In-growth Test for Monolithic 3D Integrated SRAM

Pu Pang[1], Yixun Zhang[1], Tianjian Li[1], Sung Kyu Lim[2], Quan Chen[1], Xiaoyao Liang[1] and Li Jiang[1*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

***Abstract*—Monolithic three-dimensional integration (M3I) directly fabricates tiers of integrated circuits upon each other and provides millions of vertical interconnections with inter-layer vias (ILVs). It thus brings higher integration density and communication capability compared with three-dimensional stacked integration (3D-SI). However, the Known-Good-Die problem haunting 3D-SI–a faulty tier causes the failure of the entire stack–also occurs in M3I. Lack of efficient test methodologies such as the pre-bond testing in 3D-SI, M3I may have a more significant yield drop and thus its cost may be unacceptable for main-stream adoption. This paper introduces a novel In-growth test method for M3I SRAM. We propose a novel Design-for-Test (DfT) methodology to enable the proposed In-growth test on cell-level partitioned incomplete SRAM cells. We also build a statistical model of cost and discover a prospective judgement to determine whether or not to stop the fabrication, in order to prevent from raising the cost of fabricating more tiers upon the irreparable tiers. We find that a "sweet point" exists in the judgement, which can minimize the overall cost. Experimental results show the effectiveness of our proposed test methodology.**

I. Introduction

Due to the growing area occupation of the embedded SRAMs in SoCs, it becomes increasingly important but challenging to increase the density and capacity of SRAMs. Three-dimensional (3D) integration is a promising solution. Conventional 3D stacked integration (3D-SI) relies on through-silicon vias (TSVs). TSVs have a large pitch size, and thus limits the density of vertical connections. Monolithic 3D integration (M3I) overcomes this problem by using nano-scale inter-layer vias (ILVs) [1], which are an order magnitude smaller and shorter than TSVs; they are also easier to be aligned. Thus, M3I SRAMs can provide higher vertical bandwidth, and higher capacity when compared to 3D-SI SRAM.

Despite many advantages, M3I has an Achilles' heel. The 3D integration suffers from the Known-Good-Die problem that any defective tier leads to the failure of the whole stack [2]. Fortunately, 3D-SI can rely on the pre-bond testing to screen out the defective dies and only good dies are stacked together. However, each tier directly grows upon other tiers in M3I. Even if faults in a tier can be detected by manufacturing test, this fabricated tier cannot be removed from the remaining good tiers. With respect to cost, it is preferable to discard the faulty M3I with few tiers in fabrication process, rather than discarding the whole M3I chip after some tiers are found faulty and irreparable in the final test. In a 4-tier M3I SRAM, for instance, if the 1st tier is fabricated with too many faults to be reparable, it would rather stop the fabrication process than keep fabricating the three remaining tiers. Because the additional fabrication and test processes are in vain.

However, it is difficult to adopt the above "early stop" in conventional cell-level partitioned M3I SRAM. The main reason is that the SRAM cells are divided, where nMOS and pMOS transistors are split into adjacent tiers. There are two challenges: 1) A SRAM cell is incomplete before two tiers are fabricated. It is hard to test these incomplete SRAM cells when the bottom tier is fabricated but the top tier is not. 2) A redundant row/column is used to replace a row/column of complete SRAM cells. Given the fault map of the bottom tier, it is challenging to judge whether to "early stop" or not, because new faults may emerge and redundancies may be insufficient after the top tier is fabricated.

To resolve the two challenges, we propose *In-growth test* that consists of a novel DfT method to test incomplete cells and a statistical *judgement factor*. The proposed DfT method can temporally compose complete SRAM cells on a single tier and thus we are able to test cells on the bottom tier. The DfT infrastructure requires negligible area overhead, and one additional metal-etching process, which is a compatible and robust fabrication process. Based on the test results, the judgement factor serves as the statistical "knob" to guide the In-growth test and fabrication process. Based on the judgement factor, we can decide the "early stop" of the fabrication process to minimize the overall cost. We provide mathematical and experimental proofs that a "sweet point" always exists in the judgement of "early stop", which can minimize the overall cost. In-growth test can be applied in the fabrication process of M3I SRAM.

