Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, eds

OPTIMIZED INSPECTION CAPACITY FOR OUT OF CONTROL DETECTION IN SEMICONDUCTOR MANUFACTURING

Israel Tirkel

Ben-Gurion University of the Negev
P.O.B. 653
Beer-Sheva 84153, ISRAEL

ABSTRACT

In-line inspection is designated to detect out-of-control (OOC) performance in order to increase quality output, and thus profit. Since inspection capacity is costly, it raises the question of how much capacity should be acquired to minimize OOC. Clearly, lower OOC requires higher capacity. This paper suggests a model that optimizes inspection capacity and OOC tradeoff. It is based on a typical process step monitored by an inspection facility. Processed items are sampled and inspected, considering inspection rate and response time, in order to minimize OOC rate at a given capacity. The inspection operating curve is established for demonstrating the tradeoff between OOC rate and inspection capacity. It exhibits that OOC decreases at a reduced rate with increasing capacity. Fab specific financials can provide the cost ratio between capacity and OOC for determining the preferred working point on the inspection operating curve.

1 INTRODUCTION

Production management strives to optimize the profit by maximizing the quality performance, while simultaneously minimizing the cost associated with inspection. This work originates in wafer fabrication and studies the relationship between quality performance, as measured by out-of-control (OOC), and the inspection capacity required to obtain it. It considers in-line inspection of processed items versus end-of-line inspection. In-line inspection concerns statistical design of sampling for process control, and suggests the selection of sample size, sampling frequency, and control limits (Montgomery 2009, Nahmias 2009). OOC implies that the processing machine is performing out of the control limits and requires repair, and that the item's quality might be lower. This is contrary to out-of-specification which indicates the item should be discarded (or reworked). We focus on investigating the sampling frequency and scheduling, by applying an inspection scheme for OOC detection in a simplified production stage model. The goal of this study is to suggest an approach and a model to optimize the inspection capacity and OOC tradeoff.

The work is motivated by the need to maximize quality performance using inspection capacity effectively (Mittal and McNally 1998). The production model assumed is based on a wafer fabrication machine, which randomly and independently deteriorates from in-control (IC) to OOC. The machine is inspected via items (i.e. wafers) it produces, where an inspection triggers a machine's repair, if needed. The focus is on improving the machine's performance rather than the process technology. The inspection scheme is applied for minimizing OOC rate given inspection capacity. No items are discarded since defective output (i.e. die) is combined with potentially good output on the same item (i.e. defected dies and good dies on the same wafer).

Wafer fabrication studies frequently rely on queueing models (Hopp et al. 2002, Zisgen et al. 2008). Since the early work of Duncan (1956) production systems quality performance models have been investigated (Liano 2007), and production systems maintenance has been intensively studied (Wang 2002). Inspection strategies in manufacturing relate to inspection-oriented quality-assurance for driving minimum production cost (Mandroli, Shrivastava and Ding 2006). They concern in-line inspection of semi-finished items and discrete-parts in a serial production system. Their definitions only partially

978-1-4673-4781-5/12/$31.00 ©2012 IEEE

apply here, since the inspection is performed to examine the state of the machine while considering work-in-process and inspection response time. Thus, the model here is unique.

SEMI E35 (1995) is the standard metrics for equipment cost of ownership, commonly used for semiconductor equipment acquisition (Sohn and Moon 2003). It assumes that the cost of machines comprises most of the fabrication cost, and that the machine's quality performance can be indicated per machine. Activity based costing extends cost of ownership with further cost analysis (Block and Carr 1999, Miraglia 2002, Scholtz 2002, Fandel and Wright 2008). This work relies on cost of ownership and activity based costing methods and assumptions in considering wafer fabrication capacity cost and inspection cost.

Section 2 describes the production system and the inspection scheme models, Section 3 explains the optimization process, Section 4 presents the results, and Section 5 draws the conclusions.

