cadence®

Modeling Stress-Induced Variability Optimizes IC Timing Performance

By Nishath Verghese, Ramez Nachman, Philippe Hurat

Intentionally induced stress techniques can provide great benefits to design teams working on advanced integrated circuit (IC) process nodes, improving drive current by up to 20 percent. However, to take full advantage of these benefits design teams must model the variability induced by these techniques. Today, many foundries include proximity effects in their circuit models and rely upon electronic design automation (EDA) tools. Cadence offers a comprehensive solution for modeling and managing stress-induced variability.

Contents

Introduction

Many design teams migrate to advanced IC process nodes to increase performance while reducing area and power. Timing performance and predictability can be compromised, however, if there's too much systematic variability. Fortunately, systematic variability can be modeled and mitigated if one understands the causes. A leading cause of systematic variability at 45 nm and below is the application of mechanical stress to transistors.

Stress is commonly used to enhance performance in CMOS ICs, but it also causes variability that can make it difficult to close timing. At 45 nm and below, in fact, stress-induced variability can result in unexpected timing variations of 15 percent or more. Designers considering advanced nodes therefore need to be aware of the causes and impacts of stress, and understand approaches for modeling and mitigating stress-induced variability.

The variability of stress is the problem – not stress itself. In fact, stress is, for the most part, intentionally induced through various techniques that change the lattice spacing inside transistor channels. These techniques increase electron mobility or hole mobility (absence of electrons) and thus boost performance. But stress is also unintentionally and non-uniformly applied due to the constraints of the layout, and is unintentionally induced through various technologies such as shallow trench isolation (STI), a widely used technique that uses oxide to isolate transistors. Thus, IC design teams still need to model stress even if a silicon process doesn't intentionally induce stress.

However, evaluating stress and its variability is not straightforward, because the evaluation must consider "proximity effects." This means that designers can't just look at transistors in isolation—the location and dimensions of neighboring layout features change the surrounding stress, and therefore the timing performance. At 45 nm and below, for example, varying STI width (also called active-to-active spacing) impacts stress. This width can be known only if the location and dimensions of nearby features are taken into account.

Today, most IC design teams handle stress-induced variability by simply applying guardbands. If design teams model stress-induced variability, however, they can increase device performance, reduce excessive guardbands, and avoid unexpected parametric silicon failures. To help with this task, EDA software needs to comprehend stress effects on timing, and offer capabilities that help optimize placement in order to mitigate variability.

This whitepaper discusses sources of transistor stress, the effects of stress, and stress modeling and mitigation for library development and post-route analysis. It shows how IC design teams can model stress-induced variability and reduce the need for guardbands that might negate the performance advantages that stress can bring to advanced process technologies.

Sources of Stress

Intentionally induced stress

Intentional stress techniques apply tensile (pulling) stress to NMOS devices, or compressive (pushing) stress to PMOS devices. Because tensile stress increases the electron mobility and NMOS current flow is determined by electron flow, increasing electron mobility improves NMOS performance. In contrast, compressive stress makes holes move faster. Since PMOS current flow is determined by hole flow, compressive stress improves PMOS timing performance.

Applying tensile stress to NMOS devices, and compressive stress to PMOS devices, allows silicon process engineers to boost drive current, and therefore improve chip performance. Note, however, that threshold voltage might also change with stress, which is one more reason for modeling in order to predict stress effects on timing.

Two commonly used techniques for intentionally inducing stress are:

- Strained silicon through SiGE stress layer
- Stress liner applied through SiN capping layer

Strained silicon techniques apply a silicon germanium (SiGE) stress layer under n-channels to induce tensile stress, or embed the layer in the source and drain to induce compressive stress in p-channels. Today, embedding SiGE in the source and drain is easier to do and is more common than placing it under n-channels. SiGE has a larger lattice dimension than silicon. Therefore, when SiGE is embedded in silicon, the mismatch between the crystal dimensions causes compressive stress. Because diffusion can increase compressive stress, SiGE modeling must consider the amount of diffusion around transistors.

Stress liner techniques place a "capping layer," usually a silicon nitride (SiN) film, over transistors. A dual capping layer is often applied, with one layer over NMOS transistors and one layer over PMOS transistors. The recipe of each capping layer is such that when placed on top of PMOS devices, the layer is compressive, and when placed on top of NMOS devices, the layer is tensile. With single stress liner (SSL), the layer is either all compressive or all tensile. With dual stress liner (DSL), the layer is compressive for PMOS devices and tensile for NMOS devices (Figure 1).

