Reducing the Leakage and Timing Variability of 2D ICs Using 3D ICs

Sherief Reda
Division of Engineering
Brown University
Providence, RI 02912
sherief_reda@brown.edu

Aung Si
Division of Engineering
Brown University
Providence, RI 02912
aung_si@brown.edu

R. Iris Bahar
Division of Engineering
Brown University
Providence, RI 02912
iris_bahar@brown.edu

ABSTRACT

This paper examines the ramifications of using 3D integration technology on the leakage and timing variability of integrated circuits. We develop models that estimate the outcome of mapping a 2D design onto a 3D stack from a process variation perspective. We statistically prove and experimentally demonstrate that 3D integration is a useful technique to combat process variations even if the die/wafers layers involved in 3D stacks are integrated blindly without any parametric tests prior to integration. We further show that if individual die parametric testing information is available, then it is possible to drastically reduce the impact of process variations. We develop fast, near optimal integration strategies based on recursive matching techniques. Our results show that 3D integration can reduce the variability in leakage and timing of planar ICs by around 50% without any testing and by more than 90% with additional test requirements.

Categories and Subject Descriptors: B.7.1 [INTEGRATED CIRCUITS]: Types and Design Styles—*Advanced technologies*.
General Terms: Performance, Design
Keywords: 3D integrated circuits, leakage, timing, variability.

1. INTRODUCTION

Advanced sub-wavelength semiconductor fabrication techniques have resulted in substantial amounts of process variations. A myriad of physical phenomena contribute to the observed variations [1, 5]. Process variations translate to variations in the key electrical parameters (i.e., speed and power) of the devices and interconnects of an integrated circuit. These variations increase the uncertainty in the outcome of the manufacturing process of integrated circuits (ICs), which could lead to significant differences in speed and power from the nominal values [2]. Predictive technology models expect that the severity of process variations, as well as the sources that contribute to their presence, will further increase in future semiconductor manufacturing technologies.

Three-dimensional integration with thorough-silicon vias (TSVs) is an emerging technology that promises to improve the performance and form-factor of integrated circuits. In 3D ICs, multiple layers (or wafers) are manufactured separately and then stacked and bonded vertically. The inter-die communication is carried out by TSVs that span the die of the 3D stack. There are three main bonding techniques for 3D integration: wafer-to-wafer, die-to-wafer, and die-to-die. All these bonding techniques offer some flexibility that can be exploited to devise integration strategies that maximize the yield of 3D ICs [3, 6, 7].

This paper provides another strong incentive to switch from planar technology to 3D integration by showing that partitioning and mapping a 2D design into a 3D integrated stack, it will be possible to significantly reduce the impact of process variations on leakage and timing. The contributions of this paper are as follows.

- We propose a generic method in Section 2 to realize process corner simulation for 3D ICs in order to predict the behavior of a 3D design at various process variation corners.
- Using statistical analysis in Section 3, we prove a key result that shows that 3D integration provides a convenient method to reduce the impact of process variations as measured by the 6σ spread around the leakage/timing mean. Our result assumes no information of parametric testing is available.
- In case parametric test information for the individual die is available, we provide strategies in Section‘4 to combine the die/wafers during 3D integration to dramatically reduce the impact of process variations. For two layers, we propose an optimal technique with a runtime of $O(n \log n)$, where n is the number of die, and we extend it heuristically for k layers to run in $O(kn \log n)$.
- We provide strong experimental results from a simulated design and a production chip to clearly support our argument that 3D integration is a viable method to reduce manufacturing variability.

2. PROCESS CORNER SIMULATION FOR 3D ICS

One of the main techniques to characterize the impact of process variations, or in general on-chip process variations, is *process corner simulation*. Given a characterized standard cell library from a vendor at various corners (e.g., gate lengths, threshold voltages, temperatures), circuit designers simulate their design using various corner combinations from the library. The outcomes from the corner simulation process provide a good overview of the variability in speed and power that could impact the design. Typically timing and leakage power are strongly correlated [2], where fast ICs have high leakage, and ICs with low leakage have slow timings. That is, there is a “sweet spot" where the best power/performance trade-off is met. Integrated circuits with leakage and timing worse than the design specifications must be discarded leading to a reduction in yield. Reducing the 6σ of timing/leakage reduces the fabrication variability and improves the yield by reducing the number of ICs with slow timing and the number ICs with high leakage.

