Transistor Flaring in Deep Submicron – Design Considerations

Vipul Singhal, Keshav C.B., Sumanth K.G., P.R. Suresh

Abstract - *The deep sub-micron regime has brought-up several manufacturing issues which impact circuit-performance and design. One such issue is flaring of transistors which causes the channel length to vary along the transistor width. It seriously impacts self-alignment of the CMOS process, causes transistor-matching issues, and makes accurate transistor modeling difficult. It can have significant adverse yield impact. In this paper, the effect, its consequences, and different solutions are examined. A methodology for modeling the effect of flaring on transistor performance is introduced. Different solutions for high density and high performance designs are proposed.*

Keywords: Design for Manufacturability (DFM), Deep Submicron (DSM), pullback, photolithography, Subwavelength-lithography, Optical Proximity Correction (OPC), SPICE–models, standard-cell library.

1. **Introduction**

The need for faster and cheaper electronic devices is pushing the VLSI industry towards fabrication of devices with the smallest possible geometries. However, as the geometries shrink, a number of manufacturing and device modeling issues are surfacing. A significant number of manufacturing issues are related to the limitations of the photolithographic process which is used to pattern the geometries onto silicon. Flaring of transistors is one such emerging issue. In the lithographic process the concave regions of a geometry (like inner corners) tend to remain under-etched due to optical intensity effects and lesser availability of etchants. This makes it nearly impossible to create sharp angles. Fig.1 below depicts one such instance.

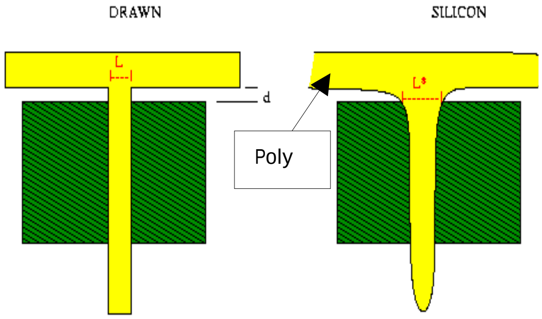
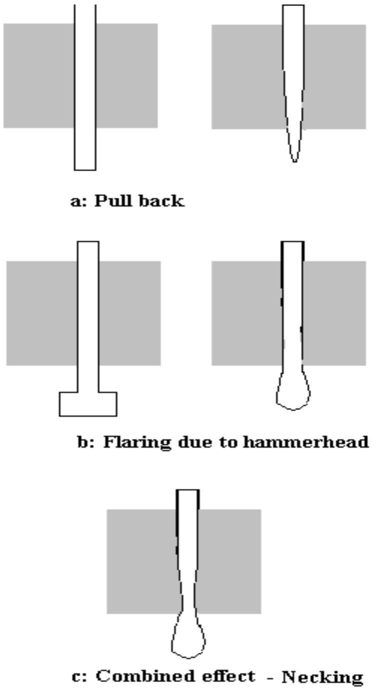


Fig 1. Transistor Flaring: Layout Vs Silicon

The channel length L increases drastically ("flares-out") towards the upper edge of the transistor. This effect is called transistor flaring. Since poly is commonly used for routing within standard cells, poly bends near transistor gates are common, and hence flaring is a very commonly occurring phenomenon in Deep Submicron Lithography.

Similar manufacturing effects related to the photo-lithographic process are pullback and necking. Pullback is shrinkage of geometries due to over-etching at corners and narrow ends (Fig 2a). To reduce pullback optical proximity correction (OPC) techniques such as creating hammerheads (Fig 2b) are used. In [1], an overview of OPC and types of OPC are presented. Necking (Fig 2c) is a reduction in the channel length in a portion of the transistor, due to combined effect of pullback and flaring at the hammerhead. In the rest of this paper the term flaring is used to include necking effects.


a: Pull back

b: Flaring due to hammerhead

c: Combined effect - Necking

Fig 2. Pullback and Necking.

In this paper we focus on flaring since it is a commonly occurring phenomenon and has considerable impact on electrical performance of circuits. The impact of flaring on circuit performance is discussed in section 2. Section 3 summarizes the previous work done in this area and why we need to find a different solution. Two approaches towards a solution – one for high performance designs and one for high-density designs – are presented in section 4. The yield impact of the proposed solutions is discussed in section 5. We conclude in section 6 by summarizing the results and their significance.

