Wafer Cut and Rotation for Compound Yield Improvement in 3D Wafer-on-Wafer Stacking

Bei Zhang, Vishwani D. Agrawal
Department of Electrical and Computer Engineering
Auburn University, Auburn, AL 36849, USA
bzz0004@auburn.edu, vagrawal@eng.auburn.edu

Abstract—Three-dimensional IC (3D IC) exhibits various advantages over traditional two-dimensional IC (2D IC), including heterogeneous integration, reduced delay and power dissipation, smaller chip area, etc. Wafer-on-wafer stacking is most attractive for 3D IC fabrication, but it suffers from low compound yield. To improve the compound yield, two efforts have been done in this work. First, a hybrid wafer-on-wafer stacking procedure is proposed which incorporates rotationally symmetric wafers and running repository based best-pair matching algorithm. Further, a novel manipulation of wafer cut and rotation (CR) is proposed. In this method, wafers with rotational symmetry are either cut to 4 or 2 sectors. These sectors are then rotated and used to replenish the repositories. The CR method is further combined with best-pair matching algorithm for compound yield evaluation. Simulation results show that for wafers with nine different defect distributions, plain rotation of wafers offers trivial benefits in yield, which suggests that the proposed hybrid stacking procedure is not as beneficial as what it is expected to be. Experiment results on CR method however shows significant higher yield than existing methods. The advantage becomes even more obvious with the increase of repository size and the number of stacked layers. For 3D IC with 7 layers, the relative yield increase of CR method over existing method can reach as high as 189%.

Keywords: 3D IC, wafer-on-wafer stacking, compound yield, hybrid, cut and rotation

I. INTRODUCTION

Despite many challenges, the three-dimensional integrated circuit (3D IC) is a hot topic in semiconductor industry these days [3], [8], [10]. To achieve higher levels of integration, multiple layers of active electronic component are stacked vertically in a 3D IC. Connections between layers are provided by through silicon vias (TSVs) [1], [6], [9], [13]. TSVs are short and reduce the need for long interconnects as required on planar ICs, thus reducing the delay and power consumption [3], [11]. Another promising aspect of 3D IC is the heterogeneous integration, which means that dies in the stack can be fabricated by different vendors and can be optimized according to their own technologies [3], [8], [10]. Moreover a 3D IC offers smaller device footprint, which is desirable in hand-held devices.

Currently, there are three types of layer stacking methods in 3D IC fabrication, namely, die-on-die, wafer-on-wafer and die-on-wafer. Among these methods, wafer-on-wafer stacking is most attractive. It offers the highest production throughput since each stack bonding produces a large number of stacked ICs. Other advantages of wafer-on-wafer stacking include smallest die sizes, thinnest wafers, and highest TSV densities [14], [18], [24]. Although the other two stacking methods offer higher final yield, they are harder to handle, stack, and process, besides being expensive [5], [16].

On the other hand, a bottleneck in the wafer-on-wafer stacking is its relatively low compound yield, especially for large number of stacked layers and low wafer yields. To improve the yield or more accurately predict the actual yield in industry, two kinds of efforts are worth mentioning:

1) Various matching algorithms have been proposed so as to select the best matching wafers to stack instead of stacking them randomly [14], [17], [18], [21], [22].
2) Exploiting more practical defect distribution model (like considering wafer maps with radially clustered defects [17]) or specially designing wafers (like fabricating wafer with rotational symmetry such that two wafers can be matched in more ways than one [16]) have been proposed.

Our effort in this work has three parts.

1) A hybrid wafer-on-wafer stacking procedure is proposed in Section IV, which has the advantages of several up-to-date methods [16], [21], [22]. By doing so, we hope to get improved compound yield. Analysis of the performance of this hybrid procedure is given in the experiment section. As a combination of the best practices in existing works, the hybrid procedure serves as a comparison to the novel cut and rotation method, which is the core contribution of this work.

