Cost and Yield Comparison of Wafer-to-Wafer, Die-to-Wafer, and Die-to-Die Bonding

Amy Palesko, Chet Palesko
SavanSys Solutions LLC
10409 Peonia Court
Austin, TX, 78733, USA
Ph: 512-402-9943
Email: amyp@savansys.com

Abstract

There are multiple process options and technologies to consider when creating a product that requires 3D integration. Bonding two wafers (or two die) can be accomplished in various ways, such as with thermocompression, fusion, or adhesive bonding. However, the best assembly process cannot be determined by only studying the pros and cons of the bonding technology itself. There are also three main process flows to consider when pursuing 3D assembly: wafer-to-wafer, die-to-wafer, and die-to-die bonding. This paper will compare the cost and yield for each of these process flows, all of which have advantages and disadvantages depending on the application.

Activity based cost modeling will be used to construct three basic process flows, one for each bonding option. The process flows will each be divided into a series of activities, and the total cost of each activity will be accumulated. A variety of cost and yield trade-offs will be conducted using these process flows. The variables the trade-offs will focus on will include die size, throughput, incoming wafer cost, incoming wafer defect density, and residual die defect density. The goal of this analysis is to understand the variables that impact cost and yield when bonding wafers and/or die for a product that requires 3D assembly.

Key words

Bonding, Cost, Die-to-die, Wafer-to-die, Wafer-to-wafer, Yield.

I. Introduction

This study focuses on the three main process flows for 3D assembly: wafer-to-wafer, die-to-wafer, and die-to-die. Before discussing the process flows in detail, it should be noted as an introduction that there are a variety of factors to consider that are not easily included in the scope of the cost models used in this study. One example of the variety of choices to be made beyond the selection of one of these process flows is the bonding method. The models in this paper focus on thermocompression bonding. Both flows use this method to enable as much of an apples to apples comparison as possible. In reality, there are other bonding options available, such a fusion bonding or bonding that utilizes adhesives (e.g. BCB). Reflow bonding may also be an option in some cases.

Activity based cost modeling was used to construct basic process flows for this study. In activity based cost modeling, the process flows are divided into a series of activities, and the total cost of each activity is accumulated. The cost of each activity is determined by analyzing the following attributes: time required, amount of labor required, cost of material required (consumable and permanent), tooling cost, depreciation cost of the equipment, and yield loss associated with the activity.

II. Overview of Process Flows

In this section, each 3D assembly process flow is discussed at a high level—including considerations not easily reflected within the scope of this study's cost models—before the comparison method and cost models are introduced in detail.

A. Wafer-to-wafer

Wafer-to-wafer process flows generally have the highest throughput of the three flows addressed in this paper. However, this high throughput is accompanied by several limitations. Since the wafers are being bonded prior to dicing, this means that there is an obvious limit on the size of the die that can be bonded together: the top and bottom die will always be the same size. Yield is also an important factor to consider in this flow. With wafer-to-wafer bonding, there is no ability to match known good locations to other known good locations. Consequently, there is yield fall-out as good locations that would result in good die on one wafer are matched to bad locations on the other wafer. There are topography requirements to consider as well, because interconnect has to be formed over the entire wafer at once. It is also important to have a good CTE match between the top and bottom wafers [1]. Lastly, there will be some processing/handling limitations due to the fact that two wafers are being worked with and handled.

Despite the many limitations listed above, there are benefits as well. In addition to the fact that wafer-to-wafer bonding maintains the highest throughput of the three process flows, it also allows for the most accurate alignment between the two die (in wafer form) to be bonded. Required bonding overlay accuracy is an important point to consider. This is heavily dependent on application, and although it is not included in the scope of the models for this study, it is important to keep in mind when facing real technology decisions. The International Technology Roadmap for Semiconductors (ITRS) states in their Global Interconnect Level 3D-SIC/3D-SOC Roadmap that the bonding overlay accuracy in 2009-2012 should be on the order of 1.0 to 1.5um, and 0.5 to 1.0um for the 2012 to 2015 timeframe [2].

In summary, accuracy and throughput are two major benefits to wafer-to-wafer bonding. Taking this into consideration, when bonding small die to small die, wafer-to-wafer is often preferred [3].

