Toward Through-Process Layout Quality Metrics

Fook-Luen Heng[a], Jin-Fuw Lee[a], Puneet Gupta[b]
[a]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[b]UCSD, now Blaze DFM Inc., Sunnyvale CA 94089

ABSTRACT

Quality of a layout has the most direct impact in the manufacturability of a design. Traditionally, layout quality is ensured in the first order by design rules, i.e. if a layout is free of design rules violation, it is a good layout. It is assumed such a layout will be fabricated to specification. Moreover, a design rule clean layout also ensures the electrical performance of the circuit it represents. There are other layout quality measures, e.g. random defects yield of a layout is modeled by critical area, systematic defects yield is sometime measured by a weighted score of recommended design rules. All the traditional layout quality measures are computed with drawn layout shapes.

In the advent of low K_1 lithography and the increasing variability of process technologies beyond 90nm, nominal layout quality measures need to be revisited. Traditionally, nominal electrical properties such as L-eff and W-eff are extracted from drawn layout, and the corner cases are estimated with worst case process conditions. Most of these parameters are layout pattern dependent. As a matter of fact, they can be systematic through process and can have large impact in the modeling of circuit parameters [1].

In this paper, we investigate a through process layout quality measure, in which we extract through process electrical parameters from simulated through process resist contours. We showed a mechanism to compute a statistical model that predicts through process electrical parameters from the process parameter variation. We demonstrated that such computation is practical.

1. INTRODUCTION

Due to manufacturing process not keeping up with the scaling of layout geometries, discrepancy between shapes drawn by the designer and those printed on wafer is growing. As a result, modeling of and accounting for these process variations becomes an important component of current and future design flows. A large fraction of variability is systematic and predictable [1]. Several sources of variation impact the layout and hence designers' intent in a predictable pattern-dependent way. Examples of such sources include focus, exposure dose, misalignment, chemical mechanical planarization. With existence of such pattern dependent variations, manufacturable and variation-robust layout becomes very important [2,3]. A methodology to evaluate layout quality in terms of yield, manufacturability, performance and power metrics is absolutely essential for such layout design. Such layout Quality of Result (QOR) metrics can be a key component to qualify all design for manufacturing flows.

Electrical properties of a circuit layout are given as nominal values, or approximated by a distribution. Critical electrical parameters such as gate length, contact resistance are very geometry dependent and vary systematically. Other yield determinants such as critical area for defects are also very pattern dependent. With accurate through process wafer shape simulation, these critical parameters can be characterized more accurately. Accurate QOR can enable (1) layout quality evaluation; (2) layout yield optimization; (3) fine tuning of layout as technology becomes more mature; (4) reduced guard banding in electrical modeling. Contributions of this work include the following.

- Methodology to compute various circuit parameters such as gate length, gate width and contact resistance from simulated resist contours.
- A flow to quantify the impact of varying process parameters such as focus and misalignment on the design parameters and to obtain realistic distributions for them.

The organization of the paper is as follows. Next section describes the components of layout QOR. Section 3 describes the flow for estimation of QOR specifically for cell layouts. Section 4 gives the experimental description and results. We conclude with ongoing work in Section 5.

Design and Process Integration for Microelectronic Manufacturing III,
edited by Lars W. Liebmann, Proceedings of SPIE Vol. 5756
(SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.601842

2. QOR COMPONENTS

To assess layout quality, all geometry dependent parameters which impact power, performance or yield need to be computed with varying process conditions. The list of such parameters is long and includes gate length, gate width, contact/via resistance, metal resistance and capacitance, critical area for defects, electromigration reliability, etc.

2.1. QOR Metrics

The four axes for QOR measurements are as follows.

