Joint-Optimization for SRAM and Logic for 28nm node and below

Staf Verhaegen[*][a], Michael C. Smayling[b], Peter De Bisschop[a], Bart Laenens[a]
[a]imec, Kapeldreef 75, Leuven, Belgium B-3001
[b]Tela Innovations, Inc., 655 Technology Pkwy., Suite 150, Campbell, CA, USA 95008

ABSTRACT

In current and next generation nodes lithography is pushed to low k_1 lithography imaging regimes. A gridded design approach with lines and cuts has previously been shown to allow optimizing illuminator conditions for critical layers in logic designs.[1] The approach has shown good pattern fidelity and is expected to be scalable to the 7nm logic node. [2]

A regular pattern for logic makes the optimization problem straightforward if only standard cells are used in a chip.[3,4] However, modern SOC's include large amounts of SRAM as well. The proposed approach truly optimizes both, instead of the conventional approach of sacrificing the SRAM because of logic layouts with bends and multiple pitches.

The biggest problem in co-optimizing logic cells and SRAM bit cells is the orientation of critical layers. For SRAMs, the gate and metal1 layers have lines in parallel directions, while in standard cells they are perpendicular. This would require abandoning dipole illumination for the combined optimization, and at best using some form of quadrupole.

The alternative is to design the logic and SRAMs to be unified from the beginning. In this case, critical layer orientations as well as pitches could be matched and each of the layers optimized for both functional sets of patterns. Choices of patterns can be made to achieve DSMO (Design-Source-Mask-Optimization).

In the 28nm to 22nm logic nodes – with contacted pitches from 110nm to 90nm and metal1 pitches from 90nm to 70nm – one of the questions to answer is when and for which layers double patterning is needed. The limit of single patterning immersion lithography can only be explored through a smart combination of restricted designs and powerful source-mask optimization tools. In this paper a 28nm SRAM block with bit and word line periphery will be used to look at choices for Design-Source-Mask-Optimization.

Keywords: Low k_1, gridded design rules, restricted design rules, Design-Source-Mask Optimization

1. INTRODUCTION

Although immersion scanners extended the resolution limits of optical lithography [5], no improvement in fundamental resolution has been available since 1.35 NA immersion scanners were introduced in 2007. With λ/NA = 143nm, the Rayleigh equation CD = k_1 λ/NA requires that further improvements in resolution depend on k_1. This will require RET (resolution enhancement techniques) such as OAI (off-axis illumination), OPC (optical proximity correction), and a-PSM (attenuated phase shift masks). SMO (source-mask optimization) is being used to co-optimize the lithography process conditions such as illuminator shape and OPC treatment.

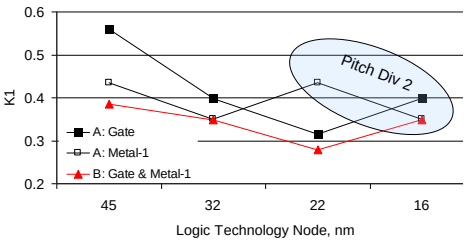


Figure 1: k_1 trend for sub-65nm logic technology nodes

k_1 has been decreasing for recent logic technology node as shown in Figure 1. To maintain pattern fidelity at k_1 values below ~0.6, resolution enhancement techniques (RET) have been introduced. Two approaches to scaling are shown in Figure 1. The "A" path has a tighter metal1 pitch and a relaxed gate Pitch compared to path "B," which has the same pitch for both layers. Pitch division is needed for the "A" scaling path metal1 at 22nm, and for both gate and metal1 for both "A" and "B" scaling paths at 16nm.

[*] staf.verhaegen@imec.be; phone +32 16 281783

As k_1 decreases, "practical limits" are imposed by the design style.[6] 2D layouts with bent polygons are limited to ~0.35. A 1D layout style with parallel straight lines looking much like a grating pattern, and has a limit of ~0.28. Extensive efforts are being made to define "restricted design rules" which allow bends but with constraints on widths or spaces.[7] A 1D layout style with further requirements for keeping lines on a regular grid permits using a simplified set of design rules described as "gridded design rules" (GDR).[8] 1D-GDR using SDP (spacer double patterning) to pitch-divide the lines and a conventional cut mask has been used to demonstrate 16nm logic patterns.[9]

2. METHODOLOGY, WORKFLOW AND TOOLS

2.1 Test case: SRAM block with periphery

The SRAM block used for the simulations is shown in Figure 2. The block includes word-line drivers, bit cells, and column multiplexers. Metal1 fill is not shown. The metal1 pitches are identical in each of the functional parts of the block. Standard cells use the same metal1 and gate pitches.