2. Impact of transistor flaring

The foremost electrical fallout of flaring is degradation in performance that is difficult to quantify. One reason is that the electrical characteristics of the flared transistor are more difficult to model compared to the straight rectangular transistor. The second reason is that the flaring makes the transistor performance dependent on the extent of poly-to-diffusion misalignment. A certain amount of misalignment between two layers is an accepted limitation in any wafer fabrication process. In the case of an unflared transistor, a minor misalignment between poly and diffusion does not change the shape of the channel. On the other hand, in case of the flared transistor the geometry of the channel is highly dependent on the extent of misalignment. This is shown in Fig 3 below. Here as the poly shifts down ("south"), the extent of flaring on the N transistor increases, and flaring on P transistor decreases. Hence Dp:Dn ratio increases, where Dp is the drive strength of P transistor and Dn is the drive strength on N transistor. Similarly when the poly shifts up ("north"), the Dp:Dn ratio decreases. Thus the Dp:Dn ratio of the circuit is now a function of the extent and direction of misalignment.

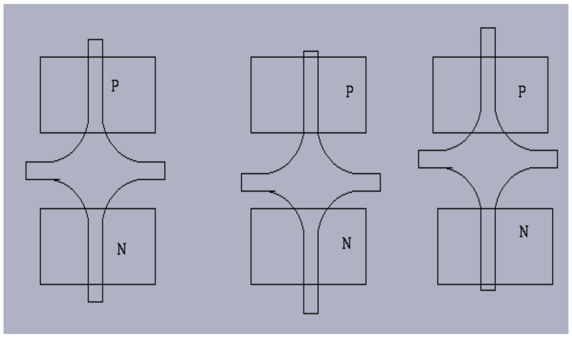


Fig 3. Illustrating the effect of misalignment with flaring: The figure on left shows perfectly aligned case. The central figure has poly misaligned south, and the rightmost has poly misaligned north.

The timing inaccuracy resulting from flaring could have significant yield impact. Our analysis with statistical simulation results showed that we could have upto 4% yield impact due to timing failures associated with flaring.

3. Previous work

A number of solutions have been proposed to correct the lithographic distortion that occurs during mask manufacturing. Most OPC based corrective schemes, which involve manipulation of layout geometries to counter the non-ideal lithographic effects. Simple OPC techniques reduce flaring and other similar distortions to some extent, and are universally used. However, to pattern transistors with shape and electrical characteristics very close to the intended, very complex OPC techniques have to be employed [1]. The penalties associated with complex OPC are increased reticle inspection time, increased complexity of reticle inspection algorithms and increased complexity of data management resulting in much higher costs associated with mask manufacturing[2]. As the device dimensions shrink with successive technologies, the increasing mask manufacturing and inspection costs have increased. All this has made aggressive OPC techniques a very expensive option [3], not usable in normal ASIC designs.

4. Proposed Solutions

The possible solutions exist in four different areas – manufacturing, OPC, design, and transistor modeling. Improved lithography is the most obvious solution, but there is no technique immediately in sight. Optical proximity correction (OPC) techniques like creating notches and serifs can reduce flaring but are limited by mask manufacturing and inspection issues. Using non-patternable geometries for intensity modulation might be another useful OPC technique but involves the danger of over-etching and creating open-circuits. Due to these manufacturing and OPC limitations, we suggest a combination of design and modeling techniques in the following subsections.

4.1 Designing out flares– A solution for high performance applications

In very high performance designs timing accuracy is critical and usually die area is of secondary importance. Hence some area can be sacrificed to adhere to a set of layout guidelines or spacing rules that can help avoid significant transistor flaring. In this section we describe a methodology to determine such guidelines/rules.

Since flaring is caused where there is a poly turn near diffusion of the transistor, it would be desirable to avoid such a situation. This could be done by-

(A.) Avoiding poly routes connecting to transistor gate poly at right angles. This is illustrated in Fig4. Here significant reduction in flaring has been achieved by using metal route instead of poly for connecting gates of parallel transistors.