2 THE MODEL

This section presents methodology and explains the assumed production system and inspection models, which are mostly based on Tirkel and Rabinowitz (2011).

2.1 The Production Model

Wafer fabrication lines include hundreds of steps where in current technologies almost half are allocated for inspections, such as: defect detection, surface uniformity, geometric and electric measurements. Consequently, the average process-to-inspection step ratio is assumed 1:1. In practice, the case where a single process step is individually inspected is known as a station monitor. For simplicity, the case of integrative monitors, where a sequence of inspection steps follows a sequence of process steps, is disregarded. Similar to practice, it is assumed that some of the items are inspected while other items continue directly to the following process step.

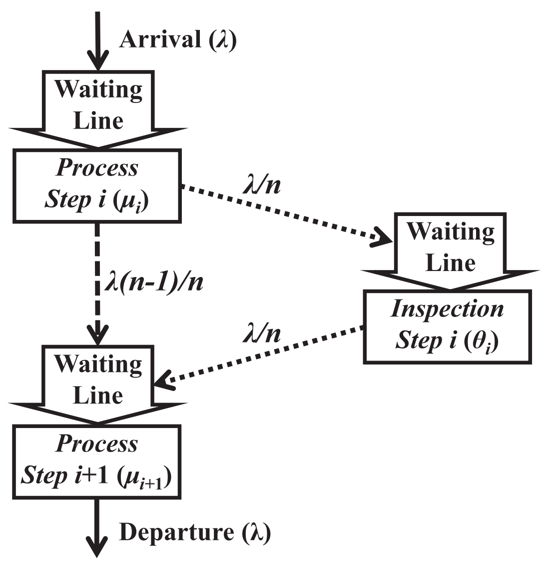

Figure 1: The typical production stage model

The production model assumes a series of consecutive, independent and repeating stages, where a typical stage structure is illustrated in Figure 1 (Tirkel, Reshef and Rabinowitz 2009). An item departing *Process Step i* can either continue directly to *Process Step* i+1, or be initially routed to *Inspection Step i* and only then to *Process Step i+1*, pending on the sampling scheme. The model determines the machine's state when it starts processing, as either: (a) stay IC with a known transition probability p, (b) transfer from IC to OOC with probability 1-p, or (c) stay OOC with probability 1. The purpose of *Inspection Step i* is to assess the item's quality, and as a result conclude the *Process Step i* machine's state as OOC or IC. The machine's state detection is assumed to be with no errors; if OOC is detected, the respective machine is immediately repaired to IC, at the first instance it is not processing. All items accumulate flow-time in the process steps, yet only some in the inspection steps. It is assumed that items are not reworked or discarded, and do not acquire flaws in the inspection but only in the processing steps. It is also assumed that all process steps in the production line are independent in their states. Each step is a single queue with a single waiting-line and a single server (FIFO). It is assumed that each process step is a G/M/1 queue that can be approximated by an M/M/1 queue, based on empirical results that show insignificant statistical gap of the flow-time distribution. Each inspection step is an E_n/M/1 queue, which represents an Erlang distributed arrival with parameter *n* (explained in the inspection model's section). The following definitions apply in Figure 1:

λ – *Process Step i* arrival/departure rate, representing the total throughput [items/time-unit],

μ_i – *Process Step i* service rate, representing process capacity [items/time-unit],

θ_i – *Inspection Step i* service rate, representing inspection capacity [items/time-unit].

The illustrative example results, presented in Section 4, rely on the following model's parameters,

$p = 0.99,\ \lambda = 1, \mu_i = \mu_{i+1} = 1.11, \theta_i = range\ [0.11, 1.11]$

The fundamental definitions of the production line performance measures are,

The expected flow-time of *m* items (indexed *j*=1, 2, …, *m*) processed in any *Step i*,

$$FT_i = \frac{1}{m}\sum_{j=1}^{m}\left(CompleteTime_{ij} - StartTime_{ij}\right) \tag{1}$$

where, $StartTime_{1,j} = 0$, and $StartTime_{i,j} = CompleteTime_{i-1,j}$.

