2004


Shanrui Zhang

Minsu Choi
Missouri University of Science and Technology, choim@mst.edu

Nohpill Park

Fabrizio Lombardi

Follow this and additional works at: http://scholarsmine.mst.edu/faculty_work

Part of the Electrical and Computer Engineering Commons

Recommended Citation
http://scholarsmine.mst.edu/faculty_work/377

This Article - Conference proceedings is brought to you for free and open access by Scholars' Mine. It has been accepted for inclusion in Faculty Research & Creative Works by an authorized administrator of Scholars' Mine. For more information, please contact weaverjr@mst.edu.
Cost-Driven Optimization of Fault Coverage in Combined Built-In Self-Test/Automated Test Equipment Testing

Shanru Zhang1, Minsu Choi1*, Nohipill Park2 and Fabrizio Lombardi3
1 Dept of ECE, University of Missouri-Rolla, Rolla, MO 65409-0040, USA, {sz2k2,choim}@umr.edu
2 Dept of CS, Oklahoma State University, Stillwater, OK 74078, USA, npark@cs.okstate.edu
3 Dept of ECE, Northeastern University, Boston, MA 02115, USA, lombardi@ece.neu.edu

Abstract - As the design and fabrication complexities for the instrumentation-on-silicon systems intensify, optimization of combined Built-In Self-Test (BIST) and Automated Test Equipment (ATE) testing becomes more desirable to meet the required fault-coverage while maintaining acceptable cost overhead. The cost associated with combined BIST/ATE testing of such systems mainly consists of the following: 1) the cost induced by the BIST area overhead and 2) the cost induced by the overall testing time. In general, BIST has faster testing speed than ATE, while it can provide only limited fault-coverage and driving higher fault-coverage from BIST means additional area cost overhead. On the other hand, higher fault-coverage can be usually achieved from ATE, but excessive use of ATE results in additional test time cost. Fault-coverage of BIST and ATE plays a significant role since it can affect the area overhead in BIST and test time in BIST/ATE. This paper is to propose a novel numerical method to find an optimized fault-coverage implemented in BIST and ATE so that a minimum cost can be achieved. The proposed method, then, is applied to two parallel combined BIST/ATE testing schemes to assure its technical validity.

Keywords - BIST, ATE, combined BIST/ATE, yield, fault coverage, optimization.

1. INTRODUCTION

In the last decade, instrumentation-on-silicon technology has been developed rapidly which results in highly integrated and complex chips. So the testing is becoming more and more difficult with the comparatively old test machines. Usually Automatic Test Equipment (ATE) is used to drive the test patterns to the device-under-test (DUT) and then strobe the output from it to see the test result is a pass or fail. The testers used in industry usually have up to 1024 channels [12]. Each channel can drive or strobe signal from the 1 or more DUT pins and can measure the electrical parameters of the DUT. Compared with the current GHz chips, the most testers are working at 100MHz level [12]. The full speed test machine is too expensive for production so that some other testing techniques are used to complement ATE testing [13]. BIST is widely used for this purpose nowadays [6], [11]. The main advantage of BIST is the fast testing speed which is almost at the full speed of DUT [6] regardless the speed of the test machine, its speed is kind of tester independent. So for the high speed BIST testing takes the place of ATE for speed up. The disadvantage is although there are lots of sophisticated way to generate test patterns for BIST, still many random-resilient faults can not be detected while those faults can be detected by deterministic fault-oriented patterns in ATE testing [6], [11]. Considering test time is one of the most contributor to test cost, it is obvious that using BIST test to achieve a relatively acceptable fault coverage which can result in a test time reduction because of the higher testing speed, then using the deterministic patterns to achieve the required fault coverage by slower ATE testing. On the other hand, the higher BIST fault coverage will result in a larger BIST area overhead which may significantly increase the cost of designing and silicon fabricating. So these two portions of the test cost must be taken into consideration for the overall cost, and this paper will discuss this issue later. Parallel testing is another factor which should be taken into consideration for the purpose of the overall test cost reduction. The typical tester in manufacturing industry has 256 to 512 channels so that it is possible to put multiple DUTs onto the test head to do the testing in parallel, meaning that a set of DUTs tested at the same time. This kind of method is widely used for memory chips like SDRAM, Flash because those kind of chips usually have less pins so that testers have enough channels to support like 32 or 64 DUTs parallel testing [12]. For some very complex devices like CPUs or chipsets, which have hundreds of signal pins, a single test head can not support parallel testing due to the limitation of channels. Some advanced testers can support more than one test head so that parallel testing can be applied as well. So the parallel testing can be used in more complex chip testing as a way to increase the test yield and efficiency. As discussed in [1], during parallel test, most of the resources of the ATE such as memory, test channels, power supply, are shared among DUTs. Therefore, when determined as faulty, a DUT can not be replaced until the test process for all the DUTs completed [12]. So in this process, the channels assigned to those DUTs, which have already been diagnosed as faulty, are idling until the test finish. The idle time is a function of the yield and faulty coverage [1] and contributes to the total cost of the overall test process which will be analyzed in this paper.
II. PROBLEM DEFINITION

