Predictive Yield Modeling for Reconfigurable Memory Circuits

Dennis J. Ciplickas, Xiaolei Li and Rakesh Vallishayee
PDF Solutions, Inc.
San Jose, CA 95110

Andrzej Strojwas
Elec. and Comp. Eng. Dept.
Carnegie Mellon University
Pittsburgh, PA 15213

Randy Williams and Michael Renfro
Intel Corp.
Albuquerque, NM, 87124

Raman Nurani
KLA-Tencor Corp.
San Jose, CA 95161

Abstract—

This paper presents a novel approach to the modeling of defect related yield losses in reconfigurable memory circuits. The proposed approach is based on the critical area extracted from the memory layout and the in-line defect inspection data. A complete chip level yield model that takes into account the actual redundancy scheme, is presented with the demonstration of excellent accuracy between the model prediction and bitmap data from an actual Flash memory product manufactured by Intel Corporation.

I. INTRODUCTION

With today's complex VLSI designs and manufacturing processes, predictive yield modeling is indispensable for rapid yield learning. Without predictive yield models based on in-line defect measurements, it can be difficult to disaggregate the root causes of yield loss without extensive, time-consuming failure analysis (FA). Furthermore, even given expensive FA, empirically-based yield models rarely achieve sufficient accuracy to estimate and prioritize the likely yield impact of potential improvement opportunities. This paper presents a predictive yield modeling methodology which circumvents these difficulties and achieves remarkable accuracy through a combination of product layout, in-line measurement and micro-level & macro-level yield loss modeling.

To demonstrate the full capability of the proposed methodology, reconfigurable Flash memory devices were chosen as the vehicle for this research. These devices often allow highly flexible repair programs and have equally complex repair constraints. Furthermore, unlike logic devices, memory arrays provide high resolution signatures of failure events simply from bitmap probe testing. Detailed analyses of block, row, column, bit and other failure signatures in these bitmaps have been used extensively in the past for yield improvement. These data-driven efforts, however, have enjoyed comparatively little support and utilization from predictive yield models. In fact, “macro” yield predictions using the critical area of the entire chips or design blocks are usually too coarse to provide sufficient insight into the physical mechanisms driving particular failure signatures. Consequently, significant quantities of in-line inspection and end-of-line test data have historically been necessary to empirically model the relationships between failure event signatures and physical failure mechanisms.

Once obtained, however, relationships between physical failure mechanisms and functional product failures are extremely valuable for such tasks as defect budgeting, evaluation of in-line inspection efficiency, and evaluation or optimization of the redundancy scheme used to reconfigure the memory circuit. Thus, a clear opportunity exists for a yield modeling methodology which allows fast creation of predictive models for failure event signatures and immediate correspondence between end-of-line test results and physical failure mechanisms. The research presented in this paper targets this primary goal.

This paper is divided into four sections: 1) micro-yield loss event modeling, 2) chip-level redundant yield modeling, 3) in-line data modeling, and 4) end-of-line model validation.

Section II opens the discussion with a detailed description of the micro-yield event extraction. Within the context of this paper, “micro-yield events” are analogous to bitmap failure signatures, although they are much more precisely defined than the figures typically generated by bitmap analysis software. Section II presents an efficient extraction technique for the determination of micro-yield event critical areas based on a standard, geometrical oversizing method used to extract traditional critical areas.

Section III describes the process in which micro-yield events, and other yield events (such as chip periphery failures) are combined to form a chip-level redundant yield prediction model.

Section IV outlines the steps required to transform raw in-line defect measurements into analytical defect size distributions. As is evident from the results shown in Section V, careful analysis of the measured defect data can compensate for insufficient capture rates and sensitivity of the inspection equipment for smallest defects. Such issues are unavoidable in current inspection technologies and will become even more important as critical dimensions continue to shrink.

