Assessing the Quality Level of Digital CMOS IC's under the Hypothesis of Non-Uniform Distribution of Fault Probabilities

F. Corsi, C. Marzocca
Dipartimento di Elettrotecnica ed Elettronica
Politecnico di Bari
Via Orabona 4, 70125 Bari, Italy

S. Martino
Tecnopolis CSATA Novus Ortus
Laboratorio di Microelettronica
Strada Prov. Casamassima km. 3 70010
Valenzano - Bari, Italy

Abstract

An extension of the well known defect level model by Williams and Brown has been proposed, to account for non-uniform distribution of fault occurrence probabilities. The field experiment reported by Maxwell and Aitken is interpreted in terms alternative (which may also be considered complementary) to those provided by the model by Agrawal, Seth and Agrawal. Some simulation experiments show that, for circuits implemented in standard cell style, there is a correlation between occurrence probabilities and testability values of SSA and BRI faults.

1 Introduction.

The quality level Q of a production lot of digital CMOS IC's is usually assessed by means of evaluation of the "defect level" (DL), defined as the percentage of bad chips which pass all phases of production testing. In practice, DL is expressed as the number of defective parts per million (ppm) of those which pass as good. The current trend, for high quality productions, is in the range of a few hundred ppm. Hence Q is defined as the complement of DL: $Q = 1 - DL$.

Many attempts have been made to accurately model the quality level of a production lot, most of which are reported and compared in [1].

Besides the early work of Wadsack [2], which has gained renewed interest with the advent of recent technologies [1], one of the most popular models is that by Williams and Brown (WB) [3], in which the defect level is also predicted in terms of product yield and fault coverage. Although the original work concerned single stuck-at (SSA) faults, there is no particular need to confine the model’s validity to these faults. For example, extension to bridging (BRI) faults has actually been done [1, 4], the only requirement being to preserve consistency between the model variables. This means that if a functional test is employed which provides some amount of coverage of SSA and BRI, then the yield must be taken as the functional yield referred to the same faults, and DL must be expressed in terms of the percentage of parts which are defective due to the presence of any of the modelled faults.

It has been observed [5] that this popular model has often been misused, by referring, for example, to the total yield (both functional and parametric), while expressing the fault coverage in terms of SSA faults only, which leads to too conservative estimates of DL.

The model which is presently most referred to in the literature is the one developed by Agrawal, Seth and Agrawal (ASA) [6], which can be considered an extension of Wadsack’s. It includes the possibility that a single defect may produce more than one fault, and introduces a new parameter which represents the average number of faults per faulty chip. With this parameter, it is possible to provide reasonable fitting to field experiments [5]. Note that the results of field experiments are expressed in terms of a conceptually different quantity: the "field reject rate", defined as the percentage of defective parts which are, for any reason, returned from the field and which therefore may include reliability problems or even incorrect handling and use of the component.

WB and ASA models share some common assumptions:
- no defect clustering is considered;
- faults are considered as independent random events;
- fault occurrences are assumed to be equiprobable events.

All these assumptions are open to question and render evaluation of the quality level somewhat approximative. In particular, fault independence is definitely false when multiple faults originated by a single defect are present. It is clear that the assessment of high values of Q relies essentially on the effectiveness of the test which, however, is prone to a number of difficulties arising...
from the need to introduce simplifying assumptions during the process generating the test stimuli.

In this perspective, all the possible actions at design level tending toward a more effective testing have a great impact on the quality of the final product.

In this work we wish to emphasize the consequence of eliminating the assumption of equiprobable faults in DL models. We suggest that, by associating the occurrence probability to the detection probability of each fault, it is possible to highlight possibly critical situations of faults with high occurrence probabilities and low detectability or, conversely, to disregard low probability faults for which the detectability is not determinant for the quality level. In practice, we shall show that individual fault occurrences and detection are not uncorrelated events as is usually assumed at a first approximation.

Those characteristics can also be exploited at design level by introducing a kind of design for testability which consists in either making more testable faults with high occurrence probabilities or, by acting at layout level, reducing the occurrence probabilities of hard to detect faults [7, 8].

