Predictive Models and CAD Methodology for Pattern Dependent Variability

Nishath Verghese, Richard Rouse and Philippe Hurat
Cadence Design Systems
555 River Oaks Parkway
San Jose, CA 95134

Abstract:

Lithography, etch and stress are dominant effects impacting the functionality and performance of designs at 65nm and below. This paper discusses pattern dependent variability caused by these effects and discusses a model-based approach to extracting this variability. A methodology to gauge the extent of this pattern dependent variability for standard cells is presented by looking at the difference in transistor parameters when the cell is analyze in different contexts. A full-chip methodology that addresses the delay change due to systematic varation has been introduced to analyze and repair a 65nm digital design.

I. Introduction

As designs migrate to 65nm process technology and below, the impact of systematic manufacturing effects on design functionality and performance is increasing substantially. Manufacturing design layout features of sub-wavelength feature sizes requires a continual increase in the use of optical proximity correction (OPC) and resolution enhancement techniques (RET). Manufacturers today use sophisticated simulation software and model-based methods to perform OPC and RET in mask-making. However, designers of chip layout are still using rules (to mimic these models) provided to them by manufacturers - for chip routing, DRC, layout parameter extraction (LPE) and parasitic extraction (RCX). As geometries shrink, the impact of neighboring layout patterns on the printability and electrical characterisitics of a drawn shape can no longer be captured effectively using rules. Models of printability and performance impact on transistors and wires due to layout patterns must be made available to designers. "Model-based DFM" tools provide such capabilities to chip designers. Using models calibrated to manufacturing processes, these tools allow designers to predict the systematic impact of pattern-dependent variability on the functionality and performance of their chips. They can be used in conjunction with existing routers, extractors and timing tools to provide insight into the impact of manufacturing effects like lithography, etch and stress. This paper presents model-based design analysis and optimization methodology for pattern-dependent variability.

II. Pattern-Dependent Litho/Etch Variability

In modern projection lithography, the numerical aperture (NA) of the scanner optical system is increasing and is above 1 in the case of immersion lithography. Simultaneously, the k1 factor (normalized line width k1=CD/(λ/NA), where CD is minimum feature size, λ the wavelength) is steadily decreasing from generation to generation. As a result of smaller k1 (k1~0.3 at 45nm node), pattern fidelity detiorates resulting in an increased number of "hotspots" or critical locations in layout that have a critically small process window even though layouts may be design rule compliant. Some of these hotspots can lead to catastrophic failures, such as predictable opens and shorts while others can be marginal, impacting yield due to process variations. Even with more aggressive OPC/RET, the number of hotspots tends to increase as technology scales [1].

Likewise, the increased variability from 65nm to 45nm leads to stronger manufacturability impact. There are larger intra-pattern and pattern-to-pattern variations, especially for immature litho, etch processes and OPC. Several fundamental litho choices impact pattern fidelity and variability – immersion/dry; lithography settings and illumination techniques; selected design rules; RET/OPC aggressiveness, CD control on mask, etc. For instance, gate CD variability can increase ~1.5-2 times from a regular pattern (with restricted design rules) to random/variable pattern (flexible design rules) [1]. It's interesting to note that, according to ITRS, CD control < 4nm (3σ) has NO known solutions today [2]. This will lead to increasing but predictable variability in transistor current depending on poly and active patterns.

Fabless designers who jointly work with the fab are increasingly required to perform early analysis of CD variability. This "DFM" information must be fed back to the design house for design optimization. DFM requirements are getting more complex due to the non-linearity and increased complexity of process-to-design and design-to-process interactions. They must therefore be addressed with models and simulation.

978-1-4244-1922-7/08/$25.00 ©2008 IEEE

III. Strained Silicon and Pattern-Dependent Stress

Increasingly, in nanometer processes, strained silicon is used to increase the switching speeds of transistors by enhancing the carrier (electron/hole) mobility. This occurs due to the lowering of the effective mass of the carrier under strain:

$$\mu = \frac{q\tau}{m^*} \qquad (1)$$

where $1/\tau$ is the scattering rate and m^* is the carrier effective mass. Mobility directly modifies the carrier velocity, υ according to the applied electric field, E as:

