Enhanced Design Flow and Optimizations for Multi-Project Wafers

Andrew B. Kahng Ion I. Măndoiu Xu Xu Alex Z. Zelikovsky

***Abstract*— The aggressive scaling of VLSI feature size and the pervasive use of advanced reticle enhancement technologies leads to dramatic increases in mask costs, pushing prototype and low volume production designs at the limit of economic feasibility. Multiple project wafers (MPW), or "shuttle" runs, provide an attractive solution for such designs, by providing a mechanism to share the cost of mask tooling among up to tens of designs. However, MPW reticle floorplanning and wafer dicing introduce complexities not encountered in typical, single-project wafers. Recent works on wafer dicing adopt one or more the following assumptions to reduce the problem complexity, (i) assuming equal production volume requirement for all designs, (ii) assuming that the same dicing plan is used for all wafers or for all rows/columns of reticle images on a wafer, (iii) assuming unrealistic wafer models such as a rectangular array of projections and (iv) assuming fixed wafer shot-map. Although using one or more of the above assumptions makes the problem solvable, the performance of the solutions is degraded.**

In this paper we propose a comprehensive MPW flow aimed at minimizing the number of wafers needed to fulfill given die production volumes. Our flow includes two main steps: (1) multi-project reticle floorplanning, and (2) wafer shot-map and dicing plan definition. For each of these steps we propose improved algorithms as follows. Our reticle floorplanner uses hierarchical quadrisection combined with simulated annealing to generate "diceable" floorplans observing given maximum reticle sizes. Our dicing planner allows multiple side-to-side dicing plans for different wafers as well as different reticle projection rows/columns within a wafer, and further improves dicing yield by partitioning each wafer into a small number of parts before individual die extraction. We also propose to employ the dicing plan definition heuristic as a procedure for the wafer shot-map definition in order to fully utilize round wafer real estate by extracting the maximum number of functional dies from both fully and partially printed reticle images. Experiments on industry testcases show that our methods outperform significantly not only previous methods in the literature, but also reticle floorplans manually designed by experienced engineers.

Copyright (c) 2006 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

Work supported in part by Cadence Design Systems, Inc., the MARCO Gigascale Silicon Research Center, NSF Grant CCR-9988331. A preliminary version of the results in this paper has appeared in [9].

ABK is with the Departments of Computer Science and Engineering, and of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0114. E-mail: abk@ucsd.edu.

IIM is with the Computer Science and Engineering Department, University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155. E-mail: ion@engr.uconn.edu.

XX is with the Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093-0114. E-mail: xuxu@cs.ucsd.edu.

AZZ is with the Computer Science Department, Georgia State University, University Plaza, Atlanta, Georgia 30303. E-mail: alexz@cs.gsu.edu.

I. Introduction and Motivation

With the shrinking of VLSI feature size and the pervasive use of advanced reticle enhancement technologies such as Optical Proximity Correction (OPC) and Phase Shifting Masks (PSM), mask costs are predicted to reach $10 million

by the end of the decade. These high mask costs push prototyping and low volume production designs at the limit of economic feasibility since the costs cannot be amortized over the volume. Multiple Project Wafers (MPW), or "shuttle" runs, provide an efficient method to reduce the cost [11]. Thus, MPW has now become a commercial service offered by both independent providers such as MOSIS and CMP and semiconductor foundries such as the TSMC and IBM. An overview of related multi-layer mask technologies, which rely on sharing the reticle space between multiple layers of the *same* design, typically via blading, is given in [3].

Most previous papers on MPW reticle floorplanning rely on an "ideal-dicing" model which assumes either zero dicing loss [4] or arbitrary margins in the floorplan formulation. Chen and Lynn [5] considered the problem of finding the minimum area slicing floorplan, with 90 degree chip rotation allowed. Xu et al. [14] studied the MPW mask floorplanning under die-alignment constraints imposed by the use of die-to-die mask inspection. All these approaches assume that all dies can be obtained, which is impractical for current side-to-side wafer dicing technology. A grid-packing formulation for MPW mask floorplanning is proposed in [1] and [2], with the assumption that arbitrary blank area can be left on a die. However, in practice, arbitrary margins can not be tolerated due to package requirement.[1]

Side-to-side dicing based floorplanners consider the constraints imposed by the current side-to-side dicing technology. Due to the complexity of the general dicing problem, it is crucial to simplify the dicing problem and use a fast yet accurate wafer cost evaluator in the floorplanners. According to the different dicing simplification methods, the current reticle floorplanners can be divided into three categories.

- *Single wafer dicing plan (SWDP) assumption based floorplanners* assume that all wafers share the same dicing plan. Kahng et al. [8] were the first to consider the side-to-side wafer dicing problem with SWDP assumption. They propose three optimal integer linear programming (ILP) solutions and a fast heuristics for wafer cost evaluation. The fast wafer cost evaluator is used in a sequence pair based simulated annealing floorplanner. Recently, Kahng et al. [10] proposed a grid floorplanner. The wafer cost of grid floorplans can be directly calculated with a close form formula. Therefore, it is even practical to apply branch and bound algorithm to exhaustively search the whole solution space for small testcases. However, the close-form wafer cost calculation also depends on the impractical assumption that a wafer is a rectangular array of projections. Also the runtime of the proposed branch and bound algorithm may explode for large testcases.
- *Single row and column dicing plan (SRCDP) assumption based floorplanners* employ the assumption that all rows and columns of reticle images within a wafer are diced using the same set of cuts. Xu et al. [15], [16] formulate the dicing problem as a minimum coloring problem. Wu et al. [12] extend the min-coloring based dicing approach by proposing three ILP formulations for optimal minimum coloring. In [13], they propose to perform chip replication and give integrated ILPs for simultaneous floorplanning and dicing which are impractical even for small testcases due to large runtime.

