Reprinted from

EUROPEAN
JOURNAL
OF OPERATIONAL
RESEARCH

European Journal of Operational Research 80 (1995) 308–327

The impact of defects on a process with rework

Saligrama R. Agnihothri *, Ron S. Kenett

School of Management, Binghamton University, Binghamton, NY 13902-6000, USA

N·H

ELSEVIER

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH (EJOR)

Editorial policy
Operational Research (OR) involves the application of scientific methods to the management of complex systems of people, machinery, materials, money and information. It seeks to produce an understanding of managerial problems and to develop models which will enable the consequences of decisions to be investigated. OR methods have already had widespread application and success in many areas of business, industry and government, and their use is rapidly increasing in social systems and in the international arena. The European Journal of Operational Research (EJOR) is intended to strengthen these advances by publishing high-quality, original papers that contribute to the practice of decision making, within and beyond Europe, irrespective of whether their content describes an application or a theoretical development. A special call is made for case study papers showing OR in action in firms and other organizations. Mathematical or computer-science contributions should be capable of being used in the foreseeable future.

EJOR has regular, special and feature issues. Almost all issues contain a leading invited review explaining the developments in an OR topic over the past five years to the general OR reader. In addition to the other five types of contribution described in the Instructions to Authors at the end of this issue, EJOR includes OR software news (including ORSEP) and reviews of books published recently. For further information, see the Instructions to Authors at the end of each issue. Authors of contributions are asked to read these Instructions before sending four copies to any of the editors at the address above.

Publication information. *European Journal of Operational Research* (ISSN 0377-2217) is sponsored by the Association of European Operational Research Societies within IFORS. For 1994 volumes 72–79 are scheduled for publication. Subscription prices are available upon request from the publisher. Subscriptions are accepted on a prepaid basis only and are entered on a calendar year basis. Issues are sent by surface mail except to the following countries where air delivery via SAL is ensured: Argentina, Australia, Brazil, Canada, Hong Kong, India, Israel, Japan, Malaysia, Mexico, New Zealand, Pakistan, P.R. China, Singapore, South Africa, South Korea, Taiwan, Thailand, U.S.A. For all other countries airmail ratesare available upon request. Claims for missing issues must be made within six month of our publication (mailing) date. Please address all your requests regarding orders and subscription queries to: Elsevier Science B.V., Journal Department, P.O. Box 211, 1000 AE Amsterdam, Netherlands. Tel.: 31-20-5803642, fax: 31-20-5803598. Advertising rates are available upon request. Mail orders and inquiries to: Elsevier Science B.V., Advertising Department, P.O. Box 211, 1000 AE Amsterdam, Netherlands.

US MAILING NOTICE – *European Journal of Operational Research* (0377-2217) is published semi-monthly by Elsevier Science B.V. (Molenwerf 1, Postbus 211, 1000 AE Amsterdam). Annual subscription price in the USA US$ 2002.00 (US$ price valid in North, Central and South America only), including air speed delivery. Second class postage paid at Jamaica, NY 11431.
USA POSTMASTERS: Send address changes to: *European Journal of Operational Research*, Publications Expediting, Inc., 200 Meacham Avenue, Elmont, NY 11003. Airfreight and mailing in the USA by Publications Expediting.

Published semimonthly 0377-2217/94/$07.00 Printed in the Netherlands

ELSEVIER

European Journal of Operational Research 80 (1995) 308–327

EUROPEAN
JOURNAL
OF OPERATIONAL
RESEARCH

The impact of defects on a process with rework

Saligrama R. Agnihothri *, Ron S. Kenett

School of Management, Binghamton University, Binghamton, NY 13902-6000, USA

Abstract

In this paper, we attempt to quantify the impact of defects on various system performance measures for a production process with 100% inspection followed by rework. In the literature, the number of defects (feedback loops) is assumed to be a random variable having a geometric distribution. We model the number of defects as a random variable having any general discrete distribution, and investigate the impact of the defect distribution on system performance measures such as yield, production lead time, and work-in-process inventory. We provide management guidelines for short term control decisions such as identifying potential bottlenecks under increased workloads and allocating additional resources to release bottlenecks. In order to meet the long term goal of continuously decreasing defect levels, we propose a budget allocation method for process improvement projects.

1. Introduction

Although ‘doing it right the first time’ is an important goal to pursue, defects and rework are common occurrences in a manufacturing process. The impact of this rework on the entire production process is, in general, not well understood. In this paper we will examine how the pattern of defects affects the performance of a process with 100% inspection followed by rework. Specifically, we investigate the behavior of production lead time, work-in-process inventory, yield, and material handling costs for various defect levels in the manufacturing phase.

As a concrete example, we consider a situation encountered in the production floor of a printed circuit board manufacturing industry. We consider a wave soldering process followed by a testing machine. In the wave soldering process, printed circuit boards are passed over a solder bath to attach previously assembled components to the board (see Figure 1 for a schematic representation). After soldering, the boards are tested on an In-Circuit-Tester machine in order to detect possible defects. There are typical defect categories such as solder bridges, wrong component, missing component, and

---

* Corresponding author.

SSDI 0377-2217(94)00173-A

defective component. Each class of defects is diagnosed separately and is corrected at the rework center. After fixing all the defects in a defect class, the printed circuit board goes back to the test center for further testing. Thus, there is one rework loop per class of defects. If defects are discovered one at a time, defect classes consist of individual defects and the number of rework cycles equals the number of defects.

Because the wave soldering process is not perfect, it generates defects and hence waste due to inspection, rework, scrap, yield loss, additional material handling costs, and excessive production delays. One of the direct consequences of production delays due to rework is that the feedback necessary for process control (to control the defect level) and process improvement takes too long to reach its destination so that it looses any effective value. In the electronics industry costs due to such defects amount to 10–25% of total sales. Typical numbers in the industry for the percent of boards found defective range from 1–30%. Soldering defects are usually assessed in defects per million solder points and typical numbers for defect levels range from 5 to 5000 parts per million (ppm). Reductions in these numbers directly increase revenues and improve a company's competitive position.

Our objective in this paper is to investigate the impact of defects generated by the wave soldering machine on the entire manufacturing process. In the literature, the uncertainty in the defect level is modelled by assuming that the number of defects (rework or feedback loops) follows a geometric distribution, that is, after every test the probability that the product exits the system without having to be reworked is assumed to be a constant p, independent of the number of earlier feedback loops. We model the number of defects as a random variable having a general discrete distribution and quantify the impact of the defect distribution on the system performance measures such as yield and production lead time. The objective is to derive qualitative insights on the problem.

Our main them is that defect level is the leading variable in a company's short term and long term performance. We make the distinction between short term control and long term planning. For short term control, we show how to identify potential bottlenecks that will develop under increased workloads and determine how to release such bottlenecks by adding test equipment or rework stations. In the long term we can decrease the defect levels by activities such as redesigning the product or initiating a collaborative program with suppliers. Assuming that different products with varying defect levels are produced on the same machine, we need to create a priority list for process improvement projects. We propose a method for allocating a given budget for process improvements between such projects.

Several authors have studied queues with delayed feedback while modelling computer networks. Foley and Disney (1983) develop a model by assuming that the probability of a unit feeding back after service depends only on the queue lengths at the manufacturing and rework centers and the service time at the manufacturing center. Queues with Bernoulli feedback (where probability of feeding back is p) is studied by Disney and Hannibalsson (1977) and others. Seidmann and Nof (1985) consider a unitary manufacturing with Bernoulli feedback. In a unitary cell, a unit may recirculate between manufacturing and rework stations. Here, rework is an integral part of the main process. Their results are further extended to consider batch production (Seidmann et al., 1985) and to include the case when the feedback probabilities are unequal and depend on the number of earlier feedbacks (Seidmann and Nof, 1989). Agnihothri (1987) studied the impact of an imperfect server on a service system. In an assembly line setting, Robinson et al. (1990) study the impact of different rework strategies (such as on-line vs off-line rework) on system performance measures such as work-in-process inventory and throughput rate.

The next section presents the model and provides a summary and interpretations of the results obtained. This section also discusses the qualitative implications of our analysis for short term control and long term planning, and simulation results to test the sensitivity of some model assumptions. Mathematical derivations are provided in Section 3 and Section 4 summarizes the paper.

2. Qualitative results

2.1. The model

The basic manufacturing system is illustrated in Fig. 1. The system consists of three work centers. A manufacturing center (which is the soldering center in our case) denoted by M, a test center denoted by T, and a rework center denoted by R. The manufacturing process is not perfect. Errors can be made during this process. The total number of defects generated during the manufacturing process is a random variable. The products are tested at the work center (WC) T for possible defects. If a defect is found, the product moves on to a rework station denoted by R where the defect is corrected, and returns the test center for further testing. Here the defects are defined and classified in terms of the rework cycles. That is, all defects which can be diagnosed by one passage through the test center are grouped togegther and are fixed simultaneously as one defect class. We make the following assumptions:

- The defect classes are detected one at a time.
- The number of defects generated in WC M, denoted by N, is assumed to have a discrete probability distribution.
- The manufacturing process M is under statistical control and hence defects are independent of each other.
- The parts arrival rates to WC i is denoted by λ_i, $i \in G = \{\mathrm{M, T, R}\}$, the set of work centers. For notational convenience, we assume $\lambda_{\mathrm{M}} = \lambda$. In addition, the arrival process to work center M follows a Poisson process. This is a reasonable assumption, and is observed to be true when printed circuit boards of different types arrive in batches to be soldered.
- Each work center can consist of multiple parallel servers. The number of servers in WC i is denoted by c_i where $i \in G$. Without loss of generality and for the sake of convenience we assume $c_{\mathrm{M}} = 1$. Here, c_{T} and/or c_{R} could be ∞ in which case we will have an ample server system.
- The service times in WC i, denoted by S_i, are assumed be independent random variables with exponential distributions with means $\overline{S}_i = 1/\mu_i$, $i \in G$. This assumption approximates the actual situation where the service times in centers M and T are nearer to deterministic distribution and service times in rework center may depend on the type of the defect class. Hence the results obtained by this model provides an upper bound to the exact results, as confirmed by simulation.
- A first-come-first-served (FCFS) service discipline is used in all work centers.
- There is ample waiting room capacity (unlimited buffer) in each work center.
- The servers at the test and rework centers are assumed to be perfect. That is, the test identifies all defect classes correctly and no new defects are introduced while testing. Similarly, rework center corrects all defects identified in the current cycle and does not introduce new ones.
- The system is under steady state.