Section V outlines the results of applying these modeling methods to a state-of-the-art Flash memory product at Intel Corporation. The predictive nature of our yield model is validated by comparing predictions based solely on in-line inspection data and product layout with measured yields extracted from the end-of-line sort data. Both virgin and redundant yields are analyzed.

Finally, Section VI summarizes the conclusions reached as a result of this research and suggests promising applications based upon the results of this study.

II. MICRO-YIELD EVENTS

"Micro-yield loss events" are roughly equivalent to the "failure event signatures" observed in bitmap test measurements. Examples of micro-yield loss events include: single bit, bit pair and bit quad failures; two, three or more adjacent row failures; two, three or more adjacent column failures; and row+column failures.

Micro-yield loss events, as defined here, are subtly different from measured failure events. Micro-yield loss events have a well-defined root cause; *e.g.*, extra-material causing shorts between distinct electrical nodes. Root causes of measured bitmap failures are often much more difficult to infer, even with special Built-in Self Test (BIST) and data analysis techniques. This distinction is important for two reasons. First, it places strict requirements on the in-line inspection data used for the yield predictions; *e.g.*, only shorts-related defect distributions should be used in the model. Second, it allows the micro-yield predictions to serve as a hypothesis test for potential root causes of observed bitmap failures. Simply put, if a set of micro-yield predictions match the equivalent set of observed failure event probabilities, the modeled and actual root causes are likely to be similar. This level of confidence is difficult to achieve with the more traditional macroscopic yield predictions.

Micro-yield predictions are computed using a modified Poisson model [7]:

$$Y_e = \prod_{\forall l} \exp\left(-\int_{x_0}^{\infty} [CA_{e,l}(x)][DSD_l(x)]dx\right) \quad (1)$$

e = event type (e.g. row pair short)
l = material layer
x_0 = minimum feature size
x = defect size
$CA_{e,l}(r)$ = critical area of event e in layer l
$DSD_l(r)$ = density of defects in layer l with size r

Defect density functions $DSD(r)$ are estimated using in-line defect inspection data as discussed in Section IV. Critical area functions for each event $CA_{e,l}(r)$ are computed using the product layout and a modification of the traditional oversizing algorithm for critical area extraction [5]. During the extraction, each critical area polygon is categorized by the event type corresponding to the set of electrical nodes participating in each unique geometrical overlap. This concept is illustrated in Figure 1. It is important to note that in this algorithm, all event critical areas are disjoint by construction. This permits simple product terms in the chip-level yield models.

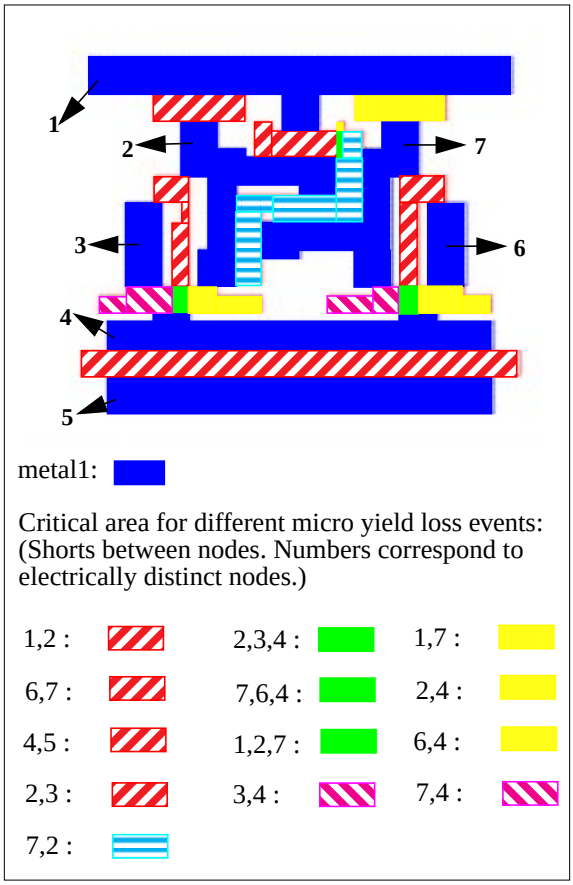

Figure 1. SRAM cell metal1 (Critical Area Categories) (Shown for illustrative purposes)

