Implementing Defect Tolerance in 3D-ICs by Exploiting Degrees of Freedom in Assembly

M. Tauseef Rab, Asad Bawa, and Nur A. Touba

Computer Engineering Research Center
Department of Electrical and Computer Engineering
University of Texas, Austin, TX 78712-1084
Email: {tauseefrab, bawa}@utexas.edu, touba@ece.utexas.edu

Abstract

When assembling a three-dimensional integrated circuit (3D-IC), there are several degrees of freedom including which die are stacked together, in what order, and with what rotational symmetry. This paper describes strategies for exploiting these degrees of freedom to reduce the cost and complexity of implementing defect tolerance. Conventional defect tolerance schemes involve bypassing defects by reconfiguring the circuitry so that system operation is performed using defect-free circuitry. Explicit reconfiguration circuitry is required to perform the reconfiguration, and the power distribution network must be designed to support all redundant elements. The schemes proposed in this paper use the degrees of freedom that exist when a 3D-IC is assembled at manufacture time to implicitly bypass manufacturing defects without the need for explicit reconfiguration circuitry. Defects are identified during manufacture test, and the 3D-ICs are assembled in a way that avoids the use of the defective circuitry. It is shown that leakage power and performance overhead for defect tolerance can be significantly reduced.

Keywords: defect tolerance; fault tolerance; three-dimensional ICs; reconfiguration

1. Introduction

Three-dimensional integrated circuits (3D-IC) using through-silicon vias (TSVs) is an important new technology that provides a number of significant advantages including increased functional density, shorter interconnect, higher performance, and lower power. Stacking in a 3D-IC can be done wafer-to-wafer (W2W), die-to-wafer (D2W), or die-to-die (D2D). W2W allows higher manufacturing throughput, but achieving a good compound yield is difficult. In D2W and D2D, pre-bond testing can be used to screen out defective die and use only known-good die (KGD) when constructing the stack.

As technology continues to scale to smaller dimensions, with increasingly complex manufacturing processes, the ability to reliably manufacture 100% defect-free circuitry becomes a significant challenge. It is widely recognized that future technologies will need to rely on defect tolerance techniques. The cost benefits of using defect tolerance techniques to boost yield are becoming increasingly compelling. Conventional defect tolerance involves adding redundant elements and switches in the design. After manufacture, testing is performed to locate defects, and then reconfiguration using the switches is performed to bypass the defects so that system operation is performed using defect-free circuitry [Koren 98].

When assembling 3D-ICs, there are several degrees of freedom including which die are stacked together, in what order, and with what rotational symmetry. The idea proposed here is to implement defect tolerance by using these degrees of freedom to implicitly bypass defects without the need for explicit reconfiguration circuitry. For D2D stacking, the dies that are used to assemble the 3D-IC are manufactured and tested resulting in a repository of dies in which the location of the defects is known. From this repository, the goal is to assemble the maximum number of 3D-ICs (i.e., maximize yield) in which the defects are bypassed so that defect-free system operation can be performed. The key idea is to make the design such that the computing elements that are used differ depending on which layer or which rotational symmetry is used. So a die with a defect in a certain location may be usable in certain layers or certain rotational symmetries, but may not be usable in others. Since the repository contains a large selection of die with different defects, the dies can be matched together in an optimal way when assembling the 3D-ICs so as to maximize the overall yield of 3D-ICs. The advantages of this approach include eliminating the need for switches and explicit reconfiguration circuitry and its associate overhead. Another advantage is the power distribution network in the surrounding layers need only be designed to support the computing elements being used and not the redundant elements. This is a big advantage as it avoids leakage current in the redundant elements and allows for a less expensive power/ground distribution network to be used.

While the proposed defect tolerance techniques can be used for memories, they are more suited for logic elements. Because memories have a very regular structure, very efficient defect tolerance techniques based on using spare rows and columns are already available [Schuster 78], [Zorian 03]. Recent work in [Chou 09, 10, 11], [Jiang 10] and [Rab 12] has proposed techniques for using unused spares in one die to help in repairing another die in an adjacent layer in 3D-ICs. This approach allows memory die to be matched up in a way that maximizes yield when assembling 3D-ICs.

