Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller

SETTING QUALITY CONTROL REQUIREMENTS TO BALANCE BETWEEN CYCLE TIME AND YIELD INASEMICONDUCTOR PRODUCTION LINE

Miri Gilenson
Liron Yedidsion

Technion, Israel Institute of Technology
Haifa 32000, ISRAEL

Michael Hassoun

Ariel University
Ariel 40700, ISRAEL

ABSTRACT

We consider a semiconductor production line in which production stations are afflicted by a defect deposition process and immediately followed by an inspection step. We propose to integrate operational aspects into quality considerations by formulating a Cycle Time (CT) versus Yield trade off. We connect the two performance measures through the determination of the limit for defects at the inspection step.We extend former results to a tandem production line and present an optimal greedy algorithm that provides the Pareto-optimal set of Upper Control Limit (UCL) values for the line. The obtained model enables decision makers to knowingly sacrifice Yield to shorten CT and vice versa.

1 INTRODUCTION

Traditionally, the semiconductor industry emphasizes quality over Cycle Time (CT). Yield engineers usually set quality control requirements based on targeted Yield, basically leaving the industrial engineers to struggle for the best possible CT under these requirements. Lately, following a growing demand for products (typically memories) for which CT is of tremendous importance, manufacturers have started to reconsider this state of things. Meyersdorf and Yang (1997), as well as Khetan and Fowler (1995) present some aspects of the Yield-CT trade-off without quantifying them. More recently, Tirkel, Reshef and Rabinowitz (2009) proposed a dynamic monitoring policy instead of the traditional constant sampling. Such dynamic policies are not new, but were so far used to improve Yield (Dauzère-Pérès et al. 2010). In a similar vein, Goren and Rabinowitz (2011) suggest a model for efficient integration of Yield and CT under a combined in-line inspection and repair policy. They suggest random inspection and repair times, as well as finite queues while analyzing a queuing network model performance with the decision variable being the inspection rate.

In a former publication, Gilenson, Hassoun and Yedidsion (2012) have formulated the trade-off existing at a single station level through the determination of the control limits. Tightening the Statistical Process Control (SPC) requirements results in a better yield at the expense of a higher frequency of false alarms and superfluous machine stoppage, thus increasing CT. In the current research, we consider a multi-station tandem line and extend our previous results by finding the Pareto-optimal set of control limit values that optimize the balance between Yield and CT for the whole production process. Before doing so, in the next section, we summarize the results for the single station case (some are slightly different from Gilenson et al.2012)

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

2 SINGLE STATION SYSTEM

In the framework of this paper we model a tandem production line (Li and Meerkov 2008). Stations are connected by exactly one input and one output. In this section, we consider one of the stations in isolation and study the impact of the control limits on both Yield and CT. As illustrated in Figure 1, each production station in the network is followed by a metrology step in which the items are examined for defects through the use of SPC charts, and the decision whether or not to let the station continue producing, is taken. The capacity of the metrology stations is supposed to be infinite and there are no queues forming in front of the metrology stations. Each station behaves as a single first-come-first-serve waiting line with a single server.

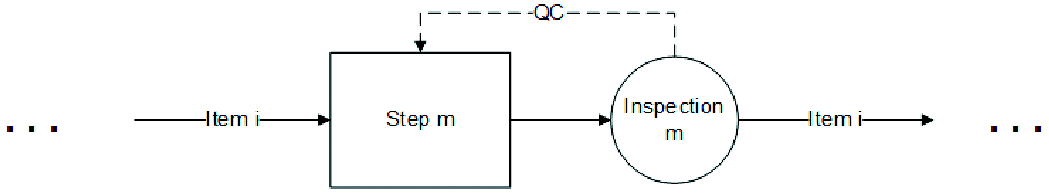

Figure 1: Illustration of an arbitrary production step m.

In our model, defects are device killers, independently of their exact location. However, the definition of a defect, regarding its size or any other characteristic, can be different at each station (in practice, certain operations are more sensitive than others).