II. Preliminaries

The cell-level partitioned design in M3I SRAM is widely adopted. Lim et al. [3] present various M3I SRAM designs, showing both 2P4N (two pMOS transistors and four nMOS transistors) and 3P3N designs can outperform the 2D SRAM design in terms of area and performance. Yu et al. [4] present the design of M3I SRAM with InGaAs/Ge MOSFETs.

Due to the Known-Good-Die problem, 3D integration suffers from a high failure rate, causing a higher cost to satisfy the target yield. This is why an efficient test methodology is necessary for 3D integration. However, the conventional cell-level partitioned design brings challenging problems in M3I testing, which are described in details as follows.

Testability Problem – The principle of the SRAM test, e.g., March test [5], is to write and then read the test vector upon the SRAM cells in sequence. Only a self-contained SRAM cell with complete peripheral circuits can provide the read/write functionality. However, before the top tier is fabricated, the

This research was partially supported by National Natural Science Foundation of China (Grant No. 61602300) and Shanghai Science and Technology Committee (Grant No. 15YF1406000). Corresponding author is Li Jiang.

978-3-9819263-0-9/DATE18/©2018 EDAA

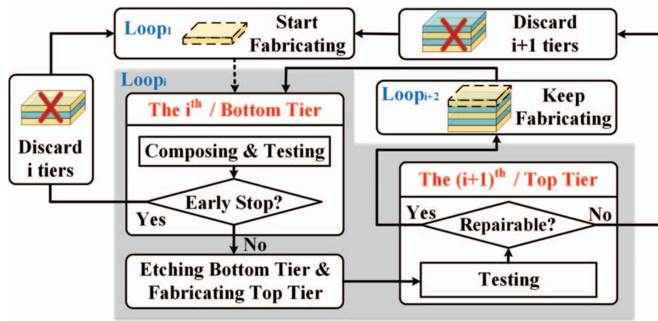

Fig. 1: The flow of the fabrication process integrated with In-growth test.

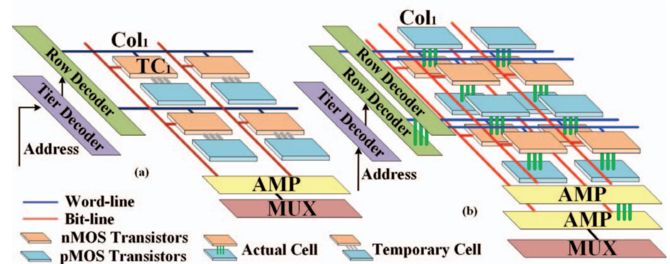

Fig. 2: The DfT method of incomplete cells. (a) Temporary cells and the layout of the bottom tier. (b) Actual cells and the layout of two tiers.

M3I SRAM cells are incomplete and not testable. Zhao et al. [6] solve a similar problem in 3D-SI by adding redundant circuits and buffers to design a pre-bond testable clock tree. Unfortunately, we cannot solve this problem by adding redundant transistors. This is because the same number of redundant transistors are required as that of the original transistors, which leads to significant area increase.

Reparability Problem – 3D SRAMs ask for a strong capability of fault tolerance due to the high failure rate. Conventionally, redundancies are deployed into the SRAM to replace the rows/columns containing fault cells. The inter-layer redundancy-sharing techniques [7] are proven to be particularly efficient in 3D memory. However, these techniques are impossible for M3I SRAMs. Because ILVs bound the corresponding transistors on adjacent tiers together and it is impossible to use a redundant transistor to replace a faulty transistor on the same tier. We cannot judge whether the M3I SRAM is repairable after testing incomplete cells on the bottom tier, so it's challenging to judge whether to "early stop" or not before the top tier is fabricated.

Problem Solutions – In this paper, we propose a novel DfT method to solve the testability problem, by means of redesigning the layout and peripheral circuits and composing temporary complete cells. For the reparability problem, this paper uses the redundancy requirements as judgement factor, and it can be derived from the test results of the bottom tier. To be specific, we set a threshold value of the requirements to judge whether to "early stop" or not. By building a statistical model of cost, we provide a best threshold setting (denoted as "sweet point"), which reaches the minimum overall cost.