We denote the mathematical expectation (mean) of random variable (r.v.) X by $\overline{X}$. The probability density function (p.d.f.) of a continuous r.v. X is denoted by $f_X(\cdot)$. Since we are interested in measuring the effect of defects introduced at M, we will be following the product from the time it leaves M until it leaves the system. That is, we will be concentrating on the activities of the test and rework centers and consider these two together as a subsystem denoted by $G_s = \{\mathrm{T, R}\}$. Fig. 1 illustrates the flow of parts in the system where λ is the input rate of parts into the system. Waiting lines can form in each WC. Under steady state, the input rate equals the output rate, so that after initial transient effects are damped, we have that the flow through the rework loop, λ_{R}, is determined by the defects generating process modeled by a random variable N. We will provide results both for the general case and for specific distribution functions of N.

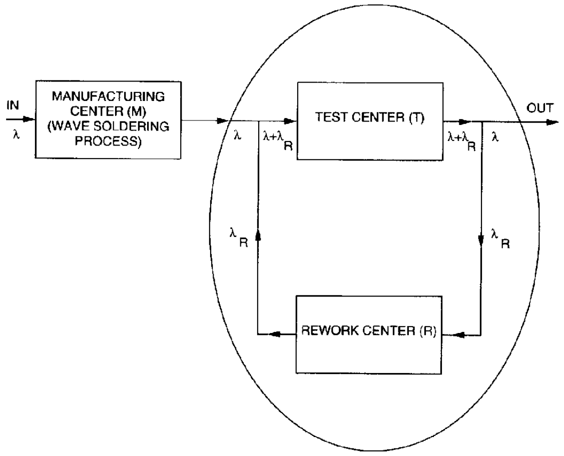

Fig. 1. Schematic representation of the process.

2.2. *Summary results and interpretations*

In this section, we define some performance measures of interest and summarize the main results which provide operational insights to the problem. The mathematical derivation of the results are provided in Section 3. As mentioned before, it is assumed throughout our analysis that the system is in steady state.

Distribution free measures

We first define measures which depend solely on the average number of defects and do not depend on the distribution of number of defects. Results on these measures will hold for any work-conserving service discipline used in work centers, and are therefore not restricted to FCFS discipline.

Flow rate through rework centers: If $\bar{N}$ is the average number of defect classes per part, then the flow rate through the rework center is $\lambda_R = \lambda \bar{N}$ and through the test center is $\lambda_T = \lambda(\bar{N} + 1)$. This is quite intuitive since λ is the external arrival rate to the subsystem and each board makes on an average $\bar{N}$ visits to the rework center. Similarly, the flow rate through the test center is the sum of the external arrival rate and the arrival rate from the rework center. Note that the flow rate through the test and rework centers only depends on the average number of defects. The proof of this is given in Proposition 3.

Average material handling cost per unit time: Material handling cost is the cost of moving units between the test center and rework center. Although we did not model the material handling time between test and rework centers, our model could be modified to incorporate this. The material handling cost is affected by the flow rate between test and rework centers. If we assume that there are ample material handling units, that is, the delay due to unavailability of material handling units is negligible, then, the material handling cost per unit time is proportional to $\lambda \bar{N}$, since the flow rate into and out of the rework center is $\lambda \bar{N}$.

Test equipment yield: The test equipment yield, y_T, defined as the flow of defect free boards per unit time that leave the test center (and hence the subsystem) relative to the work arrival rate to the center, is given by $y_T = \lambda/(\lambda + \lambda_R) = 1/(\bar{N} + 1)$. Note that the test center's workload consists of new boards that just completed the wave soldering process and circuit boards that come from the rework centre. The yield y_T can also be interpreted as the proportion of entering units leaving the subsystem. Once again,

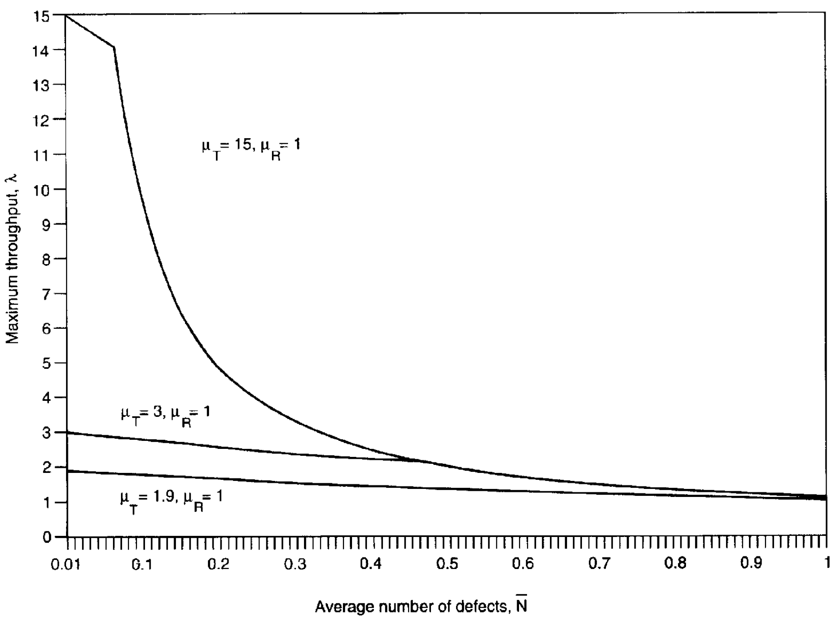

Fig. 2. Impact of average number of defects on the maximum system throughput. Both test and rework centers are assumed to have single servers.

the test equipment yield depends only on average number of defects and not on the variance, or the number of servers in WC T and WC R. As $\bar{N}$ increases, yield decreases very rapidly.

System throughput: In steady state, the system throughput, defined as the number of defect free units leaving the system per unit time, is equal to λ, the arrival rate to WC M. However, in order for the system to be stable, λ should be less than

$$\operatorname{Min}\left(\mu_M, \frac{c_T\mu_T}{(\bar{N}+1)}, \frac{c_R\mu_R}{\bar{N}}\right).$$

The subsystem will be stable only if

$$\lambda < \operatorname{Min}\left(\frac{c_T\mu_T}{(\bar{N}+1)}, \frac{c_R\mu_R}{\bar{N}}\right).$$

Fig. 2 shows the effect of $\bar{N}$ on the maximum throughput. Here, we assume that both WC T and WC R have single servers. From Fig. 2, we see that when $\bar{N}$ is small, the rework center has a small work load. Even though μ_R is smaller than μ_T, WC T is the bottleneck, and the upper bound on the throughput is due to the limited capacity in WC T. But as $\bar{N}$ increases, WC R becomes the bottleneck and the throughput is determined by the service rate at WC R. Thus, in order to increase the throughput beyond these limits, one has to either increase the service capacity or decrease the average number of defects.

Average subsystem lead time (LT): The subsystem lead time (overall subsystem wait), denoted by OW, is defined as the time between entering and leaving the subsystem. During this time, each unit visits the test center $N+1$ times and the rework center N times. This is reflected in (14) (Section 3) which gives the average subsystem lead time. As long as we assume exponential servers in both test and rework centers, the mean wait in test and rework centers per visit depends only on the average number of defects but not on its variance. Thus $\overline{\mathrm{OW}}$ depends only on $\bar{N}$ and is independent of the distribution of the number of defects. Let $\mathrm{OW}(c_T, c_R)$ denote the subsystem wait when there are c_T servers in WC T and c_R servers in WC R. If the system parameters are fixed except the number of servers, then one can prove that

$$\mathrm{OW}(\infty, \infty) \prec \mathrm{OW}(c_T, c_R) \prec \mathrm{OW}(1, 1)$$

where $\mathrm{OW}(\infty, \infty)$ corresponds to the situation where there are ample servers in both test and rework centers, and $X \prec Y$ implies that r.v. X is stochastically smaller than Y. Thus ample and single server cases provide upper and lower bounds for the overall subsystem waits. In particular,

$$\overline{\mathrm{OW}}(\infty, \infty) < \overline{\mathrm{OW}}(c_T, c_R) < \overline{\mathrm{OW}}(1, 1).$$

Average subsystem work-in-process inventory: The subsystem WIP, denoted by OL, is the number of units waiting to be processed within the subsystem. Since, from equation (19), $\overline{\mathrm{OL}} = \lambda\overline{\mathrm{OW}}$, the observations made on $\overline{\mathrm{OW}}$ above also apply to $\overline{\mathrm{OL}}$.

Distribution dependent measures

We now define performance measures which depend on the probability distribution of the number of defects, N. We consider a few well known discrete distributions for N, and find the effect of the distribution of N on the defined performance measures. In particular, we look at the effect of $\overline{N}$ and $\mathrm{Var}(N)$ on the performance measures. We first define some discrete distributions.

