Advancements in Diagnosis Driven Yield Analysis (DDYA)

A survey of state-of-the-art scan diagnosis and yield analysis technologies

Yu Huang, Wu Yang and Wu-Tung Cheng
Mentor Graphics Corporation, 8005 S.W. Boeckman Road, Wilsonville, OR 97070, USA

1. Introduction

For advanced technology, the current industry is seeing very complicated silicon defect types and defect distribution. It is very common for the yield of a new design to start at a very low level and it’s critical to ramp up yield as quickly as possible to meet the market window. During manufacturing stage for volume production, yield excursions do occur and the yield needs to be improved to a stabilized level. It is also of a great interest for people to increase the stabilized manufacturing yield as high as possible so that they can improve the profit margin. In all these scenarios, silicon defect diagnosis and statistical analysis can be of great help to identify the root cause of the defects, which is critical for the success of yield improvements.

One of the major challenges in the yield analysis process is to identify the systematic issue, and find its root cause. Furthermore, it is very important to find the right die candidate for physical failure analysis (PFA) and provide enough information about the targeted suspects to reduce the time consuming and expensive PFA efforts. Conventionally, the fault isolation technique such as bitmapping [1] for SRAMs has been effective to drive down the defect count of a manufacturing line, thanks to the critical dimensions in high-density SRAM cells and its regular structures. For advanced technologies, being able to manufacture irregular logic standard cells and largely random back-end-of-line (BEOL) metal layers has become very challenging. A popular methodology called diagnosis driven yield analysis (DDYA) is illustrated in Figure 1. It consists of two major parts. One is that using various diagnostic technologies can provide fast and accurate volume diagnosis data; and the other part is that applying the subsequent statistical analysis on the volume data can quickly identify systematic defects and root causes. This gives failure analysis engineers and yield engineers a very fast and highly effective way of defect localization and identification, complementing their traditional, hardware-based methods.

We will survey the recent advancements in DDYA in this paper, which is organized as follows. In Section 2, we will first review the advancement of diagnostic technologies based on various fault models. It also covers related technologies to improve the diagnosis accuracy and performance. Our focus here is to classify the diagnostic technologies rather than giving algorithmic details of each paper. A comprehensive list of references is given at the end of this paper for your own further study. In Section 3, we will discuss statistical analysis technologies based on volume diagnosis results. Various technologies such as zonal analysis, Root Cause Deconvolution (RCD), and DFM related analysis will be addressed. Correspondent case studies will be given. Finally we conclude the paper in Section 4.
2. State-of-the-Art Scan Diagnosis

Different defects can be diagnosed with different fault models and different diagnosis algorithms. The diagnosis technologies can be roughly classified based on the fault models.

2.1 Scan chain faults diagnosis

Scan-based testing has proven to be a cost-effective method for achieving good test coverage in digital circuits. The Achilles heel in the application of scan-based testing is the integrity of the scan chains. The amount of die area consumed by scan elements, chain connections, and control circuitry varies with different designs. Scan elements and clocking can occupy nearly 30% of a chip’s area. The percentage of scan chain defects also varies with different designs. In some cases, chain failures account for almost 50% of chip failures and some systematic defects can be learned from the scan chain diagnosis. Therefore, scan chain failure diagnosis is important in yield improvement. [2] - [10] are major references for scan chain diagnosis. In [2], a comprehensive survey was given for this area. Different chain diagnosis techniques can be classified into different categories, as illustrated in Table 1. Tester-based diagnosis techniques use a tester to control scan chain shift operations, and PFA equipment to observe defective responses at different locations and identify failing scan cells. Hardware-based methods use special scan chain and scan cell designs to facilitate diagnosis. These techniques are effective in isolating scan chain defects but with extra hardware cost. Software-based techniques use algorithmic diagnosis to identify failing scan cells. Compared with hardware-based methods, software-based techniques are more widely applied in industry for general designs, because no design modification is required. Software-based chain diagnosis can be further classified as simulation-based, probability-based and dictionary-based, as illustrated in Table 1.

