SAND2010-1138C

Yield Modeling of 3D Integrated Wafer Scale Assemblies

David V. Campbell
Sandia National Labs
Albuquerque NM, 87185
dvcampb@sandia.gov

Abstract

3D Integration approaches exist for wafer-to-wafer, die-to-wafer, and die-to-die assembly, each with distinct merits. Creation of "seamless" wafer scale focal plane arrays on the order of 6-8" in diameter drives very demanding yield requirements and understanding. This work established a Monte Carlo model of our exploratory architecture in order to assess the trades of the various assembly methods. The model results suggested an optimum die size, number of die stacks per assembly, number of layers per stack, and quantified the value of sorting for optimizing the assembly process.

Driver Application

We have investigated various 3D assembly processes as part of an internal Laboratory Directed Research and Development (LDRD) exploratory effort for the development of wafer-scale focal planes. The notion assumes construction of seamless focal planes of at least 6" in diameter, scalable to even larger sizes. Seamless means a focal plane effectively having no gaps in the detector field. Further, the complexity of the pixel used in of the application — ~100's of transistors in support of per pixel A/D and other specialized functionality— over constrained the available 2D area.

This infers multiple issues: 1) the solution requires a 3-4 layer 3D assembly to meet the 2D area constraint, and 2) yield likely precludes a realistic single-layer monolithic focal plane of this size, which infers an even less likely task of yielding a multi-wafer stack. Given the seamless construction and yield limit, it was realized early in development that the focal plane would require unique assembly of closely spaced 4-side abuted 3D stacks. Still yield would need to be well understood and modeled for successful results. If developed early, such a yield model would be effective in making assembly trades. Although we developed a model specifically for the focal plane application, the approach and results suggest broader application to 3D assembly in general.

Monte Carlo analyses have long been used to effectively model statistical processes such as the variance of production builds. [1] For this approach one must establish a set of variables that define the variance of the build and then statistically establish a large population of samples numerically representative of the build as described by the randomness of the defined variances. Then apply one or more deterministic methods, assembly processes in this case, to create a resultant population that can be further analyzed for trends and variances. The beauty of this procedure for a production build is that one can simulate many times the volume of the production without ever building a single part.

Model Definition

The baseline architecture assumed that multiple layers would be assembled to create what would be referred to as 3D stacks. Each stack represents a modular unit that includes all of the functionality for possessing a standalone 2D block of pixels. These stacks would then be assembled presumably by the same or similar 3D technique onto a "motherboard" carrier in close proximity to one another. Figure 1 illustrates a demonstration version of this assembly.

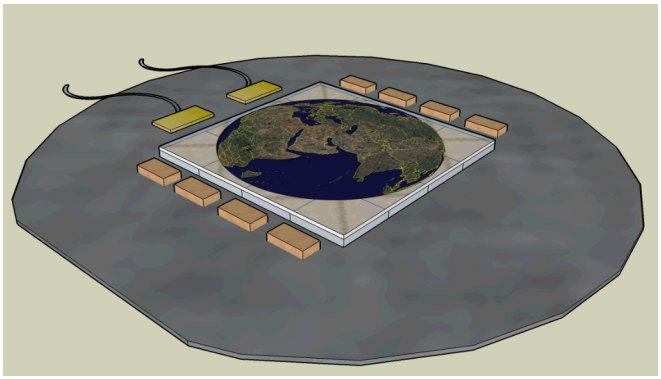

Figure 1: Wafer Scale Focal Plane Demo Cartoon

This approach assumes that by keeping the motherboard simple, high wafer scale yields could be achieved. It also assumes adequate testing to ensure known good die stacks for the assembly. The small 4x4 array chosen for the demonstration simplifies the overall task, but retains proof of all of the needed processes, including stack construction, stack assembly to wafer, a 4-side abutable design with close proximity placement, and a scaleable approach.

Key to the success of this or any other assembly depends on the determination of the following factors:

- **What wafer process yields are required?**
- **What is the optimum die size?**
- **How many layers could be stacked?**
- **Would test coverage is needed?**
- **Will the final assembly have acceptable yield?**
- **How extendable is the concept?**

Model Variables

Model variables were defined for each of the factors of interest in the model definition.

Wafer Defect Density, D_0

Two models are commonly considered to provide good first order estimates for determining die yields, given in Equation 1 and Equation 2.

$$Y_D = e^{-\sqrt{A}D_0}$$

Equation 1: Seeds Model

$$Y_D = \left(\frac{\left(1 - e^{-AD_0}\right)}{AD_0}\right)^2$$

Equation 2: Murphy Model

The Seeds model is empirically derived and generally considered a more conservative estimate. The more advanced Murphy model has lower yield at larger D0 values and greater yield at lower D0 representative of yield learning that comes with process maturity, typical of process production. Both models rely on knowledge of die area and process defeat density.

