J Electron Test (2016) 32:369–375
DOI 10.1007/s10836-016-5585-1

Real-Time Adaptive Test Algorithm Including Test Escape Estimation Method

Christian Streitwieser[1]

Received: 14 January 2016 / Accepted: 18 April 2016 / Published online: 25 April 2016

Abstract This work aims to reduce the test time per chip, but affect the quality of the products only as little as possible compared to the conventional semiconductor test. For that purpose, a simulation software has been developed that performs adaptive test simulations based on real data and test results of conventional production testing. Products has been analyzed and both the potential for test time savings and the resulting loss of quality are determined. Finally, it is shown that by the intelligent omission of tests, compared to conventional testing procedures, test time reduction of up to 50 % can be achieved. In contrast, there are only few defective components that are not detected by the adaptive test procedure and which reduced the quality of products. In best case, the fraction of test escapes was under 40 ppm. To handle the potential risk of test escapes this paper also introduces a statistical estimation method.

Keywords Adaptive test (AT) · Test time reduction (TTR) · Statistical process control (SPC) · Part average testing (PAT)

1 Introduction

Traditionally, all ICs are tested identically during each backend test insertion and the list of tests, test limits, test flow and test conditions are always the same when it comes to conventional semiconductor testing. Whenever a change or optimization of this conventional test procedure is required it's done offline, which makes it impossible to adjust the way how a device is tested during production. Despite its advantages, conventional semiconductor testing is a fixed and rigid process and human involvement is needed for any adaption and improvement. In contrast, the applied adaptive test approach can be seen as an extension of conventional test process.

As shown in Fig. 1 the approach is a test process with a closed loop control system extended by controlled statistical analysis of the measurements. An adaptive test (AT) controller makes predictions about the behavior of the device under test (DUT) based on statistical data analysis and afterwards triggering targeted test sets for the current DUT. By that directed decisions for the next tests on the same die or for the next devices could be derived and tests applied on each die are individually. With the method it's possible to dynamically adapt for maximum cost effectiveness by tuning the closed loop system in real-time, which means after every single measurement [2, 5].

To sum up, the purpose of adaptive testing in this work is to reduce test costs through test time reduction without any significant losses in quality by optimizing the defect detection probability.

2 Method

2.1 Adaptive Test Algorithm

The algorithm is capable of adapting the test content based on prior measurements of the same test session as well as measurements on the current DUT. Thus the scope of the algorithm is limited to the lifetime of a test program which is for instance from the beginning to the end of a wafer sort. The

Responsible Editor: V. D. Agrawal

✉ Christian Streitwieser
christian.streitwieser@ams.com

[1] Technology/Corporate Test Development, ams AG, Unterpremstaetten, Austria

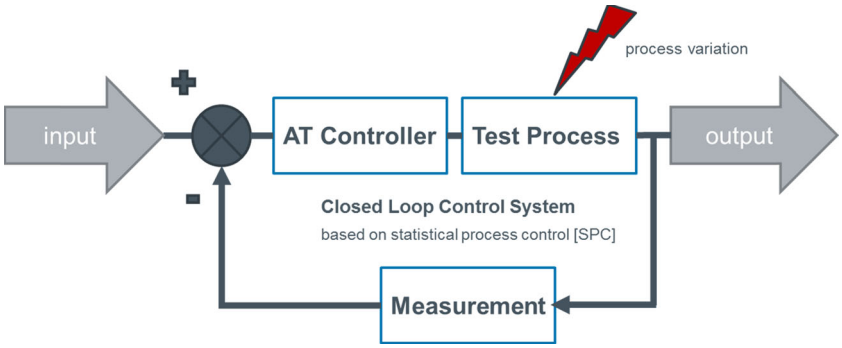

Fig. 1 Adaptive test closed loop control system

basic concept of the introduced adaptive algorithm, as shown in Fig. 2, is described in this section.