Previously published methods required polynomial execution time for such detailed critical area extraction and were therefore limited to node-pair interactions within small circuits [1]. In this study, a software tool suite called pdEx [6] was enhanced to perform the critical area categorization "on the fly" during a conventional Boolean AND operation. Since this Boolean AND operation is already required for "macro" critical area extraction, no execution time penalty is incurred for the micro-yield event extraction. These algorithms implemented in the software system are essential in a practical application of the redundant yield modeling techniques presented here.

The style and extent of selected event types is governed by the repair scheme of the memory device. Only those events which are repairable must be modeled individually. All other failure events are lumped as an "unrepairable event."

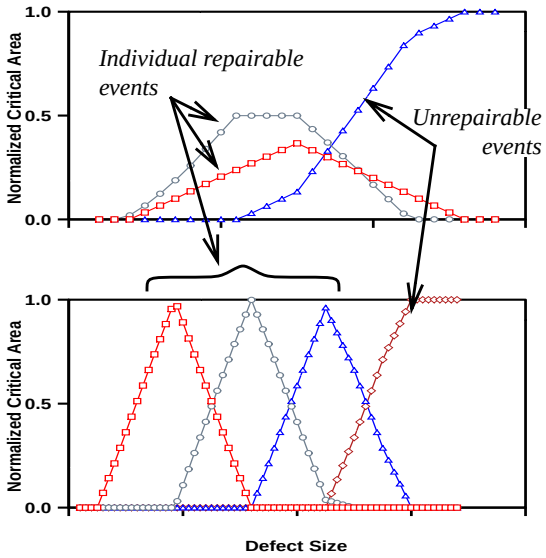

Figure 2. Typical critical areas for micro-yield events.

Critical area curves were extracted for the *poly1*, *poly2*, *metal1* and *metal2* layers from the Flash device in this study. For increased accuracy in predicting coplanar, inter-layer shorts, the *contact* layer and *poly2* layer were combined for the *poly2* extractions [4]. Examples are shown in Figure 2. Finally, micro-yield predictions for each event type (*e.g.*, single bit, double row) were formed by combining individual layer yields according to Equation (1).

III. CHIP-LEVEL REDUNDANT YIELD MODELING

A chip-level yield prediction is achieved by combining a hierarchy of micro-yield loss events. Previously published research in this area has been restricted to straightforward breakdowns of unrepairable and independently repairable sets of blocks (*e.g.*, one of rows, columns or cells) [1]. A Flash device was chosen for this research because of the particularly flexible and complex repair constraints available in this family of memory devices. For example, even though repair options often exist at the block, column and row levels, due to the nature of the Flash CAMs employed to store the "repair program," certain combinations of damaged rows and columns are often unrepairable. Furthermore, the total number of repairable rows and columns are also often limited at the *chip* level rather than the *block* level (as is common in DRAM and SRAM memory devices).

In theory, it is possible to assign event classes to all possible combinations of nodes on an entire chip and generate a complete set of micro-yield events using the algorithm outlined in Section II. These micro-yield events are then systematically combined to model all possible repair scenarios. In practice, however, this method is both infeasible and unnecessary. Depending on the repair options, a natural and sufficiently accurate yield model hierarchy often presents itself. Figure 3 illustrates the general structure of the Flash device analyzed in this study. The corresponding hierarchical yield model is shown below:

$$\begin{aligned} Y_{\text{chip}} &= Y_{\text{periph}}(Y_{\text{plane}})^2 \\ Y_{\text{periph}} &= \prod_{\forall l} \exp\left(-\int_{x_0}^{\infty} CA_{\text{periph},\,l}(x) DSD_l(x)dx\right) \\ Y_{\text{plane}} &= Y_{\text{plane-nr}} Y_{\text{plane-r}} \\ Y_{\text{plane-nr}} &= \prod_{\forall l} \exp\left(-\int_{x_0}^{\infty} CA_{\text{plane-nr},\,l}(x) DSD_l(x)dx\right) \end{aligned} \tag{2}$$

