algorithms,” *Proc. of ICCAD*, pp. 176-179, November 1991.

[7] Hassoun S., Sasao T., editors, Logic Synthesis and Verification, Kluwer Academic Publishers, 2002.

[8] Kahng, A., and Pati, Y., “Subwavelength optical lithography: challenges and impact on physical design,” *Proceedings of ISPD*, pp. 112-115, 1999.

[9] Nassif, S., “Delay Variability: Sources, Impact and Trends,” *Proc. of ISSCC*, pp. 368-369, 2000.

[10] Nassif, S., “Within-chip variability analysis*,” IEDM Technical Digest*, pp. 283-286, 1998.

[11] Gattiker, A. et al, “Timing yield estimation from static timing analysis,**”** *Proc. of ISQED*, pp. 437-442, 2001.

[12] Nassif, S., “Statistical worst-case analysis for integrated circuits,” Statistical Approaches to VLSI, Elsevier Science, 1994.

[13] Orshansky, M., Spanos, C., and Hu, C., “Circuit Performance Variability Decomposition”, *Proc. of IEEE International Workshop on Statistical Metrology for VLSI Design and Fabrication*, p.10-13, Kyoto, Japan, 1999.

[14] Orshansky, M., and Keutzer, K., “A Probabilistic Framework for Worst Case Timing Analysis,” *Proc. of DAC,* pp. 556-561, 2002.

[15] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 2001.

[16] Stine, B. et al., “Inter- and intra-die polysilicon critical dimension variation,” *Proc. of SPIE,* pp. 27-35, 1996.

[17] Sylvester, D. “BACPAC: Berkeley Advanced Chip Performance Calculator”, 1999.

[18] Sylvester, D., Keutzer, K., “Getting to the bottom of deep-submicron,” *Proc. of ICCAD,* pp. 203-211, 1998.

[19] PrimeTime ™, (a trademark of Synopsys), 2002.

[20] Vrudhula, S., Blaauw, D., Sirichotiyakul, S., “Estimation of the Likelihood of Capacitive Coupling Noise,” *Proc. of DAC*, pp. 653-658, 2002.

[21] Zanella, S. et al, "Statistical Timing Macromodeling of Digital IP Libraries", *Proc. of Fifth International Workshop on Statistical Metrology*, pp. 76-79, 2000.

[22] Devadas, S. et al, “Computation of floating mode delay in combinational logic circuits: Theory and algorithms,” *IEEE Transactions on Computer-Aided Design*, 12(12): 1913-1923, December 1993.

[23] Devadas, S. et al, “Computation of floating mode delay in combinational logic circuits: Practice and implementation,” *IEEE Transactions on Computer-Aided Design*, 12(12):1924-1936, December 1993.

[24] Handel Jones, International Business Systems, personal communication.