2) A novel manipulation of wafers is introduced. To be specific, wafers fabricated with rotational symmetry are cut into either 2 or 4 sectors (called subwafers). Subwafers are rotated correspondingly to make all of them look identical. In this work, we refer to these subwafers as identical if they have the same die distribution and die orientation. These identical subwafers are then used to replenish the repository. A repository consists of wafers corresponding to a
specific layer of a 3D IC stack. The simulation results show that a simple cut and rotation method produces much higher compound yield than that of existing methods [14], [16], [17], [18], [21], [22], [24]. For example, compared with the work in [17], [21], [22], the relative improvement of compound yield can reach as high as 189%.

3) We compared the compound yield of different stacking procedures under various defect distribution models.

The rest of this paper is organized as follows. Section II introduces the background and motivation for this work. Section III discusses various aspects of wafer-on-wafer stacking which are closely related with this work. A hybrid wafer-on-wafer stacking procedure considering rotational symmetric wafer and best-pair matching algorithm is proposed in Section IV. Section V further proposes a novel manipulation to the wafer, i.e., wafer cut and rotation. Experimental results and comparisons to related work [17], [21], [22] are presented in Section VI. Section VII concludes the paper.

II. BACKGROUND AND MOTIVATION

Since a bottleneck in wafer-on-wafer stacking is the low compound yield, many researchers have proposed optimal matching algorithms to break this bottleneck. Smith et al. [18] stack wafers with same or similar wafer maps from two different repositories. Reda et al. [14] propose several matching algorithms including a globally greedy matching, an iterative matching heuristic (IMH), and global optimal matching based on integer linear programming (ILP). Verbree et al. [24] propose an iterative greedy matching algorithm, which applies globally greedy matching to only two repositories at a time. Both proposals [14], [24] are based on static repository which means none of the repositories will be replenished until they run out of wafers. Contrary to the static repository scheme, Taouil et al. [21], [22] propose the concept of running repository. After a wafer leaves its repository, a new wafer immediately enters the repository so the repositories will always be full of wafers. This new scheme of replenishing repository is proven to offer higher compound yield, and more importantly, lower run time complexity than static repository.

All of the above works consider only wafer maps with uniform defect distribution, which is not a good model for describing defect distribution in the real world. It is well known that the defects on wafers have clustering effect [7], [23], [25], [26]. A clustered defect distribution model was first considered by Singh [17] in the application to wafer-on-wafer stacking. Singh’s work shows that with the same stacking procedure, more practical wafer maps generate higher compound yield than wafer maps with uniformly distributed defects. However, the matching algorithm [17] is based on a static repository instead of running repository, which may not fully exploit the advantage of the clustered defect model. Also, in industry, there may be various kinds of defect distributions on the wafer.

Singh [16] also proposes a way to fabricate wafers with rotational symmetry. With such a characteristic, two wafers can be matched four ways to find the best match. The rotational symmetry offers a new characteristic of what matching algorithm is used. With rotations the repository size is virtually increased by the rotation number, which is helpful for the improvement of compound yield. Weaknesses of the contribution [16] include an impractical assumption of uniform defect distribution and the use of only a static repository.

Figure 1 shows four different aspects of the stacking procedures: defect distribution models, wafer manipulations, repository replenishment schemes, and mixing algorithms. Notice that these aspects are generally independent of each other and any alternative can be generally selected for one aspect without interfering with choices on others. In Figure 1, top to bottom path shows the choices made by the referenced works. For example, the leftmost path means that [13] uses a uniform defect distribution model, takes no action for wafer manipulation, and proposes greedy, IMH, and an ILP matching algorithms based on static repository replenishment scheme. Horizontally, Figure 1 organizes the existing works in an ascending order of their publication date.

To express the two motivating factors of the present work, let us examine the two rightmost paths in Figure 1:

1) If we consider the manipulation of wafer rotation [16], and wafer matching based on a running repository scheme [21], [22], the compound yield should be improved. That motivates the hybrid scheme.