Since the modeling objective is to predict trends versus absolute numbers, the actual defect yield model contributes little significance to the actual results. Comparisons were made between the two, but the simpler Seeds model was found to be sufficient for the purposes of this work.

More important to the model outcome is the determination of D_0 requirements. It is easy to assume overly optimistic wafer yields given modern processing facilities. But wafer yields depend strongly on the design being fabricated. [2] DRAM facilities enjoy the highest yields since their process ideally targets a single repetitive design per fab line. A process line for a single irregular design, such as a high-volume X86 processor, would generally have a degraded defect density. An ASIC process line typically suffers greater degraded processing due to variances in the design pushing manufacturing tolerances. Development facilities, common to the processing of prototype or exploratory designs such as the focal plane of discussion, typically experience even worse defect density. Therefore, for comparison, the model used a range of values to envelop everything from the most optimistic processing to the realistic scenario of development facilities.

Die Size

Die Size impacts yield in multiple ways making its optimization one of the harder determinations without a model. Picking too small a die size increases die yield but simultaneously drives up the number of die stacks per assembly and number of assembly operations. In contrast selecting too large a die would limit die and stack yields. Die size also impacts the number of die per wafer. Die sizes ranging from medium to very large were selected for the model. The notion of small die was eliminated from consideration on the basis of extending the design to larger scale.

Number of Layers Stacked

Obviously the number of layers stacked directly impacts yield, but the number of layers also directly correlates to circuit design partitioning, number of interconnects, and other circuit design optimization factors. So knowing an optimum for the number of layers early in the design helps determine such circuit design trades.

Stack Sort Methods

Similar to the investigations of others [3], of primary interest in model development was knowing whether some form of optimization of assembly would significantly impact yield. Secondly, what level of difficulty is associated with the optimization. To help determine this, three methods of wafer sort were modeled.

1. ***Random Stacking.*** Random stacking represents completely arbitrary production assembly. No testing, expense, or effort is made to favor yield. Since the model generates wafer maps in a random fashion, random assembly can be implemented simply by stacking wafers in sequential order. That is, wafer stack 1 = wafer 1 of layers 1-4, wafer stack 2 = wafer 2 of layers 1-4, etc. This is just one random case. Many cases could be computed and averaged. This was not done since a statistically significant number of lots are included in the analysis.
2. ***Basic Sorted Stacked.*** Random stacking does not take into account zero or low yielding wafers. Sorted assembly applies the simplest of testing, expense, and effort to sort wafers in order of yield for assembly. For basic sorted assembly, the wafers are sorted by yield, highest to lowest. The highest yielding wafer of level 1 is then paired with the highest yielding wafer of level 2, and so forth for the appropriate number of layers. This should favor yield with a minimum of effort.
3. ***Optimized Sort.*** An actual ideally optimized sort appears analogous to the problem of optimizing a "traveling salesman" route. It requires exhaustive matching and comparison of wafers, which quickly exceeds reasonable resources relative to any possible gains. For purposes of modeling an alternate definitive method was suggested for the optimized sort, as follows:
    i. Sort all the wafers by yield, highest to lowest.
    ii. Compute all the possible stack yields for the first layer 1 wafer against all the layer 2 wafers.
    iii. Pick the highest maxima of wafer results or first maxima if more than one of equal value.
    iv. Eliminate those wafers from the mix.
    v. Repeat from step 2 until a stack is created for all wafers in a lot.

    This is a non-exhaustive solution as any given set of wafers could yield more than one maximum, but it is assumed good enough, particularly relative to the low level of resource required to perform the optimization.

Testing

The certainty to which each die is known good or known good die (KGD) statistics certainly strongly influence assembly yield. Also, assembling multiple parts together becomes the multiplicative result of the yields for each process step involved.

Assuming that all die have equal yield certainty and that the yield of each process step is well known, the formula of Equation 3 defines the aggregate assembly yield.

$$Y_A = {Y_D}^n \cdot Y_{STEP1} \cdot Y_{STEP1} \cdot \cdots$$

Equation 3: Aggregate Assembly Yield

This information is necessary in order to make absolute yield assessment. However, to simply assess relative differences, the impact of KGD yield and assembly operations was assumed to be constant per a given implementation and therefore was not directly included it in the model. That is, an assembly with *n* die on a motherboard yields based on the number of die and certainty of goodness per die independent

of whether the die is a single-layer die or a multi-layer stack. The assumption here is that die stacks can be tested as thoroughly as individual die. This factor becomes not just an assumption, but a necessity in terms of yielding die stacks. This emphasizes the importance of testing for yielding a finished assembly and bounds the scaling of the assembly.