B. Die-to-wafer

One major benefit to this process flow is that the yield is higher than that of the wafer-to-wafer flow, due to the use of known good die. Bad die can be discarded before being bonded to the base wafer. An example of a yield comparison follows. If three wafers that each have an incoming yield of 80% are stacked, and the yield of each bonding process is 95%, the resulting wafer-to-wafer yield is calculated to be 46%. However, for the die-to-wafer process, a KGD test can be performed. In this example, the test has a 90% fault coverage. This does not only apply to the two wafers that are diced, but it also applies to the base wafer; known good locations where the good die can be stacked are identified. After this test is performed and the known good die are discarded (and the known good locations on the wafer noted), if the bonding process is kept at 95%, the final yield becomes 85% [4].

Similar to the wafer-to-wafer process flow, there are still limits to the type of processing that can be done during the die-to-wafer flow due to the fact that there is still a full wafer to being handled. On the other hand, there is also a benefit in this case because one wafer can have many types of die. The limitation of having to bond same size die to each other has been removed.

Some consider the die-to-wafer process flow to be the best of both worlds [4]. One of the trade-offs to consider with this process flow is between alignment accuracy and throughput: the higher the placement alignment required, the lower the throughput [5]. Die-to-wafer is generally viewed as a good choice for low volume/high mix applications. When bonding larger die to larger die, unless a very stringent alignment accuracy is required, die-to-wafer is generally preferred [3].

C. Die-to-die

The die-to-die option generally has the slowest throughput of the three process flows. Similar to the die-to-wafer process flow, there are benefits from a yield perspective because known good die can be combined with known good die [6]. This process flow, like die-to-wafer, is generally a good choice for low volume/high mix applications, but if accuracy is not important, it can also be a high volume process. There is also more flexibility in terms of processing choices and handling options, since two die are being worked with and no full wafers have to be handled.

From a cost modeling perspective, the die-to-wafer and die-to-die bonding flows are similar and will have similar cost drivers. Therefore, wafer-to-wafer and die-to-wafer are the two process flows discussed and modeled in the limited scope of this study. It will largely be application differences that separate the die-to-wafer and die-to-die options.

D. Comparison Method

As discussed in the previous sections, there are numerous factors to consider when selecting a process flow, many of which are related more to application than to cost. There can also be many variations within any given process flow. Therefore, it is important to take note of the details surrounding the particular cost models used in this paper. Those assumptions are discussed in this section.

The two cost models used for this study are simplified versions of generic wafer-to-wafer and die-to-wafer thermocompression bonding processes. The purpose of using simplified process flows is to allow for general conclusions to be drawn about the cost drivers for each process flow. In other words, the goal of this study is to determine the major cost drivers that exist regardless of bonding method choice. More robust, detailed models would have to be constructed to examine features of specific bonding scenarios and technologies.

The cost models are utilized in two ways in this study. They are primarily used for sensitivity analyses, in which both flows are tested to determine the impact of different variables on the total cost. Their secondary purpose is to study which method is more cost effective depending on particular variables.

These cost models were designed to make all comparisons as apples to apples as possible. Because of the high number of variables that may affect both process flows, the relative cost changes are the most interesting results of this study. However, the absolute values are not negligible. Both cost models were calibrated against numbers from SUSS Microtech which indicate approximate bonding costs for die of different sizes at different alignment accuracies. This data was used to check the baseline assumptions of this paper's generic cost models [3].

The primary variables addressed in this study are the cost of the incoming wafers (pre-dicing), the defect density of the incoming wafer (or residual defect density of an incoming die, which is dependent on that wafer's starting defect density), the throughput of the bonding step, and the size of the die.

III. Process Flow and Cost Models

The two cost models used in this study are described below.

The wafer-to-wafer model begins with steps that allow for the cost and defect density of the top and bottom wafers to be set. The wafers go through a variety of preparation steps before proceeding through a thermocompression bonding process. They are then diced, and the final yield is taken into account before the process flow concludes.