2.1 The Control Limit's Impact on CT

In a tandem production line, the average arrival rate is equal at all stations and we denote it λ. The service duration (processing time) is a constant t_m. At the metrology station, part of each item's surface is sampled. Let us denote the sample area for station $m$by A_m; that is the proportion of sampled dice on the item. When the number of defects on the sampled area exceeds a predefined Upper Control Limit for station m (UCL_m), the station is said to be Out Of Control (OOC), and production is interrupted. Otherwise, the station is said to be In Control (IC). The all-target value for defects is obviously zero; therefore, we disregard here any type of lower control limit. Without any loss of generality, we assume the number of defects added to the sampled area of a specific item at process step m, denoted x_m, to be a Poisson process with parameter μ_m. The station is described as a two-state station, and its defect deposition rate can either be low $\underline{\mu}_m$ or high $\overline{\mu}_m$: $\overline{\mu}_m > \underline{\mu}_m$.

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

$$P(OOC_m) = 1 - P(x_m \leq UCL_m) = 1 - \sum_{k=0}^{UCL_m} \frac{(\mu_m)^k e^{-\mu_m}}{k!}$$

where$\mu_m \in \left\{\underline{\mu}_m, \overline{\mu}_m\right\}$. The inspection process is subject to errors. We denote by α_m the probability that a monitor exceeds UCL_m when the defect deposition rate is low (type 1 error), and by β_m the probability of a monitor to remain below the UCL_m when deposition rate is, in fact, high (type 2 error). We consider a sample to be IC if $x_m \leq UCL_m$ and OOC otherwise. Accordingly

$$\begin{aligned}\alpha_m = P\left(x_m > UCL_m | \underline{\mu}_m\right) &= \sum_{k=UCL_m+1}^{\infty} \frac{(\underline{\mu}_m)^k e^{-\underline{\mu}_m}}{k!} \\ &= 1 - \sum_{k=0}^{UCL_m} \frac{(\underline{\mu}_m)^k e^{-\underline{\mu}_m}}{k!}\end{aligned} \tag{1}$$

$$\beta_m = P\left(x \leq UCL_m | \overline{\mu}_m\right) = \sum_{k=0}^{UCL_m} \frac{(\overline{\mu}_m)^k e^{-\overline{\mu}_m}}{k!}. \tag{2}$$

We model the evolution of a single station over time with four states (also depicted in Figure 2):

1. The defect deposition rate is low $\left(\underline{\mu}_m\right)$ and the monitor indicates that the process is IC;
2. The defect deposition rate is low $\left(\underline{\mu}_m\right)$ and the monitor indicates that the process is OOC (type 1 error);
3. The defect deposition rate is high $\left(\overline{\mu}_m\right)$ and the monitor indicates that the process is IC (type 2 error);
4. The defect deposition rate is high $\left(\overline{\mu}_m\right)$ and the monitor indicates that the process is OOC.

We denote the probability of the deposition rate to switch from $\underline{\mu}_m$ to $\overline{\mu}_m$by p_m. Once a monitor is OOC, production is stopped and the station undergoes an inspection and, if needed, a repair. Such repairs are assumed to always be successful and inevitably bring the station back to state 1. In addition, once the deposition rate has risen, it will not go back to $\underline{\mu}_m$ unless a repair is conducted. Under these assumptions, we can now present the station as a Markov Chain, as shown in Figure 2:

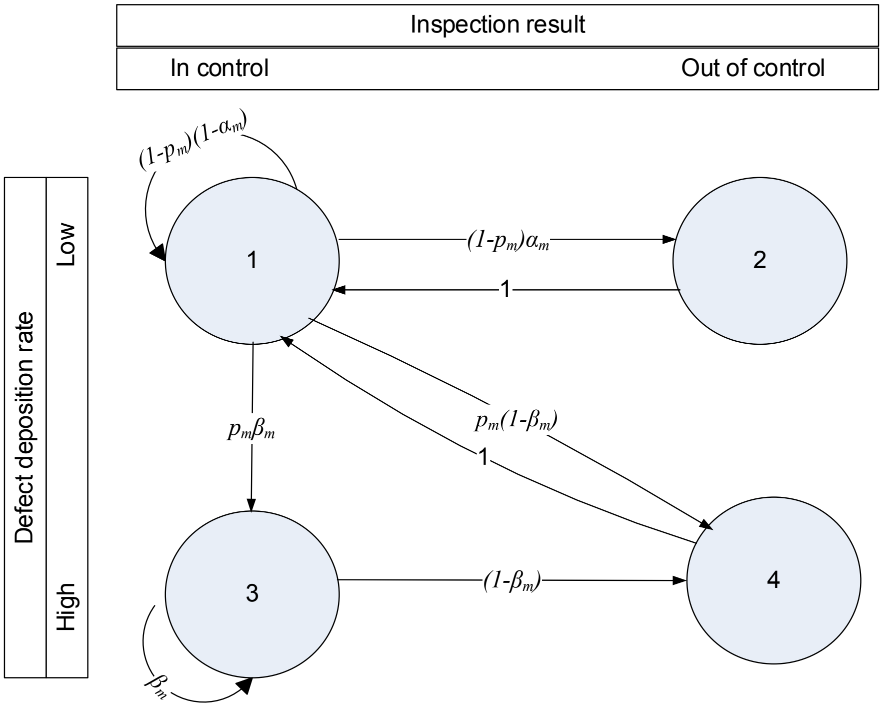

Figure 2: Station m as a four-state Markov chain.