Assuming that there exist redundant blocks within each memory plane, the redundant plane yield can be described as:

$$\begin{aligned} Y_{\text{plane-r}} &= \sum_{i=0}^{N_{br}} \binom{N_b}{i} (Y_{\text{block}})^{N_b - i} (1 - Y_{\text{block}})^{i} \\ Y_{\text{block}} &= \prod_{\text{unrepairable events}} Y_{e_i} \end{aligned} \tag{3}$$

where N_b is the total number of blocks/array plane and N_{br} is the total number of redundant blocks/array plane. There are three notable features of Equation (3).

First, the block-level repair is handled by a conventional combinatoric method [2]. This method is used because all possible block repair scenarios in Flash memories are simple to enumerate.

Second, intra-block repair is simply handled with a product of all unrepairable intra-block micro-yield events. Intra-block repair scenarios would have been infeasible to enumerate with a conventional combinatoric method given the complex row/column repair dependences in the Flash device being analyzed. This fact highlights an essential aspect of the micro-yield

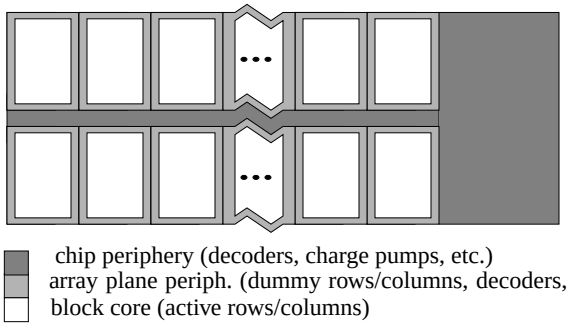

Figure 3. Typical Flash memory device structure

event methodology. Namely, that the combinatoric complexities of conventional redundant yield models are pushed backwards into geometrical critical area extraction. Thus, rather than enumerating the yields of all possible scenarios, a Boolean-AND merges and extracts only the relevant scenarios while conveniently lumping the remaining scenarios.

Third, no cross-level redundancy constraints are shown in Equation (2). For example, a maximum amount of row or column repair per chip places a constraint between the chip and block levels. In general, such constraints must be handled by enumerating all such possible scenarios and including a conditional term in the yield model. For the Flash device analyzed here, the probability of such events was exceedingly small. This was supported by both simple micro-yield predictions and actual manufacturing data. Thus, these terms have been excluded from this presentation for clarity.

Finally, it should be noted that the preceding yield model assumes that only a single defect may occur within the total critical area of each micro-event type. This is a valid assumption given the block sizes of the chip at hand and random defect densities present in a typical manufacturing site. If block sizes were sufficiently large, or defect densities sufficiently high, however, multiple defects per critical area could occur. This is handled by a straightforward re-classification of micro-yield event types and the use of a binomial model in equation (1) rather than the conventional Poisson approximation presently employed.

IV. IN-LINE DATA ANALYSIS

In-line data analysis is a critical component of the predictive yield modeling methodology presented here. This data complements the critical area extractions and, taken together, provides the essential information necessary to cross-validate likely hypotheses for the end-of-line yield loss mechanisms.

Careful filtering and smoothing of in-line data is required to avoid encountering the same capture rate and sensitivity pitfalls suffered by empirically-based defect models (*e.g.* kill ratios [3]). While previous generations of in-line inspection equipment made this task prohibitively difficult, recent advances in defect inspection technology have enabled successful predictions based on in-line inspection data. In particular, the KLA213X equipment allows for extraction of defect size information and has sufficient sensitivity to be used for predictive yield modeling purposes.

