

SETTING QUALITY CONTROL REQUIREMENTS TO BALANCE CYCLE TIME AND YIELD – THE SINGLE MACHINE CASE

Miri Gilenson,
Liron Yedidsion

Technion, Israel Institute of Technology
Haifa 32000, ISRAEL

Michael Hassoun

Ariel University Center
Ariel 40700, ISRAEL

ABSTRACT

Control limits in use at metrology stations are traditionally set by yield requirements. Since excursions from these limits usually trigger machine stoppage, the monitor design has a direct impact on the station's availability, and thus on the product cycle time (CT). In this work we lay the foundation for a bi-criteria trade-off formulation between expected CT and die yield based on the impact of the inspection control limits on both performance measures. We assume a single machine plagued by a particle deposition process and immediately followed by a monitor step. We explore the impact of the upper control limit on the expected final yield on one hand, and on the distribution of the station time between consecutive stoppages on the other. The obtained model enables decision makers to knowingly sacrifice yield to shorten CT and vice versa.

1 INTRODUCTION

In the challenging semiconductor manufacturing process, the connection between quality and flow time, or as it is called in the fab jargon “Cycle Time” (CT), is an important issue. The race for ever smaller, higher density devices from the same silicon wafer and the periodic upgrades in wafer size drive the industry through very fast obsolescence cycles.

In this environment, perfect quality is all but impossible, and a certain portion of devices inevitably fails the functional tests that conclude the fabrication process. In addition to each production step's quality requirements, the issue of protecting the wafer from being contaminated by particles the size of a few tenths of a micron is present all along the hundreds of steps necessary to building the device. Should a particle large enough fall on a device, this device will not function. The numerous devices lying on a wafer being physically bound together, the portion of functional devices at the end of the process is a crucial indicator of the technology health. This performance measure is denoted “die yield”. One aspect of the relation binding Yield and CT has been considered by Wein (1992) and Cunningham and Shanthikumar (1996) among others, who studied the adverse effects of prolonged lot stay on particle contamination. We propose to study the opposite problem of the impact of quality-related decisions on CT. To maintain a high die yield, line quality control is of utmost importance and allows taking two types of decisions: First, should the sampled wafers be concerned with a quality issue, and their expected yield compromised, the management may decide to scrap them, depending on how much work has already been invested in these wafers and how much is still to be done. Second, monitoring wafers allows postulating on the condition of the machine or machines on which the last operations were performed, and, if needed, to stop production in order to bring the machine back to specifications. In this regard, Munga et al. (2011) have proposed a solution to optimize the root problem identification once an excursion has occurred at a metrology station controlling not one, but a series of operations, while assuming that for productivity reasons, production is not stopped until the machine responsible for the excursion is identified.

This constant monitoring takes its toll on the product's CT. The additional time spent by the wafers at metrology stations, as well as the machine stoppages following an out-of-spec monitor, both slow down the work-in-process (WIP) flow. Therefore, there is a trade-off between the design of a quality policy and CT. Raising the frequency of monitors along the process, the number of wafers sampled, or the number or size of the sampled sites on these wafers naturally increases yield while stressing the metrology tools' capacity, thus adding waiting time. This trade-off has been well described in Colledani (2008), and Colledani and Tolio (2006; 2011).

The logic device (CPU) industry is characterized by a high benefit margin and the willingness of customers to wait for the right product. As such, it has naturally shifted towards a decision process that first sets the quality control requirements based on the targeted yield, basically leaving the industrial engineers to struggle for the best possible CT under these requirements. Lately, under the formidable increase in demand for memory devices that are not bound by the same market rules, short CT has become paramount for semiconductor companies' survival. As a consequence, setting a quality policy while considering its impact on CT seems to offer an opportunity to managers. Meyersdorf and Yang (1997), as well as Kethan et al. (1995) present some aspects of the trade-off specific to fabs without quantifying it. But more recently, Tirkel et al. (2009) quantified the CT-Yield trade-off, while proposing a dynamic monitoring policy instead of the traditional constant sampling broadly in use (such dynamic policies are not new, but were so far mainly used to improve Yield. See Dauzère-Pérès et al. (2011) for an example).