The number of production cycles between two consecutive deviations from UCL_m is known as the Average Run Length (ARL) and was found to be (Gilenson et al.2012):

$$ARL_m = \frac{1 + p_m \frac{\beta_m}{1-\beta_m}}{1 - (1-p_m)(1-\alpha_m)}. \tag{3}$$

At each production cycle, the probability for the station to be OOC and consequently stopped for inspection is given by:

$$p_m^{Stop} = \frac{1}{ARL_m} = \frac{1 - (1-p_m)(1-\alpha_m)}{1 + p_m \frac{\beta_m}{1-\beta_m}} \tag{4}$$

In our model, the service time, denoted S_m accounts for both the processing time and repair duration (it is often referred to as "effective service time" or "effective processing time", (see Hopp and Spearman 2008). We have:

$$S_m = \begin{cases} t_m + V_m & \text{with probability } p_m^{Stop} \\ t_m & \text{with probability } 1 - p_m^{Stop} \end{cases}$$

where the vacation duration, V_m, is a state-dependent random variable. With the stationary probability vector for station *m*

$$\vec{\pi}^m = \begin{pmatrix} \dfrac{1}{1+(1-p_m)\alpha_m+\frac{p_m}{(1-\beta_m)}}, \dfrac{(1-p_m)\alpha_m}{1+(1-p_m)\alpha_m+\frac{p_m}{(1-\beta_m)}}, \\ \dfrac{\frac{p_m\beta_m}{(1-\beta_m)}}{1+(1-p_m)\alpha_m+\frac{p_m}{(1-\beta_m)}}, \dfrac{p_m}{1+(1-p_m)\alpha_m+\frac{p_m}{(1-\beta_m)}} \end{pmatrix}$$

the first two moments of the vacation duration distribution function are:

$$v_m = E[V_m] = vi_m + \frac{\pi_4}{\pi_2+\pi_4} vr_m \tag{5}$$

$$w_m = Var[V_m] = vi_m^2 + \left(\frac{\pi_4}{\pi_2+\pi_4}\right)^2 vr_m^2. \tag{6}$$

The resulting service time does not meet any familiar distribution. Moreover, the arrival process characteristics to a random station within the network are a priori, unknown. Following these two understandings, we tackled the determination of the expected value of the station CT by applying a $G/G/1$ queuing model. We denote the expectancy of the CT by CT_m and the expectancy of waiting and service times by $E[W_m]$ and $E[S_m]$ respectively. We have:

$$CT_m = E[W_m] + E[S_m]. \tag{7}$$

According to Kingman's Bound (Kingman 1962) the mean waiting time in a $G/G/1$ queue is approximately

$$E[W_m] = \frac{\rho_m}{1-\rho_m}\frac{CA_m^2+CS_m^2}{2}E[S_m], \tag{8}$$

where CA_m and CS_m are the station's inter-arrival and service times coefficient of variation (CV), and ρ_m is the station's utilization that is proportional to both the arrival rate to the system and the service time:

$$\rho_m = \lambda E[S_m]. \tag{9}$$

The effective service time squared CV is found using

$$\begin{gathered} E[S_m] = t_m + p_m^{Stop} v_m \\ Var[S_m] = \left(2p_m^{Stop} - \left(p_m^{Stop}\right)^2\right) w_m \end{gathered} \tag{10}$$

whereq_m^{Stop} is the station's probability being IC and p_m^{Stop} is the (complementary) probability of the station being OOC:

$$CS_m^2 = \frac{Var[S]_m}{E^2[S]_m} = \frac{\left(2p_m^{Stop} - \left(p_m^{Stop}\right)^2\right) w_m}{\left(p_m^{Stop} v_m + t_m\right)^2}, \tag{11}$$