2) The motivation for a cut & rotation strategy comes from the realization that compared with die-on-die and die-on-wafer stacking, the low yield of wafer-on-wafer stacking is due to the restriction of keeping all dies on a
wafer together. In other words, dies on the same wafer are all bonded together during the wafer matching process. If the wafer can be cut into several smaller sectors, the restrictions among the dies on a wafer can be reduced. Moreover, if the wafer being cut is fabricated with rotational symmetry, then each sector (subwafer) can be rotated correspondingly to make all subwafers from the same wafer look identical. As will be illustrated in this paper, greater matching flexibility and higher compound yield can be achieved.

III. PRELIMINARIES

A. Different defect distributions on the wafer

A negative binomial defect distribution model [2], [15], [19], [20] has been widely used to predict the die yield from defect density, die area, and a clustering parameter. However, this model does not provide enough information for us to generate wafer map with different defect distributions. Also because of the strict confidentiality of the industry, analytical data on real wafers are not always exposed. This explains why most publications assume a uniform distribution of defects [14], [16], [18], [21], [22], [24]. The problem with this assumption is that the unrealistic uniform distribution model always leads to a pessimistic compound yield in the wafer-on-wafer stacking procedure [17].

Based on previous literature [4], [12], nine patterns of wafer maps corresponding to different defect distributions are generated for the yield analysis of wafer-on-wafer stacking procedure. The spatial probability functions of these nine patterns are shown in figure 2 where different gray levels correspond to different levels of yield ranging from zero (black pixel) to one (white pixel).

A brief introduction of these nine patterns is:
1) A large central spot showing the low yield at the center of the wafer.
2) A shifted semi-ring showing higher yield at the lower right corner of the wafer.
3) A slightly shifted small spot.
4) A very thin centered ring showing radial yield degradation.
5) A mixed pattern of repetitive rows and shifted semi-ring.
6) A shifted semi-ring showing higher yield level at the right corner of the wafer.
7) A shifted semi-ring showing higher yield level at the lower part of the wafer.
8) A thick centered ring showing radial yield degradation.
9) A relatively thin centered ring showing radial yield degradation.

In this paper, experiments are conducted based on the wafer maps generated from Figure 2.

B. Wafers with rotational symmetry

An example of a wafer with rotational symmetry is illustrated in Figure 3(a) where the die distribution on the wafer is symmetric with respect to both the horizontal and vertical lines. The die orientation has 90° difference between adjacent quadrants. If all wafers in all repositories have this characteristic, then any pair of wafers drawn from two different repositories can be matched in any one of the four 90° rotational positions. This virtually enlarges the physical repository size four times. The wafer map introduced in [16] is only capable of fourfold rotation, here we also consider wafers capable of double rotation. As shown in Figure 3(b), the wafer will look identical after each 180° rotation if the die distribution is anti-symmetric across the vertical line, i.e., two halves of the die are oriented with 180° rotation.
C. Running repository based best-pair matching algorithm

Running repository scheme is considered in all experiments in this paper since it provably produces higher yield and lower run time complexity than the static repository. Based on such a scheme, the matching algorithm is chosen as the best-pair based algorithm [21], [22] due to its high yield. Thus, wafers from the first two repositories are matched without any restriction, and the pair producing maximum yield is selected (best-pair match). Then the pair of wafers as a whole is matched with every wafer from the next repository to find the best one (best-one match), and the same process iterates until the last repository. After one complete stack is formed, each repository is replenished immediately. This process is repeated until the production size (total number of stacks fabricated in production) is reached. Note that in the matching algorithm, the matching criterion can be of multiple choices.

D. Matching criteria

The purpose of wafer matching is to get the maximum final compound yield for a given production size. Given two pre-bond tested wafers, there are basically three criteria to find how well they match [21], [22]: (1) the number of matching good dies (MGD); (2) the number of matching bad dies (MBD); (3) the number of unmatched faulty dies (UFD). UFD is formed either by a good die overlapping a bad die or a bad die overlapping a good die. Like most publications on wafer matching consider only MGD as the criteria [14], [16], [17], [18], [24] this paper also considers MGD since evaluating the best matching criterion is not our focus here.