In the current work, we tackle the question of the control limits determination as one of those factors that impacts both yield and CT in a semiconductor environment. Indeed, setting tighter limits will allow for better yield levels and quicker responses to quality drifts. The price paid will be in the form of more frequent machine stoppages that inevitably impact the stations' availability and thus CT. In this work, we do not address the evident question of capacity reduction at bottleneck stations, although, clearly, it is an important consideration, but rather focus on the general slowing down of the flow due to the stoppage frequency increase. We analyze the impact of the control limits on a single machine CT and on the expected yield in Sections 2 and 3, respectively. In Section 4, numerical results of the trade-off between the two metrics are presented.

2 IMPACT OF CONTROL LIMIT ON CT

In the framework of this paper we model a single machine, followed by a metrology step in which the lots are examined for defects, and the decision to let the station continue producing or not, is taken. The inspection time is assumed to be negligible.

Wafer lots arrive at the station at rate λ . The service duration is a constant t .

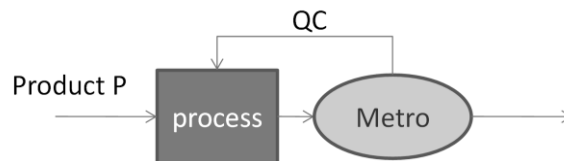


Figure 1: Single station scheme.

The assumption made about defects appearance being a Poisson process may better correspond to the particle contamination process, yet it is generic enough to be relevant to other types of defects as well.

Tens to hundreds of microelectronic devices are built layer upon layer on the silicon base. Defects appearing at any stage of the process may, or may not, destroy the device functionality. In our model, defects are device killers, independently of their exact location. However, the definition of a defect, regarding its size or any other characteristics, can be different at each station (in practice, certain operations are more sensitive than others). At the metrology station, part of the lot/wafer surface is sampled. When the

number of defects on the sampled area exceeds a predefined Upper Control Limit (UCL), the station is said to be Out Of Control (OOC), and production is interrupted. Otherwise, the station is said to be In Control (IC). We assume the number of defects added to the sampled area of a specific wafer, denoted x , to follow a Poisson process with parameter μ . The station is described as a two-state machine, and its deposition rate can either be high ($\bar{\mu}$) or low ($\underline{\mu}$).

The probability for a monitor to exceed the control-limits can be obtained by:

$$P(OOC) = 1 - P(x < UCL) = 1 - \sum_{k=0}^{UCL} \frac{(\mu)^k e^{-\mu}}{k!}, \text{ where } \mu \in \{\bar{\mu}, \underline{\mu}\} \quad (1)$$

We denote α the probability that a monitor exceeds the UCL when the deposition process rate is low (type 1 error), and β the probability of a monitor to remain under UCL when the deposition process is, in fact, high (type 2 error). We consider a sample to be IC if $x \leq UCL$ and OOC otherwise. We have:

$$\alpha = P(x > UCL | \underline{\mu}) = \sum_{k=UCL+1}^{\infty} \frac{(\underline{\mu})^k e^{-\underline{\mu}}}{k!} = 1 - \sum_{k=0}^{UCL} \frac{(\underline{\mu})^k e^{-\underline{\mu}}}{k!} \quad (2)$$

$$\beta = P(x \leq UCL | \bar{\mu}) = \sum_{k=0}^{UCL} \frac{(\bar{\mu})^k e^{-\bar{\mu}}}{k!} \quad (3)$$

We model the evolution of a single station over time with four states:

1. The deposition rate is low ($\underline{\mu}$) and the monitor result is *IC*;
2. The deposition rate is low ($\underline{\mu}$) and the monitor result is *OOC* (type 1 error);
3. The deposition rate is high ($\bar{\mu}$) and the monitor result is *IC* (type 2 error);
4. The deposition rate is high ($\bar{\mu}$) and the monitor result is *OOC*.