III. Propose Techniques

A. Fabrication Process Integrated with In-growth Test

Fig. 1 shows the flow of the M3I SRAM fabrication process integrated with our proposed In-growth test. We denote the fabrication process of the ith tier (the bottom tier) and the $(i+1)^{th}$ tier (the top tier) as $Loop_i$ with $i = 1, 3, ..., 2N-1$. In $Loop_i$, the ith tier is fabricated first. Then, two adjacent incomplete cells on the ith tier are composed into a temporary complete cell using our DfT method (addressed in III-B), and thus the conventional SRAM test method can be applied. After testing the cells on the ith tier, it's time to judge whether or not to stop the fabrication (i.e. "early stop", which is addressed in III-C and III-D). If "early stop", all i tiers are discarded; otherwise, the additional metal layer used to compose temporary cells are etched away and the $(i+1)^{th}$ tier will be fabricated. After fabricating the $(i+1)^{th}$ tier, it is necessary to test all actual cells, because new faulty cells may emerge. With the fault map of actual cells, we can repair the faults (if repairable) and keep fabricating ($Loop_{(i+2)}$); otherwise, all $(i+1)$ tiers are discarded.

B. DfT method of Incomplete Cells

Most M3I SRAM designs split nMOS and pMOS transistors into two adjacent tiers, each of which contains same type transistors. Consequently, the cells on one tier are incomplete and not testable unless the other tier is fabricated. To solve this testability problem, we need to fabricate nMOS and pMOS transistors on the same tier. This is possible because CMOS technology allows that nMOS and pMOS transistors are fabricated on the same tier and the 2D CMOS-based SRAM does so. In M3I, [8] presents a process flow, where nMOS and pMOS transistors are fabricated on one tier.

Using the 6T (2P4N) SRAM cell [3], Fig. 2(a) shows the DfT of incomplete cells on the bottom tier, where four nMOS and two pMOS transistors are arranged alternatively. In this design, adjacent six transistors (2P4N) can be composed into a temporary SRAM cell. Additional metal wires are required to connect the transistors. The number of wires is equal to the number of ILVs linking the transistors between two tiers in the original M3I SRAM cell. It should be noted that the temporary cells are organized along the direction of redundancies (suppose redundant columns are used in this paper); the reason will be explained later.

Before fabricating the top tier, the additional metal wires are etched away. In M3I, this process can be concealed in the following fabrication process for the second tier; the overhead can thus be omitted. The layout of two tiers is shown in Fig. 2(b), wherein the nMOS and pMOS transistors on the top tier are arranged oppositely. Thus, ILVs between these two tiers can connect the two partial cells and form a complete SRAM cell. In addition, the decoder and the group of sense-amplifiers are partitioned to two tiers; this design style is also commonly adopted in 3D memory design.

C. Judgement Factor of "Early Stop"

Since whether a SRAM is repairable depends on whether all faulty cells can be replaced by redundancies, it's reasonable to take the redundancy requirements as the judgement factor

of "early stop". In this paper, we suppose the column-based redundancy architecture is used, which is suitable for SRAMs. As mentioned above, the temporary cells on the bottom tier are organized along the column. Hence, when a temporary cell is detected as a fault (e.g., TC_1 in Fig. 2(a)), it's certain that all this column (Col_1) will be replaced and one redundant column is needed, despite that the top tier is not fabricated yet. Based on above all, we can get to know the requirements Q_i according to the test results of the bottom tier (i^{th} tier).

Obviously, if Q_i is greater than or equal to all available redundant columns (denoted as R), it's confirmed that we should "early stop". The challenge emerges when $Q_i < R$. A static judgement (e.g., when $Q_i > 0.5R$ we should "early stop") may not lead to the best cost-effectiveness because we should be more conservative to stop the fabrication when many tiers have already been fabricated.