<table>
<thead>
<tr>
<th>Tester-based</th>
<th>Hardware-based</th>
<th>Simulation-based</th>
<th>Probability-based</th>
<th>Dictionary-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3]</td>
<td>[4]</td>
<td>[5], [6], [7]</td>
<td>[8],[9],[10]</td>
<td>[11]</td>
</tr>
</tbody>
</table>

2.2 Logic faults diagnosis and layout aware logic diagnosis

The logic diagnosis algorithms can be divided into two major categories. The first category applies the cause-effect principle [12], and the second category traces the effect-cause dependencies [13] [14]. The methods in the first category build the simulation response database for the modeled faults and compare this database with the observed failure responses to determine the possible cause of the observed failures. These methods are sometimes referred to as the fault dictionary methods. The effect-cause-based algorithms analyze the actual responses and determine which fault(s) might have caused the observed failure effects. An effect-cause algorithm typically consists of three steps. The first step selects initial fault suspect list by using, for example, the path-tracing algorithm [15]. The purpose of this step is to quickly narrow down the fault suspects by tracing backward from the failure scan cells and primary outputs (POs). In the second step, each suspect fault is injected into the circuit and simulated forward to determine if it matches the responses observed from scan cells and POs. In the third step, a ranking mechanism is applied to identify faults that are more likely to happen. The popular publications for logic diagnosis are [16] - [21].

While logic diagnosis works with a logic-level abstraction of the design, the use of additional design layout information during diagnosis can lead to much better localization of defects and reduce the searching area. Layout-aware diagnosis provides not only the locations such as physical bridges and open net segments, but also their associated properties as cell area, critical area, bridge types and via macro names. The physical call-outs provide valuable classifications for statistical learning and directly results in more die becoming suitable for PFA [22] - [24].

2.3 Diagnosis with embedded compactors

A large number of different compression techniques have been proposed in the literature and have been used in the real designs. Typically, a hardware block called compactor is placed along the scan path on the output side. This block transforms the data captured in a large number of internal scan chains into a data stream that is delivered on a few scan output channels for comparison with the expected values by the ATE. To diagnose failures after compaction, one straightforward approach is to completely bypass the on-chip compression hardware. Doing so provides direct access to the scan cells enabling the application of the well-established standard fault diagnosis. However, one of the major drawbacks of such an approach is the requirement for two separate test sets – a compressed test set for volume production testing purposes and an uncompressed test set for fault diagnosis purposes. In addition, such an approach does not facilitate on-line diagnosis, which is fast diagnosis based on the volume production test results. Thus, it is desirable to
enhance the fault diagnosis techniques to be applicable in the context of compressed test sets.

Another class of diagnostic methodologies through compactor is called indirect diagnosis, which performs diagnosis for compactor-based designs through two phases. In the first phase, the scan cells which should observe failures are identified mathematically from the compactor outputs collected on ATE. In the second phase, with information about which scan cells observe failures from phase 1, any diagnostic algorithm that was originally target at circuits without compactors could be applied. Examples of indirect diagnosis are proposed in [25] for X-Compactor and in [26] for Convolutional Compactor.

The third class of methodology diagnosis methodology is called compactor independent direct diagnosis [27] [28]. Direct diagnosis based on technologies such as circuit transformation, pseudo-scan chains and pseudo-scan cells can be applied for performing fault diagnosis in the context of any compression schemes with any diagnosis algorithms. The proposed methodology enables seamless reuse of the existing standard scan-based diagnosis infrastructure with compressed test data.

2.4 Diagnosis based on other fault types

In addition to the chain and logic stuck-at fault models mentioned above, transition fault diagnosis [29] [30], logic hold-time diagnosis [31] [32], Iddq failure diagnosis [33] [34], compound defect diagnosis [35] [36] [37], scan enable and clock signal diagnosis [38] [39] are also discussed in prior publications. These techniques are useful for handling specific defect types.

Recently cell-aware diagnosis becomes popular. [40] [41]. A general flow would start with the layout extraction step to identify realistic possible defect locations based on physical parameters (resistors and capacitors, for example) in the extracted netlist. This step is followed by an analog fault simulation, and a synthesis step to create the cell-aware library models. The cell-aware library models will be used for cell-aware diagnosis to identify the physical defects within the cells even by using regular stuck-at or other types of production ATPG patterns. If these cell-aware fault models are used by ATPG [42] tools to generate high quality cell-aware test pattern, it can significantly increase the chance catching real silicon bugs and manufacturing defects. Using these cell-aware test patterns for cell-aware diagnosis can help improve diagnosis resolution. At 20nm and higher technologies, PFA is typically done with a dual-beam focused ion beam (FIB) and a scanning electron microscope (SEM). In this process, the image basically bounces back at you, allowing you to incrementally cut and inspect. You can inspect one layer, cut to the next, inspect, etc., until you see the problem. If in diagnosis get resolution down to approximately 1 cell, or about 2 micron (2um), you are in good shape. In contrast, at 14nm, Transmission Electron Microscopes basically require you to access/prepare on both sides. It means that you only have a single chance to inspect the failing die. Therefore, you need significantly better diagnosis resolution. Cell-aware diagnosis is an important technology in this background.