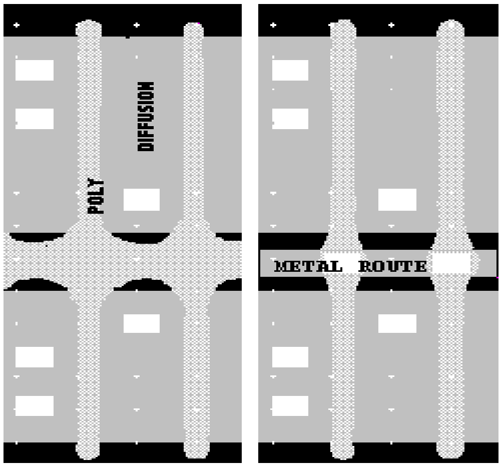


Fig 4. A: Original (Poly routed) B: Improved (Metal routed)

(B.) If the routing resources or other constraints within the cell-layout do not permit solution A, use a larger spacing (“d”) between the poly-turn and the diffusion of the transistor as shown in Fig5. Here the transistor is formed away from the flared region of the poly, and this results in a uniform channel length that is close to the designed channel length. (This is to be compared with small “d” and large flaring seen in Fig. 1)

It is important to come up with a suitable value of the poly-turn to diffusion distance (marked “d” in Fig 5), so that there is significant reduction in flaring at a minimal area cost. The rest of this subsection describes our methodology for finding the optimal value of “d”. We start with the assumption that the spacing “d” should be the minimum that will allow for a delay degradation of not more than (say) 5% over that of an unflared transistor.

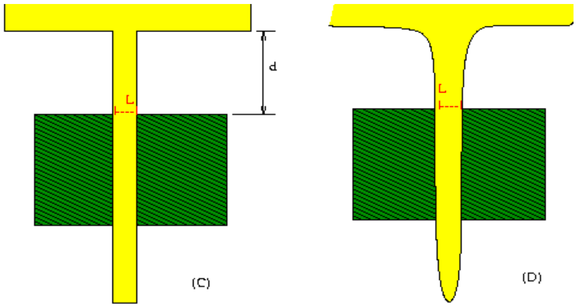


Fig 5. A: Drawn B: Silicon

The steps that were followed to get this value are as follows-

1. Start with layouts of a number transistors having different widths (W), and different poly turn to diffusion spacing (d).

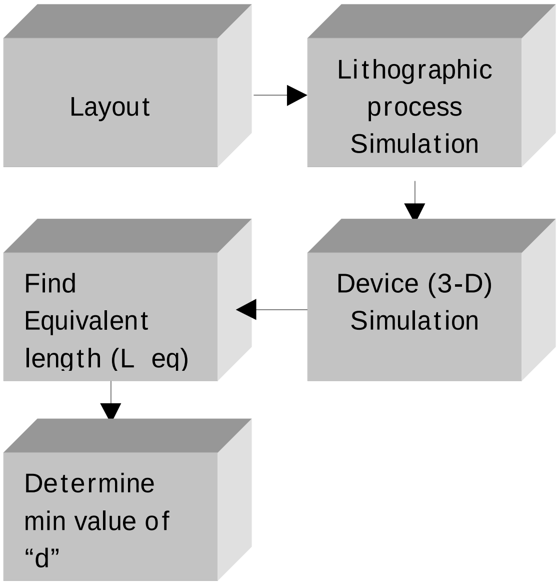

Layout
Lithographic process Simulation
Device (3-D) Simulation
Find Equivalent length (L_eq)
Determine min value of “d”

Fig 6. Methodology to find the minimum value of Poly-turn to-diffusion spacing “d”.

2. Run lithographic process simulations to find the expected on-silicon shapes of the transistors. (We used process simulation tool Prospector form Avant!). Since the electrical impact of flaring depends on misalignment, process simulations need to be run assuming the worst-case misalignment, in order to obtain the worst-case electrical performance models.
3. Take measurements of the flared channel geometries from the process simulation results.
4. Create corresponding geometries in a 3-D device simulator (Da Vinci) [4] and find the drain current at a given bias condition.
5. Find-out the channel-length at which an unflared transistor gives the same current as the current measured for the flared transistor in previous step. This length would then be the equivalent channel length (L_eq) for the flared transistor. This means that an unflared transistor of length L_eq can replace the given flared transistor for the purpose of simulation.
6. Using an inverter circuit constructed from transistors similar in size to the transistor under consideration, find the variation of delay with channel length. Since equivalent channel length (L_eq) has previously been correlated to the distance “d”, we can now plot a curve mapping the delay to d. we already have mapping of the transistors (The curve in Fig 7).