$$\upsilon = \mu E \qquad (2)$$

Since carrier velocity is directly proprtional to the transistor switching speed, mobility directly influences it. In the nanometer regime, biaxial stress has been the conventional method to strain the transistor channel. A widely adopted method to introduce wafer-based biaxial stress to enhance CMOS performance is practiced by growing a silicon film atop relaxed SiGe virtual substrate. For standard layout and wafer orientations, NMOS performance (electron mobility) is enhanced by uniaxial longitudinal tensile stress while PMOS performance (hole mobility) is enhanced by uniaxial longitudinal compressive stress. Interestingly, both NMOS and PMOS performance are enhanced by biaxial tensile stress [3]. However, due to recent studies that have revealed its advantages uniaxial process strained silicon is being adopted in nearly all high-performance logic technologies. A predominant method of introducing uniaxial longitudinal stress is by deposition of CVD silicon nitride film on the device structure. This enhances electron mobility, thereby improving NMOS performance. A compressive silicon nitride capping layer can generate uniaxial longitudinal compressive stress in the PMOS channel enhancing hole mobility. Such a dual stress liner process architecture with tensile and compressive silicon nitride layers over NMOS and PMPOS respectively is illustrated in Fig 5. Alternatively, p-channel device performance is enhanced by using selective SiGe layer as source/drain regions.

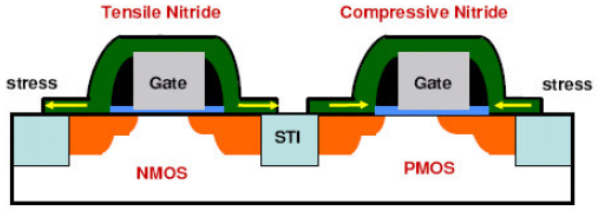


Figure 1: Introduction of uniaxial longitudinal stress in CMOS devices. [4]

Note that as the STI width is modulated, the transverse tensile channel stress is also modulated. Higher STI width results in a larger amount of STI to pull the channel as shown in Figure 2. This change in applied transverse stress can be brought about by varying diffusion topologies (north/south) and this will cause a direct change to the mobility of the transistor in consideration. Likewise, diffusion to the left and right of a given transistor will change its longitudinal stress.

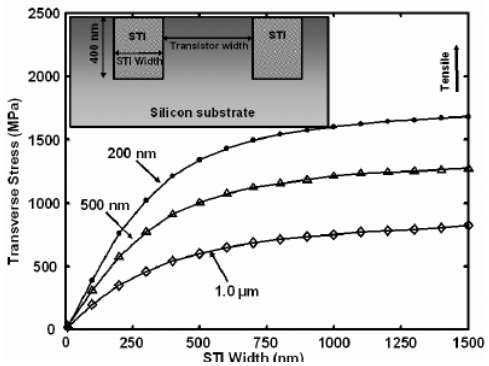


Figure 2: Effect of varying STI width on channel stress for different transistor widths [3].

Figure 3 shows simulations of the lateral compressive stress in the channel as a function of poly pitch and film stress, where the poly pitch effect is enhanced slightly as the film stress increases. This stress reduction is a result of poly gate disrupting the stressor vector in the lateral direction [5]. Figure 4 contains experimental data showing the poly pitch effect as the PMOS Idsat-Ioff improvement decreases as poly pitch reduces. Results for both a contact etch stop layer (cESL) and a dual Etch Stop Layer (dESL) integration scheme (where both tensile and compressive films are used for simultaneous NMOS and PMOS improvement) are shown. It is clear that stress impact of neighboring poly gates significantly exacerbates the poly pitch related variability that are attributed to lithography and etch.

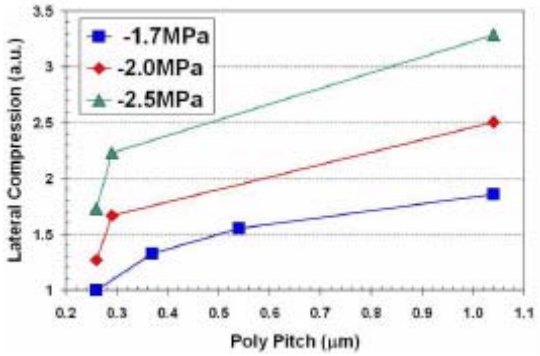


Figure 3: Mechanical stress simulations for lateral compressive stress in the PMOS channel versus gate poly pitch as a function of film stress [5].