ISLPED'09, August 19–21, 2009, San Francisco, California, USA.

Our objective is to extend the process corner simulation procedure to handle 3D ICs. One of the interesting aspects of 3D technology is that the different layers that belong to the same 3D stack are manufactured separately and then bonded afterwards. Thus, when a 2D design is partitioned and mapped into a 3D IC, each partition could be impacted differently and independently by process variations. Essentially 3D technology substitutes inter-die variation trends for intra-die variation trends.

When a design is partitioned among two or more layers in a 3D stack, it would be necessary to simulate the corners of each design partition, and use the results to *compose* the final corners of the 3D stack, where each design partition has M corners and thus the 3D stack has M^2 corners. For k layers, the number of corners is M^k. Simulating each design partition separately is feasible by substituting all TSVs by their equivalent RC model. Throughout this paper, we assume a $5\mu m \times 5\mu m$ TSV with 43 $m\Omega$ and 40 fF. Composing the final 3D corners depends on the interactions between the various design layers. For example, consider a 64-bit ripple-carry adder design. The adder is a simple design that we will use frequently throughout this paper to illustrate the impact of some of the proposed ideas. It has the advantage of being modular and straightforward to split across multiple layers. We execute a process corner simulation of the adder using SPICE with a 90 nm library. We analyze the process corners of this design in three 3D configurations: a 2-layer 3D stack where each layer has a 32-bit adder, a 4-layer 3D stack where each layer has a 16-bit adder, and an 8-layer 3D stack where each layer has an 8-bit adder. The corners of the individual layers are first obtained, and then the corners of the 3D design are calculated by considering all corner combinations from every layer. We plot the corners of the three 3D design configurations together with the corners of the planar design in Figure 1. From the results, as the number of layers in the 3D stack increases, the total number of corners increases. Furthermore, it appears from the results that the process variability is "tightening" in both leakage and timing as the number of layers in the 3D stack increases. This observation is formalized in the next section.

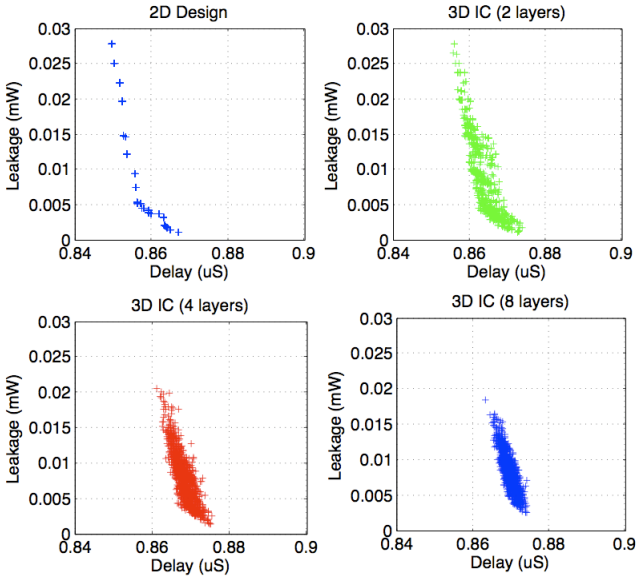


Figure 1: Design corners for 64-bit adder in four configurations: a traditional planar configuration; a 2-layer 3D stack where each layer has a 32-bit adder; a 4-layer 3D stack where each layer has a 16-bit adder; and an 8-layer 3D stack where each layer has a 8-bit adder.

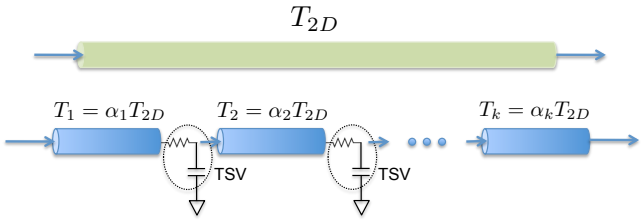


Figure 2: Critical path in 2D and 3D designs.