and the CT_m is found by substituting $\vec{\pi}^m$ and (8)–(10) into (7):

$$CT_m = \frac{\rho_m}{1-\rho_m}\frac{CA_m^2+CS_m^2}{2}E[S_m] + E[S_m]$$

$$= \left(t_m + p_m^{Stop} v_m\right) \left[1 + \frac{\lambda\left(t_m + p_m^{Stop} v_m\right)}{1 - \lambda\left(t_m + p_m^{Stop} v_m\right)} \left(\frac{1}{2} CA_m^2 + \frac{\left(2p_m^{Stop} - \left(p_m^{Stop}\right)^2\right) w_m}{2\left(p_m^{Stop} v_m + t_m\right)^2} \right) \right]. \tag{12}$$

The explicit form of CT_m is found by substituting (4) into (12).

2.2 The Control Limit's Impact on Yield

In this section we formulate the impact of UCL_m on Yield expectancy. In the single station case, the die Yield at station m, denoted by $yield_m$, is the portion of dice that were not contaminated during process step m only. The defect deposition over A_m is a Poisson process with rate $\mu_m \in \left\{\underline{\mu}_m, \overline{\mu}_m\right\}$. Assuming the defects to be uniformly scattered, the deposition process over the entire item is Poisson too with rate $\frac{\mu_m}{A_m}$.

To find the Expectancy of the Defect deposition Rate at station m, denoted EDR_m, we have to find whether the proportion of time defects' deposition rate is either low or high respectively.Using $\vec{\pi}^m$ we obtain that:

$$EDR_m = \frac{(\pi_1^m + \pi_2^m)\underline{\mu}_m + (\pi_3^m + \pi_4^m)\overline{\mu}_m}{A_m} = \frac{(1 + (1 - p_m)\alpha_m)\underline{\mu}_m + \left(\frac{p_m}{1-\beta_m}\right)\overline{\mu}_m}{A_m\left(1 + (1 - p_m)\alpha_m + \frac{p_m}{1-\beta_m}\right)}. \tag{13}$$

Clearly, the value of UCL_m, by affecting the sensibility to defects, will also affect the stationary probabilities of $\vec{\pi}^m$, and thus EDR_m. We denote the number of devices on a single item by ND. The probability of a defect to destroy any specific device on the item is $\frac{1}{ND}$ while the probability for a device to stay functional after the deposition of a single defect is $\left(1 - \frac{1}{ND}\right)$. The station Yield is:

$$yield_m = \left(\frac{ND - 1}{ND}\right)^{EDR_m} = \left(\frac{ND - 1}{ND}\right)^{\frac{(1+(1-p_m)\alpha_m)\underline{\mu}_m + \left(\frac{p_m}{1-\beta_m}\right)\overline{\mu}_m}{A_m\left(1+(1-p_m)\alpha_m+\frac{p_m}{1-\beta_m}\right)}}. \tag{14}$$

2.3 Yield to CT Trade-off for a Single Station

Both CT_m and $yield_m$ are monotone decreasing functions of UCL_m, and therefore bijective. Even though there is no natural limitation on the UCL_m values, we keep them within the range $\left[\left\lceil\underline{\mu}_m\right\rceil, \left\lfloor\overline{\mu}_m\right\rfloor\right]$ outside of which the monitoring becomes almost ineffective. Accordingly, for any given combination of arrival rate λ and a station's parameters, we can generate the trade-off curve of $CT_m$to $(1 - yield_m)$. To keep things intuitive, we chose to use $(1 - yield_m)$ instead of $yield_m$, so that the trade-off occurs between two competing minimization measures.

Since UCL_m values are discrete, then plotting CT_m and $(1 - yield_m)$ by changing the UCL_m values only, gives us a discrete set of points, denoted by Ω_m, on the CT_m to $(1 - yield_m)$ plane, each point associated with a unique UCL_m value.

Due to the complexity of both CT_m and $yield_m$ functions, we have not been able to demonstrate that set Ω_m is convex. We conducted an intensive numerical study and could not find a single instance to contradict this assumption. Yet, in case a setting in which a trade-off curve that is not convex does exist, we propose to use a convex hull, thus replacing concave points by a linear combination of their framing points, instead of the original curve (the algorithm used to generate this new curve is rather straightforward and thus omitted from this paper).