In this paper we propose a comprehensive MPW flow aimed at minimizing the number of wafers needed to fulfill given die production volumes. Our flow includes two main steps: (1) multi-project reticle floorplanning, in which the reticle floorplan is designed for the given list of dies with fixed shot-map and simplified dicing cost evaluation; and (2) wafer shot-map and dicing plan definition, in which the

[1] No margins are allowed for our industry testcases from CMP.

exact dicing plan and wafer center location is determined for the floorplan generated in Step 1. In Step 2, the dicing plan generation algorithm is included in the wafer shot-map definition algorithm for accurate wafer cost calculation. Our contributions are as follows. For the first flow step, we propose an algorithm based on fixed hierarchical quadrisection structure which is suitable for fast wafer cost evaluation with simulated annealing to generate "diceable" floorplans observing given maximum reticle sizes. Our algorithm leads to an average reduction of 10-20% in the required number of wafers compared to reticle floorplans manually designed by experienced industry engineers. For the second step, we give an integer program which can be used to find in practical time the *optimal* dicing plan under the SRCDP assumption. We also give a two-level optimization algorithm that simultaneously allows multiple dicing plans for different wafers and for different reticle projection rows/columns within a wafer. We also show the advantages of partitioning each wafer into a small number of parts before individual die extraction. For a fixed reticle floorplan, the two-level optimization algorithm is shown to give an average reduction in the required number of wafers of 42% without wafer partition, and of 47%, respectively 63%, when partitioning into 2 or 4 parts. Finally, we propose to include wafer shot-map definition, which has not been previously considered in the context of MPW, in order to fully utilize the real estate on round wafers by extracting the maximum number of functional dies from both fully and partially printed reticle images. This optimization is shown to yield an average reduction of 13.6% in the required number of wafers for a reticle floorplan.

The rest of the paper is organized as follows. In the next section we describe the basics of MPW with side-to-side wafer dicing. In Section III, a novel hierarchical quadrisection method is presented for reticle floorplanning. Section IV describes the multiple dicing plan (MDP) advantages and give a novel two-level optimization algorithm. Section V combines the wafer shot-map definition with dicing plan definition for further wafer cost reduction. Finally, in Section VI we give experimental results comparing proposed methods on industrial testcases.

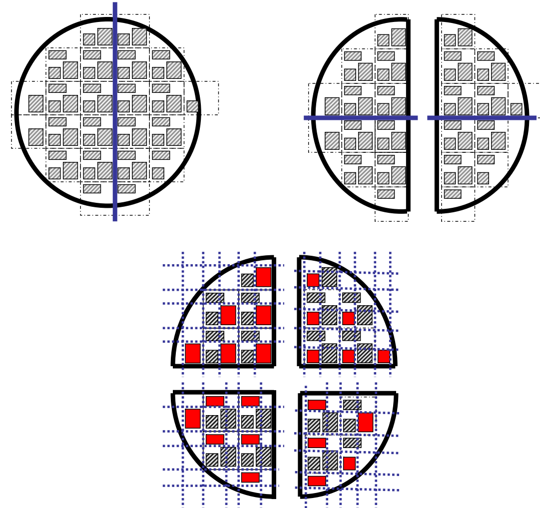

Fig. 1. Four quadrant dicing: the wafer is first divided into four quadrants, then each quadrant is diced independently using side-to-side cuts.

II. Preliminaries

A wafer consists of a number of reticle projections arranged in a number of reticle image *projection rows* and *projection columns*. Each projection is a copy of the same reticle image. In the prevalent "side-to-side" wafer dicing technology, the diamond blades can not stop at arbitrary points during cutting; consequently, all projections in the same projection row (or column) will share the same horizontal (or vertical) cutlines. In this paper, we extend side-to-side dicing to allow preliminary partitioning of each wafer into a small number of parts (e.g., halves or quarters)as shown in Figure 1 so that the side-to-side dicing plans for the parts can be independent from each other.

Following [8], two dies D and D' on a reticle are said to be in *vertical (resp. horizontal) dicing conflict* if no set

of vertical (resp. horizontal) cuts can legally dice both D and D'. Let $\mathcal{D}$ denote the set of dies on a given reticle. The *vertical reticle conflict graph* $R_v = (\mathcal{D}, E_v)$ is the graph with vertices corresponding to the dies and edges connecting pairs of dies in vertical dicing conflict. The *horizontal reticle conflict graph* $R_h = (\mathcal{D}, E_h)$ is defined similarly. As usual, a set of vertices in a graph is called independent if they are pairwise nonadjacent. A *maximum horizontal (or vertical) independent set* is a subset of $\mathcal{D}$ which can be sliced out by a set of horizontal (or vertical) cutlines; the set of cutlines used for a wafer are called as a *wafer dicing plan*. The *dicing yield* of a die D is defined as the number of legally diced copies of D divided by its volume requirement. The *wafer dicing yield* is defined as the minimum dicing yield over all dies $D \in \mathcal{D}$, which needs to be at least 1.

III. Reticle Floorplanning

In this section, we focus on the following MPW reticle floorplanning problem: Given a maximum reticle size, and the size and required volume for each die, find a reticle floorplan (allowing die rotations) and a wafer dicing plan minimizing the number of used wafers.

Compared with other floorplanning problems, the main difficulty of the MPW reticle floorplanning problem lies in the wafer cost calculation. To simplify and speed up the estimation of wafer cost and dicing plan yield, we use hierarchical quadrisection-based floorplanning. The reticle floorplan is based on a hierarchical quadrisection mesh which is constructed in the following recursive way.

- At Level 1, the reticle area is divided into 4 regions with one horizontal line and one vertical line: $R(1,1), R(1,2), R(1,3)$ and $R(1,4)$, where $R(i,j)$ is the jth Region for Level i.
- At Level $i+1$, each region at Level i, $R(i,j)$, is divided into 4 regions with one horizontal line and one vertical line: $R(i+1, 4^{j-1}+1), R(i+1, 4^{j-1}+2), R(i+1, 4^{j-1}+3)$ and $R(i+1, 4^{j-1}+4)$.

Finally, there are 4^l regions at Level l. Figure 2(a) shows a mesh of Level 2. The constructed mesh is "soft" since the dimensions of regions are determined by the dies within the regions. The number of level l is chosen such that 4^l is greater than the number of dies.[2] Then we place the dies in the regions of Level l mesh such that each region $R(l,j)$ $(j = 1...4^l)$ contains at most one die. Different die placements lead to different reticle floorplans. Figure 2 (b) and (c) show two different reticle floorplans for a set of 10 dies based on the same mesh in Figure 2(a). A simulated-annealing based algorithm is used to find the best die placement.

We denote the width of the region $R(i,j)$ as $W(R(i,j))$ and the height as $H(R(i,j))$. The hierarchical quadrisection allows computing height and width in a bottom-up manner.