Earlier work in defect tolerance for logic circuits has investigated using arrays of regular processing arrays

978-1-4673-3044-2/12/$31.00

including linear arrays, rectangular arrays, and binary tree architectures in which spares are included [Singh 88]. The schemes described in this paper can be applied for these types of architectures where the processing elements are distributed across multiple die. This can be done in a way that avoids the need for explicit reconfiguration circuitry.

The paper is organized as follows: Section 2 describes defect tolerance based on layer ordering and discusses the improvements in area and power that can be gained. Section 3 describes the idea of using rotational symmetry to improve defect tolerance. Section 4 discusses how multiple rotatable die can be used in a 3D-IC. Section 5 shows experimental results, and Sec. 6 is a conclusion.

2. Defect Tolerance Based on Layer Ordering

Consider a 3D-IC where the two middle layers contain identical die, as shown in Fig 1. The two middle die can be designed with one or more sub-modules duplicated. The other layers can be designed so that it always interfaces to and uses copy *A* of all duplicated sub-modules in the 3rd layer die, and always uses copy *B* of all duplicated sub-modules in the 2nd layer die. If a die has a defect in copy *A* of some duplicated sub-module, that die could be stacked so that it is in the 2nd layer where copy *B* is used, as shown in Fig 2. If a die has a defect in copy *B* of some duplicated sub-module, that die could be stacked so that it is in the 3rd layer where copy *A* is used.

After manufacture, the dies are tested and the defects located. The repository of die would include some die that are defect-free, some that have a defect in copy *A*, and some that have a defect in copy *B*. When assembling the 3D-ICs, the dies with defects would be placed in the appropriate layer where the defect is bypassed. The defect-free dies could be place in either layer. In this manner, the overall yield of operational 3D-ICs that can be constructed from the die can be maximized. Yield would only be lost for die which had defects in both copy *A* and copy *B*.

Experimental results in Sec. 5 highlight the reduction in area and power by taking advantage of the layer ordering for redundant logic.

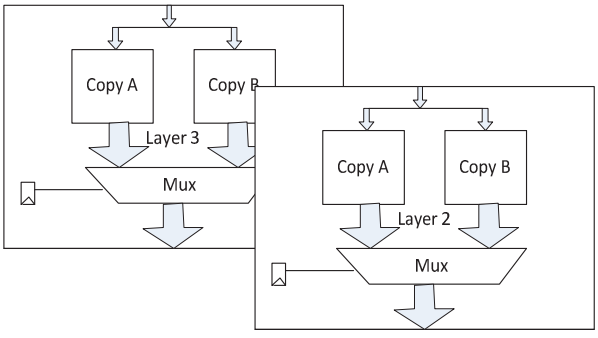

Figure 1. Conventional Defect Tolerance using a Powered Spare with a MUX to reconfigure

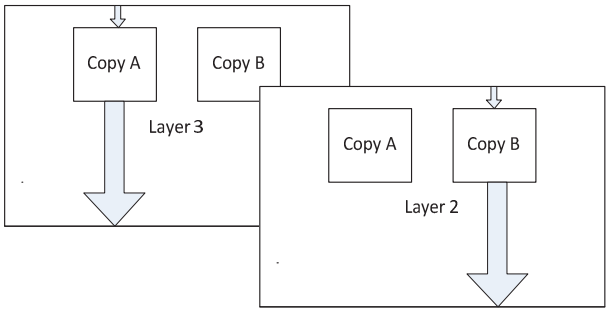

Figure 2. Proposed Defect Tolerance where in Layer 2, Copy B is used, and in Layer 3, Copy A is used. Spare is not powered and No Reconfiguration Circuitry is required

3. Defect Tolerance Based on Rotational Symmetry

The idea of using rotational symmetry was proposed in [Signh 11] at the wafer level for W2W stacking in 3D-ICs. The idea is to place die symmetrically on the wafer so that the wafer can be rotated 90 degrees at a time yielding 4 different rotational symmetries. This increases the number of ways that wafers can be matched together in W2W stacking. In this paper, we propose to use rotational symmetry for D2D stacking as illustrated in Fig. 3.