3 MULTI-STATION SYSTEM

In this section we extend our results to deal with a multi-station tandem line. First we show that both CT and yield are separable and find the marginal contribution of each station to the line's CT and yield respectively. We then show that the marginal contribution of each station to the CT is convex to its contribution to the line's yield. Since these marginal changes are caused by adjusting the UCL which is discrete, then the change in both CT and yield is discrete as well. Accordingly, we devised a greedy algorithm to construct the trade-off curve between CT and yield by choosing to adjust, in each step, the UCL of the station for which we get the most efficient CT to Yield marginal change.

3.1 CT Expectancy in a Multi-Station System

Due to the dependency between the stations induced by the items' flow, calculating the CT expectancy for multiple stations is more complex than its counterpart in the single station case. In a tandem production line, the arrival rate of station m equals the departure rate of the station $m - 1$. Let us denote the squared CV of inter-departure time from station $m$by CD_m^2. We have:

$$CA_m^2 = CD_{m-1}^2. \tag{15}$$

Marshall (1968) approximates the inter-departure square CV time by:

$$CD_m^2 \approx (1 - \rho_m^2)CA_m^2 + \rho_m^2 CS_m^2 \quad . \tag{16}$$

By recursively applying this approximation to stations in our system we get:

$$CT_m = \left(t_m + p_m^{Stop} v_m\right) * \left[1 + \frac{\lambda\left(t_m + p_m^{Stop} v_m\right)}{1 - \lambda\left(t_m + p_m^{Stop} v_m\right)} \frac{\sum_{j=1}^{m-1}\left(\prod_{i=j}^{m-1}\left(1 - \lambda^2\left(t_i + p_i^{Stop} v_i\right)^2\right)\right)\left(\lambda^2\left(2 - p_{j-1}^{Stop}\right)p_{j-1}^{Stop} w_{j-1}\right)}{2} + \frac{\lambda^2\left(2 - p_{m-1}^{Stop}\right)p_{m-1}^{Stop} w_{m-1}}{2} + \frac{\left(2 - p_m^{Stop}\right)p_m^{Stop} w_m}{2\left(t_m + p_m^{Stop} v_m\right)^2}\right. \tag{17}$$

Which leads to the entire line CT, denoted by CT, by:

$$CT = \sum_{m=1}^{M} C\,T_m \tag{18}$$

3.2 The Yield Expectancy in a Multi-Station System

In the multi station case, the Yield is the portion of dice undamaged throughout the whole process. The defect deposition rate at each station is an independent random variable; consequently $YIELD$ is the product of the probabilities that a die survives all stations $m = 1, \ldots, M$:

$$YIELD = \prod_{m=1}^{M} y\,ield_m = \left(\frac{ND-1}{ND}\right)^{EDR_1} \left(\frac{ND-1}{ND}\right)^{EDR_2} \cdots \left(\frac{ND-1}{ND}\right)^{EDR_M}$$
$$= \left(\frac{ND-1}{ND}\right)^{\sum_{m=1}^{M} EDR_m}. \tag{19}$$

3.3 Yield to CT Trade-off for a Multi-Step System

As the CT and $YIELD$ for the multi-station system are well defined, we present a greedy algorithm to create the CT to a $(1 - YIELD)$ trade-off curve for the Pareto-optimal set of the entire production line. This trade-off curve allows decision makers to maintain a balance between CT and $YIELD$ by controlling the UCL_ms of the inspection process.

Let us define POS as the Pareto-Optimal Set of ordered pairs $(CT, 1 - YIELD) \in R^2$ each of which represents a point on the CT to a $(1 - YIELD)$ trade-off curve. Initially, $POS = \{\emptyset\}$. The algorithm adds points to POS one at a time and constructs the efficient frontier of CT to a $(1 - YIELD)$ trade-off by connecting every two consecutive points. But first we have to define all the relevant UCL_mpoints. We define a set of only the UCL_m values associated with Pareto optima points left on the convex hull (after filtering out all the non-convex ones). We denote this set by J'_m. Since the size of each set is unique we define an index i_m for the elements of the J'_m set.