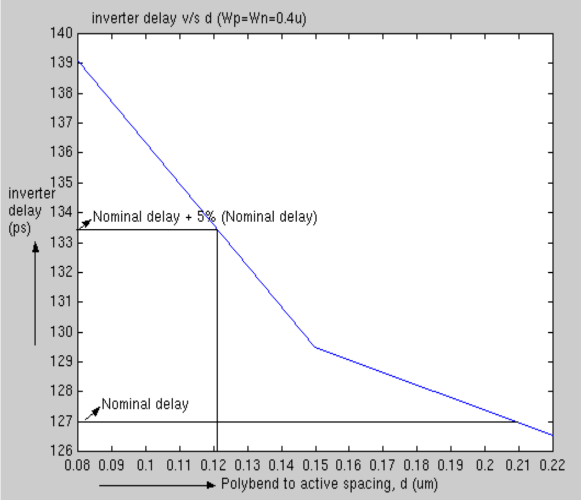


Fig 7 : Delay variation as function of Poly-turn to moat spacing for an inverter.

7. From the graph, find the value of “d” corresponding to 5% delay degradation from the delay of an unflared nominal transistor. The value thus obtained is the minimum spacing guideline for the poly-turn to diffusion distance.

Example – Fig 7 shows the case for a 0.4-micron wide transistor in 95nm drawn gate length technology. From the curve, corresponding to a 5% delay degradation we get a poly-bend-to-diffusion spacing of 0.12microns.

Thus we have demonstrated methodology for finding a minimum poly-turn to diffusion spacing for any given technology and transistor size. One example of effectiveness of the layout guidelines and methodology proposed here is the correction for flares in cell-layouts from a high performance 0.095µm production cell-library . An analysis similar to the one described above showed that a poly-turn to diffusion spacing of 250nm would sufficiently reduce flaring on transistors. An example is shown in Fig 8.

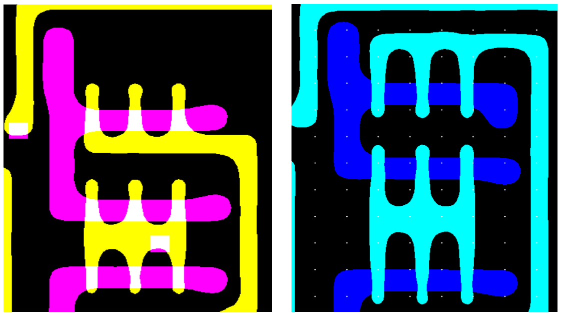


Fig 8 A: Original B: Improved

Here, both the poly and diffusion have been reshaped, maintaining the same transistor sizes, but increasing the poly–turn to diffusion spacing. The reduction in flaring and the straightening of the transistor channels is visible from lithographic-process simulation results shown in Fig 8. The curved regions of poly have moved away from the transistor channel region after layout modification.

In simple cells the propagation delay is seen to vary as quadratic function of the average channel length. This function can be determined by curve fitting from the available data and then used heuristically as a figure of merit (FOM) for the propagation delay. Fig 9 illustrates the overall improvement in the delay FOM variance caused by the modifications made in the layouts of the worst-hit library cells. The horizontal axis represents the variation in performance (delay FOM) caused by flaring as a percentage of the nominal delay FOM. The plots in the figure show frequency distribution of the variation in delay FOM of transistors, for two libraries – ‘initial’ and ‘pass2’. The initial library (before layout modifications, red curve in the picture), has a long “tail” of very highly flared transistors. This tail is eliminated after the corrections (pass2, blue curve). This shows that the layout verification on worst hit cells has been very effective in reducing the performance variation due to flaring.