3. VARIABILITY REDUCTION USING 3D ICS

In this section we prove that by splitting a 2D design into multiple die that are manufactured independently and then stacked together, it is possible to reduce the timing and leakage variability as measured by the manufacturing 6σ. This result occurs naturally without the need to execute any form of testing or application of optimized integration strategies. Our result is generic for any design and a proof is as follows.

Theorem 1. Assume a given 2D design with critical path delay of mean T_{2D} and timing variance $\sigma^2_{2D\_T}$, and a total leakage of L_{2D} and leakage variance $\sigma^2_{2D\_L}$. If the design is partitioned and mapped to k layers suitable for 3D integration, then (1) the timing variance of the 3D ICs, $\sigma^2_{3D\_T}$, is at least equal to $\frac{\sigma^2_{2D\_T}}{k}$, and at most equal to $\sigma^2_{2D\_T}$ (i.e., $\frac{\sigma^2_{2D\_T}}{k} \leq \sigma^2_{3D\_T} \leq \sigma^2_{2D\_T}$); and (2) the leakage variance of the 3D ICs, $\sigma^2_{3D\_L}$, is at least equal to $\frac{\sigma^2_{2D\_L}}{k}$ and at most equal to $\sigma^2_{2D\_L}$ (i.e., $\frac{\sigma^2_{2D\_L}}{k} \leq \sigma^2_{3D\_L} \leq \sigma^2_{2D\_L}$).

Proof: First we prove the timing variability part. Assume that the critical path is divided into k sections, such that $\alpha_i T_{2D}$ gives the fraction of delay of the section that straddles layer i. Figure 2 provides an overview of the critical path partitioning. Then the delay of the critical path section at layer i is given by

$$\forall i = 1 \ldots k : T_i = \alpha_i T_{2D} + D_{TSV}, \text{such that} \sum_{i=1}^{k} \alpha_i = 1, \quad (1)$$

where D_{TSV} is the constant delay introduced by a TSV. Because the k layers are manufactured separately (i.e., the corners of different die in the stack can be considered as independent), then we can calculate the mean and variance of the 3D IC timing T_{3D} as follows:

$$E[T_{3D}] = E[\sum_{i=1}^{k} \alpha_i T_{2D} + D_{TSV}] = T_{2D} + kD_{TSV} \quad (2)$$

$$\sigma^2_{3D\_T} = \text{Var}[T_{3D}] = \text{Var}[\sum_{i=1}^{k} \alpha_i T_{2D} + D_{TSV}] = \sigma^2_{2D\_T} \sum_{i=1}^{k} \alpha_i^2 \quad (3)$$

Note that we assumed that the variance in the TSV delay is negligible; such an assumption is quite reasonable, as the dimensions of TSVs are in the μm range and are not impacted by typical subwavelength photolithographic ailments. It is easy to show that the quantity $\sum_{i=1}^{k} \alpha_i^2$ under the constraint $\sum_{i=1}^{k} \alpha_i = 1$ attains a minimum value of $\frac{1}{k}$ for balanced partitions and a maximum value of 1 for completely imbalanced partitions. Thus, we reach the first conclusion that

$$\frac{\sigma^2_{2D\_T}}{k} \leq \sigma^2_{3D\_T} \leq \sigma^2_{2D\_T}.$$

For the leakage part, the same argument can be followed. Assume for the k partitions that $\beta_i L_{2D}$ gives the leakage of the subcircuit mapped to layer i and $\sum_{i=1}^{k} \beta_i = 1$. Because the k layers

are manufactured independently, then we can calculate the mean and variance of the 3D IC leakage L_{3D} as follows:

$$E[L_{3D}] = E[\sum_{i=1}^{k} \beta_i L_{2D}] = L_{2D} \quad (4)$$

$$\sigma^2_{3D\_L} = \text{Var}[L_{3D}] = \text{Var}[\sum_{i=1}^{k} \beta_i L_{2D}] = \sigma^2_{2D\_L} \sum_{i=1}^{k} \beta_i^2. \quad (5)$$

Note that for Equation (4), we assumed that the 3D IC is going to operate at the same temperature as the 2D IC. While 3D stacking is widely believed to increase the junction temperatures due to the increases in power densities, our technique will reduce leakage variability which helps to counter balance the increase in power density. From Equation (5), we reach the second conclusion that

$$\frac{\sigma^2_{2D\_L}}{k} \leq \sigma^2_{3D\_L} \leq \sigma^2_{2D\_L}. \quad \blacksquare$$