- At Level l, if there is a die in the region $R(l,j)$, $W(R(l,j))$ is equal to the width of the die and $H(R(l,j))$ is equal to the height of die; otherwise, $W(R(l,j)) = H(R(l,j)) = 0$.
- At Level i, $W(R(i,j)) = Max(W(R(i+1,4^{j-1}+1)), W(R(i+1,4^{j-1}+4))) + Max(W(R(i+1,4^{j-1}+2)), W(R(i+1,4^{j-1}+3)))$.
  $H(R(i,j)) = Max(H(R(i+1,4^{j-1}+1)), H(R(i+1,4^{j-1}+2))) + Max(H(R(i+1,4^{j-1}+3)), H(R(i+1,4^{j-1}+4)))$.

There are two main advantages of the proposed floorplan structure. First, the structure is suitable for conflict elimination since there are no conflicts between dies located in diagonal regions. Second, the wafer cost can be easily evaluated with the following lemma.

[2]It is sufficient to choose $l = 3$ in practice since the case of putting more than 64 dies in one reticle is very rare, although we may choose $l = \lceil log_4 Number\ of\ dies \rceil$ if the number of dies is larger than 64.

Lemma 1: All dies can be divided into at most 2^l conflict-independent sets of dies for the floorplan in a Level l mesh such that any two dies in the same set are not in conflict.

Proof: The lemma is true for $l = 1$ since the dies in $R(1,1)$ and $R(1,3)$ are not in conflict and the dies in $R(1,2)$ and $R(1,4)$ are not in conflict.

Suppose the lemma is true for $l = i$, for $l = i+1$: the reticle is first divided into four regions $R(1,1), R(1,2), R(1,3)$ and $R(1,4)$ and each region is further divided into a Level i mesh. Since the lemma is true for $l = i$, there are at most 2^i conflict-independent sets for each of the four regions. We denote the kth conflict-independent set of the region $R(1,j)$ $(j = 1..4)$ as $S(1,j,k)$. Since any die in $R(1,1)$ is not in conflict with any die in $R(1,3)$, we can have the 2^i combined conflict-independent sets $S(1,1,k) \bigcup S(1,3,k)$ $(k = 1..2^i)$. Similarly, we can have another 2^i combined conflict-independent sets $S(1,2,k) \bigcup S(1,4,k)$ $(k = 1..2^i)$. Therefore, there are at most 2^{i+1} conflict-independent sets. □

It is obvious that all copies of the dies in the same conflict-independent set can be simultaneously sliced out since they are not in conflict. If we assume that only the dies of one conflict-independent set are obtained for each wafer, the *wafer requirement* for a conflict-independent set S is $MAX_{D \in S}(\lceil \frac{N(D)}{Q(D)} \rceil)$, where $N(D)$ is the volume requirement of the die D and $Q(D)$ is the number of copies of die D per wafer.[3] The total wafer requirement is the sum of the wafer requirements of all the conflict-independent sets.

We give a generic simulated annealing placement algorithm for finding reticle floorplan in Figure 3. The algorithm starts with the floorplan with each die randomly placed in the 4^l regions as its initial placement. First, the algorithm try to minimize the floorplan area in order to find a feasible solution. After a feasible solution is found, the objective switches to minimize the *total wafer requirement* whose calculation is specified in Lemma 1 and the paragraph following the proof of Lemma 1. Note that for speeding-up the algorithm, quadrisection floorplan evaluation does not include the dicing plan and the shot-map definition. At each step we find a neighbor solution based on the following moves:

- Region exchange move, which exchanges the dies in two regions if at least one of the regions contains a die;
- Orientation move, which rotates one die by 90 degrees if the width and height of the die are different.

Each generated solution is evaluated and kept with a probability dependent on the current temperature (see Figure 3). Finally, we may inset additional copies of dies if the reticle dimension is not increased (Lines 12-15).[45]

IV. Multiple-Dicing-Plan Dicing

The following problem has been introduced in [8].

Side-to-Side Wafer Dicing Problem (SSWDP). Given a reticle floorplan with dies $D = \{D_1, \ldots, D_n\}$, required production volume for each die $N(D_i)$, $i = 1, \ldots, n$, and the positions of the reticle projections of the wafer, find the minimum number of wafers N_w and the corresponding dicing plan for each wafer such that the wafer dicing yield is at least 1.

[3] In order to speedup the wafer cost evaluation in the floorplanning step, we fix the wafer center at the point $(0,0)$ and set $Q(D)$ to the number of dies D on the wafer (see Section V).

[4] Whether a die D can be inserted is decided by finding a free room for D on the reticle : we place the left-bottom corner of D and its 90° rotation at the corners of each die in the reticle and check whether D overlaps with other dies in the reticle.

[5] In practice, there is not too much empty space left in the reticle even the number of dies is substantially smaller than the grid-number of the mesh due to two reasons: (1) if there is no die in a region of Level l, the region area is zero; (2) if a floorplan has too much empty area, its dimension will exceed the maximum reticle dimension and this floorplan will be discarded.

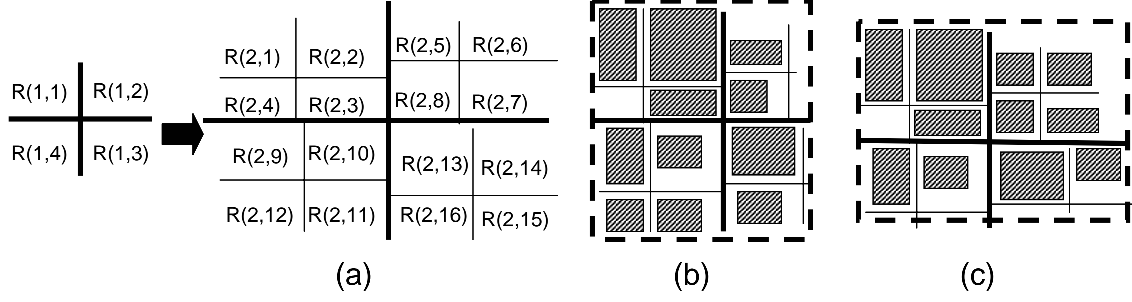

Fig. 2. Two-level Hierarchical Quadrisection Floorplan.