4.2 Modeling flares– A solution for high-density applications.

In high-density applications it is not practical to

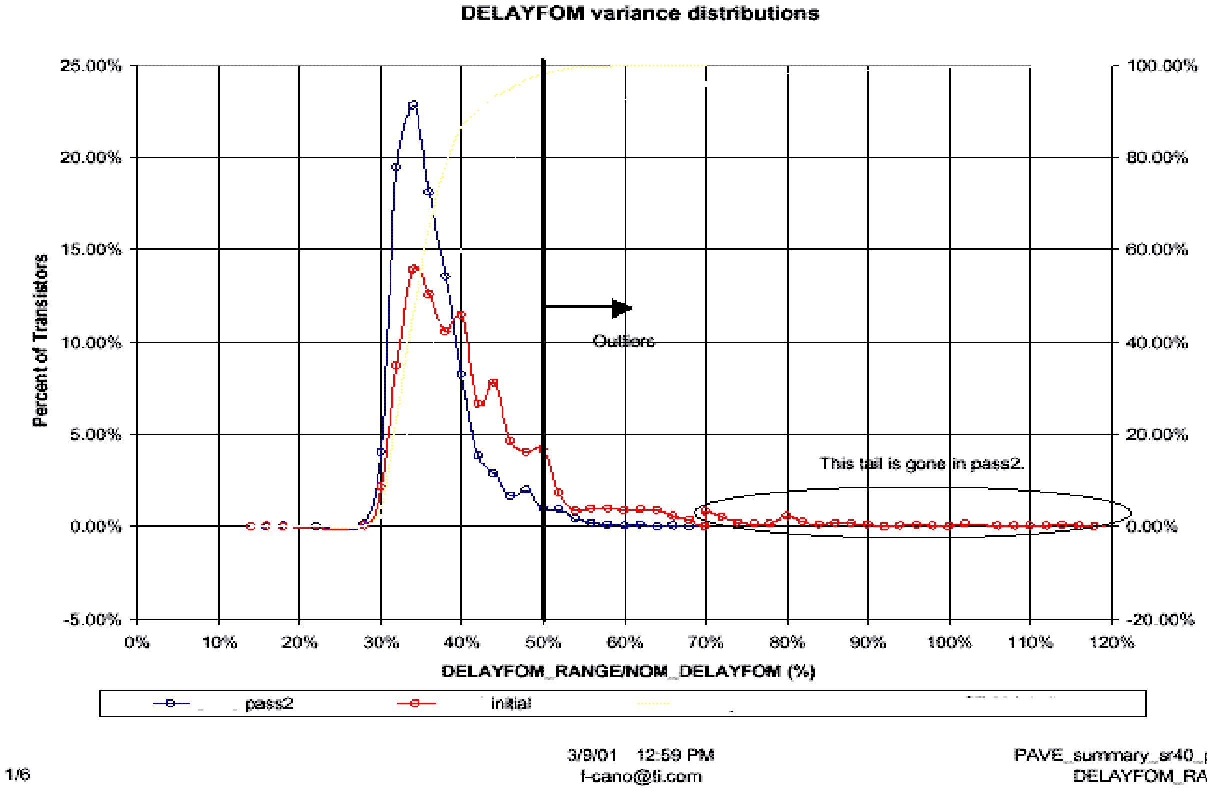


Fig 9: Impact of flare-cleanup in an ASIC library

conform to a large poly-turn to diffusion spacing rule since the area hit could be unacceptably large. Also, using the guideline about avoiding poly-routes may not be possible, since the layouts of smaller standard cells tend to get metal-routing congested, and poly-routes have to be used frequently in order to complete the cell-internal routing. Typically a highly density-oriented design is not targeted at very high frequencies, hence the performance loss due to flaring might be acceptable. However, it would still be needed to know how much maximum delay change is possible on each cell so that chip level timing constraints can be analyzed and the designer can be sure of setup/hold time being met. Hence the uncertainty introduced in the delays (or setup/hold times) of cells due to flaring is a bigger concern here than the performance loss. A solution acceptable for high-density designs would be to let the flaring be, if the L_eq of each flared transistor can be determined at the time of netlist extraction from layout. In this solution, transistors would be allowed to remain flared, but the extraction tool would populate the (spice) netlist with L_eq instead of the nominal channel length. Hence the circuit simulation would represent the on-silicon behavior more accurately.

For this solution to be feasible it should be possible for the SPICE-netlist extractor to determine the equivalent length of the transistor in terms of the drawn geometries in and around the transistor. We know that flaring decreases as spacing of the poly-turn from the diffusion ("d") increases. Also, transistors with very small width (W) are most susceptible to flaring effects since the flared region forms a greater fraction of the total channel width. We propose that L_eq be modeled as a function of W and d. The rest of this subsection discusses this type of modeling. (We have not considered the drawn channel length L as a variable here, assuming that minimum allowable length has been used throughout).