We will refer to the original model by WB. In this context, we will include in the fault model multiple faults, treated as complex events for which an occurrence probability value and fault coverage must be evaluated individually, the only requirement being that the complex event can be handled by the fault simulator. For example, for a short in a metal bus involving more circuit nets, a single fault primitive is used and a single occurrence probability is computed.

Section 2 focuses on the extension of the WB model obtained by introducing fault probabilities. Section 3 is devoted to the definition and evaluation of the weighted fault coverage for non-equiprobable faults, including fault detectability. Section 4 deals with some particular properties of this proposed DL model. In section 5, some numerical results are reported for both simulation experiments and the field experiment by Maxwell and Aitken [5] and finally, in section 6, some general comments and remarks conclude the work.

2. Defect level for non-uniform probability distribution of fault occurrences

Within the framework of the theory of IFA [9], we distinguish physical defects (or simply defects) which originate from imperfections in the manufacturing process, from faults which are the manifestation of these defects at the functional level. More specifically, a defect is an alteration of the desired physical structure of the sandwich of conducting, semiconducting and insulating layers which constitutes an IC [9]. In practice, we may classify defects as follows:

a) defects which do cause no fault;

b) single defects producing single faults;

c) single defects producing multiple faults;

d) multiple defects producing single faults;

e) multiple defects producing multiple faults.

Here we will disregard type a) (non critical) defects and consider only those causing at least one fault (critical defects, or simply defects in this context). Thus, a fault may originate from one or more defects and, following IFA, for each of them a critical area [9, 10] can be defined as that area within which the occurrence of at least one spot defect of suitable size will cause a fault [11].

In principle, it is possible to evaluate a critical area for a fault which originates from each of the four cases: b) to e). In this perspective, we consider a multiple fault as a particular fault occurrence which is not necessarily equivalent to several single faults, but which is individually taken into account as a single fault of complex type. Clearly some of these faults will be equivalent to SSA’s.

We refer to a test for a digital CMOS IC of either functional (voltage) or IDDQ (current) type in which a sequence of v test vectors applied at the primary inputs (PI’s) of the chip is able to detect m out of the n total hypothesised faults, producing a fault coverage Θ=m/n.

Here we distinguish Θ from fault coverage T which regards only SSA faults.

Following the original derivation of the DL model by WB, we define:

A: as the event in which none of the n hypothesized faults is present on the chip;

B: as the event in which none of the m out of n faults is present on the chip.

Then, from the Bayes theorem, the conditional probability is:

$$P(A|B) = P(B|A) \frac{P(A)}{P(B)} = \frac{P(A)}{P(B)}$$

since $P(B|A)$ is clearly 1.

Then DL, defined as the probability of a bad chip passing the test but that would have failed had the fault coverage been higher [1], is:

$$1 - DL = \frac{P(A)}{P(B)}$$

Assuming independence between faults and equal probability of occurrence for each of them, we obtain:

$$1 - DL = \frac{(1-p)^n}{(1-p)^m} = (1-p)^{\frac{m}{n}} = \Theta^{(1-\Theta)} \quad (1)$$

which is the classical WB model.
Removing the restriction on the equiprobability of faults leads to [7]:

\[ 1 - DL = \prod_{i=1}^{m} (1 - p_{ij}) \]

Hence, the probability of occurrence of fault i can be expressed as:

\[ p_i = 1 - Y_i = 1 - e^{-\lambda_i} \]

With this assumption:

\[ \Omega = \sum_{j=1}^{n} \lambda_j \]

\[ Y_i = e^{-\lambda_i} \]

3. Weighted fault coverage for non-equiprobable faults

Let us now derive some alternative and more meaningful expressions of \( \Omega \). We define the failure rate \( \lambda_i \) as the quantity:

\[ \lambda_i = \sum_{j=1}^{n} A_{ij} D_j \]

where \( A_{ij} \) is the critical area for fault i due to defect mechanism j, \( D_j \) is the defect density for defect type j and we assume that h different defect mechanisms can produce the fault i. The failure rates are thus dependent on the chip layout and on the fabrication process.