The bit cell array has one row of dummy cells at the bottom, just above the word line drivers. There are two columns of dummy cells to the left of the array, adjacent to the bit-line multiplexers.

For a contacted gate pitch of 110nm and a metal1 pitch of 90nm, the bit cell size was 0.119μm^2. The size of the cell is 2 times the gate pitch in horizontal direction and 6 times the metal1 pitch in the vertical direction. This height is partly determined by the extra space that is foreseen between the diffusion areas of the n-MOS and the p-MOS transistors for formation of the n- and p-wells. If a SOI technique would be used this extra space may not be needed and the height of the cell could be reduced to 5 times the metal1 pitch. Of course then also the layout of the bit-line multiplexers may need to be adapted. It is a topic for further investigation if the height of this block can be reduced also to 5 times metal1 pitch or if it has to be laid out with 10 times metal1 height and two cells next to each other to drive the bit lines of every two rows in the array.

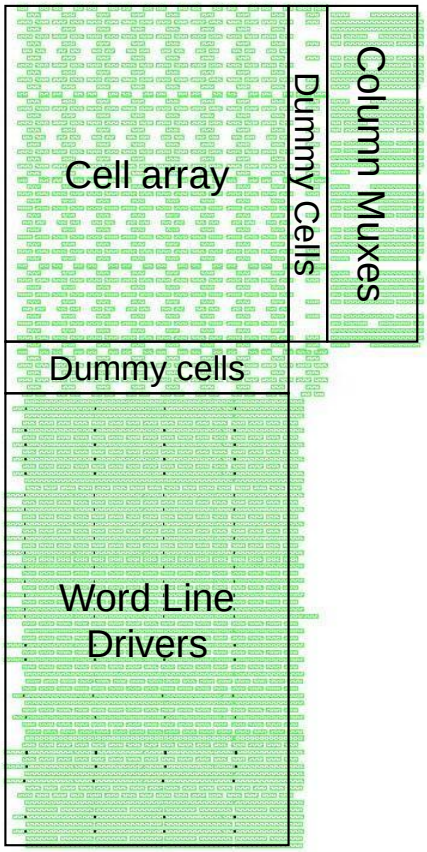


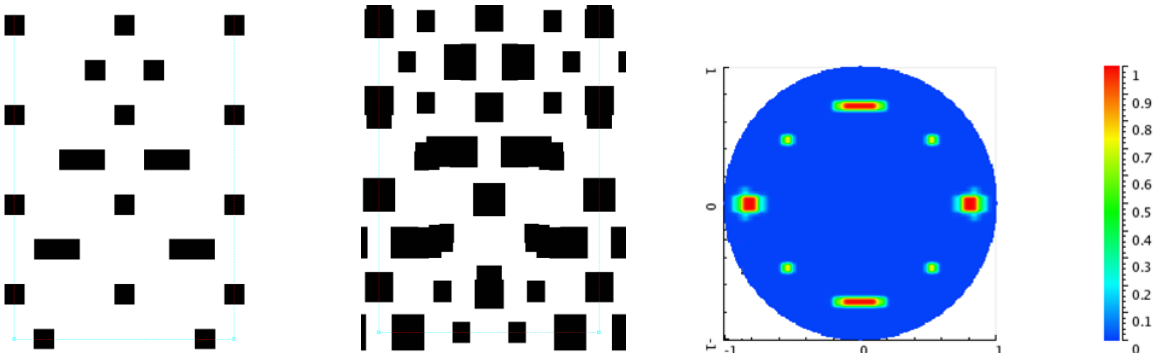
Figure 2: SRAM block layout

5 critical layers were investigated during the optimization flow in order to not create a problem in one of the layers when making design changes in another layer. The layers that were investigated are:

- active layer; often also called the diffusion layer
- gate
- local interconnect where applicable; in our case it is only used for connecting active area regions of the source and/or drains of n- and p-MOS transistors.
- contact
- metal1

2.2 Source-mask optimization

To get the best printing performance the mask patterns and the illumination source shape have to be optimized concurrently. In this paper the Tachyon SMO software from Brion Technologies Inc has been used for this purpose. As input you provide a target layout (Figure 3a) and a litho process and as output you get an optimized mask layout (Figure 3b) and a corresponding illuminator shape (Figure 3c). The optimized illumination shape can either be a pixelated source as shown in the figure or be a standard diffractive optical element (DOE). In [9] you will find an overview of the latest results achieved with this technique.



a) input contact hole layout

b) output mask layout (OPC and sub-resolution assists)

c) output illuminator shape

Figure 3: Example input and outputs of Tachyon SMO flow

2.3 Design style and rules

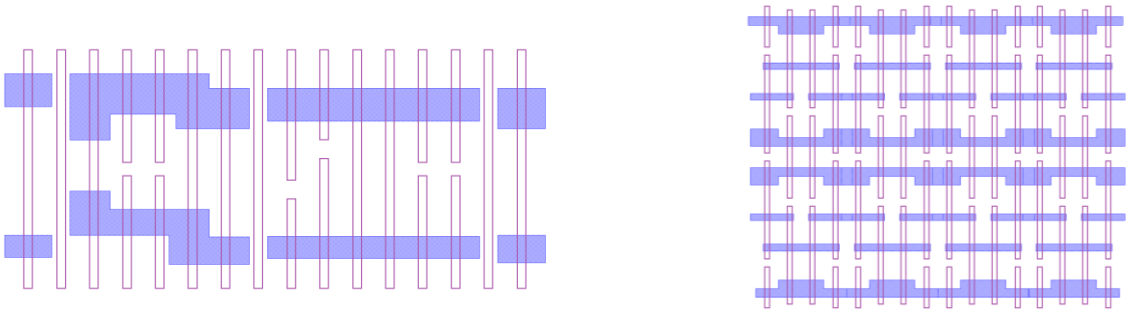
2.3.1 Conventional approach

A typical SRAM cell is laid out with the gate and metal1 features in roughly the same direction, and metal2 perpendicular to the gate electrodes. These layers in a conventional logic standard cell usually are not drawn like this, and in fact include bends and jogs.

Because of the 2D nature of the standard cell layouts, the illuminator choices for those layers are limited to annular or at best quadrupole configurations. This means that the typical SMO is often constrained before the optimization even begins. New, very flexible illuminators are underutilized for random 2D patterns.

2.3.2 Unified layout approach

The layout for a typical standard cell using Gridded Design rules is shown in Figure 4a. The gate electrodes are oriented in the vertical direction, with active regions running horizontally. Figure 2b shows a group of SRAM bit cells designed to be compatible with the logic cell. The gate orientation and pitch are the same for both the standard cell and the SRAM bit cell. The SRAM periphery is also constructed with the same pitch and orientation.



a) Standard cell active and gate layers.

b) SRAM active and gate layers.

Figure 4: Unified layout; example active and gate layer layouts

Both the standard cells and the SRAM bitcell and periphery can be laid out using a local interconnect layer. The benefits for standard cells are discussed further elsewhere in this conference.[11]

2.3.3 Design restrictions

Even with a unified layout and aggressive lithography optimization sometimes hot spots may remain in the design. Re-laying-out the design so that these hot spots are avoided is one of the possible solutions. In the TELA approach only known-good topologies are used for generating layouts from netlists for logic [12]. This approach allows thus to remove

topologies that are limiting the lithography process window. Ideally this reduction of used topologies only has minimal impact on the area of the laid-out design. The art of Design-Source-Mask optimization then consists of finding the right set of topologies so that an optimal trade-off can be found between area and lithography process window.

2.4 Negative tone lithography process

Negative tone lithography processes can give better lithography performance for certain applications. In imec a Fujifilm process with a solvent based developer is investigated to look at performance of negative tone lithography processes. An overview of results is presented for this process in [13, 14]. This process is taken into account in our study by looking at an optimization on both light field and dark field masks for the poly cut, the contact and the metal1 layer.