The initial steps to be followed in modeling the L_eq of the flared transistor are the same as steps 1-5 in section 4.1 above. We get L_eq vs. "d" plots as shown in Fig 11. The plot shown is for a transistor with designed channel length of 0.15 microns. For small values of d the curve may not be monotonic because of variations in OPC with d.

Having obtained the L_eq data, the next step is to model the L_eq as a function of spacing "d" and width "W". This function can be obtained by curve-fitting using values of L_eq for different values of d and W. However, our studies revealed that the variations in OPC make it very difficult to get a single accurate function for all ranges of d and W. One solution to this problem is to use different functions for different regions in the (d, W) space. We have done this by binning transistors into different width-ranges. For each range of widths, L_eq is modeled as a function of "d" the poly-turn to gate distance.

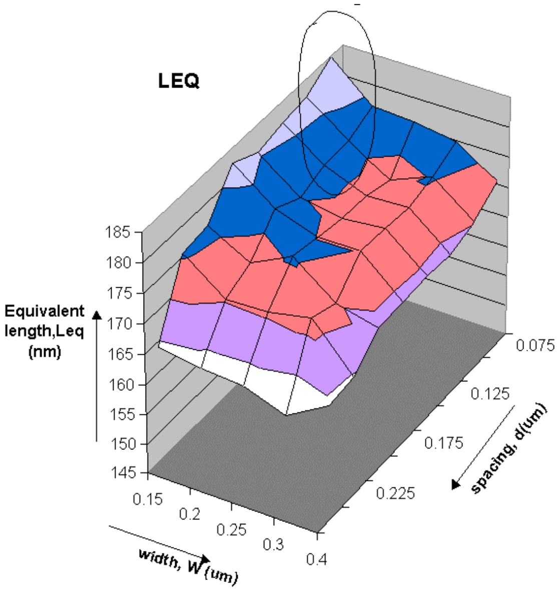


Fig 10. Variation of L_eq with poly-turn to diffusion spacing and transistor width.

Fig 11 below shows an example of modeling L_eq as a function of "d" for transistor channel width of ~0.2um. The designed value of channel length is 150nm. The L_eq data is obtained through DaVinci simulations as described earlier, and then a polynomial function is computed that gives the best fit for the points.

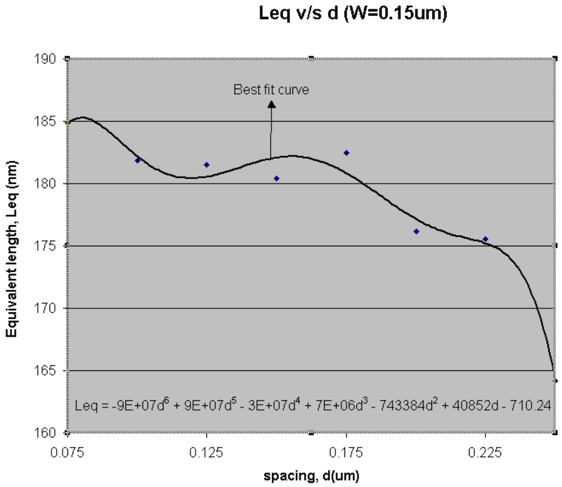


Fig. 11 Curve fitting to get L_eq as a polynomial function of 'd' at a given width W.

L_eq is now a known function of poly-turn to diffusion spacing (d) for a given width (W). This function is valid for all transistors in all cells. Thus we have demonstrated a complete flow from optical/process simulation, through 3-D device simulation, to electrical simulation, for modeling the performance of the flared transistor in terms of simple layout parameters like channel width and poly-turn spacing from diffusion. The advantage of this type of modeling flow is that it has to be run only once per technology, and not for every circuit/layout. Given any specific layout, since d and W are readily extracted from the layout, it is possible for the extraction tool to compute L_eq for each flared transistor and use it in creating the (SPICE) netlist for the circuit.

5. Failure/Yield impact

The timing-inaccuracy caused by flaring can lead to functional failures. In this section we have made an attempt to get a rough estimate of the yield impact these failures can lead to. For this we have taken the help of statistical simulation.