In this project, approximately 50 lots of in-line inspection data from KLA 213X and Tencor SS7X00 equipment were collected from two fabs at all critical levels (after key patterning and deposition/via formation steps). An extensive analysis of the inspection recipes was performed to maximize the capture rate of the dominant defects while minimizing the nuisance counts.

For each layer modeled by critical area extractions, a defect size distribution was extracted from in-line KLA 213X data. Histogram-based fitting methods were found to be unreliable and a more robust mean/variance matching method was used instead. A complete description of this fitting method is beyond the scope of this paper.

However, it is well known that the capture rate of the optical inspection tools is low for the smallest defect sizes. Hence, the reported defect density is smaller than the true value. To compensate for this equipment limitation, we have introduced a scaling function f(x) which measures the amount of extrapolation between the modeled and observed defect densities for a full range of defect sizes. Hence, the defect density used in yield model is given by:

$$DSD(x) = D_0 f(x) \frac{k}{x^p}$$

$$D_0 \text{ is the total measured defect density}$$

$$f(x) \text{ is the scaling function} \qquad (4)$$

$$\int_{x_0}^{\infty} \frac{k}{x^p} = 1.0 = \text{size distribution function,}$$

The distribution type is described by the parameter p which is extracted from the measured defect data for regions where the capture rate is sufficiently high. Then, this distribution type is used to extrapolate the defect sizes down to the minimum size of defect that can cause a chip failure.

It was found that using an alternate defect size definition $\sqrt{XY}$ provided more reliable yield predictions than the standard KLA "DSIZE" definition of $\min(X, Y, \sqrt{A})$, where X and Y are the horizontal and vertical dimensions of the bounding box containing an

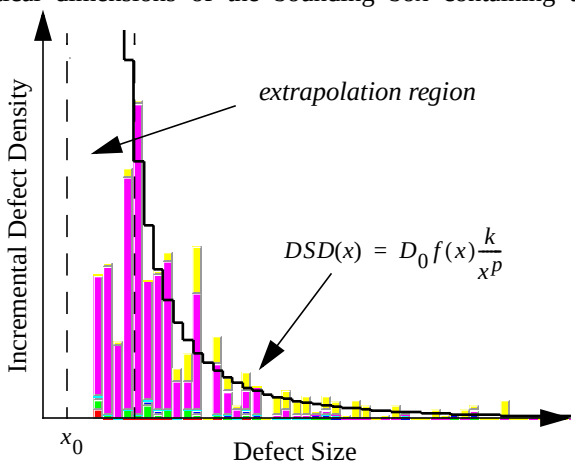

Figure 4. Typical DSD fit to KLA 213X data.

actual defect. This is a natural consequence of the assumptions present in critical area yield modeling. Namely, that all defects are both assumed to be symmetric, and that true defects tend to cause shorts along the maximum dimension rather than the minimum dimension.

A typical example of a fit of the defect size distribution (DSD) is shown in Figure 4. The extrapolated densities are typically much greater than the measured ones. Typical extrapolation factors (ratio of extrapolated to measured defect densities) in this project ranged from 2 to 6, depending on the sensitivity of the inspection recipe and equipment used. More important than the value of the extrapolation factor, however, is the fraction of inferred yield impact. A typical yield impact curve corresponding to the DSD fit in Figure 4 is shown in Figure 5. As is evident in the figure, up to 25% of the predicted yield impact, including the peak yield impact region, may be accumulated in the extrapolated portion of the DSD curve. Furthermore, as shown below, the accuracy of the overall yield prediction indicates that such extrapolations are valid extensions of the measured defect density at larger defect sizes. Observe, however that for process control purposes, the measured defect densities for the range where the capture rate is very high are typically sufficient.

V. MODEL VALIDATION

The micro-yield predictions and methods described above are validated below through comparison with measured bitmap yields. Rather than comparing aggregate chip yields, in which many confounding factors can dilute the validation, the more stringent test of *individual* event yield comparisons was utilized. In this manner, both micro- and macro-verification is achieved.