1. *Performance*. For devices, performance comes from gate length and gate width. Since, printed wafer shapes are not rectilinear or uniform as drawn; therefore estimation of these parameters after wafer shape simulation is important. For vias and contacts, resistance is the performance parameter. For metal interconnect, both resistance and capacitance need to be computed from wafer shape contours. Global and semi-global interconnect is more critical to performance than local interconnect. Moreover, as a result of varying process, these design parameters and hence delays of input to output paths are distributions rather than single values.
2. *Power*. Similarly, for power whether leakage or dynamic power, gate length and gate width are key parameters. Due to difference in nature of dependence of power metrics such as leakage on layout parameters such as gate length, the ``averaged'' measurements for the layout parameters would differ from metric to metric.
3. *Functional Yield*. Low catastrophic or functional yield loss is very important especially during the process ramp-up phase. Since wafer shapes look different than the drawn shapes, critical area analysis needs to be done on simulated wafer shapes rather than drawn shapes. Another mechanism for functional yield loss is opens and shorts through process due to imperfect lithography. For instance, with varying focus, lines may grow or shrink resulting in opens or shorts. There can be several other mechanisms for yield loss in sub-90nm processes. Another example is resist pattern collapse due to long high aspect ratio lines [4].
4. *Reliability*. Mean time to failure (MTTF) due to effects such as electromigration is another QOR metric for layouts. Electromigration is heavily dependent on current density which in turn depends on the cross-sectional area of the wire. Since the final printed cross section of a wire changes with process, MTTF is also a distribution. Similar reliability issues are present for vias and contacts.

Given probability distributions of power/performance, another aspect of QOR computation is a quality metric of the distribution itself. For instance, questions like whether it is good to tradeoff variance for mean of a Gaussian become relevant in a statistical analysis and optimization regime. Similar questions about correlation between different cells within the same library are also possible. For instance, if one wants all cells in the library to be correlated in gate length, the percentage contributions of each of sources of variation to gate length variation should be similar. This makes layout quality a metric of the entire cell library rather than a standalone cell layout. In this work we will focus on QOR for single cells rather than a whole design. Moreover, we will discuss only the power and performance aspects of through process QOR and leave functional yield for future work.

2.2. Sources of Pattern Dependent Variation

Several basic sources of process variation impact drawn patterns in a predictable and systematic way. These are the sources of variation whose impact can be simulated. Examples include focus, exposure dose, chemical mechanical planarization, flare, misalignment errors, etc. In the current work we take into account focus and alignment errors.

Variation in focus can come from systematic topography variation on wafer, lens aberrations and random equipment errors [5]. In the current work, we model focus as a Gaussian random variable with zero mean. As shown in [1], systematic variations of electrical parameters can result with a random focus variation.

A misalignment or overlay error between masks of different layers is another source of variation which is typically handled in design rules. For gates, the relevant overlay is between active and poly masks while for contacts or vias three masks are involved. We model overlay errors between two layers as the perturbation (in polar coordinates) of one layer with respect to other. Every misalignment is denoted by a (r, θ) pair with (0, 0) means perfect alignment: r is taken to be normally distributed around 0 while θ is assumed to have a uniform distribution on $[0, 2\pi]$.

3. THE QOR FLOW

The generic flow for computing QOR metrics is outlined in Figure 1. A canonical environment for the cell layout as in [6] is constructed around every cell. The layout is then made to go through the standard mask data preparation process which includes Optical Proximity Correction (OPC), Phase Shift Mask (PSM), etc. Each layer of the corrected layout is then simulated at different defocus levels to obtain wafer shape contours at different focus points. To emulate misalignment, layers are translated in x direction by $r\cos\theta$ and in y direction by $r\sin\theta$ with respect to the reference layer. In this section we describe our methods of computing various circuit parameters from wafer shape contours.

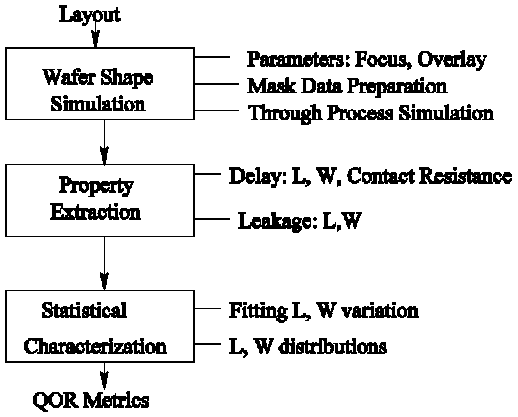

Figure 1: Flow for calculation of through process layout quality

3.1. Calculation of Gate Width

To compute effective gate width we first shrink the active region contour to within the misalignment tolerance with respect to poly. We then approximate the non-rectilinear contour with an equivalent rectangular active region which has the same area. The flow is depicted in Figure2.