N-Deterministic: When N has a deterministic distribution, we know the number of defects with probability 1. That is, if $P[N = \varepsilon] = 1$, $\varepsilon = 0, 1, 2, \ldots,$ then there are exactly ε defects. In this case $\overline{N} = \varepsilon$ takes only discrete values. However, for the sake of convenience and ease of comparison, we graphed the values of $\overline{N}$ in continuum. Note that $\mathrm{Var}(N) = 0$.

N-Bernoulli: $P[N = 1] = p$; $P[N = 0] = 1 - p$. Here, the maximum number of defects is equal to 1. That is, from the rework point of view, there is only one class of defects. The test center finds all the defects simultaneously and they are all fixed in the rework center. The proportion of units with one defect is p, and with zero defects is $(1 - p)$. $\overline{N} = p$; $\mathrm{Var}(N) = p(1-p) = \overline{N}(1 - \overline{N})$; and $0 \le N \le 1$; $0 \le \mathrm{Var}(N) \le 0.25$.

N-Binomial: $P[N = r] = \binom{n}{r} p^r (1-p)^{n-r}$, $r = 0, 1, 2, \ldots, n$; $0 \le p \le 1$. In this case, there could be a maximum of n defects. Each defect can occur with a chance of p, $0 \le p \le 1$. Here $0 \le \overline{N} \le n$.

N-Geometric: $P[N = n] = p(1-p)^n$, $n = 0, 1, 2, \ldots, 0 < p < 1$. The maximum possible number of defects is unlimited, although the chance of finding n defects decreases geometrically as n increases. This distribution has a lack-of-memory property. That is, knowing that at least m defects have been observed so far, the conditional probability of finding at least n more defects is the same as the unconditional probability of finding at least n defects. Thus the number of defects is completely random; without any patterns. Here, $0 < N < \infty$.

N-Poisson: $P[N = n] = e^{-\varepsilon} \varepsilon^n / n!$, $\varepsilon > 0$, $n = 0, 1, 2, \ldots$. Once again the maximum possible number of defects is unlimited. The chance of finding too small or too large numbers of defectives (relative to the mean) is small. Here $0 < N < \infty$.

N-Uniform: $P[N = n] = 1/(a + 1)$; $n = 0, 1, \ldots, a$; $a \ge 0$. In this case, the maximum possible number of defects is equal to a, a finite number. The chances of finding n defects is equally likely, for any n between 0 and a. $\overline{N}$ $(= \frac{1}{2}a$, where a is an integer) takes only discrete values. For the sake of convenience, we graphed the values of $\overline{N}$ in continuum.

We now define the subsystem performance measures.

First pass yield: The first pass yield, y_F, is the proportion of the throughput leaving the subsystem without going through the rework loop. This is the same as the proportion of boards that leave the wave soldering machine per unit time that are defect free, and is equal to $P[N = 0]$. Fig. 3 graphs y_F as a function of $\overline{N}$ for various defect distributions. From this, we see that y_F decreases as $\overline{N}$ increases, as

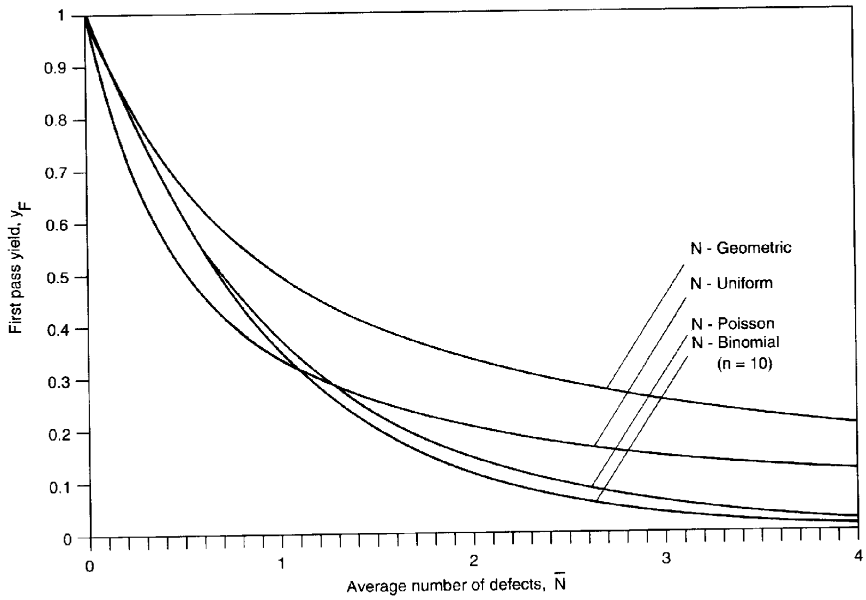

Fig. 3. Impact of average number of defects on the first pass yield.

expected. Among the Geometric, Binomial, Poisson, and Uniform distributions for N, y_F is uniformly higher when N has geometric distribution.

Other related measures of subsystem lead time: An important objective in manufacturing is to reduce the production lead time. An increase in the number of defects increases LT, if defects are reworked. We have already measured the impact of defects on the average subsystem lead time, $\overline{\mathrm{OW}}$. We now look at the variance of OW. Fig. 4 shows the impact of $\bar{N}$ on the variance of OW. Here, we assume single servers in both work centers, with parameters $\mu_T = 15$, $\mu_R = 5$, $\lambda = 1$. Even when N has a deterministic distribution (Var(N) = 0), Var(OW) increases rapidly as $\bar{N}$ increases. From (18) we see that for a given $\bar{N}$, μ_T, μ_R, and λ the rate of increase in Var(OW) per unit increase in Var(N) is constant and is equal to $(\bar{W}_T + \bar{W}_R)^2$, where $\bar{W}_i$ is the mean wait per visit in WC i, $i \in G_s$. The larger the Var(N), the larger the Var(OW) and vice versa. In order to decrease Var(OW), we need to first reduce the average number of defects and then the variance of the number of defects. When both tests and rework centers have ample servers, the overall pattern of the impact of $\bar{N}$ on Var(OW) remains the same, although the magnitude is smaller. Hence, we do not discuss this case again.

Coefficient of variation of OW (denoted by CV(OW)) is defined as [s.d. (OW)]/$\overline{\mathrm{OW}}$. When N has a Poisson distribution, Fig. 5 compares the CV(OW) for two extreme cases: when the work centers T and R have single servers and ample servers respectively. We see that for small $\bar{N}$, CV(OW) is smaller when both the work centers have single servers. This is quite counter intuitive. The same phenomenon is observed even when N has a distribution other than Poisson. We now look at r, the correlation coefficient between the total time spent in T and in R by each circuit board. From (27) we see that when service time is exponential, r is non-negative, depends only on the mean and variance of N, and is independent of the number of servers and the time spent per visit in each work center.

We now analyze the impact of the defect distribution on the probability density function of the overall subsystem lead time. Fig. 6 gives the p.d.f. of OW, $f_{\mathrm{OW}}(t)$, when N has a Bernoulli distribution, with different values of $\bar{N}$. When $\bar{N} = 1$, N has a deterministic distribution with one defect. As $\bar{N}$ varies from 0.1 to 1, $f_{\mathrm{OW}}(t)$ changes from unimodal to bimodal, and back to a unimodal distribution. Note that for a

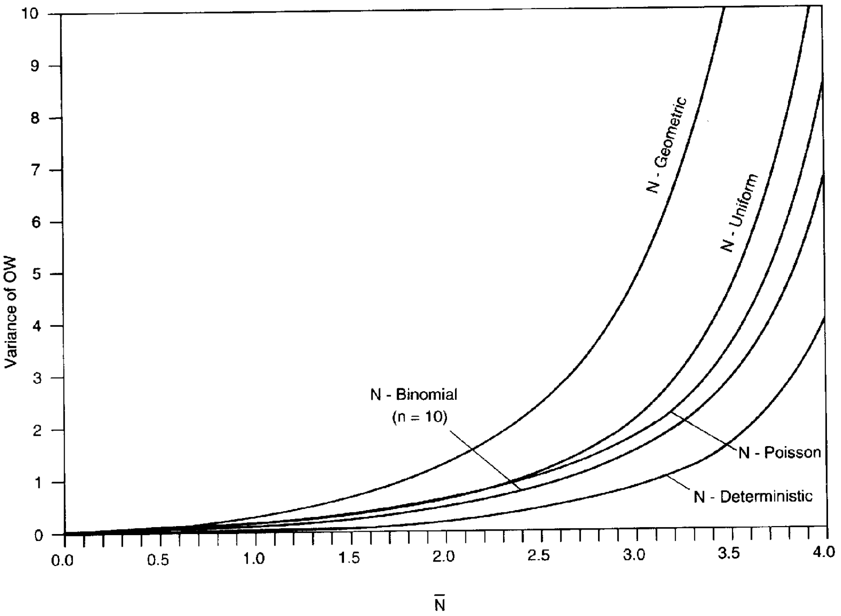

Fig. 4. Impact of average number of defects on the variance of subsystem lead time. Both WC T and WC R are assumed to have single exponential servers with parameters $\mu_T = 15$, $\mu_R = 5$, $\lambda = 1$.