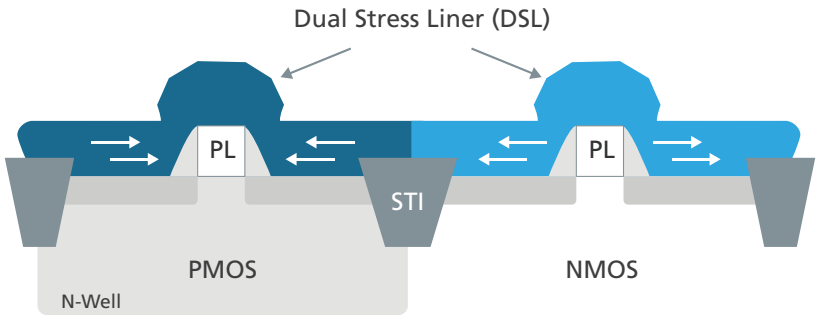

Figure 1: Dual stress liner (DSL) applies a SiN film in order to create tensile stress on NMOS devices, and compressive stress on PMOS devices.

The boundary between tensile and compressive regions is usually derived from the location of the n-well layer. As a result, the distance between a transistor and the n-well boundary has stress implications and should be considered in modeling. Stress can change with the distance from the boundary. Thus, DSL modeling must take proximity effects into account.

Since placing SiGE under n-channels is not easy, an alternative is to apply a SiGE stress layer to p-channels only, and to use SSL or DSL to apply tensile stress to the n-channels. This can be done by applying either the tensile SSL or tensile/compressive DSL layer over the entire wafer, and then etching the film away from the p-channel transistor area.

Polysilicon pitches (transistor gate pitches) disturb the stress liners when contacts punch through it, and thus lower the amount of stress provided by the liner. When contacts and poly are closer to a gate, they will have more influence. This is another factor to consider when modeling stress (Figure 2).

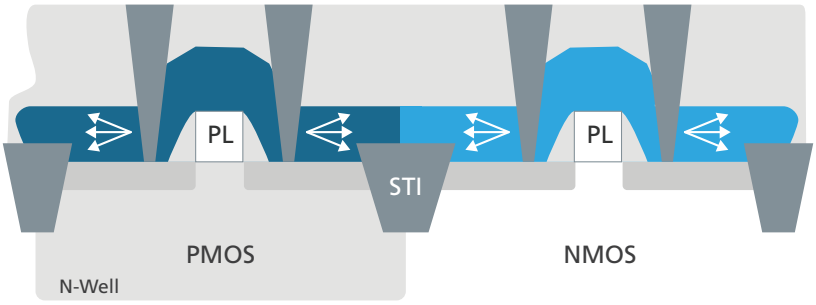

Figure 2: Contacts break the continuity of a DSL, reducing the amount of stress.

Even though techniques such as SiGE, SSL, and DSL induce variability, there is a strong motivation to use them, since they can improve drive current by as much as 20%.

Unintentional stress – shallow trench isolation

STI uses silicon dioxide (SiO2) to isolate transistors and separate NMOS and PMOS regions. This has been the standard device isolation scheme since the 250 nm process node. STI results in SiO2 abutting against silicon, and the resulting stress generally becomes compressive after processing, due to the thermal expansion mismatch between Si and SiO2. The width and thickness of the STI determines how much oxide is next to transistors, and consequently how much stress is applied.

At 65 nm and above, designers cared only about the distance between a transistor and the nearest STI edge. At 45 nm and below, the width of the STI channel has become an important concern, given that STI-induced stress can change IC drive current (for better or worse) by as much as 10 percent. Thickness does result in stress, but it's a fixed amount per die with no proximity effects. STI width – or active-to-active spacing – is dependent on the placement of transistors, and thus has a significant proximity effect.

When transistors are laid out very close, the amount of STI is small. When transistors are far apart, the amount of STI is larger. Since it's compressive, wider STI will increase PMOS performance and degrade NMOS performance, while narrower STI will accomplish the opposite.