It has been shown [12] that, under the assumption of Poisson distribution of defects, the elementary subyield \( Y_i \) referred to fault i and defined as the probability of non-occurrence of fault i, is:

\[ Y_i = e^{-\lambda_i} \]

Thus, a high level of quality, assessed by low values of DL (DL \( \equiv 100 \div 200 \) ppm), is related mainly to the degree of testability of those faults with high occurrence probabilities.

\[ DL = 1 - Y^{(1-\Omega)} \]

which is formally identical to (1) but \( \Omega \) now has the meaning of a weighted fault coverage.

Also, we can express:

\[ 1 - DL = \prod_{i=1}^{m} (1 - p_{ij}) \]

which emphasizes that DL depends only on the probability of occurrence of those faults not covered by the \( v \) vectors applied.

Thus, a high level of quality, assessed by low values of DL, is related mainly to the degree of testability of those faults with high occurrence probabilities.

4. Weighted fault coverage for non-equiprobable faults

Let us now derive some alternative and more meaningful expressions of \( \Omega \). We define the failure rate \( \lambda_i \) as the quantity:

\[ \lambda_i = \sum_{j=1}^{n} A_{ij} D_j \]

where \( A_{ij} \) is the critical area for fault i due to defect mechanism j, \( D_j \) is the defect density for defect type j and we assume that h different defect mechanisms can produce the fault i. The failure rates are thus dependent on the chip layout and on the fabrication process.

It has been shown [12] that, under the assumption of Poisson distribution of defects, the elementary subyield \( Y_i \) referred to fault i and defined as the probability of non-occurrence of fault i, is:

\[ Y_i = e^{-\lambda_i} \]

Hence, the probability of occurrence of fault i can be expressed as:

\[ p_i = 1 - Y_i = 1 - e^{-\lambda_i} \]

With this assumption:

\[ \Omega = \sum_{j=1}^{n} \lambda_j \]

\[ Y_i = e^{-\lambda_i} \]

Now, if we denote by \( \bar{\lambda}_\Theta \) the average of the terms in the sum at the numerator, and by \( \bar{\lambda} \) the average of those at the denominator,

\[ \Omega = \frac{n \Theta \bar{\lambda}_\Theta}{n \bar{\lambda}} = \Theta \frac{\bar{\lambda}_\Theta}{\bar{\lambda}} = \Theta \alpha(\Theta) \]

yields the relationship between \( \Omega \) and \( \Theta \). In other words, \( \alpha(\Theta) \) is the ratio of the mean failure rate of the detected faults, at fault coverage \( \Theta \), to the mean failure rate of all the faults.

Let us now define a kind of detectability indicator for each fault i: \( \psi_i^v \).

\( \psi_i^v = 0 \) if fault i is not detected by the \( v \) vectors producing the fault coverage \( \Theta \);

\( \psi_i^v = 1 \) if fault i is detected by the \( v \) vectors producing the fault coverage \( \Theta \).

Then, the weighted fault coverage corresponding to \( \Theta \) is:

\[ \Omega(\Theta) = \frac{\sum_{i=1}^{n} \lambda_i \psi_i^v}{\sum_{i=1}^{n} \lambda_i} \]

It is immediately verified that:

\[ \Theta = \frac{\sum_{i=1}^{n} \psi_i^v}{n} \]

Therefore:

\[ \Omega(\Theta) = \Theta n \frac{\sum_{i=1}^{n} \lambda_i \psi_i^v}{\sum_{i=1}^{n} \lambda_i \sum_{j=1}^{m} \psi_j^v} = \Theta \frac{\sum_{i=1}^{n} \lambda_i \psi_i^v}{\bar{\lambda} \sum_{i=1}^{n} \psi_i^v} \]
or:

\[ \Omega(\Theta) = \Theta \frac{\lambda \psi}{\lambda \Psi} \]

This is an alternative way to highlight that, when \( \Omega \neq \Theta \) some correlation exists between the occurrence probabilities of faults and their detectabilities (i.e. the probabilities of \( \psi_i = 1 \) after the application of \( v \) vectors taken at a random).

