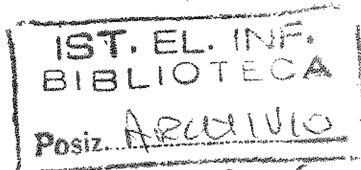
A2-04
2000

GMM

ITG



12th Workshop
Testmethods and Reliability of Circuits and Systems

March 19 – 21, 2000, Grassau, Germany

GI FA 3.5 / ITG FA 8.2 / GMM FB 8
Kooperationsgemeinschaft Rechnergestützter Schaltungs- und Systementwurf
IEEE – TTTC, Test Technology Technical Council

Wafer-Scale VLSI Testing

Antonio Caruso[2], Stefano Chessa[1], and Piero Maestrini[1,2]

[1] Istituto di Elaborazione dell'Informazione del CNR, Pisa, Italy, Via S. Maria 46, 56126, Pisa, Italy.

[2] Dipartimento di Informatica, University of Pisa, Corso Italia 40, 56125, Pisa, Italy.

1 Introduction

In the manufacturing of VLSI integrated circuits (ICs), it would be desirable that the ICs undergo an accurate test directly on the wafer, before they are cut and packaged. This would save the cost of bounding and packaging the faulty ICs, which may be a relatively large fraction. However, accurate testing on the wafer can hardly be supported by the consolidated technology of testing machines. For this reason, the test of ICs on the wafer is usually limited to providing power, ground and a few signals, and to checking some basic electrical properties. This results in a test with limited coverage, which is unable to identify all faulty ICs. The ICs which pass the preliminary test (which may include a relatively large fraction of faulty ICs) must undergo a more accurate test and are possibly discarded after they have been bounded and packaged, an this may contribute heavily to the overall cost of the manufacturing process. Another drawback is that the test on the wafer proceeds sequentially, since the test machine must step over individual IC, and this implies that the time needed to test the entire wafer is quite long.

A different approach to the functional manufacturing test of ICs on the wafer can be based upon concepts inherited from the *system-level diagnosis theory* [1], which was introduced to diagnose systems composed by units able to perform mutual test. A test involves two units and, in principle, proceeds as follows: the *testing unit* provides a test input sequence to the *tested unit*, which returns an output sequence. The testing unit compares the latter sequence with the expected output sequence and defines a *test outcome*, which is binary: 0 if the test passes and 1 otherwise. The set of all test outcomes (called *syndrome*) is input to a controlling computer (called *diagnoser*), which executes a *diagnosis algorithm* aimed at identifying the faulty units.

The diagnosis algorithm partitions the set of all units into set K of units declared good, set F of units declared faulty, and set S of units declared *suspect*, i.e. those units which the algorithm is unable to classify as either good or faulty. The diagnosis is said to be *correct* if units assigned to set K are actually fault-free and those assigned to set F are actually faulty. The diagnosis is said to be *complete* if every unit is classified as either good or faulty, i.e. set S is empty.

A well-known diagnosis model, called PMC model [1], assumes that the test outcome of a test executed by a fault free unit is always reliable, and it is completely unreliable otherwise. This invalidation rule is reported in Table 1.

The problem of wafer-scale testing of ICs closely resembles system-level diagnosis under many aspects. In fact, individual ICs can be arranged on the wafer as a grid, where every IC (with the exception of those in the border) is connected to a constant number of neighbors. Depending on the number of interconnection to

Table 1: Invalidation rule of the PMC model.

Testing unit	tested unit	Test outcome
fault-free	fault-free	0
fault-free	Faulty	1
Faulty	fault-free	0 or 1
Faulty	Faulty	0 or 1

neighbors (3, 4, 6 or 8), the grid is called triangular, square, hexagonal or octagonal. Interconnections are used to perform mutual tests, as needed for the purpose of self-diagnosis.

With this methodology, hereafter referred to as *autodiagnostic wafer-scale testing*, all ICs on the wafer undergo an intensive test before they are cut, bounded and packaged, and those that failed to pass the test are immediately discarded. The test is executed at the operating speed of ICs, or to a comparable speed, thus contributing to the accuracy. Essentially, the tests of all the ICs in the wafer proceed in parallel and the testing machine does not need to step over the individual ICs, thus reducing the time and the cost to perform the test.