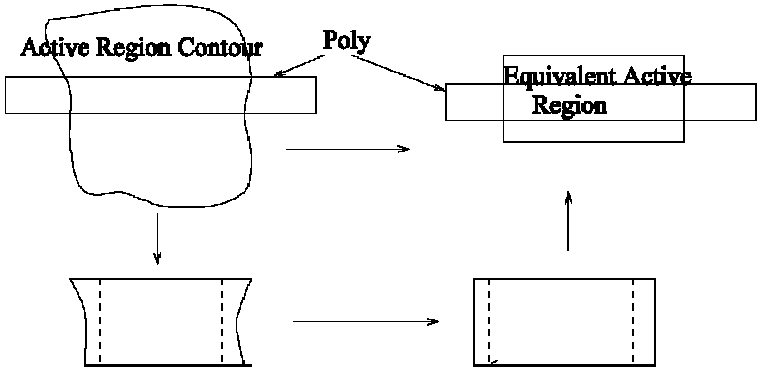

Figure 2: Construction of equivalent active region from an active region contour

3.2. Calculation of Gate Length

Calculation of gate length is more involved due to heavy dependence of leakage and delay on it. The basic flow is shown in Figure 3. The gate shape contour (derived from intersection of poly layer and equivalent active region) is made rectilinear to yield (W, L) pairs.

We pre-construct a look-up table of currents and current slopes yielding $I(W, L)$, $\frac{\Delta I}{\Delta V_d}(W,L)$, $\frac{\Delta I}{\Delta V_g}(W,L)$, where $I(W, L)$ corresponds to saturation current (for computing equivalent gate length for delay) or off current (for computing equivalent gate length for sub-threshold leakage) and the other two entries are slopes of the current with respect to drain and gate voltages respectively. Since the granularity of width for which this look-up table needs to be constructed (20nm in our experiments), is much smaller than required for validity of typical short-channel device models, we compute the current values for a ``20nm wide device'' as the difference between currents values of a 1000nm and 1020nm devices. This also avoid repeated counting of ``end effects'' such as width correction.

The equivalent gate length is computed to minimize

$$[((I(W_{equiv}, L_{equiv}) - \sum I(W_i, L_i))^2 + (\frac{\Delta I}{\Delta V_d}(W_{equiv}, L_{equiv}) - \sum \frac{\Delta I}{\Delta V_d}(W_i, L_i)))^2 + (\frac{\Delta I}{\Delta V_g}(W_{equiv}, L_{equiv}) - \sum \frac{\Delta I}{\Delta V_g}(W_i, L_i)))^2]$$

where (W_i, L_i) constitute the width, length pairs along the gate width $W_{equiv}(= \sum W_i)$. In the look-up table length is varied at the granularity of 1nm while width is linearly interpolated. Note that current values have to be looked up from the appropriate look-up table constructed for each different kind of device used in the layout.

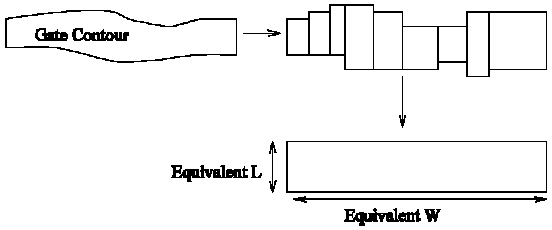

Figure 3: Calculation of equivalent gate length

3.3. Line End Shortening

With process variation, some gates may undergo line end shortening resulting in short gates as shown in Figure 4. Modeling such short gates as devices with resistance in parallel, we see that 3-4nm of shortening is sufficient to make the device stop functioning. Therefore, we treat such line end shortening as a fault.