To obtain statistics for each test it is necessary to collect data at the beginning of the session for all available test for the first devices. Therefore, initially a statistical control sample has to be drawn which is determined by the control sample size (CSS) parameter and performed with the full test list for each DUT. A full test list is executed in the flow chart as long as the DUT counter is smaller than the predefined CSS parameter. After the initial sample, if the information content and the statistical significance are considered to be large enough, the tests which are most likely to fail should never be skipped. Tests which seem to be under statistical control are allowed to be omitted with the aim of saving test time. Nevertheless, they are not skipped entirely during the further testing process, but sampled sometimes to verify the assumption that they are not likely to fail. This is achieved with another decision branch and a predefined SAMPLE_RATE parameter in the graph.

The question is now, which tests are likely to fail and which are not? This classification is based on a fail pareto analysis which is updated dynamically during the whole test session. Through a pareto analysis it is possible to identify the most frequently occurring test that fails. So it simply represents a frequency distribution of discrete data arranged by category. The pareto method does not automatically identify the most important tests, but rather only those that fail most frequently [4]. The threshold within the algorithm for the decision of whether or not a test is allowed to be skipped is very strict: Once a test shows an error, it is no longer allowed to skip it in further test session. On the other hand, only tests which never show fails in the prior executions are allowed to be skipped. To sum up, the algorithm until now achieves its goal of test time reduction due to a test sampling procedure based on a dynamic test fail pareto analysis. The aim of the sampling procedure is to improve test efficiency by only testing as much as is required to meet the outgoing quality target given the incoming manufactured quality level [4]. It is clear that two wafers of the same product with two different yields have the same requirements concerning outgoing quality. The algorithm's task is now to increase the test intensity on the wafer with the worse manufacturing quality compared to the other one to reach this goal.

An improvement concerning fault coverage, which means minimizing the number of test escapes, is added by extending the algorithm with a statistical outlier monitoring method. The

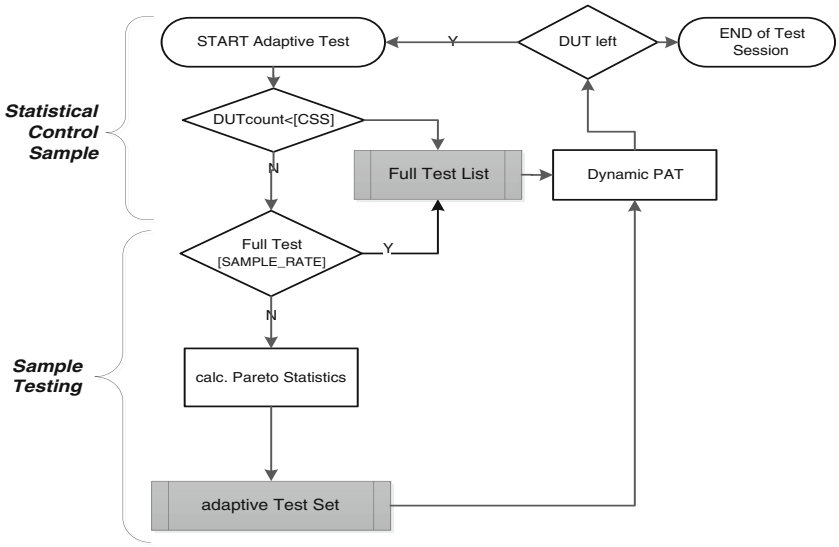

Fig. 2 Adaptive test algorithm

basic concept here is that if a test differs significantly from an expected behavior but is still inside of the specified test limits, that part is perhaps bad. So when each part originates from the same production process it is assumed that if a device differs from its population a defect may be the reason for its suspicious behavior. Specifically, the algorithm developed uses a technique based on the so-called dynamic part average testing (PAT) method from the Automotive Electronics Council (AEC) [1]. The main concept of dynamic part average testing is to derive suitable limits on measurement parameters based on a moving window of most recently tested parts.

In the end, based on the different statistical post-processing methods, either targeted adaptive test content or a full test set is executed on the DUT. To find adequate values for the 2 mentioned predefined parameters of the algorithm, an exhaustive search, also known as the brute force method in computer science, is used. In general, the exhaustive search method is a method for optimization and problem solving. The method is based on the trying of all possible or, at least, of many feasible cases to find a satisfying result to the problem. [3] The best case for the adaptive test simulation would be a big test time reduction without any impacts on the quality of test. This implies the same fault coverage as the conventional test would be desirable. In practice a trade-off between fault coverage and test time reduction (TTR) will be needed.