Input: Dimensions of n dies, β: $0 \leq \beta < 1$

Output: Reticle floorplan and wafer dicing plan

1. Construct the hierarchical quadrisection floorplan mesh
2. Assign the n dies to regions at random
3. If (floorplan width and height smaller than maximum reticle dimensions) then $FoundFeasible \leftarrow True$
4. Else $FoundFeasible \leftarrow False$
5. While (not converged and # of moves $< Move_Limit$)
6. Pick a move at random
7. If (floorplan width and height smaller than maximum reticle dimensions) then
8. $FoundFeasible \leftarrow True$; $\delta \leftarrow$New Wafer requirement - Old Wafer Requirement
9. Else, if ($FoundFeasible = False$) then $\delta \leftarrow$ New Area - Old Area, else $\delta \leftarrow \infty$
10. If ($\delta < 0$) then accept the move, else accept the move with probability $e^{-\frac{\delta}{T}}$
11. $T \leftarrow \beta T$
12. While ($\exists$ a die that can be inserted)
13. Sort all dies that can be inserted in descending order of $N(D)/A(D)$
14. For each die D_i do
15. If (D_i can be inserted) then insert it

Fig. 3. Hierarchical Quadrisection Floorplan.

In [8] and [10], the authors adopt SWDP assumption, which limits the solution space. The IASA method proposed for SDP in [8] can be extended to solve MDP by placing N_w wafers into one "super-wafer" whose row (column) number is N_w times the initial row (column) number as shown in Figure 4. However, the runtime will increase rapidly when N_w is large since we need to check all rows and columns of the "super-wafer" in each iteration. The large runtime makes it unsuitable to be used in our proposed flow since the wafer shot-map definition step requires the accurate wafer cost calculation for each candidate wafer center location.

A. Integer Linear Program for Restricted MDPs

In [15], the authors assume that each wafer uses exactly one horizontal dicing plan and one vertical dicing plan for all projection rows/columns within a wafer. This assumption allows them to use a coloring based heuristic which gives good results for testcases with large volume requirement. In this section we give an integer linear programming formulation which allows finding optimal MDPs restricted in this way.

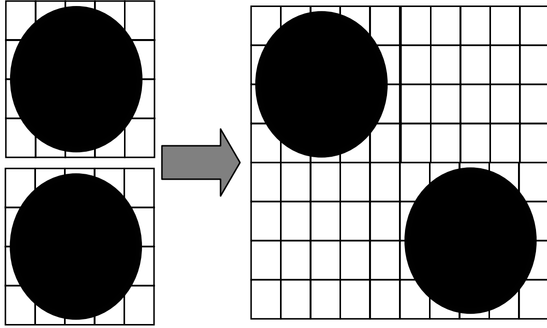

Fig. 4. Placing two wafers on one "super-wafer".

As in [15], two dies D and D' on a reticle are said to be in *dicing conflict* if they are either in horizontal dicing conflict or vertical dicing conflict. The *conflict graph* $R_c = (\mathcal{D}, E_c)$ is the graph with vertices corresponding to the dies and edges connecting pairs of dies in dicing conflict. A *maximum conflict independent set* is a subset of $\mathcal{D}$ which can be sliced out by a set of horizontal and vertical cutlines. We use *MCIS* to denote the set of all maximal independent sets in the conflict graph.[6] For each independent set $C \in MCIS$, let f_C denote the number of wafers which use the dicing plan defined by C, MDP can be formulated as the following integer linear program:

Minimize N_w (ILP1)

subject to

$$\sum_{D \in C} Q(C,D) f_C \geq N(D), \qquad \forall D \in \mathcal{D}$$
$$\sum_C f_C = N_w$$
$$f_C \in \mathbb{Z}_+, \qquad \forall C \in MCIS$$

where $Q(C,D)$ is a constant which represents the number of copies of die D obtained from a wafer diced according to C. The ILP can be optimally solved in a short time since there are only $|MCIS|$ variables and $|\mathcal{D}|+1$ constraints. As shown in Section 6, the runtimes of ILP are within 0.03 second in all the experiments on industry testcases with up to 30 dies.

[6] *MCIS* can be found as follows. We denote the *MCIS* for i dies as $MCIS(i)$.

1) Sort all dies according to max_x.
2) $MCIS(1) \leftarrow \{D_1\}$.
3) For $(i = 2; i \leq n; i++)$
4) Find the last die D_j which satisfies $max_x(D_j) < min_x(D_i)$
5) Add D_i to every set in $MCIS(j)$ and $MCIS(i) \leftarrow MCIS(j) \cup MCIS(i-1)$.

B. Two-level Optimization Algorithm for MDP

Although the ILP method can solve the MDP problem quickly, its performance will be degraded for the small volume requirement cases. Extended IASA for MDP can produce a good solution but suffers from large runtime with large N_w. In order to rapidly find a near optimal solution for MDP, we propose the Two-level Optimization (TLO) heuristic shown in Figure 5. We first solve ILP1 to obtain an upper bound on N_w. Then we gradually reduce the number until the yield is smaller than 1. In Lines 04-08, we assume all rows (columns) of each wafer using the same horizontal (vertical) dicing plan. The dicing plan for each wafer are obtained by solving the following ILP:

Minimize Y (ILP2)

subject to

$$N(D) - \sum_{D \in C} Q(C,D) f_C \leq y_D, \qquad \forall D \in \mathcal{D}$$
$$\sum_C f_C = N_w$$
$$\sum_D y_D = Y$$
$$f_C \in \mathbb{Z}_+, \qquad \forall C \in MCIS$$
$$y_D \in \mathbb{Z}_+, \qquad \forall D \in \mathcal{D}$$

where Y is the total number of unsatisfied volume requirement and y_D is the number of unsatisfied volume requirement for the die D. Since one maximal conflict independent set may belong to several maximal horizontal (vertical) independent sets, we use y_D as the weight of D and choose the maximal horizontal (vertical) independent set with the maximum total weight for each wafer. Then we perform the "row-and-column level" dicing plan replacement in Lines

10-13 to improve the yield.[7] A *candidate pool* is employed to speed up the process. Since the wafer yield depends on the dies with the minimum dicing yield, the dicing plans which can sliced out at least one of these dies are put into the candidate pool. Only the dicing plans in the candidate pool will be tried in each iteration. The candidate pool will be updated with the change of min-yield dies. This process is greedy which requires the yield increase with each dicing plan replacement. If a die D does not belong to any chosen horizontal or vertical dicing plan, we need to simultaneously change a horizontal and a vertical dicing plan to obtain one copy of D and increase the yield. Therefore, a *cross selection* process in Lines 14-17 is used to choose the dicing plans for one row and one column simultaneously. Since the "cross selection" process is extremely time-consuming, we do it only for the center row and column of each wafer.