In order to create the POS we need to calculate the values of $CT$and $(1 - YIELD)$; therefore the UCL_m value for each station in the network should be defined. Let us define the set of all $UCL_{i_m,m}$ values for all stations in the system by:

$$UCLS = \left\{UCL_{i_m,m} \colon m = 1, \ldots, M, i_m = 1, \ldots, |J'_m|\right\}.$$

The UCL_m values for each station vary within $\left[\left\lceil \underline{\mu}_m \right\rceil, \left\lfloor \overline{\mu}_m \right\rfloor\right]$. The algorithm initiates when the UCL_m are initialized to $UCL_{1,m} \left(= \left\lceil \underline{\mu}_m \right\rceil\right)$for $m = 1, \ldots, M$. Thus, the $(1 - YIELD)$ is minimized while the CT is maximized within all Pareto-optimal points. The algorithm progresses iteratively by choosing, at each step, the station for which increasing $UCL_{i_m,m}$ to $UCL_{i_m+1,m}$ gives us the lowest CT to a $(1 - YIELD)$ ratio among all the stations in the system.

Let us denote the marginal change in CT_m and $yield_m$ occurring due to a change in $UCL_{i_m,m}$ by:

$$\varepsilon_{i_m,m} =: \left(CT_m | UCL_{i_m+1,m}\right) - \left(CT_m | UCL_{i_m,m}\right)$$

and

$$\xi_{i_m,m} =: \left(1 - yield_m | UCL_{i_m+1,m}\right) - \left(1 - yield_m | UCL_{i_m,m}\right),$$

respectively. The marginal change in the ratio between CT and $(1 - YIELD)$ due to the marginal change in CT_m and $yield_m$ is:

$$\frac{[CT_1 + \ldots + CT_m + \ldots + CT_M] - \left[CT_1 + \ldots + \left(CT_m - \varepsilon_{i_m,m}\right) + \ldots + CT_M\right]}{[1 - yield_1 \cdot \ldots \cdot yield_m \cdot \ldots \cdot yield_M] - \left[1 - yield_1 \cdot \ldots \cdot \left(yield_m - \xi_{i_m,m}\right) \cdot \ldots \cdot yield_M\right]}$$

$$= \frac{\varepsilon_{i_m,m}}{\frac{-\xi_{i_m,m}}{yield_m} YIELD}. \tag{20}$$

In each iteration of our algorithm we look for the change in one of the $UCL_{i_m,m}$ that would result in the steepest decline on the CT to a $(1 - YIELD)$ trade-off curve, i.e., the most efficient change. Since in each point in the algorithm we compare the marginal change in CT and $(1 - YIELD)$ with respect to the same Pareto-optimal point (the last Pareto-optimal point that was added to POS), then the $YIELD$ in Eq. (20) could be considered as a constant, and thus omitted from the comparison. We define $\theta_{i_m,m}$ as the marginal ratio between CT and $(1 - YIELD)$ for two consecutive $UCL_{i_m,m}$ if there are at least two arguments in J'_m that were not examined yet; otherwise, $\theta_{i_m,m}$ receives the value zero.

$$\theta_{i_m,m} = \begin{cases} -\dfrac{\varepsilon_{i_m,m} yield_m}{\xi_{i_m,m}} & \text{for } i_m < |J'_m| \\ 0 & \text{otherwise} \end{cases}. \tag{21}$$

Finally, in order to extract the UCL_m for each station given a specific point in POS on the Pareto-optimal curve we define set Ψ. Each element in Ψ is comprised of an ordered pair $(CT, 1 - YIELD)$ of POS, and of a vector $UCLS$. That is: $\Psi = \{CT, 1 - YIELD, UCLS\}$. Accordingly, each point on the Pareto-optimal curve is associated with a specific setting of UCL_m for each station.
The algorithm input is: λ, ND and $\{\mu_m, p_m, t_m, v_m\}$ for $m = 1, \dots, M$.

Algorithm 1. Constructing the CT to a $1 - YIELD$ trade-off curve

```
Initialization: Set $UCLS = \{UCL_{i_m,m}: m = 1,\dots,M, i_m = 1,\dots,|J'_m|\}$;
        $\vec{I} = (i_1,\dots,i_M) = (1,\dots,1)$; $\Psi = \emptyset$; and $POS = \emptyset$.
Step 1: Calculate $\alpha_m, \beta_m$ probabilities for station $m = 1,\dots,M$ using Eqs. (1) and (2) for $UCL_{1,m}$.
Step 2: Calculate $CT_m$ using Eq. (17) and $(1 - yield_m)$ using (14) for $m = 1,\dots,M$.
Step 3: Calculate $CT$ using Eq. (18).
        Calculate $(1 - YIELD)$ using Eq. (19).
        Set $POS = POS \cup (CT, 1 - YIELD)$.
        Set $\Psi = \Psi \cup (CT, 1 - YIELD, UCLS)$
Step 4: Set $m = j = 1,\dots,M \arg\min\left(\theta_{i_j,j}\right)$, where $\theta_{i_j,j}$ is calculated using Eq. (21) and $i_j = \vec{I}(j)$.
Step 5: If $\theta_{i_m,m} = 0$ then stop. Otherwise set $\vec{I}(m) = \vec{I}(m) + 1$, calculate $\alpha_m, \beta_m$ probabilities for
        station $m$ using Eqs. (1) and (2) for $UCL_{i_m,m}$, where $i_m = \vec{I}(m)$.
        Calculate $CT_m$ using Eq. (17) and $(1 - yield_m)$ using (14) and return to Step 3.
```