This paper deals with the cost-driven optimization of combined BIST/ATE testing. The total test cost is comprised of several factors, such as yield, fault coverage for different test stage, test sequence, test process modeling and so on. A novel method to optimize the factors in order to minimize the total cost is to be proposed in this paper. As a few works have been done to analyze the parallel testing in [1], similar approaches are used throughout the paper, yet more practical factors are considered to model the test process with respect to the overall test cost.

1. BIST/ATE in separate stages: the DUTs go through the BIST test first. DUTs which have been found faulty by BIST are screened out (i.e., will not go to the next ATE test stage) by delayed replacement. The DUTs which can pass BIST test go to ATE test in another stage.
2. BIST/ATE in same stage: the DUTs go through the BIST test first, too. No matter what is the test result of BIST, all the DUTs will go to ATE test without being purged off from test head.

Compared with ATE stage, the tester used in BIST stage is not required to be so advanced, and usually has slower working frequency, small memory, etc. So the price of the tester is cheaper as well as the test time cost. In this case, we can combine the BIST with some comparatively simple DC testing such as open/short test, IDDQ/leakage test in the low end tester and in the later discussion in this paper, a cost saving can be reached by using this scheme in some situation. After this test stage, pass those good DUTs to the following ATE stage. In this stage, usually a high-end tester with higher speed and memory is needed to drive test patterns to DUTs and observe the results. So, the cost per unit-time is different in these two stages from the machine utilization point of view. An additional touchdown between these two stages should be taken into consideration as well.

For the second approach, BIST and ATE are utilized in the same test stage. So the tester used for this approach is usually a high-end tester, which is more costly than the tester for the first approach. The advantage of this scheme is, because the DUTs which failed in BIST test do not need be removed from the test head, no additional touchdown is required so that it can help saving test time cost.

As already shown in [1], the test time is a function of fault coverage. For BIST testing, additional circuitry should be added to the chip to get a certain desired level of fault coverage. Compared with ATE, which uses Automatic Test Pattern Generation (ATPG) tools to generate patterns to achieve a desired level of fault coverage, the fault coverage achieved by BIST is much more expensive. Because of the very fast application time in BIST test, the test time in BIST stage is much lower than in ATE which can contribute to the cost reduction. So this paper will find a balance point to achieve minimum cost.

A novel cost-driven optimization technique for combined BIST/ATE testing is proposed, and then validated through parametric simulations in this paper.

III. PRELIMINARY

A few numerical models for test time of parallel testing have been reported in [1] and the proposed cost-driven optimization technique is based on the models.

The following notation will be used throughout this paper.

1. \( t_p \): the expected time required for a DUT to pass the entire test process
2. \( t_f \): the expected time required to diagnose a faulty DUT
3. \( y \): the yield
4. \( V \): the number of total test vectors
5. \( n_0 \): the expected number of faults per faulty chip
6. \( f_M \): required fault coverage for the whole testing process
7. \( f_E \): fault coverage achieved by ATE testing
8. \( N_x \): the number of patterns in the minimal test set applied by ATE
9. \( T_{avg} \): the average test time for parallel testing

In order to characterize the test time in parallel testing, several terms have been defined in the paper. \( y \) is the yield (i.e., the number of good DUTs at the end of testing divided by the total input DUTs in percentage), \( t_p \) is the test-time-good which is the expected time required for a good DUT passing the whole test process, and \( t_f \) is test-time-bad which is the expected time required to diagnose a faulty device. So the expected test time of a DUT, which is designated by \( t \), can be calculated by

\[
t = t_p \times y + t_f \times (1 - y)
\]

