Optimizing the Quality Control Station Configuration

Michal Penn and Tal Raviv
Faculty of Industrial Engineering and Management
Technion, Haifa, Israel

October 2004, Revised April, September 2006

E-mail: mpenn@ie.technion.ac.il, talraviv@eng.tau.ac.il

Abstract

We study unreliable serial production lines with known failure probabilities for each operation. Such a production line consists of a series of stations; existing machines and optional quality control stations (QCSs). Our aim is to decide on the allocation of the QCSs within the assembly line, so as to maximize the expected profit of the system. In such a problem, the designer has to determine the QCS configuration and the production rate simultaneously. The profit maximization problem is approximated assuming exponentially distributed processing times, Poisson arrival process of jobs into the system and the existing of holding costs. The novel feature of our model is the incorporation of holding costs which significantly complicated the problem. Our approximation approach uses a branch and bound strategy that employs our fast dynamic programming algorithm for minimizing the expected operational costs for a *given* production rate as a subroutine. Extensive numerical experiments are conducted to demonstrate the efficiency of the branch and bound procedure for solving large scale instances of the problem and for obtaining some qualitative insights.

"Ever increasing quality is mandatory-not only for corporate profitability-but also for corporate survival"

Inman, Blumenfeld, Huang and Li [10].

1 Introduction

A multistage manufacturing system integrates several successive manufacturing stages (machines) to fabricate products. Producing high quality products at low cost is always

one of the concerns, and thus production costs and rate as well as inspection costs should receive high attention.

Inspection of a product is performed at various stages of its manufacture to assure, or increase, the quality of the product before it is used in final applications. The general inspection allocation problems in manufacturing systems contain several sub-problems such as; optimal allocation of inspection stations within a manufacturing system, finding optimal percentage of the total number of components to be inspected, finding the best action among several options such as rework, repair or scarpping, finding the proper inspection limits to determine the conforming (non-conforming) products, and others. Many researches have studied various inspection effort allocation problems assuming certain conditions about the production system and a characteristics of the inspection process. Models and optimization algorithms for the problem of installing inspection stations are dated back to 1965, see Lindsay and Bishop [14]. A survey on the problem of optimal allocation of inspection stations (referred to also as quality control stations (QCSs)) in multistage systems, appears in Raz [20].

Strategic allocation of inspection stations in multistage production systems was studied before under various assumptions; some of these assumptions are listed below. Some studies focused on serial production processes (e.g Chakravarty and Shtub [4], Shiau [21], and Bowling et al. [3]) while others on non-serial ones (e.g. Elmaghraby [7]). Some papers assumed a constant known acceptance probability at each stage of the assembly (e.g. [4], [7], Kogan and Raz [13]) while others studied nonconstant probabilities, such as in a deteriorating process (e.g. Ben-Daya and A. Rahim [2]). Some of the models assume versatile inspection stations, in particular the case where each station can detect defects from all preceding operations (e.g. [14], Yum and McDowell [23], Bai and Yun [1]) while others assume that each inspection station can detect defects coming only from the immediately preceding operation (e.g. Rebello et al. [19], Kakade et al. [11]). In addition, inspection can be an either/or decision (e.g. [4], [7]), or one with varied inspection limits (e.g. Chen and Thornton [5], [21]). The applied inspection policy can be 100% inspection (e.g. Bowling [3]) or a sampling inspection (e.g. [11]). Furthermore, the corrective action used rework, scrap or others can be modeled. In some works such as [5], [21] and [11], the corrective actions used are rework or scrap, while in others only the scarpping action is allowed (e.g. implicitly in [13]). Also, some papers (e.g. [21]) assume some limitations and constraints on the inspection availability, while in many other papers these limitations are not taken into consideration. Some papers assume error-free inspection (e.g. [14], [11]) while others assume Type I error, (conforming part is rejected) and Type II error (non-conforming part is accepted) (e.g. [21], [2], [1]).

In our model we assume a serial production process with exponentially distributed processing and inspecting times and Poisson arrival process of jobs into the system. Also, at each installed inspection station an either/or decision is to be made, and that the scrap

action is the only selected corrective action allowed. We further assume a 100% inspection policy with error-free inspections (in high volume production, inspections are generally performed by automated visual inspection systems which are highly reliable, consistent and accurate, and thus such an assumption is reasonable) and that each inspection station can detect defects from all preceding operations. In addition, a constant known acceptance probability at each stage is assumed. Such a probability can be estimated using some reliability engineering techniques. See, for example, Elsayed [8] for various methods for estimating such probabilities.