Note that, unlike the \( \lambda_i \) which depend on the chip layout and technology, the \( \psi_i \) depend on the testability properties of the circuit implementation.

Moreover, from eq.(3) we have:

\[ \alpha(\Theta) = \frac{1}{\Theta} \sum_{i=1}^{n} \lambda_i / \Lambda \]

where: \( \Lambda = \sum_{i=1}^{n} \lambda_i = -\ln Y \).

It can be inferred that:

\[ 0 < \alpha(\Theta) \leq \frac{\lambda_{\text{max}}}{\lambda} \]

Therefore, an experimental observation of \( \lim_{\Theta \to 0^+} \alpha(\Theta) = \alpha_0 > 1 \) (see section 5) is an indication that the first test vectors applied catch the faults with higher occurrence probabilities. In the best case, when the first test vector detects just the faults with highest probabilities, \( \alpha_0 \) will approximate the ratio between the modal and the mean value of the \( \lambda_i \)'s pdf. That is:

\[ \alpha_0 \equiv \frac{\lambda_{\text{max}}}{\lambda} \]


We may rewrite eq. (2) as:

\[ \text{DL} = 1 - Y^{(1-\alpha(\Theta)\Theta)} \tag{4} \]

Although the functional dependence of \( \alpha(\Theta) \) has not been shown in closed form, we may try to obtain some general indication about the shape of DL(\( \Theta \)).

First, we know that, for equiprobable faults, \( \Omega = \Theta \), hence \( \alpha(\Theta) = \text{const} = 1 \), so reducing to the WB model.

Then, by taking the derivative of DL with respect to \( \Theta \), for \( \Theta \to 0 \), we get:

\[ \lim_{\Theta \to 0^+} \frac{d\text{DL}}{d\Theta} = \alpha_0 Y \ln Y \]

This represents the slope of DL(\( \Theta \)) at \( \Theta = 0 \). We see that for \( \alpha_0 > 1 \), the initial slope of DL is greater than the corresponding one of the WB model, as qualitatively shown in fig.1.

![Fig.1 Initial behaviour of DL(\( \Theta \)) for different values of \( \alpha_0 \)]

In other words, for \( \alpha_0 > 1 \) we are in the condition in which the first test vectors applied to the chip detect a set of faults whose average failure rate \( \lambda_{\Theta} \) is greater than \( \lambda \), by a factor which tends to \( \lambda_{\text{max}} / \lambda \).

We have now an alternative interpretation of the initial slope of DL(\( \Theta \)) with reference to the ASA model in which this slope is directly related to the number of faults per faulty chip.

This can be intuitively explained by the fact that faults having large critical areas are also more likely to be caught by the first test vectors. In fact, a large critical area implies circuit nets which span over a large portion of the chip layout with consequent high likelihood to be excited and observed by a test vector taken at random.

An obvious consequence of having an initial slope greater than that of the WB model is that, for high values of \( \Theta \), the slope becomes smaller that in WB. This is because high probability faults have already been caught and further increase of \( \Theta \) will get only the lowest contributes to \( \Omega \). In this way, the rate of decrease of DL(\( \Theta \)) gets smaller and we can stop the test process for a value of \( \Theta \) substantially smaller than in the WB model to guarantee a given value of DL.

We have already shown that the WB model can be obtained by simply assuming \( \alpha(\Theta) = \text{const} = 1 \). In the same way, we can also derive the ASA model by assuming for \( \alpha(\Theta) \) the following functional dependence:

\[ \alpha(\Theta) = \Theta \ln \left( Y + (1-\Theta)(1-Y) e^{-(n_1-1)\Theta} \right) \]

For \( Y \to 1 \), \( \alpha(\Theta) \to \frac{1}{\Theta} \).
Another interesting feature of this model is that SSA faults are generally a subset of all faulty conditions accounted for by $Θ$ and the value $T=1$ occurs generally before $Θ=1$, therefore, if we report $T=1$ in the graph of $DL(Θ)$, we observe a $DL(T=1)$ which, for $α_0>1$, can be much lower than in WB model.