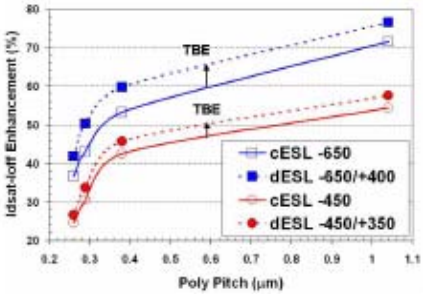


Figure 4: PMOS Idsat-Ioff enhancement versus gate poly pitch and as a function of channel stress and single cESL or dESL integration [5]

IV. Predictive Litho/Etch Simulation

To bring these systematic process effects in design, model-based solutions are needed. To address litho effects, a predictive litho and etch simulation methodology utilizes a fast, accurate and secure model, which captures the entire RET/OPC manufacturing flow, including retargeting, assist-feature insertion, PSM, OPC and etch information released by the designer-specific target fab. From a simulator using this secure model, silicon contours for shapes drawn on diffusion, poly, and metal layers are determined as shown in Fig 5 [6]. The secure model characterizes process performance and predicts the effects of RET, OPC, litho, etch and mask effects on design shapes without having to run the RET/OPC flow. Designers can apply this model to their design databases to accurately predict the silicon contours across the process window, for shapes drawn on diffusion, poly and metal layers and detect printability hotspots such as bridging (short), pinching (open), contact overlap, etc.

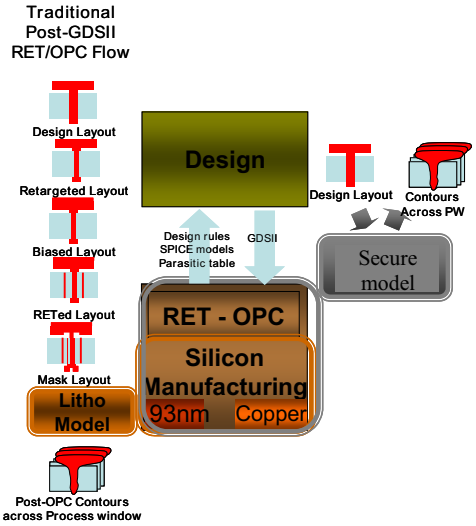


Figure 5. The secure model shown on the right encapsulates the entire fab mask-making flow shown on the left, to predict silicon contours from drawn layout by simulation.

The model is fast enough to enable simulation across the process window of a full-chip database in hours [6], and therefore is usable in a design environment.

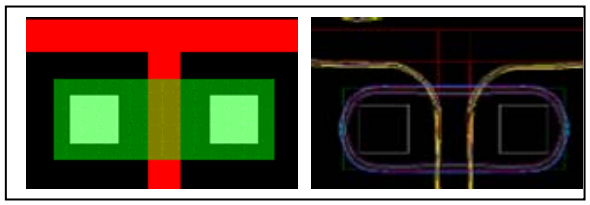


Figure 6: Diffusion and poly gate contour simulation across process window using 65nm secure model [7]

Figure 6 shows the silicon contour prediction using such a secure model at different process points (focus, exposure) of poly and active layers on a 65nm process. This systematic shape variation on silicon can lead to changes in the drive current of a transistor which must be predicted for accurate circuit simulation.

V. Litho/Stress Aware Electrical Models

Since the silicon rounding effects (flaring, necking) on poly and active layers significantly change the shape of the transistor channel, accurate circuit simulation requires a contour-based extraction based on actual drive current in the channel. A current density model to predict the drawn drive current must include narrow width effects [7]. For narrow width devices, more (or less) current may flow at the edge of the device, due to corner rounding, stress, or non-uniform dopant distributions. A physically-based analytical current density model can be derived from the currents for a range of device widths and lengths. This model can thus easily be calibrated to silicon data or SPICE models. Average gate length models do not account for narrow width effects, which can contribute up to 25% more (or less) current per unit width than wide transistors.