2.5 Diagnosis performance and accuracy enhancement

The continuously increasing size of modern designs poses two primary challenges to achieving high diagnosis throughput with traditional scan diagnosis methods. The first issue is that the time for diagnosing a single failing dies keeps increasing for larger designs. The second issue is the extremely high computational memory required for processing today’s monster designs with billions of transistors.

One solution is to improve volume diagnosis throughput by using a server to manage parallel processors on one or more workstations, as illustrated in Figure 2 [43]. The diagnosis server can automate and distribute diagnosis jobs to multiple processors called Analyzer. It can automatically monitor failure datalogs fed into the server, track the diagnosis history and output the diagnosis results based on the failure datalog directory structure and create database for yield learning later.

![Figure 2: Server-based distributive diagnosis](image)

On a single diagnosis process, many works have been published on improving the performance for diagnosis algorithm using various techniques, such as pattern sampling [44], fault dictionary 0-0, machine learning 0 and GPU-based simulation 0. The high memory requirement for diagnosing very large designs is addressed by static partition [51] or dynamic partition based diagnosis [52] [53].
For core-based SoC testing, ATPG at core-level and hierarchical pattern retargeting at the SoC level can significantly reduce the memory footprint and CPU time compared to running ATPG at the SoC level. For the same reason, hierarchical diagnosis can reverse map failure data logs at the SoC level to the core, and run diagnosis at core-level, which also significantly reduces the memory footprint and CPU time [54]. Note that for identical core instances, the scan stimuli broadcast to all instances of the core, which makes the diagnosis a little bit tricky. In [55] and [56], different on-chip embedded comparators are proposed to facilitate the diagnosis for identical core instances.

To enhance the diagnostic accuracy, several diagnostic ATPG methods were proposed in the past. These algorithms can be classified as (A) for chain diagnosis [57] - [61]; (B) for logic diagnosis [62] - [66]; and (C) for compound defect diagnosis [67].

3. Statistical Yield Analysis and Case Studies

The various scan-based diagnosis reviewed in the previous section can provide valuable volume diagnosis data such as defect classifications, localization and attributes. Statistical analysis based on the data gives people different and actionable methods to identify the root causes of yield loss.

3.1 Best industrial practice in DDYA

In [68], some industrial experiences are shared on how to set up the DDYA flow for scan based logic designs and what are the best practice to lead to the desirable results. The recommended best practices include:

(A) To catch more physical defects, in addition to traditional stuck-at patterns, high quality patterns are required to be generated by technologies such as timing-aware ATPG, power-aware ATPG, layout-aware ATPG and cell-aware ATPG.

(B) In order to have good diagnosis resolution and accuracy, the minimum requirement for collecting failures is as follows:

- (B.1) For chain diagnosis: to collect all chain failure patterns and at least 32 failed scan patterns.
- (B.2) For logic diagnosis: 256 failing cycles can provide a good diagnosis resolution.
- (B.3) In the early yield learning and excursions, it’s recommended to collect 100-200 failing die.
- (B.4) During the yield ramp stage, 500-1000 failing die need to be collected.
- (B.5) For finding subtle systematics during the mature yield stage, 1000-2000 of failing device data logs are needed.

3.2 Zonal Analysis

Out of many different ways for yield analysis is to do zonal analysis based on diagnosis data. Figure 3 shows one of the typical yield signature identification and root-cause analysis methods in DDYA. The initial wafer map shows a random defect distribution. The volume layout-aware diagnosis discloses some die having bridge suspects on different metal layers, from M1 to M5.

We divide the wafer map into three radial zones (center, middle, and outer). If the defects are randomly distributed, the distributions of the bridge defect pareto should be consistent across the radial zones for different layers. If bridge M2 has higher occurrences in the outer zone than others, the random distribution trend is broken and a systematic yield signature bridge based on M2 at the outer zone is identified. Further analysis of suspect types for the dice in outer zone shows the side-by-side M2 bridging defect is the root cause. However, either signature or root cause was hidden from the initial wafer map.

PFA candidates are picked for a few dice with high diagnosis resolution and best representing the yield signature and root cause, i.e. the side-by-side M2 bridge. PFA in this flow is performed only on a few die to validate the finding. After this bridge yield signature analysis, you can filter out the related die and continue analyzing the remaining die. This can be a recursive process until all the significant signatures are found and root causes are analyzed.