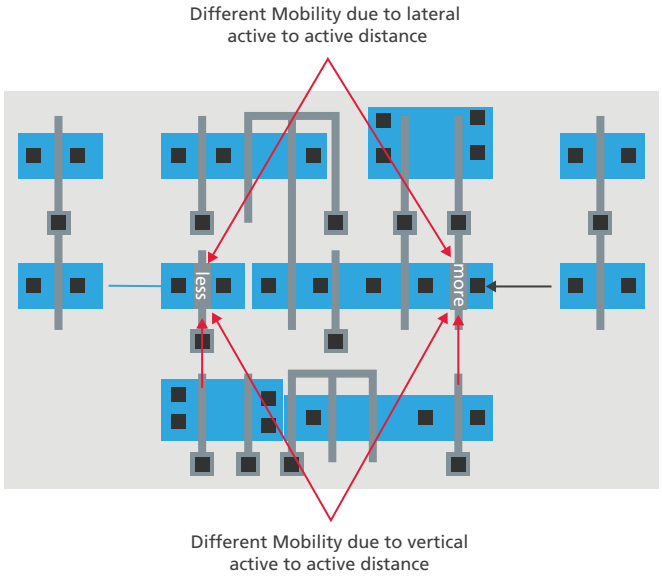

Figure 3: STI width is determined by active-to-active spacing. Greater distances will result in wider STI and more compressive stress.

An issue that's closely related to STI is length of diffusion (LOD). This is the distance from an n-channel or p-channel to the STI oxide edge. As LOD grows shorter, more STI stress is placed on transistor gates. From a standard cell point of view, LOD is not necessarily context dependent (dependent on proximity effects), because it usually does not extend beyond the cell boundary, and its effect is already accounted for when the cell is characterized. However, STI width is context dependent because it depends on the distances of devices in neighboring cells, as shown in figure 3.

BSIM4 models for SPICE simulation provide two LOD parameters. These are sa, which specifies distance to the left of the channel, and sb, which specifies distance to the right of the channel. Standard BSIM4 models do not, however, model STI width, which is a more significant concern at 45 nm and below.

Well proximity effect

The well proximity effect (WPE) results from the location of well boundaries with respect to transistors. WPE is not a stress effect, but it does impact mobility and threshold voltage, and it is a proximity effect that impacts the stress liner. As previously noted, stress liner boundaries usually coincide with well boundaries. Therefore, WPE is typically modeled along with stress.

Impacts of Stress

Transistor stress affects the transistor mobility, saturation velocity, and threshold voltage (Vt). This in turn affects performance and timing. Designers typically care most about variability in timing. The problem with modeling stress-induced timing variations is that identical standard cells may have different timing characteristics due to proximity effects. Most design tools assume that a given standard cell will always have the same timing, regardless of where it is placed in the layout.

In modeling stress, the most important consideration is capturing changes in current. The chart below shows a decrease in NMOS and PMOS saturation current (Isat) of up to 25 percent with closer horizontal well boundaries due to DSL. This could translate to as much as a 25 percent timing difference depending on the type of cell that is implemented.

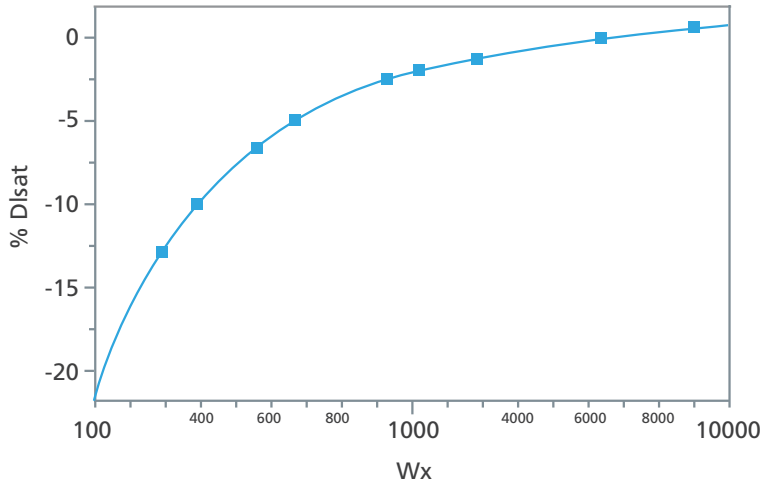

Figure 4a: Isat degrades with closer horizontal well boundaries (DSL). Wx is the distance in nm.

The following chart shows how NMOS and PMOS Isat degrades with closer surrounding poly. Here we see a decline of over 10 percent as poly spacing drops below 250 nm.