2.5 Lithography performance validation metrics

For validating and comparing different solutions several metrics have been used in this study. A first metric are classic process windows using Bossung curves and the corresponding MEEF value. For this validation several cut lines have been defined in the layout (Figure 5). The SPEC put on the cut lines is 10% of the target CD value. The result of this metric is thus also dependent on the choice of cut lines.

Another metric used in the flow are so-called PV-bands (Figure 6). These are bands that indicate the outer and inner boundary of all printed contours through process window. In this study these bands have mostly been used for qualitative verification; e.g. to find hot spots with big PV-band. Superposition of PV-bands of different layers is also used to find regions that would be sensitive to overlay variations; again no thorough quantitative validation has been done here.

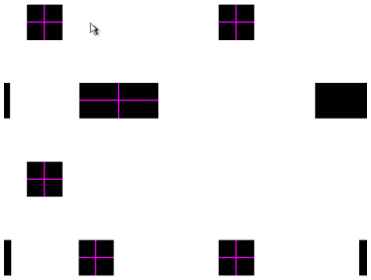


Figure 5: Cut line placement

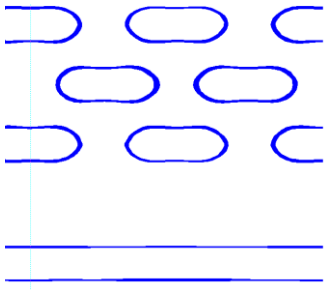


Figure 6: PV-bands

Currently an extension of the metrics is being investigated to correlate better the lithography performance with electrical performance. This includes translating litho printing contours to effective L and W of transistors in a netlist to simulate the circuit performance. The yield of the connection formed by two overlapping layers like contact to metal1 or gate is correlated to the area of the overlap of these two layers. If one would be able to compute the overlapping area of two layers as a function of dose, focus, mask CD and overlay error, a yield related metric could be introduced.

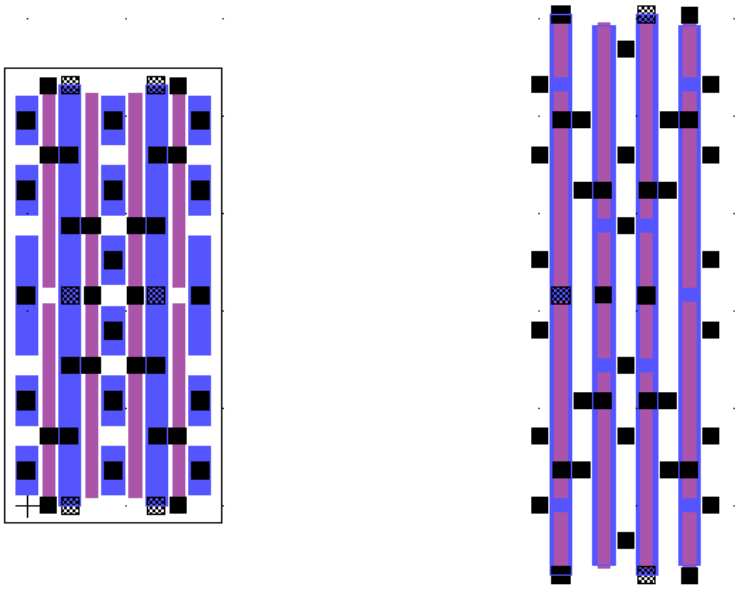
3. SRAM BLOCK AND PERIPHERY LAYOUT VARIATIONS

In this chapter an overview will be given of all the steps that have been performed on the SRAM block layout to optimize the area and the lithography performance.

3.1 Unified SRAM and periphery layout and metal2-gate phasing

At advanced technology nodes, the contacted gate pitch and metal2 pitch are not necessarily related by a simple ratio of integers. For example, at 90nm, an efficient ratio was 3:2, since the metal2 pitch could be much smaller than the gate pitch. In the SRAM work presented here, the ratio could be 11:9 if we chose the metal2 pitch to be the same as the metal1 pitch. However, this kind of ratio is not efficient for SRAM array design which needs regular patterns within the cell array. For a unified design, a 1:1 ratio was chosen for the bit-cells, the SRAM periphery and the logic cells.