(1)

As per the discussion in [2], \( n_0 \) is the possible average number of faults per faulty chip, so the probability of a faulty DUT to be detected by test vector \( v \) (among the total of \( V \) test vectors) can be given by

\[
p_f = \frac{(V - v + 1)^{n_0} - (V - v)^{n_0}}{V^{n_0}}
\]

(2)

From [1] and [2], the expression can be explained as follows; there are total of \( V^{n_0} \) combinations in which the \( V \) available vectors can detect the \( n_0 \) faults. The number of combinations in which only a vector \( v \) detects the fault and none of the previous vectors (1 to \( v-1 \)) detects any of the \( n_0 \) faults is \((V - v + 1)^{n_0} - (V - v)^{n_0} \). Thus the expected test-time-bad for a faulty device with \( n_0 \) faults is given by

\[
t_f \propto \sum_{v=1}^{V} \frac{(V - v + 1)^{n_0} - (V - v)^{n_0}}{V^{n_0}} \times v \approx \frac{V}{n_0 + 1}
\]

(3)

Also we define \( C \) as the number of DUTs can be put onto test head simultaneously for parallel testing. \( D \) as the number of total input DUTs. As shown in [1], the average test time of a parallel testing is

\[
T_{avg} = (1 - y)^C \times t_f + (1 - (1 - y)^C) \times t_p
\]

(4)
In [4], an almost linear relationship between test set size and fault coverage has been found in semi-logarithmic scale. In other words,\[ f_E = \frac{f_M}{\log_{10} N} \times \log_{10} V \] (5)

where \( f_E \) is the fault coverage achieved by ATE testing, \( f_M \) is the required fault coverage for the whole testing process (combined BIST/ATE), \( V \) is the test vector size in ATE and \( N \) is the number of patterns in the minimal test set which are applied by the ATE [1], [4].

IV. ANALYSIS ON TEST TIME COST

In the manufacturing factory, test time can significantly affect most of the costs such as machine utilization, direct people resource, relevant material storage (auxiliary machines), etc. Thus, if test time can be reduced, the total testing cost can be cut significantly. Considering the last testing speed of BIST, as the BIST test coverage increases, to reduce the ATE test coverage to speed up the overall test process, the total test time can be significantly reduced.

The following notation will be used throughout the paper, in addition to the notation defined in the previous section.

1. \( \lambda \): a constant to model the random-resilience of faults
2. \( f_B \): test coverage achieved by BIST testing
3. \( y_B \): input yield of BIST stage
4. \( y_E \): input yield of ATE stage
5. \( t_{PB} / t_{PE} \): the test time a DUT passing/failing in BIST test
6. \( t_{PB} / t_{PE} \): the test time a DUT passing/failing in ATE test
7. \( T_{avgB} \): the average test time for BIST testing

By the Williams’ model shown in [3], the BIST coverage after applying \( N \)th vector can be expressed as follow
\[ f_B = f_M \times (1 - e^{-\lambda \times \log_{10} N}) \] (6)

where \( \lambda \) denotes a constant that is used to model the random-resilience of faults. The larger \( \lambda \) is, the easier the fault can be detected by random test patterns. Since the test time is proportional to the number of test vectors, we can solve \( N \) from the equation and then get the test time as follows
\[ t_{PB} \propto N = 10^{\frac{-1}{\lambda}} \times \ln(1 - \frac{f_B}{f_M}) \] (7)

Because the passing/failing status can be known for a DUT only after all the test vectors are executed, the test-time-good and test-time-bad are the same in BIST. So we have \( t_{PB} = t_{PB} \).

From equation (4), the average test time \( T_{avgB} \) in BIST stage for total \( D \) DUTs can be calculated as follows
\[ T_{avgB} = D/C \times t_{PB} \] (8)

Using the same method, we can solve \( V \) from equation (5), and
\[ f_E = f_M - f_B \] (9)

also the test time good in ATE is proportional to the test vector size, and can be expressed as follows
\[ t_{PE} \propto V = 10^{\frac{-1}{\lambda} \times \ln(1 - \frac{f_E}{f_M})} \times \log_{10} N \] (10)

From equation (3) and (10), it can be observed that the \( t_{PB} \) proportional to the function of \( f_B \). Now let us scale the test time and let \( \alpha \) denote the relative speed of BIST over ATE [4], [5], so that we have
\[ t_{PB} = 10^{\frac{-1}{\lambda} \times \ln(1 - \frac{f_B}{f_M})} \times \log_{10} N \] (11)
\[ t_{PE} = \alpha \times t_{PB} \] (12)
\[ t_{PE} = \frac{\alpha}{1 + \ln(1 - \frac{f_E}{f_M})} \times \log_{10} N \] (13)

In the first approach, the detected faulty DUTs by BIST should be removed from the total \( D \) DUTs. In order to get the parallel test time in ATE stage, we need calculate the remaining \( T \) DUTs after BIST.