Technically, the problem is formulated as follows: Given 1) *The SRAM technology*, including the fault model and failure rate, 2) *Design parameters*, including the number of SRAM columns L, the number of SRAM tiers $2N$ and the number of redundancies R, 3) *Fabrication parameters*, e.g., the average fabrication and test cost of the bottom tier C_b and the top tier C_t, we try to find a threshold T_i for the ith tier–"early stop" the fabrication when $Q_i \leq T_i$–to minimize the overall cost. We denote such T_i as "sweet point". To find such a sweet point, we first build a statistical cost model, which is described in the next section.

D. Statistical Cost Model

The key idea is to derive the functional relation between the cost and T_i, which naturally lead to the "sweet point".

Preparations – We first consider $Loop_i$. When the ith tier (the bottom tier) has just been fabricated, there are three possible events: 1) $B_{(i,1)}$: Keep fabricating the $(i+1)^{th}$ tier (the top tier) and the actual cells are repairable. 2) $B_{(i,2)}$: "Early stop". 3) $B_{(i,3)}$: Keep fabricating but the actual cells are irreparable.

Let $E(X)$ be the expected cost of the event X. $E(B_{(i,2)})$ can be derived as shown in Fig. 3. When $B_{(i,2)}$ happens, $(i-1)/2+1$ bottom tiers and $(i-1)/2$ top tiers have already been fabricated. But all these tiers have to be discarded. We can derive $E(B_{(i,3)})$ under the same principle. Both $B_{(i,2)}$ and $B_{(i,3)}$ cause waste all the fabrication and test cost which have been spent. For convenience, we combine $B_{(i,2)}$ and $B_{(i,3)}$ in a composite event $B_{(i,4)}$. Let $P(X)$ be the probability of the event X, $E(B_{(i,4)})$ can be derived in a similar fashion as shown in the figure.

Model establishment – Next, we consider the complete fabrication process with In-growth test (denoted as an event F), using a $2N$-tier M3I SRAM as an example. Note that F contains not only the successful process but also all the possible aborted processes.

As discussed in the preparations, in $Loop_i$, the $(i+2)^{th}$ tier will be fabricated if B$_{(i,1)}$ happens. Conversely, if B$_{(i,4)}$ happens, we have to discard all tiers and fabricate a new one from the start, which leads to an additional cost $E(B_{(i,4)})$. As shown in Fig. 3, a fabrication process ultimately has two outcomes: 1) S (red line): Successful fabrication of the $2N$-tier

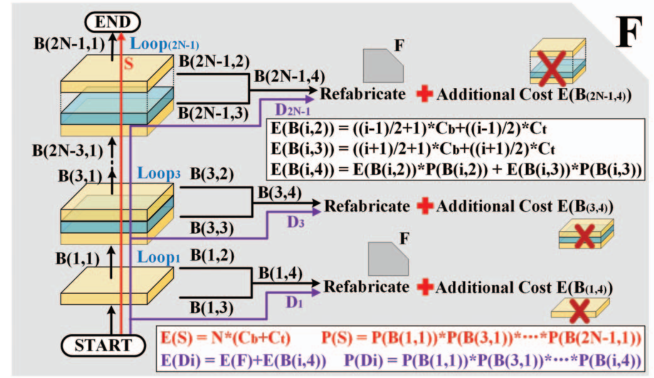

Fig. 3: The diagram of the 2N-tier M3I SRAM fabrication process integrated with In-growth test.

SRAM (in one time), resulting from a chain of events $B_{(1,1)} \rightarrow B_{(3,1)} \rightarrow ... \rightarrow B_{(2N-1,1)}$; 2) D_i (purple line): Encountering an irreparable tier and restarting the fabrication, resulting from a chain of events $B_{(1,1)} \rightarrow B_{(3,1)} \rightarrow ... \rightarrow B_{(i-2,1)} \rightarrow B_{(i,4)}$. Thus, the expected cost of F, which is the overall cost, can be derived as follows:

$$E(F) = E(S) * P(S) + \sum (E(D_i) * P(D_i)) \quad (1)$$

wherein $E(S)$ is the expected cost of a 2N-tier SRAM successfully fabricated in one time, S happens when all $B_{(i,1)}$ consecutively happen, $P(S)$ refers the probability of even S, $E(D_i)$ consists of the new fabrication cost F and the additional cost, and $P(D_i)$ is the probability that event D_i happens when $B_{(i,4)}$ happens in the end of the chain of events.