2.2 Test Escape Risk Management

In the following an approach for calculating and estimating the lot fraction defective of adaptive testing is introduced. Lot fraction defective is a statistical term and refers to the portion of faulty ICs among all the others. It's obvious that adaptive testing represents a type of a statistical sampling procedure. If a wafer is not tested with all the tests on every die, it is clear that there can be no 100 % assurance that all the imperfect parts are detected. But one thing that can be done is to estimate the fraction of defects that may occur in the untested units of the wafer, based on information from the previously tested units. The accuracy of this estimation can be refined by specifying its uncertainty using a proper statistical confidence interval. These on inferential statistics based calculations are then used to specify and derive the lot fraction defective caused by the adaptive test algorithms.

In Fig. 3, on the left side, a pass/fail wafer map of a conventional wafer sort is shown in which one square represents one DUT. Each DUT is subjected to a test set with four tests which are always executed in sequence, beginning with test 1 and ending with test 4. The detailed considerations for example single DUT results from this testing are shown on the right in Fig. 3. In situation A, the die has succeeded in all the tests and is marked as a pass. The next situation, Situation B, shows a die, which has failed the second test after previously passing the first test beforehand. After failing this test 2, the IC is not tested further, and test 3 as well as test 4 will be never executed because of the stop on fail (SOF) method of conventional production testing. The last example, C, again shows a faulty die which has already failed the first test. Testing is stopped after this point. This means that each IC is subjected to the full test set, and once only one test fails in the row is the whole die marked as fail (red color). Only if an IC passes the entire test set, it can be marked as faultless, which is then displayed in green. To put it more generally, it can be said that an IC consists of a number of subunits which can take three information conditions. These three conditions are pass, fail and untested, whereas an IC only works if all the subunits are free of errors.

In contrast, in Fig. 4 on the left side of the drawing the same wafer is shown. This time the wafer was tested with an adaptive algorithm. For demonstration purposes the same dies as in the other wafer map are divided into four subunits and again each subunit refers to a test. Because of the adaptive test algorithm it is now the case that not every die will be tested with all the tests available, so that some subunits of the chip will be left untested. For these kinds of subunits, no pass/fail information as in the conventional SOF method is available and they are displayed as white squares in the example wafer map. The final decision as to whether a device is faulty or faultless is the same; once a subunit fails the overall die is declared to be faulty. The difference to the conventional test is that untested subunits can also occur between other tests and are not always lined up at the end of the testing sequence. So it is also possible that an IC is considered as error-free if not every subunit was tested – situation A. C differs from B because test 2 was left out in-between, which is also quite possible. However, both are considered to be faulty because test 3 failed each time. Here it is again clear that untested units may result in undetected faults.

Now the basic idea is to use the information about the tested subunits, which includes the corresponding test result and based on that, estimate the appearance of defects in the fraction of untested subunits. Afterwards the estimated portion of faulty subunits is used to derive an estimate for the lot fraction defective of ICs resulting due to adaptive testing.

The first assumption for this failure estimation model is that every subunit is statistically independent, which means that the outcome of each subunit does not depend in any way on the outcome of another subunit. Every subunit has the same possibility either to be faulty or not. Therefore the tests are also assumed to be independent of each other, and in the example of Fig. 4 one die exists of four subunits.