In this work, each pre-bond tested wafer is represented by a square matrix where element “1” indicates a good die, and element “0” indicates either a bad die or a non-existing die. A simple “AND” followed by “SUM” operation between any two matrices will yield the number of matching good dies between two wafers.

IV. A HYBRID WAFER-ON-WAFER STACKING PROCEDURE

We introduce a hybrid wafer-on-wafer stacking procedure, which incorporates the rotational symmetry of wafers [16] and running repository based best-pair matching algorithm [21], [22]. This procedure combines the merits of the best practice from the aspects shown in Figure 1.

It has been proven [16] that by simple rotation the compound yield can be improved. The reason is quite straightforward: each rotation of a symmetric wafer actually produces a new wafer map, and the repository size is virtually enlarged by how many times the wafer can be rotated (Figure 3). Accordingly, a larger repository provides more choices for wafer matching algorithms. Therefore we choose rotationally symmetric wafer in this work. We further select the running repository scheme and best-pair based matching algorithm for wafer-on-wafer stacking.

Our initial expectation from this hybrid stacking procedure is that it would produce a considerable compound yield improvement. However, detailed experimental results for this procedure in Section VI actually show only trivial improvement. That motivated the work in the next section.

V. WAFER CUT AND ROTATION FOR YIELD IMPROVEMENT

The hybrid procedure of Section IV still does not break the restrictions among all dies on a single wafer. In this section, a novel manipulation of wafer cut and rotation is presented which helps break the restrictions among the sectors of dies from a single wafer.

A. Wafers cut into sectors

Compared with just rotating the wafer, a more flexible manipulation is to cut each individual wafer to several sectors (called subwafers). If all wafers can be cut to subwafers, then a subwafer can match with any subwafer cut from the same wafer location in another repository. Previously, all subwafers of a wafer were kept together during wafer matching. By cutting the wafer, the restriction of matching all sectors of one wafer with all sectors of another wafer is eliminated.

Cutting the wafer to several sectors offers an adaptive method between wafer-on-wafer stacking and die-on-die stacking. It adaptively gains the advantages of these two different stacking methods. Comparing with die-on-die stacking, the throughput is largely increased because now each stack produces a sector of 3D ICs. Comparing with wafer-on-wafer-stacking, the yield should be improved because of the reduced restrictions between sectors on an individual wafer.

It is quite obvious that extreme cutting will produce individual dies, and it becomes die-on-die stacking. In this way, the stacking has most of the flexibility and thus produces the highest yield. However, the throughput is largely reduced, and the stacking and bonding of individual dies requires a large amount of efforts. Another extreme is not to cut at all. In that case it becomes wafer-on-wafer stacking, which produces higher throughput but with lower compound yield.

Figure 4 shows four 90° sectors cut from a wafer where the arrow indicates the die orientation within a sector. Similarly, we can cut the wafer into halves (180° sectors) or any number of sectors.

B. Subwafer rotation

After cutting the wafers into subwafers, each subwafer can only be matched to another subwafer located at the same position within the wafer. For example, the top-left subwafer (second subwafer in Figure 4) from repository 1 can only be matched to the top-left subwafer from repository 2. If all subwafers look identical, the restrictions of subwafer
locations are eliminated and the matching will be more flexible. The idea to obtain identical subwafers from an individual wafer is straightforward. If the subwafers are cut from a wafer fabricated with rotational symmetry, each of the subwafer can then be rotated by a certain degree to make them look identical. Figure 5 illustrates the cut and rotation manipulation to the wafer in Figure 3(a).