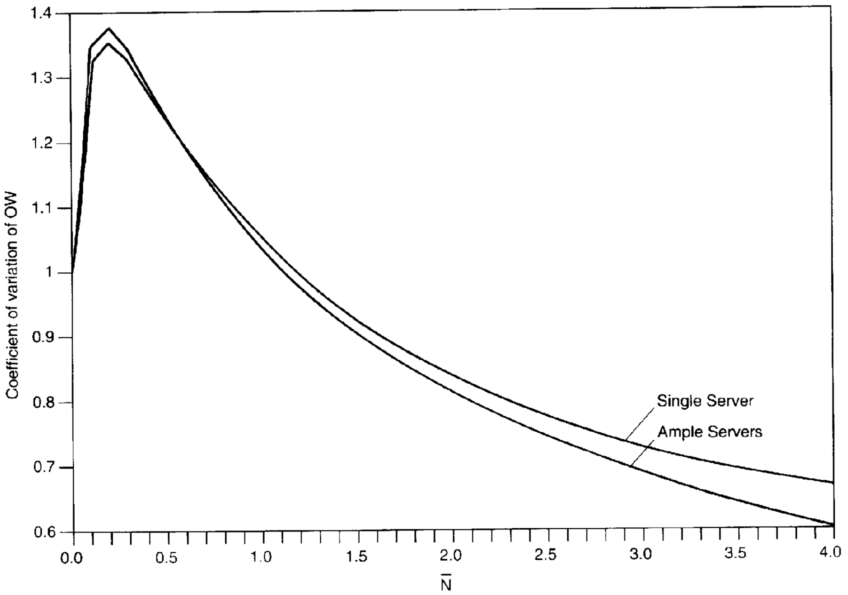

Fig. 5. Comparison of coefficient of variation of OW for two extreme situations, when the test and rework centers have (i) single exponential servers and (ii) ample exponential servers, with parameters $\mu_T = 15$, $\mu_R = 5$ and $\lambda = 1$. N is assumed to have a Poisson distribution.

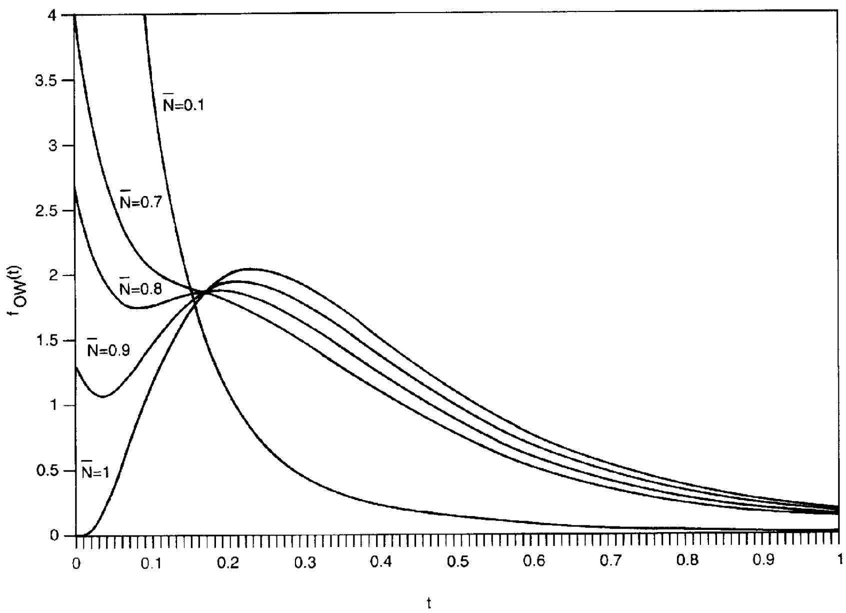

Fig. 6. Impact of $\bar{N}$ on $f_{OW}(t)$ when N has a Bernoulli distribution. Both test and rework centers are assumed to have single exponential servers with $\mu_T = 15$, $\mu_R = 5$ and $\lambda = 1$.

given t^*, $P[\mathrm{OW} > t^*]$ increases as $\bar{N}$ increases. Fig. 7 analyzes the effect of varying the service rate in the rework center on $f_{OW}(t)$. In all the above examples we assumed that both the test and rework centers have a single exponential server. When both work centers have ample exponential servers, the

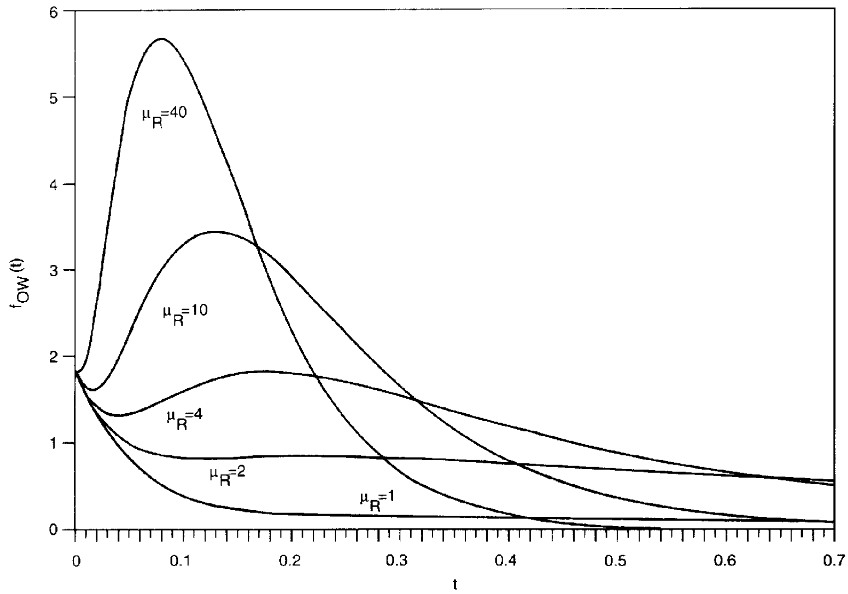

Fig. 7. Impact of varying rework center service rate on $f_{OW}(t)$ when N has a Bernoulli distribution with $\bar{N} = p = 0.9$. Both test and rework centers have single exponential servers, $\mu_T = 20$ and $\lambda = 1$.

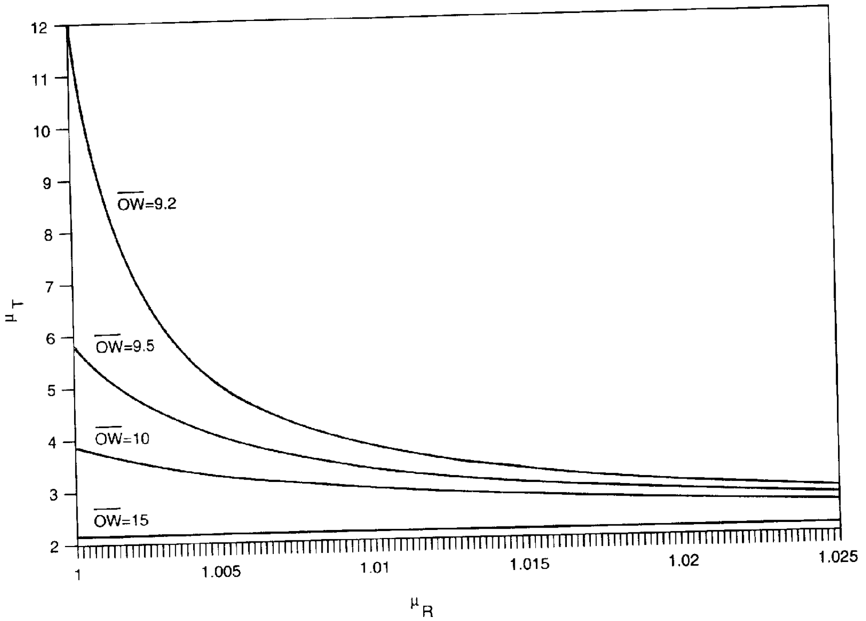

Fig. 9. Combination of μ_R and μ_T which would provide the same $\overline{OW}$ (or $\overline{OL}$). It is assumed that both test and rework centers have single exponential servers, $\lambda = 1$ and $\overline{N} = 0.9$.

for solder defects is an inadequate circuit board design. The long term project to reduce defect levels would include redesigning of the boards. Our model could be used in two ways. First, since product improvements by redesign will impact service times at WC T and WC R, as well as the mean and distribution of N, our model could be used to analyze the improvement in performance measures. In turn, we can identify cost reductions by reduced work-in-process inventory or reduced number of servers. Second, since all boards cannot be redesigned simultaneously due to budget constraint, in general, we need to identify the subset of boards that will be redesigned so as to maximize the overall benefit to the company. We now propose a budget allocation method for such improvement projects using the concept of Value Added Detractor or 'Vador' proposed by Hoadley (1986). According to Hoadley, a quality and productivity improvement strategy consists of increasing the value added by all processes that make up an enterprise. A Vador is anything that detracts from the value added of a process such as a defect, mistake, and failure. In order to eliminate Vadors, one must invest in value added improvement. Vador values are determined by considering the impact of an investment in an improvement project taking into account the period of time for which the improvement will have benefits, the number of boards manufactured during this period and the average cost per Vador. Different circuit board designs will have different Vador values. In our analysis we use the average defect level as the vador and identify which designs should be considered as candidates for redesign improvement projects.

Suppose there are m possible improvement projects. For project i, $i = 1, 2, \ldots, m$, define

$\overline{N}_i(x_i)$: Average defect level per output unit per year if we invest an amount x_i on improvement. This function is assumed to be decreasing and convex with $\overline{N}_i(0) = \overline{N}_i$, the present defect level.

H_i: Number of years for which improvement will have a benefit.

D_i: Average number of boards manufactured per year over H_i years.

C_i: Average cost per board per unit defect. This includes all relevant costs such as cost of rework loops, and costs due to increased lead time.

B: Budget available for improvement projects.