The die-to-wafer model contains similar steps, except that one of the wafers (with a cost and defect density associated with it) is diced prior to bonding and the bad die are scrapped before beginning the bonding flow. Once the bonding flow begins with the die and the wafer, the wafer preparation steps from the wafer-to-wafer flow are performed on the wafer. As with the wafer-to-wafer flow, the die and wafer go through a thermocompression bonding process, a dicing step occurs, and then final yield is taken into account as one of the final steps in the process flow.

There are a few assumptions for both process flows that must be established. Neither flow accounts for any overhead or profit margin, so the results are focused on direct cost. In terms of factory assumptions, this flows assume a well-balanced and fully utilized line. Both flows are made up of steps that are set to industry standard equipment, material, and throughput values. In all analysis done for this study, die bonded together have the same size.

All of the values associated with each step were kept static for the data collection except for the following: incoming wafer cost [7], incoming wafer defect density, and thermocompression throughput (minutes per wafer or die). The die size is changed as well, which affects multiple steps in both flows (like dicing time and number of die resulting from a particular wafer size). The variables stated above are the only changeable variables in the process flows, as well as the factors which this study focuses on.

IV. Cost and Yield Comparisons

A variety of sensitivity analyses were carried out for both flows to test the sensitivity of total cost to different variables. Note that the resulting "die cost" that is referenced in all following scenarios is the diced end product after a wafer-to-wafer or die-to-wafer process has concluded. In other words, the resulting die cost is the cost of two die bonded together.

The base values used for the variables are listed below. For the sensitivity analysis, all variables were limited to the list below except for the variable that was being tested for sensitivity.

- Die size: 10mmx10mm
- Defect density: 0.3 or 0.1 defects per square cm
- Throughput: 400 chips per hour (0.15 minutes per chip) or 4 wafers per hour (15 minutes per wafer)
- Wafer cost: $2000 (per wafer)

A. Incoming Wafer Cost

The first sensitivity analysis studies the effect of incoming wafer cost on the total die cost (Fig. 1). The relationship is linear for both the die-to-wafer and wafer-to-wafer flows. This is a straight-forward and expected result: the higher the cost of any incoming materials, the higher the price of the resulting die.

B. Throughput

The relationship between throughput and die cost is more interesting. For both the wafer-to-wafer and die-to-wafer process flows, the resulting graph is a curve. This is because the key value in the cost calculation is the minutes per wafer (or minutes per die) required for the bond step. Within the industry, bonding times are generally presented as wafers or die per hour [8]. Therefore, although the bonding time increases linearly when presented as a wafer/die per hour statistic, the actual cost is based on the time it takes per wafer/die, which changes at a different rate. In the wafer-to-wafer flow, going from 2 to 4 wafers per hour results in a change from 30 to 15 minutes per wafer. The time was cut in half. However, going from 4 to 6 wafers per hour (another increase of just 2 wafers) only represents a change from 15 minutes per wafer to 10 minutes per wafer. The impact is not linear