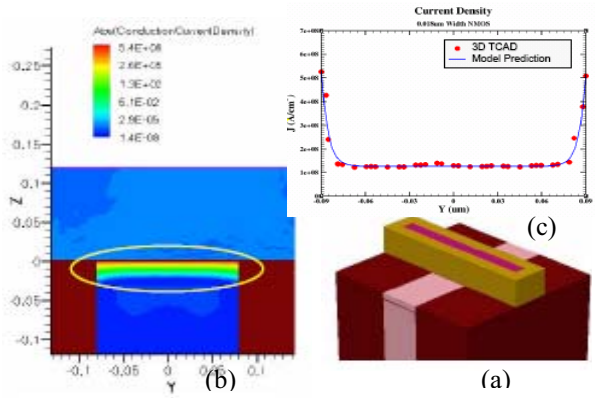
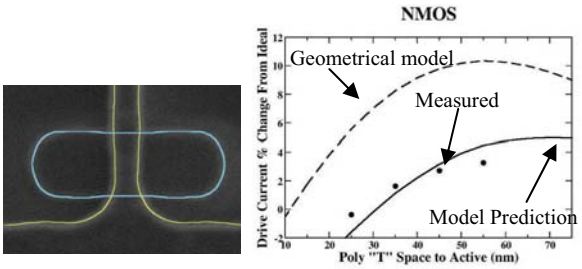


Figure 7: (a) 3-D device geometry for 50nm gate (b) Current density profile across width of device using 3-D TCAD (c) Current density profile comparison of 3-D TCAD and current density model prediction.

Figure 7 shows the prediction of current density using the analytical current density model for a 50nm transistor compared to the simulated current density from a 3-D device simulation of the same transistor using TCAD software [8]. The narrow width effect is pronounced at the edges of the device and can be seen to be accurately captured with the model. Figure 8 (a) shows the result of shape simulation using the secure model for a 65nm process compared to SEM silicon image of a T-shaped transistor structure [7]. Applying the current density model to these silicon contours, the simulated currents have been compared to measured data. The simulated vs. measured currents for varying Poly "T" to active spacings are plotted in Fig. 8 (b). The measured data represents the median of 1512 sites. The solid line shows the calculated currents using the analytical current density model. The dashed line shows the currents assuming a uniform current density across the width of the device which significantly overestimates the current.

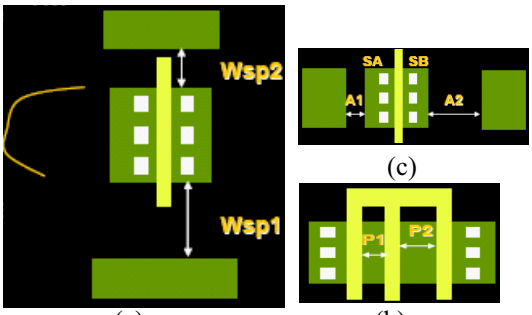


(a) (b)

Figure 8: (a) Simulated contours overlaid on silicon SEM image for a 65nm device (b) NMOS Poly "T" currents vs. poly space to active. [7]

Additional to shape effects, neighboring geometries can introduce asymmetries in the current density, due to stress effects. These effects are difficult to capture in standard compact models, which typically reflect mean values of mobility and threshold voltage. However, the current density equations can be used to account for these stress effects. Consider the case of active spacing above and below the transistor, shown in Figure 9 (a). The influence of STI stress from the neighboring active will cause the mobility to vary down the width of the device. These effects can be modeled and used to modify the analytical current density calculation. For active spacing, the current density coefficients can be modified with a power law stress relaxation equation:

$$Coeff = Coeff\left(\frac{0.5}{1+\left(\frac{D}{Wsp1}\right)^{F}}+\frac{0.5}{1+\left(\frac{D}{Wsp2}\right)^{F}}\right) \quad (3)$$



(a) (b) (c)

Figure 9: (a) Active above and below the device introduces an asymmetric mobility along the device width (b) Neighboring Poly influence the stress (c) Neighboring actives break LOD (SA/SB) symmetry.

where coeff is the required current density coefficient, D is the range, and F is the power law. Similar equations can be applied for other process induced stress effects, such as neighboring poly, dual stress liner, contact volume, and SiGe volume. The total current can then be calculated by integrating the current density along the width of the device. This current can differ significantly from the currents captured in a compact SPICE model depending on neighboring layout patterns. Most SPICE simulators allow the instantiation of parameters DELVT0 and U0 in the transistor element card of the SPICE netlist. These instance parameters modify the threshold voltage and mobility of the compact SPICE model for the given instance. The current density can be used to determine an effective DELVT0 (threshold voltage shift) and U0 (low field mobility) such that the compact model will produce the desired current. If SPICE models have been built with the LOD (length of diffusion) effect and WPE (Well Proximity Effect), they have to be modified. For example, neighboring active can introduce an asymmetry in the SA and SB parameters of the BSIM4 model. This can be corrected by placing the model in a subcircuit, for NMOS and PMOS, and adding equations that correct for the asymmetry. Additional layout parameters, such as active spacing can also be passed to this subcircuit. This combination, of the current density model and subcircuit compact model modification captures the silicon stress effects and enables designers to simulate their impact on design behavior.