Further, we may easily express, from eqns (1) and (4), $α(Θ)$ as a function of the DL values assumed in WB and in our model.

$$α(Θ) = \frac{1}{Θ \ln Y} \ln \left[ \frac{1 - DL_{WB}(Θ)}{1 - DL(Θ)} \right]$$

which may be useful to obtain $α(Θ)$ from actual field experiments.

The detectability indicators $ψ_i$, which are functions of fault controllabilities and observabilities (e.g. the detectability of a net $SA-0$ is $D(A/0)=C(A=1)|O(A))$, are computed by simulating the $ν$ test vectors to the circuit under test, much like as in the program STAFAN by Agrawal [13].

After the evaluation of all the above quantities, it is straightforward to obtain:

$$Θ(ν) = \frac{\sum_{i=1}^{n} ψ_i^ν}{n}$$

and, from eq (3) the $DL(ν)$ as:

$$1 - DL = Q = \prod_{i=m+1}^{n} (1 - p_i)$$

$$\ln Q = \ln \prod_{i=m+1}^{n} (1 - p_i) = -\sum_{i=m+1}^{n} λ_i$$

So, by exploiting the detectability indicators:

$$Q = \exp \left( -\sum_{i=m+1}^{n} λ_i \right) = \exp \left( -\sum_{i=1}^{n} λ_i (1 - ψ_i^ν) \right)$$

and:

$$DL(ν) = 1 - \exp \left( -\sum_{i=1}^{n} λ_i (1 - ψ_i^ν) \right)$$

From the above equations it is possible to evaluate, for each set of $ν$ test vectors, both $Θ$ and $DL(Θ)$.

5. Experimental results.

A simulation experiment was set up by employing a set of sample circuits suitably designed in standard cell style. The circuits are not very complex but possess testability characteristics which reproduce some extreme situations encountered in practical circuits. For each circuit, a list of SSA and non-feedback bridging faults [14] together with their occurrence probabilities were extracted by means of the software ACRIT [14]. Then the fault coverage were computed with reference to a set of input vectors generated for the SSA+BRI faults and applied in both functional and IDDQ test modes. In this way, the $DL(Θ)$ curves were obtained in the two cases. Moreover, the field experiment by Maxwell and Aitken [5] was considered and interpreted in terms of the proposed model.

For one of the sample circuits considered, Figs. 2a, 2b and 2c depict respectively the failure rate $λ_i$ and the detectability indicator $ψ_i^ν$ (functional and IDDQ test mode) for each fault extracted by ACRIT.

![Failure rate $λ_i$](image1)

![Detectability indicator (functional)](image2)

![Detectability indicator (IDDQ)](image3)
The characteristics of the DL(Θ) curves computed by using the WB and our model were compared. Fig. 3 shows the DL(Θ) curves for one of these circuits in the case of functional test.

A remarkable difference in the DL values predicted by the two models holds also for relatively high values of coverage Θ. The behaviour of DL as a function of Θ, if the IDDQ testing approach is employed, is also reported in Fig. 4.

Figs. 3 and 4 show that, for the examined circuit, considering the fault occurrences as non-equiprobable events has a great impact on the DL value. Moreover, we observe that the use of the IDDQ testing approach rather than functional one, did not produce remarkable differences unless we refer to non-equiprobable fault hypothesis.

An alternative way of representing these results is to evaluate α(Θ) for various values of the coverage Θ. In Table I these values are computed for our circuit (functional and IDDQ test mode) and also for the experiment performed by Maxwell and Aitken [5] by using eqn. (5) in section 4.

<table>
<thead>
<tr>
<th>α(Θ)</th>
<th>α0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>M&amp;A</td>
<td>3.41</td>
<td>2.88</td>
<td>2.33</td>
<td>1.81</td>
<td>1.79</td>
<td>1.62</td>
<td>1.44</td>
<td>1.28</td>
<td>1.15</td>
<td>1.08</td>
</tr>
<tr>
<td>Func.</td>
<td>1.25</td>
<td>1.14</td>
<td>1.14</td>
<td>1.15</td>
<td>1.14</td>
<td>1.13</td>
<td>1.11</td>
<td>1.09</td>
<td>1.08</td>
<td>1.06</td>
</tr>
<tr>
<td>IDDQ</td>
<td>1.37</td>
<td>1.14</td>
<td>1.22</td>
<td>1.27</td>
<td>1.23</td>
<td>1.21</td>
<td>1.18</td>
<td>1.15</td>
<td>1.12</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Table I

In all cases we observe a value of α0 > 1, which means that the effectiveness of the first test vectors applied to the CUT is generally much higher than that foreseen by the WB model.