Moreover, the fault distribution is assumed to follow a binomial model. A binomial distribution is a probability model for sampling from an infinitely large population, where p represents he fraction of defective or nonconforming items in the population. In this context, x represents the number of nonconforming items found in a random sample of n items

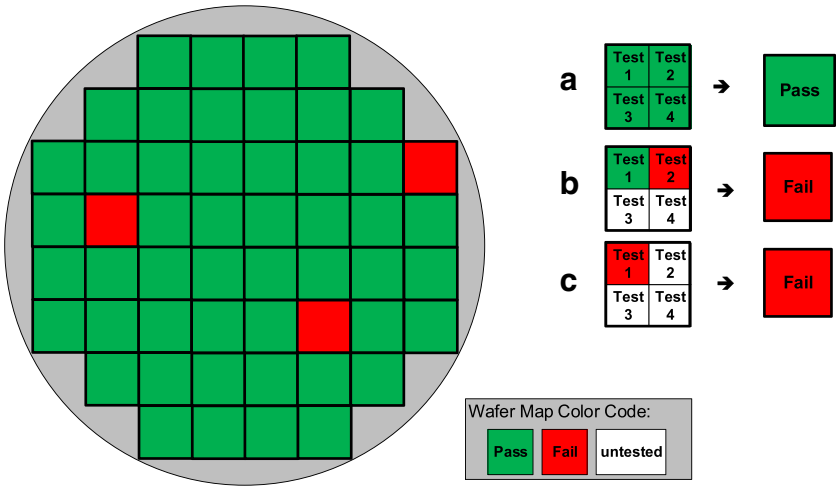

Fig. 3 Wafermap with pass and fail information content based on stop on fail (SOF) strategy of conventional test

of the population. The binomial distribution function is defined as follows in Eq. (1).

$$B_{n,p}(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

After an adaptive test session has been completed, how often each single test was executed by the algorithm can be exactly traced back. Figure 5 represents, for example, a product with 20 tests for each DUT and 3200 dies on the wafer. So after the algorithm has finished for each test, the information about its sample size n_i is available. Furthermore, the exact information how often each test failed during the simulation is also available because the fail pareto list contains these statistics. The number of faults which occurred per test is x_i. Based on this, it is now possible to estimate the lot fraction defective per test with $\hat{p}_i$, the so-called point estimator for the binomial distribution (2).

$$\hat{p}_i = \frac{x_i}{n_i} \tag{2}$$

The assumption is that one test is a binomial model with the parameter n_i, p_i (3)

$$T_i \tilde{} B(n_i, p_i) \tag{3}$$

When the number of tests is t, the model of the whole chip test T can be written as (4):

$$T = \sum_{i=1}^{t} T_i \tag{4}$$

Since the initially assumed independency of the tests, the individual defect portions p_i have to originate from a population with the same lot fraction defective. So every test sample with its size n_i and faults x_i must therefore originate from the population of the subunits with the same lot fraction defective p_{sub} – which is the unknown. This leads directly to (5)

$$T = B\left(\sum_{i=1}^{k} n_i, p_{sub}\right) \tag{5}$$

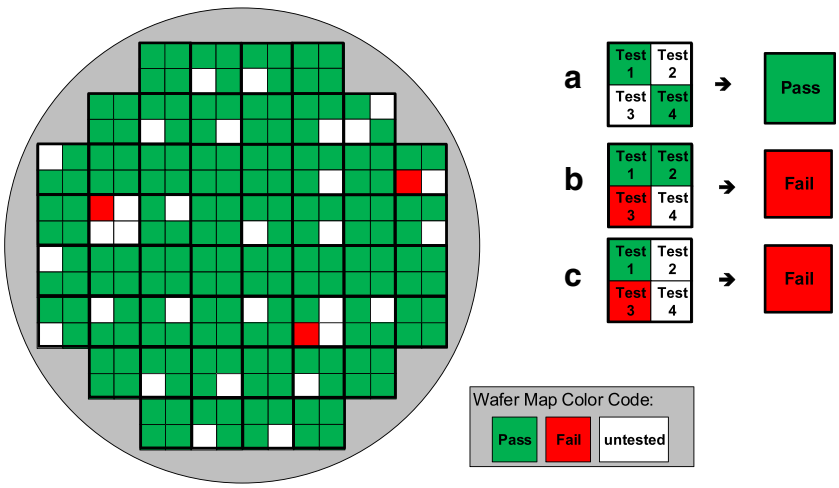

Fig. 4 Wafer map with pass and fail information content based on stop on fail (SOF) strategy of adaptive test

