Impact of lithography variability on statistical timing behavior

Christopher Progler[a], Amir Borna[b], David Blaauw[b], Pierre Sixt[a]
[a]Photronics Inc., 901 Millennium Drive, Allen, TX 75013
[b]University of Michigan, 1301 Beale Ave., Ann Arbor, MI 48109

ABSTRACT

We describe a numerical model for chip level lithography variability analysis. Gate level critical dimensions are adjusted based on lithographic variability simulations and these perturbed gate lengths are input to a chip timing analyzer. Statistical modeling studies highlight the interaction between lithography variability and chip timing performance including the role of lithography error correlation length, optical proximity effect residuals, exposure system imperfections and photomask errors. Understanding these relationships is a critical building block for lithographic error tolerancing, design manufacturability improvement and lithography limited yield enhancements on integrated circuits for which timing is a key performance metric.

Keywords: lithography variability, timing, simulation

1. INTRODUCTION

Low k factor lithography drives many new process-design interactions that must be comprehended early in the development process to ensure rapid device yield ramp and acceptable steady-state yield entitlement. Unfortunately, current electronic design automation methods and process simulation tools often fall short on adequately treating these critical process-design interactions under low k lithographic imaging conditions. Until recently, numerical silo tools were adequate for chip design, process development and yield understanding activities. That is, the electrical characteristics of the chip were often modeled against a certain worst case set of process variability assumptions and the lithographic tolerances developed with little linkage to specific chip level metrics. The net result can be chip designs with a non-optimal guard band or safety zone accompanied by lithographic tolerances in severe over- or under- specified states. The desire to improve cycle time and stability in new device yield drives the need to improve this linkage between chip modeling and lithography modeling efforts. Moreover, if such a linkage can be made accurate and productive, process-design tradeoffs and optimization may be evaluated during the design and post-layout data treatment steps. In fact, one might contemplate a fully process-aware chip design flow in which every element of the lithographic variability is captured in a design process. However, reality and efficiency appear to lean more toward the deployment of a relatively small subset of crossover analysis tools designed to highlight the process interactions and, optimally, take such interactions into account at specific and critical junctures in the chip design and manufacturing phases.

With this in mind, the question arises on exactly where to focus attention in the development of crossover simulation functionality. The control of CD (i.e., critical dimension) at the polysilicon gate level remains the single most important process parameter for CMOS logic circuit performance. Furthermore, as lithographic CD control approaches an undeliverable “red zone” on many technology roadmaps, the need emerges to understand precisely how lithographic CD variation impacts a specific chip design. The most interesting cases arise when device critical path transitions are a variable function of the gate CD signature. By linking estimates of lithographic CD control to critical path delay or, in the context of this work, chip timing we can study these interactions in a realistic way. Adding a statistical component to the simulation flow provides a closer representation of a true manufacturing induced variability.

Toward this, we describe an integrated simulation flow that propagates lithographic error sources through transistor netlist extraction, gate synthesis and, finally, timing estimation. Our flow is of sufficient productivity to allow statistical error studies on full-chip designs consisting of thousands of transistors. Such chip sizes are large enough to study the interaction between lithographic error sources and timing behavior. The lithography variability of most interest in this work is a statistical variation of systematic error components. This error subset is important in low k lithography implementation and tolerancing. For example, model based optical proximity correction (i.e., MBOPC) attempts to remove or compensate for systematic within chip errors driven by the proximity of features in a device layout (e.g., the change in critical dimension as a function of pitch). Despite best efforts in MBOPC, variability in the lithographic printing process due to say errors in setting focus or exposure will produce a residual signature in the compensated CD control. This residual signature can leave a highly correlated and hence electrically important imprint across the chip.

The remainder of this paper describes the simulation flow we have developed along with results from studies covering exposure/focus, lens aberrations, mask errors and point defects. The value in linking lithographic variability simulation to chip timing prediction is hopefully made clear and brief proposals for deploying such a capability in the overall design flow are provided.