We denote the probability of the contamination rate to change from $\underline{\mu}$ to $\bar{\mu}$ by p . Once the contamination rate has risen it will not go back to $\underline{\mu}$ unless a repair procedure is conducted. Hence, once a monitor is OOC and the station is stopped (states 2 or 4), a repair or inspection is mandatory and inevitably brings the station to state 1. Under these assumptions, we can now present the station as a Markov Decision Process, shown in Figure 2 below:

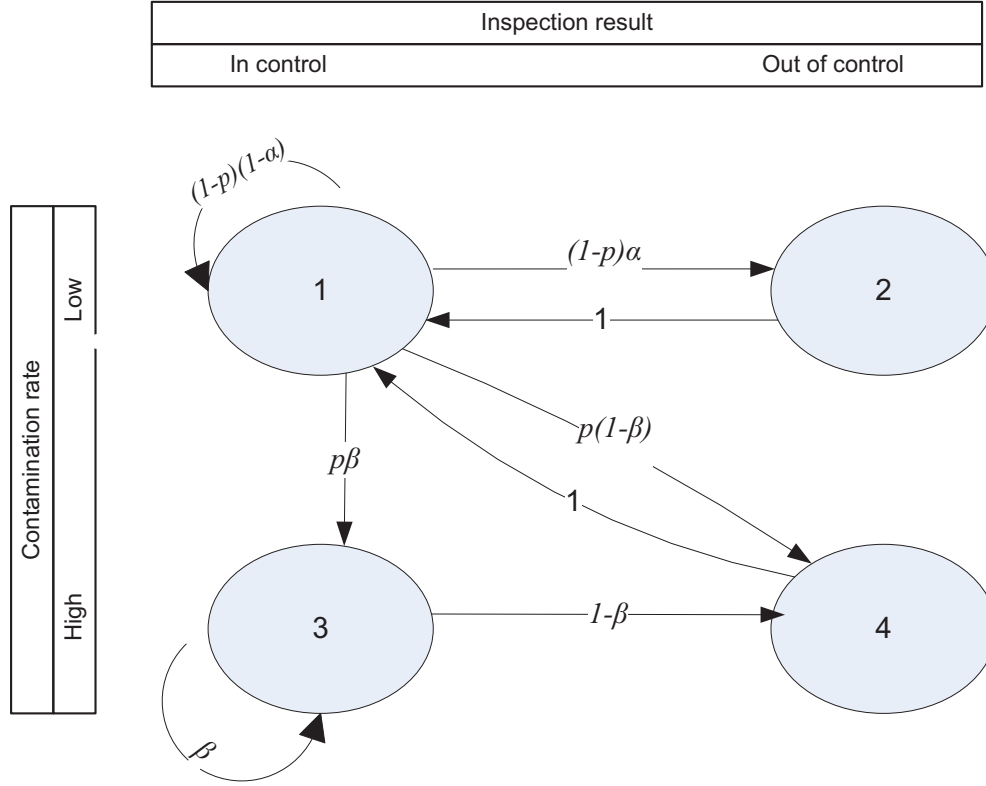


Figure 2: The station as a Markov chain.

We wish to determine the pattern of stoppages caused by excursion from the control limits. Since a repair is always successful and production is initiated with the machine in state 1, we are therefore trying to determine the distribution of the time required for the system to transfer from state 1 to either state 2 or 4. We denote the number of production cycles between two consecutive excursions from UCL, known as the Run Length by RL and the expectancy of this metric by ERL.

The distribution function of RL is obtained by calculating P_{stop}^n , the probability to stop after n process cycles. The Chapman–Kolmogorov equation (see Bolch et al. 2006) allows us to formulate the n -step transition probability from state 1 to state 2 as:

$$P_{12}^n = P_{11}^{n-1} P_{12}$$

$$P_{12}^n = [(1-\alpha)(1-p)]^{n-1} \alpha(1-p) = \alpha(1-\alpha)^{n-1} (1-p)^n \quad (4)$$

and the probability to move in n steps from state 1 to state 4 as:

$$P_{14}^n = P_{11}^{n-1} P_{14} + \sum_{k=0}^{n-2} P_{11}^k \cdot P_{13} \cdot P_{33}^{n-k-2} \cdot P_{34}$$

$$P_{14}^n = [(1-\alpha)(1-p)]^{n-1} p(1-\beta) + \sum_{k=0}^{n-2} [(1-\alpha)(1-p)]^k p\beta p^{n-k-2} (1-\beta) \quad (5)$$

