Seamless-merging-oriented parallel inverse lithography technology

Yang Yiwei(杨祎巍)†, Shi Zheng(史峥), and Shen Shanhu(沈珊瑚)

(Institute of VLSI Design, Zhejiang University, Hangzhou 310027, China)

Abstract: Inverse lithography technology (ILT), a promising resolution enhancement technology (RET) used in next generations of IC manufacture, has the capability to push lithography to its limit. However, the existing methods of ILT are either time-consuming due to the large layout in a single process, or not accurate enough due to simply block merging in the parallel process. The seamless-merging-oriented parallel ILT method proposed in this paper is fast because of the parallel process; and most importantly, convergence enhancement penalty terms (CEPT) introduced in the parallel ILT optimization process take the environment into consideration as well as environmental change through target updating. This method increases the similarity of the overlapped area between guard-bands and work units, makes the merging process approach seamless and hence reduces hot-spots. The experimental results show that seamless-merging-oriented parallel ILT not only accelerates the optimization process, but also significantly improves the quality of ILT.

Key words: lithography; parallel; inverse lithography technology; seamless merging; convergence
DOI: 10.1088/1674-4926/30/10/106002 **EEACC:** 2550G

1. Introduction

Lithography is an important step in IC manufacturing for transferring patterns from mask to wafer; however, distortion always exists in lithography due to diffraction and interference of light waves. Many traditional resolution enhancement technologies (RETs), including rule and model based optical proximate correction (OPC), sub-resolution assist features (SRAF), and various phase shift mask (PSM) techniques, are designed to improve pattern fidelity in lithography. However, these traditional RETs are gradually becoming insufficient, as the 193 nm wavelength is now applied to the 45 nm node and even beyond[1]. Unlike the traditional RETs, which mainly modify the original masks, ILT creates completely different masks in shapes based on lithography models to improve pattern fidelity, with the advantage of large process windows and more accurate mask description[2, 3].

The lithography system can be described using Eq. (1), where the function Litho(.) represents the lithography system, including resist development. The variables mask, contour and target represent lithography mask, lithography result of mask on wafer and target patterns on wafer respectively. The mathematical description of ILT is in Eq. (2).

$$\text{contour} = \text{Litho}(\text{mask}), \tag{1}$$

$$\text{mask} = \text{Litho}^{-1}(\text{target}). \tag{2}$$

ILT has been studied for decades, although the study has only become popular in recent years. According to previous work[2–7], the inverse function $\text{Litho}^{-1}(\cdot)$ dose not exist due to the ill-posedness of function $\text{Litho}(\cdot)$ and the inverse lithography problem can be translated into an optimization problem with the aim of making the distance between contour and target as small as possible, described here as ||Litho(mask)-target|| with various constraints. Various types of optimization methods, such as level-set methods[3], DCT-2 based mask synthesis[8] and gradient-related methods[9], can be used in ILT. Using whatever method, ILT is always time-consuming and in practice it costs unbearable time and unaffordable memory to get a result of the entire layout on a single processor. Parallel ILT is a feasible way to speed up ILT, and the so-called traditional parallel ILT in this paper is investigated in Ref. [10]. There are three main steps in traditional parallel ILT: the first is that the layout is partitioned into tiles called work units; in the second step each work unit is put into the ILT engine in parallel; and in the third step all work units are merged together. In the first step, a buffer called a guard-band is expanded around the work unit to include patterns in the immediate neighborhood of the work unit, thus the effect from neighboring patterns can be taken into consideration. However, the guard-band is an intuitive solution, and it does not guarantee correctness due to un-seamless merging, which will be discussed in detail in following sections.

This paper proposes a seamless-merging-oriented parallel ILT, which not only partitions the layouts to accelerate the optimization process but also adds convergence enhancement penalty terms (CEPT) in the optimization processes of each work unit to guide the convergence, and finally reduces the hot-spots caused by merging. The experimental results show that the seamless-merging-oriented parallel ILT efficiently restrains the effects induced by layout partition and merging, and improves the quality of the result.