The OOC rate is the proportion of OOC items out of the total items processed in *Process Step i*,

$$OOC_i = \frac{Items\ processed\ in\ Process\ Step\ i\ in\ OOC\ state}{Total\ items\ processed\ in\ Process\ Step\ i} \tag{2}$$

2.2 The Inspection Model

The inspection scheme follows a pre-defined sampling algorithm (e.g. decision rule). Items are sampled for inspection as determined by the inspection scheme. This work applies the most common inspection scheme in practice, noted Fixed Measure Rate. It follows the decision rule, where: if *n*-1 consecutive items depart the process step, send the n^{th} item to inspection. Inspection rate (IR) is the proportion of items inspected, determined by IR=1/*n*.

Any inspection is associated with flow-time accumulated via the items sampled. The inspection time (IT) is measured since an item departs the processing step and until it departs the respective inspection step (i.e. response time) and triggers a machine repair, if needed. During this time the machine continues to process items. Longer IT will result in more items processed by a potentially OOC machine, and will consequently drive lower OOC rate. The IT depends on the inspection scheme.

The study investigates the effects of the inspection scheme and the inspection capacity on the OOC rate. The inspection scheme is defined by IR, and drives the consequent IT. The inspection capacity is defined in percent of process capacity (i.e. 10%, 20%, …, 100%),

$$Inspection\ Capacity_i = \frac{\theta_i}{\mu_i} \cdot 100 \tag{3}$$

This definition of inspection capacity is used since the cost of process machines comprises most of the fabrication cost, where all other costs can be measured relative to it, based on semiconductors standards (SEMI E35 1995). This assumption also eliminates the need to rely on revenue and average sale price which is out of the scope for this work.

Further formulation details are summarized in Appendix A.

3 OPTIMIZATION PROCESS

This section outlines the process of optimizing the inspection capacity and OOC rate tradeoff, in three phases. Each phase relies on the previous one, yet obtains results which are individually meaningful. Following is the outline of the phases, which is later detailed in Section 4 of the results:

I. The minimum OOC rate is obtained via the inspection scheme's IR and the consequent IT, given the inspection capacity (Figure 2). This phase is based on the model described in Section 2, and is further explained via the results presented in Section 4.
II. The inspection operating curve is established using the minimum OOC rate solutions (Figure 3), each obtained via applying a different inspection capacity level, as defined in (3). The inspection operating curve is created by consolidating these solutions within a single illustration (Figure 4).
III. Once the inspection operating curve is established, the optimal working point on the curve can be obtained by relating the fab specific financial data. Since the inspection operating curve reflects the Pareto efficiency of inspection capacity versus OOC rate, the optimal working point is selected relating the OOC-to-capacity cost ratio (Figure 5), as further explained in Section 4.

4 RESULTS

This section presents the detailed results following the outline of Phases I, II, III.

4.1 Phase I

Figure 2 illustrates OOC rate versus IR as a convex curve, explained next (follow the curve in Figure 2, left-to-right). As more items are sent to inspection IR grows, which consequently drives improved machine's quality performance, as measured by the OOC rate to decrease (A). As even more items are sent to inspection the IR grows further, consequently the OOC rate reaches a minimum (B) and then starts to increase (C). This increase is explained by the growing load at the inspection facility, resulting in prolonged IT thus causing the machine to process more items until repaired, potentially in OOC state, which drives the OOC rate increase. The minimum OOC rate is 9.19% (circled) and it is obtained at 17% IR (n=6), given 30% inspection capacity.