P_{stop}^n can now be calculated as:

$$P_{stop}^n = P_{12}^n + P_{14}^n = \alpha(1-\alpha)^{n-1}(1-p)^n + p(1-\beta)\frac{\beta^n - (1-\alpha)^n(1-p)^n}{\beta - (1-\alpha)(1-p)} \quad (6)$$

The states are a closed group and the number of process cycles until we stop the station for readjustment is a discrete random variable with probabilities P_{stop}^n for $n=1, \dots, \infty$. Therefore, we have:

$$ERL = \sum_{n=1}^{\infty} n P_{stop}^n. \quad (7)$$

However, in an attempt to reach a closed form formulation of ERL, we found that for small values of α and β (it is reasonable to assume that any sample method used in the industry would guarantee that α and β do not exceed 0.2) the distribution function of the RL can be approximated by a geometric distribution whose parameters can be calculated based on the steady-state probabilities of the Markov chain. When referring to the Markov chain presented in Figure 2, it is easy to prove that it is ergodic. By solving the equilibrium equations, we obtain the stationary probability vector:

$$\vec{\pi} = \left(\frac{1}{1+(1-p)\alpha + \frac{p}{(1-\beta)}}, \frac{(1-p)\alpha}{1+(1-p)\alpha + \frac{p}{(1-\beta)}}, \frac{\frac{p\beta}{(1-\beta)}}{1+(1-p)\alpha + \frac{p}{(1-\beta)}}, \frac{p}{1+(1-p)\alpha + \frac{p}{(1-\beta)}} \right). \quad (8)$$

Consequently, if we consider m_t the monitor result at period t , the probability to stop the machine in one step is:

$$p^G = P(m_{t+1} = OOC | m_t = IC)$$

The complementary probability for a monitor to remain in control from one step to the following one is therefore:

$$q^G = (1 - p^G) = P(m_{t+1} = IC | m_t = IC),$$

which we have calculated to be:

$$q^G = \frac{\pi_1[(1-p)(1-\alpha) + p\beta] + \pi_3\beta}{\pi_1 + \pi_3} = \frac{(1-p)(1-\alpha)(1-\beta) + p\beta}{(1-\beta) + p\beta} \quad (9)$$

and:

$$p^G = 1 - q^G = \frac{(1-\beta)[1 - (1-p)(1-\alpha)]}{(1-\beta) + p\beta} \quad (10)$$

Therefore, the geometric distribution function of the probability to stop at the n^{th} step is approximately

$$P_{stop}^n \approx (q^G)^{n-1} p^G = \left(\frac{(1-p)(1-\alpha)(1-\beta) + p\beta}{(1-\beta) + p\beta} \right)^{n-1} \frac{(1-\beta)[1 - (1-p)(1-\alpha)]}{(1-\beta) + p\beta} \quad (11)$$

and the ERL can be expressed as:

$$ERL \approx \frac{1}{p^G} = \frac{1 + p \frac{\beta}{1-\beta}}{1 - (1-p)(1-\alpha)}. \quad (12)$$

The final step in this connection between UCL and CT aims at connecting the station ERL to the CT of a product. The single machine case is tractable. The model was studied by Madan and Saleh (2001) and was denoted as M/D/M/1. The model's assumptions are as follows: customers arrive at the server following a Poisson distribution with rate λ . The processing duration is deterministic and equals t . The time between consecutive vacations as well as vacation durations are exponentially distributed with parameters p^G and $\frac{1}{v}$, respectively. According to Madan and Saleh (2001), the average CT at the station can be expressed as:

$$CT = t + \frac{p^G(v + v^2)}{1 + p^G v} + \frac{\lambda(t^2 + 2p^G v(t + v))}{2(1 - \lambda t - \lambda p^G v)}. \quad (13)$$