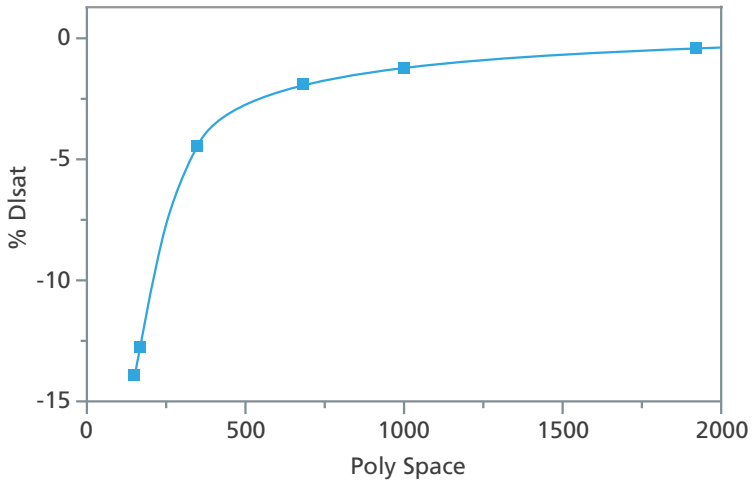

Figure 4b: Isat decreases as poly space shrinks.

The following chart provides an example of NMOS Isat degradation with closer vertical wells, and shows how the impact increases with smaller device widths. Taken together, these three charts show how dramatic proximity effects can be.

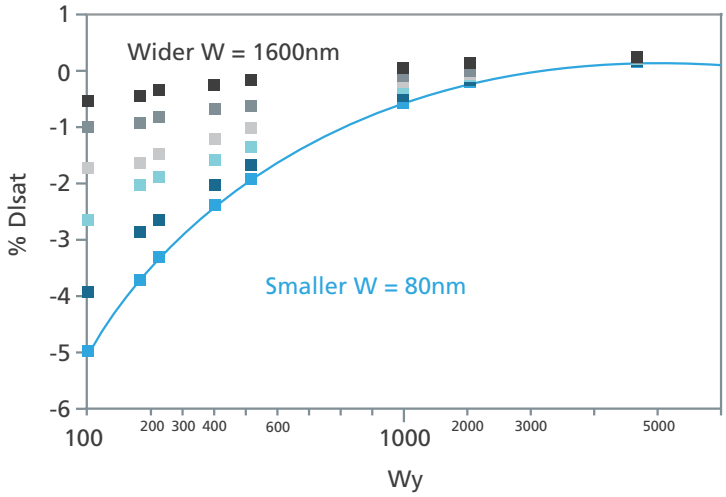

Figure 4c: With smaller device widths, closer vertical wells have a more profound impact on Isat.

Modeling and Mitigating Stress-Induced Variability

At advanced process nodes, library developers and IC designers need to understand stress-induced variability and its mitigation. This section describes stress modeling and mitigation in library design, post-route analysis, and analog custom design.

Standard cell library development

As previously noted, BSIM4 models represent LOD parameters but do not typically include STI width or other proximity effects. Thus, some foundries and IDMs are creating customized versions of BSIM models that include proximity effects such as poly spacing.

Developers of standard cell libraries need to run transistor-level simulations of stress effects, using foundry rule decks, extracted stress parameters from layout, and SPICE models. However, if one takes a standard cell and places it in a different layout context, the simulation results may be different. Ideally, designers would run many circuit simulations in many different layout contexts to figure out the possible extent of variability, and to architect the layout to minimize unwanted variability.

In practice, running so many circuit simulations becomes very difficult. In the absence of a better solution, library design teams usually try to identify a "typical" context, which is then used for the timing characterization of the entire library. This methodology has multiple difficulties. Designers need to guarantee that the chosen "typical" context is actually typical for the entire library, find the optimal context for each cell, extend the analysis to worst-case and best-case contexts, measure context variability, and validate characterization assumptions. Also, the layout team cannot run easily trade off analyses between area, delay and variability because of the lack of an easy way to measure variability. This can result in a sub-optimal layout and increased variability in design.

The Cadence® Litho Electrical Analyzer can predict electrical variation from lithography and stress effects, taking layout context into account. It is used for both library design and post-route analysis and optimization. For example, the tool provides an environment that can analyze a standard-cell library for variability over a number of different layout contexts, and provide metrics for library designers to take corrective action. The Litho Electrical Analyzer can then feed information to placement and routing tools in order to mitigate placement-induced variability.

To analyze proximity effects, the Litho Electrical Analyzer can automatically generate random layout contexts, or extract contexts from a given design. The tool internally runs the Cadence Litho Physical Analyzer to generate length/width (L/W) variations, and uses the Cadence Physical Verification and Cadence QRC Extraction tools to generate stress parameter variations. As shown in Figure 5, a traditional extraction approach would only consider L/W, whereas a Litho Electrical Analyzer extraction includes mobility (U0), saturation velocity (Vsat), and threshold voltage (Vt) parameters.