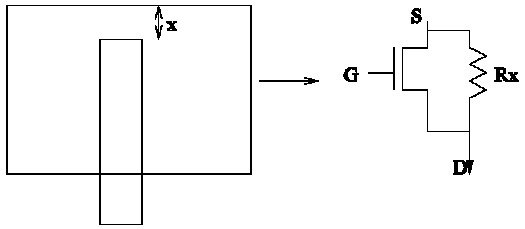

Figure 4: Line end shortening can be modeled as a resistance in parallel to the device

3.4. Generating Canonical Environments

If a small piece of layout (such as a standard cell) has to go through layout quality analysis, constructing a layout environment for it which is fairly representative of its instantiations in a real design is important. One simple way of constructing such environment was proposed in [6]. We take a slightly more involved route tied to design rules and methodology. We construct two environments for the layout: ``dense'' and ``isolated'' to emulate the two extreme

printing contexts.[1] We generate the context by tiling specially laid out minimum sized ``dense'' and ``isolated'' cells. Layouts for both context cells obey all design rules (e.g. minimum poly spacing).

We assume the layout under test to be equally likely to occur in each context. As a result, we calculate QOR numbers as average of dense and isolated contexts. Figure 5 shows a standard cell and its dense canonical environment.

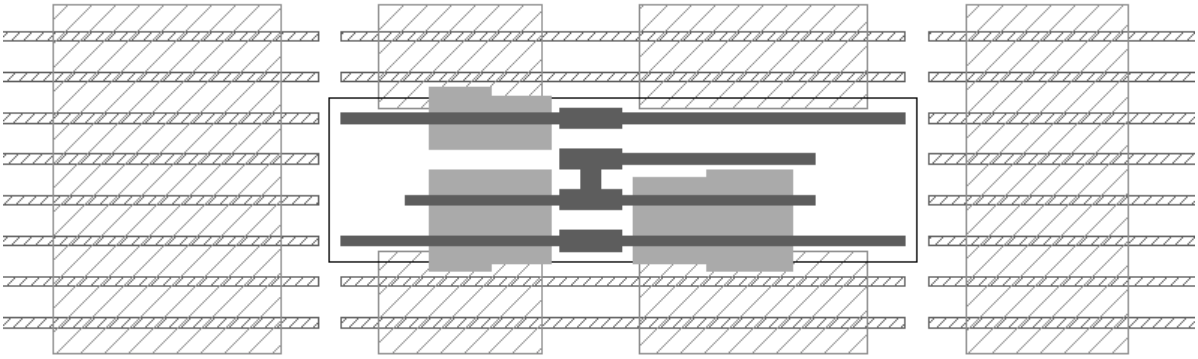

Figure 5: A standard cell and its dense environment

4. EXPERIMENTS AND RESULTS

4.1. Tables of equivalent CDs

Since it is quite expensive to generate simulated contours, we create minimum number of through-focus contours required. To generate ideal quadratic dependence on focus, at least three through focus contours are needed. To compile a reasonable table of equivalent CDs for general focus variations, we made 7 through focus contours for both poly and diffusion masks, and 25 misalignment vectors ($r\cos\theta$, $r\sin\theta$). For each combination of these process variations, W_{equiv} and L_{equiv} are computed for all FETs from the corresponding set of contours. These equivalent CDs are then stored in a table for use in the later query during the Monte Carlo process to compute the distributions of the CDs. For process variations not in the table of equivalent CD, we can either use interpolation method or formula fitting method to compute the CD value.

4.2. Formula fitting method

If the nominal focus is taken at optimal focus point, Bossung curves become symmetric with respect to the nominal focus. For a symmetric smile/frown curve, the leading term in the Taylor series expansion is quadratic for small focus variations. Under these two assumptions, we can fit W_{equiv} and L_{equiv} with quadratic formula in focus variations. Moreover, we model the dependence of CDs on the misalignment ($r\cos\theta$, $r\sin\theta$) to be linear. Therefore, the quadratic fitting formula for gate length and width is

$$\text{(Quadratic)}\quad a_0 + a_1F_{poly} + a_2F_{diff} + a_3r\cos\theta + a_4r\sin\theta + a_5F^2_{poly} + a_6F^2_{poly} + a_7F_{poly}F_{diff}$$

where F_{poly} is the defocus value of poly mask, F_{diff} is the defocus value of diffusion mask, and ($r\cos\theta$, $r\sin\theta$) is the misalignment of poly mask relative to diffusion mask.