The forward lithography model, the foundation of ILT, is introduced briefly in the second section. Traditional parallel ILT and seamless-merging-oriented parallel ILT are discussed

† Corresponding author. Email: yangyw@vlsi.zju.edu.cn
Received 17 April 2009, revised manuscript received 11 May 2009

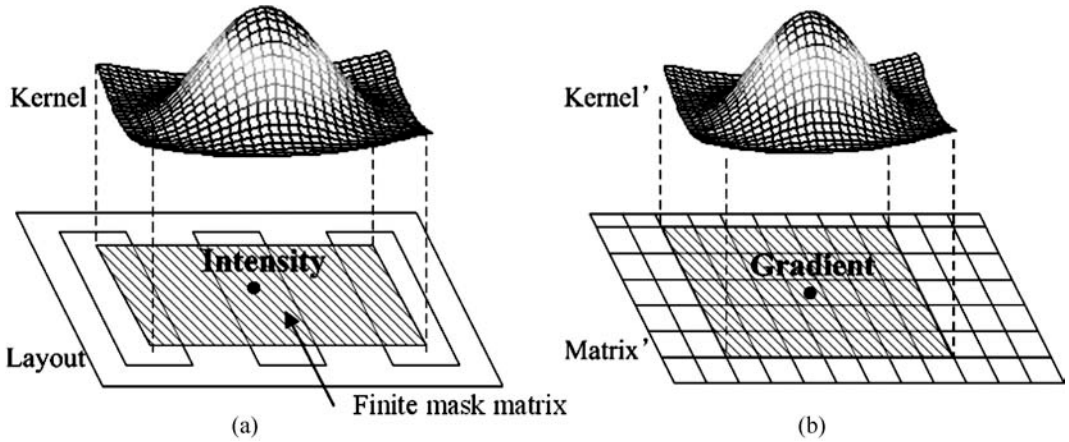

Fig. 1. Schematic of (a) Property 1 and Property 2. Property 1 and (b) Property 2 state that intensity (gradient) is determined by kernel (kernel') and finite elements of mask matrix (matrix') only.

in the third section. A discussion on the experimental results and the conclusion are presented in the last section.

2. Forward lithography model

The symbols and assumptions used in this paper are defined first. $\boldsymbol{M}$, $\boldsymbol{Z} \in \boldsymbol{R}^{P\times Q}$ are mask matrix and target matrix with P rows and Q columns respectively, and the pixels (p, q) in $\boldsymbol{M}$ and $\boldsymbol{Z}$ are termed m_{pq} and z_{pq}. $\boldsymbol{I}$ is the intensity matrix in accord with the mask matrix $\boldsymbol{M}$, and $i(p, q)$ or i_{pq} is the intensity of entry (p, q). $\otimes$ is a convolution operator. In this paper, the mask type is binary ($m_{pq} \in \{0,1\}$) for convenience, which can be easily extended to other mask types, such as attenuated PSM.

The function Litho(mask) mainly consists of an optical model and a resist development model, which are described by the SOCS (sum of coherent systems)[11] optical model and constant threshold resist models[11], respectively.

The SOCS system is described in Eq. (3), where λ_i is the amplitude of each lithography kernel h_i, and $h_i \in \boldsymbol{C}^{K\times K}$; K is the kernel ambit. The model input is the mask and the model output is intensity.

$$\begin{aligned} i(p,q) &= \sum_{i=1}^{N} \lambda_i |h_i \otimes \boldsymbol{M}|^2 \\ &= \sum_{i=1}^{N} \lambda_i [(\mathrm{Re}(h_i) \otimes \boldsymbol{M})^2 + (\mathrm{Im}(h_i) \otimes \boldsymbol{M})^2]. \end{aligned} \tag{3}$$

Property 1: Equation (3) indicates that the intensity of point (p, q) is determined by kernel and finite mask area only, as Figure 1(a) shows.

The constant threshold resist model is described by the sigmoid function. The model input is intensity and the model output indicates the resist thickness. Taking a dark field mask as an example, the constant threshold resist model is Eq. (4).

$$\mathrm{sig}(i) = \{1 + \exp(-a(i - i_{\mathrm{tr}}))\}^{-1}, \tag{4}$$

where a is the steepness factor and i_{tr} is the threshold. The curve of the sigmoid function is shown in Fig. 2.