2 Implementation Aspects

The first challenge to the autodiagnostic wafer-scale testing is the implementation of the tests. In fact the system-level diagnosis theory requires that the units (ICs) be able to perform tests on the adjacent ICs, and this condition may not be met unless ICs are processors. Diagnostic model based on comparison [2], provide solution to this problem. In principle, all ICs receive a common input sequence and every IC produces an output sequence. For every pair of adjacent ICs, the output sequences are compared by means of comparators embedded in the wafer and accumulated in a single bit register, which will eventually store comparison outcome 0 if both ICs in the pair produced the same output sequence, and 1 otherwise. This requires revised diagnostic model and invalidation rules, which account for faulty comparators. Observe that comparisons may be executed in parallel.

A naive implementation of the autodiagnostic wafer-scale testing, as above described, would require a large number of interconnections across the wafer. Beside those needed to distribute the power, ground and clock, interconnections would be required to distribute the test vectors, to read out the syndrome and to convey output sequences to comparators. All but the latter interconnections (which connect adjacent dies) would be global. The comparators and the controlling logic, as well as the interconnections, would be external to the dies, and would have to be disposed when the dies are separated. The cost of the silicon area needed to the purpose of autodiagnostic wafer-scale testing would be negligible; however, for several reasons, the above described implementation would be a challenge to the manufacturing technology.

A practical implementation should be based on test sequences generated inside individual dies or stored in a test memory. To reduce memory requirements, test sequences could be split in subsequences to be loaded repeatedly. Output vectors could be stored in the test memory and comparators could be fed serially. This strategy, while increasing the silicon area occupied by dies, would reduce to a minimum the logic and the interconnections external to dies.

With this implementation, the autodiagnostic wafer-scale testing somehow resembles a BIST technique. In

fact, the test sequences are either generated or stored internally to dies, and the comparators produce signatures of output sequences. The novelty consists in the fact that signatures are produced by combining output sequences accumulated inside adjacent ICs. While requiring flow of information between adjacent IC pairs, this technique has the advantage that individual signatures are extremely compressed (arrangement in a square grid yields a signature of 4 bits). Results reported in the next section show that, in spite of the extreme information compression, the diagnosis algorithm is able to correctly identify almost all the good ICs in most expected situations.

3 Diagnosis algorithm

To be suitable for autodiagnostic wafer-scale testing, the diagnosis algorithm should be able to provide correct and almost complete diagnosis under the fault situations occurring in wafer environments. Several diagnosis algorithms [3, 4, 5] have been proposed for this application. Although they are based on mutual tests of adjacent ICs, those algorithms could easily be reformulated to accommodate comparison test.

The algorithms proposed in [3, 4, 5] provide diagnoses which are correct and complete with high probability in case of random faults. However they may fail dramatically under systematic fault situations; that is when all ICs produce the same, wrong output in response to a common input sequence. Systematic faults cannot be neglected when testing ICs, given the probability of weak points in design which may increase the vulnerability of ICs, or as the result of the repeated exposure of flawed or misaligned masks of the dies on the wafer.

An algorithm, which is able to provide diagnosis that is provably correct and probabilistically complete, even in the occurrence of systematic faults, has been introduced in [6]. This algorithm, in a formulation based on comparison test, operates in three steps: Local Diagnosis, Fault-Free Core Identification and Augmentation. The Local Diagnosis performs a preliminary classification of ICs as either D, or Z. ICs classified as D are defined in disjoint pairs with the property that, for every pair, at least one unit is faulty. The state of ICs classified as Z remains unidentified in this step, but, in most cases, it will be determined in the following of the algorithm. Hereafter, Z and D denote the sets of units classified Z and D, respectively, and $\#Z$, $\#D$ denote their cardinalities.

In the second step, the ICs in set Z are distributed into Z-aggregates, that is, connected subgrids of ICs classified as Z. The Z-aggregates have the property that all the ICs in the same aggregate are in the same (faulty or non-faulty) state. The Fault-Free Core (denoted K) is defined as the union set of all the Z-aggregates of maximum cardinality (denoted α). A syndrome-dependent bound for diagnosis correctness, denoted T_σ, is also defined in this step as $T_\sigma = \#D/2+\alpha$. All the ICs belonging to K are actually non-faulty if $\alpha>0$ and if the cardinality of the actual fault set is below T_σ [6, 7].

The last step, called Augmentation, exploits the comparison between ICs in set K and the adjacent ICs not in K (which are classified either Z or D) to incrementally augment set K (thus identifying more fault-free ICs) and to identify some faulty ICs.

A detailed description and an evaluation of this algorithm for different grids is reported in [6, 7]. The

Table 2: Simulation results for n=256.