In [6], Williams and Brown have shown that
\[ D_L = 1 - y_B = 1 - y_1^{1-f} \] (14)

where \( y_B \) is the input yield of a test stage, \( y_0 \) is the output yield of the test stage and \( f \) is the fault coverage of the current test stage. Because the combined BIST and ATE fault coverage is \( f_M \), and the ATE coverage is \( f_M - f_B \), we can calculate the input yield of BIST \( y_B \) and input yield of ATE \( y_E \), if the final yield \( y \) is known.
\[ y_B = y_1^{1-f_M} \] (15)
\[ y_E = y_1^{1-f_{ME}} \] (16)

Since BIST and ATE are two successive stages, the \( y_E \) is also the output yield of BIST \( y_B \). Thus the number of good DUTs after BIST, which is designated by \( M \), is
\[ M = D \times \frac{y_B}{y_0} = D \times y_1^{1-f_B} \] (17)

In order to calculate the test time in ATE by equation (4), we need know the "true" yield of ATE (i.e., the number of good DUTs passing all the test process divided by the number of total input DUTs in percentage value)
\[ y_E = y_1^{1-f_{ME}} \] (18)

Now we can get the test time in ATE, which is \( T_{avgE} \)
\[ T_{avgE} = M/C \times \left[ (1 - y_1^{1-f_B})^C \times t_{PB} + (1 - (1 - y_1^{1-f_B})^C) \times t_{PE} \right] \] (19)
Since we use different testers in these two serial test stages, and ATE cost is much higher than BIST cost because the ATE testing requires more test channels to drive and strobe signals from the DUT so that the machine is more advanced. Thus, it is assumed that the cost in ATE is $\beta$ times higher than in BIST stage. Let us further assume that cost in BIST is expected as dollar per unit test time and the touchdown time cost is $\phi$ which is based on the manufacturing process. We can scale $\phi$ as a percentage of cost in ATE. So, the cost for the test process is

$$Cost_{test} = 1 \times T_{avgB} + \beta \times (1 + \phi) \times T_{avgE}$$  \hspace{1cm} (20)$$

The $\phi$ can reflect numerous factors of cost in the manufacturing line. It highly depends on the process, like how to get the DUTs out from the first tester and how to merge several subsets of DUTs into a new set and then go to the next stage. So once the manufacturing process of the factory is defined, the $\phi$ can be obtained from the empirical data. $\phi$ has a significant impact on the total cost in test as shown in figure (1).

$$t_{pBE} = t_{pB} + t_{pE}$$  \hspace{1cm} (21)$$

Assuming that all faults are equally likely, $f_B$ is the probability that a DUT fails at BIST stage and $f_M-f_B$ is the probability that a DUT fails at ATE stage. So, the fail time is

$$t_{fBE} = t_{fB} \times f_B + t_{fE} \times (f_M - f_B)$$ \hspace{1cm} (22)$$

Now $t_{pBE}$ and $t_{fBE}$ can be plugged into equation (4), and the total parallel test time can be obtained from

$$T_{avgBE} = \frac{D}{C} \times \left[ (1 - y)^C \times (t_{fB} \times f_B + t_{fE} \times (f_M - f_B)) + (1 - (1 - y)^C) \times (t_{pB} + t_{pE}) \right]$$ \hspace{1cm} (23)$$

It is notable that the tester for this testing scheme is an advanced tester and there is no additional touchdown time cost $\phi$. Now we can have the cost in this scheme

$$Cost_{testBE} = \beta \times T_{avgBE}$$ \hspace{1cm} (24)$$

Now let us compare the two different approaches. Parametric simulation results are shown in figure (2) and figure (3), where $\alpha=100$, $\beta=2$, and $\phi$ has different values 0, 0.1, 0.2, 0.3. We can see the $\phi$ plays a significant part on the difference between these two approaches, when the $\phi$ increases, the first approach shows a worse performance from the cost point of view compared with the second approach.