Input: *MHIS*, *MVIS*, *MCIS*
Output: N_w and dicing plan for N_w wafers
01. Solve ILP1 to obtain the N_w upper bound
02. while (yield$\geq$ 1)
03. $N_w - -$
04. Solve ILP2 and choose one set C for each wafer
05. Set the weight of each die D as y_D
06. **For** (each wafer)
07. Choose max horizontal (vertical) independent set
08. **While** (improve==true)
09. **While** (improve==true)
10. **For** (each row and column)
11. try other horizontal (vertical) dicing plans
12. **If** (wafer-dicing yield increases)
13. Replace the current dicing plan
14. **For** (the center row and column of each wafer)
15. *Simultaneously* try other pairs of dicing plans
16. **If** (wafer-dicing yield increases)
17. Replace the current dicing plan

Fig. 5. Two-level Optimization Heuristic

V. Wafer Shot-Map Definition

In the previous section we have fixed reticle images is order to reduce the problem complexity. However, if we allow the reticle images position to freely moved on the wafer, then the wafer cost can be reduced even more. The wafer shotmap definition step, which determines the position of reticle images printed on wafer, was previously investigated for general wafers to maximize the wafer yield [6]. However, it was ignored in the previous papers in the MPW context. In both [10] and [8], the wafer is modeled as a rectangular array of projects, which is not true for actual round wafers. This simplification may lead to wrong dicing yield estimation since (i) the projection rows (columns) do not have equal contributions to the wafer dicing yield – the rows/columns near the center contain more reticle images, and (ii) fully printed dies within partial reticle projection are ignored. For a round wafer with the radius r and the center (x_0, y_0), a die image D is *on wafer* if and only if $(x-x_0)^2+(y-y_0)^2 \leq r^2$ for all $(x,y) \in D$ (see Figure 6. Given a rectangular reticle image, a *shotmap* is a regular tiling of the plane with identical copies of the reticle. The corresponding problem of wafer position with respect to shot-map is formulated as follows:

Wafer Shot-Map Definition Problem (WSMDP). Given a projection plane and the wafer radius r, find the position of the wafer minimizing the number of wafers required to meet the given production volumes.

The periodic property of the projection plane imply following lemma:

Lemma 2: The optimal solution of WSMDP can be achieved when the location of the wafer center is restricted to be within one reticle projection L.

Proof:Let the reticle width and height be R_W and R_H and the

[7] In the process of yield and wafer cost evaluation, we may take the dicing operation setup cost and lithography cost into consideration. Here, yield improving is equal to total manufacturing cost reduction.

optimal solution of WSMDP can be achieved when the wafer center located in $(i\times R_W+x, j\times R_H+y)$, where i,j are integers and $0\le x<R_W, 0\le y<R_H$. It is obvious that for any copy of a die located in the wafer centered in $(i\times R_W+x, j\times R_H+y)$, there is a corresponding copy of the same die located in the wafer centered in (x,y) and vice versa. Therefore, the optimal solution can also be achieved when the wafer center located in (x,y). □

Therefore, the wafer center is constrained in one projection. The wafer center location is further constrained by the following lemma:

Lemma 3: The optimal solution of WSMDP can be achieved when at least two die corners located on the circular boundary of the wafer and the dies having these corners are located within the wafer.

Proof:Suppose the optimal solution of WSMDP can be achieved when the wafer center located in (x,y). Let S be the set of the four corner coordinates of all dies on the wafer. One fact is that one die is on the wafer if and only if its four corners are in the wafer, so the solution remains optimal if all the points in S are in the wafer. If there is no points in S are on the wafer boundary, then for any point $(x_i,y_i)\in S$, $(x_i-x)^2+(y_i-y)^2<r^2$. Let $t_i=x_i-x+\sqrt{r^2-(y_i-y)^2}$. It is easy to prove that $t_i>0$ and $(x_i-x-t_{i,1})^2+(y_i-y)^2=r^2$ (intuitively, this equals to move the wafer center to the right by t_i to make the point (x_i,y_i) on the boundary). Let t be the smallest value of all the t_i values and move the wafer center to $(x+t,y)$. Then at least one point in S will be located on the wafer boundary. Also if any point $(x_i,y_i)\in S$ is out of the wafer, then $(x_i-x-t)^2+(y_i-y)^2>r^2$ and $(x_i-x-t)^2+(y_i-y)^2>(x_i-x)^2+(y_i-y)^2 \Rightarrow r^2-(y_i-y)^2<(x_i-x-t)^2$ and $t>2(x_i-x)$ $\Rightarrow t_i<x_i-x+\sqrt{x_i-x-t)^2}=t$. However, this is impossible since t is the smallest value. Therefore, all points in S are still in the wafer when the wafer center is located at $(x+t,y)$.

Next, if there is only one die corner (x_1,y_1) on the boundary, we can perform coordinate transformation such that (x,y) in the original coordinates becomes $((x+t-x_1)\cos\phi+(y-y_1)\sin\phi, (y-y_1)\cos\phi-(x+t-x_1)\sin\phi)$ in the new coordinates, where $\phi=\arctan(\frac{y-y_1}{x+t-x_1})$. It is easy to prove that the points (x_1,y_1) and $(x+t,y)$ become $(0,0)$ and $(r,0)$ in the new coordinates. Let the (x'_i,y'_i) be the new coordinates of the points $(x_i,y_i)\in S$, $(x'_i-r)^2+y'^2_i<r^2$ $\Rightarrow x'^2_i+y'^2_i<2rx_i$. Let $\theta_i=\arcsin(\frac{\sqrt{x'^2_i+y'^2_i}}{2r})-\arcsin(\frac{x'_i}{\sqrt{x'^2_i+y'^2_i}})$. It is easy to prove that $\theta_i<0$ and $(x'_i-r\cos\theta_i)^2+(y'_i-r\sin\theta_i)^2=r^2$ (intuitively, this step equals to rotate the wafer center around the point $(0,0)$ by θ_i to make the point (x'_i,y'_i) on the boundary). Let θ be the largest value of all the θ_i values and rotate the wafer center around the point $(0,0)$ by θ to $(r\cos\theta, r\sin\theta)$. Then at least two points in S will be located on the wafer boundary: $(0,0)$ and the point (x'_i,y'_i) whose θ_i is θ. If any point $(x'_i,y'_i)\in S$ is out of the wafer, that is, $(x'_i-r\cos\theta)^2+(y'_i-r\sin\theta)^2>r^2 \Rightarrow \frac{\sqrt{x'^2_i+y'^2_i}}{2r}>\sin(\theta+\arcsin(\frac{x'_i}{\sqrt{x'^2_i+y'^2_i}}))\Rightarrow\theta_i>\theta$. However, it is impossible since θ is the largest value. Therefore, all points in S are still in the wafer when the wafer center is located at $(r\cos\theta, r\sin\theta)$. □

If two points (x_1,y_1) and (x_2,y_2) on the wafer boundary are known, the wafer center is located at either $(\frac{x_1+x_2-(y_1-y_2)t}{2},\frac{y_1+y_2+(x_1-x_2)t}{2})$ or $(\frac{x_1+x_2+(y_1-y_2)t}{2},\frac{y_1+y_2-(x_1-x_2)t}{2})$, where $t=\sqrt{\frac{4r^2}{(y_1-y_2)^2+(x_1-x_2)^2}-1}$.