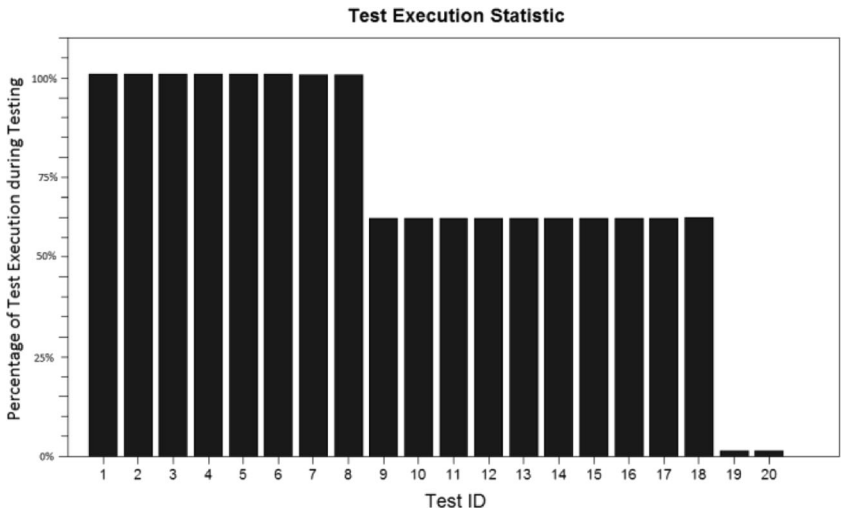

Fig. 5 Execution statistic of an adaptive test session

The point estimator for the fraction defective in the subunits can now be calculated (6):

$$\hat{p}_{sub} = \frac{\sum_{i=1}^{t} x_i}{\sum_{i=1}^{t} n_i} \tag{6}$$

The next goal is to estimate the IC fraction defective based on this subunit level estimate $\hat{p}_{sub}$. Like in the SOF method, for the estimation it is also assumed that simply one defect leads to an overall faulty die. When taking in account, that one die exists of t subunits, the point estimator for the lot fraction defective for ICs can be calculated according to (7):

$$\hat{p}_{lot} = \frac{\sum_{i=1}^{t} x_i}{\sum_{i=1}^{t} n_i} \cdot t \tag{7}$$

The expected value $E(X)$ or faulty dies out of a sample n of dies would refer to (8)

$$E(X) = n \cdot \hat{p}_{lot} \tag{8}$$

To derive a worst case scenario for the lot fraction defective which can occur in the untested subunits the confidence interval is used to indicate the reliability of the estimate. A confidence interval specifies a range within the parameter is estimated to be located. The (1-α) confidence interval for an unknown lot fraction defective p is an interval which includes with a (1-α) probability the unknown lot fraction defective p.

The confidence interval is given by (9) (10) (11):

$$p_l \leq \hat{p}_{sub} \leq p_u \tag{9}$$

$$P\{x \leq E(X_{sub})\} := \sum_{x_{sub}=0}^{E(X_{sub})} \binom{n_{sub}}{x_{sub}} \cdot {p_u}^{x_{sub}} \cdot (1-p_u)^{n_{sub}-x_{sub}} = \frac{\alpha}{2} \tag{10}$$

$$P\{x \leq E(X_{sub})-1\} := \sum_{x_{sub}=0}^{E(X_{sub})-1} \binom{n_{sub}}{x_{sub}} \cdot {p_l}^{x_{sub}} \cdot (1-p_l)^{n_{sub}-x_{sub}} = \left(1-\frac{\alpha}{2}\right) \tag{11}$$

with $x_{sub} = \sum_{i=1}^{t} x_i$ and $n_{sub} = \sum_{i=1}^{t} n_i$

So the lot fraction defective for the subunits lies in between the boundaries $p_l \leq \hat{p} \leq p_u$ with a probability of (1-α) percent. In the worst case situation the lot fraction defective is expected