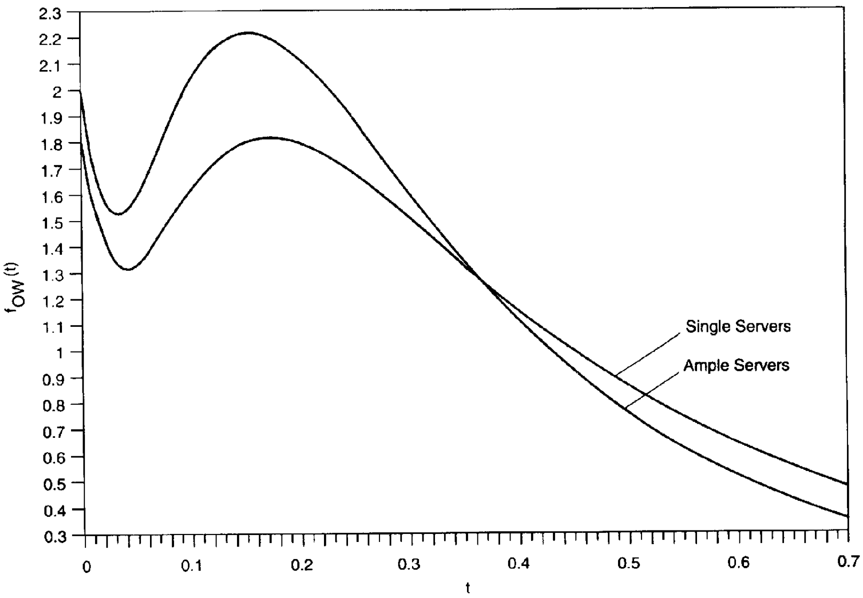

Fig. 8. Impact of number of servers on $f_{OW}(t)$. It is assumed that N has a Bernoulli distribution with $\overline{N} = 0.9$ and both work centers have single exponential servers with $\mu_T = 20$, $\mu_R = 4$ and $\lambda = 1$.

corresponding graphs look similar. Fig. 8 analyzes the impact of the number of servers in the work centers on $f_{OW}(t)$. It compares $f_{OW}(t)$ when both work centers have single servers versus ample servers, when N has a Bernoulli distribution. More graphs of different situations are provided in Agnihothri and Kenett (1991).

2.3. *Short term control*

The above results can be used in many different ways. We present here an example of short term control. We assume that the system is in steady state and that the soldering process is under statistical control. For simplicity we also assume that there is a single server in each of the work centers and we do not vary the service rate at the manufacturing center. The control parameters are the number of defects per board and the service rates at the test and rework centers. These three parameters dictate the performance measures defined earlier and, in particular, the subsystem throughput and the subsystem LT. For example, the average wait in the subsystem is limited by $\overline{N}$, μ_T and μ_R. Reducing the average number of defects per board typically requires a redesign of the board and cannot be done in a short time frame. In the short term one can increase either the service rate or the number of servers in the test or rework center. A target average subsystem lead time ($\overline{OW}$) can be achieved by choosing one of several different combination of μ_R and μ_T as shown in Fig. 9. The decision can be based on the cost of service in each center. For example, when $\overline{OW} = 9.2$, one can choose either large μ_T and small μ_R or small μ_T and large μ_R. However, when $\overline{OW} = 15$, the effect of μ_R when $\mu_R \geq 1$ seems to be negligible. Fig. 10 describes the decision making process using a flow chart.

2.4. *Long term improvements*

So far, we have tried to quantify the impact of defects given a defect pattern. In general, there are several types of boards, each having its own defect pattern. The ultimate objective is to reduce the defect levels. In our problem, the defects are generated by the soldering process. One of the possible reasons

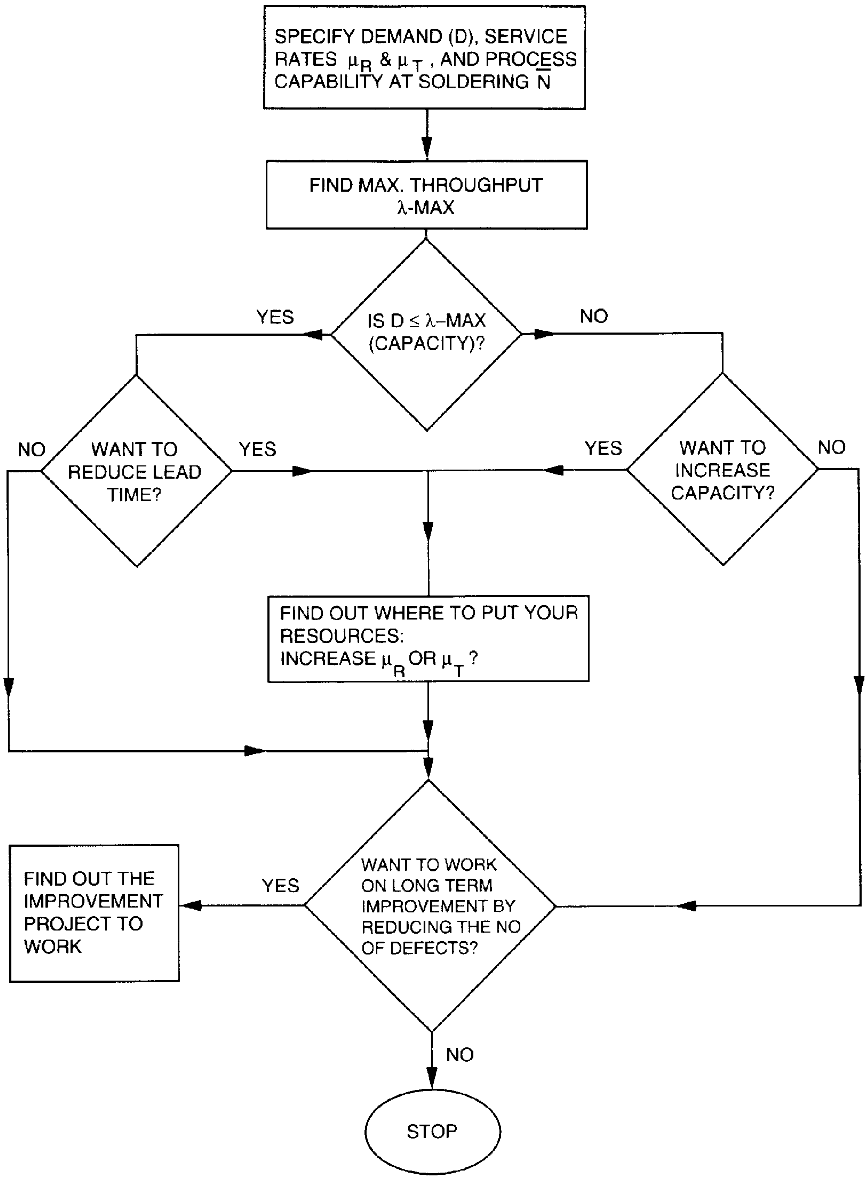

Fig. 10. Flowchart for decision-making.

Thus, the improvement derived from investing an amount x_i on design i, is $D_iC_iH_i[\bar{N}_i(0) - \bar{N}_i(x_i)]$. The budget allocation problem is to choose investments $\{x_i,\ i = 1, 2, \ldots, m\}$ so as to

$$\text{Maximize} \quad \sum_{i=1}^{m} D_iC_iH_i\left[\bar{N}_i(0) - \bar{N}_i(x_i)\right]$$

$$\text{subject to} \quad \sum_{i=1}^{m} x_i \le B.$$

An algorithmic solution to the above mathematical programming problem can be described as follows (see Hoadly, 1986 for details).

Step 1. For every project i, define an improvement index $V_i(x_i) = -D_i C_i H_i \bar{N}_i'(x_i)$, where $\bar{N}_i'(x_i)$ is the first derivative of $\bar{N}_i(x_i)$ with respect to x_i. Compute $V_i = V_i(0)$.
Step 2. Determine an initial improvement index target, V^* (which is the Lagrangian multiplier).
Step 3. Select project i for improvement if $V_i > V^*$.
Step 4. For each selected improvement project i, allocate an amount x_i^* such that $V_i(x_i^*) = V^*$.
Step 5. Compare $\Sigma_{i=1}^m x_i^*$ to B. Choose a new V^* by decreasing V^* if $\Sigma_{i=1}^m x_i^* < B$ and increase V^* if $\Sigma_{i=1}^m x_i^* > B$. Go to Step 3.

Thus, optimal value of V^* is found by a trial and error method.

A numerical example. For project i, let $\bar{N}_i(x_i) = \bar{N}_i e^{-x_i/a_i}$, where $a_i > 0$ is a constant. Note that $\bar{N}_i(0) = \bar{N}_i$. Here, a_i can be thought of a benchmark improvement cost since if we spend an amount $x_i = a_i$, then the new average defect level will be 37% ($e^{-1} = 0.37$) of the current average defect level. Let $m = 4$, $D_1 = D_3 = 10\,000$; $D_2 = D_4 = 1000$ circuit boards per year; $C_1 = C_4 = \$500$, $C_2 = C_3 = \$100$; $\bar{N}_1 = 0.4$, $\bar{N}_2 = 1$, $\bar{N}_3 = 0.5$; $\bar{N}_4 = 0.3$ defects per board; $H_i = 5$ years, $a_i = \$5000$, $i = 1, \ldots, 4$; and $B = \$10\,000$. From the algorithm, we have $V_i(x_i) = (D_i C_i H_i \bar{N}_i / a_i)\, e^{-x_i/a_i}$, and $V_i(0) = V_i = D_i C_i H_i \bar{N}_i / a_i$. Hence, $V_1 = 2000$, $V_2 = 100$, $V_3 = 500$, $V_4 = 150$. Let $V^* = 400$. Then choose projects 1 and 3 for redesign. Since $x_i^* = a_i \log(V_i/V^*)$, we get $x_1^* = \$8047$, and $x_3^* = \$1116$. But $x_1^* + x_3^* = 9163 < 10\,000$. Hence, decrease V^* and find x_i^* such that $x_1^* + x_3^* = 10\,000$. The optimal $V^* = 367.9$, $x_1^* = \$8466$, and $x_3^* = \$1534$.