We next derive the functional relation between the overall cost and thresholds, i.e. $E(F) = f(T_1, T_3, ...T_{2N-1})$, by solving the equation (1). Because the probability of events $P(B_{(i,1/2/3/4)})$ is decided by the threshold T_i. Suppose the probability distribution $p(L,k)$ denotes that k faulty SRAM columns exist in a M3I SRAM with L columns and R redundant columns. $P(B_{(i,1/2/3/4)})$ can be derived as follows:

- $P(B_{(i,1)})$: If there are k redundant columns required in the bottom tier, $B_{(i,1)}$ happens when $k \leq T_i$ and the faulty columns within the $L-k$ columns of the top tier (t) is less than or equal to $R-k$, i.e. $t \leq R-k$. Hence, $P(B_{(i,1)})$ is the cumulative probability for all possible k and t.

$$P(B_{(i,1)}) = \sum_{k=0}^{T_i} (p(L,k) * \sum_{t=0}^{R-k} p(L-k,t))$$

- $P(B_{(i,3)})$: It is similar with $P(B_{(i,1)})$ but with $t > R-k$.

$$P(B_{(i,3)}) = \sum_{k=0}^{T_i} (p(L,k) * (1 - \sum_{t=0}^{R-k} p(L-k,t)))$$

- $P(B_{(i,2)})$: The probability when the above two events do not happen.

$$P(B_{(i,2)}) = 1 - P(B_{(i,1)}) - P(B_{(i,3)})$$

More importantly, since $T_i \in N$ and $0 \leq T_i \leq R$, there must exist a set of specified values T_i to minimize the overall cost $E(F)$, which is the so-called "sweet point".

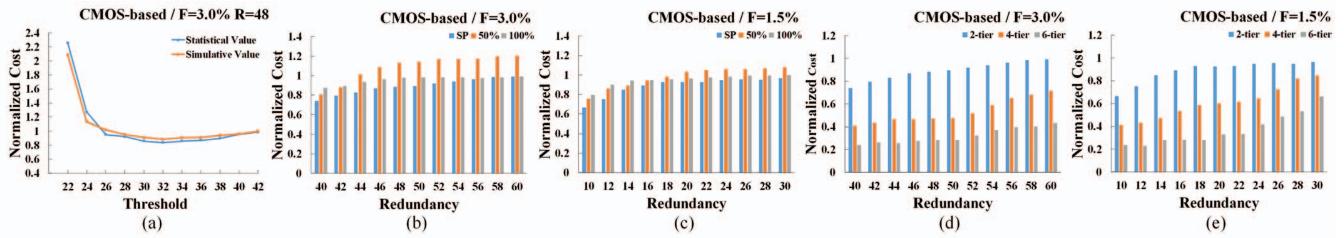

Fig. 4: Comparison between the results of (a) statistics and simulation in different settings of threshold in 2-tier case. (b)(c) three different threshold setting strategies in 2-tier case. (d)(e) M3I SRAM with different tiers.

IV. Experimental Results

A. Experimental Setup

A simulator based on 6T M3I CMOS-based SRAM is built to evaluate the effectiveness of our proposed In-growth test. The SRAM has 512x256 actual cells in every two tiers, and 256x256 temporary cells in the bottom tier, while faults are randomly injected in Compound Poisson Distribution [9].

We conduct Monte-Carlo simulation to fabricate 1000 M3I SRAMs with two different test methodologies and track their cost (including the wafer cost, the assembly cost and the test cost). The conventional test methodology tests the whole M3I SRAM after the fabrication process finishes, which serves as the baseline. However, our proposed In-growth test methodology tests the temporary cells on the bottom tier to get Q_i and then judge whether to keep fabricating or discard all tiers by comparing Q_i with T_i. It should be noted that, the cost of those discarded tiers is also taken into account. The cost of our proposed In-growth test methodology is normalized to the baseline as reported in the results.