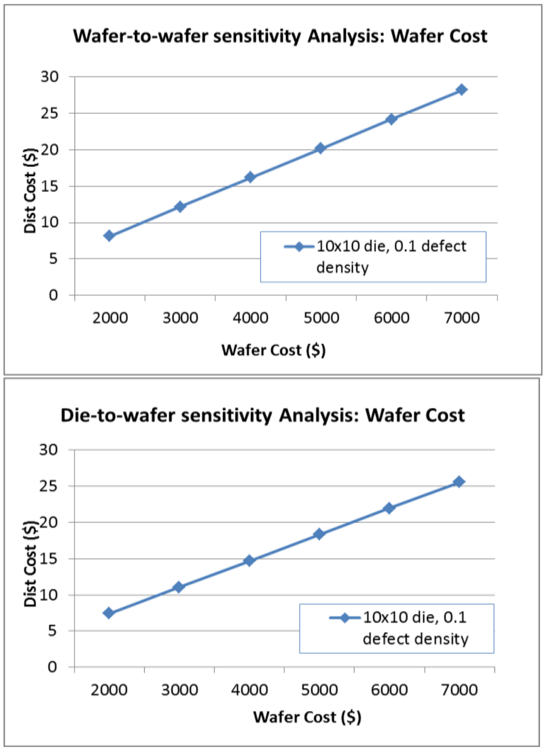

Figure 1 – Wafer Cost Sensitivity

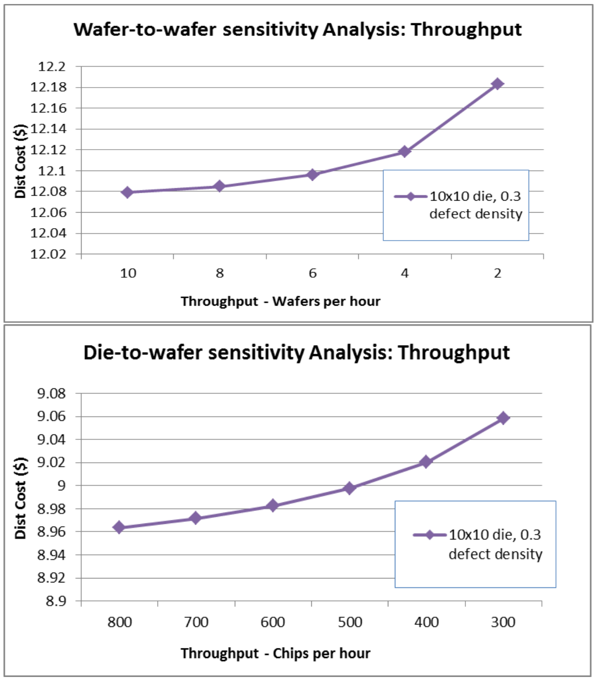

Figure 2 – Throughput Sensitivity

The most important takeaway from the graphs in Fig. 2 is that throughput has a limited effect on total die cost. The difference between bonding 10 wafers per hour or only 2 wafers per hour sounds like a large difference, but the die cost only changes by about 10 cents. The impact is just as limited for the die-to-wafer case.

C. Wafer Cost and Defect Density

The sensitivity of die cost to the incoming wafer cost was tested at multiple defect densities.

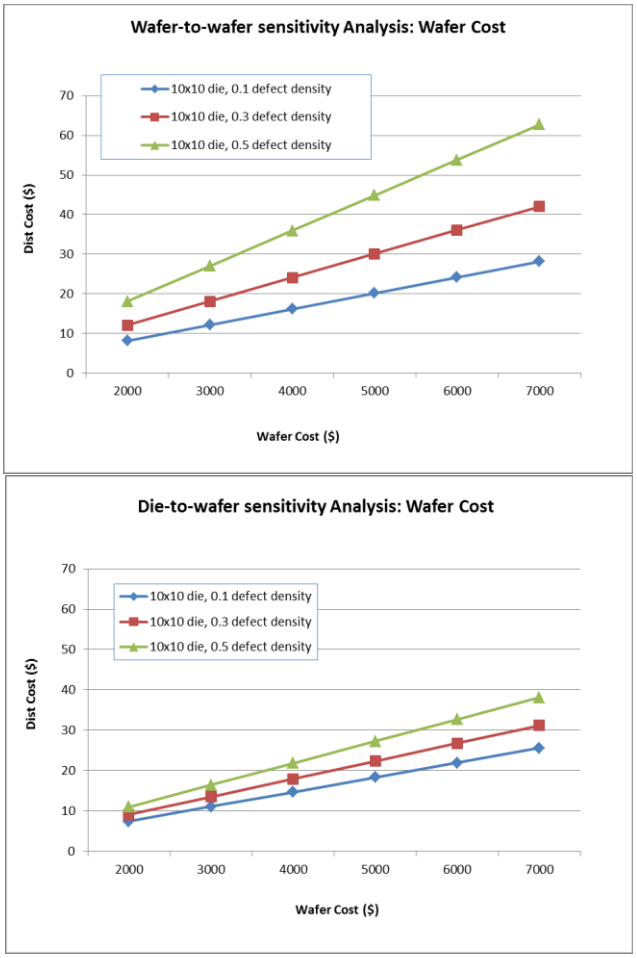

Figure 3 – Wafer Cost and Defect Density

This was done to better understand the relationship between defect density and wafer cost. For both the wafer-to-wafer and die-to-wafer flows, the slope of the curve is sharper when dealing with a higher defect density. Increasing the defect density will always increase the final cost, due to more defects causing more scrap costs; increasing the wafer cost will always increase the final cost, because the process starts with more expensive material. Increasing both at the same time essentially magnifies the effect, since more expensive material is being used from the start, and more of it is being lost due to defects. The wafer-to-wafer flow is affected more by any yield defects, given that bad die cannot be scrapped in advance. That effect can be seen in the graphs in Fig. 3.