A study of cell-libraries reveals that typically the flip-flops are the most sensitive to flaring, particularly in terms of setup and hold time variations. Also they tend have transistors with widths close to the process min width. Transistors with very small widths are most susceptible to flaring effects since the flared region forms a greater fraction of the total channel width. Hence by determining the impact of flaring on flip-flop performance we can get a worst-case estimate of the overall impact of flaring. The test bench consists of a flip-flop with data signal changing just with in the nominal setup (or hold) time of the flip-flop. The small transistors in the flip-flop (whose lengths are more susceptible to variation due to flaring) are identified. The lengths of these transistors are then varied to mimic transistor flaring. A statistical simulation tool - Circuit-surfer [5] is used to vary the lengths of transistors. The statistical simulation tools does a monte-carlo analysis by randomly varying the lengths of the specified transistors within a gaussian distribution, and running SPICE. In each SPICE simulation, the output is monitored to check if the new data is registered. Simulations which have no transitions at the output and those which result in unacceptable clock to output delay are treated as simulations which fail to meet the nominal setup time of the flip-flop.

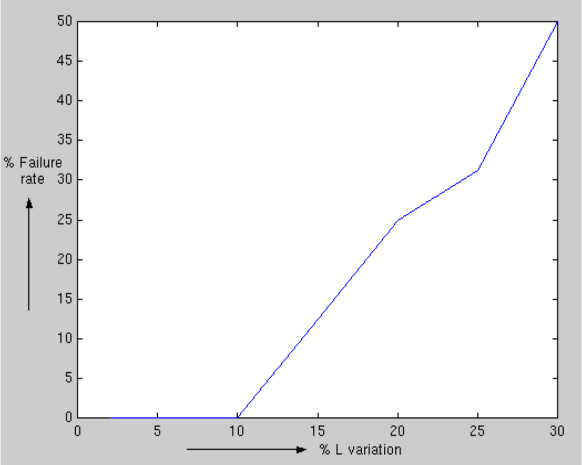


Fig. 12: Failure(%) as function of % variation in L

For a given percentage variation in transistor length L, the failure rate (failure rate = no. of simulations with setup or hold time failure / total no. of simulations) is recorded. Fig 12 shows variation in failure rate with percentage variation in transistor length. This analysis reveals that any variation in transistor length in excess of 10% can manifest as setup time failure in flip-flops. Our studies on L_eq show that upto 13% variation in L due to flaring is common. From the graph in Fig 12, this would translate into about 4% failure rate. The possibility of such a high failure rate highlights the significance of taking care of flares by either designing them out, or modeling their effects.

6. Conclusion :

In this paper we have proposed two types of solutions for transistor flaring related issues. For the high performance designs we have recommended that the layout rules/guidelines be modified to avoid flaring. We have demonstrated a method to determine the optimal poly-turn to diffusion spacing that is effective in avoiding flaring and has minimal area impact. For the high density designs we have recommended retaining flared transistors but getting better timing accuracy though a transistor modeling and circuit extraction based approach, and have demonstrated a new methodology/flow for implementing this. We have shown through statistical simulation that the flaring could cause upto 4% yield loss if not taken care of by using modeling or design techniques. The solutions described in this paper have been used on our standard cell libraries with significant improvement in functional yield.

7. Acknowledgements:

The authors are grateful to Carl Vickery and Clive Bittlestone for bringing the transistor flaring issue and its implications to our notice, and initiating the search for a solution. Thanks to Frank Cano for running Prospector and generating the delay FOM plots on the layouts modified by us for flaring-reduction. Thanks to Dibyendu Goswami for suggesting metal routes instead of poly routes as a step towards reducing flares.

8. References :

[1] N.Cobb, A.Zakhor, M.Reihani, F.Jahansooz, V.Raghavan : Experimental results on Optical Proximity Correction with variable threshold resist model. Proceedings of the SPIE, vol 3051 pp 458-468

[2] F.M. Schellenberg, Pat LaCour : Implementation issues for production OPC.
http://www.mentorg.com/dsm/tpapers/production_opc.pdf

[3] B.Bruggeman, Y.S.Lin, et.al : Micrography cost analysis.
http://www.mentorg.com/dsm/tpapers/micro_cost_analysis.pdf

[4] Davinci 3-D device simulator
http://www.avanticorp.com/Avant!/SolutionsProducts/Products/Item/1,1500,40,00.html

[5] Circuit Surfer performance yield estimator
http://www.pdf.com/services_tech.phtml