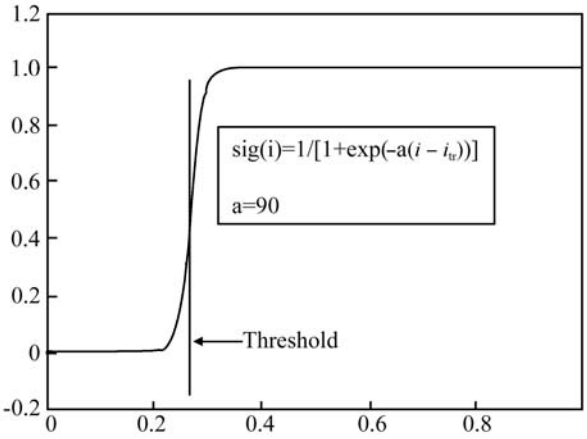

Fig. 2. Curve of sigmoid function. The larger the steepness factor a is, the steeper the curve is at the threshold.

Therefore, the lithography model can be obtained by combining the optical model and the resist model together in Eq. (5), which is a nonlinear function in matrix form.

$$\mathrm{Litho}(\boldsymbol{M}) = \mathrm{sig}(\boldsymbol{I}(\boldsymbol{M})). \tag{5}$$

3. Traditional parallel ILT and seamless-merging-oriented parallel ILT

Taking the gradient-related method as an example, this section shows how traditional parallel ILT works as well as the drawbacks of merging, and seamless-merging-oriented parallel ILT is then introduced.

3.1. Traditional parallel ILT

The cost function $C(\cdot)$ is defined as the distance between contour and target in discrete space in Eq. (6).

$$C(\boldsymbol{M}) = \|\mathrm{Litho}(\boldsymbol{M}) - \boldsymbol{Z}\| = \sum_{p=1}^{P}\sum_{q=1}^{Q} (\mathrm{sig}(i(p,q)) - z_{pq})^2. \tag{6}$$

In Ref. [9], the element m_{pq} is represented in Eq. (7), which means the variables are changed from m_{pq} (discrete variable) to θ_{pq} (continuous variable). $\boldsymbol{\theta}$ is a matrix in accord

with mask matrix M.

$$m_{pq} = (1 + \cos\theta_{pq})/2, \quad \text{for} \quad p = 1, ..., P;\ q = 1, ..., Q,$$
$$\boldsymbol{M} = (1 + \cos\theta)/2. \tag{7}$$

Then gradient-related iterative methods can be applied to Eq. (6). The kth iteration is Eq. (8).

$$\theta_k = \theta_{k-1} - sd, \tag{8}$$

where s is the step size and d is the gradient of cost function defined in Eq. (9).

$$d = \nabla C(\theta) = -\sum_{i=1}^{N} \lambda_i \text{Matrix}_i^{'} \otimes \text{Kernel}_i^{'}, \tag{9}$$

where

$$\text{Matrix}_i^{'} = 2\alpha \sin(\theta) \odot (\text{sig}(I) - Z) \odot \text{sig}(I) \odot (1 - sig(I)) \odot[(\text{Re}(h_i) + \text{Im}(h_i)) \otimes M],$$
$$\text{Kernel}_i^{'} = \text{rot180}(\text{Re}(h_i) + \text{Im}(h_i)), \tag{10}$$

where $\odot$ means multiply element-by-element, and rot180(.) means rotating the input matrix 180 degrees, which is often seen in convolution processes.

Property 2: In Eq. (9), the main operator of gradient formulation is convolution; hence gradient has the same properties as intensity, as Figure 1(b) shows. Only finite entries in the matrix $\text{Matrix}_i^{'}$ influence the gradient inferred from Property 1 and Property 2, and thus it is reasonable and necessary to add the guard-band around the work unit to compute the correct gradient.

In traditional parallel ILT, only using a guard-band is not enough to make the merged result hot-spot clean, because the work unit fringe environment before and after merging is different. As Figure 3 shows, the fringe environment of work unit A will change from guard-band A to work unit B during merging, which may induce hot-spots. Therefore, a new method is proposed in the next section to restrain the hot-spots.

3.2. Seamless merging in parallel ILT

Only using a guard-band is not enough, as stated above. The work unit fringe environment changes from the guard-band to adjacent work units during merging, which leads to potential hazards, and this is not seamless enough. Here, 'seamless' means that the guard-band is similar to the overlapping adjacent work units. As Figure 4 shows, the guard-band of work unit A and adjacent work units B in the same region in the layout, which is the shadow area, should be similar. The similarity, defined as the percentage of similar area over the overlapping area in Eq. (11), is an important index of the quality of merging, and high similarity means that the image of the work unit result changes little before and after merging. In Eq. (11) area means the area in the layout needing similarity evaluation, which is the shadow area in the layout in Fig. 4, and $\oplus$ is XOR operation.