Rotational symmetry can be used as an effective defect tolerance strategy in designs which have spares in a die. Consider a 3D-IC where one layer contains a die with four identical cores. Three functional cores are needed, and one is a spare. If the die can be layed out symmetrically as shown in Fig. 3, then it is possible to rotate the die to always place the defective core in say the southwest quadrant when bonding the die to the 3D-IC stack. The other layers can be designed knowing the core in the southwest quadrant will not be used. This eliminates the need for any explicit reconfiguration mechanism and avoids the associated power and performance overhead. Moreover, no TSVs, power distribution network, or DFT support would need to be designed on the other layers (surrounding the symmetric die) to support the core in the unused quadrant. This avoids power dissipation, including static power dissipation, in the unused core and reduces both functional as well as power/ground TSVs running to the unused core. This helps improve manufacturability and yield because there are fewer things that can fail, while still providing the desired defect tolerance coming from having a spare core.

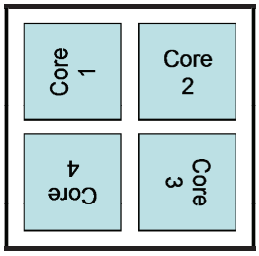

Figure 3. Symmetric Rotatable Die

A couple of issues for implementing this approach include the following. First, the die needs to be square in order to have this symmetry. Second, buses and signals that are passing through this layer to interface with other layers would need to be designed so that they have a symmetric layout so that rotating the die will not affect them (i.e., there will still be a TSV/wire connecting them regardless of the rotation).

Note that the example in Fig. 3 shows 4 cores. It is also possible to implement a symmetric rotatable die using 2 cores with one on each half of the die. In this case, the die does not have to be square as the rotation would be 180 degrees.

4. Multiple Rotatable Die

Current and future multi-core CPUs achieve high core counts by increasing the size of the processor die. As die size and transistor density grow, the susceptibility of these processors to hard faults grows as well [Powell 09]. Furthermore, ongoing research suggests that next generation designs should be optimized for performance, area and thermal issues jointly [Li, 06]. The longevity and the reliability of the cores are a function of the heat dissipation within the core. Hot spot formations and other undesirable conditions may occur due to variations in heat dissipation within the core. Hot spots can be formed, for example, when a high power core gets integrated with low power memories, thereby creating localized high power regions. It is expected that the heat flux emanating from hot spots in the next generation microprocessors may exceed 100W/cm^2, six times more than the average heat flux on the die [Yang 07]. Rotational symmetry can be applied in multi-layered multi-core 3D ICs to help address these issues. Rotational symmetry can be effectively used not only to increase the yield of the 3D stacked cores, but also to improve the reliability and the longevity of the 3D stacked ICs, as it may be used to help alleviate the thermal issues.

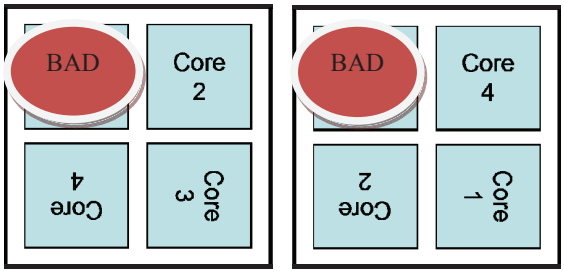

Figure 4. Stacking Two 4 Core Die where Failed Die Stack on Top of Each Other

Consider an example of a 6 core 3D IC where 2 rotatable layers are dedicated for cores, such that each layer has 4 cores; 3 functional and one redundant core. Figure 4 shows two dies with exactly one bad core in each die. Since one of the cores in each die is a spare, any die with 3 or more functional cores can be stacked together to produce a 6 core design. At design time, a stacking arrangement can be chosen based on, for example, whether the IC should be optimized for power or heat dissipation. Below we present two ways in which the 3D stacked IC may be designed, one where the design is optimized for low power and the other where the design is optimized to avoid formation of hot spots.