The result of Theorem 1 provides the main motivation of this paper as reflected by its title. Essentially, mapping a 2D design into a number of die that are manufactured independently and then later integrated randomly (or blindly) without any test into a 3D IC, can reduce the variability 6σ. This reduction in variability leads to large improvements in yield and manufacturing costs as it simultaneously reduces both the number of ICs with slow timing and the number of ICs with high leakage. Thus, we argue that another advantage for 3D ICs is to reduce the impact of process variations.

4. OPTIMAL MITIGATION OF PROCESS VARIATIONS

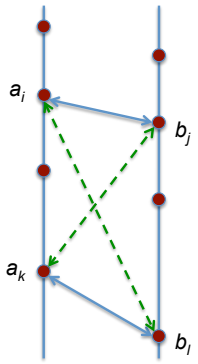
In this section we investigate the potential of further reducing the effects of manufacturing variability by using optimal integration strategies that first test the pre-bond individual die and then use the test results to stack the die optimally to reduce the variability 6σ. The flexibility of die-to-wafer and die-to-die integration allows the possibility to match die from different layers to optimize the parametric yield and/or maximize performance [3]. In this section we develop a novel insight that shows that under certain classes of functions, optimal integration can be achieved in $O(n \log n)$ where n is the number of die per layer.

First consider the *base case* where $k = 2$. In this case we can formulate the problem as a weighted matching problem on the complete bipartite graph $G = (U, V, U \times V)$, where $|U| = |V| = n$ is the number of die per layer available for 3D integration. Consider the case where a circuit is split into two partitions and mapped into two layers of a 3D IC where the leakage of the first layer is a random variable denoted by L_1 and the leakage of the second layer is a random variable L_2. The actual values taken by the random variables are obtained from the corner simulation procedure or after testing the individual die. We will denote the n instances of random variable L_k by $L_{k,1}, L_{k,2}, \ldots, L_{k,n}$. Note that the order of the integration will not impact the average value of the distribution $\mu = E[L_1 + L_2] = (\sum_{i=1}^{n} L_{1,i} + \sum_{j=1}^{n} L_{2,j})/n$. To achieve our objective of minimizing the variability, we set the weight of each edge w_{ij} connecting node $i \in U$ to node $j \in V$ to be equal to $\frac{(L_{1,i}+L_{2,j}-\mu)^2}{n-1}$. Although it might appear that a traditional matching algorithm (e.g., Hungarian algorithm) could find the optimal matching in $O(n^3)$ [4], we prove that the problem has a special structure that enables it to be solved in only $O(n \log n)$.

We associate with each node $i \in U$ a real number a_i and with each node $j \in V$ a real number b_j. For our variance-minimization formulation, we set $b_j = L_{2,j}$ and $a_i = L_{1,i} - \mu$. In this case for all i and j, the weight w_{ij} of arc (i, j) can be expressed as $\frac{(a_i+b_j)^2}{n-1}$. If we sort the nodes in U and V before matching such that $a_1 \geq a_2 \geq \ldots a_n$ and $b_1 \leq b_2 \leq \ldots b_n$, then the following statement is true.

Theorem 2. $M = \{(i, i)|i = 1, 2, \ldots, n\}$ is a minimum-variance matching.

Proof: The proof is by contradiction. If an optimal matching differs from $M = \{(i, i)|i = 1, 2, \ldots, n\}$ then it contains at least one pair of crossed arcs (i, l) and (k, j) as illustrated in the adjacent figure. Let such a matching $M_1 = M + (a_i, b_l) + (a_k, b_j) - (a_i, b_j) - (a_k, b_l)$. Crossing such pairs of arcs leads to a change Δ in the total matching weight by



$$\begin{aligned} \Delta &= \frac{(a_i + b_l)^2 + (a_k + b_j)^2}{n-1} - \frac{(a_i + b_j)^2 + (a_k + b_l)^2}{n-1} \\ &= \frac{2a_i b_l + 2a_k b_j - 2a_i b_j - 2a_k b_l}{n-1} \\ &= 2\frac{(a_i - a_k)(b_l - b_j)}{n-1} \end{aligned}$$

By construction, $a_i \geq a_k$ and $b_l \geq b_j$ and thus $\Delta \geq 0$. Consequently crossing a pair of arcs can only increase the variance. $\blacksquare$

The implication of Theorem 2 is that it is possible to obtain the minimum variance matching in $O(n \log n)$ by simply sorting the die of the first layer in descending order according to their speed or leakage and sorting the die in the second layer in ascending order according to their speed or leakage and then match in order.