The algorithm produced the POS, that is the CT to $(1 - YIELD)$ sorted set of points representing the trade-off between the two measures. To generate the convex trade-off curve, one can connect each pair of consecutive points in the set.
We illustrate the algorithm results in the following example for an arbitrary set of parameters: $\{\mu_m, p_m, t_m, vr_m, vi_m\}$ for $m = 1,\dots,5$ as follows:

Table 1: Stations' parameter data for the numerical example

Station	$\underline{\mu}_m$	$\bar{\mu}_m$	p_m	t_m	vr_m	vi_m
1	1	5	0.99	4	20	5
2	3	9	0.95	5	15	5

3	4	11	0.98	5	15	3
4	3	7	0.995	10	30	10
5	6	12	0.95	5	30	3

and present the Pareto-optimal trade-off curve for this set in Figure 4.

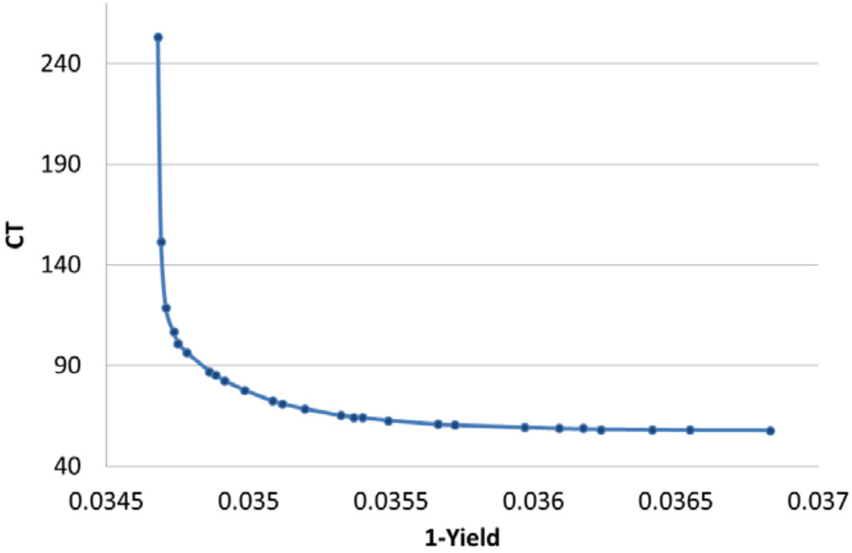

Figure 4: An example of CTtoa(1-YIELD) Pareto-optimal curve for a multi-station system.

Figure 4 describes the CT improvement in terms of $YIELD$ loss and vice versa for the five stations' tandem line. The algorithm chooses, at each step, the most efficient CT_m to a $(1 - yield_m)$ ratio among all stations; therefore the CT to a $(1 - YIELD)$ trade-off curve is in fact Pareto-optimal for it is constructed according to the functions for each of the stations that are convex.

We observe that the CT does not converge to the sum of the direct processing times, which is 24 time units, but is somewhere around 44 time units. The reason for this is obviously the additional waiting time. Similar insights are relevant to the $YIELD$ i.e., even when the $UCLS$ is set to $\underline{\mu_m}_m$ for all stations and any deviation from the UCL_m is revealed almost immediately, we do not observe a perfect $YIELD$.

4 SUMMARY

In this paper we have presented a trade-off between the CT and the Yield that allows control limits of the in-line inspection process to be considered as adjustable decision variables. This innovative approach defies the classic assumption that control limits are predetermined in accordance with the Yield requirements only and allows decision makers to knowingly sacrifice Yield in order to improve the CT or vice versa, in order to maximize their profits.

We have extended our former results to the multi-station tandem line. The overall CT was calculated using the $G/G/1$ queuing network approximation, while the final Yield was calculated from the single stations' Yields.