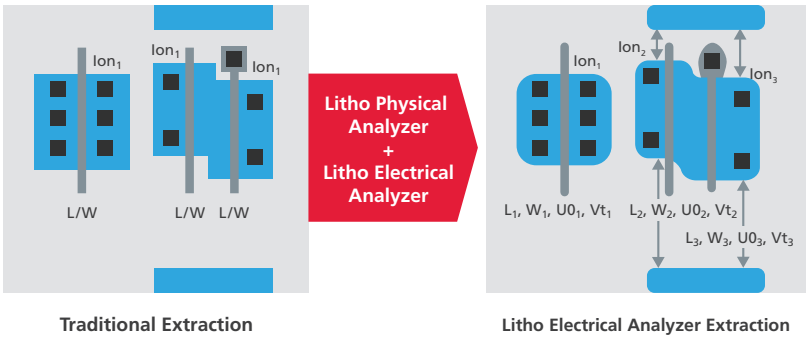

Figure 5: The Litho Electrical Analyzer extraction includes mobility and voltage threshold.

Inputs to the Litho Electrical Analyzer include a library, Litho Physical Analyzer technology file, and LVS rule decks (Figure 6). The Litho Electrical Analyzer runs Spice simulations to get timing data for each possible delay arc, and leakage power for each input combination. Once cell characterization is completed, the Litho Electrical Analyzer reports context variability statistics, including litho hotspots, gate variations, delay variations, and leakage variations. It can also dump out a modified Spice netlist for standalone simulation.

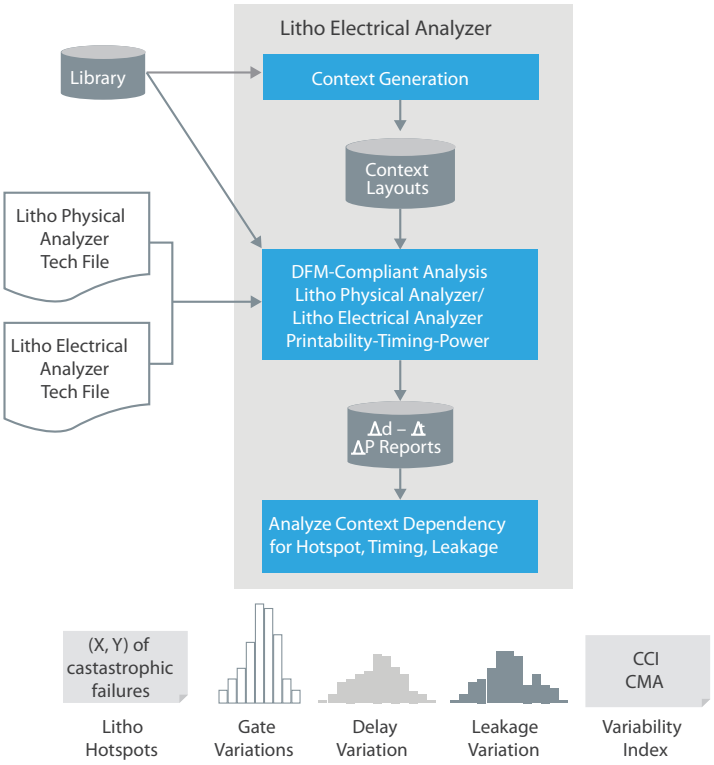

Figure 6: Library variability analysis using the Litho Electrical Analyzer.

The Litho Electrical Analyzer offers many benefits for optimizing libraries, which includes allowing users to identify typical, best-case, and worst-case contexts that should be used for accurate characterization; quantify context variability for each cell; perform quantified area/timing variability architectural and layout tradeoffs; and prioritize various layout optimizations.

The Litho Electrical Analyzer can also generate a Cell Context Index (CCI) file that combines litho and stress effects into variability "scores." This information can drive placement optimization in the Cadence Encounter® Digital Implementation System, which can swap cells in the layout to mitigate litho-induced and stress-induced variability.

The Litho Electrical Analyzer also produces a Cell Margin Adjustment (CMA) file for derating standard cells during static timing analysis to reduce excessive design margins. Timing analysis typically uses an on-chip variation (OCV) margin to account for random or non-modeled systematic variations, such as context-induced variations. With Litho Electrical Analyzer, users can evaluate the sensitivity of each cell of the library, and the OCV can be adjusted to reflect this sensitivity. A cell with a large sensitivity should have an increased OCV, while a very robust cell should use a smaller OCV.