Another parametric simulation results are shown in figure (4) and figure (5), where $\phi=20\%$, $\beta=2$, and $\alpha$ has different values from 1, 10, 50, 100.

The $\alpha$ is the BIST/ATE speed ratio. In a certain manufacturing process, the parameters such as $\beta$, yield, $\phi$ are known, so that if $\alpha$ is small, and pretty high $f_B$, the first approach has some advantages over the second one. Also, there are many other factors which can affect the cost results of those two approaches. For a certain manufacturing process, those relevant parameters can be plugged into the equations (20) and (24) and then choose the less costly one of these two different schemes. Those parameters can be either obtained from factory processing or empirical data. So, we can use the same analysis method above to decide which approach is more suitable for a certain case.
To implement BIST, we need to build additional circuitry on the chip to realize the testing function; meanwhile an additional cost is induced. Compared with the test time cost in the manufacturing process, the physical chip fabrication cost is much higher. Also, the higher fault coverage means a higher cost. Although as shown in figure (1) the cost in test time decreases when fault coverage in BIST increases, we cannot let the fault coverage to be unrealistically high because the additional cost due to BIST area overhead should be considered at the same time.

Let us scale the cost in BIST overhead by unit-test-time cost in BIST. Assuming the unit overhead cost is $\gamma$ times higher than unit-test-time cost in BIST, the area overhead cost of $D$ DUTs is

$$Cost_{area} = \gamma \times D \times Area$$

The $Area$ is BIST circuit in square cm divided by total chip circuit in square cm in percentage.

VI. OVERALL COST COMBINED TEST TIME COST AND AREA OVERHEAD COST

From the above analysis, the overall cost can be obtained by adding those two separate costs together

$$Cost_{all} = Cost_{test} + Cost_{area}$$
Drawing a plot of $Cost_{total}$ in dollar value in Y axis versus BIST fault coverage $f_B$, in percentage value in X axis, we can see that there is a minimum point in the curve which is the minimum cost point we need achieve.

In order to draw figure (6), several parameters are predefined. The required fault coverage $f_{MT}$ is equal to 98%, the overall yield is equal to 95% which we can obtain from industry empirical data; $\alpha$ equal to 100 which means BIST test speed is 100 times faster than ATE; $\beta$ is equal to 2 which means the tester in ATE stage is 2 times more expensive than the one in BIST stage; $\phi$ is equal to 20%, and those three parameters $\alpha, \beta, \phi$ are based on which kind of manufacturing process is employed. $N_a$ is 10 as well as $\lambda$ which are from empirical data.

There are five curves in the graph: BIST area overhead cost curve, test time cost curves for the two approaches and total cost curves for the two approaches respectively. Both of the total cost curves have the minimum point at $f_B=70\%$, which means we can achieve minimum cost if we design the BIST circuit to get 70% fault coverage in the conditions we described in above paragraph. But, obviously, in this case, the second approach costs less than the first one. Thus, in this kind of manufacturing process, we can feedback to design engineers to develop BIST with 70% fault coverage and generate ATE test patterns to achieve the rest $f_{MT}, f_B=28\%$ fault coverage; and use the second approach which means put BIST and ATE in one test stage (i.e. same tester) then can achieve the minimum cost.

The proposed optimization model is flexible enough to accommodate different sets of parameters to find the optimal cost and BIST fault coverage combinations.

VII. CONCLUSION

In this paper, a method to optimize the fault coverage in BIST combined with ATE testing to achieve a minimum cost is proposed. Based on significant testing parameters such as yield, touchdown time $\phi$, BIST/ATE speed ratio $\alpha$, and the relation table of area overhead versus BIST fault coverage $f_B$ to determine which parallel testing scheme to use and how high BIST fault coverage $f_B$ should be maintained in order to achieve a minimal cost. The main purpose of this paper is to find a optimal point of $f_B$ at which a minimum cost of the whole processing in a certain manufacturing processing system can be achieved. The cost saving efficiency of these two approaches is analyzed in different situations and the way to make selection between those two approaches is proposed. Parametric simulation results assure that the proposed cost optimization technique is simple and effective to find the optimized parameters.

REFERENCES