Once the gate / metal2 ratio is set, then the phasing between the two layers needs to be considered. Figure 7 shows examples of two bit-cell mini-arrays. The 8-track cells on the right side have gate and metal2 “in-phase” or co-linear. The 6-track cells on the left side have gate and metal2 180 degrees out of phase, with metal2 lines running parallel to gate but centered between gate lines.



a) 6 metal1 tracks high cell with out-of-phase gate-metal2 patterns

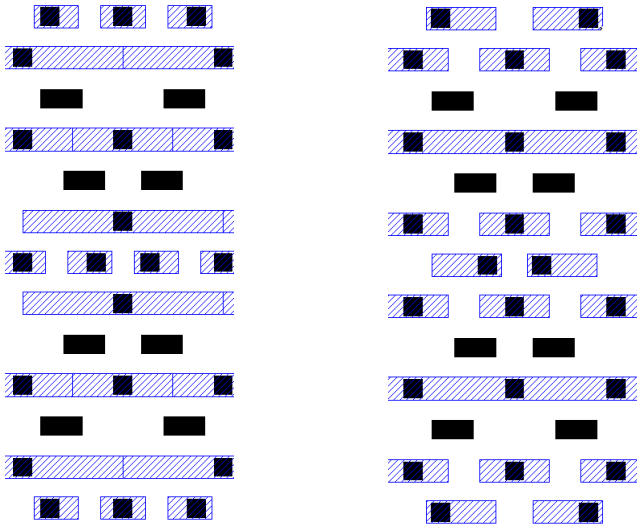
b) 8 metal1 tracks high cell with in-phase gate-metal2 patterns

Figure 7: Example SRAM cell layouts with changing gate-metal2 phasing

The 8-track cell uses extra metal1 lines for horizontal global connections, so metal2 resources are not needed for feed-throughs to upper metal layers. The vertical word line connections to alternate gate electrodes can be done with alternating stacked vias and metal1 jumpers. To achieve smaller cell area, as in the 6-track cell, global routing was pushed up to metal3 and metal4. Because of this, more metal2 feed-throughs are needed, and putting metal2 over the diffusion contacts allows stacked vias. Having metal2 centered between the gates provides the best combination.

3.2 SRAM bit cell layout variation for ground connection

As said in the introduction a design of the SRAM with and without local interconnect was investigated. But even then different design layout variations are possible in the SRAM cell and we will take here the design with local interconnect as example. In Figure 8 the contact and metal1 layers are given for different layouts in the ground connection. On the left you find a case where bit lines are in a continuous line and the ground connection is in between the word line connection. On the right you find a design where the metal1 bit line connection is cut in pieces and the ground connection is in between them; the latter thus needs to move the bit line connection up to a higher level of metal.



a) Variation 1 b) Variation 2

Figure 8: Contact (black) and metal1 (blue) layout for two variations on the ground connection layout using local interconnect

In Table 1 the worst process window numbers are given over different cut lines drawn in the design. This also means that the worst EL number may come from another cut line than the worst DOF or the worst MEEF value. Numbers are given for both variations of the contact and metal1 layer and for both light field (LF) and dark field (DF) imaging of the patterns. For numbers below or above a certain threshold, the cell background is made gray; this is 6% for EL, 100nm for DOF and 4X for MEEF.

For the contact layer, MEEF is one of the important parameters and here only the light field version for variation 2 of the design has a MEEF value in SPEC. For both cases the light field imaging seems to perform better than the dark field imaging. For the metal1 layer the picture is less clear. For variation 1 of the design there is a spot in the design where the dark field case is performing better than light field for a spot in the design. For variation 2 of the design it is less clear which is the preferred solution. The light field case has better MEEF and DOF and the dark field case has better EL. In the end variation 2 of the design was chosen due to MEEF requirements for the contact hole layer. For contact holes light field imaging is the clear winner and for metal1 also light field imaging was chosen mainly due to lower MEEF value but the DOF is also much bigger.