Rotational symmetry of the dies provides a degree of freedom in designing the surrounding layers in terms of where the defective unused cores on each of the layers will be located. For each of rotatable layers, at design time, the quadrant where the unused defective core is located can be selected. If the 3D IC is designed for low power, then the arrangement in Fig. 4 can be used where both unused defective core are placed in the same quadrant (e.g., northwest as shown in Fig. 5). The advantage of putting them both in the same quadrant is that it simplifies the voltage regulator and simplifies the routing of clock and power lines.

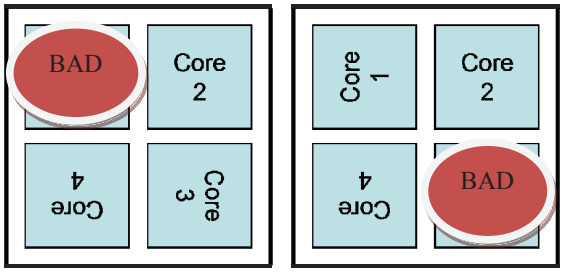

Figure 5. Stacking Two 4 Core Die where Failed Die Do Not Stack on Top of Each Other

Another alternative stacking arrangement is shown in Fig 5. In this case, the other layers are designed so that the unused defective cores are placed in different quadrants. The advantage of this stacking arrangement is that heat dissipation gets spread out better and makes hot spot formations less likely. The two stacking arrangement options are summarized in Table 1.

Based on the stacking arrangement selected, the other layers are designed so that they can interface with the two rotatable cores. The unused cores are not powered, so there is no need for power/ground distribution, clock, or DFT support for those quadrants from the other layers. Based on how the design is done, after manufacture, testing is done to locate any defective cores, and then the 3D-ICs are assembled by rotating the rotatable die to put the defective cores in the unused quadrants. Note that earlier defect tolerance techniques based on arrays of processing elements with spares [Singh 88] can be implemented using multiple rotatable die. Each die effectively contains 3 processing elements and one spare, and a set of such die can construct an array of the desired size.

Table 1. Tradeoffs for Stacking Arrangement

Stacking Type	Advantage	Dis Advantage
Same Quadrant (Fig. 4)	Low Power	Hot Spot Formation
Different Quadrant (Fig. 5)	Heat Balancing	More Complex Routing

5. Experimental Results

Experiments were conducted to show the area and power savings when employing defect tolerance based on layer order. A cluster (module) of logic from an industrial design was synthesized in two different ways. In the first case the conventional defect tolerance approach was implemented where reconfiguration switches were provided to allow explicit reconfiguration (as shown in Fig. 1). The design was synthesized using a standard TSMC library. Next, the logic for the proposed method where no explicit reconfiguration logic was required (as shown in Fig. 2) was synthesized.

This was done for two different blocks, and Tables 2 and 3 below compare area and power between the two approaches. As can be seen from the results, leakage power is reduced by more than a factor of 2 because of the fact that the unused copy is not connected to the power distribution network. Area is reduced because the reconfiguration circuitry is no longer needed.

Table 2. Comparison of Block 1 Implemented with Conventional Defect Tolerance versus Proposed Defect Tolerance based on Layer Order

	Conventional	**Proposed**
Input Ports	1345	1345
Output Ports	1183	1183
Flip Flops	17825	17824
Combo Cells	118672	116922
Area	92926	92068
Leakage	25.49 (uW)	12.63 (uW)

Table 3. Comparison of Block 2 Implemented with Conventional Defect Tolerance versus Proposed Defect Tolerance based on Layer Order

	Conventional	**Proposed**
Input Ports	612	612
Output Ports	558	558
Flip Flops	14531	14530
Combo Cells	273978	272940
Area	252339	251590
Leakage	44.5623(uW)	22.215 (uW)

6. Conclusions

As manufacturing and design complexity continue to increase, the cost benefits of using defect tolerance will become increasingly compelling. It was shown how the degrees of freedom that exist when assembling a 3D-IC can be used to implement defect tolerance without the need for explicit reconfiguration circuitry. These techniques reduce the power, area, and performance overhead for defect tolerance.