As was previously observed in several studies, the allocation of inspection stations along a production line (termed a QCS configuration) affects the throughput of the system. This observation was pointed out in Rebello et al. [19] where two objective functions of minimizing the cost and maximizing the yield, that is, the percentage of fault-free units leaving the system, were combined in two different ways. First by maximizing the yield under budget constraints and then by minimizing the ratio of the total cost to yield. The first problem is NP-hard and the authors solved it using enumerative techniques, while for the second problem they developed a polynomial time algorithm. Kakade et al. [11] and [1], under different models than ours, assumed that if the inspection operation is a bottleneck in the assembly line, than the inspection cost is due to the reduction of the throughput rate of the line. In [11] the aim is to optimize the cycle time and used simulated annealing procedure to solve the problem.

We observe that under our assumptions, if the first machine is a bottleneck, then the installation of any QCS along the line has no effect on the throughput of conforming parts, but it still may reduce the production cost. On the other hand, if the first machine is not a bottleneck, then installing some QCSs along the line may increase the rate of the throughput of conforming parts, and thus may increase the profit. This last observation was utilized in the algorithms developed in this paper and the ones we have presented in [18].

Furthermore, it turns out that the chosen QCS configuration substantially affects the quantity of work in process (WIP) within the system and thus the actual costs. This phenomenon was first pointed out in a descriptive manner by Drezner, Gurnani and Akella [9] but, to the best of our knowledge, was never incorporated into an optimization algorithm. Note that WIP is defined as the number of items already started their process but not yet finished their last operation, i.e., raw material waiting for production in front of the first workstation is not considered as WIP.

To demonstrate the effect of QCS configuration on WIP, consider for example a small serial production line with 4 independent machines with identical production rate $\mu = 1$ and an identical success probability of $p = 0.95$. That is, the overall success probability of the system is $0.95^4 \approx 0.815$. If the desired production rate, of conforming products, is 0.774 then the first machine should process new jobs in rate of $\lambda = 0.95 \approx \frac{0.774}{0.815}$. We

assume that the arrival process of raw material to the first machine is a Poisson process. That is, in a steady state, the expected length of time a product is within the system is

$$1 + 3 \times \frac{1}{\mu(1-\rho)} = 1 + 3 \times \frac{1}{1-0.95} = 61$$

and so by Little's law the expected WIP is $L = W \times \lambda = 61 \times 0.95 = 57.95$. Now, the installation of a QCS between the first and the second machine will reduce the arrival rate to the remaining machines on the line to $0.95 \times 0.95 = .9025$. Suppose the inspection rate of the installed QCS is 2, then the expected time an item stays in the system is reduced to

$$1 + \frac{1}{2(1-\frac{0.95}{2})} + 3 \times \frac{1}{1-0.9025} \approx 32.82.$$

Thus, the expected WIP in the system is reduced by some 46% to $32.82 \times 0.95 = 31.18$. Another benefit gained by removing non-conforming items from the line is saving of processing costs. On the other hand, clearly, the inspection device incurs it own costs and so this trade off should be considered.

In this paper we define and solve a QCS configuration model of a serial production line where QCSs are to be installed along the line, and present a method to analyze and optimize the performance of such a system. Two optimization problems are considered: minimization of the expected operational cost under a given production rate and maximization of the expected profit where the QCS configuration and the production rate are to be determined concurrently. The novel feature of this study is the introduction of holding cost into the optimization problem. We note that this substantially increases the difficulty of the problem.

In [18] Penn and Raviv have presented an $O(n^4)$ algorithm, where n stands for the number of machines along the line, to maximize the profit, in a steady state, of a production system with inspection stations. The algorithm in [18] works for any stationary stochastic arrival process, any processing and inspection time distributions and under the assumption that **no** holding costs incurred by work in process. The basic idea behind the polynomial algorithm in [18] is the observation that the size of the set of possible values of optimal production rates is relative small and an efficient method to identify this set. Unfortunately, the above idea fails to hold if holding costs are incorporated into the system, and thus different techniques had to be utilized for solving the more complicated problem discussed in this paper.