With predictive litho/etch simulation followed by litho/stress aware electrical current calculation, transistor parameters can be extracted and updated in the SPICE netlist. Designers can then simulate the updated transistor-level netlist and perform an accurate timing simulation or leakage estimation. Chip designers using pre-characterized standard cells in an implementation flow based on synthesis and place-and-

route tools and timing sign-off flow based on static timing analysis can also apply this methodology.

VI. Cell Library Analysis

When designing a standard cell or characterizing its electrical behavior prior to use in a chip implementation flow, it is important to know the impact of layout context-dependent manufacturing variations on the cell's parametric views such as timing and leakage. During standard cell design, this context-dependent parametric analysis can be used to improve the cell layout and architecture so that variability is minimized. During chip implementation, the context-dependent variability is used to drive implementation tools (place-and-route) to limit the use of highly variable cells in timing-critical parts of the design and avoid leakage hotspots due to placement-related variability [9].

A flow for context-dependent library analysis is shown in Fig. 10, where other cells of the library are used to generate contexts for each standard cell of interest. The contexts for the cell are randomly generated using user-controlled parameters for context halo, spacing, filler cell occupancy rate, number of contexts and so on. Once the contexts for a cell are created, each context is analyzed with silicon contour simulation followed by contour-based extraction of transistor (and interconnect) parameters, and subsequently an analysis of the variation in timing and leakage from nominal (or drawn layout) for each cell context. A histogram of the variability in transistor/interconnect parameters and timing/leakage characteristics of the cell with the number of contexts is obtained. Statistics of the spread in variability provide insight into the cell's vulnerability to placement. From a histogram of the variation in the cell's transistor parameters (L, W, Vt0, U0 etc.) as shown in Fig. 11, or its performance attributes, metrics are derived from each histogram to determine a variability index for the cell. The mean and spread of the histogram are used to derive the variability index such that the higher the index, the more robust a cell is to placement-related variability. The variability index for each cell offers a mechanism to compare and contrast the relative robustness of different cell types or of different cell layouts of the same type to placement change. Such an index can be utilized by place-and-route tools to direct placement of cells in timing-critical areas of the design. [9]

VII. Full Chip Analysis

To determine the impact of manufacturing variations in the context of a placed and routed design layout, the timing behavior of each instance of a standard cell, given its input waveforms and wire load, is determined using a fast simulation of that cell's transistors with device parameters appropriately modified after contour simulation and parameter extraction. The design is traversed in a topological breadth-first order, and each instance simulated with the cell transistors.

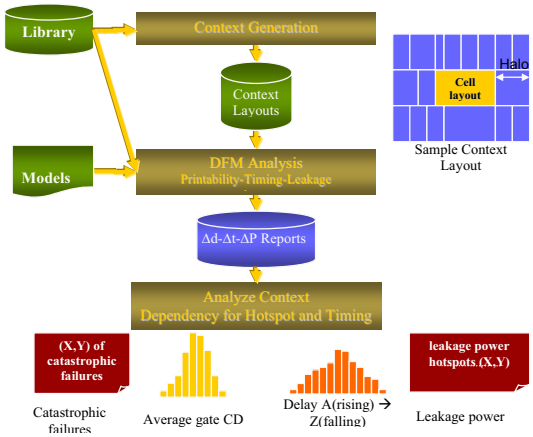


Figure 10: Context generation and analysis for placement-dependent timing, leakage and manufacturing hotspot variability.