D. Die Size and Defect Density

This section addresses the sensitivity of total bonded die cost as it relates to both die size and defect density. The relationship is not linear, as it was in the previous section when wafer cost was the main variable in question. This is because multiple process steps are affected by changing the die size. For example, changing the die size changes the number of die that will result from one wafer, and that change in the number of die is not linear. Furthermore, the dicing step parameters change because although the dicing speed is held constant for both process flows, the number of cuts changes depending on the die size.

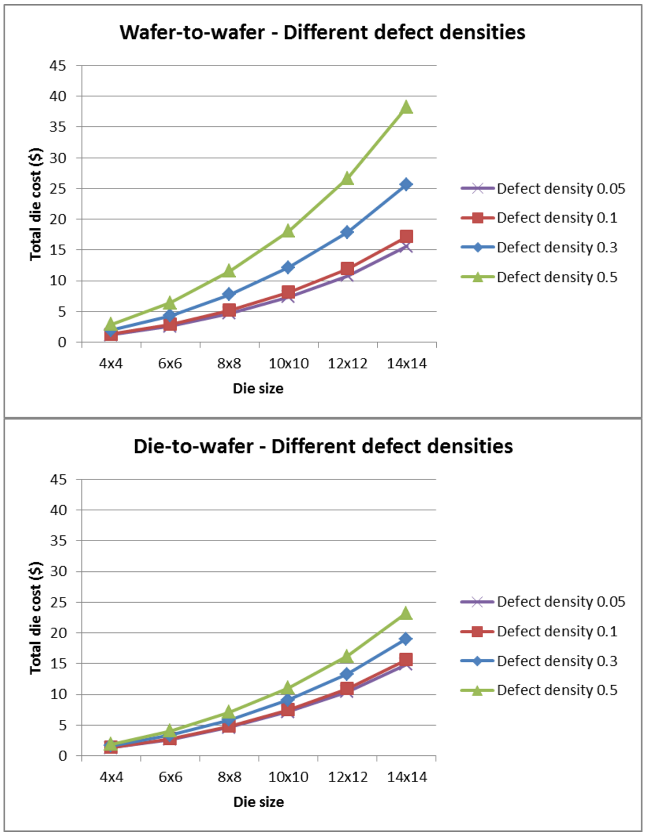

Figure 4 – Die Size and Defect Density

The most important takeaway from both graphs in Fig. 4 is that the higher the defect density, the more quickly the total cost goes up. In both graphs, all curves start near the same point. However, as the defect density is increased, the curve climbs more sharply. As demonstrated in both this and the previous section, defect density is a key factor—defects coming in with the base, top, or both wafers will result in costly scrap steps either during the flow, or worse, at the end of the flow.

E. Defect Density – Direct Comparison

This final analysis shifts the focus from a sensitivity analysis of each flow to a direct comparison between the flows. The curve in Fig. 5 charts data collected at 0.05 defects per square cm (the same defect density for both the top and bottom incoming wafers) for the wafer-to-wafer and die-to-wafer flows. Based on the parameters of this study, it is only at this low level of defect density that wafer-to-wafer appears to become a cost effective option when compared to die-to-wafer. The crossover point in Fig. 5 indicates that fact.