A possible intuitive interpretation of the results is suggested by the fact that all the considered circuits are implemented in standard cells style. Therefore they are composed by a set of logic gates whose layouts are rather compact and a number of routing channels which allocate the interconnection metal buses which are rather long lines. This produces SSA and bridging faults in the routing channels which have large critical areas. On the other hand, faults within the cells possess much lower values of critical area. Hence, if we distinguish between faults internal to logic gates (Ifaults) and faults between logic gates (Efaults) we observe in general a much higher occurrence probability of the latter respect the former [8]. Table II refers.

<table>
<thead>
<tr>
<th>fault type</th>
<th>percentage of total #of faults [%]</th>
<th>weighted coverage [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efaults</td>
<td>SSA</td>
<td>26.60</td>
</tr>
<tr>
<td></td>
<td>BRI</td>
<td>39.90</td>
</tr>
<tr>
<td>Ifaults</td>
<td>SSA</td>
<td>8.37</td>
</tr>
<tr>
<td></td>
<td>BRI</td>
<td>2.96</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td>22.17</td>
</tr>
</tbody>
</table>

Table II

Moreover the detectabilities of the internal faults are generally lower than those of the external faults which are more likely to be excited and observed directly from PI’s and PO’s. In summary, we got just what we experimentally found: faults with high occurrence
probabilities are also more easily detected than low probability ones. This produces the observed DL values which are generally lower than those obtained when fault equiprobability is assumed.

6. Conclusions.

An extension of the well known defect level model by Williams and Brown has been proposed to account for non-uniform distribution of fault occurrence probabilities.

An interpretation of the field experiment reported by Maxwell and Aitken has been given in terms which are alternative (and may be also considered complementary) to those provided by the model by Agrawal, Seth and Agrawal.

Some simulation experiments show that, for circuit implemented in standard cell style, there is a direct correlation between high occurrence probabilities and high testability values of SSA and BRI faults. This produces an initial slope of \( DL(\Theta) \) more pronounced than in the WB model.

For the sample circuits of our experiment, IDDQ testing provides better weighted fault coverage and hence lower values of DL for relatively low values of \( \Theta \). But at high values of \( \Theta (\Theta>0.95) \), this is not verified any more, and it is rather the functional testing which provides higher values of \( \Omega \) and hence lower values of DL.

As regards to random testing, it has been found that vectors generated for the SSA faults generally behave somewhat better than random ones in the detection of BRI at high values of \( \Theta \).

Further, from the above results, we may draw some final remarks for what concerns the DL model more specifically.

First of all, modelling the DL at high levels of accuracy is a very hard problem. In fact, it is not easy to foresee what will be the effective fault coverage with respect to all the realistic non-modelled faults, provided by the assumed fault model, even in the case of the most refined fault models which include, for example, also delay faults.

Moreover, usually the assumption of independent faults may result rather unrealistic in practical cases, especially when multiple faults are to be accounted for as equivalent to more simple faults. On the other hand, when very high quality levels (DL<200 ppm) are to be guaranteed, the fault coverage of all modelled faults must be well besides the SSA 100% coverage.

The main conclusion of this contribute is that accounting for a non-uniform probability distribution of faults on the chip has the effect of providing less conservative estimates of DL, particularly in the case of MOS IC’s manufactured in standard cell style.

Last, and perhaps most important aspect to be considered, is that obtaining a product characterised by a very low DL does not depend uniquely on the testing strategy adopted, but it is at design stage that the chip quality is first assessed. One possible action at layout level which may be undertaken, in fact, is to reduce the critical areas of faults which exhibit poor testabilities with the selected testing strategy.

Bibliography.