B. Results and Analysis

1) Accuracy of the Model: As shown in Fig.4(a), we evaluate the model with 3.0% failure rate (denoted as F), 48 redundant columns (denoted as R) and varying thresholds. There are two curves: one represents the statistical value derived by solving the equation (1) using MATLAB, and the other represents the simulative value derived from the Monte-Carlo simulation. It can be observed that the statistical value perfectly fits the simulative value. Thus, our statistical model can be easily adopted to get the "sweet point" (32 in Fig.4(a)).

In Fig. 4(a), we can observe that In-growth test achieves a smaller cost compared with the conventional methodology, when the threshold is set between 26 to 40. Also, the cost becomes larger if the threshold is too small. Because the judgement of "early stop" becomes too aggressive, which means we insist on "early stop" even if keeping fabricating will bring a repairable SRAM. When the threshold is close to or even equal to R, the cost is close to the conventional test methodology, since very few "early stop" will be allowed.

2) Effectiveness of the "sweet point": Fig.4(b) and (c) compare the normalized cost derived by three different threshold setting strategies with varying failure rates and redundancies. We set the threshold to the "sweet point" derived by our statistical cost model, as well as 50% and 100% of the redundant column count, respectively. We denote the three strategies as SP, 50% and 100%. As shown in the figure, SP outperforms other two strategies in all cases. We can also observe that the fewer redundant columns are deployed, and more cost can be saved in SP case. This is because the yield of the SRAM with fewer redundant columns is lower, which makes "early stop" more profitable. An interesting observation happens in the 50% strategy. As the failure rate of two tiers are nearly the same, 50% seems an intuitive and natural strategy for "early stop". However, it can be worse than the baseline when many redundant columns are deployed. This is because when there are enough redundancies and the yield is high, 50% will become too aggressive.

In summary, it's hard or impossible to use a static threshold to achieve the minimum cost, for the cost is deeply influenced by the technologies and design parameters. With necessary of fabrication and design parameters, the proposed DfT and cost model can significantly reduce the cost of M3I SRAM.

3) Applying to Multi-tier Case: Fig.4(d) and (e) present the normalized cost of the M3I SRAM with 2, 4 and 6 tiers. As can be seen, the cost reduction of In-growth test improves significantly as the number of tiers increases. The primary reason is that the yield of the M3I SRAM decreases dramatically as the number of tiers increases, i.e., the Known-Good-Die problem. Generally, we have a stronger motivation to apply In-growth test for the M3I SRAMs with more tiers.

V. Conclusion

The M3I SRAM is a promising alterantive for the future VLSI prodcuts. However, it suffers from more significant yield loss compared with 3D-SI. This paper proposes a novel In-growth test method to reduce the manufacturing cost of M3I SRAMs. Experimental results show the effectiveness of the proposed test method.

References

[1] Ebrahimi M. S, et al. Monolithic 3D integration advances and challenges: From technology to system levels. *Soi-3d-Subthreshold Microelectronics Technology Unified Conference*, 2014.

[2] Lee H. H. and Chakrabarty K. Test Challenges for 3D integrated circuits. *Design and Test of Computers*, 2009.

[3] C. Liu and S. K. Lim. Ultra-high density 3D SRAM cell designs for monolithic 3D integration. *Interconnect Technology Conference*, 2012.

[4] Yu K. C., et al. Evaluation of monolithic 3-d logic circuits and 6t srams with ingaas-n/ge-p ultra-thin-body mosfets. *Journal of the Electron Devices Society*, 2016.

[5] Goor A. J. V. D., et al. March tests for word-oriented memories. *Design Automation and Test in Europe Conference*, 1998.

[6] Zhao X., et al. Pre-bond testable low-power clock tree design for 3D stacked ICs. *International Conference on Computer-Aided Design*, 2009.

[7] Li T., et al. Fault clustering technique for 3D memory BISR. *Design Automation and Test in Europe Conference*, 2017.

[8] Batude P., et al. Advances in 3D CMOS sequential integration. *International Electron Devices Meeting*, 2009.

[9] Koren I. and Koren Z. Defect tolerance in VLSI circuits: Techniques and yield analysis. *Proceedings of the IEEE*, 1998.