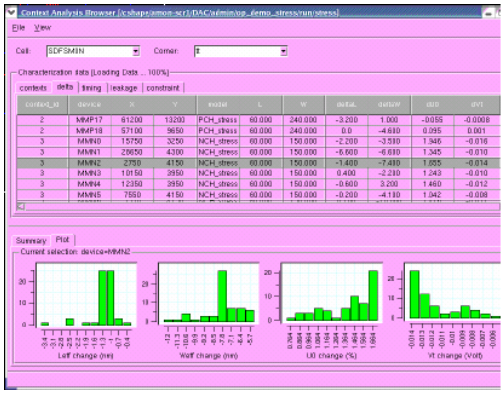


Figure 11: Library variability browser showing histograms of variation in transistor parameters (L, W, U0, Vt) due to random con text.

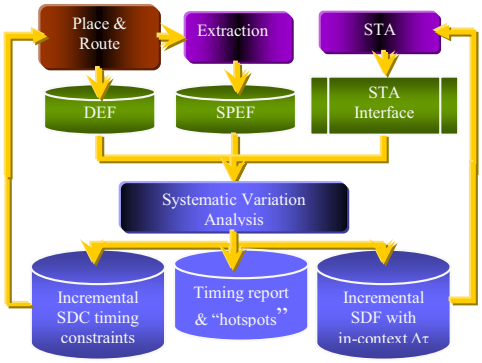


Figure 12: Full-chip methodology for systematic variation analysis and repair of parametric failures.

The change in timing due to manufacturing variation is calculated as a change in delay, $\Delta\tau$, between the cell with nominal device parameters and wire parasitics and

the cell with modified transistor parameters and wire parasitics. This is calculated as:

$$\Delta\tau = t_{d1} - t_d \quad (4)$$

and written out as an incremental delay file which can be imported into a static timing analyzer. Note that t_{dl} is the delay with manufacturing variations and t_d is the delay of the drawn netlist in (4). A design methodology incorporating systematic variation analysis is shown in Fig. 12. Design data, parasitic extraction file and a static timing report are read in and systematic variation analysis is performed. For both critical instances and nets of the design (defined by timing slack) or for all nets and instances, this analysis consists of contour simulation followed by parameter extraction and computation of delay change as in (4). The resulting delay changes are back-annotated into a static timing analyzer to report new critical paths and violations (timing "hotspots") due to systematic variation. Figure 13 shows the timing report of a failing critical path under systematic variation in a 65nm SoC designed for a clock period of 3.25ns. Each timing point in the critical path is identified and the corresponding incremental delay, the "variability delta and the total path delay up to that timing point is listed [10].

```
Start Point : i_tv80_core/ISet_reg_1_/CK
End Point   : i_tv80_core/IncDecZ_reg/D
Path Type   : max
Clock period : 3.250
-------------------------------------------------------------------------------
Point                                        Incr      Variability  Path
                                                         Delta
-------------------------------------------------------------------------------
i_tv80_core/ISet_reg_1_/CK (HSDFM1N)         0.000       0.000      0.000 r
i_tv80_core/ISet_reg_1_/Q (HSDFM1N)          0.151       0.011      0.162 f
i_tv80_core/ISet_reg_1_ASTi/A (CKBUFM2N)     0.002      -0.000      0.163 f
i_tv80_core/ISet_reg_1_ASTi/Z (CKBUFM2N)     0.137       0.010      0.311 f
i_tv80_core/i_mcode/U2039/A (OR2M2N)         0.000      -0.000      0.311 f
i_tv80_core/i_mcode/U2039/Z (OR2M2N)         0.091       0.009      0.411 f
...
i_tv80_core/U1959/Z (OR4M6N)                 0.093       0.011      3.284 f
i_tv80_core/U1958/B (CKAN2M12N)              0.000      -0.000      3.284 f
i_tv80_core/U1958/Z (CKAN2M12N)              0.036       0.004      3.325 f
i_tv80_core/U1943/B (OR2M8N)                 0.000      -0.000      3.325 f
i_tv80_core/U1943/Z (OR2M8N)                 0.035       0.005      3.364 f
i_tv80_core/U3723/B (MUX2M3N)                0.000      -0.000      3.364 f
i_tv80_core/U3723/Z (MUX2M3N)                0.043       0.005      3.412 f
i_tv80_core/IncDecZ_reg/D (HSDFM1N)          0.000       0.000      3.412 f
-------------------------------------------------------------------------------
data arrival time                                                   3.412
total path delay                                                    3.412
total delta delay                                        0.300
```