![Figure 3: An example of zonal analysis](image)
Case Study 1: Volume diagnosis yield analysis for ramp-up based on zonal analysis.

A product with double digits of yield loss was analyzed. The issue had been chased for several months using traditional method without a success. For DDYA flow, around 1150 failure files were data logged from 4 lots (49 wafers). Each failure file is related to one failed die. The layout-aware diagnosis was run and the diagnosis results were loaded into the yield analysis system. The full statistical analysis flow is illustrated in Figure 4.

Once all the diagnosis reports were selected into a population, zonal analysis was performed on all die for all signature types. A platform needs to have commonly used signatures. For example “bridge layout: count of die” is to analyze bridge suspects based on layers for different zonal types. It’s also important to have custom signatures so that users can have their own signature focus. The zonal analysis (Figure 4 step1) showed a R (radio zone) type yield sensitive signature in the suspect region. The suspect region analysis (Figure 4 step2) showed a red hot spot. The related dice associated with the hot spot were filtered out and it disclosed a dominating cell type of failure (Figure 4 step3). We checked both the suspect location Pareto (Figure 4 step4) and suspect cell Pareto (Figure 4 step5), and both indicated this signature is related to cell_type1. Once a cell signature is identified as the root cause of the problem, we selected 28 die with signature of cell_type1 and the next step is to pick the die for PFA. Here are the criteria (Figure 4 step 6):

- Make sure to select a die containing the yield signature (in this case cell) we want to explore
- Select the die with the minimum number of suspects and make sure the selected die does contain the signature cell related suspects.
- From the selected die, select the suspects with highest diagnosis score.
- If there are more die to choose from, select the suspects with smallest enclosing circle, and
- PFA for the selected die.

Based on above procedure and existing silicon material, there were two dice selected for PFA focusing on location in step4 and cell_type1 in step5. The PFA found both dice had poly shorts (Figure 4 step7) and the process tuning returned the yield to the baseline. The whole process took about a couple weeks to drive the yield to the baseline and step1 to step6 only took about hours.
Case Study 2: DDYA for mature yield improvement [69] [70]

A high volume design was in final production and had a high percentage mature yield from wafer scan test. The goal was to improve the yield by even a fraction of one percent, which would translate to big cost savings. The focus was on the scan test failure since it contributed to a significant amount of yield loss. The initial analysis with traditional yield methods showed that the current mature yield loss was caused by random (baseline) defects, and any further analysis proved to be very time consuming and expensive.

To perform DDYA experiment, 1300 failure files were collected out of several lots. Each failure file represented a scan test failing die. Some failed on the scan chains and others on logic only. Each logic-only-failure file had up to 256 failing cycles and more failing cycles were collected for scan chain diagnosis. When the volume diagnosis was done, the results were loaded into the yield analysis platform. Initial wafer maps illustrated a random failure distribution.

The diagnosis resolution based on single symptom and single suspect was checked as good since it was better than the 50% and 10% threshold respectively seen across industries. The first focus was on 500 or so dice failed chain test only. The chain failure wafer map (Figure 5-2) showed two areas of interest but area1 didn’t have materials so that area2 was analyzed further. It showed that middle band dice of Y zonal type had a higher chain defect distribution. These dice were filtered into the analysis console to perform chain failure related analysis and a signature was identified (Figure 5-2). From its top feature Pareto chart (Figure 5-3), drill down to the center of the wafer map helped pick two dice (Figure 5-5 red spot) and were labeled as die1 and die2.

After completion with the chain failure yield analysis, we selected the remaining logic-only-failure dice (around 700) and analyze all signatures across all the zonal types. When the analysis was complete, the possible yield signatures were reflected in the console area (Figure 6-1) with different colors. In this case, the “Suspect Region: Count of Die” was flagged as a higher possible yield signature in zonal type Y. It implied that defect locations were sensitive to a specific layout region (Figure 6-2).

There were 30 dice associated with this suspect region. The majority of them were due to open layers (including vias) ranged from M1 to M5 based on open layer signature. The wafer map of these 30 dice showed that the top Y region Figure 6-3 was highly correlated with the hotspot in Figure 6-2. Die picking criteria was used here again and 3 dice (die3, die4 and die5) were picked as PFA candidates. The layout view of the die with open suspect is shown in Figure 6-4.