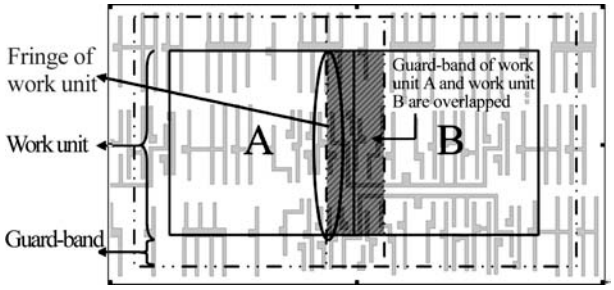

Fig. 3. Two work units A and B (solid frame) and their corresponding guard-bands (shadow frame).

$$\text{similarity} = \left(1 - \frac{\iint\limits_{(x,y)\in\text{area}} |\text{iltmask1} \oplus \text{iltmask2}|\,\text{d}x\text{d}y}{\iint\limits_{(x,y)\in\text{area}} \text{d}x\text{d}y}\right) \times 100\%. \tag{11}$$

To approach seamless merging, the new method proposed in this paper adds CEPT to the traditional parallel ILT process which means that the guard-band has two tasks: (a) to provide an environment to ensure that the lithographical result of the work unit it surrounds is close to the target; (b) to make the distance between the guard-band and the overlapping area of adjacent work units as small as possible; in other words, the guard-bands and the overlapping area of work units have high similarity. For example, in Fig. 4, in the top picture the shadow area in the layout has two roles: it is part of the guard-band of work unit A and at the same time it is part of work unit B; in the bottom left picture, without CEPT in traditional parallel ILT, the shadow area in the guard-band of work unit A just provides the environment of work unit A and it has low similarity with the shadow area in work unit B; in the bottom right picture, with CEPT in seamless-merging-oriented parallel ILT, the shadow area in the guard-band of work unit A not only provides the environment of work unit A but also has a smaller distance between itself and the shadow area in work unit B; it therefore has high similarity with the shadow area in work unit B. Specifically, in the parallel ILT process, CEPT will be added to the gradient d in Eq. (9), with the form $w \times \partial \sum_{(p,q)\in G}(m_{pq} - \text{target}_{pq})/\partial\theta_{pq}$, where target is the area in the work unit which overlaps with the guard-band (generally 8 work units are involved), w is weight and G is guard-band area. In this paper, the expression for CEPT is given in Eq. (12), where $\boldsymbol{W}$, Mask and Target are weight matrix, mask matrix and target matrix, respectively. $\odot$ means multiply element-by-element.

$$\text{CEPT} = \boldsymbol{W} \odot \sin(\theta) \odot (\text{Mask} - \text{Target}). \tag{12}$$

The flow of seamless-merging-oriented parallel ILT is shown in Fig. 5; GDS parsing is the first step for loading the layout. Hierarchy analysis analyzes the target layout and finds repetitive patterns so that only a single instance of each pattern is processed. After hierarchy analysis, the remaining instances are partitioned into work units with guard-bands. In parallel