As in Figure 7, when the wafer center is constrained in one projection L, all dies within Region 1 can be on the wafer. All dies within Region 2, which is the intersection of four cycles with radius of r whose centers located at the four corners of L, will be within the wafer no matter where within L the wafer center is. We define the set S as the set of dies which are in Region 1 and are not in the Region 2. From Lemma 3, it is sufficient to consider the points in L which is determined by at least two corners of dies in S when the corners are on the wafer boundary. The number of these points are at most $O(|S|^2)$. An optimal solution can be achieved by checking all these points.

However, obtaining the optimal solution is impractical due to the large number of points to be checked (as shown in Table IV) and the relatively long runtime of wafer cost calculation procedure TLO. Therefore, we propose a hierarchical wafer shot-map definition algorithm as summarized in Figure 8, which only calls TLO dicing procedures at a few hierarchically selected locations. We first divided the projection L into several grids, then run TLO when wafer

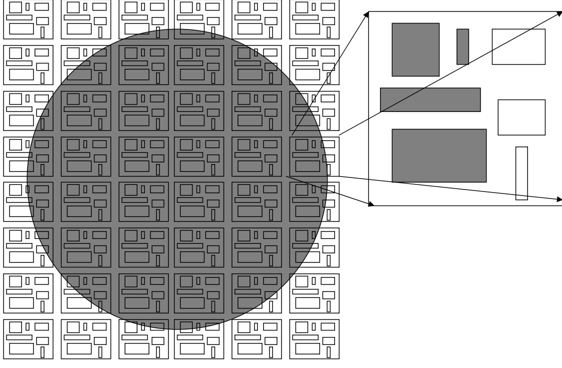

Fig. 6. A periodic shot-map with dark circular wafer. A partially printed reticle contains dark completely printed projects.

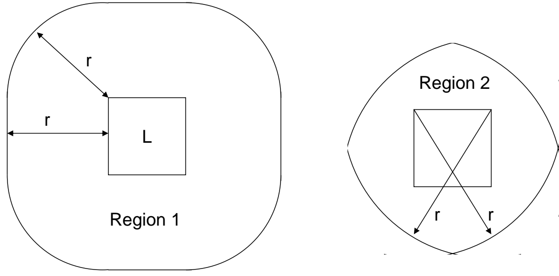

Fig. 7. Region 1 and Region 2 for the projection L.

Input: wafer radius r, reticle dimensions, one projection L
Output: wafer center location minimizing manufacturing cost
1. $L_0 \leftarrow L$ 2. **For** (level =0; level $< l$; level++) 3. Divide L_0 into $k \times k$ uniformly-spaced grid 4. **For** (all the grids) 5. choose the grid center p as the wafer center 6. **If** ($F(p)$ not included in any stored feasible set) 7. calculate N_w with TLO, store $F(p)$ 8. Find the grid g with the minimum N_w 9. $Min_N_w \leftarrow N_w$; $L_0 \leftarrow g$;

Fig. 8. Hierarchical Wafer Shot-Map Definition Algorithm

centering at each grid center. The "best" grid will be chosen for the next iteration. The following trick is employed in the algorithm to speed up the process. For each candidate wafer center location p, there is a *feasible set* of dies, $F(p)$, on the wafer when the wafer center is at p. Obviously, the wafer cost will not be reduced if $F(p)$ is a subset of one feasible set whose wafer cost is already calculated. We store all feasible sets whose wafer costs are calculated for comparison. In Line 6, if $F(p)$ is included in any stored set, p will be skipped to avoid redundant wafer cost calculation.

VI. Experimental Results

We used six industry testcases from CMP [17] to evaluate the performance and scalability of the proposed algorithms, each having between 12 and 31 dies with varying sizes and production volume requirements. For the wafer shot-map and wafer dicing problem, we used the reticle floorplan of the actual industry MPW runs which were manually designed by an experienced engineer. The basic parameters of the six testcases are listed in Table I.

Reticle Floorplanning. We implemented our hierarchical quadrisection floorplan algorithm in C++. The maximum reticle dimension is set as 2cm. After the placement, we use a fixed wafer shot-map and TLO dicing method to generate the dicing plans for all the wafers. The reticle floorplan results are summarized in Table II. Here "CMP" denotes the original industry floorplan used by CMP, "IASA+SA" is the SDP driven floorplanner used in [8] and HQ is our proposed hierarchical quadrisection floorplan algorithm. The results show that our proposed hierarchical quadrisection floorplan can save the wafer cost by 9.1%, 23.5% and 16.1% for one part, two parts and four parts compared with the original industry floorplan. On the other hand, "IASA+SA" increases the wafer cost by 18.2%, 14.7% and 17.8%, which indicates that "IASA+SA" is not a good choice for MDP on round wafers.

Wafer Dicing. We implement the wafer dicing algorithms in the C++ language. We set the wafer diameter as six inch and use a fixed wafer shot-map for all testcases. The number of wafers used (N_w) and runtime of four methods are shown in Table III, where IASA is the SDP method used

Cases	# dies	Total volume	Max Vol.	Min Vol.	Die area(cm^2)	$\|MCIS\|$	$\|MHIS\|$	$\|MVIS\|$
Ind 1	12	330	40	25	1.13	19	32	36
Ind 2	14	275	25	6	1.36	19	15	50
Ind 3	24	775	67	25	1.82	56	280	200
Ind 4	31	755	30	8	1.62	242	450	1008
Ind 5	14	250	25	12	0.86	18	63	40
Ind 6	24	625	35	25	2.26	127	588	1080

TABLE I

CMP TESTCASE PARAMETERS.

in [8], E-IASA is the extended IASA in Section 3.1, ILP is the integer linear programming restricted MDP method specified in Section 3.2 and TLO is the proposed two-level MDP optimization method. Each method was run without any wafer partition and with wafer partition into 2 or 4 parts prior to dicing. The results show that compared with the original IASA with one part, the wafer cost can be reduced by 34.2% by using four parts. E-IASA can reduce the wafer cost by 39.5% for one part at the expense of long runtime. ILP can reduce the cost by 5.3% for one part and can reduce the cost by 57.9% for four parts. Therefore, ILP is more efficient for multiple part dicing. TLO achieves the best solution quality in a short time. TLO reduces the wafer cost by 63.2% for four parts.