3 IMPACT OF THE CONTROL LIMIT ON YIELD

In our single-machine model, the impact of the control limit on the average yield is straightforward. Considering a single process step with contamination $\mu \in \{\underline{\mu}, \bar{\mu}\}$, the Expectancy of the Contamination Rate denoted by ECR is:

$$ECR = (\pi_1 + \pi_2)\underline{\mu} + (\pi_3 + \pi_4)\bar{\mu} \quad (14)$$

Clearly, the value of UCL, by affecting the sensibility to defects, will also affect the stationary probabilities of the Markov chain, and thus the ECR . We assume the defects to be uniformly scattered over the wafer's surface and their deposition to follow a Poisson process. We define the number of dies on a single wafer by N . The probability of a particle to fall on a specific die is thus $\frac{1}{N}$. This probability also determines the proportion of un-damaged dies, namely the die yield:

$$E(yield) = \left(\frac{N-1}{N} \right)^{ECR}. \quad (15)$$

4 NUMERICAL STUDY

We performed an experimental analysis of the single station model. Several variable sets were considered and the CT and yield were plotted as functions of the UCL. We present here an example that exhibits values typical for real manufacturing data (i.e., high yield and low contamination rate).

The defects' deposition rates, $\underline{\mu}$ and $\bar{\mu}$, are 5 and 15 defects per sampling area, respectively; the probability to change from low to high contamination rate is $p = 0.1$. Wafer arrival rate is $\lambda = 0.05$ parts per time unit, the service duration is $t = 5$ time units, and the mean vacation period is $v = 10$ time units. The number of dies on a single wafer is $N = 100$.

Figure 3 describes the impact of UCL on both CT and yield. Figure 4 plots the CT improvement in terms of yield loss and vice versa. As one can observe, both the CT and the yield decrease while control limits are broadened. One should recall that a UCL lower than $\underline{\mu}$ will lead to an almost immediate stop (probability greater than 0.5) even when the process is in control, and a UCL greater than $\bar{\mu}$ tends to yield little to no stoppage even when the process is out of control and a repair is called for.

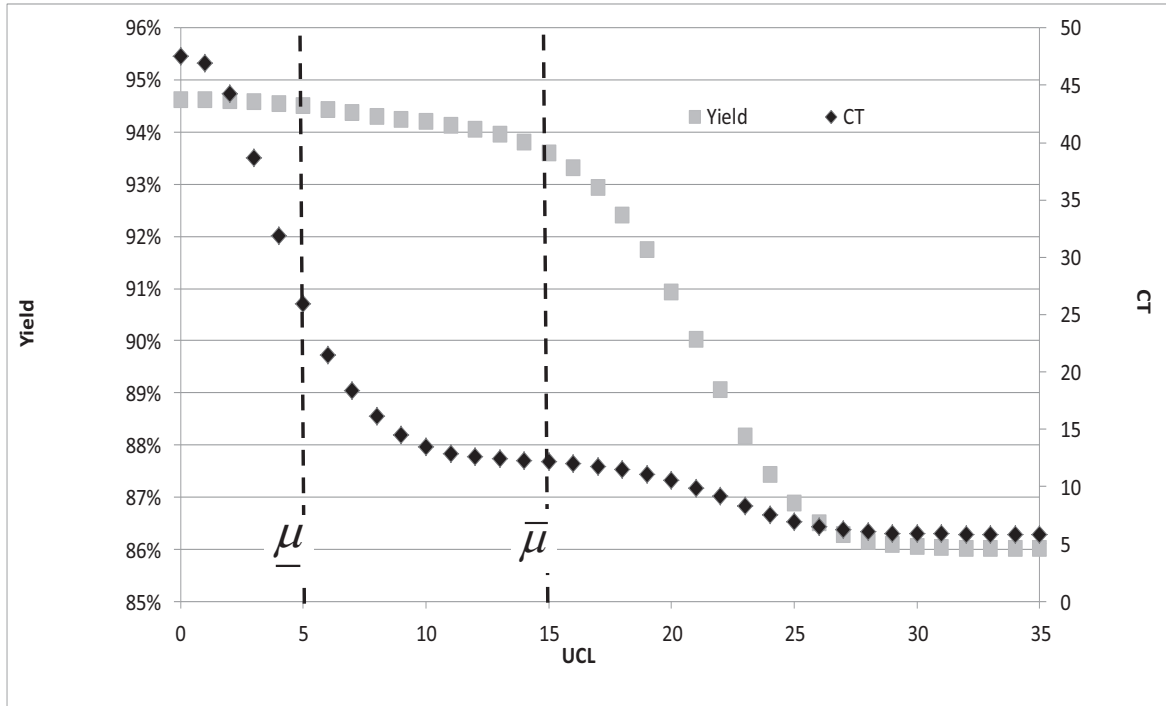


Figure 3: CT and yield as a function of UCL.