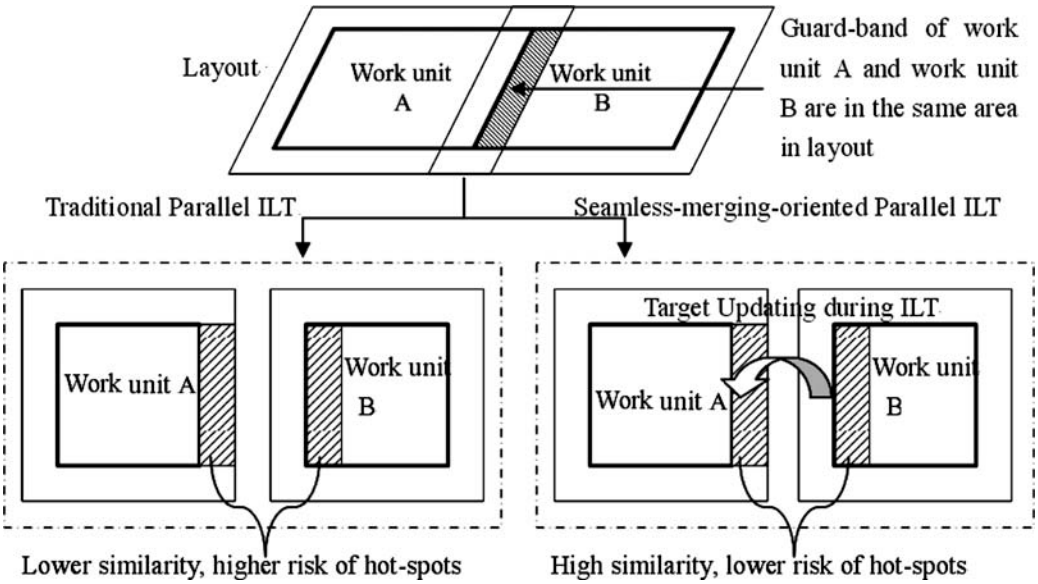

Fig. 4. Target updating schematic. The shadow area in the guard-band of work unit A and the shadow area in work unit B are in the same region in the layout. The shadow area of work unit B is the target of the shadow area in the guard-band of work unit A.

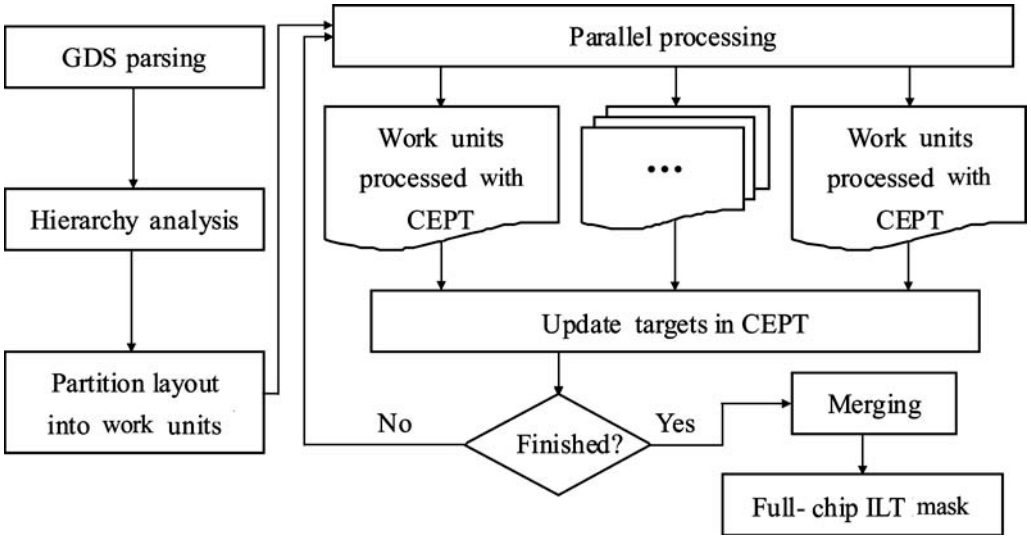

Fig. 5. Flow chart of seamless-merging-oriented parallel ILT.

processing, all work units are processed by certain iterations first; after that the targets of all work units are updated and parallel ILT processing continues. These two steps are repeated until the ILT process is finished. The target updates ensure that the guard-bands have high similarity with their adjacent work units. The guard-bands take the effects from neighboring patterns into account as well as the effects from changes of neighboring patterns, and so are more accurate.

The additional costs of this seamless-merging-oriented parallel ILT are extra computations for calculating CEPT and time for updating the targets.

4. Experiment discussion and conclusion

Experiments were designed to test the quality of seamless-merging-oriented parallel ILT. The parameters of the experiment, including model parameters and optimization parameters, were only designed for the experiments.

The layout is a poly-silicon layer of a logic layout in 130 nm node and the parameters of the lithography system are as follows: numerical aperture is 0.85, wave length is 193 nm, traditional illumination with $s = 0.78$, threshold is 0.3, kernel ambit is 15 pixels, guard-band width of work unit is 60 pixels and one pixel is 10×10 nm^2.