The proposed matching algorithm can be heuristically extended to any number of layers $k \geq 2$ using a recursive framework as outlined by the algorithm in Figure 3. Given the leakage results of every die in the k-layer stack, the algorithm recursively finds the best integration of die in the first $k - 1$ layers (Step 3) and then sorts the resultant stacks in descending order (Step 4). Then the algorithm combines the sorted stacks in order with the die in layer k after sorting in ascending order (Step 1). The overall runtime is $O(kn \log n)$. The extended algorithm is not optimal and can be only regarded as a "good" heuristic. As an example of its non-optimality, consider a 3D integration process with three layers where there are two die available for each layer. Consider the following (albeit artificial) leakages, layer 1: 1 W and 2 W; layer 2:

Procedure: MATCH($\cdot$)
Input: Die leakage/timing characterization results for some k layers $(L_1, L_2, \ldots, L_k)$
Output: Combined die in optimal order.

1. sort the die in layer k, $L_k = \{L_{k,1}, \ldots L_{k,n}\}$, in descending order according to their leakage/timing corners.
2. if $n = 1$ then return L_k.
3. R=MATCH$(L_1, L_2, \ldots, L_{k-1})$.
4. sort R in ascending order.
5. return $R + L_k$.

Figure 3: Procedure MATCH($\cdot$) to minimize the variability.

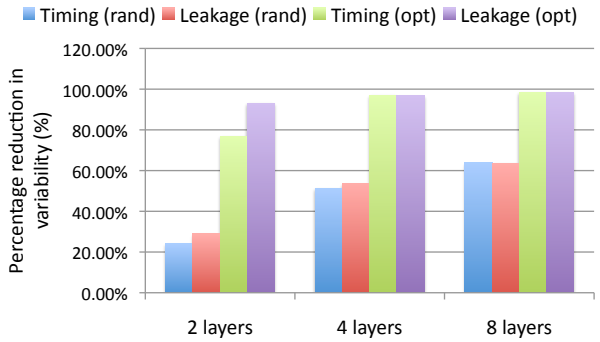


Figure 4: Reduction in timing and leakage variability over 2D design for the 64-bit adder design.

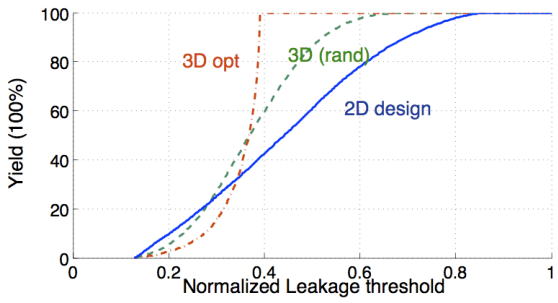


Figure 5: 2D versus 3D yield fabrication as a function of the leakage threshold using leakage characterization data from a production chip.

1 W and 0.01 W; and layer 3: 2 W and 4 W. The minimum-variance combinations are: $1 + 0.01 + 4 \approx 5$ W and $2 + 1 + 2 = 5$ W. On the other hand, the minimum variance combination for the first two layers is: $1 + 1 = 2$ W and $2 + 0.01 = 2.01$ W.

For our illustrative 64-bit adder design, we calculate the percentage reduction in variability of both timing and leakage (as computed by 6σ) using 3D integration in contrast to a 2D design. We evaluate both blind and optimal integration as computed using the algorithm MATCH$(\cdot)$ in Figure 3. In the bar-plot of Figure 4, we report the results for 3D stacks with 2-layers, 4-layers, and 8-layers. The results show that as a 2D design is partitioned and mapped into a 3D stack then the variability of the leakage and timing decreases as the number of layers increases even if no special test or integration strategies are deployed.