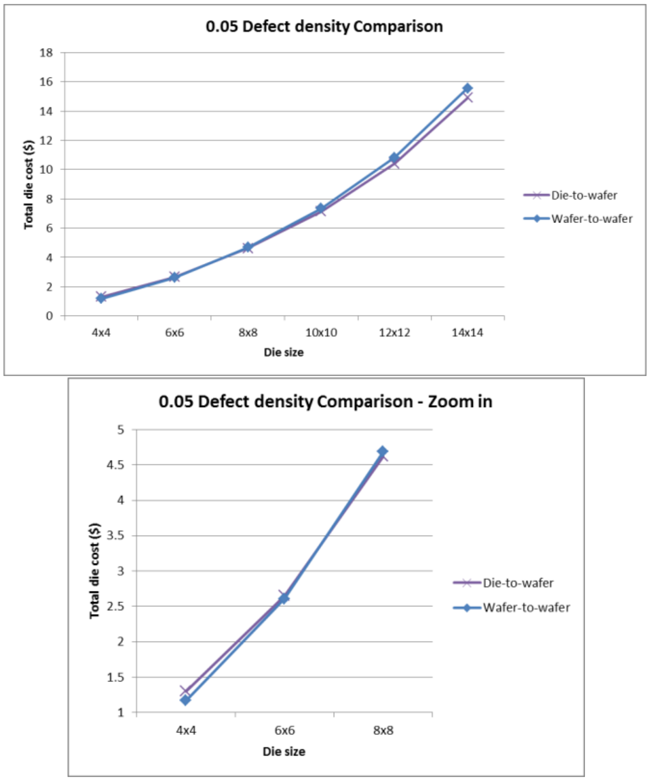

Figure 5 – Die-to-wafer and Wafer-to-wafer Comparison

The second part of Fig. 5 zooms in on the first three data points. The crossover can more easily be seen in that image. At a 4mmx4mm and 6mmx6mm die size, the data reveals that wafer-to-wafer is the cost-effective choice. Die-to-wafer only becomes cost effective in this case at an 8mmx8mm die size and above.

V. Conclusion

The original goal of this study was to focus on multiple wafer-to-wafer and die-to-wafer comparisons, highlighting which option was more cost-effective in different scenarios. However, after calibrating both basic process flows to industry data, it became clear that, given the parameters of this study, there are very few cases in which wafer-to-wafer is the most cost-effective choice. The best case for wafer-to-wafer can be made when the die sizes are small and the defect density is low.

There are numerous factors that are outside the scope of these basic models. Application will play a major part in selection of wafer-to-wafer, die-to-wafer, and die-to-die technology. Placement accuracy is also a major consideration. Those factors should be considered in a real-world application in addition to the cost study in this paper. Within the scope of this study, there was nevertheless valuable information gained. Sensitivity analysis revealed that some relationships are linear and expected (e.g. the impact of incoming wafer cost), while others are less straight-forward. Sensitivity analysis also showed that a factor like throughput does not have a large impact on the total cost even when a throughput increase appears on the surface to be a large change.

Yield (as it relates to defect density) was identified as one of the most important considerations in both the wafer-to-wafer and die-to-wafer flows. It was clearly shown that wafer-to-wafer is more sensitive to yield issues than die-to-wafer, but the impact of yield on die-to-wafer is also notable.

References

[1] E. Pabo, "Enabling Technologies for 3D Integration Aligned Permanent Bonding and Temporary Bonding," EV Group, 3D System Integration Workshop, Georgia Tech, 2011.

[2] ITRS, "Interconnect", 2009.

[3] S. Farrens, "Wafer and Die Bonding Technologies for 3D Integration," SUSS MicroTec, MRS Fall 2008 Proceedings, 2008.

[4] P. Soussan, "3D Heterogeneous Integration: Convergence Between Die Stacking and Wafer Bonding, a Fabrication Perspective", IMEC, ICRA 2013.

[5] "Chip Stacking for 3D IC", EV Group, http://www.evgroup.com/en/solutions/3d-ic/chip_stacking, 2014].

[6] Q. Chen, "A novel chip-to-wafer (C2W) three-dimensional (3D) integration approach using a template for precise alignment," Microelectronic Engineering, 2011.

[7] S. Jones, "A Simulation Study on 450mm Wafer Fabrication Costs," IC Knowledge LLC Presentation, 2010.

[8] M. Wimplinger, "Wafer-to-wafer and Chip-to-Wafer Bonding for 3D Integration," EV Group, EMC 3D 2007.