As can be seen from Figure 5, the four subwafers are first cut apart and then rotated anticlockwise by 0°, 270°, 180°, and 90°, respectively. Thus, we get four identical subwafers. Similarly, the wafer can be cut to halves and with the left-hand subwafer rotated by 180° to get two identical subwafers. After rotation, any subwafer from one repository can be matched to any subwafer from another repository. The cut and rotation method provides more choices for subwafer stacking in matching algorithms.

C. Discussion on the number of cuts

It is quite natural to think about cutting wafers with rotational symmetry to more sectors than just 2 or 4. However, if a wafer is cut to either 3 or more than 4 sectors, new challenges will appear. We make two observations. One, dies on the wafer cannot be arranged as compact as is the case for 2 or 4 sectors. In other words, there will be space wasted at the edges of each sector due to the square shape of the chip. Two, cutting a wafer to too many small sectors will generate a circular area of certain radius inside which chips cannot be printed, i.e., the area within the circle will be too small to accommodate a complete die.

Figure 6 illustrates this point where the wafer is divided into 6 equal sectors. The dotted areas indicate where there is not enough space to accommodate a full die. As can be seen, these areas are either at the edge of the sector or near the center of the wafer. All dotted central areas form a small circle where no single die can be placed within a sector.

In other words, cutting a wafer into either 3 or more than 4 sectors will waste considerable amount of space and reduce the number of available dies on a wafer. Correspondingly, the cost of producing a 3D IC will increase which may not be compensated even if the stacking yield is increased by cut and rotation. To decide how many dies are actually lost caused by sector-cutting, we need to do geometrical calculations considering the number of cuts (or sectors), the shape of the die, area of the wafer, etc. For a given number of wafers, it may be possible to find an optimal cut number (other than 4) that can produce largest number of good 3D ICs. However, to decide the optimal cut number is beyond the scope of this paper.

In this work, we only consider cutting wafers into 2 or 4 sectors depending on the manufacturer’s priority. Intuitively, if wafer throughput is the major concern cutting in 2 pieces is preferred since matching subwafers with larger size can improve the throughput. Otherwise, if compound yield is to be optimized, cutting in 4 pieces will be preferred.

Figure 7 shows the complete stacking procedure of the cut and rotation method applied to an example of 3 stacking levels. In Figure 7, initially all repositories are filled with subwafers. For a given repository size \( m \), there will be either \( 2m \) or \( 4m \) subwafers within each repository, depending on whether a wafer is cut into 2 or 4 pieces. The best-pair match between the first two repositories and the best-one match for the rest of the repositories are conducted afterwards. Consider for now that the matching is with respect to sub-
TABLE I: Different wafer manipulation methods.

<table>
<thead>
<tr>
<th>Names</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>Two wafers are matched directly</td>
</tr>
<tr>
<td>Rotation4</td>
<td>Two wafers can be matched in 4 different ways due to rotational symmetry</td>
</tr>
<tr>
<td>Rotation2</td>
<td>Two wafers can be matched in 2 different ways due to rotational symmetry</td>
</tr>
<tr>
<td>Cut and Rotation4</td>
<td>Each wafer is cut to 4 sectors and with each sector rotated for matching</td>
</tr>
<tr>
<td>Cut and Rotation2</td>
<td>Each wafer is cut to 2 sectors and with each sector rotated for matching</td>
</tr>
</tbody>
</table>

wafers instead of wafers. For each repository replenishment, there is a back-up wafer which is cut and rotated. As one subwafer leaves a repository, a new subwafer from the back-up wafer will replenish the repository, immediately. Once the back-up wafer is used, a new back-up wafer will replace it. Since running repository based best-pair matching algorithm is used in Figure 7, the run time complexity is $O((c \times m \times p \times s))$ [21], [22] where $c$, $m$, $p$ and $s$ represent number of cuts, repository size, production size and number of stacked layers, respectively.

Till now, five different manipulations of wafers are proposed. They are summarized in Table I.