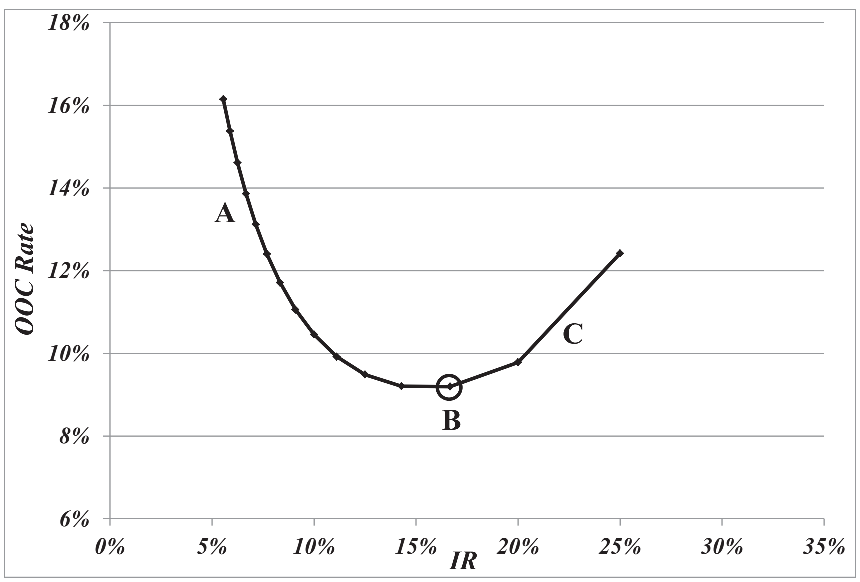

Figure 2: Minimum OOC Rate given 30% Inspection Capacity

4.2 Phase II

Figure 3 illustrates OOC rate versus IR for three different inspection capacity levels of 20%, 30%, and 40%, as defined by (3). Notice all curves exhibit convex shaped curves, in accordance with Phase I. The curves comparison demonstrates that higher inspection capacity generates lower OOC rates, given the same IR. This is explained due to shorter IT, due to lower load on the inspection facility. At 20% inspection capacity, minimum OOC rate is 12.75%, at 11% IR (n=9); at 30% inspection capacity, minimum OOC rate is 9.19%, at 17% IR (n=6); and at 40% inspection capacity, minimum OOC rate is 7.25%, at 20% IR (n=5). This demonstrates that increased inspection capacity enables higher IR, which results obtaining lower minimum OOC rates. Figure 3 illustrates the minimum OOC rate point on each curve (circled).

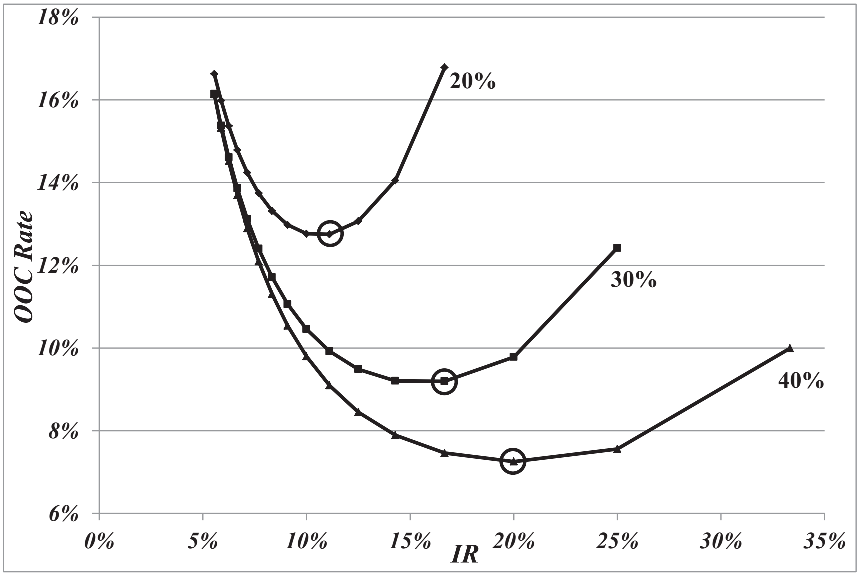

Figure 3: Minimum OOC Rate given 20%, 30%, and 40% Inspection Capacity

Figure 4 illustrates the inspection capacity in the range of 10% through 100% of process capacity, as defined by (3), versus the corresponding minimum OOC rate obtained at each capacity level. It reflects that OOC rate monotonically decreases at a reduced rate, versus increasing inspection capacity. This curve defines the inspection operating curve. It demonstrates the tradeoff between the inspection capacity and the accordant OOC rate (as a function of IR). Thus, the inspection operating curve describes the operating relationship between inspection capacity and the OOC rate.