The cost minimization problem is solved using a simple fast polynomial time dynamic programming algorithm assuming exponentially distributed processing times and Poisson arrival process of jobs into the system. We note that the assumption of the Poisson and exponential distributions are commonly used in Queuing Theory (e.g., [12]) and can also be served as good approximations for other processing times and arrival process distributions in serial production lines (e.g. [16], [22]). The profit maximization problem is

approximated under the same assumptions using a branch and bound strategy that employs the dynamic programming algorithm as a subroutine. The main contributions of this paper are twofold: Incorporation of holding costs in the objective function of the minimization and maximization problems and the efficient algorithms developed for solving the above problems; especially the branch and bound method for the maximization problem. Extensive computational experiments were conducted to demonstrate the efficiency of the proposed procedures. These experiments also show the advantage of the above method as a good heuristic for non-Poisson arrival processes and, in particular, for the deterministic arrival process (constant inter-arrival times), which is a more applicable dispatching policy for serial production lines. Also, additional computational experiments were conducted for a qualitative analysis of the system. These experiments show the effect of the production and inspection rates on the number and locations of the installed QCSs as well as pointing out on some interesting manegerial insights.

The paper is organized as follows: In Section 2 the cost minimization problem is defined and a method for calculating the operational cost of a system with a *given* QCS configuration is presented. Section 3 presents a simple polynomial time dynamic programming algorithm to obtain a minimal cost QCS configuration for a given *pre-specified* production rate under the assumption of Poisson arrival process of jobs into the system and exponential distribution of the processing times. Section 4 is devoted to the main contribution of the paper, the definition of the profit maximization problem and the description of the developed branch and bound approximation strategy for solving it. In Section 5 we present our numerical experiments that show the practical efficiency of our algorithms, demonstrate their applicability to the case of constant inter-arrival times and describe some qualitative insights.

2 The Model

In this section we start with the definition of a QCS system and then, given such a system, we find out some simple conditions, in terms of the arrival rates, under which the QCS system is stable, as well as determine the operational cost of the system.

2.1 System Description

Our assumptions and notations are summarized below.

1. A stream of identical jobs arrives at the first machine, according to a stationary stochastic Poisson process.

2. Each machine starts processing its next operation as soon as it turns ready and a job is available in its feeding buffer.

3. Each machine has its own exponential distributed processing times of the jobs on that machine, and has its own operational cost for processing each job.
4. The probability of producing a non-conforming product on machine i, given it arrives conforming to that machine, is known and constant.
5. The success and the failure events on each machine are independent.
6. A QCS can be assigned immediately after each machine to perform 100% inspection.
7. For each QCS, the inspection operation times are exponential i.i.d random variables and each job requires the same inspection cost on a given QCS.
8. A QCS after any machine can detect any of the previously caused defects, if such exist.
9. Each inspection process is error-free.
10. A non-conforming product will be discarded if realized as such.

The following types of cost are taken into consideration:

1. Variable cost per each operation of the machines and the installed QCSs.
2. Capital cost of the installed QCSs[1].
3. Holding cost of the work in process.
4. Penalty cost of delivered non-conforming products.

2.2 Notations

N The number of machines in the serial production line.

M_i The i^{th} machine in the production line.

QC_i The Quality Control Station (QCS) that is installed immediately after M_i.

x_i Mean processing time of machine M_i.

x'_i Mean inspection tine of QC_i.

p_i Probability of a conforming job entering machine M_i to remain conforming after leaving the machine.

[1]The capital cost of the machines are considered as sunk costs and thus are not incorporated in our optimization problem.

q_{ij} Probability that a **conforming** job leaving machine i remains conforming after leaving machine j. We assume the q_{ij}'s are known for any pair of indices i and j. By convention $q_{ii} = 1$. Clearly, if we assume independence of the p_is then $q_{ij} \equiv \prod_{l=i+1}^{j} p_l$.

q_{0i} Unconditional probability that a conforming job remains conforming by the time it leaves machine i.

c_i Cost of performing an operation on machine M_i (variable cost).

c'_i Cost of performing an operation on QC_i, if such installed (variable cost).

f'_i Fixed (capital) cost per unit of time of operating QC_i.

h_i Holding cost per unit of time an item waiting to be processed on M_i.

h'_i Holding cost per unit of time an item waiting to be inspected by QC_i.

r_B Penalty cost of non-conforming product delivered from the system.

$r(x)$ Revenue function, the total revenue per unit of time for production rate x.

$\mathbf{Y}$ - A QCS configuration. Sometimes refereed to as a characteristic vector, with $Y_i = 1$ (resp., $Y_i = 0$) indicating that QC_i is (resp., is not) installed. In other cases, we refer to it as a set of locations $Y \subseteq \{1, \ldots, N\}$ with $i \in Y$ implying that QC_i is installed.