For timing analysis of large macro or chip, one often need a simplified Gaussian distribution to describe the probability of delay variation, such as required in statistical timing analysis tool [7]. Such an approximate Gaussian distribution can be achieved with the linear fitting formula:

$$\text{(Linear)}\qquad a_0 + a_1F_{poly} + a_2F_{diff} + a_3r\cos\theta + a_4r\sin\theta$$

[1] The actual meaning of dense or isolated will depend on the lithography process being used. For instance, in case of a SRAF-based flow, the smallest and largest pitches may not be the two extremes of printing due to SRAF insertion.

In reality, the fabrication process becomes more and more difficult with 65nm and 45nm technologies. It is hard to control the nominal focus at the best focus point, and keep the 3-sigma focus variations small as well. To account for the deviation from the fitting formula, the general interpolation method is needed.

4.3. Interpolation method

The table of equivalent CDs forms a four dimensional grid, $y(x_1, x_2, x_3, x_4)$ where $y=L_{equiv}$ or W_{equiv}, $x_1=F_{poly}$, $x_2=F_{diff}$, $x_3=r\cos\theta$, and $x_4=r\sin\theta$. If the desired point does not fall on the grid, we need to find the 16 closest grid points on the 4-dimensional cube around that point, and then apply multidimensional interpolation to find estimate of y. The basic idea of multidimensional interpolation is to break into a succession of one-dimensional interpolations. First, the 16 grid points are grouped into 8 pairs of points with the same x_1, x_2, and x_3 coordinates. For each pair, one-dimensional interpolation is applied in the x_4 direction. This way, eight x_4-interpolated values are obtained at grid points on the 3-dimensional cube of x_1, x_2, and x_3. Iterate this procedure for x_3, x_2, and x_1, and we will get the final interpolated value at desired point. The multidimensional interpolation method will give the exact CD for points falling on the grid. For CD at the point not on the grid, we can achieve very good accuracy if we have a fine mesh of grids.

Both interpolation and formula-fitting methods have been implemented. We ran some standard cell books through. With 40000 Monte Carlo simulations, using Gaussian distributions of focus and misalignment, the probability distributions of L_{equiv} and W_{equiv} are computed. The probability distribution from interpolation method is shown in Figure 6a. The probability distributions from quadratic and linear fitting are respectively shown in Figure 6b and 6c. The quadratic fitting distributions do keep the general skew shape of the interpolation distribution, while the linear fitting distribution becomes a symmetric Gaussian curve.

The runtime for extracting the L/W distribution of a simple NAND gate with 4 transistors are 18 sec., 13 sec., and 13 sec., respectively for the interpolation, quadratic and linear method, on an IBM Power 270 workstation.