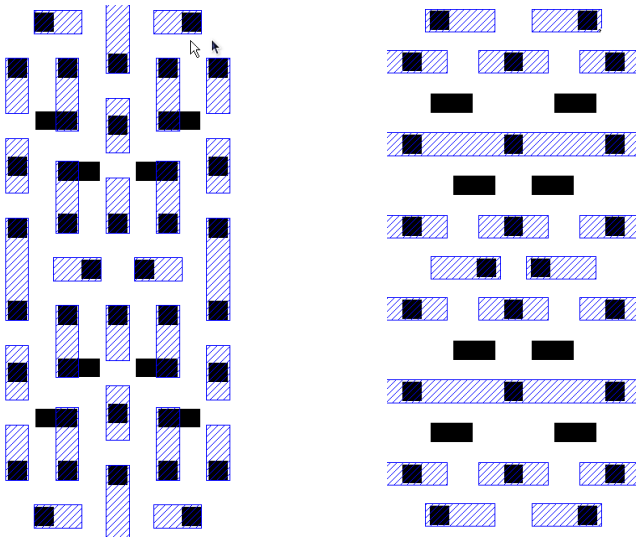
	Contact				**Metal1**			
	Var1		**Var2**		**Var1**		**Var2**	
	LF	**DF**	**LF**	**DF**	**LF**	**DF**	**LF**	**DF**
Worst EL [%]	10.4	4.0	9.8	7.9	5.1	7.0	7.4	8.0
Worst DOF@6%EL [nm]	110	-	132	88	-	156	192	131
Worst MEEF[X]	4.2	6.6	3.7	5.2	4.2	3.8	4.4	4.9

Table 1: Process windows for contact and metal1 layer for two variations in layout for ground connection (LF: light field imaging, DF: dark field imaging, Var1: Figure 8a, Var2: Figure 8b)

3.3 With and without local interconnect for SRAM bit cell

In this paragraph the design of the SRAM bit cell with and without local interconnects will be compared. In Figure 9 you can see the difference in the design without local interconnects on the left and with local interconnects on the right for

the contact and metal1 layers. The design without local interconnects needs both horizontal and vertical trenches on the metal1 layer and the one with local interconnects has only horizontal trenches.



a) Without local interconnect

b) With local interconnect

Figure 9: Contact (black) and metal1 (blue) layer for SRAM cell with and without local interconnect

In Table 2 the worst metrics are given for both designs. Here always light field imaging is used. For both contact and metal1 the design with local interconnects performs better than the design without local interconnects. For the rest of this paper the design with local interconnects was used. The interconnect layer itself has huge process windows and acceptable MEEF values even when using standard illumination shapes and will not limit the scaling.

	Contact		**Metal1**	
	No LI	**LI**	**No LI**	**LI**
Worst EL [%]	7.5	9.8	7.5	7.4
Worst DOF@6%EL [nm]	116	132	150	192
Worst MEEF[X]	4.6	3.7	5	4.4

Table 2: Process windows for contact and metal1 layer for layout with and without local interconnect layer

3.4 Gate layer performance

The gate layer can be patterned in one step where both lines and butting line-ends have to printed. Alternatively it can be printed using a double patterning technique; in a first step long lines are printed and in a second step trenches are printed to open the gaps in the lines to form the butting line-ends. In Table 3 an overview is made of the process window metrics for both cases. During optimization a pixelated source was used for the single patterning case and standard source shapes for the double patterning case. For the single patterning case there is a high MEEF value; this is caused by a high sensitivity to mask variations for printing the butting line-ends. One could live with this high value as long as there is no bridging of the line-end through process windows and the CD control is acceptable where the gate is crossing over the active area; e.g. where a transistor is formed. In Figure 10 the PV-

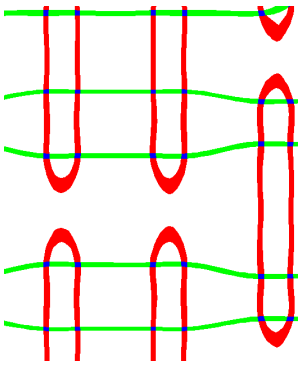


Figure 10: PV-bands of active (green) and gate layer (red)

bands for optimized single patterning is given; for the bands +/-55nm focus, +/- 3% dose and +/- 1nm maskCD(1X) is used. Due to the rounding of the gate patterning at the line-end the transistor performance will be highly influenced by overlay errors between active and gate layer. For the single patterning case the overlay has to be controlled very strictly or the design may have to be relaxed somewhat. For the double patterning case using a cut mask technique huge process windows and low MEEF values are present. Especially the improved EL and lower MEEF will make the control of the transistor length much better. Therefore the double patterning case is the preferred technique.