From an operations standpoint the same holds. That is, the yield of the operation falls out as a common factor when comparing yields for die size and sorting methods. In this sense the model treats the assembly operations as unity factors. Again for determining absolute yields this is not valid. But the initial assessment of interest in the model was to determine if we could adequately yield assemblies at all. If one can not yield to ideal operations there is no reason to determine assessment of operation yields.

Other Factors

A number of other factors play into the model and could easily be modified. Two of these include wafer size and lot size. Both of these were fixed for all generated data runs. Single nominal fixed-values for each of these terms was deemed sufficient, since Monte Carlo analysis provides good indicators for trends. Larger lot sizes and wafer sizes only support assumptions for design scaling. Wafer size was constrained to 6" based on certain program drivers that dictated a specific 6" process line. Likewise, the lot size was set at 10 wafers based on the same specific process line.

Monte Carlo Analysis and Parameters

As outlined the objective was to create a Monte Carlo analysis based on the parameters and assumptions stated above.

The analysis was performed using a Perl script written specifically for the task. The basic script simply builds a population of wafers based on the input parameters, which follow:

- 100 lots of 10 wafers ea for each of 4 stack layers (4000 total wafers per case).
- 8 Die sizes and quantities assuming 6" wafers – 10:120, 15:45, 20:24, 25:14, 30:9, 35:7, 40:4, 45:4, where 10:120 means 120 die per wafer @ 10 mm.
- D_0 values of 0.01, 0.03, 0.05, 0.1, 0.3, 0.5, and 1.0 defects per square cm.
- 3 stack methods: RANDOM, SORTED, OPTIMUM.
- Stacks of 2, 3, and 4 layers.

The script first builds a database of wafer yield maps, equivalent to what might be generated from testing wafers. All wafer maps are independently and randomly generated for uncorrelated results, representative of production. The wafer map creation does not take into account systematic yield issues such as center to edge wafer gradients, etc. The analysis parameters create 4000 total wafers for each die case. With 8 die sizes and 7 defect densities that equates to 224,000 total wafer maps and 6,356,000 die. This is considered very large for the expected low volume focal plane builds. However, such quantities may be insufficient for other applications.

The script saves the wafer maps in an SQLite database for reuse. All stacking, whether 2, 3, or 4 layers and regardless of sorting method use the same wafer maps for a direct comparison of results. This correlation equates to performing multiple production runs with different processes while using the same pieceparts, thus giving an objective comparison of the processes.

From this database of wafer maps, the script assembles 1,589,000 individual die stacks for each of 2, 3, and 4 layer stacks by each of three stacking methods, ~14.3M total.

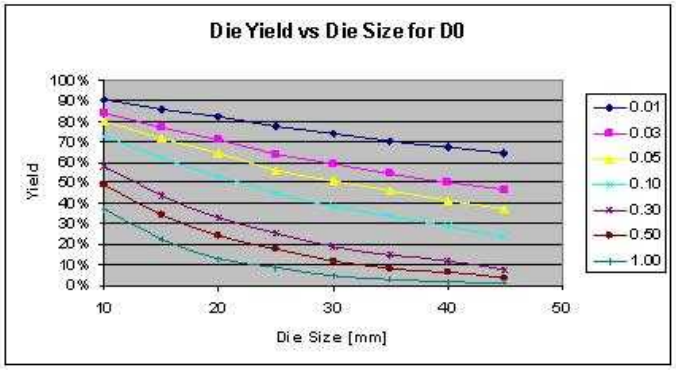

Figure 2: Die Yield vs Die Size and Defect Density

Observations

Before assessing model results a number of checks were made on the database to ensure integrity and serve as a sanity check and validation. For example, Figure 2 shows die yield as a function of die size and defect density that is consistent with Seeds model.