Wafer Shot-Map Definition. Our algorithm for the wafer shot-map definition problem is implemented in C++.

The wafer cost and runtime results are summarized in Table IV. "# points" is the number of possible wafer center locations to be checked to obtain an optimal solutions. l is the number of levels and k is the grid size used in each level. Compared with the fixed shotmap, the wafer cost can be reduced by 9.1% by using 10×10 grid at the expense of increased runtime. Runtime will significantly increase when N_w is reduced since there will be more dicing plan replacement iterations in TLO procedure. Although using 100×100 grid can further reduce the wafer cost (13.6%), the runtime becomes impractical (over $100X$). However, a good tradeoff between solution quality and runtime can be achieved by using our proposed hierarchical wafer shotmap definition algorithm with $l = 3$ and $k = 10$. The wafer cost is reduced by 13.6% while the runtime is within $2.5X$ compared with using 10×10 grid.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we proposed improved algorithms for multi-project reticle floorplanning, wafer shot-map definition, and wafer dicing. Experiments on industry testcases show that our methods outperform significantly previous methods in the literature as well as floorplans manually designed by experienced engineers. Our methods can also be extended to handle additional constraints such as die-alignment constraints imposed by the use of die-to-die mask inspection [14] by merging two copies of a die in a single "super-die". In ongoing work we investigate the use of multiple die copies on the reticle and multi-layer reticles for further reductions in the manufacturing cost of given die production volumes.

REFERENCES

[1] M. Andersson, C. Levcopoulos and J. Gudmundsson, "Chips on Wafers", *Proc. WADS (Workshop on Algorithms and Data Structures)*, August 2003.

Cases	# part	CMP		IASA+SA			HQ		
		N_w	area	N_w	area	CPU(s)	N_w	area	CPU(s)
Ind 1	1	3	1.13	3	1.58	24.2	3	1.42	0.00
Ind 2	1	3	1.36	3	1.83	39.2	2	1.65	0.00
Ind 3	1	4	1.82	7	1.96	1031	4	2.26	0.01
Ind 4	1	4	1.62	5	2.72	2351	4	1.82	0.01
Ind 5	1	2	0.86	2	1.77	51.7	2	1.19	0.00
Ind 6	1	6	2.26	6	3.60	795	5	2.66	0.01
Total		22		26			20		
Red.(%)				-18.2			9.1		
Ind 1	2	2	1.13	2.5	1.58	24.2	1.5	1.42	0.00
Ind 2	2	2	1.36	2	1.83	39.2	1.5	1.65	0.00
Ind 3	2	3	1.82	4	1.96	1031	3	2.26	0.01
Ind 4	2	3.5	1.62	3.5	2.72	2351	2.5	1.82	0.01
Ind 5	2	1.5	0.86	1.5	1.77	51.7	1.5	1.19	0.00
Ind 6	2	5	2.26	6	3.60	795	3	2.66	0.01
Total		17		19.5			13		
Red.(%)				-14.7			23.5		
Ind 1	4	1.5	1.13	1.75	1.58	24.2	1.25	1.42	0.00
Ind 2	4	1.5	1.36	1.75	1.83	39.2	1.5	1.65	0.00
Ind 3	4	2.75	1.82	4	1.96	1031	2.75	2.26	0.01
Ind 4	4	2.75	1.62	3.25	2.72	2351	2.25	1.82	0.01
Ind 5	4	1	0.86	1.25	1.77	51.7	1	1.19	0.00
Ind 6	4	4.5	2.26	4.5	3.60	795	3	2.66	0.01
Total		14		16.5			11.75		
Red.(%)				-17.8			16.1		

TABLE II

RETICLE FLOORPLAN RESULTS FOR SIX INDUSTRY TESTCASES. CMP IS THE ORIGINAL INDUSTRY FLOORPLAN USED IN CMP, "IASA+SA" IS THE SDP DRIVEN FLOORPLANNER USED IN [10] AND HQ IS OUR PROPOSED HIERARCHICAL QUADRISECTION FLOORPLAN ALGORITHM.

[2] M. Andersson, C. Levcopoulos and J. Gudmundsson, "Chips on Wafers", *Computational Geometry - Theory and Applications*, 30(2), 2005, pp. 95-111.

[3] A. Balasinski, "Multi-layer and multi-product masks: cost reduction methodology", *Proc. 24th BACUS Symp. on Photomask Technology*, Proc. SPIE, Vol 5567, 2004, pp. 351–359.

[4] A. Caprara, A. Lodi, and M. Monaci, "An Approximation Scheme for the Two-Stage, Two-Dimensional Bin Packing Problem", *Lecture Notes in Computer Science 2337, Springer-Verlag*, pp. 320-334.

[5] S. Chen and E. C. Lynn, "Effective Placement of Chips on a Shuttle Mask", *Proc. SPIE*, Vol 5130, 2003, pp. 681-688.

[6] Chien, C., and J. Deng, "Optimization of Wafer Exposure Patterns Using A Two-dimensional Cutting Algorithm", *Intl. Trans. in Operational Research*, Vol 8(5), 2001, pp. 535-545.