The failing information (location along with failing type) was given to the foundry for PFA and there are 3 PFA hits out of 5 die total Figure 7 and the remaining two die were not performed further due to yield improvement. The foundry made a corresponding process correction based on the three PFA results, and as a result, the mature yield was improved by 3 times of the goal. All this analysis and process correction was done in a few weeks, a result which traditional yield analysis was unable to achieve.
3.3 ROOT CAUSE DECONVOLUTION (RCD)

Root Cause Deconvolution (RCD) is a statistical enhancement technology towards the next step in diagnosis resolution enhancement. It works by analyzing a reasonable amount (usually one thousand to get started) of layout-aware diagnosis reports together to identify the underlying defect distribution (root cause distribution) that is most likely to explain this set of diagnosis results. While layout-aware diagnosis points to a suspect across multiple physical layers, RCD can isolate a particular root cause in that suspect. This increase in the failure analysis relevance and success rate dramatically reduces the PFA cycle time from months to days. RCD also enables “virtual FA” [71], the ability to determine defect distribution for a population of failing devices before any failure analysis is performed.

Case Study 3: Use RCD to enhance diagnosis resolution

As illustrated in the flow shown in Figure 8, layout-aware diagnosis is performed on a set of die that failed manufacturing test (1). Each diagnosis result contains a set of root causes that are potential explanations for the failure (2). RCD processes the diagnosis reports, eliminates the noise, and identifies the underlying distribution of root causes that best explain the set of diagnosis results. In this example, RCD determines that a distribution containing three root causes best explain these diagnosis reports (3). From this distribution, the user can focus on the root cause of interest. This may be the most significant root cause, or one that is deemed interesting for other reasons, for instance one that has not been seen before and/or one may have zonal sensitivity. For each suspect in each diagnosis report, RCD calculates and assigns a probability for the suspect being explained by each of the root causes in the RCD distribution. This means that the user can easily identify the die that has the highest probability of representing a particular root cause, and use that as guidance to select die for FA (4).
analysis to identify a critical ‘T’ feature in product that was caused by a combination of OPC and lithographic effects. Correlating features extracted from a design to test results of the associated silicon enables a method to rapidly identify systematic critical features and often with minimal reliance on subject matter experts. In [77], a novel flow to correlate diagnosis and DFM is proposed, which include the following steps: (A) Locating a defect’s location in silicon based on layout aware scan diagnosis; (B) Locating a critical feature; (C) Zonal analysis; (D) Quantifying the fallout.

Other researches are done to correlate the layout diagnosis results with extracted features and then use RCD to give user a defect or root cause distribution pareto based on hybrid features and conventional signatures such as cell area, cell type and via. RCD is also used to pick the high confidence die and suspect correspondent to a root cause for PFA.

Case Study 4: Critical feature identification [77]

A failing 28nm device submitted to PFA was found to contain a Metal-2 open defect that occurred on a signal line sandwiched between two other lines at minimum spacing. It was hypothesized that one or more of the minimum spaced neighbors could be inducing the open defect. In order to validate these hypotheses the yield engineers could have submitted several devices diagnosed with Metal-2 open defects to FA and see if they all had a similar topology. Alternatively, it is possible to leverage the ~5,000 diagnosis results to uncover systematic DFM hits.

Eight hypotheses were generated based on length of the sandwich between 2 neighbors or length of projection along 1 neighbor (Figure 9). The eight hypotheses were generated by sorting the sandwich length from lowest to highest and assigning the first third of the violations to a category name “small sandwich”, the middle third to a category named “medium sandwich”, and the longest third to a category named “large sandwich”. Additionally there is a category for “length independent” and all four categories are to be considered as rules. All four rules are repeated for only a single minimum space neighbor. Then the eight rules were individually overlaid with the diagnosis results to come up with the DFM Hits.

In order to validate the hypothesis that the sandwich rule is systematic, the expectation is that the rule would be an outlier in some zone. In this case wafers from four different lots had been diagnosed and overlaid with DFM, but only one of the lots had a shift in the defect distribution toward Metal-2 opens Therefore the ‘zones’ in this case each represent data from different lots. The results of intersecting the diagnosed defect locations of these ~5,000 die with the 8 hypotheses were analyzed. From test data alone, the Large Sandwich with two neighbors is identified as the best description of the true root cause for the open defect. This conclusion has been corroborated by further PFA.

![Figure 9: Critical feature rules](image)

4. Conclusions

In this paper, we surveyed the recent advancements in DDYA, which includes scan-based diagnosis technologies and diagnosis driven yield analysis. Multiple industrial cases studies are given to illustrate the values of the DDYA flow. By using the advanced DDYA, diagnosis can be more accurate, fast and informative. More importantly it can help improve yield by finding the systematic defect, identifying the root causes, correlating diagnosis results with DFM and picking highly possible die and suspect for PFA to validate all the findings

References:


