A Quality-Cost Model of In-line inspections for excursion detection and reduction

Gavin D. R. Hall, Roger Young, Matt Dunne and Mina Muro
Defect & Device Yield Engineering
ON Semiconductor
23400 NE Glisan, Gresham, OR 97030-8411
Email: Gavin.Hall@onsemi.com

ABSTRACT

A methodology for optimized defect excursion (EXR) monitoring is proposed using an economic statistical process control (SPC) model for defect limited yield. The cost of in-line defect inspections is increasing at an exponential rate, particularly for a 300mm fabrication facility. Therefore, optimized in-line inspection schemes have become more critical for controlling the costs of semiconductor manufacturing. In order to minimize the inspection costs while maintaining acceptable yield, a cost function which incorporates the power of the inspection, the interval between inspections, and the yield impact (cost) is optimized for all inspection locations in a given process flow with a fixed sampling budget. This methodology can be used to allocate inspections based upon the risk of yield excursions at defect limited process layers. This model can also be used to establish quantitative estimates of return on investment (ROI) of inspections to inform decisions regarding purchase of in-line monitoring (ILM) tools or sampling adjustment. A quality-cost model has been derived using the theory of economic SPC, and has been implemented in a high volume CMOS fabrication facility with a high degree of success.

[*Keywords*: Advanced SPC techniques, Defect Inspection, Cost Reduction, Cost of Ownership, Yield Enhancement, Defect-to-Yield correlation]

INTRODUCTION

The risk of adding or removing an inspection per unit time at a specific point in the flow depends on the information gained per inspection. If the inspection gives information about an assignable cause of yield loss that has occurred on a tool and the information can be used in a timely corrective action, the inspection can be said to provide "excursion control." If the source of defects is an isolated event, such as certain sources of CMP scratches, which only affects a single wafer or lot and does not persist on further processing, then the inspection only gives information about the defect when the defect is caught. However, some isolated events may be indicative of incipient problems of longer lasting character. Thus, defects must be observed with care. As time progresses the information obtained from in-line data can be used to establish a pareto of defect sources which can be used to determine yield models, monitor the health of the line, and prioritize process engineering efforts based on defect-limited-yield.

The major considerations for an ILM engineer in the sampling allocation problem are: (1) the total number of layers to be inspected and their locations in the flow; (2) the rigor of the inspections; and, (3) the interval between inspections at each layer (measured in time, or equivalently, material) .

The allocation of an inspection point in the flow is dictated by engineering insight into possible failure modes, the amount of time to the end of line sort or acceptance sub-sampling, and other considerations. In this investigation the location of sampling layers will be predetermined and fixed by previous cycles of yield- learning.

The rigor of the inspections is determined by evaluating the errors and the sensitivity in the inspections' recipes which includes the pixel size of a SEM review, the threshold sensitivity, and the hypothesized minimal killer defect size. For example, defects deposited on a metal resist pattern are only of concern if they are large enough to cause a blocked metal etch. However, even if these considerations are met, there may still occur smaller than critical size that cluster to a degree which still creates a metal bridge. There are also sources of risk due to the control level, L, of the defect monitoring chart, and other statistical sampling errors accounted for in terms of Type I (α) and Type II (β) risk. In this investigation, it is assumed that the ILM recipes are optimized to the degree that defective material can be detected with high accuracy and the main source of error is statistical.

The interval between inspections determines the risk due to possible low yield on unmonitored material. This question is answered by addressing the relative risk of yield-limiting defects, and the value of the information gained from standard inspections' baseline defect rate. The relative risk is obtained by using an Automated Defect Classification scheme (ADC) and a yield-to-defect/kill-ratio methodology to calculate the killer defect density (KDD) [1], [2]. Once this is done, different inspection layers can be quantitatively ranked relative to each other and a low-risk sampling allocation can be obtained.

Defect limited yield can be subdivided into categories of yield-loss-risk as shown in Fig 1(a). Low frequency high impact (LFHI) events are referred to as defect excursions (EXR), and high frequency low impact (HFLI) events that are the baseline defectivity. HFHI events are not indicative of a

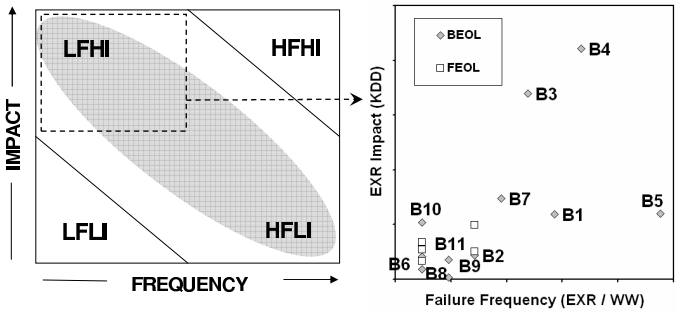

Fig. 1. Impact vs. frequency matrix showing region of concern for ILM sampling of a stable process. HFLI is the baseline defectivity and LFHI is high impact defect excursions. (b) The EXR impact of various FEOL and BEOL layers is plotted against frequency. Note that FEOL layers have low impacts and frequencies, while BEOL layers are scattered at high frequencies and impacts.

stable process, and LFLI events do not justify corrective action are not considered.

Fig 1(b) shows various inspection layers ranked by KDD and EXR frequency. BEOL layers are labeled B1-11 for reference. As an example, one can readily see that for similar frequency, B1 and B3 have very different KDD impacts, and for similar KDD impacts, B1 and B5 have very different frequencies. Even though it is evident which layers pose the highest risk to yield impact, it is not readily apparent from this picture how to allocate inspections in order to minimize the exposure to all these risks at once. In order to answer this question a model is derived which establishes the cost of risk due to exposure to defect-limited yield excursions. This function is then minimized to find the optimal sampling allocation.

Defect Data Methodology

Wafers are sampled at standard inspection points (called "layers") throughout the flow based on technology inspection requirements established in early product development and yield learning. Defect information found by inspection tools is fed into an ADC (auto-defect classification) system which determines the defect type. The defect is then binned by category (i.e. particles, blocked etch, stringers, scratches etc.) called a "fine-bin". This data along with the defect images, size and coordinates are stored in a DDMS (defect data management system). The average kill-ratio is assigned to a fine-bin based upon overlay analysis with respect to end of line wafer-probe data.

Each defect inspection layer in the flow is under SPC control, which is characterized by the control level, L, the interval between inspections, I, the number of wafers/lot per inspection, n, the review fraction of the ADC system, f, and other factors. Defect excursion limits are established for defect count, defect density, characteristics, or spacial patterns. Multiple control levels for these different failure modes act collectively to define an effective SPC control level. The interval between inspections is controlled by a scheduler, in order to maximize the utilization of the inspection tool. When

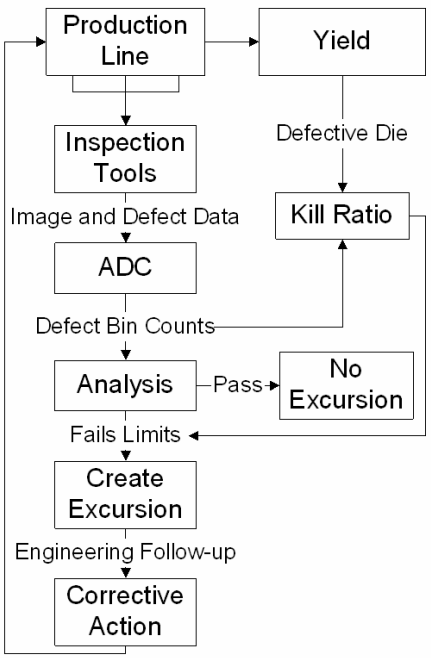

Fig. 2. In-line inspection flow schematic.

a lot violates the control level, immediate follow-up actions are initiated to isolate the source of the defects, contain the problem, determine root-cause, fix the problem, verify that the defectivity level is no longer above established limits, and then return to monitoring the layer for reoccurance. This process is shown schematically in Fig 2.

When a defect excursion occurs, the event is recorded and documentation is generated that gives the excursion visibility throughout the organization, which is also associated with the lot for follow-up correlation analysis. Several aspects of the failure including at-risk material (uninspected lots since last-known-good inspection), containment time, and inspection point (layer), are recorded. Chronic failure modes are assigned to case studies that are further investigated by engineering groups, and receive more visibility and resources from the organization for long-term corrective action and continuous yield improvement.

An accurate yield-to-defect model is pre-supposed for the following development of the cost-of-risk model. More refined models will create better estimates of the cost-of-risk at a given layer, however, as long as the yield model is consistent from layer-to-layer, the sample allocation should be unaffected. Errors in the sample allocation are caused mainly by large errors in the relative impact of a layer in comparison to the rest of the line. These errors can be reduced by pursuing an iterative approach to the yield-to-defect correlations, reducing the error through continuing refinement. There is a great deal of literature on yield-to-defect, yield models and methodology [1]–[4], so this aspect of the problem will not be treated here.

MODEL DEVELOPMENT

The following model development proceeds along the lines of standard "economic" or cost-based models of optimized SPC, as formulated by A. J. Duncan [5], Lorenzen & Vance [6], and Moskowitz *et al* [7]. Further details can be found in these references.

Each inspection layer is characterized by the control level, L, the number of wafers/lot, n, and the interval between inspections, I. The probability of an inspection not detecting an EXR, given that one has occurred — known as the β risk— is calculated as a function of the EXR magnitude, Δ, the defect distribution, $\Phi(\cdot)$, the control level, L, and the number of wafers, n. The review fraction, f, is used in a calculation of a β-risk specific to ADC, by Shindo *et al* [8]. Without loss of generality, we will not be using the review fraction in the demonstration of the model in this paper.

The β-risk is used to estimate the average number of intervals which will pass before a defect control level is violated given that an EXR has occurred. The number of intervals is called the out-of-control average run length (ARL), which is given by $ARL = (1-\beta)^{-1}$. The in-control average run length (ARL0) is the average length of time between false-alarms and is calculated through the α risk, which is simply a function of the in-control defect distribution and the limit, L. Constraints on the α risk can help minimize false alarms.

The process shown in Fig. 2 can be modeled as a re-generative system, with the estimated EXR frequency λ as the generalized transition rate [6]. Thus, the renewal-reward theorem states that the average cost per unit time is given by the expected costs over a renewal cycle divided by the renewal cycle time. The total cost includes the yield loss and scrap loss costs, the variable costs of inspection, fixed cost of equipment, the costs of false alarm, the cost of corrective actions, engineering support and other fixed costs.

The optimal sampling allocation will not depend on additive fixed costs as these will not change with the input parameters. The inspection equipment is assumed to be fully utilized, therefore the variable costs of inpection become fixed to a reasonable appoximation. The cost of corrective action will not be considered in calculations due to the difficulty of defining this cost objectively. Corrective action cost can be considered another fixed cost burden which may consist of additional inspection tool volume for ad hoc follow-up inspection(s), and test wafers. The cost of false alarm will be kept small through a constraint on the α risk. Due to these considerations, the cost function will not contain these terms and will only contain terms which pertain to the cost of yield loss. For this reason, this simpler cost function will be called the *cost-of-risk*.

The expected cost is given by

$$E[C] = C_0/\lambda + C_1(\Delta)[I(ARL) - \tau + \gamma + T] \qquad (1)$$

where C_0 is the baseline yield loss cost, C_1 is the cost of an EXR, which depends on the EXR magnitude Δ, τ is the average time of EXR occurence, γ is the queue-time at the inspection tool, and T is the average time until tool/process shut-down for corrective action. The inspection interval,I, and the ARL were explained earlier. The average time of EXR occurence is calculated by A. J. Duncan [5]. For EXR control, the baseline cost, C_0 is not considered and so this term is left out of the cost of risk.

The cycle-time is given by

$$E[T] = \lambda^{-1} - \tau + I(ARL) + \gamma + T \qquad (2)$$

and as an additional approximation the average queue-time at the inspection tool, γ, is taken out of the cycle time since it is assumed to be constant from layer-to-layer and will not affect the allocation.

The full expression for the cost of a production cycle is given in Lorenzen & Vance [6]. The cost of risk (as defined above) is then given by

$$C = C_1(\Delta)\frac{e^{-\lambda I} + \lambda I ARL + (1-ARL)\lambda I e^{-\lambda I}}{\lambda I(e^{-\lambda I} + ARL(1-e^{-\lambda I}))} \qquad (3)$$

This equation gives the total defect limited yield loss risk given an EXR of Δ standard deviations above baseline for a given layer. To obtain the total risk across all layers, we will add the risk from each layer. Given its complexity, a linear combination of equation 3 is difficult to optimize. A first approximation is to assume the statistical uncertainty is vanishingly small ($ARL = 1$, or $\beta = 0$). This leads to the expression, for the ith layer:

$$C(\beta = 0) = C_i(e^{-\lambda_i I_i} + \lambda_i I_i - 1)/\lambda_i I_i \qquad (4)$$

which is a standard result for the material at risk cost for an inspection monitored system with certain detectibility of failure [9], [10]. In what follows, this approximation will be called the *certain detection approximation* (CDA). To second order, this function can be written as

$$C_i(\beta = 0) \cong \frac{1}{2} C_i \lambda_i I_i + \mathcal{O}(\lambda I)^2 \qquad (5)$$