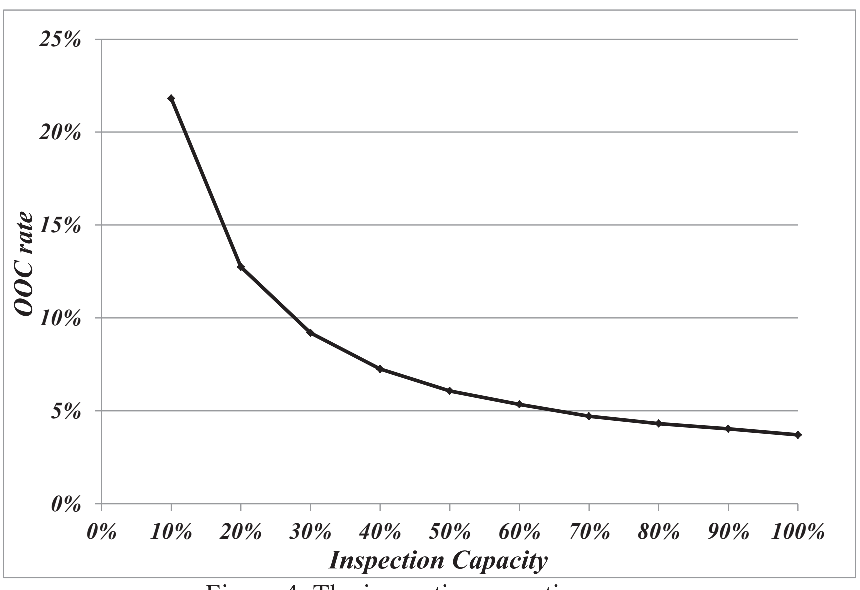

Figure 4: The inspection operating curve

4.3 Phase III

Figure 5 illustrates the preferred working point given the OOC-to-capacity cost ratio. It is based on the fab's specific financial data. The proposed cost ratio reflects the cost of 1% OOC rate (e.g. associated with output loss cost) relative with cost off 1% inspection capacity (e.g. associated with equipment acquisition cost), given by,

$$Cost\ Ratio = \frac{Cost\ of\ 1\%\ OOC\ rate}{Cost\ of\ 1\%\ inspection\ capacity} \tag{4}$$

Once the cost ratio is given, the preferred working point can then be selected. This is exhibited via the two following examples and Figure 5:

a. Cost ratio at 0.5 – reflects OOC-to-capacity cost ratio which is relatively high. It is illustrated in Figure 5 with a steep 0.5 gradient (dotted line). The preferred working point indicates inspection capacity acquisition of 20% and an OOC rate of 12.75% (IR=11%).
b. Cost ratio at 0.1 – reflects OOC-to-capacity cost ratio which is relatively low. It is illustrated in Figure 5 with a mild 0.1 gradient (dotted line). The preferred working point indicates inspection capacity acquisition of 50% and an OOC rate of 6.07% (IR=25%).

Clearly, higher inspection capacity-to-OOC cost ratio drives lower inspection capacity and higher OOC rate, and vice-versa.