Figure 6 is the pattern target on the wafer. It is partitioned into two work units: work unit A and work unit B. Area *a* is the guard-band of work unit A, while area *b* is the guard-band of work unit B. Two methods, traditional parallel ILT and seamless-merging-oriented parallel ILT, are used to get the result of this layout. In traditional parallel ILT, work units A and B are put into the ILT engine separately, and after the ILT optimization process work units A and B are extracted and put together. However, in seamless-merging-oriented parallel ILT, work units A and B are put into the ILT engine; the targets of work units A and B are updated every 60 iterations during the ILT optimization process and after the ILT

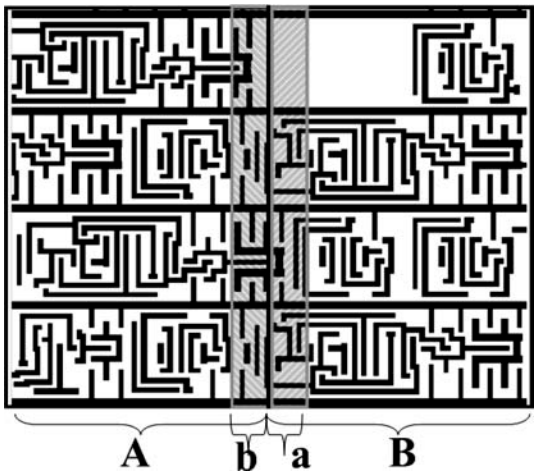

Fig. 6. Schematic of mask partition in experiment. Areas A and B are two work units and shadow areas *a* and *b* are guard-bands of work units A and B respectively.

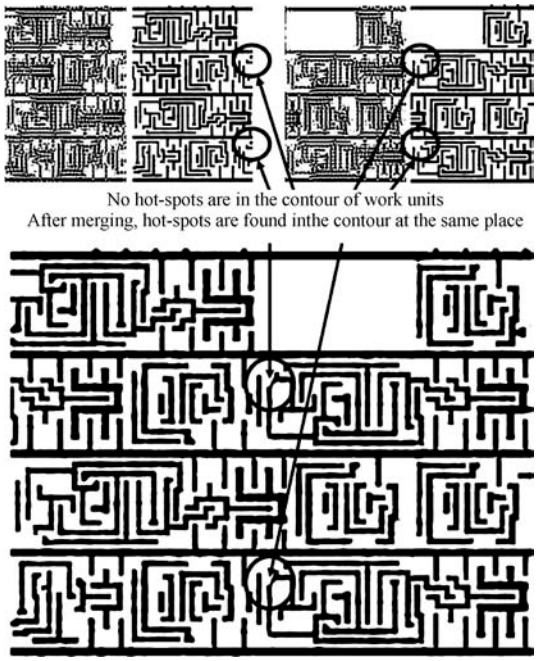

Fig. 7. Traditional parallel ILT result. Top-left is the result of work unit A, top-right is the result of work unit B and the bottom is the contour of merging the results of work units A and B. There are two hot-spots in the contour labeled by circles.

optimization process work units A and B are extracted and put together.

In Fig. 7 there is no hot-spot in work units A and B. However, after merging, there are two bridge errors in the middle of the contour. The reason for this is that during merging the guard-band of work unit A is replaced by work unit B and the guard-band of work unit B is replaced by work unit A, thus the fringe environments of work units A and B are changed by 2.83% and 2.26% respectively.

In Fig. 8, hot-spots exist neither in the results of work unit A and work unit B, nor in the merged result. From Tables 1 and 2, similarity is significantly improved by using seamless-merging-oriented parallel ILT. Higher similarity indicates that the environment of the work unit changes little before and after