	SP	DP	
	Gate	Lines	Cut mask
Worst EL [%]	7.8	18.5	22
Worst DOF@6%EL [nm]	114	278	220
Worst MEEF[X]	5.5	1	2.7

Table 3: Process windows gate layer with single and double patterning; pixelated source is used for single patterning and standard DOEs for double patterning

3.5 SRAM periphery layout optimization

Even when using compatible unidirectional layout for the periphery of the SRAM block hot spots may be present. To investigate this we took our preferred SRAM cell design as in Figure 9b with the optimized source for each of the layers. Using this illumination condition we applied a mask-only optimization on the periphery and looked at the process window metric for a whole bunch of cut lines to see if there are any hot spots. In Figure 11 the histograms are shown for the contact hole periphery. The numbers are distributed around the worst values that were found earlier for the SRAM cell (the second column of Table 2). The worst numbers in these histograms reduce the process window somewhat more: EL from 9.8% to 8.5%, DOF from 132nm to 98nm and MEEF from 3.7 to 5.5. If this higher MEEF value is not acceptable one could investigate if a redesign can solve the problem; alternatively an additional source-mask optimization could be performed with the periphery hot spots included. After inspecting the PV-bands the derived results were considered acceptable for the found hot spots.

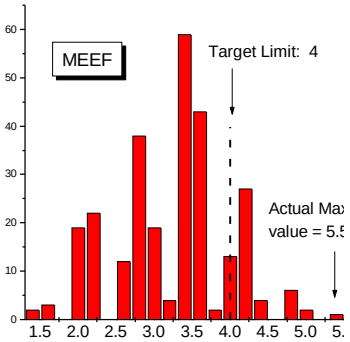
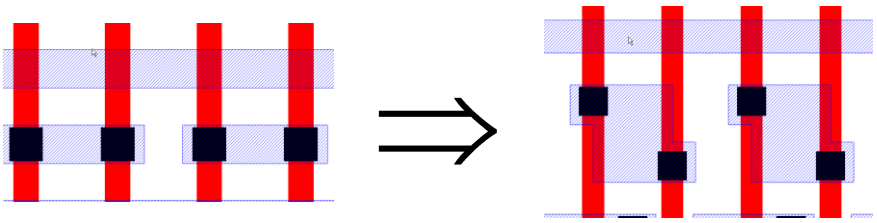
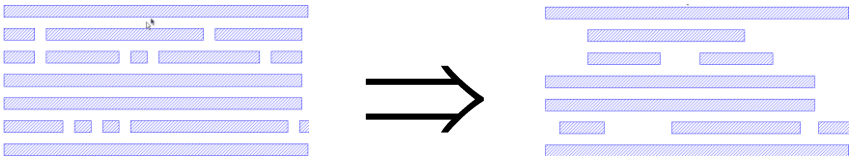


Figure 11: Contact hole periphery process window metric histograms

For the metal1 layer a strict 1D layout has advantages, for example allowing straightforward splitting for multiple patterning. However, for the case in which the patterning can be done with a single exposure, there are instances in which using short, wide pieces of metal1 can avoid the use of small gaps in metal1, extra vias and metal2. Some adjustment of the length and end-spacing can give a pattern which has good printability using the optical settings optimized for the rest of the SRAM block. In Figure 12 an example is given of the design with small gaps and trenches present together with some re-laid out versions using either some rerouting on other layers or use of bigger metal1 pads. With the design changes and the resulting improved process window numbers should allow to print the metal1 layer single patterning for our SRAM block. Analog to the contact hole layer also here spots were found with higher MEEF values but were not considered killers after inspecting the PV-bands.