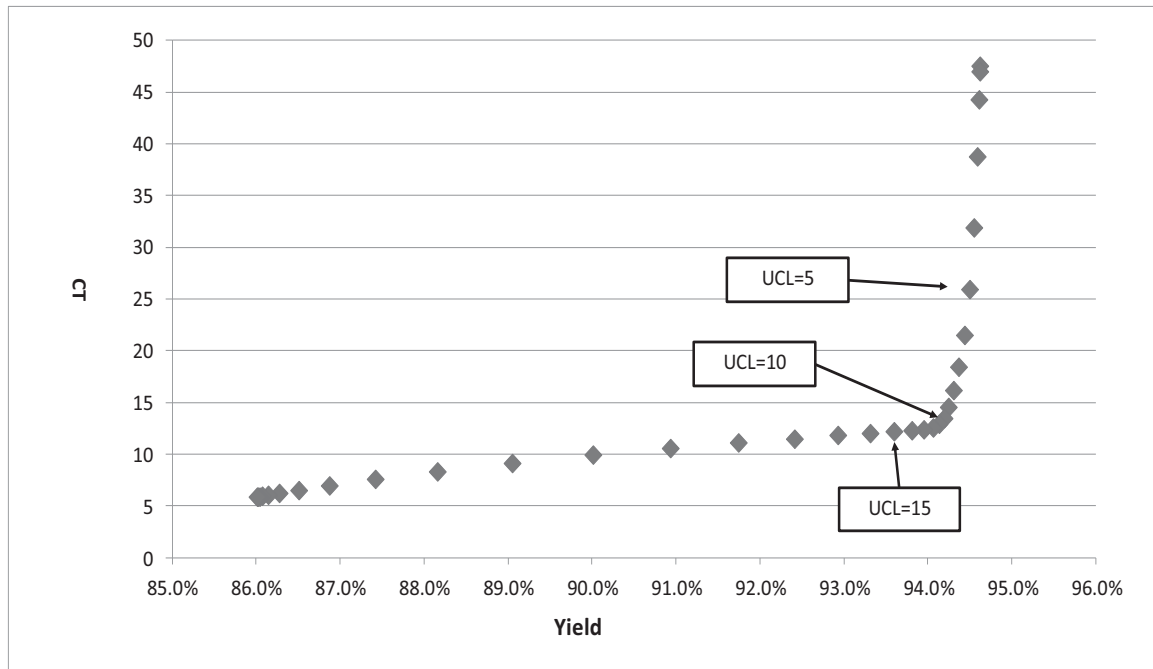


Figure 4: CT vs. yield.

From a large number of similar analyses, we have noticed that the shape of the CT-Yield graph (Figure 4) is typical of any experiment presenting a significant gap between $\bar{\mu}$ and $\hat{\mu}$. Moreover, the “knee”

of these graphs invariably takes place for UCL values in the vicinity of the middle between $\bar{\mu}$ and $\underline{\mu}$. Of course, depending on the estimated relative cost of yield and CT, setting the UCL at the mid-value between the two deposition rates may not be optimal. Nevertheless, in our model, and without any further information, they represent a “good choice”. Departing from it seems to quickly increase CT with only small Yield improvement in one direction, and to quickly decrease Yield with only small CT reduction in the other.

5 CONCLUSION

This research presents a new trade-off between CT and yield that allows control limits of the in-line inspection process to be considered as adjustable decision variables. This novel approach defies the classic assumption that control limits are predetermined according to yield requirements only. The present work is but a first step in this direction, allowing us to consider in a future work the cumulative effect of a queuing network of several process operations such as the one described herein. While we plan to make our study more applicable by considering, for example, different distributions of parameters such as the defect deposition or the stoppage process, or by assigning a metrology tool to a segment of operations instead of a single one, these new directions are expected to limit the ability of analytical tools to solve the problem, and we shall rely more heavily on simulation tools to do so.

All these models assume static policies, but we can also extend the ideas developed here to dynamic schemes, where the control limits may be adapted within a certain range, depending on the congestion status of the related segment of operations.