[7] B. Courtois, "Infrastructure for Wducation and Research: from National Initiatives to Worldwide Development". Paper presented at Tech. Univ. Darmstadt Infrastructure Overview, available at `http://vlsi1.engr.utk.edu/ece/msn/courtois.pdf`

[8] A. B. Kahng, I. I. Mandoiu, Q. Wang, X. Xu, and A. Zelikovsky, "MultiProject Reticle Floorplanning and Wafer Dicing", *Proc. Intl.*

Cases	# part	IASA		E-IASA		ILP		TLO	
		N_w	CPU(s)	N_w	CPU(s)	N_w	CPU(s)	N_w	CPU(s)
Ind 1	1	4	0.9	3	21.4	6	0.0	3	0.14
Ind 2	1	3	0.9	3	20.9	5	0.01	3	0.18
Ind 3	1	9	4.8	5	617	5	0.03	4	4.59
Ind 4	1	7	26.1	4	1631	8	0.03	4	73.6
Ind 5	1	2	1.9	2	15.5	4	0.0	2	0.21
Ind 6	1	13	13.2	6	2634	8	0.00	6	3.57
Total		38		23		36		22	
Red.(%)				39.5		5.3		42.1	
Ind 1	2	3	2.6	2.5	37.0	3	0.0	2	0.05
Ind 2	2	3	2.3	2	18.8	2.5	0.0	2	0.06
Ind 3	2	7	16.8	4.5	1485	3.5	0.01	3	3.98
Ind 4	2	5	76.9	3.5	3041	4	0.02	3.5	0.76
Ind 5	2	2	5.7	1.5	17.7	2	0.0	1.5	0.21
Ind 6	2	9	37.4	5	4457	5	0.02	5	0.04
Total		29		18.5		20		17	
Red.(%)		23.7		51.3		47.4		55.3	
Ind 1	4	2	6.5	1.75	31.4	1.75	0.01	1.5	0.02
Ind 2	4	2	6.3	1.75	29.9	2.25	0.0	1.5	0.02
Ind 3	4	7	44.8	3.75	2246	3	0.01	2.75	0.17
Ind 4	4	4	225	3	6176	3.25	0.03	2.75	0.72
Ind 5	4	1	13.6	1	17.9	1	0.0	1	0.01
Ind 6	4	9	91.6	4.75	10606	4.75	0.02	4.5	0.82
Total		25		16		16		14	
Red.(%)		34.2		57.9		57.9		63.2	

TABLE III

WAFER DICING RESULTS FOR SIX TESTCASES. IASA IS THE ALGORITHM PROPOSED IN [10]; E-IASA IS OUR EXTENDED IASA HEURISTIC; ILP IS THE PROPOSED INTEGER LINEAR PROGRAMMING APPROACH; TLO REFERS TO OUR TWO LEVEL OPTIMIZATION ALGORITHM.

Symp. on Physical Design, pp. 70-77, April 2004.

[9] A. B. Kahng, I. I. Mandoiu, X. Xu, and A. Zelikovsky, “Yield-driven multi-project reticle design and wafer dicing”, *Proc. BACUS*, October 2005.

[10] A. B. Kahng and S. Reda, “Reticle Floorplanning With Guaranteed Yield for Multi-Project Wafers”, *Proc. International Conference On Computer Design*, pp. 106-110, October 2004.

[11] R. D. Morse, “Multiproject Wafers: not just for million dollar mask sets”, *Proc. SPIE*, Vol 5043, 2003, pp. 100-113.

[12] M.-C. Wu and R.-B. Lin, “A Comparative Study on Dicing of Multiple Project Wafers”, *Proc. Symposium on VLSI (ISVLSI)*, 2005.

[13] M.-C. Wu and R.-B. Lin, “Reticle Floorplanning and Wafer Dicing for Multiple Project Wafers”, *Proc. Intl. Symposium on Quality Electronic Design*, 2005.

[14] G. Xu, R. Tian, D.F. Wong, and A. Reich, “Shuttle Mask Floorplanning”, *Proc. SPIE*, Vol 5256, pp. 185-194.

[15] G. Xu, R. Tian, D. Z. Pan and M. D. F. Wong “A Multi-objective Floorplanner for Shuttle Mask Optimization”, *Proc. SPIE*, Vol 5567, 2004, pp. 340-350.

[16] G. Xu, R. Tian, D. Z. Pan and M. D. F. Wong “CMP Aware Shuttle

Cases	# part	# points	Fixed		$l=1,k=10$		$l=1,k=100$		$l=3,k=10$	
			N_w	CPU(s)	N_w	CPU(s)	N_w	CPU(s)	N_w	CPU(s)
Ind 1	1	1023940	3	0.14	3	13.7	2	2496	2	31.7
Ind 2	1	1226510	3	0.18	2	32.9	2	2790	2	68.5
Ind 3	1	2513061	4	4.59	4	442.7	4	39763	4	1069
Ind 4	1	5433742	4	73.6	4	7455	4	635342	4	18674
Ind 5	1	1717290	2	0.21	2	20.9	2	1762	2	48.3
Ind 6	1	2144469	6	3.57	5	1024	5	96521	5	2581
Total			22		20		19		19	
Red.(%)					9.1		13.6		13.6	
Ind 1	2	281189	2	0.05	2	4.87	2	372	2	9.73
Ind 2	2	341819	2	0.06	2	5.73	2	461	1.5	46.8
Ind 3	2	685663	3	3.98	3	376	3	28909	3	937
Ind 4	2	1491077	3.5	0.76	3	1937	3	93402	3	4852
Ind 5	2	463405	1.5	0.21	1.5	17.6	1	3594	1	79.8
Ind 6	2	605159	5	3.57	4	479	4	38721	4	971
Total			17		15.5		15		14.5	
Red.(%)					8.8		11.8		14.7	
Ind 1	4	56205	1.5	0.02	1.5	1.76	1.25	634	1.25	16.8
Ind 2	4	75445	1.5	0.02	1.25	3.51	1.25	337	1	39.1
Ind 3	4	152140	2.75	0.17	2.75	19.4	2.5	45827	2.5	673
Ind 4	4	316718	2.75	0.72	2.5	173	2.5	14523	2.5	483
Ind 5	4	98545	1	0.01	1	0.97	0.75	1877	0.75	13.5
Ind 6	4	133111	4.5	0.82	4	567	4	30469	4	1235
Total			14		13		12.25		12	
Red.(%)					7.1		12.5		14.3	

TABLE IV

COST EFFICIENCY OF WAFER SHOT-MAP DEFINITION STEP FOR SIX INDUSTRY TESTCASES. "# POINTS" IS THE NUMBER OF POSSIBLE WAFER CENTER LOCATIONS TO BE CHECKED TO OBTAIN AN OPTIMAL SOLUTIONS. l IS THE NUMBER OF LEVELS AND k IS THE GRID SIZE USED IN EACH LEVEL.

Mask Floorplanning", *Proc. Asia South Pacific Design Automation Conference (ASPDAC)*, 2005.

[17] Curcuits Multi-Projects, `http://cmp.imag.fr/CMP`.