This a linear model of the cost, EXR frequency, and interval between inspections. This model will be called the *linearized certain detection approximation* (LCDA). Equation 5 will be fairly accurate for EXRs with large Δ, however, it will deviate greatly from the "exact" model (equation 3) at intermediate EXR levels. These intermediate EXR levels have a greater contribution to the cost (in frequency of occurence) than the extreme levels, so they must be accounted for. As an approximation in this intermediate zone, we set

$$C_i \cong \frac{1}{2} C_i \lambda_i I_i (ARL_i) \qquad (6)$$

This form will be used as the cost-of-risk model for a given layer. The cost, C_1, is estimated using a yield model of the defect event. In this investigation, the Poisson yield model is used (with unclustered defect density).

The three approximations are plotted in Fig. 3, as a function of EXR level. As can be seen, for large Δ, CRM is higher than the exact model by 10%. For intermediate Δ, the CRM follows the exact model better than the other approximations,

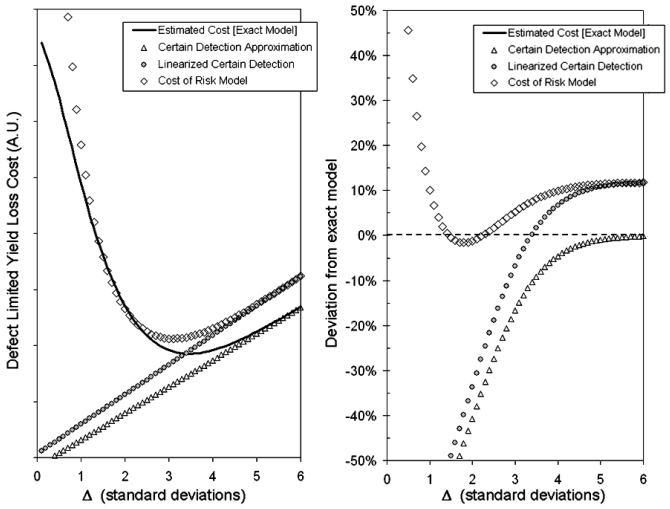

Fig. 3. (LHS) The defect limited yield loss cost as a function of the EXR magnitude Δ (standard deviations). The bold line is the exact model as given in Lorenzen & Vance (LV) [6]. For comparison is shown the certain detection approximation ($\triangle$), the linearized certain detection approximation ($\circ$), and the cost of risk model ($\diamond$). (RHS) The deviation (error) from LV for the three approximations: CDA, LCDA, and CRM.

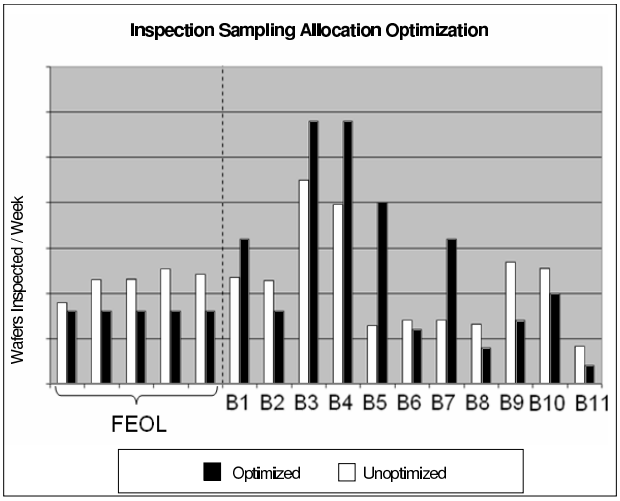

Fig. 4. ILM allocation (wafers / WW) comparing the prior (unoptimized) allocation to the optimized allocation using the allocation model given by Eq.. Due to low risk, FEOL layers have reduced sampling. Note the dramatic reallocations in B5 and B7 reflecting their high frequency and impact, whereas B9-B11 are reduced.

and toward small Δ all approximations diverge. As a weighting function for the EXR distribution, the CRM provides a reasonable risk metric which has the added benefit of being linear.