a) Use of metal1 pad



b) reroute signals and remove dummy metal1 trenches

Figure 12: Example layout with small metal1 gaps and adapted layout

4. CONCLUSIONS

An SRAM designed to be compatible with logic standard cells was studied. Using advanced Design-Source-Mask optimization techniques solutions were found for all critical layers of a 28nm SRAM block including bit line and word line periphery. For the gate layer double patterning using a cut mask approach is preferred. The other layers can be done using single patterning if the right design restrictions are applied to the layout; MEEF and thus mask CD control was found to be one of the critical parameters to watch out for in this technology node.

ACKNOWLEDGEMENTS

This work would not have been possible without the support from a lot of people. Firstly the Tela Innovation staff that has been working on the layout of the SRAM block. Additionally support from Brion Technologies on their Tachyon SMO software and hardware; especially Paul Van Adrichem, Stephen Hsu and Ziphan Li were helpful to reach the presented results. Last but not least the paper would have not been possible without the imaging and processing expertise of the whole imec lithography group.

REFERENCES

[1] M. C. Smayling, V. Axelrad, "32nm and below Logic Patterning using Optimized Illumination and Double Patterning," Proc. of SPIE, vol. 7274-19 (2009).

[2] Y. Borodovsky, "Lithography 2009: Overview of Opportunities," SemiCon West (2009).

[3] S. Verhaegen, A. Nackaerts, V. Wiaux, E. Hendrickx, "Impact of lithography on the design of next generation logic cells," Semicon Japan (2006).

[4] M. C. Smayling, H. Y. Liu, L. Cai, "Low k_1 logic design using gridded design rules," Proc. of SPIE vol. 6925-68 (2008).

[5] B.J. Lin, "Immersion lithography and its impact on semiconductor manufacturing," Proc. of SPIE vol. 5377-3 (2004).

[6] W. Arnold, "Lithography for the 32nm Technology Node," IEDM 32nm Technology Short Course (2006).

[7] L. Capodieci, "From Optical Proximity Correction to Lithography-Driven Physical Design (1996-2006): 10 years of Resolution Enhancement Technology and the roadmap enablers for the next decade," Proc. of SPIE vol. 6154-1 (2006).

[8] M. C. Smayling, "Gridded Design Rules – 1-D Design Enables Scaling of CMOS Logic," Nanochip Technology Journal, vol. 6(2) (2008).
[9] C. Bencher, H. Dai, Y. Chen, "Gridded Design Rule Scaling: Taking the CPU toward the 16nm node," Proc. SPIE Microlithography, vol. 7274-14 (2009).
[10] J. Bekaert, B. Laenens, S. Verhaegen, L. Van Look, D. Trivkovica, F. Lazzarino, G. Vandenberghe, P. Van Adrichem, R. Socha, S. Baron, M.C. Tsai, K. Ning, S. Hsu, H.Y. Liu, M. Mulder, A. Bouma, E. van der Heijden, O. Mouraille, K. Schreel, J. Finders, M. Dusa, J. Zimmermann, P. Graüpner, J.T. Neumann, C. Hennerkes "Freeform illumination sources: An experimental study of source-mask optimization for 22 nm SRAM cells," Proc. of SPIE vol 7640-7 (2010).
[11] M. C. Smayling, Mircea Dusa, Robert J. Socha, "22nm Logic Lithography in the Presence of Local Interconnect," Proc. of SPIE 7640-43 (2010).
[12] http://www.tela-inc.com/
[13] J. Bekaert, L. Van Look, V. Wiaux, V. Truffert, M. Maenhoudt, G. Vandenberghe, M. Reybrouck, S. Tarutani, "Printing the contact and metal layers for the 32 and 22 nm node: Comparing positive and negative development process," 6th International Symposium on Immersion Lithography Extensions (2009).
[14] L. Van Look, J. Bekaert, V. Truffert, M. Reybrouck, S. Tarutani, M. Maenhoudt, G. Vandenberghe, "Printing the metal and contact layers for the 32- and 22-nm node: comparing positive and negative tone development process," Proc. of SPIE vol. 7640-69 (2010).