VI. EXPERIMENTAL RESULTS

A. Experimental setup

We consider 200-mm wafers with edge clearance set as 5 mm. The square die we considered has a size of 31.8 mm². For the selected wafer size and die area, the number of dies per wafer is 804. Figure 2 is used to generate the nine different patterns of wafer maps. If not specified explicitly, a production of 100,000 3D ICs is targeted in the experiments for each pattern. All the experiments are repeated 1,000 times with results averaged to remove noise.

The running repository based best-pair matching algorithm is utilized in this section. In this algorithm, initially, $n^2$ $(n^2 = c \times m)$ comparisons are conducted to find the match information for all the wafer (subwafer) pairs from the first two repositories. To speed up the matching algorithm, a heap structure is used to store the match information in our simulation. Each time a pair of wafers (subwafers) leaves the first two repositories, the corresponding elements are pruned from the heap. As two new wafers (subwafers) enter the first two repositories, their relationships with the existing wafers (subwafers) are constructed and added to the heap. Once the heap is constructed, only $2n^2 - 1$ comparisons are needed each time to replenish the heap.

In this section, the five manipulations of Table I are used with running repository based best-pair matching algorithm. The names of these manipulations refer to the complete stacking procedures depending on the context.

B. Comparison of various stacking procedures for different defect distributions

In this section we examine the compound yields of different stacking procedures under nine different defect distribution models. Initially, the yield of Basic procedure with repository size 1 is calculated for nine types of patterns, respectively. Then for each type of pattern, the yield for all procedures is normalized with respect to its calculated constant. The normalized yield versus repository size for different stacking procedures and defect distributions are shown in Figure 8 with stacked layer number of 3.

The legend of Figure 8 indicate different stacking procedures. For example, Basic means the procedure uses running repository based best-pair matching algorithm, without any manipulation to wafers. Note that Rotation represents the hybrid stacking procedure proposed in section IV. The Basic Procedure here is actually a hybrid of the methods in [17] and [21], [22].

In this paragraph, we compare the performance of different stacking procedures. Regardless of what defect model is used, yield of CR procedure is always much higher than the others. The superiority of our CR procedure is because in CR the restrictions among subwafers are broken while in Rotation and Basic all subwafers are bonded together. In CR, subwafers selected from the same repository are not necessarily from the same wafer. The differences between CR and Rotation become even more obvious as the repository size grows from 1 to 50. As shown in Figures 8, there are up to 50% differences in normalized yield between CR4 and Rotation when repository size reaches 50.
Next, we evaluate the impact of cut number on the yield of the CR procedure. It is obvious from Figure 8 that CR4 constantly offers higher yield than CR2. The reason for the yield difference between these two is the greater flexibility provided in CR4. In CR4, each wafer is cut to 4 pieces reducing restrictions between subwafers and produces a virtual repository twice the size of the virtual repository of CR2.

We further evaluate the impact of rotation number on the yield of proposed Hybrid procedure. As can be seen from Figure 8, for pattern 1, 4, 8 and 9, yield of Rotation4 is better than that of Rotation2 and Basic, but the improvement is slight. Why does larger rotation number not help the hybrid procedure significantly? A possible explanation is that under pattern 1, 4, 8 and 9, bad dies are already clustered either at the center or near the edge of wafers, in which case rotating the wafer does little for aligning good dies. For the rest of the patterns, we can see the yield of Rotation and basic are the same. To explain this phenomenon, let’s re-examine the nine patterns. Of the nine patterns, only four of them (namely, pattern 1, 4, 8, 9) are symmetric to the wafer center while the rest of them are all shifted by some amount. It is obvious that given two wafer maps with the same probabilistically non-symmetric defect distribution, the best way to match them is not to rotate them at all. So even the wafer maps used in our experiments have the capability of four-fold rotation, the Rotation method will automatically avoid any rotation. Our observations suggest that for practical wafers with various defect distributions, benefits gained from simple rotation are pretty trivial.