$L_i(Y)$ - The location of the last installed QCS before machine M_i in configuration Y, or 0 if there is no such QCS. ($L_i(Y) \equiv \max\{j \in Y \cup \{0\} | j < i\}$).

The tuple $(\mathbf{p}, \mathbf{x}, \mathbf{x}', \mathbf{c}, \mathbf{c}', \mathbf{f}', \mathbf{h}, \mathbf{h}', \mathbf{r_B})$ with an arrival rate of value a is referred to as a *QCS(a) cost minimization problem*. A system defined by a $QCS(a)$ problem and a given configuration Y is denoted by $(QCS(a), Y)$. Given a $(QCS(a), Y)$ system, $A_i(Y, a)$ stands for the arrival rate of the jobs into machine i.

2.3 Preliminaries

A (QCS(a),Y) system is said to be *stable* if the expected amount of work in process in each of the system's buffers converges to some constant as t, the time the system operates, goes to infinity. The following two conditions for stability can be easily derived.

Observation 2.1 *Let (QCS(a), Y) be a given system and assume the following inequalities hold*

$$a \cdot q_{0,L_i(Y)} < \frac{1}{x_i} \qquad \forall i = 1, ..., N \tag{1}$$

and

$$Y_i \cdot a \cdot q_{0,L_i(Y)} < \frac{1}{x'_i} \qquad \forall i = 1, ..., N, \tag{2}$$

then the system is stable and for each i, $A_i(Y, a) = a \cdot q_{0,L_i(Y)}$.

Observation 2.2 *Given a* $(QCS(a), Y)$*, the system is stable if and only if*

$$a < \min \left\{ \min_i \frac{1}{x_i \cdot q_{0,L_i(Y)}}, \min_{i:Y_i=1} \frac{1}{x'_i \cdot q_{0,L_i(Y)}} \right\}.$$

Note that Observations 2.1 and 2.2 hold for any stationary arrival process. However, under exponential processing and inspection times and Poisson arrival process our system can be modelled as a tandem Jackson Network, implying the arrival process at any station is a Poisson process. Thus, the queue length process in front of any station is as in an $M/M/1$ system with arrival rate $a \cdot q_{0,L_i(Y)}$ and service rate $\frac{1}{x_i}$. The expected waiting time (including the service time) is,

$$\frac{1}{\frac{1}{x_i} - a \cdot q_{0,L_i(Y)}}.$$

Hence, the total expected cost per item handled by a machine for a given QCS configuration Y and arrival rate a is explicitly given by,

$$\mathcal{C}_i(Y, a) = c_i + \frac{h_i}{\frac{1}{x_i} - a \cdot q_{0,L_i(Y)}}. \tag{3}$$

Similarly, the expected cost incurred by the QC_i is

$$\mathcal{C}'_i(Y, a) = Y_i \left(c'_i + \frac{h'_i}{\frac{1}{x'_i} - a \cdot q_{0,L_i(Y)}} \right). \tag{4}$$

Hence, the expected operating cost of the system per time unit is given by

$$\begin{array}{rl} C(Y, a) \equiv & a \cdot \sum_{i=1}^{N} q_{0,L_i(Y)} \cdot [\mathcal{C}_i(Y, a) + \mathcal{C}'_i(Y, a)] + \sum_{i=1}^{N} Y_i f'_i \\ & +(1 - Y_N) \cdot a \cdot q_{0,L_Y(N)} \cdot (1 - q_{L_Y(N),N}) \cdot r_B. \end{array} \tag{5}$$

Note that $\mathcal{C}_i(Y, a)$ is defined only for arrival rates $a < (x_i \cdot q_{0,L_i(Y)})^{-1}$ and thus the queue in front of machine i is finite. Similarly, if $Y_i = 1$, then $a < (x'_i \cdot q_{0,L_i(Y)})^{-1}$ is the domain of $\mathcal{C}'_i(Y, a)$, otherwise the domain is $\mathbb{R}_+$. Clearly, the domain of $C(Y, a)$ is the intersection of these domains. Note that if no confusion arises, we assume that the above three functions are defined over $\mathbb{R}_+$ and obtain the value ∞ outside their actual domains.