2. DESCRIPTION OF THE MODELS

Our lithography variability analysis flow consists of 4 key modules marked in Figure 1. In Module 1, lithography simulation is used to estimate the change in critical dimension due to a specific lithography error source. The perturbed layout is then passed into a netlist extraction module along with the starting unperturbed layout. Here, a transistor level netlist is produced with all gate lengths adjusted in accordance with the lithography error simulation. Next, the gate generator in Module 3 splits this flat design netlist into logic modules which become inputs for Module 4: the timing analyzer. In this module, a static timing analyzer generates timing estimates on all the perturbed circuits critical paths. This overall flow is repeated a number of times with different realizations of lithography error thus building statistics between lithography variability and chip timing.

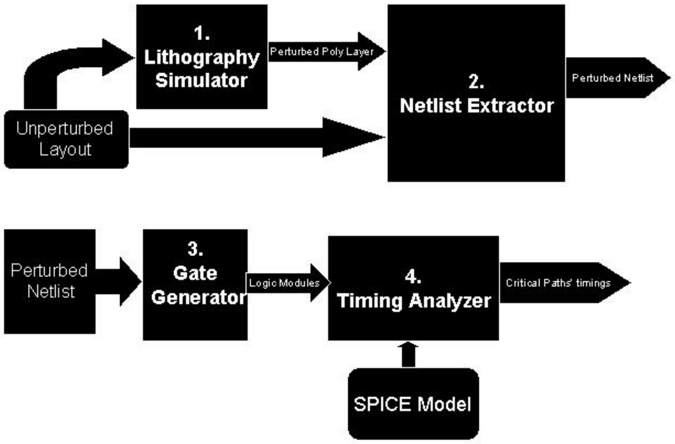

Figure 1: Variability analysis flow

2.1 Lithography simulator

Figure 2 shows the primary elements of the lithography simulation engine custom built for this work. The gate and active layers of a full design are extracted, flattened and a new layer is created from the merged gate/active levels. If the subsequent layer is too large to handle in one simulation (i.e. number of polygons drives either a physical memory or complexity bottleneck) then a block processing step is applied where a smaller block is automatically extracted and a new subset of polygons established within this new block. The full layer is built up with a sequential sum of these smaller blocks. Within the block, a full area lithography simulation is performed at the end points of the lithography error sampling space using a standard Hopkins formulation [1]. The resulting image is interrogated for transition points within the overlapping active-gate region. Transition points are highlighted by taking the irradiance difference function

$$\Delta I = I_{EP1} - I_{EP2} \qquad (1)$$

where EP refers to the error tolerance endpoints. From here, positions along the *ΔI* difference contour showing the largest variation inside the active region are tagged for subsequent computation during Monte Carlo analysis.

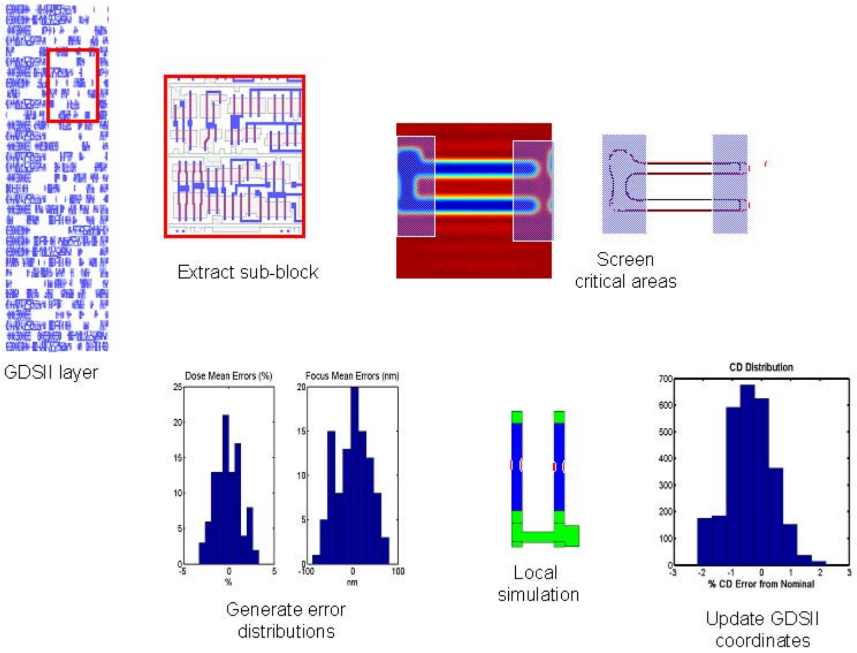

Figure 2: Lithography simulation flow

After the critical area tagging, error distributions are constructed from a pre-defined set of lithography variability components driven by focus/dose changes, projection lens effects, photomask errors and point defects. For each single realization of the error components (i.e., presumably one error realization corresponds to one chip in a multi-chip fabrication process), a local simulation is performed within the tagged areas of the device. To compute the partially coherent aerial image at a single resolution point in the layout, the following equation may be employed