2.5. Simulation results

Our model assumes that the service times in all three work centers are exponentially distributed. However, this is a restrictive assumption. In order to check the sensitivity of this assumption to the

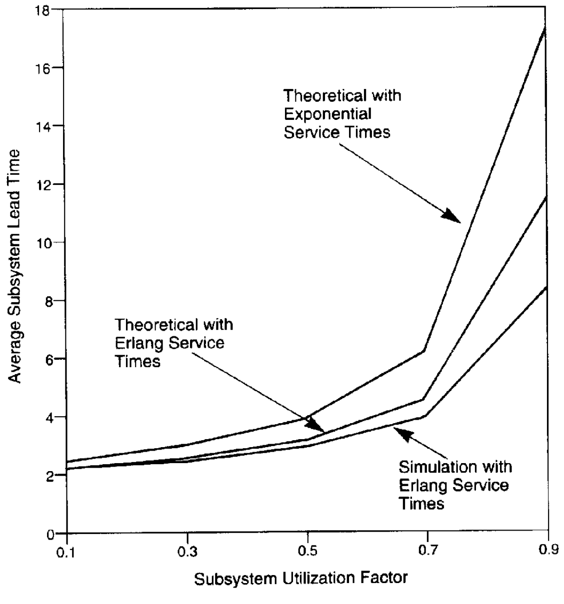

Fig. 11. Average subsystem lead time as the utilization factor varies. It is assumed that all three work centers have the same service time distribution, with $\mu_S = 0.625$, $\mu_T = 1$ and $\mu_R = 1.25$ and N has a Bernoulli distribution with $p = 0.6$.

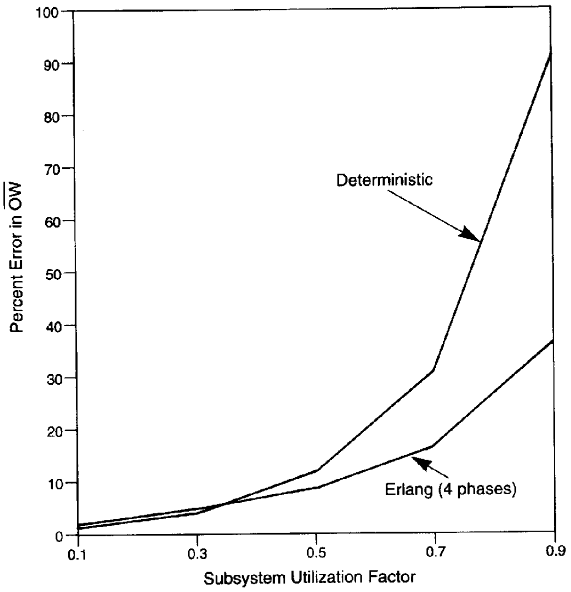

Fig. 12. Percent error in $\overline{OW}$ using theoretical model as utilization factor varies, when service times at all the work centers have deterministic and Erlang distribution respectively. The values of the parameters are the same as in Fig. 11.

theoretical results obtained, we developed a simulation in Siman language. We assumed that there is a single server at each work center and the external arrival process to the work center M is Poisson. We wanted to evaluate the impact of the distribution of N on the mean subsystem lead time when the service time distribution at all three work centers are non exponential. We tested both deterministic and Erlang distributions for service times. In order to obtain $\overline{OW}$ using our theoretical model when service times are non-exponential, we first obtained $\overline{W}_T$ and $\overline{W}_R$ by using an M/G/1 model at WC T and R with arrival rates given by (3) and then used (14) to get $\overline{OW}$.

From the simulation, we see that the theoretical results obtained by exponential service time assumption provide an upper bound in all the cases as conjectured before. An example of the difference between theoretical and simulation results is given in Fig. 11. Also, in all the cases the error in estimating the average subsystem lead time using the theoretical model increases as the sub system utilization increases. Finally, the percent error increases as the service time distribution changes from Erlang to deterministic distribution (with coefficient of variation of service time decreasing from one to zero), as indicated in Fig. 12. Hence, the error is largest when the service time distribution is deterministic. It can be noted that the error in most cases is not quite significant. The average error is less than 15% when the service time distribution is Erlang, and less than 25% when it is deterministic.

3. Mathematical derivation

In the following analysis we assume that the system has reached steady state. Let $P[N=n]=e_n$, $n=0, 1, 2, \dots$, be the probability distribution of the r.v. N, the number of defects in a particular unit passing through M. Let V_i denote the number of visits by a specific unit to the WC i, $i \in G_s$.

Proposition 1.

$$V_{\mathrm{R}} \equiv N \text{ and } V_{\mathrm{T}} \equiv N+1, \qquad (1)$$
$$P[V_T = k] = e_{k-1}, \quad k = 1, 2, \ldots .$$

Proof. Follows from the model's assumptions. □

Proposition 2. *The departure process of type r unit from WC* M *is a Poisson process with rate* λe_r, $r = 0, 1, \ldots$.

Proof. By assumption, WC M is an M/M/1 queueing system, with arrival rate λ. Let $D = \{D(t),\ t \geq 0\}$ be the number of the departures from WC M $\in (0, \varepsilon]$. Then D, under steady state, is a Poisson process with rate λ (Burke's theorem).

Let

$$X_n = \begin{cases} 1 & \text{if } n\text{-th departing unit is a type } r \text{ unit,} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\{X_n,\ n = 1, 2, \ldots\}$ is the embedded Bernoulli process, independent of D, with $P[X_n = 1] = e_r$. If we let $M_r(t)$ and $L_r(t)$ denote total number of type r units, and total number of non-type r units in $(0, t]$ respectively, then $M_r(t) = \Sigma_{n=1}^{D(t)} X_n$, and $L_r(t) = D(t) - M_r(t)$. We can show (see Çinlar, 1975, p. 89) that $\{M_r(t),\ t \geq 0\}$ and $\{L_r(t),\ t \geq 0\}$ are independent Poisson Processes with respective rates λe_r and $\lambda(1 - e_r)$. □

Let λ_i be the arrival rate to WC i, $i \in G_s$. Then we have the following:

Proposition 3. For $i \in G_s$, *the arrival rates are given by*

$$\lambda_{\mathrm{R}} = \lambda \overline{N} = \lambda \overline{V}_{\mathrm{R}} \text{ and } \lambda_{\mathrm{T}} = \lambda(\overline{N} + 1) = \lambda \overline{V}_{\mathrm{T}} \qquad (3)$$

respectively.

Proof. The steady state arrival rate to the subsystem is λ and so is the departure rate from the subsystem. Since $P[V_{\mathrm{T}} = k] = e_{k-1}$, $k = 1, 2, \ldots$, the rate at which units with exactly k visits to WC T depart the subsystem is λe_{k-1}. In order to find λ_{R}, imagine for a minute, that the WC T has infinite parallel servers. All units visiting WC T for the k-th time are served by server k. Then, the rate of departure out of the system from server k is λe_{k-1} and the arrival rate to serve k is the arrival rate of units with at least $k - 1$ defects which is $\Sigma_{r=k-1}^{\infty} \lambda e_r$.

Hence the total arrival rate of units to be reworked is

$$\lambda_{\mathrm{R}} = \sum_{k=2}^{\infty} \sum_{r=k-1}^{\infty} \lambda e_r = \lambda \sum_{k=1}^{\infty} \sum_{r=k}^{\infty} e_r = \lambda \overline{N}$$

(after changing the order of summation). Now $\lambda_{\mathrm{T}} = \lambda + \lambda_{\mathrm{R}} = \lambda(\overline{N} + 1)$. □

Proposition 4. *The entire system is stable if*

$$\lambda < \mu^* = \mathrm{Min}\left\{ \mu_M, \frac{c_T \mu_T}{1 + \overline{N}}, \frac{c_R \mu_R}{\overline{N}} \right\}. \qquad (4)$$

Proof. Note that WC i is stable if $\lambda_i < c_i\mu_i$, $i \in G$. Substituting for λ_i, and simplifying, we see that WC M is stable if $\lambda < \mu_{\mathrm{M}}$, WC T is stable if $\lambda < \mu_{\mathrm{T}}/(1+\overline{N})$, and WC R is stable if $\lambda < \mu_{\mathrm{R}}/\overline{N}$. Hence,

$$\lambda < \mu^* = \operatorname{Min}\left\{\mu_M,\ \frac{c_T\mu_T}{1+\overline{N}},\ \frac{c_R\mu_R}{\overline{N}}\right\}. \qquad \square$$

Since the external arrivals are Poisson and the service times are exponentially distributed, we have a Jackson network. Hence, each node $i \in G_s$ can be viewed as an independent M/M/c queue with parameters λ_i and μ_i.

For $i \in G$, define W_i^q and W_i to be the queueing time and wait time in WC i per visit, and L_i be the queue length in WC i.