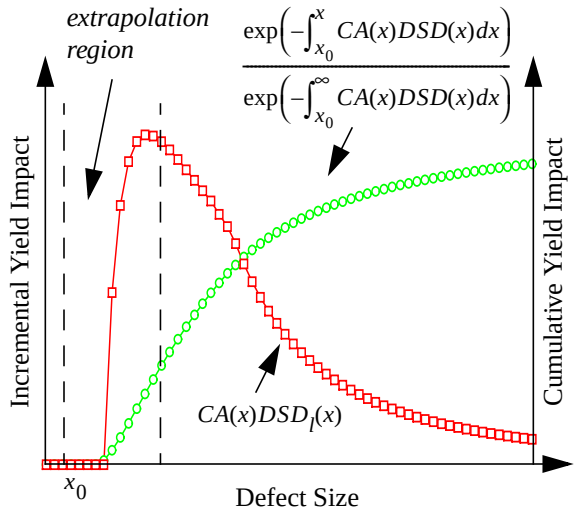

Figure 5. Extrapolated portion of micro-yield impact curves

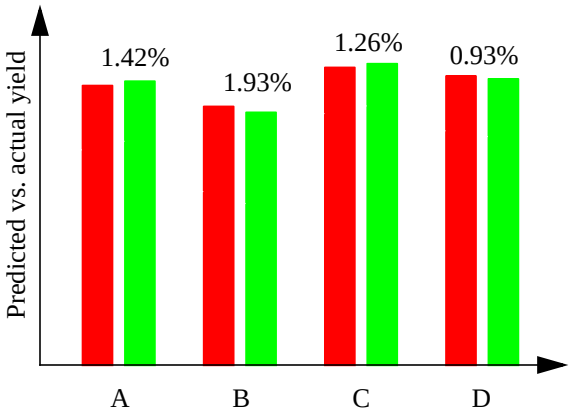

Figure 6. Actual vs. predicted pre-repair yields for the four dominant array failure mechanisms. The numbers above the bars represent the absolute errors in yields.

The wafer-probe testing method used in this project allowed six different array events to be observed. Of these six events, four dominant array yield losses were modeled by the methods presented here. Numerous micro-yield events were required to model each of the four bitmap events (*e.g.*, both *poly1* and *metal1* shorts are required to model bitline failures). The remaining two events concerned inter-layer shorts and neither contributed significantly to actual yield loss.

The four micro-yield events which dominated yield losses were intra-layer shorts at two poly levels and two metal levels. Since their occurrence in the array was observable in the sort test results, we were able to compare directly our model accuracy for both the virgin (raw) and repaired yield results. These comparisons were performed for a development fab (Fab A) and a volume production fan (Fab B). In both cases, the accuracy of the model was extremely good, which is demonstrated in Figure 6 for the pre-repair yield for the four dominant events. Maximum error for event yield predictions was 1.93% in Fab A and 0.8% in Fab B. The accuracy of the overall yield prediction for the memory array (2 planes) was also excellent, namely 0.4% in both Fab A and Fab B.

To verify our redundant yield model accuracy, we have compared the redundant yields for the row and column events (by grouping the appropriate poly and metal level events and taking into account the redundancy scheme in the actual Flash memory product). The results for Fab A are shown in Fig. 7 together with the overall repaired yield prediction error (the accuracy for the latter was 1.95%). For Fab B, similarly good results were obtained and the overall redundant yield prediction error was also less than 1.99%).