$$I(u_1, v_1) = \sum_{source} \left| \iint h(u,v) A(u - u_1, v - v_1) \exp i2\pi(ux + vy) du dv \right|^2 \qquad (2)$$

where *h* is the complex amplitude response function of the imaging system and *A* is the object field centered on the specific computation point of interest. It is straightforward to extend Eq. 2 to a cluster of points within a tagged region for rapid local imaging calculations. Our concept of a local calculation also permits productive resist and vector simulations in the model should they be necessary.

While Figs. 2 along with Eqs. 1 and 2 summarize one method to productively manage the image computation step, many other possibilities exist to efficiently compute images. Alternatives may provide faster results under various conditions and approximations. The more critical aspect of the lithography simulation step is the ability to accurately and flexibly model multiple lithography error sources in a spatially (i.e., across field, across wafer etc.) correct way.

2.2 Timing analyzer

In order to analyze the impact of the resulting gate length variation on circuit performance we consider critical path delay as the metric for timing variation. In this case, the path delay (D) is the sum the gate delays (d_k) located on the path

$$D = \sum d_k \; . \tag{3}$$

The gate delay (d_k) can be modeled using a simplified inverter based delay equation

$$d = \frac{C_L V_{DD}}{2I} \tag{4}$$

where C_L is the load capacitance and I is the saturation current of the pull up or pull down network. In order to distinguish between pull up and pull down saturation currents, Eq. 5 [2] should be applied

$$d = \frac{C_L V_{DD}}{3.7}\left(\frac{1}{I_{dn}} + \frac{1}{I_{dp}}\right) \tag{5}$$

where I_{dn} is the pull down network saturation current and I_{dp} is that of the pull up network.

The ultra-short channel lengths in deep submicron devices drive a large electric field for most of the carriers. As a result, carriers pass the channel with maximum velocity (v_{sat}) which theoretically makes the saturation current independent of channel length [3]

$$I_{dsat} = WV_{sat} C_{ox} (V_{DD} - V_{th}) \tag{6}$$

Experiments show that the saturation current in deep submicron devices is not completely independent of channel length and the saturation current can be described by the universal empirical equation

$$I_{dsat} \sim L_{eff}^{-0.5} T_{OX}^{-0.8} (V_{DD} - V_{th}) \, . \tag{7}$$

To simplify the analysis we assume $L_{eff} \approx L_{gate}$ and $C_L \cong L_{gate} W C_{OX}$. Combining (5) and (6) we link gate delay with the gate length (L_{gate}) [4]:

$$d = K L_{gate}^{1.5} \tag{8}$$

where *K is* a process related constant.

Although the above timing formulas are derived for an inverter, the gate length can be modified to adopt these formulas to any gate configurations. In our timing analysis step, the delay computation begins from the logic modules (i.e., gates generated by the gate generator module) connected to primary inputs and propagates toward the output nodes. For each node in the netlist, low-to-high and high-to-low delays are computed and the related waveforms are stored in piece-wise-linear format to be used for the delay computation of successive nodes.

Once all the inputs of a gate have known delays, single-transient combinations of inputs are simulated using the spectre (Cadence) circuit simulator. The gate output delay is then computed for each combination of inputs using

$$outDelay = i\_k\_Delay + gateDelay \tag{9}$$

where *i_k_Delay* is the delay of the input #k and *gateDelay* is the logic module's 50% delay due to this input combination. The maximum *outDelay* is selected as the output delay of the gate and the output waveform is stored and used in subsequent delay computations as the signal propagates. The above process continues until all the primary outputs have known delays

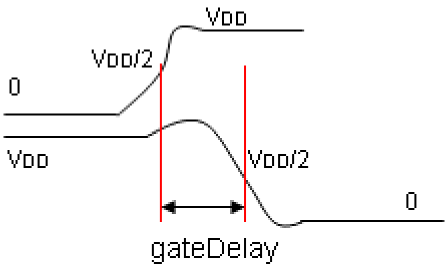

Figure 3: Graphical description of gate delay

.

3. LITHOGRAPHY VARIABILITY STUDIES

The variability simulation flow described in section II was tested on a 16-bit adder device constructed from the Artisan design library using TSMC 0.18 technology files. To highlight lithography variability driven by optical effects, such as proximity interactions, a lithography based k-factor of 0.40 was applied (i.e., as opposed to shrinking the device to a smaller dimension). The adder device had a total of 2826 gates over an area of roughly 180 by 50 microns. The following sections summarize the primary error components considered. Three sections follow that highlight the individual errors of dose/focus, lens aberrations and mask point defects. The last section combines all errors with mask CD control into a larger Monte Carlo study.

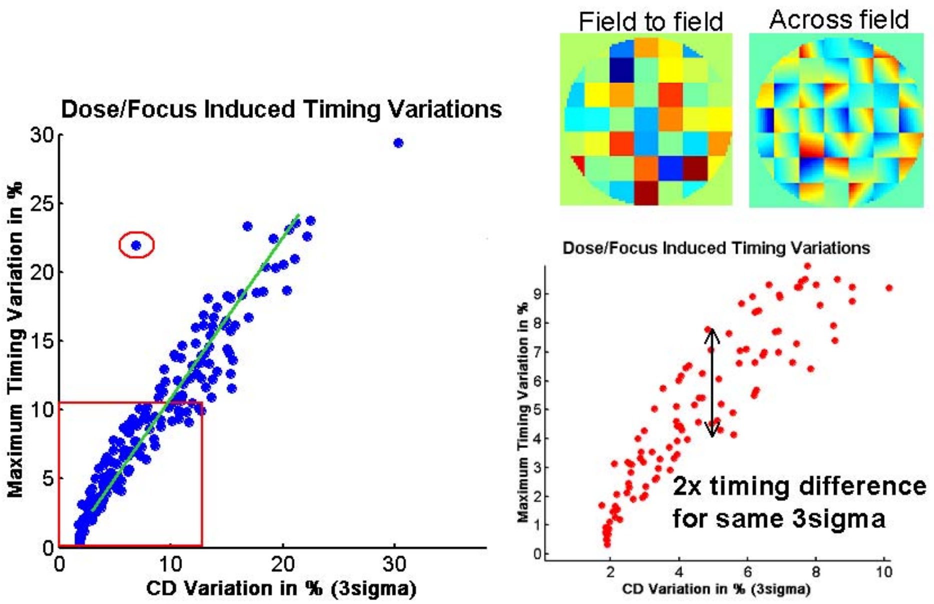

Figure 4: Result of first variability induced timing study covering field to field and across field dose/focus errors.

3.1 Dose/focus variation

Figure 4 (upper right) shows the errors considered in the first variability study: a field to field variation in exposure/focus setting (i.e., mean change) along with a low order across field variation in exposure/focus setting [6]. These dose and focus variations were adjusted through a statistical Monte Carlo simulation within the bounds of normal process variations to induce the CD error across gates of the adder device.

This same Figure summarizes the output of the simulation flow after timing errors are calculated for each simulation of a 200 point Monte Carlo run. The horizontal axis indicates the resulting percentage gate length variation for the gate regions over active area. That is, each data point is generated from the 3sigma value of the 2826 gates for each chip. The vertical axis indicates the maximum timing error for each chip. The plot in the lower right expands a portion of the main plot .

The linear trend line in the main plot highlights the strong correlation between maximum timing error and across chip linewidth variation within the active regions. However, for a given 3sigma CD variation, even a relatively modest 200 point Monte Carlo run highlights as much as a 2X difference in timing for the same across chip variation as annotated in the lower right plot. A clear outlier in the trend is also noted. Table 1 below compares the residual errors after fitting a linear trend line between maximum timing error and a number of common CD control metrics.

CD Control Measure	Norm of residuals after linear regression
3sigma (all gates)	48.7
Mean + 3sigma (active gates)	17.6
3sigma (active gates)	26.7
Mean (active gates)	13.1

Table 1: Comparing standard CD control metrics against timing statistics for 200 chip dose/focus Monte Carlo simulation

It is clear from Table 1 that dose and focus variations which generate well correlated active gate length changes are more critical for timing error control. Moreover, the correlation must be considered on the specific collection of active gates in the device.

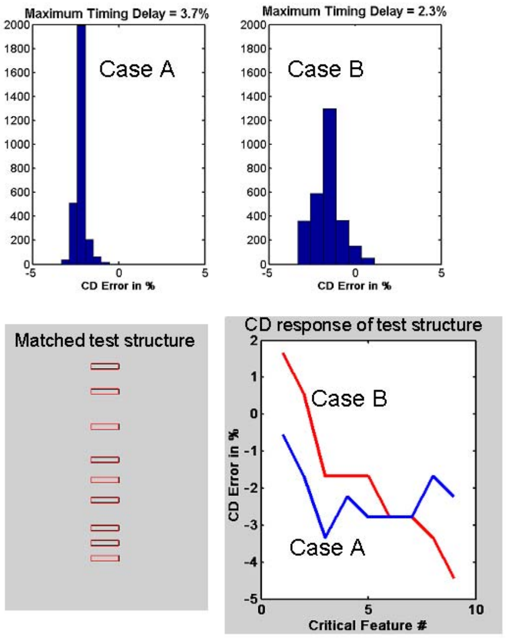

Figure 5: CD histograms extracted from an additional Monte Carlo run exercising dose/focus errors

Figure 5 shows an example along the same theme from a second dose/focus variability study using smaller perturbations of the dose/focus error. This figure summarizes only CD histograms of two specific chips (cases A and B) after the Monte Carlo run.

In Case A, the error has a very narrow spread with a slightly larger mean error while in Case B the distribution is somewhat broader with a lower mean value. Table 2 summarizes the key metrics for the two specific chips highlighted in Figure 5.

Metric (active gates)	**Case A Value**	**Case B Value**
CD Mean Error (%)	2.2	1.5
Mean+3Sigma (%)	3.3	4.2
Maximum Timing (%)	3.7	2.3

Table 2: Two additional chips run with smaller single dose/focus perturbations

It is interesting to note for the very tight CD distribution in Case A (e.g., 3sigma error is only 1.1%), the residual mean error drives a larger timing error than in Case B. That is, the larger population of correlated critical gates drives a larger maximum timing error even though the total spread of the data is larger in Case B. The implications for this and MBOPC are important. For example, if dose and focus variations (or any variation for that matter) drive an OPC residual that is well correlated or "resonates" additively with the active gate spacings/sizings in the device then this lithography error is of paramount importance. The trend toward IC layouts with a high degree of design rule restriction ironically drives a greater tendency for correlated CD errors when the OPC or RET implementation breaks down.

Moreover, using generic CD through pitch structures to calibrate and stress OPC models may be inadequate to protect against variability induced detuning of the OPC model as new layout constructs are exercised against the same model. Taking this one step further, Figure 5 (bottom) shows an OPC test structure derived directly from the critical gate spacings and widths in the gate level layout (as opposed to a generic CD through pitch or similar structure). Figure 5 right shows that this test module is able to resolve the dose/focus induced interaction in the device by emphasizing the well correlated variations driven by focus/dose condition A.

3.2 Lens aberrations

Prior work linked specific lens aberration signatures to lithographic variation [7, 8]. Lens aberrations will also detune optical proximity correction and RET models as a user attempts to port a device across multiple exposure systems. In the ideal case, the lens user drives exposure system defects to a level where a model developed for one system remains correct for alternate systems or perhaps critical layers, such as gate, are dedicated to well-matched exposure systems.

Such goals may not be practically achievable and simulating the impact of aberration induced CD residuals on timing of a given design is useful for assessing manufacturability concerns, setting tool matching specifications and technology transfer strategies. It may also be useful in debugging timing related yield failures.

Figure 6 upper left shows a calibration simulation where the OPC model quality is initially established using a given exposure tool error signature. In the calibration step, correction quality is varied in equal steps from no correction at the extreme right end of the plot to a well-corrected state at the extreme left. As an aside, such a simulation exercise can be used to set OPC aggressiveness (i.e., edge segmentation, feature count, mask resolution) linked to chip timing needs however that is not the intent for this particular study. A specific OPC model is then frozen to give a residual timing error of approximately 5%. Next this model is deployed across 7 alternate lens signatures extracted from production grade exposure systems with RMS wavefront errors between 22 and 37 milliwaves. The tool to tool mean CD control is held to better than 1% for this exercise thus minimizing effects from an aberration driven mean shift in linewidth. The lower left graph of Figure 5 highlights a range in maximum timing errors of roughly 2X across the 7 exposure systems.

On the right of Figure 6, we see that timing performance is only weakly correlated with RMS wavefront error in this case. However, after running a Zernike coefficient sensitivity study, certain coefficients are noted to drive strong timing error while other coefficients show little impact. In fact, 4 Zernike terms show timing correlation levels in excess of 90% for this device creating a valuable link between design and manufacturing steps.

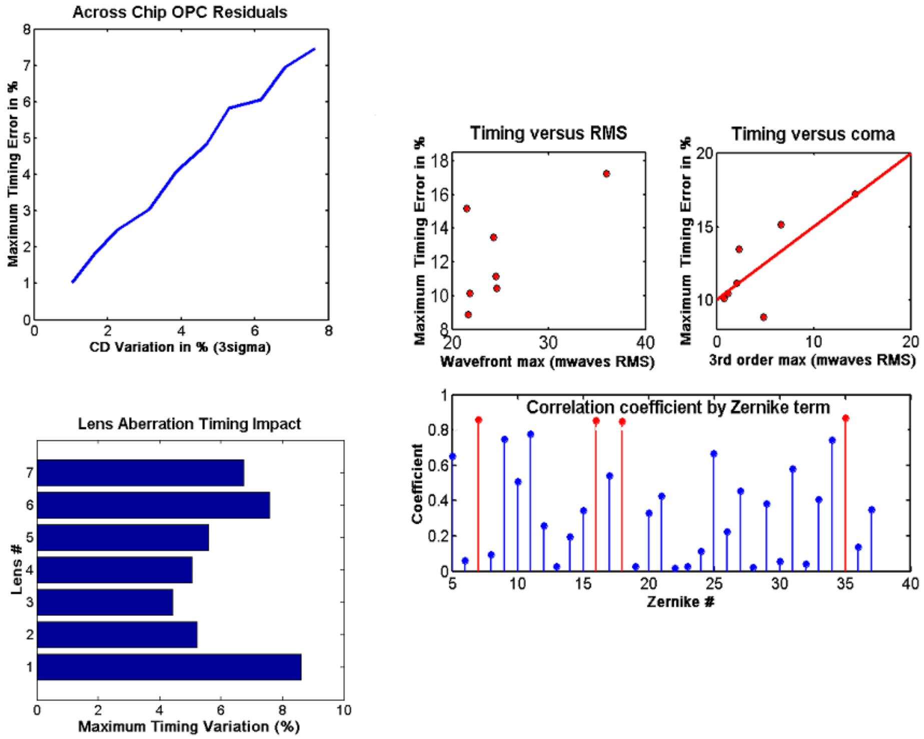

Figure 6: Lens aberration matching linked to timing analysis

3.3 Point defects

Point defects in the context of this work refer to very localized CD errors due to perhaps a local contamination (i.e., immersion lithography bubbles, added particles, processing perturbations) as opposed to a CD variation with broader spatial extent as described in the previous sections. Point defects are interesting as part of a timing study: though the local CD variations may be large, they will not be correlated over long distances in the device and hence should drive less timing variation.

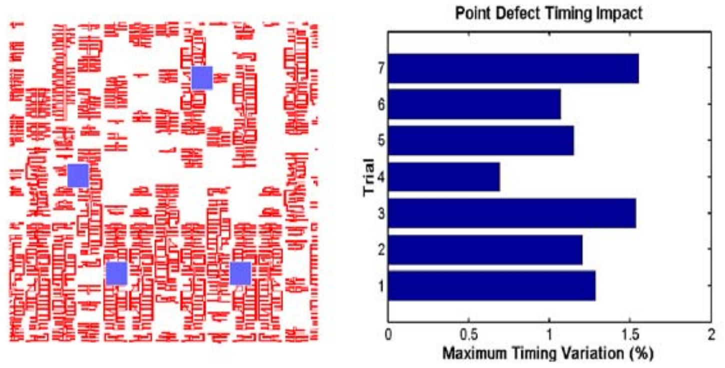

Figure 7: CD point defect example for timing impact

Figure 7 shows a layout where the local CD's are perturbed by 5% in small isolated blocks randomly placed throughout the gate layer. The final 3sigma CD variation across all gates is approximately 4% after block perturbation and the mean error is less than 1%. Figure 7 (right) shows the result of 7 point defect perturbations (i.e., random placements of the CD error blocks) on timing performance. Even though the CD variation as measured by 3sigma is significant, the timing impact is very small due to the uncorrelated nature of these CD point defects.

3.4. Multi-error example

The final study combines multiple errors into one Monte Carlo run. Figure 8 shows the error tree designed for this specific timing sensitivity evaluation. Three exposure systems (i.e., lens aberration signatures) are considered along with 3 mask CD signatures. In addition, dose/focus variations and point defects are applied to each chip across the run. During execution, the 200 chip Monte Carlo evaluation takes each field in turn by applying a statistical variation of dose, focus and point defects on top of a fixed lens and mask signature corresponding to one of the three tool/mask combinations. Mask CD and exposure tool signatures applied here are derived from the spatial shape and magnitude found on production grade masks and exposure systems consistent with a process operating at a lithographic k factor of 0.40. The right of Figure 8 separates the histogram by summarizing the Monte Carlo simulation into 3 mask/tool groupings. As shown, the unique signatures driven by each mask/tool grouping drive specific CD versus timing behaviors. For example, mask/tool set 3 shows a highly non-linear variation in timing for small perturbations in CD while mask/tool set 1 gives the largest CD/timing errors and the steepest trend slope.