		.1n	.2n	.3n	.4n	.5n
Triangular	*%incomplete*	64.1	100	100	100	100
	$E[\#N_d]$, *(sqm)*	2.51 (2.06)	15.9 (16.27)	88.23 (33.2)	125.29 (16.87)	128.16 (9.11)
	$E[T_\sigma]$, *(sqm)*	250.1 (3.31)	226.46 (20.77)	154.46 (28.24)	134.2 (12.93)	139.95 (6.44)
Square	*%incomplete*	10.88	62.19	98.04	100	100
	$E[\#N_d]$, *(sqm)*	1.33, (0.92)	2.54, (2.93)	13.79, (12.99)	61.12, (27.921)	96.3, (15.97)
	$E[T_\sigma]$, *(sqm)*	253.72, (1.43)	249.84, (3.6)	230.68, (16.94)	182.34, (26.49)	160.72, (13.88)
Hexagonal	*%incomplete*	5.33	26.23	62.03	92.25	100
	$E[\#N_d]$, *(sqm)*	1.22, (0.59)	1.454, (1.04)	2.618, (2.76)	6.076, (5.75)	27.538, (21.23)
	$E[T_\sigma]$, *(sqm)*	255.004, (1.07)	253.624, (1.78)	251.104, (3.67)	245.144, (7.77)	221.92, (20.6)
Octagonal	*%incomplete*	0.43	4.9	23.76	64.27	97.28
	$E[\#N_d]$, *(sqm)*	1.25, (0.62)	1.459, (1.24)	2.137, (2.02)	3.859, (4.43)	15.843, (15.83)
	$E[T_\sigma]$, *(sqm)*	255.692, (0.57)	255.372, (0.83)	254.62, (1.87)	252.184, (4.02)	238.7, (16.19)

expected values of T_σ and of the degree of diagnosis completeness were evaluated by means of simulation experiments. Results obtained with different grids interconnecting n=256 units are reported in Table 2. The faults, whose number ranged from $0.1n$ to $0.5n$, were distributed uniformly over the grids, and the outcomes of comparisons involving pairs of faulty units were generated assuming probability 0.5.

Table 2 reports the percentage of incomplete diagnoses (entry *%incomplete*), the average number of suspect units (entry $E[\#N_d]$) and the average value of T_σ (entry $E[T_\sigma]$). The standard deviation (entry *sqm*) of $E[\#N_d]$ and $E[T_\sigma]$ is also reported.

4 Conclusions

We have presented a methodology, based on the theory of system-level diagnosis, to execute the wafer-scale test of ICs. The proposed methodology provides correct and almost complete identification of good ICs and covers most systematic fault situations. It is expected that the new methodology could considerably reduce the manufacturing cost, by avoiding the cost of bounding and packaging of most faulty ICs, and by avoiding the need of sophisticated test machines. The time needed to execute the test could also be greatly reduced since all ICs in the same wafer can be tested in parallel.

5 References

[1] F.P. Preparata, G. Metze, and R.T. Chien, "On the Connection Assignment Problem of Diagnosable Systems". *IEEE Trans. on Comp.*, vol. EC-16, pp. 848 - 854, Dec. 1967.

[2] Malek, M., "A Comparison Connection Assignment for Diagnosis of Multiprocessor Systems", *Proceedings of the 10th Symposium on Computer Architecture*, May 1980, pp. 31-35.

[3] S. Rangarajan, D.,Fussel, and M. Malek, "Built-in Testing of Integrated Circuit Wafers", *IEEE Trans. on Comp.*, vol. 39 n. 2, pp. 195 - 205, February 1990.

[4] L.E. LaForge, K. Huang, and V.K. Agarwal, "Almost Sure Diagnosis of Almost Every Good Element", *IEEE Trans. on Comp.*, vol. 43 n. 3, pp. 295 - 305, March 1994.

[5] K. Huang, V.K. Agarwal, L. LaForge, and K. Thulasiraman, "A Diagnosis Algorithm for Constant Degree Structures and Its Application to VLSI Circuit Testing", *IEEE Trans. on Parallel and Dist. Syst.* vol. 44 n.4, pp.363 - 372, April 1995.

[6] S. Chessa, *Self-Diagnosis of Grid-Interconnected Systems, with Application to Self-Test of VLSI Wafers*, doctoral dissertation, University of Pisa, Italy, Department of Computer Science, TD-2/99, March 1999.

[7] P. Santi, *Evaluation of a Self-Diagnosis Algorithm for Regular Structures*, PhD Thesis, University of Pisa, Italy, Department of Computer Science, November 1999.