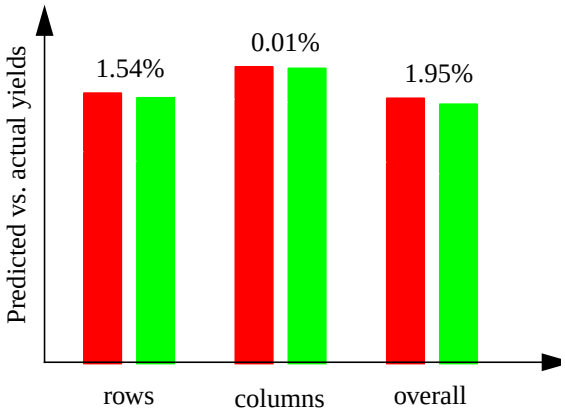

Figure 7. Actual vs. predicted post-repair yields. The numbers above the bars represent the absolute errors in yields.

We also computed the overall defect limited yield for the entire chip including the non-redundant chip periphery and were able to explain a vast majority of the total yield loss and therefore extract the systematic yield component.

VI. CONCLUSIONS

This paper presents a novel approach to modeling the defect related yield losses in reconfigurable memory circuits. Our methodology is based on the accurate modeling of the micro-yield loss events using the concept of critical area, combining these events in a chip-level combinatoric model and providing accurate yield loss prediction based upon the in-line defect inspection data. Extremely good prediction results have been achieved by not only accurate modeling of the yield loss mechanisms, but also by a novel approach to the fitting of the defect density and size distribution to the in-line inspection data, and more adequate modeling of defect sizes available from the KLA-Tencor inspection equipment. This project has demonstrated the potential of linking the in-line defect inspection to yield impact, and therefore has opened many exciting applications for both design and manufacturing. These applications include layout design rule and redundancy optimization, as well as defect budget specifications, inspection plan development and even wafer disposition based upon in-line inspection results.

VII. ACKNOWLEDGMENTS

This paper is a result of a two year joint project between Intel Corporation, PDF Solutions, Carnegie Mellon University and KLA-Tencor Corporation. The authors would like to express their thanks to many individuals who made this project possible.

More specifically, we would like to thank John Kibarian and Kimon Michaels from PDF Solutions for their support and many fruitful discussions. We would also like to acknowledge Thomas Waas, Michael Kramhoeller and Restu Ismail for their help in implementing the redundancy models in pdEx.

We would also like to thank Owen Jungroth from Intel Corporation for making the Flash memory design available and for explaining the redundancy scheme that was utilized.

We are also grateful to Dave Fletcher from KLA-Tencor Corporation for initiating the joint project and Dave Joseph from KLA-Tencor Corporation for his support throughout the project.

VIII. REFERENCES

[1] J. Khare, D. B. I. Feltham and W. Maly, “Accurate Estimation of Defect-Related Yield Loss in Reconfigurable VLSI circuits,” *IEEE Journal of Solid State Circuits*, vol. 28, no. 2, pp. 146-156, February 1993.

[2] T. L. Michalka, R. C. Varshney and J. D. Meindl, “A Discussion of Yield Modeling with Defect Clustering, Circuit Repair, and Circuit Redundancy,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 3, no. 3, pp. 116-127, August 1990.

[3] P. Mullenix, J. Zalnoski and A. J. Kasten, “Limited Yield Estimation for Visual Defect Sources,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 10, no. 1, pp. 17-23, February 1997.

[4] R. K. Nurani, A. J. Strojwas, W. P. Maly, C. Ouyang, W. Shindo, R. Akella, M. G. McIntyre and J. Derrett, “In-Line Yield Prediction Methodologies Using Patterned Wafer Inspection Information,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 11, no. 1, pp. 40-47, February 1998.

[5] C. Ouyang and W. Maly, “Efficient Extraction of Critical Areas in Large VLSI IC’s,” in *Proc. Int. Symp. Semiconductor Manufacturing (ISSM 96)*, Tokyo, Japan, 1996, pp. 301-304.

[6] pdEx Users Manual, PDF Solutions Inc., 1998.

[7] C. H. Stapper, “Modeling of Integrated Circuit Defect Sensitivities,” *IBM J. Res. Develop*, vol. 27, no. 6, pp. 549-557, November 1983.