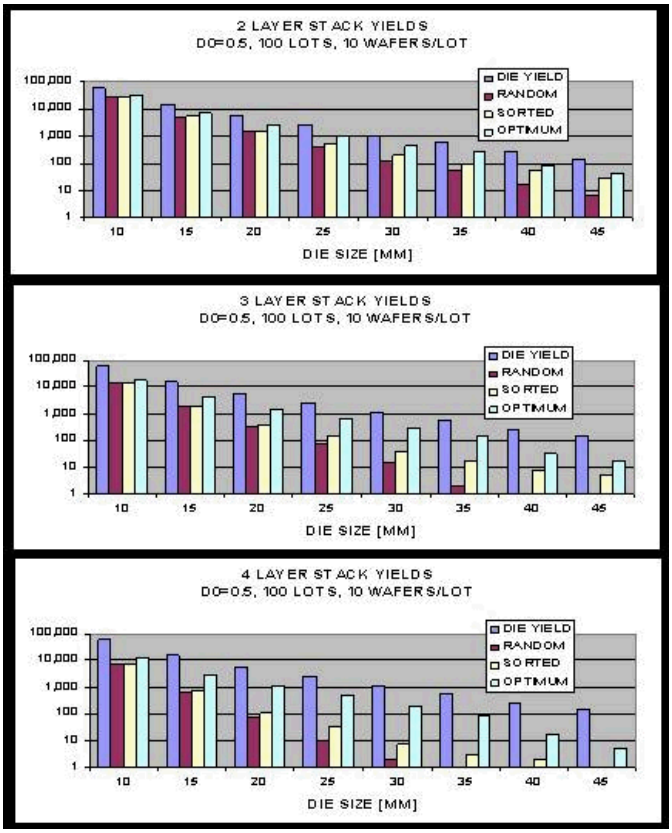

Figure 3: Stacking Method, D0=0.5

Primary interest for the model was to determine value of presorting wafers before assembly. Figure 3 shows a summary of 2, 3, and 4 layer stacks for a range of die sizes and side by

side comparison of the 3 sort methods. This case is for a D_0 of 0.5 defects per cm^2. For large die sizes Optimized Stacking clearly yields better than Sorted Stacking that clearly yields better than Random Stacking. For 2 layer stacks and 45 mm die the difference is nearly an order of magnitude. At small die size all methods yield nearly equivalent. For nominal sizes, the ratio of Optimized to Random increases roughly by the power of the number of layers stacked.

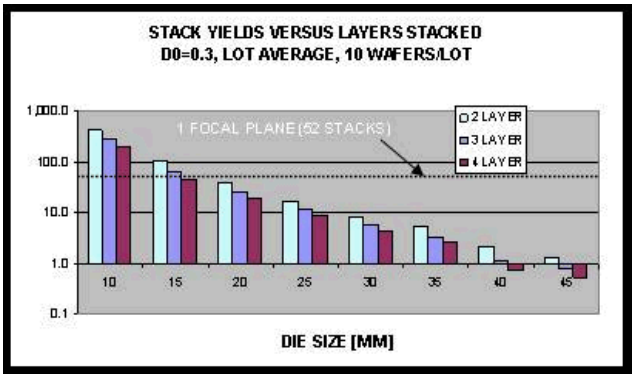

Figure 4: Stack Yields Per Lot vs Die Size

Figure 4 pictures a typical example case that shows stack yield is fairly independent of the number of layers. Note, as stated, this assumes ideal assembly operations and ideal KGD yields, which must be taken into account. The point here is the weak impact of die yield with all other factors being equal. Even without assembly process considerations the model determined that a complete baseline assembly would require a minimum of two product lots for each die layer in order to complete 1 focal plane assembly. This factored into economic and planning decisions for the program.[4]

A significant finding for the model involved determining minimum necessary processing defect level requirements for the baseline design. At a die size of `20 mm on 6" wafers a defect density of 0.175 defects per cm^2 translates to only 1 successful assembly per lot on average, which does not take into account lot-to-lot variances.

Conclusions

The described Monte Carlo analysis proved useful in determining yields and trades in the fabrication of wafer scale focal plane assemblies. Observations from the model suggest significant value in implementing a straightforward wafer sorting process based on yield in order to increase overall focal plane array. These observations further influenced factors such as baseline design die size and the number of layers of circuit partitioning for ensuring a successful focal plane result.

Acknowledgments

Special thanks to Subhash Shinde, 3D Integration Project Lead at Sandia for support of this work.

Additional thanks to Randolph "Rex" Kay, Principle Investigator for the focal plane research project. Rex also suggested the process for optimally mating wafers.

This research is supported by the Laboratory Directed Research and Development program at SNL. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

References

1. Metropolis, N.; Ulam, S. (1949). "The Monte Carlo Method". Journal of the American Statistical Association, Vol. 44, pg 335–341.
2. Web published Technology and Defect Trends, IC Knowledge.
3. Gregory Smith, Larry Smith, "Maximizing the Functional Yield of Wafer-to-Wafer 3-D Integration", IEEE Trans on Very Large Scale Integration Systems, Vol 17, No. 9, pg 1357-1362, September 2009.
4. P. Mercier, S.R. Singh, K. Iniewski, B. Moore, P. O'Shea, "Yield and Cost Modeling for 3D Chip Stack Technologies", IEEE 2006 Custom Integrated Circuits Conference, pg 357-360.