Proposition 5. *For* $i \in G$,

$$w_i(t) = \frac{\mu_i\,\mathrm{e}^{-\mu_i t}(\lambda_i - c_i\mu_i + \mu_i W_i^q(0)) - (1 - W_i^q(0))(\lambda_i - c_i\mu_i)\mu_i\,\mathrm{e}^{-(c_i\mu_i-\lambda_i)t}}{\lambda_i - (c_i - 1)\mu_i}, \quad t > 0, \tag{5}$$

where

$$W_i^q(0) = 1 - \frac{c_i(\lambda_i/\mu_i)^{c_i}}{c_i!(c_i - \lambda_i/\mu_i)} p_{0i},$$

$$p_{0i} = \left\{\sum_{n=0}^{c_i-1} \frac{1}{n!}\left(\frac{\lambda_i}{\mu_i}\right)^n + \frac{1}{c_i!}\left(\frac{\lambda_i}{\mu_i}\right)^{c_i}\left(\frac{c_i\mu_i}{c_i\mu_i - \lambda_i}\right)\right\}^{-1},$$

$$\overline{W}_i^q = \left(\frac{(\lambda_i/\mu_i)^{c_i}\mu_i}{(c_i - 1)!(c_i\mu_i - \lambda_i)^2}\right) p_{0i},$$

and

$$\overline{W}_i = \overline{W}_i^q + \overline{S}_i. \tag{6}$$

Also

$$P[L_i = n] = \begin{cases} \{\lambda_i^n/(n!\mu_i^n)\}p_{0i}, & 1 \le n \le c_i, \\ \{\lambda_i^n/(c_i^{n-c_i}c_i!\mu_i^n)\}p_{0i}, & n \ge c_i, \end{cases} \tag{7}$$

and

$$\overline{L}_i = \lambda_i \overline{W}_i, \tag{8}$$

from Little's formula.

Proof. Obvious. $\square$

In order to analyze the impact on the system, of defects generated at WC M, we now concentrate on the subsystem G_s consisting of WC T and WC R. Define OW_i to be the overall time a unit spends in WC i including all revisits to the WC i, before leaving the system. Let $\mathrm{OW} = \mathrm{OW}_{\mathrm{T}} + \mathrm{OW}_{\mathrm{R}}$ = total time spent in the subsystem, and OL = overall number of units in the subsystem. We denote the Laplace transform of a continuous r.v. X by $\tilde{X}(s) = \int_0^\infty \mathrm{e}^{-st} f_X(t)\,\mathrm{d}t$. The probability generating function (pgf) of a discrete r.v. X is denoted by $\hat{X}(z) = \sum_{k=0}^\infty P[X = k]z^k$.

Proposition 6.

$$\widetilde{\mathrm{OW}}_{\mathrm{T}}(s) = \tilde{W}_{\mathrm{T}}(s)\hat{N}\big(\tilde{W}_{\mathrm{T}}(s)\big), \tag{9}$$

$$\widetilde{\mathrm{OW}}_{R}(s) = \hat{N}\big(\tilde{W}_{\mathrm{R}}(s)\big), \tag{10}$$

$$\widetilde{\mathrm{OW}}(s) = \tilde{W}_{\mathrm{T}}(s)\hat{N}\big(\tilde{W}_{\mathrm{T}}(s)\tilde{W}_{\mathrm{R}}(s)\big), \tag{11}$$

$$\overline{\mathrm{OW}}_{\mathrm{T}} = (\bar{N}+1)\bar{W}_{\mathrm{T}}, \tag{12}$$

$$\overline{\mathrm{OW}}_{\mathrm{R}} = \bar{N}\bar{W}_{\mathrm{R}}, \tag{13}$$

$$\overline{\mathrm{OW}} = (\bar{N}+1)\bar{W}_{\mathrm{T}} + \bar{N}\bar{W}_{\mathrm{R}}, \tag{14}$$

$$\mathrm{Var}(\mathrm{OW}_{\mathrm{T}}) = (1+\bar{N})\ \mathrm{Var}(W_{\mathrm{T}}) + \big(\bar{W}_{\mathrm{T}}\big)^2\ \mathrm{Var}\ N, \tag{15}$$

$$\mathrm{Var}(\mathrm{OW}_{\mathrm{R}}) = \bar{N}\ \mathrm{Var}(W_{\mathrm{R}}) + \big(\bar{W}_{\mathrm{R}}\big)^2\ \mathrm{Var}\ N, \tag{16}$$

$$\mathrm{Cov}(\mathrm{OW}_{\mathrm{T}}, \mathrm{OW}_{\mathrm{R}}) = \bar{W}_{\mathrm{T}}\bar{W}_{\mathrm{R}}\ \mathrm{Var}\ N, \tag{17}$$

$$\mathrm{Var}(\mathrm{OW}) = \big(\bar{W}_{\mathrm{T}} + \bar{W}_{\mathrm{R}}\big)^2\ \mathrm{Var}\ N + (\mathrm{Var}\ W_{\mathrm{T}} + \mathrm{Var}\ W_{\mathrm{R}})\bar{N} + \mathrm{Var}\ W_{\mathrm{T}}, \tag{18}$$

$$\overline{\mathrm{OL}} = \lambda\overline{\mathrm{OW}}. \tag{19}$$

Proof. For $i \in G_s$ let W_i^k = waiting time in WC i during the k-th visit, $k = 1, 2, \ldots$. Then

$$\mathrm{OW}_i = \sum_{k=1}^{V_i} W_i^k \tag{20}$$

Since $W_i^{(k)}$ and V_i are independent of each other, we have

$$\widetilde{\mathrm{OW}}_i(s) = \hat{V}_i\big(\tilde{W}_i(s)\big). \tag{21}$$

Then (9) and (10) follows from (21) and (1) by noting that

$$\hat{V}_{\mathrm{T}}(z) = z\hat{N}(z) \text{ and } \hat{V}_{\mathrm{R}}(z) = \hat{N}(z).$$

To prove (11), note that

$$\mathrm{OW} = \sum_{k=1}^{N+1} W_{\mathrm{T}}^k + \sum_{k=1}^{N} W_{\mathrm{R}}^k,$$

$$\widetilde{OW}(s) = \sum_{n=0}^{\infty} e_n \tilde{W}_{\mathrm{T}}(s)\big(\tilde{W}_{\mathrm{T}}(s)\tilde{W}_{\mathrm{R}}(s)\big)^n \qquad \text{(by conditioning on } n\text{)}$$

$$= \tilde{W}_{\mathrm{T}}(s)\hat{N}\big(\tilde{W}_{\mathrm{T}}(s)\tilde{W}_{\mathrm{R}}(s)\big).$$

Also from (20) (See Feller, 1968, p. 301)

$$\overline{\mathrm{OW}}_i = \bar{V}_i\bar{W}_i$$

$$\mathrm{Var}(\mathrm{OW}_i) = \bar{V}_i\ \mathrm{Var}(W_i) + \big(\bar{W}_i\big)^2\ \mathrm{Var}(V_i).$$

Substituting for V_i from (1), in the above equations, we get the equations (12) through (16). (18) could be obtained from (11) by noting that

$$\mathrm{Var}(\mathrm{OW}) = \left.\frac{\mathrm{d}^2\widetilde{\mathrm{OW}}(s)}{\mathrm{d}s^2}\right|_{s=0} - \left(\left.\frac{\mathrm{d}\widetilde{\mathrm{OW}}(s)}{\mathrm{d}s}\right|_{s=0}\right)^2.$$

(17) could be obtained from (15), (16), and (18). Since $\overline{\mathrm{OL}} = \bar{L}_{\mathrm{T}} + \bar{L}_{\mathrm{R}}$, (19) follows from (8) and (14) or directly from Little's formula. □

In what follows, we obtain the upper and lower bounds on the mean and variance of the overall wait in the subsystem. The upper bound is obtained by assuming that there is a single exponential server in both WC T and WC R. The lower bound is obtained by assuming there are ample exponential servers in both WC T and WC R. The subsystem performance measures should always lie between these two. Assuming single exponential server at WC i, $i \in G_s$, we first obtain mean and variance of the overall wait. We then investigate the impact of the actual distribution of the defect on the overall subsystem wait.

Proposition 7. *When there is a single exponential server in WC i, $i \in G_s$, W_i, the wait per visit in WC i, has exponential distribution with rate α_i where:*

$$\alpha_{\mathrm{T}} = \mu_{\mathrm{T}} - \lambda(1 + \bar{N}), \tag{22}$$

$$\alpha_{\mathrm{R}} = \mu_{\mathrm{R}} - \lambda \bar{N}, \tag{23}$$

$$\tilde{W}_i(s) = \alpha_i / (\alpha_i + s), \quad i \in G_s. \tag{24}$$

In this case

$$\overline{\mathrm{OW}}^{\mathrm{U}} = (1 + \bar{N})/\alpha_{\mathrm{T}} + \bar{N}/\alpha_{\mathrm{R}}, \tag{25}$$

$$\mathrm{Var}(\mathrm{OW})^{\mathrm{U}} = \bar{N}(1/\alpha_{\mathrm{T}}^2 + 1/\alpha_{\mathrm{R}}^2) + \mathrm{Var}\,N(1/\alpha_{\mathrm{T}} + 1/\alpha_{\mathrm{T}})^2 + 1/\alpha_{\mathrm{R}}^2, \tag{26}$$

and Correlation Coefficient of OW_{T} *and* OW_{R}

$$= r(\mathrm{OW}_{\mathrm{T}}, \mathrm{OW}_{\mathrm{R}}) = \frac{\mathrm{Var}\,N}{\sqrt{(\bar{N} + \mathrm{Var}\,N)(\bar{N} + \mathrm{Var}\,N + 1)}}. \tag{27}$$

Proof. Follows from (5) with $c_i = 1$, $i \in G_s$. Eqs. (25) and (26) follow form (14) and (18) with $\bar{W}_i = 1/\alpha_i$ and $\mathrm{Var}\, W_i = \bar{W}_i^2$, $i \in G_s$. Eq. (27) follows from (17), (15), and (16). □

In order to get lower bound for the overall wait in the subsystem, we will assume that WC i has ample exponential servers, $i \in G_s$.