In order to compare different layers, the cost-of-risk function must be averaged over the distribution of Δ beginning at some high threshold Δ_{th} which is consistent from layer to layer. Once this averaging is done, the cost of risk of all layers is given by

$$C_R = \frac{1}{2}\sum_{i=1}^{N}\langle ARL_i C_i\rangle \lambda_i I_i \tag{7}$$

where $\langle\cdot\rangle$ stands for an average over the historical EXR distribution.

The linear form of C_R allows for optimization by the Lagrange multiplier method. In order to facilitate this optimization, the layer level wafers/lot are fixed at n_0 which is the same for all layers. With the layer-level constraint that the interval be no less than the average inspection queue time, and no greater than the time-to-sort, the total constraint that the production inspection volume is constant, the total number, n_i, of inspections per layer per unit time is given by

$$n_i = \left(\frac{n_t}{n_0}\right)\frac{\sqrt{\lambda_i\langle ARL_i C_i\rangle}}{\sum_{j=1}^{N}\sqrt{\lambda_j\langle ARL_j C_j\rangle}} \tag{8}$$

where n_t is the total available inspection tool volume (in wafers per unit time).

Discussion

The simple form of equation 7 allows for an analytic expression for the sample allocation. Other allocations can be derived by optimizing the exact model (eq. 3), the CDA (eq.4), or the LCDA (eq. 5). However, these other formulations either require a computerized search algorithm, or place too much weight on the extremes of the distribution. It also may be a case of diminishing returns for the added accuracy of the non-linear model in equation 3.

The form of eq. 7 is also intuitive considering the fact that it is comprised of frequency, detectibility, and impact. Risk metrics such as the RPN are constructed similarly [10]. In the absence of ESPC theory, if we simply chose the simplest function which contained the features of the problem (Frequency, detectability, and impact), and which was also of correct dimensions (Frequency has units of inverse time, detectability is unitless, and impact has units of cost), the immediate result would be eq. 7, however, the exact forms of the terms would be uncertain and the model derivation sheds light on what these terms should be.

Other risk metrics exist and these can be used to establish a optimal allocation. Construction of an equation involving the defectivity variance (as analogous to the cost) can provide similar results. The reason ESPC was used was to investigate the impact of detectibility on the cost, and find an approximation that can take this feature into account.

Equation 8 gives the optimal allocation of inspections given a minimized CRM. In order to compare yield loss costs with the fixed (additive) costs of inspection, and engineering support, the allocation should be used with a more accurate model such as equation 3, or the full LV model [6]. The full LV model also includes the queue-time, drill-down, and false alarm costs. The total cost of the inspection scheme can be integrated into a cost-of-ownership formulation [11], [12], in order to establish estimates of the return on investment of additional inspection capacity (ROI).

The queue-time of the inspection tool, and the drill-down time can be re-incorporated into the cost-of-risk allocation model by assuming that the average cost impact due to these factors is on the order of $C_i\lambda(\gamma + T)$. If the estimated cost

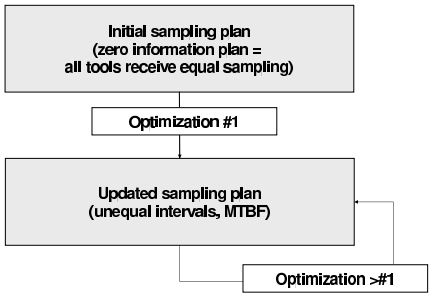

Fig. 5. Initial allocation will be uniform in absence of prior information, however, as more data is collected some layers will dominate the yield loss risk, and thus will demand a higher inspection rate

includes these factors, the correspondence is very close using this additional term. However, since the additional cost due to queue-time does not include the interval, this term does not contribute to the optimized sampling plan and this is not a part of the CRM.

Implementation of Model

The CRM was used to optimize the inspection allocation for a high volume CMOS manufacturing line. The value of λ is obtained from the DDMS database, along with the distribution of EXRs and their yield-impacts. The $\langle ARL \rangle$ is determined from the individual SPC parameters at a given layer. The results are shown in Fig. 4

The sampling allocation should be reviewed periodically and adjustments should be made if the EXR distributions have shifted significantly. One such time to make an evaluation is after a major process improvement change. After an initial proof test (extended observation) the additional data should be used to re-evaluate the allocation, and free up volume for more at-risk layers. This adaptive sampling process is illustrated in Fig.5.