To formulate the Pareto-optimal CT versus Yield trade-off curve of the multi-station tandem line we have presented an optimal greedy algorithm that recommends a set of upper control limits for each point on a Pareto-optimal curve. This technique allows us to recommend to decision makers the best policy of setting the control limits to balance CT and Yield.

In order to better adapt this model to the actual semiconductor manufacturing environment, several important extensions need to be applied to our model: The capacity of the metrology tools, whose prices have skyrocketed in the last years, cannot be dismissedanymore. Also, there is a clear need to represent their assignment to a segment of operations instead of a single one.One could also consider different types of stoppages and their approximated impact on CT (Wu, 2014). A vacation process additional to the repair process may also be considered.That would represent the high level of tool sharing and re-entrance typical in fabs. These new directions might limit the ability of analytical tools to solve the problem, and one shall rely more on simulation tools to do so.

In a different vein, very recent advances in the field of tandem queue analysis (Wu et al. 2013) offer new opportunities for more accurate approximations of the solution to our problem in its current form.

REFERENCES

Buzacott, J.A., Shanthikumar, J.G. 1993. *Stochastic models of manufacturing systems*. Prentice Hall.

Dauzère-Pérès, S., Rouveyrol, J., Yugma, C., Vialletelle, P., 2010. "A smart sampling algorithm to minimize risk dynamically." In *Proceedings of the 2010 Advanced Semiconductor Manufacturing Conference (ASMC), 2010 IEEE/SEMI, IEEE*, pp. 307–310.

Gilenson, M., Hassoun, M., Yedidsion, L. 2012. "Setting quality control requirements to balance cycle time and yield the single machine case." In *Proceedings of the 2012 Winter Simulation Conference,* edited by C. Laroque, J.Himmelspach, R.Pasupathy, O.Rose, and A.M.Uhrmacher, 1-9, Berlin, Germany. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Goren, S., Rabinowitz, G. 2011. "Combined Yield and flow time evaluation in production queuing network models." In *Proceedings of the 5th Israeli Industrial Engineering and Management Research Meeting.*

Hopp, W.J., Spearman, M.L. 2008. *Factory physics*. McGraw-Hill Irwin.

Khetan, S., Fowler, P. 1995. "Managing high IC Yields with short cycle times." In *Gallium Arsenide Integrated Circuit (GaAs IC) Symposium, 1995.Technical Digest 1995, 17th Annual IEEE,* pp. 119–123. IEEE,1995.

Kingman, J.F.C. 1962. "Some Inequalities for the Queue GI/G/.".*Biometrika* 49: pp. 315–324.

Li J., Meerkov S. M. 2008. *Production Systems Engineering*.Springer.

Marshall, K.T. 1968." Some inequalities in queuing." *Operations research*,16, pp. 651-668.

Meyersdorf, D., Yang, T. 1997."Cycle time reduction for semiconductor wafer fabrication facilities." In *Proceedings of the 1997 Advanced Semiconductor Manufacturing Conference and Workshop, 1997. IEEE/SEMI, IEEE*: pp. 418–423.

Tirkel, I., Rabinowitz, G. 2011. "Quality performance modeling in a deteriorating production system with partially available inspection." In *Proceedings of the Operations Research 2010: Selected Papers of the Annual International Conference of the German Operations Research Society,* Springer, pp. 397-402.

Tirkel, I., Reshef, N., Rabinowitz, G. 2009. "In-line inspection impact on cycle time and yield." *IEEE Transactionson Semiconductor Manufacturing,* 22: pp. 491–498.

Wu, K. 2014. "Classification of queuing models for a workstation with interruptions: a review." *International Journal of Production Research*,52, pp. 902-917.

Wu, K. and McGinnis, L. 2013. "Interpolation approximations for queues in series." *IIE transactions*,45, pp. 273-290.

AUTHOR BIOGRAPHIES

MIRI GILENSON received her MSc degree in Industrial Engineering from the Technion – Israel Institute of Technology. This article is part of her thesis work toward her MSc degree. Miri is currently aPh.D. student in IE at the Faculty of Industrial Engineering and Management at the Technion – Israel

Institute of Technology. Her research deals withdue-date assignment using Robust Optimization techniques. Her e-mail address isgilenson@tx.technion.ac.il.

MICHAEL HASSOUN is a lecturer in the Industrial Engineering Department at the Ariel University, 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.