Post-layout analysis

In addition to library variability analysis, the Litho Electrical Analyzer can be used prior to tapeout to run a post-route litho and stress variability analysis. Chip designers can use the Cadence Encounter Timing System to run static timing analysis with cell-based derating factors from the CMA file. They can also analyze critical paths that might be especially sensitive to proximity effects, and optimize those critical paths to reduce sensitivity to variations.

The diagram below (Figure 7) shows two flows supported by the Litho Electrical Analyzer. The first flow, labeled "1", is used by library designers to analyze cells and create views required for placement and routing of these cells. The second flow, labeled "2", is used by chip designers post-layout to analyze critical paths and optimize them. In this second flow, the tool generates an incremental Standard Delay Format (SDF) file that can be fed to the Encounter Digital Implementation System for post-route optimization.

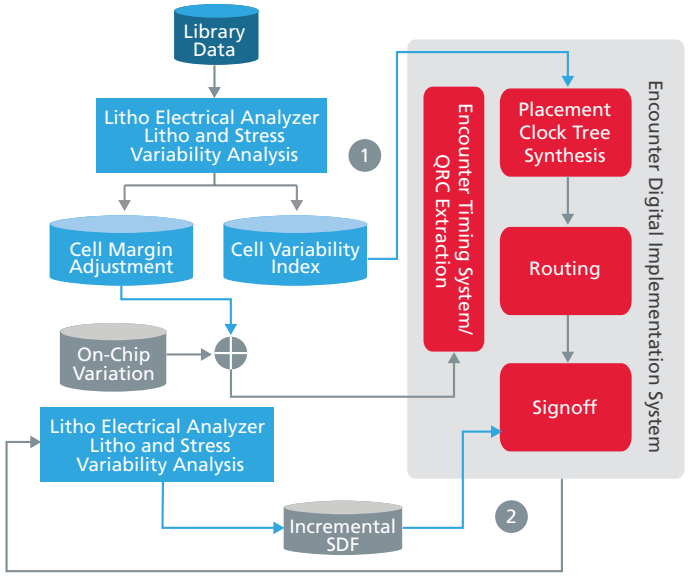

Figure 7: The Cadence Litho Electrical Analyzer supports pre-route and post-route variability-aware implementation.

Analog custom design

In analog custom design, the layout should be as close to the schematic as possible. But with the advent of stress techniques, the impact of layout proximity on transistor performance is considerable. Consequently, it is easy to end up with post-layout simulation mismatches between the layout and the schematic, resulting in long iteration loops and delayed time-to-market.

Today, if an analog designer finds that the layout is diverging from the schematic, his or her only recourse is to extract stress parameters and run a full simulation. Instead, designers need an analysis that is readily available during the creation of the layout, but does not require a complete layout and a full simulation. A commercial solution in this area is not yet available.

Conclusion

Intentionally induced stress techniques such as SiGE and DSL can improve drive current by as much as 20 percent. To fully take advantage of that benefit, design teams need to model the variability these techniques induce. Even without these intentional techniques, STI applies stress and can cause significant timing variations. Because STI stress is compressive, which helps PMOS performance but degrades NMOS performance, it could have a positive or negative effect on chip timing. These effects are traditionally ignored during timing analysis.

The most common practice for dealing with stress-induced variability today is applying guardbands. That works for most design teams at 65 nm and above, but at 45 nm and below, stress-induced timing variation is becoming more pronounced. It can already be as high as 15 percent at 45 nm. Applying excessive margins to contain this variability will degrade the very performance advantages that stress techniques were designed to bring.

Predicting stress-induced variability without modeling is hard because the variability depends on layout proximity effects. The variability of a standard cell will differ according to what's located near it. Standard characterization approaches, which assume that a cell will behave identically no matter where it's placed in the layout, don't work. And standard BSIM4 models don't consider proximity effects.

Today, many foundries and IDMs are including proximity effects in circuit models. EDA tools can be helpful as well. The Cadence Litho Electrical Analyzer, for example, can predict timing variations based on litho and stress effects. It is used both for library variability analysis and post-route analysis and optimization.

The first step in managing stress-induced variability is an awareness of the problem. The next step is to work with foundries and EDA suppliers to find solutions. With the Litho Electrical Analyzer and Encounter Digital Implementation System, Cadence offers a comprehensive set of solutions for modeling and managing stress-induced variability.