Table 1 Simulation results compared to Conv. Test

	Product Name:	A	E
Product Information	Number of Tests per Die: *# tests*	20	107
	Number of Dies per Wafer: *# parts*	32000	4694
	Number of Wafer: *# wafer*	25	25
	Yield of conventional Test: %	99.93	92.97
Test Escape	Number of Test Escapes: *# parts*	0.12	3.52
	Number of Test Escapes: *ppm*	37.67	749.89
	Fault Coverage: %	97.70	99.07
Test Time Reduction	Test Time Reduction: %	50.10	17.46
Lot Fraction Defective Estimation	Real Lot Fraction Def. p_{lot}: %	0.07	7.03
	Estimated Lot Fraction Def. $\hat{p}_{lot}$: %	0.10	7.97
	Worst Case Lot Fraction Def. $\hat{p}_{wc,lot}$: %	0.16	8.06

[1] Definition: "The AQL represents the poorest level of quality for the supplier's process that the consumer would consider to be acceptable as a process average." [3, Montgomery, 2005, site 654]

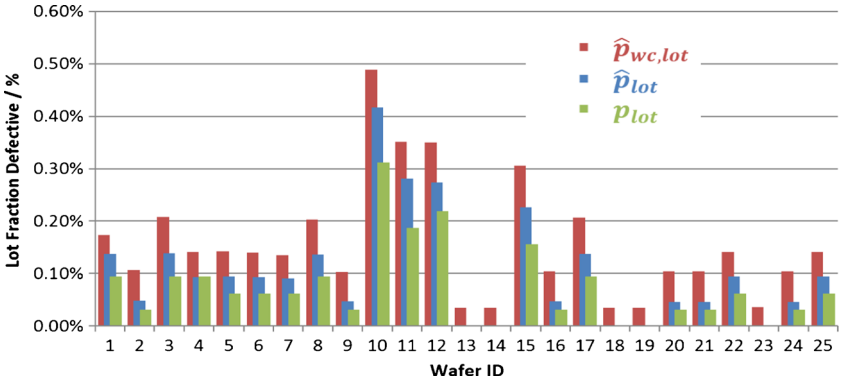

Fig. 6 Lot fraction defective estimation of product A

to be p_u. If a wafer has N dies on it, the number of untested subunits u_i is also known (12):

$$u_i = N - \mathrm{n_i} \tag{12}$$

and the number of defects in the untested can be estimated as $\hat{x}_{wc,usub}$ (13):

$$\hat{x}_{wc,usub} = p_u \cdot \sum_{i=1}^{t} u_i \tag{13}$$

When assuming that every defective subunit results in a faulty die and it is not considered that more than one defective subunit could take place within one IC, the worst case error contribution of dies can be written as (14) which represents the worst case number of estimated test escapes due to adaptive testing.

$$\hat{x}_{wc} = \hat{x}_{wc,usub} = p_u \cdot \sum_{i=1}^{t} u_i \tag{14}$$

Finally, the point estimator for test escapes can be written as (15):

$$\hat{x} = \hat{p}_{lot} \cdot \sum_{i=1}^{t} u_i \tag{15}$$

Additionally, the worst case lot fraction can also be calculated (16):

$$\hat{p}_{wc,lot} = \frac{\hat{x}_{wc}}{N} \tag{16}$$

3 Results and Discussion

The following results were derived with product specific parameter sets for the introduced algorithm with the aim to reduce test costs due to test time reduction. Table 1 shows that for the high fault coverage of about 98 % (in average 0.12 DUTs per wafer), still approximately 50 % test time reduction could be reached for the product A. Test time reductions of around 17 % where the results for product E, whereupon the fault coverage with circa 99 % is considered as good enough. Product E initially has a lower yield than A and this is a reason for the low reduction concerning test time, because for a higher failure rate the algorithm increasingly prohibits the omitting of tests. The initial lower yield also results in a in a test escape rate of 3.52 ICs per wafer. It has to be mentioned here, that the balance between fault coverage and test time reduction could always be modified by another parameter set of the algorithm and can be tailored to the economic interests. The Table 1 also shows estimated values for the lot fraction defective estimation. This allows to number the probably undetected dies of the adaptive test process.

The data of one lot consisting of 25 wafer, Fig. 6, is displayed and the lot fraction defective estimation results for the product A, separate for each wafer, are plotted.