Modeling and Input Errors

Errors due to the assumptions that led to the CRM have been discussed, however, there are other errors to consider due to the uncertainty of the inputs (C, λ, and the form of the β risk which leads to the ARL). The exact determination of λ depends on a long lead time prior to optimization, however, sometimes a decision must be made with very few (or zero) failures. In this case, it may be advisible to use the lower confidence limit for the MTBF given zero failures the values of which are given in many reliability statistics textbooks.

The cost depends on the type of yield model used, and the accuracy of the yield to defect model—whether it is an overlay analysis, critical area distribution, or some other analysis methodology. A more accurate yield model will lead to a better understanding of both the cost of risk, and the most appropriate allocation of inspection volume.

It is well established that, for an a uniform failure rate λ, the optimal distribution of inspections in time $(t_1, t_2, \ldots)$ is also uniform and equal [13]. Thus, maintaining a fixed interval, I, helps to keep the risk low. However, it is often difficult to maintain such a uniform interval, and variations will occur. The exact magnitude of this error is not known by the authors. It suffices to say that keeping the variance of the sampling interval to a minimum is a goal of the in-line manufacturing systems.

These uncertainties and errors are unavoidable in a manufacturing environment. This implies that the linear form of the CRM might be most robust against fluctuations of the input parameters as compared to more non-linear models such as eq. 3 or the full LV model. Future directions of research may include a full sensitivity analysis comparing LV, eq. 3 and the CRM.

Summary & Conclusion

In summary, a linear cost of risk model (CRM) was derived by approximating the full economic SPC of an in-line inspection layer. This model was compared with the exact model, and other approximations. More precise functions may exist, however, the CRM developed here has the attractive feature of being linear in all inputs. This allows for easy optimization, and gives an analytic form for the inspection allocation. This allocation has been used to optimize inspection volume in a high volume CMOS production line. The cost advantages to this lie in the systematic reduction of exposure to the highest risk layers based on historical data.

References

[1] P. Mullenix, J. Zalnoski, and A. J. Kasten, "Limited yield estimation for visual defect sources," *IEEE Trans. Semi. Manuf.*, vol. 10, no. 1, p. 17, 1997.

[2] L. S. Milor, "Yield modeling based on in-line scanner defect sizing and a circuit's critical area," *IEEE Trans. Semi. Manuf.*, vol. 12, p. 26, 1999.

[3] C. Stapper, "Lsi yield modeling and process monitoring," *IBM J. Res. Devel.*, vol. 20, p. 112, 1976.

[4] A. Ferris-Prabhu, *Introduction to Semiconductor Device Yield Modeling*. Artech House, 1992.

[5] A. J. Duncan, "The economic design of $\bar{X}$-charts when there is a multiplicity of assignable causes," *J. Amer. Stat. Assoc.*, vol. 66, p. 107, 1971.

[6] T. Lorenzen and L. Vance, "The economic design of control charts: a unified approach," *Technometrics*, vol. 28, p. 3, 1986.

[7] H. Moskowitz, R. Plante, and Y. H. Chun, "Effect of quality loss functions on the economic design of $\bar{x}$ process control charts," *European J. Oper. Res.*, vol. 72, p. 333, 1994.

[8] W. Shindo, E.H. Wang, R. Akella, A.J. Strojwas, W. Tomlinson, and R. Bartholomew, "Effective Excursion Detection by Defect Type Grouping in In-line Inspection and Classification," *IEEE Trans. Semi. Manuf.*, vol. 12, no. 1, p. 3, 1999.

[9] J. Sarkar and S. Sarkar, "Availability of a periodically inspected system under perfect repair," *J. Stat. Plan. Inf.*, vol. 91, p. 77, 2000.

[10] T. Bedford and R. Cooke, *Probabilistic Risk Analysis*. Cambridge Univ. Press, 2001.

[11] R. Carnes and M. Su, "Long Term Cost of Ownership: Beyond Purchase Price," *IEEE/SEMI Int'l Semi. Manuf. Sci. Symp.*, p. 39, 1991.

[12] D.L. Dance, T. DiFlorida and D.W. Jimenez, "Modeling the Cost of Ownership of Assembly and Inspection," *IEEE Trans. Comp. Packag. Manuf. Tech.*, vol. 19, p. 57, January 1996.

[13] J.B. Keller, "Optimum Inspection Policies," *Manage. Sci.*, vol. 28, no. 4, p. 447, 1982.