Another interesting phenomenon is that the yield for all stacking procedures increases as repository size gets larger. This indicates relatively large repository is preferable for yield improvement. The explanation is that larger repository size provides more candidates for matching algorithms thus increasing the compound yield. Considering the extremely small repository with size 1, the wafers are stacked without any freedom for selection. However, larger repository will
C. Impact of number of stacked layers on the compound yield

In this section the impact of number of stacked layers on final compound yield is studied. The experimental results are shown in Figure 9 where y axis indicates normalized yield with respect to the yield of the Basic procedure under the same condition. In Figure 9, the number of stacked layers varies from 2 to 7, and the repository size is set to 50.

Though not shown in Figure 9, the compound yields of all procedures decrease for larger number of stacked layers. However, as can be seen, higher improvement is gained for CR4, CR2 over Rotation4, Rotation2, and Basic. CR4 and CR2 always outperform Rotation4, Rotation2 and Basic especially for situations where compound yield becomes poorer (Figure 9(b) is an exception). For example, in Figure 9(a), for 7-level stacks the normalized yield increases from 1.00 for Basic [17], [21], [22] and 1.25 for Rotation4 to almost 2.89 for CR4, which indicates 189% and 131% relative increases, respectively. Note again, compared with basic the Rotation procedure does not help at all for pattern 2, 3, 5, 6 and 7 regardless of the number of stacked layers.

D. Impact of production size on compound yield

Since running repository scheme is utilized in our work, repository pollution is unavoidable [21], [22]. Figure 10 shows how the yield decreases as the production size increases for different types of patterns. The repository size is set to 50 and the stacked layer number is selected as 3.

Initially, the yield of CR4 procedure with production size 1 is calculated respectively for different types of patterns. Then for each type of pattern, the yield for all procedures is normalized with respect to its calculated constant. In Figure 10, as the production size increases, the yield for all procedures decreases and finally stabilizes. Interestingly,
though yields of CR4 and CR2 still outperform Rotation4 and Basic, the yield advantages become less obvious for larger production size, especially for patterns with non-symmetric probability distributions. To effectively eliminate pollution and better utilize the CR method, a new mechanism to force the unattractive wafers to leave the repository in a timely manner is needed.

VII. CONCLUSION

This paper deals with the problem of low compound yield in wafer-on-wafer stacking for 3D IC fabrication. A hybrid wafer-on-wafer stacking procedure which combines the merits of several existing methods is first proposed. We further propose a manipulation method involving wafer cut and rotation (CR). In this manipulation method, each wafer is cut to either 4 or 2 sectors. Each sector is then rotated and used to replenish the repository for matching. By wafer cut and rotations, the matching restrictions for dies on a wafer are reduced and correspondingly the compound yield is improved. Extensive experiments are conducted to compare the compound yield of the proposed hybrid and CR procedures with existing works under various defect distributions. Although the yield improvement gained from hybrid procedure is trivial, the CR procedure improves the compound yield significantly regardless of the defect distributions.

The reported experiments assume wafers used in the same stack all have the same kind of defect distribution. This may not be the case in practice since wafers from different vendors may be used for 3D stacking. Even for the same manufacturer, the fabricated wafers may have different defect distributions. Because of so, more experiments are needed to study the compound yield of stacking wafers with different defect distributions. Also, the reported experiments cut the wafer to either 4 or 2 sectors for rotation. In general, there may be some optimal number of sectors for which the compound yield would be improved most. To find this optimal number, calculations based on wafer geometry are
needed. Our ongoing research explores this aspect. Another direction of future research is to develop a mechanism that can effectively force the unattractive wafers to leave the repositories so as to reduce repository pollution. Once the problem of pollution can be solved, the CR procedure is likely to reveal larger advantages.

ACKNOWLEDGMENTS

This research is supported in part by the National Science Foundation Grant CCF-1116213. The authors would like to thank Jiao Yu for her useful comments. A detailed manuscript describing this work has been submitted for a possible presentation and publication at ITC 2013.

REFERENCES