Proposition 8. *When there are ample servers in WC i, then W_i has exponential distribution with rate μ_i, $i \in G$.*

$$\tilde{W}_i(s) = \mu_i / (\mu_i + s), \quad i \in G_s. \tag{28}$$

In this case,

$$\overline{\mathrm{OW}}^{\mathrm{L}} = (1 + \bar{N})/\mu_{\mathrm{T}} + \bar{N}/\mu_{\mathrm{R}} \tag{29}$$

$$\mathrm{Var}(\mathrm{OW})^{\mathrm{L}} = \bar{N}(1/\mu_{\mathrm{T}}^2 + 1/\mu_{\mathrm{R}}^2) + \mathrm{Var}\,N(1/\mu_{\mathrm{T}} + 1/\mu_{\mathrm{R}})^2 + 1/\mu_{\mathrm{T}}^2. \tag{30}$$

Proof. Since the WC i can be modeled as an M/M/∞ queueing system, W_i is nothing but the service time S_i, which has an exponential distribution with rate μ_i, $i \in G_s$. The mean and variance of OW can be obtained by replacing α_i by μ_i in (25) and (26). □

Thus $\overline{\mathrm{OW}}^{\mathrm{L}} \le \overline{\mathrm{OW}} \le \overline{\mathrm{OW}}^{\mathrm{U}}$ and $\mathrm{Var(OW)}^{\mathrm{L}} \le \mathrm{Var(OW)} \le \mathrm{Var(OW)}^{\mathrm{U}}$.

We will now obtain the probability density function of the overall waiting time in the subsystem assuming a particular defect distribution, for two extreme cases. When WC T and WC R have single exponential servers, α_{T} and α_{R} are given by (22) and (23). When WC T and WC R have ample servers, the corresponding expressions can be obtained by replacing $\alpha_{\mathrm{T}} = \mu_{\mathrm{T}}$ and $\alpha_{\mathrm{R}} = \mu_{\mathrm{R}}$. The proofs are omitted here and could be found in Agnihothri and Kenett (1991).

Proposition 9. *When N has deterministic distribution with $P[N = \varepsilon] = 1$, the p.d.f. of the overall wait in the subsystem is given by*

$$f_{\mathrm{OW}}(t) = \frac{\alpha_{\mathrm{T}}^{\varepsilon+1}\alpha_{\mathrm{R}}^{\varepsilon}\,\mathrm{e}^{-\alpha_{\mathrm{R}}t}}{\varepsilon!}\sum_{j=0}^{\infty}\frac{[(\alpha_{\mathrm{R}}-\alpha_{\mathrm{T}})t]^{j}(j+\varepsilon)!}{j!(j+2\varepsilon)!}, \quad t>0.$$

Proposition 10. *If N has geometric distribution with $P[N = n] = p(1-p)^n$, $n = 0, 1, 2, \ldots$, then*

$$f_{\mathrm{OW}}(t) = \left(\frac{a-br_2}{r_1-r_2}\right)\mathrm{e}^{-r_2t} - \left(\frac{a-br_1}{r_1-r_2}\right)\mathrm{e}^{-r_1t}, \quad t>0,$$

where $a = \alpha_{\mathrm{T}}\alpha_{\mathrm{R}}p$, $b = \alpha_{\mathrm{T}}p$, $r_1 = \frac{1}{2}(\alpha_{\mathrm{T}} + \alpha_{\mathrm{R}} + K)$, $r_2 = \frac{1}{2}(\alpha_{\mathrm{T}} + \alpha_{\mathrm{R}} - K)$ and $K = \sqrt{\alpha_{\mathrm{T}}^2 + \alpha_{\mathrm{R}}^2 + 2\alpha_{\mathrm{T}}\alpha_{\mathrm{R}}(1-p)}$.

Proposition 11. *If N has Bernoulli distribution with $P[N = 1] = p$ and $P[N = 0] = 1-p$, $0 < p \le 1$, then*

$$f_{\mathrm{OW}}(t) = A\,\mathrm{e}^{-\alpha_{\mathrm{T}}t} + BC\,\mathrm{e}^{-\alpha_{\mathrm{R}}t} + Bt\,\mathrm{e}^{-\alpha_{\mathrm{T}}t}, \quad t>0,$$

where

$$A = \left((1-p)\alpha_{\mathrm{T}} - \frac{p\alpha_{\mathrm{T}}^2\alpha_{\mathrm{R}}}{(\alpha_{\mathrm{R}}-\alpha_{\mathrm{T}})^2}\right), \quad B = \frac{p\alpha_{\mathrm{T}}^2\alpha_{\mathrm{R}}}{\alpha_{\mathrm{R}}-\alpha_{\mathrm{T}}} \text{ and } C = \frac{1}{\alpha_{\mathrm{R}}-\alpha_{\mathrm{T}}}.$$

Proposition 12. *If N has Binomial distribution with $P[N = r] = e_r = \binom{n}{r}p^r(1)p)^{n-r}$, $r = 0, 1, 2, \ldots, n$; $0 \le p \le 1$, then*

$$f_{\mathrm{OW}}(t) = \sum_{r=0}^{n}\frac{e_r\alpha_{\mathrm{T}}^{r+1}\alpha_{\mathrm{R}}^{r}t_{2r}\,\mathrm{e}^{-\alpha_{\mathrm{R}}t}}{r!}\sum_{j=0}^{\infty}\frac{((\alpha_{\mathrm{R}}-\alpha_{\mathrm{T}})t)^{j}}{j!}\frac{(j+r)!}{(j+2r)!}, \quad t>0.$$

Finally, we consider a case when N depends on the service time S_{M} at WC M. Let $\mathscr{E}(t)$ be the number of defects found in time $(0, t]$. We will assume that $\mathscr{E}(t)$ is a Poisson process with rate ε_1. Then $N \equiv \mathscr{E}(S_{\mathrm{M}})$. Since the service time S_{M} has exponential distribution,

$$\hat{N}(z) = \tilde{S}_M[\varepsilon_1(1-z)] = \frac{\mu_M}{\mu_M + \varepsilon_1(1-z)} = \frac{\mu_M/(\mu_M+\varepsilon_1)}{1-\left(\dfrac{\varepsilon_1 z}{\mu_M+\varepsilon_1}\right)}.$$

Hence N has a geometric distribution with $p = \mu_{\mathrm{M}}/(\mu_{\mathrm{M}} + \varepsilon_1)$. The corresponding p.d.f. $f_{\mathrm{OW}}(t)$ can be obtained from Proposition 10 by replacing $p = \mu_{\mathrm{M}}/(\mu_{\mathrm{M}} + \varepsilon_1)$.

4. Summary and conclusions

In this paper, we have considered a production process with 100% inspection followed by rework. We have modelled the number of defects generated by the manufacturing process as a random variable with any general discrete distribution, and investigated the impact of defect levels on system performance measures. We have defined measures which depend only on the average number of defects and are independent of the distribution of number of defects. For the performance measures such as first pass yield and production lead time, we have studied, in detail, the impact of the distribution of the number of defects on these measures. For short term control, we show how to identify potential bottlenecks that will develop under increased workloads and determine how to release such bottlenecks by adding test equipment or rework stations. In order to meet the long term goal of continuously decreasing defect levels, we propose a budget allocation method for improvement projects such as the redesigning the circuit boards.

Acknowledgements

The motivation for this work was originated on the production floor of Tadiran's Telecommunications Group in Petah Tiqva, Israel. The first author would like to thank the School of Management, Binghamton University, for providing summer support for this project. Thanks also go to a referee for useful comments, Professor Jeongseob Kim for his assistance in simulation, and Mark Vroblefski for his help in typing the manuscript.

References

Agnihothri, S.R. (1987), "A queueing system with an imperfect server", ORSA/TIMS National Meeting, St. Louis, MO.

Agnihothri, S.R., and Kennet, R. (1991), "The impact of defects on a process with rework", Working Paper No. 91-210, School of management, Binghamton University.

Çinlar, E. (1975), *Introduction to Stochastic Processes*, Prentice-Hall, Englwood Cliffs, NJ.

Disney, R.L., and Hannibalsson, I. (1977), "Network of queues with delayed feedback", *Naval Research Logistics Quarterly* 24, 281–291.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, Wiley, New York.

Foley, R., and Disney, R. (1983), "Queues with delayed feedback", *Advances in Applied Probability* 20, 162–182.

Hoadley, B. (1986), "A zero defects paradigm", ASQC Quality Congress Transaction, Anaheim.

Robinson, L.W., McClain, J.O., and Thomas, L.J. (1990), "The Good, the Bad and the Ugly: Quality on an assembly line", *International Journal of Production Research* 28, 963–980.

Seidmann, A., and Nof, S. (1985), "Unitary manufacturing cell design with random product feedback flow", *IIE Transactions* 17, 188–193.

Seidmann, A., and Nof, S. (1989), "Operational analysis of an autonomous assembly robotic station", *IEEE Transactions on Robotics and Automation* 5, 4–15.

Seidmann, A., Schweitzer, P., and Nof, S. (1985), "Performance evaluation of flexible manufacturing cell with random multiproduct feedback flow", *International Journal of Production Research* 23, 1171–1184.