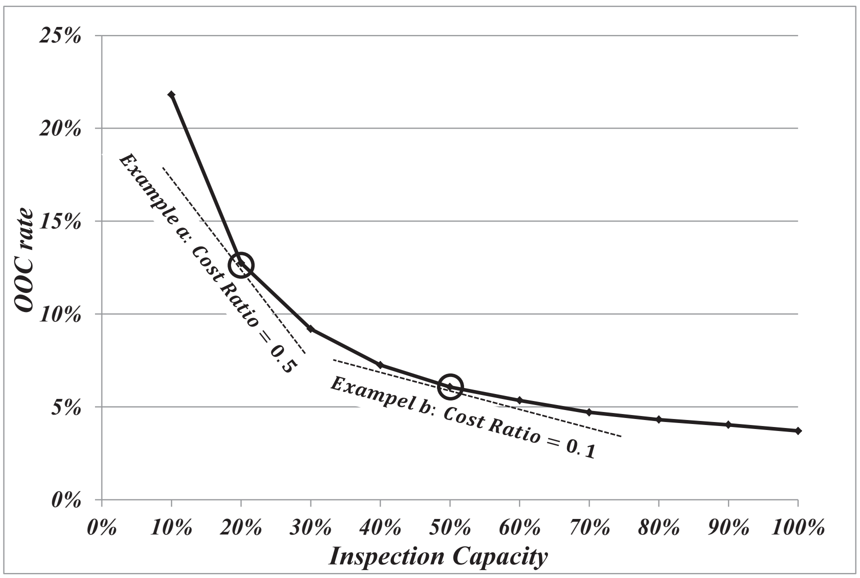

Figure 5: The preferred working point on the inspection operating curve

5 CONCLUSIONS

This work exhibits that OOC rate versus IR illustrates a convex curve, given a fixed inspection capacity. OOC rate decreases with increasing IR until a minimum is reached and then starts to increase with further increasing IR, explained by prolonged IT causing higher OOC rate. Higher inspection capacity generates lower OOC rates, given the same IR. The inspection operating curve was established illustrating that OOC rate decreases at a reduced rate, with increasing inspection capacity. Assuming OOC-to-capacity cost ratio is given, the preferred working point on the inspection operating curve can be obtained. Thus, the model suggests the optimal combination of inspection capacity and OOC rate, determined by the IR, for achieving the minimum cost associated with quality performance. The originality of this work is in modeling the relationship between OOC ratio and IR given inspection capacity, the impact analyses of inspection capacity levels on quality performance as measured by OOC rate, the establishment of the inspection operating curve which defines the tradeoff between inspection capacity and OOC rate, and in the method presented for selecting the preferred working point given the fab cost ratio.

A APPENDIX

Following is the summary of the flow-time and OOC rate analytical formulation based on Tirkel and Rabinowitz (2011).
The expected flow-time of a single inspection step (IT), is given by,

$$FT_I = 1/\theta(1-\omega), \tag{A.1}$$

where,
θ is an inspection step service rate, and
ω is the probability that an arriving item finds an inspection step busy.

The probability ω given an Erlang arrival at the queue, is given by,

$$\omega = \left(n\rho_I/\left(n\rho_I + (1-\omega)\right)\right)^n, \tag{A.2}$$

where ρ_I is the inspection step traffic intensity rate.

The expected OOC approximation of a single process step given *n*, is given by,

$$OOC_n \cong \sum_{d=0}^{\infty} OOC_{n,d} P(FD = d), \tag{A.3}$$

where,
$OOC_{n,d}$ is the expected OOC rate given *n* and *d*,
d is the number of items that start service (process) at the process step during FT_I.

The distribution of *d*, is given by,

$$P(d) = \begin{cases} \lambda^{d-1} u(\lambda + u\rho_P)/(\lambda + u)^{d+1}, & if\ d = 1,2 \ldots, \\ u(1-\rho_P)/(\lambda + u), & if\ d = 0. \end{cases} \tag{A.4}$$

where,
λ is the arrival rate at a process step,
ρ_P is the process step traffic intensity rate,
and $u = 1/FT_I$.

REFERENCES

Block, B., and L. C. Carr. 1999. "Activity Based Budgeting at Digital Semiconductor". *International Journal of Strategic Cost Management*, spring 1999, pp. 17-31.