3 Optimal QCS Configuration - The Known Arrival Rate Case

In this section we turn to solve the combinatorial optimization problem of determining an optimal QCS configuration that minimizes the cost function over all 2^N possible configurations. Not surprisingly, the dynamic programming approach is suitable for solving the minimization problem and thus it was chosen. The dynamic programming algorithm proposed solves the problem in time complexity of $O(N^2)$ and is used as a subroutine in our main algorithm for solving the maximization problem. The optimal value of the minimization problem is denoted by

$$C^*(a) = \min_{Y \in \{0,1\}^N} C(Y, a)$$

and an optimal QCS configuration that materializes this cost is denoted by

$$Y^*(a) = \text{argmin}_{Y \in \{0,1\}^N} C(Y, a).$$

If there exists more than one QCS configuration that minimizes the expected cost, then $Y^*(a)$ represents any arbitrarily chosen optimal configuration. The algorithm presented below follows similar lines of some previously presented algorithms such as Lindsay and Bishop (see [14]), but in addition accommodates arrival rates and WIP costs.

Algorithm 3.1 ***The QCS(a) Dynamic Programming Algorithm***
Input: A $QCS(a)$ problem defined by $(\mathbf{p}, \mathbf{x}, \mathbf{x}', \mathbf{c}, \mathbf{c}', \mathbf{f}', \mathbf{h}, \mathbf{h}', \mathbf{r_B})$ and an arrival rate a. The recursion function $g_i(L_i; Y_i)$ denotes the total cost that incurred by the tail of the system that begins at machine i, assuming QC_{L_i} is the last installed control station before machine i, and given the existence ($Y_i = 1$) or absence ($Y_i = 0$) of QC_i. Here, L_i is a state variable and Y_i is a decision variable. For all $i = 1, \ldots, N-1$, the function g_i is constructed by the following recursive relation:

$$g_i(L_i; Y_i) = a \cdot q_{0,L_i} \cdot \left[c_i + \frac{h_i}{\frac{1}{x_i} - a \cdot q_{0,L_i}} + \left(c_i' + \frac{h_i'}{\frac{1}{x_N'} - a \cdot q_{0,L_i}} \right) \cdot Y_i \right] + f_i' Y_i + g_{i+1}^* \left(L_{i+1}(L_i, Y_i) \right) \quad (6)$$

for $a \cdot q_{0,L_i} \in [0, \min\{\frac{1}{x_i}, \frac{1}{x_i'} + (1 - Y_i) \cdot \infty\})$, and $g_i(L_i; Y_i) = \infty$ otherwise. We use the following transition function:

$$L_{i+1}(L_i, Y_i) = \begin{cases} L_i & Y_i = 0 \\ i & Y_i = 1. \end{cases} \quad (7)$$

The initial condition for g_N is

$$\begin{aligned} g_N(L_N; Y_N) = \; & a \cdot q_{0,L_N} \cdot \left[c_i + \frac{h_N}{\frac{1}{x_N} - a \cdot q_{0,L_N}} + \left(c_N' + \frac{h_N'}{\frac{1}{x_N'} - a \cdot q_{0,L_N}} \right) \cdot Y_N \right] + \\ & f_N' Y_N + (1 - Y_N) \cdot a \cdot q_{0,L_N} \cdot (1 - q_{L_N,N}) \cdot r_B \end{aligned} \quad (8)$$

for $a \cdot q_{0,L_N} \in [0, \min\{\frac{1}{x_N}, \frac{1}{x_N'} + (1 - Y_N) \cdot \infty\})$ and $g_N(L_N; Y_N) = \infty$ otherwise.

The function g_i^* is constructed by

$$g_i^*(L_i) = \min_{Y_i} g_i(L_i; Y_i). \tag{9}$$

If, at any stage, $g_i^*(L_i) = \infty$ for all $L_i = 0, \ldots, i-1$, then the arrival rate a is not feasible for the problem and the algorithm terminates. The optimal decision at each step i (whether to install a QCS at position i or not) is determined by

$$Y_i^*(L_i) = \text{argmin}_{Y_i} g_i(L_i; Y_i). \tag{10}$$

■

We note that the $QCS(a)$ problem can also be formulated as the shortest path problem on a directed graph, with a complete underlying graph, on $N+2$ nodes denoted by $\{0, ..., N+1\}$. The cost c_{ij} of an edge (i, j), $i < j$, $\{i, j\} \subset \{1, ..., N\}$, is the total cost incurred by processing and by the inventory on all the machines indexed by $i+1, ..., j$ and QC_j; assuming QC_i is the last QCS before QC_j. Similarly, c_{0j} denotes the costs incurred by machines $M_1, ..., M_j$ and by QC_j, assuming QC_j is the first installed QCS. In addition, $c_{i,N+1}$ denotes the total cost, including penalty for non-conforming product, incurred by machines $M_{i+1}, ..., M_N$, assuming QC_i is the last QCS on the line. We observe that a shortest path from node 0 to node $N+1$ induces an optimal QCS configuration. Based on the above, we conclude that Algorithm 3.1 is merely an implementation of Dijkstra's Algorithm simultaneously with calculation of the edge costs.