Figure 13: Timing report showing "variability delta" or delay change on a critical path.

```
Start Point : i_tv80_core/ISet_reg_1_/CK
End Point   : i_tv80_core/IncDecZ_reg/D
Path Type   : max
Clock period : 3.250
-------------------------------------------------------------------------------
Point                                        Incr      Variability  Path
                                                         Delta
-------------------------------------------------------------------------------
i_tv80_core/ISet_reg_1_/CK (HSDFM1N)         0.000       0.000      0.000 r
i_tv80_core/ISet_reg_1_/Q (HSDFM1N)          0.156       0.011      0.168 f
i_tv80_core/ISet_reg_1_ASTi/A (CKBUFM2N)     0.002      -0.000      0.170 f
i_tv80_core/ISet_reg_1_ASTi/Z (CKBUFM2N)     0.146       0.011      0.326 f
i_tv80_core/i_mcode/U2039/A (OR2M8N)        -0.001       0.001      0.327 f
i_tv80_core/i_mcode/U2039/Z (OR2M8N)         0.074       0.009      0.410 f
i_tv80_core/i_mcode/U1704/A (INVM2N)         0.000      -0.000      0.410 f
i_tv80_core/i_mcode/U1704/Z (INVM2N)         0.370       0.034      0.814 r
...
i_tv80_core/U1943/B (OR2M8N)                -0.000       0.000      3.137 f
i_tv80_core/U1943/Z (OR2M8N)                 0.035       0.005      3.177 f
i_tv80_core/U3723/B (MUX2M3N)                0.000       0.000      3.177 f
i_tv80_core/U3723/Z (MUX2M3N)                0.043       0.005      3.224 f
i_tv80_core/IncDecZ_reg/D (HSDFM1N)          0.000      -0.000      3.224 f
-------------------------------------------------------------------------------
data arrival time                                                   3.224
total path delay                                                    3.224
total delta delay                                        0.305
-------------------------------------------------------------------------------
```

Figure 14: Timing report showing delay change on the critical path of Fig. 13 after the violation is fixed.

Once new timing violations due to systematic variation are identified, repair directives are issued to the place and route tool. The timing "hotspots" repair directives are modified timing constraints that tighten the timing requirement between failing pins. By adhering to the new timing constraints, the place and route tool is able to fix the systematic timing violations [10]. Figure 14 shows the report from systematic variation analysis of the fixed design. Comparing Fig. 14 to Fig. 13 shows that the critical path is now within spec.

VIII. Conclusion

To address systematic variability at 65nm and below, a model-based DFM methodology is required. Due to the increased complexity of process-to-design and design-to-process interactions, simple rule-based techniques are not sufficient anymore. Lithography, etch and stress effects, and their impact on transistor/interconnect parameters must be well understood during design. In standard cell design for digital ICs, the impact of layout proximity on parametric variation due to litho, stress and etch can be captured using random context-based analysis. The results can be compiled into a cell variability index to drive place and route tools. For full-chip designs, a methodology to detect and repair "timing hotspots" has been presented

IX. References

[1] P. Rabkin, DFM for Advanced Technology Nodes: Fabless View, *Future Fab Intl.* Vol. 20

[2] ITRS press conference. Dec. 1, 2004, Japan.

[3] N. Shah, *"Stress Modeling of Nanoscale MOSFET,"* Master's Thesis, University of Florida.

[4] N. Mohta and S.E. Thompson, "Strained Si- The Next Vector to Extend Moore's Law, *IEEE Circuits and Devices Magazine*, 2005

[5] P. Grudowski et. al., "1-D and 2-D Geometry Effects in Uniaxially-Strained Dual Etch Stop Layer Stressor Integrations," *Proc. VLSI Technology Symposium*, 2006.

[6] J Brandenburg et. al., "A Genuine Design For Manufacturing Checker for Integrated Circuit Designers," *Proc. SPIE,* 2006.

[7] T. Devoivre et. al., "Modeling and Validation of Silicon Contour-Based Extraction and Simulation of Non-Uniform Devices," *Proc. CICC*, 2007.

[8] *ISE TCAD 8.5 User Guide*, Synopsys.

[9] D. Tsien et. al, "Context-specific leakage and delay analysis of a 65nm standard cell library for lithography induced variability," *Proc. SPIE 2007.*

[10] P. Wang et. al., "Addressing Parametric Impact of Systematic Pattern Variations in Digital IC Design," *Proc. CICC 2007.*