REFERENCES

- Bolch, G., Greiner, S., de Meer, H., Trivedi, K.S. 2006. *Queuing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. Second edition, New York.
- Colledani, M. 2008. "Integrated Analysis of Quality and Production Logistics Performance in Asynchronous Manufacturing Lines." IFAC World Congress, 1-7, Seoul, South Korea.
- Colledani, M., Tolio, T. 2011. "Joint design of quality and production control in manufacturing systems." *CIRP Journal of Manufacturing Science and Technology* 4:281-289.
- Colledani, M., Tolio, T. 2006. "Impact of quality control on production system performance." *CIRP Annals - Manufacturing Technology* 55(1):453-456.
- Cunningham, S.P., Shanthikumar, J.G. 1996. "Empirical results on the relationship between die yield and cycle time in semiconductor wafer fabrication." *IEEE Transactions on Semiconductor Manufacturing* 9(2):273-277.
- Dauzère-Pérès, S., Rouveyrol, J., Yugma, C., Vialletelle, P., C. 2011. "A Smart Sampling Algorithm to Minimize Risk Dynamically." *Advanced Semiconductor Manufacturing Conference (ASMC), 2010 IEEE/SEMI*, 307-310, San Francisco, CA.
- Khetan, S., Fowler, P. 1995. "Managing high IC yields with short cycle times". *Gallium Arsenide Integrated Circuit (GaAs IC) Symposium, 17th Annual IEEE*, 119 – 123, San Diego, CA.
- Madan, K.C., Saleh, M. F. 2001. "On an M/D/1 queue with deterministic server vacations." *Systems Science Journal* 27:107-118.
- Meyersdorf, D., Yang, T. 1997. "Cycle time reduction for semiconductor wafer fabrication facilities." *Advanced Semiconductor Manufacturing Conference and Workshop, IEEE/SEMI, TEFEN, Foster City, CA*, 18-423.
- Munga, J.N., Vialletelle, P., Crolles, Yugma, C. 2011. "Optimized management of excursions in semiconductor manufacturing." *Proceedings of the 2011 Winter Simulation Conference*, Edited by S. Jain, R.R. Creasey, J. Himmelpach, K.P. White, and M. Fu, 2100-2107, France.

- Tirkel, I., Reshef, N., Rabinowitz, G. 2009. "In-line inspection impact on Cycle Time and Yield." *IEEE Transactions on Semiconductor Manufacturing* 4(22):491-498.
- Wein, L.M. 1992. "On the Relationship between Yield and Cycle Time in Semiconductor Wafer Fabrication." *IEEE Transactions on Semiconductor Manufacturing* 2:156-158.

AUTHOR BIOGRAPHIES

MIRI GILENSION received her BSc degree in Industrial Engineering together with Teacher Certification in Science and Technology from Ort Braude College of Engineering, Israel. During her undergraduate studies she carried out research on technological innovation measurement. Miri is currently an MSc student in IE at the Faculty of Industrial Engineering and Management at the Technion – Israel Institute of Technology. This article is part of her thesis work. Miri's e-mail address is gilenson@tx.technion.ac.il.

MICHAEL HASSOUN is a lecturer at the Industrial Engineering Department at the Ariel University Center, Israel. His research interests focus on modeling and management of production systems, with a special interest in semiconductor manufacturing. He earned his PhD and MSc in Industrial Engineering from Ben-Gurion University of the Negev, Israel, and his BSc in Mechanical Engineering from the Technion, Israel. In 2009, he was a Post Doc fellow at the Electrical Engineering and Computer Science Department of the University of Michigan, USA. His e-mail address is michaelh@ariel.ac.il.

LIRON YEDIDSION is a lecturer at the Faculty of Industrial Engineering and Management at the Technion – Israel Institute of Technology. His research interests lie at discrete optimization, NP-hard problems, and Approximation algorithms. He did his PhD at Ben-Gurion University of the Negev, Israel, and his Post Doc at MIT – Massachusetts Institute of Technology. His e-mail address is lirony@ie.technion.ac.il; his web page is <http://ie.technion.ac.il/Home/Users/lirony0.html>.