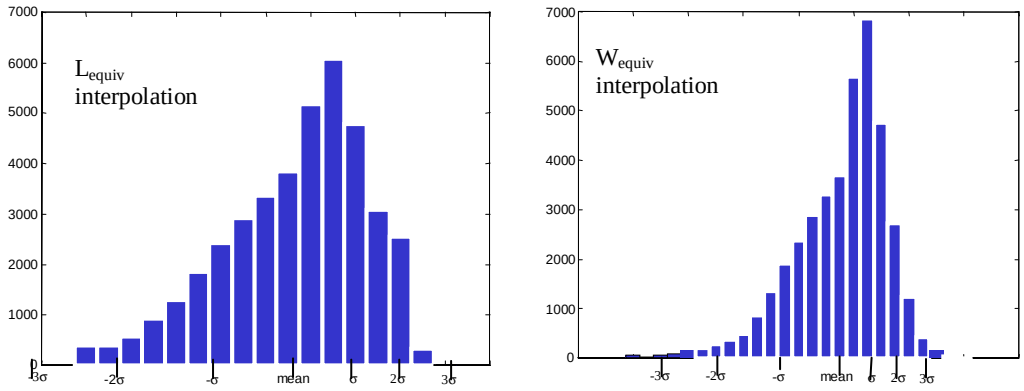

Figure 6a: The probability distribution of CDs

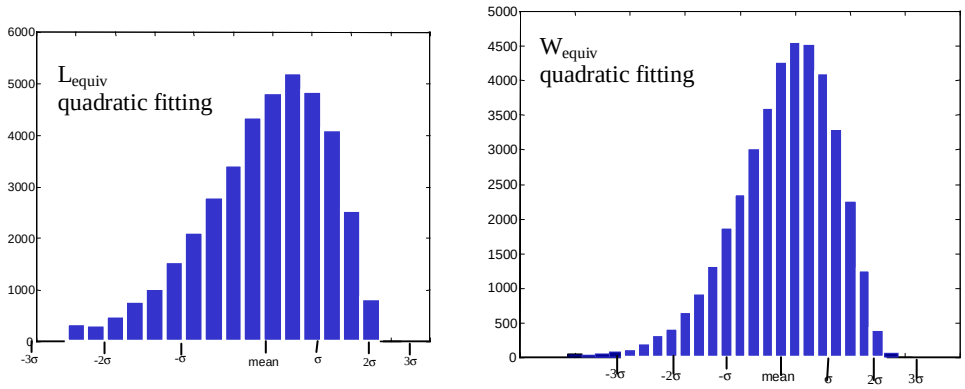

Figure 6b: The probability distribution of CDs

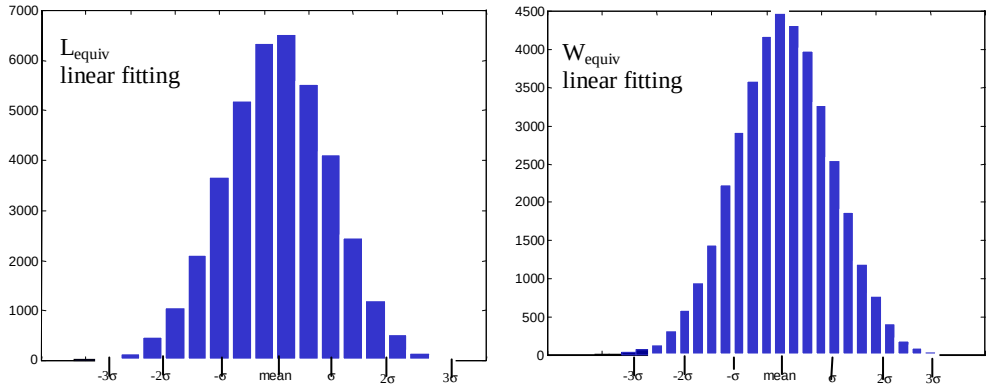

Figure 6c: The probability distribution of CDs

5. CONCLUSION

We proposed a methodology to convert simulated through process resist contours into electrical parameters. We demonstrated that quadratic method when compare with the interpolation of generated data points gives reasonable accuracy. The interpolation method runs slightly slower than the formula fitting method. In the leading edge and future technologies, the through focus variations will continue to increase relative to the feature size. The interpolation method can provide the best way to estimate probability distribution.

6. REFERENCE

1. F-L Heng, et al.``Taming Pattern and Focus Variation in VLSI Design", *Proc SPIE Conf. on Design and Process Integration for Microelectronic Manufacturing*, Feb 2004 pp 139-148.
2. Capodieci, P. Gupta, A. B. Kahng, D. Sylvester and J. Yang. ``Toward a methodology for manufacturability-driven design rule exploration". *Proc. DAC*, 2004 pp. 331-316.
3. Lars W. Liebmann, Greg A. Northrop, James Culp, Leon Sigal, Arnold Barish, and Carlos A. Fonseca, `` Layout optimization at the pinnacle of optical lithography" *Proc. SPIE Int. Soc. Opt. Eng. 5042, 1*, 2003.
4. G. Czech, E. Richter and O. Wunnicke, ``193nm Resists: A Status Report", *Future Fab Intl. Volume 12*, 2002.
5. D.G. Flagello, H.V. Laan, J.V. Schoot, I. Bouchoms and B. Geh,``Understanding Systematic and Random CD Variations Using Predictive Modeling Techniques", *Proc. SPIE Conference on Optical Microlithography XII*, 1999, pp. 162-175.
6. P. Gupta, F.-L. Heng, M. Lavin, ``Merits of Cellwise Model-Based OPC", *Proc. SPIE Conf. on Design and Process Integration for Microelectronic Manufacturing*, Feb 2004, pp182-189.
7. C. Visweswariah, K. Ravindran, K. Kalafala, S.G. Walker and S. Narayan,``First Order Incremental Block-Based Statistical Timing Analysis",*Proc. DAC*, 2004, pp. 331-336.