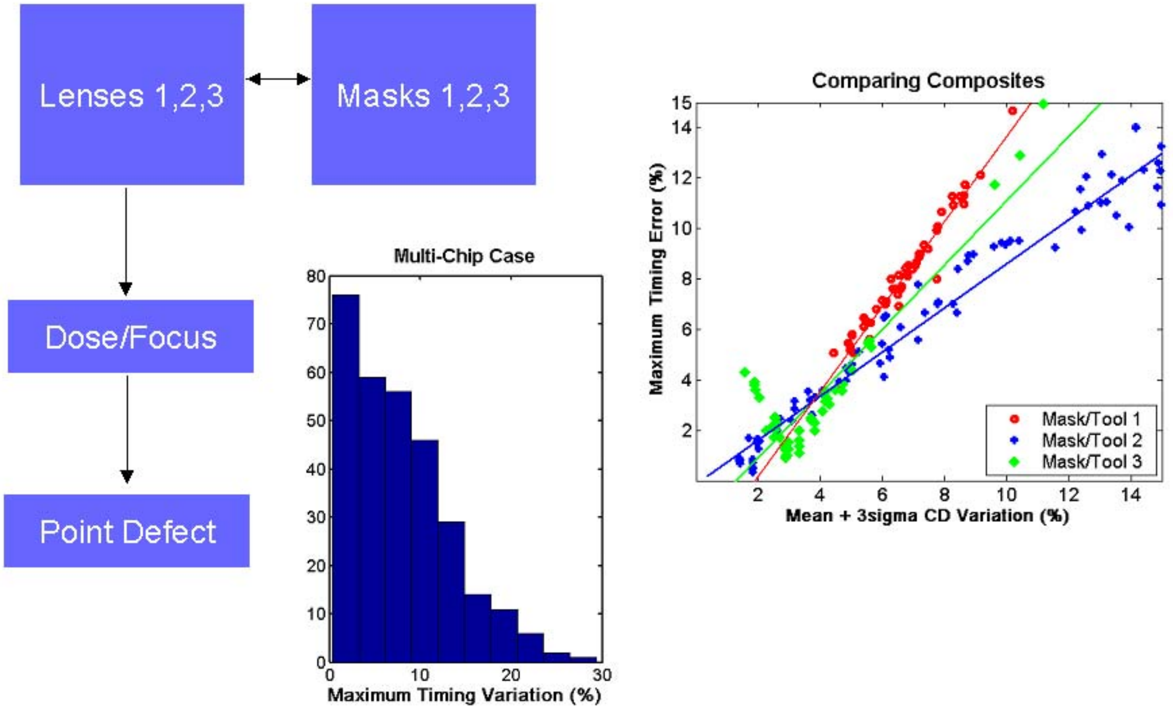

Figure 8: Multi-error 200 chip Monte Carlo run

4. CONCLUSIONS

The use of physical and statistical simulation techniques appears advantageous for mapping out the critical interaction space between lithography variability and chip timing. Toward this, we have demonstrated a productive simulation tool linking multiple sources of lithography variability to chip timing error. We find that generic mean/sigma CD control metrics may be inadequate to fully comprehend the relationship between lithography and chip timing. This seems particularly important when developing RET strategies for a given device family, fine tuning error budgets for manufacturability and driving yield improvement under marginal processes situations. Sub-component specifications (e.g., OPC residuals, process variations, mask, defects) should consider the correlation length of errors and the specific relationship between error sources and critical chip performance metrics. When such a simulation link is made, a variety of design for manufacturability advances are also enabled. For example, our simulation flow can be used to set OPC aggressiveness targets via a direct link to chip timing or screen cells for manufacturability against a specific capability index.

Regarding optical proximity correction or resolution enhancement data treatment in general, our work shows that the interaction of the RET model with both electrical chip metrics and statistical variability must be well understood for optimum chip and lithography design functions. Simulating stressors that detune the correction or compensation assumptions is important and such simulation activities should map spatially accurate variability into the chip performance domain. If such stressors drive well correlated failure in the RET model this will drive a large corresponding impact on the timing behavior. Moreover, it may not be obvious how the lithography variability links into the critical gate distribution using generic test and ACLV measurement structures. With this in mind, exploration of improved spatial metrology methods that calibrate variability to critical gate transitions appears to be worthwhile. Improved model calibration structures more adequately representing the specific device behavior may be useful as well. In this work, we propose a quasi-chirped design that accurately mapped the timing versus CD control error space using a limited number of measurements. Our timing simulation tool allowed these designs to be explored with a link to final chip performance metrics.

Regarding design flow options for the described simulation tool, one potential insertion point is the cell place and route step. Here, timing versus lithography variability criteria may be used to govern cell and critical path placements and block orientations. A more proactive approach might leverage simulation to avoid error prone or highly sensitive cells in the logic synthesis for critical paths. In the same way, spatial correlation distance should be used as a metric to avoid critical paths with highly correlated CD variations. This can be an important part of design rule restriction exercises in early lithography development.

Finally, under the cost versus performance umbrella, simulation flows as described here may be useful in deciding on the most appropriate CD control versus chip yield tradeoffs which are a function of the device application, complexity and desired cycle time.

REFERENCES

[1] H.H. Hopkins, “Image formation with coherent and partially coherent light”, J. Soc. Photo. Sci. and Eng. (3), Vol. 115 (22).
[2] C. Yu et al, “Use of short-loop electrical measurements for yield improvement” *IEEE* Trans. on semiconductor manufacturing, vol. 8, no. 2, May 1995.
[3] D. Sylvester, K. Keutzer, “Getting to the bottom of deep-submicron,” Proc. IC-CAD 1998
[4] M. Orhanskey et al, “Impact of systematic spatial intra-chip gate length variability on performance of high-speed digital circuits,” Computer Aided Design, 2000. ICCAD-2000. IEEE/ACM International Conference on, 5-9 Nov. 2000, Pages: 62 – 67
[5] V. Mehrotra et al, “Modeling the effects of manufacturing variation on high-speed microprocessor interconnect performance**,”** Electron Devices Meeting, 1998. IEDM '98 Technical Digest., International , 6-9 Dec. 1998, Pages:767 – 770
[6] C. Progler, “Simulation-enabled decision making in advanced lithographic manufacturing”, SPIE Vol. 4404 (2001).
[7] C. Progler**,** A. Wong, “Zernike coefficients: Are they really enough,” Proc. SPIE Vol. 4000 (2000).
[8] C. Progler, D. Wheeler, “Optical lens specifications from the users perspective”, Proc. SPIE Vol. 3334 (1998)