The real fraction defective p_{lot} derived from the conventional test is compared with the point estimation $\hat{p}_{lot}$ as well as the worst case estimation $\hat{p}_{wc,lot}$ from the simulation. It turns out that the calculated value fits very well to the real one and is a useful value to estimate the potential test escape risk.

For instance, the customer accepts a certain defect level of the purchased product, which is normally fixed per contract, the estimated value can be used for acceptance sampling plans. Therefore a certain AQL[1] value is agreed upon and the manufacturer only applies a sample test instead of a 100 % inspection of the delivered ICs. The sampling plan includes the size of the sample which has to be drawn as well as the acceptance number. The acceptance number is the maximum number of nonconforming units that are allowed. This means that once the number of faulty ICs within the drawn sample exceeds this number the whole sample is rejected. The consequence would be a for instance a 100 % full test of the whole wafer [1].

The estimation in the bar plot as well as in Table 1 also indicates that the point estimation $\hat{p}_{lot}$ of the developed method for each product exceeds the real occurring value p_{lot}. Obviously also the worst case estimation $\hat{p}_{wc,lot}$ based on a

[1] AQL or acceptable quality level represents the poorest level of quality for the supplier's process that the consumer would consider acceptable as a process average [4].

90 % confidence interval is always greater than the reality. So it can be said that the lot fraction estimation method consequently overestimates the real fraction of occurring faults, which makes it quiet safe concerning the test escape risk statement. A good fitting estimator would lead to randomly distributed results above as well as below the real values with a very small deviation from reality. From a theoretical and mathematical perspective, there are several reasons for this behavior. As pointed out in the methodology, there were a lot of assumptions made in order to make the estimation possible. The hypothesis that each subunit is independent of each other and in further consequence that the tests are statistically independent is probably one of the weakest points of the estimation method. Thus correlations of tests are common and as a consequence not the whole number of tests is responsible for the real occurring fraction defective. This would lead to a decreased in (7), which further results in a smaller fraction defective estimation. Another reason for the overestimation could be that the tests – subunits t - have a different likelihood to detect a defect, which means they have different test coverage.

4 Conclusion

In this work a new test method is presented which falls under the category of "adaptive test" methods. The approach aims to reduce the test time per chip, but affect the quality of the product as little as possible or at least provides a statistical statement of the potential test escape risk.

For that purpose, a simulation software has been developed that performs adaptive test simulations based on real data and test results of the conventional production testing. The outcome of the simulation covers both, the potential test time savings as well as the resulting quality loss in *ppm*. The underlying algorithms are based on statistical methods and dynamically decide which tests per chip compared to the conventional test methods can be omitted, or not. Furthermore a statistical method for a proper lot fraction defective estimation was introduced which enables a feasible test escape risk management.

Finally, by the intelligent omission of tests, compared to conventional testing procedures, test time reduction of up to 50 % can be achieved. In contrast, there are only few unrecognized but manageable defective components (test escapes in ppm) that are not detected by the adaptive test procedure and reduce the quality of manufacture.

References

1. Automotive Electronics Council (AEC) (2003) Guidelindes for part average testing (AEC - Q001 Rev-C), USA: automotive electronics council component technical committee
2. International Technology Roadmap for Semiconducto (ITRS) (2011) International technology roadmap for semiconductor 2011 Edition, Executive summary, ITRS
3. Kalyanmoy D (2004) Optimization for engineering design: algorithms and examples. Prentice-Hall of India Pvt. Ltd, Kanpur, India
4. Montgomery DC (2005) Introduction to statistical quality contorl, 5th edn. Arizona, John Wiley & Sons, Ine
5. O'Neill PM (2009) Adaptive test, "Avago Technologies"

Christian Streitwieser is a Test Development Engineer. He was born in Salzburg, Austria, in 1986. After the High School Diploma in electrical engineering (2005), he received the Master degree in biomedical engineering from the Technical University of Graz in 2012. In the course of his academic education, Christian got already in touch with ams AG and during his master thesis he worked in close cooperation with the Test Development department. 2012 he joined ams AG as a Corporate Test Development Engineer and is now focusing on R&D as well as production testing topics.