Acknowledgements

This research was supported in part by National Science Foundation under Grant No. CCF-1217750.

References

[Chou 09] Y.-F. Chou, D.-M. Kwai, and C.-W. Wu, "Memory Repair by Die Stacking with through Silicon Vias", *Int. Workshop on Memory Technology, Design, and Testing*, pp. 53-58, 2009.

[Chou 10] C.-W. Chou, Y.-J. Huang, and J.-F. Li, "Yield-Enhancement Techniques for 3D Random Access Memories", *Int. Symp. on VLSI Design Automation and Test (VLSI-DAT)*, pp. 104-107, 2010.

[Chou 11] Y.-F. Chou, D.-M. Kwai, and C.-W. Wu, "Yield Enhancement by Bad-Die Recycling and Stacking with Through-Silicon Vias", *IEEE Trans. on VLSI Systems*, Vol. 19, Issue 8, pp. 1346-1356, Aug. 2011.

[Jiang 10] L. Jiang, R. Ye, and Q. Xu, "Yield Enhancement for 3D-Stacked Memory by Redundancy Sharing across Dies", *Int. Conf. on Computer-Aided Design (ICCAD)*, pp. 230-234, 2010.

[Koren 98] Koren, I., and Z. Koren, "Defect Tolerance in VLSI Circuits: Techniques and Yield Analysis", *Proc. of IEEE,* Vol. 86, No. 9, pp. 1819-1836, Sept. 1998.

[Li 06] Yingmin Li, Benjamin Lee, David Brooks, Zhigang Hu and Kevin Skadron, "Impact of Thermal Constraints on Multi-Core Architectures", *Proc. of Thermal and Thermomechanical Phenomena in Electronics Systems,* pp. 132-139, 2006.

[Powell 09] Michael D. Powell, Arijit Biswas, Shanatnu Gupta and Shubhendu S. Mukherjee, "Architectural Core Salvaging in a Multi-Core Processor for Hard-Error Tolerance", *Proc. of Int. Symposium on Computer Archecture (ISCA)*, pp. 93-104, 2009.

[Rab 12] M. Tauseef Rab, Asad A. Bawa and Nur A. Touba, "Using Asymmetric Layer Repair Capability to Reduce the Cost of Yield Enhancement in 3D Stacked Memories", *Proc. Of International Conference on Very Large Scale Integration (VLSI-SoC),* Oct 2012

[Schuster 78] S.E. Shuster, "Multiple word/bit line redundancy for semiconductor memories", *IEEE Journal of Solid-State Circuits*, Vol. SC-13, pp. 698-703, 1978.

[Singh 88] Singh, A., "Insterstitial Redundancy: An Area Efficient Fault Tolerance Scheme for Large Area VLSI Processor Arrays", *IEEE Trans. on Computers*, Vol. 37, No. 11, pp. 1398-1410, Nov. 1988.

[Singh 11] Singh, E. "Exploiting Rotational Symmetries for Improving Stacked Yield in W2W 3D-SICs", *Proc. of VLSI Test Symposium*, pp. 32-37, 2011.

[Yang 07] Bo Yang, Peng Wang and Avram Bar-Cohen, "Mini-Contact Enhanced Thermoelectric Cooling of Hot Spots in High Power Devices", *IEEE Transactions on Components and Packaging Technologies*, Vol. 30, No. 3, pp. 432-438, Sep 2007.

[Zorian 03] Y. Zorian, and S. Skoukourian, "Embedded-Memory Test and Repair: Infrastructure IP for SOC Yield", *IEEE Design & Test of Computers*, Vol. 20, Issue 3, pp. 58-66, May 2003.