Duncan, A. J. 1956. "The Economic Design of X-Charts used to Maintain Current Control of Process". *Journal of the American Statistical Association*, 51(274), pp. 228-242.

Fandel, D., and R. Wright. 2008. "300mm Productivity Detractors Mitigation Cost Analysis". *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 222-227.

Hopp, W. J., M. L. Spearman, S. Chayet, K. L. Donohue, and E. S. Gel. 2002. "Using an Optimized Queueing Network model to Support Wafer Fab Design". *IIE Transactions*, 34, pp. 119-130.

Iwata, Y., and S. C. Wood. 2002. "Simple Cost Models of High-Process-Mix Wafer Fabs at Different Capacities". *IEEE Transactions on Semiconductor Manufacturing*, 15(2), pp. 267-273.

Liao, G. L. 2007. "Optimal Production Correction and Maintenance Policy for Imperfect Process". *European Journal of Operational Research*, 182, pp. 1140-1149.

Mandroli, S. S., A. K. Shrivastava, and Y. Ding. 2006. "A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes". *IIE Transactions*, 38, pp. 309-328.

Mittal S., and P. McNally. 1998. "Line Defect Control to Maximize Yields". *Intel Technology Journal*, 4(2).

Miraglia, S., C. Blouin, G. Boldman, S. Judd, T. Richardson, and D. Yao. 2002. "ABC Modeling: Advanced Features". *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 336-339.

Montgomery, D.C. 2009. *Introduction to Statistical Quality Control*. 6th Edition. Wiley.

Nahmias, S. 2009. *Production and Operations Analysis*. 6th Edition. McGrow Hill.

SEMI E35, 1995. "Guide to calculate cost of ownership (COO) metrics for semiconductor manufacturing equipment". *SEMI Standards*, USA, CA, http://www.semi.org

Sohn, S. Y., and H. U. Moon. 2003. "Cost of Ownership Model for Inspection of Multiple Quality Attributes". *IEEE Transactions on Semiconductor Manufacturing*, 16(3), pp. 565-571.

Scholtz III, R. L. 2002. "Strategies for Manufacturing Low Volume Semiconductor Products in High Volume Manufacturing Environment". *M.Sc. Thesis at MIT*, MA, USA.

Tirkel, I., N. Reshef and, G. Rabinowitz. 2009. "In-line Inspection Impact on Cycle Time and Yield". *IEEE Transactions on Semiconductor Manufacturing*, 22(4), pp. 491-498.

Tirkel, I., and G. Rabinowitz. 2011. "The Relationship between Yield and Flow Time in a Production System under Inspection". *International Journal of Production Research*, on-line 2011, hardcopy 2012.

Wang, H. 2002. "A Survey of Maintenance Policies of Deteriorating Systems". *European Journal of Operational Research*, 2002, 139, pp. 469-489.

Zisgen,H., B. R. Wheeler, I. Meents, and T. Hanschke. 2008. "A Queueing Network Based System to Model Capacity and Cycle Time for Semiconductor Fabrication". *Proceedings of the 40th Winter Simulation Conference*, Miami, FL, pp. 2067-2074.

AUTHOR BIOGRAPHY

ISRAEL TIRKEL is a faculty researcher-lecturer in the department of Industrial Engineering and Management at Ben-Gurion University of the Negev, Israel. He has worked for Intel Corporation, in Israel and the USA, for twenty-three years in senior management positions of Fab Operations and Program Management. He received his B.Sc. with distinction at 1983, M.Sc. with distinction at 2009, and Ph.D. at 2011 in Industrial Engineering and Management, from Ben-Gurion University of the Negev. His areas of specialization are production and operations analysis and management, and project management, which he formerly practiced and is now investigating and lecturing. He is an Associate Editor in IEEE Transactions on Semiconductors Manufacturing. His e-mail address is tirkel@bgu.ac.il