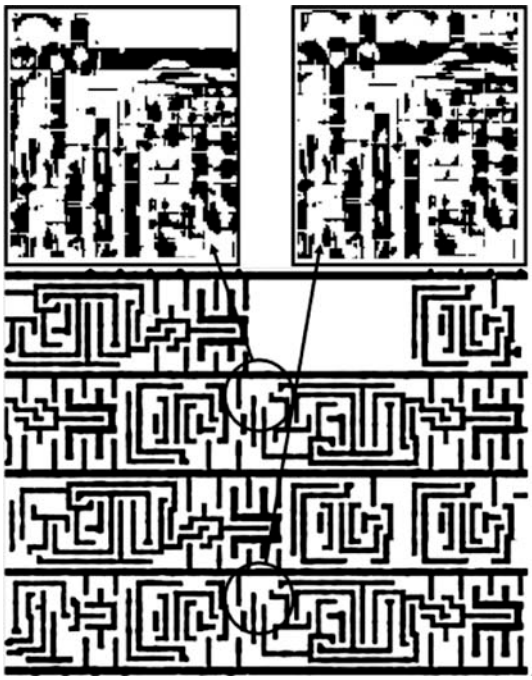

Fig. 8. Seamless-merging-oriented parallel ILT result. The top two figures represent the merged mask tiles of the two hot-spots. The bottom is the contour of merging the results of work units A and B. There are no hot-spots in the contour.

Table 1. Similarity of guard-bands of work units A and B by using traditional parallel ILT.

Table of similarity	Work unit A	Work unit B
Similarity	97.17%	97.74%
1- Similarity	2.83%	2.26%

Table 2. Similarity of guard-bands of work units A and B by using seamless-merging-oriented parallel ILT.

Table of similarity	Work unit A	Work unit B
Similarity	99.16%	99.48%
1- Similarity	0.84%	0.52%

Table 3. Average EPE and Max EPE of the two methods (pixel).

	Traditional parallel ILT	Seamless-merging-oriented parallel ILT
Average EPE	1.4	1.2
Max EPE	20	3

merging, thus the merged result is hot-spot clean. Dynamically updating the target in the CEPT takes effects from neighboring patterns into consideration as well as effects from changes in neighboring patterns of the guard-band, which greatly improves the quality of the merged results.

Table 3 clearly shows that the maximum EPE (edge placement error) is significantly reduced by using seamless-merging-oriented parallel ILT. Table 4 shows the time cost of each ILT method and seamless-merging-oriented parallel ILT is about 5% greater than traditional parallel ILT, which is acceptable.

Table 4. Time spent for the two methods (second).

	Traditional parallel ILT	Seamless-merging-oriented parallel ILT
Time	1514	1596

In conclusion, the experimental results show that seamless-merging-oriented parallel ILT outperforms traditional parallel ILT. Seamless-merging-oriented parallel ILT inherits the speed merits of traditional parallel ILT while increasing the similarity of overlapping areas in the layout to significantly improve ILT quality at low cost.

References

[1] Schellenberg F M. The next generation of RET. Proc SPIE Advanced Microlithography Technologies, 2005, 5645: 1

[2] Hung C Y, Zhang B, Pang L, et al. Pushing the lithography limit-applying inverse lithography technology (ILT) at the 65 nm generation. Proc SPIE Optical Microlithography XIX, 2006, 6154: 61541M

[3] Pang L, Liu Y, Abrams D. Inverse lithography technology (ILT) a natural solution for model-based SRAF at 45 nm and 32 nm. Proc SPIE Photomask and Next-Generation Lithography Mask Technology XIV, 2007, 6607: 660739

[4] Saleh B E A, Sayegh S I. Reductions of errors of microphotographic reproductions by optical corrections of original masks. Opt Eng, 1981, 20: 781

[5] Nashold K M, Saleh B E A. Image construction through diffraction-limited high-contrast imaging systems: an iterative approach. J Opt Soc Am A, 1985, 2: 635

[6] Granik Y. Fast pixel-based mask optimization for inverse lithography. J Microlith Microfab Microsyst, 2006, 5(4): 043002

[7] Abrams D S, Pang L. Fast inverse lithography technology. Proc SPIE Optical Microlithography XIX, 2006, 6154: 61541J

[8] Shen S, Yu P, Pan D Z. Enhanced DCT2-based inverse mask synthesis with initial SRAF insertion. Proc SPIE Photomask Technology, 2008, 7122: 712241

[9] Poonawala A, Milanfar P. Mask design for optical microlithography—an inverse imaging problem. IEEE Trans Image Processing, 2007, 16(3): 774

[10] Lin B, Ho J, Lin O, et al. Inverse lithography technology at chip scale. Proc SPIE Optical Microlithography XIX, 2006, 6154: 615414

[11] Cobb N, Zakhor A, Miloslavsky E A. Mathematical and CAD framework for proximity correction. Proc SPIE Optical Microlithography IX, 1996, 2726: 208