In a final confirmation to our results, we evaluate leakage variability effects for a production chip to be implemented using 3D technology. Our original data set comes from leakage characterization measurements of 244 wafers manufactured with 90 nm technology. We split the 244 wafers into four sets of 61 wafers each. Then we simulate the impact of integrating the die from the four sets together to form 4-layer 3D ICs. For each configuration setup, we consider both random/blind integration assuming parametric testing is not conducted, and optimal integration in case leakage characterization results are available. We assume die-to-wafer or die-to-die integration. To demonstrate the impact of leakage variability reduction, we set a leakage threshold above which an IC is discarded. We vary the leakage threshold and compute the yield for each threshold. We plot our results in Figure 5, where we give yield curves for the 2D design (blue solid line) and the mapped 3D IC design. We plot the results under both random integration with no test information available (dashed green line) and using optimal integration with available test information (red lines). As expected from the discussions and the theoretical analysis, 3D integration reduces the leakage variability and the yield. For example at a normalized leakage threshold of 0.5, the yield of 2D ICs is around 60%. The yield of the equivalent 3D ICs with blind integration is 81%, and the yield of the 3D ICs with optimal integration given test information is 100%.

Our results have significant implications as they demonstrate that 3D technology has the potential to drastically reduce the impact of manufacturing variability. This reduction adds to existing touted benefits of 3D technology such as performance, hybrid integration, and form factor. The reduction of the impact of process variations directly leads to a big improvement in the yield. Our results also show the importance of assessing the yield improvement attained from knowing parametric leakage/timing test information versus the costs incurred by testing [7]. As testing costs for 3D ICs are likely to be more expensive than 2D ICs, these costs have to be evaluated against any potential yield improvement. Regardless, blind integration gives a "free" significant yield improvement.

5. CONCLUSIONS

In this paper we have advocated using 3D integration technology to combat leakage and timing variability in 2D ICs. We have extended the process corner simulation procedure to handle 3D technology. Our corner simulation method transparently handles variations in both leakage and timing. We have also proved using statistical analysis that blind integration will reduce the variability in timing and leakage without the need to incur any extra testing costs. If testing information is available, then further significant reductions are attainable. We have proposed fast optimal integration techniques based on recursive matching, and we have quantified the impact of this improvement on the yield using realistic experimental models. Our results provide an additional compelling incentive to switch to 3D technology.

Acknowledgments

S. Reda would like to thank Sam Gu and Matt Nowak from Qualcomm corporation for the useful discussions on this topics. The work of S. Reda and A. Si were supported by a generous gift from Qualcomm Corporation.

6. REFERENCES

[1] D. Boning and S. Nassif, "Models of Process Variations in Device and Interconnect," in *Design of High-Performance Microporcessor Circuits*, 1st ed., A. Chandrakasan, W. J. Bowhill, and F. Cox, Eds. IEEE Press, 2001, pp. 98–115.

[2] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter Variations and Impact on Circuits and Microarchitecture," in *Proc. Design Automation Conference*, 2003, pp. 338–342.

[3] C. Ferri, S. Reda, and R. I. Bahar, "Strategies for Improving the Parametric Yield and Profits of 3D ICs," in *Proc. International Conference on Computer Aided Design*, 2007, pp. 220–226.

[4] E. Lawler, *Combinatorial Optimization: Networks and Matroids*. Holt Rinehart and Winston, New York, 1976.

[5] M. Orshansky, L. Milor, and C. Hu, "Characterization of Spatial intrafield Gate CD Variability, Its Impact on Circuit Performance, and Spatial Mask-Level Correction," *IEEE Transactions on Semicondictor Manufacturing*, vol. 17(1), pp. 2–11, 2004.

[6] S. Reda, G. Smith, and L. Smith, "Maximizing the Functional Yield of Wafer-to-Wafer 3D Integration," *IEEE Transactions on VLSI Systems*, to appear, 2009.

[7] L. Smith, G. Smith, S. Hosali, and S. Arkalgud, "3-D Integration: It All Comes Down to Cost," in *3-D Architectures for Semiconductor Integration and Packaging*, 2007.