Proposition 3.1 *The time and the space complexity of Algorithm 3.1 is* $O(N^2)$.

Proof. Note that calculating all edge costs can be done in $O(N^2)$ time and space. This coupled with the analogy between the $QCS(a)$ problem and the shortest path problem as well as between Algorithm 3.1 and Dijkstra Algorithm [6], imply the correctness of the proposition. ■

4 The Profit Maximization Problem

In this section we extend the problem to capture the case when the arrival rate is a decision variable rather than part of the input. Thus, our aim is to optimize the QCS configuration and the production rate simultaneously. The proposed branch and bound strategist partitions the domain of the arrival rate in search for an optimal arrival rate and utilizes the $QCS(a)$ Dynamic Programming Algorithm as a subroutine.

4.1 Problem Definition

We consider a QCS problem defined by $(\mathbf{p}, \mathbf{x}, \mathbf{x}', \mathbf{c}, \mathbf{c}', \mathbf{f}', \mathbf{h}, \mathbf{h}', \mathbf{r_B})$ coupled with a revenue function $r(x)$. The function $r(x)$ describes the expected revenue per time unit as a function of the departure rate of the conforming products from the last machine. If the firm plays in a competitive market, then this function is linear and homogeneous as the production rate of the firm admits no influence on the market price. For the discussion below we use a weaker assumption that the revenue function is K-Lipschitz continuous over the relevant domain. That is, it is continuous and differentiable almost every where with its derivative bounded above by some finite constant K. This assumption is not very restrictive since in practice the average revenue is hardly affected by small changes in the supply.

Our extended profit maximization problem, denoted by QCS, is to determine the arrival rate and the QCS configuration simultaneously in order to maximize the expected profit per time unit from the system in steady state. For a given QCS configuration Y and an arrival rate a, the total profit per time unit is,

$$P(Y, a) = \quad r(q_{0,N} \cdot a) - C(Y, a) \tag{11}$$

Now, the optimal profit for a given arrival rate a is just

$$P^*(a) = r(q_{0,N} \cdot a) - C^*(a)$$

and hence can be easily calculated using Algorithm 3.1. Therefore, our extended parametric problem can be formulated as

$$\max_a P^*(a) \tag{12}$$

and in this form, it is reduced to an optimization problem in a single continuous variable. Clearly, (12) always admits a finite optimal solution since it is always feasible for $a = 0$ and a is bounded above for any possible QCS configuration, see Observation 2.2.

4.2 Some Properties of the Objective Function

Note that in general, $P^*(a)$ is not concave or unimodal, hence standard line search techniques will not solve (12). This statement holds even for linear $r(a)$. In the sequel we explore some useful properties of $C^*(a)$ and $P^*(a)$ that form the basis for our approximation method. The proofs of Lemmas 4.1 and 4.2 are rather technical and lengthy, thus we have chosen to present them in the Appendix.

Lemma 4.1 *The function $C^*(a)$ is continuous, piecewise convex and piecewise differentiable with respect to a.*

Lemma 4.2 *For any pair of points $a_1 \geq a_0$ in the domain of $C^*(a)$, if $C^*(a)$ is differentiable at a_1 then its derivative is bounded below by*

$$\frac{\partial C^*(a)}{\partial a}(a_1) \geq \sum_{i=1}^{N} \left\{ q_{0,i-1} \left(c_i + \frac{h_i}{\frac{1}{x_i} - a_0 \cdot q_{0,i-1}} \right) + \frac{a_0 \cdot h_i \cdot q_{0,i-1}^2}{\left(\frac{1}{x_i} - a_0 \cdot q_{0,i-1} \right)^2} \right\} \equiv \zeta_{a_0}.$$

Clearly, the slope of the function $P^*(a_1)$ for any point $a_1 > a_0$ is bounded above by the Lipschitz constant K minus the lower bound on the slope of $C^*(a_0)$ obtained by Lemma 4.2. Thus we have the following corollary which is essential for our branch and bound procedure.

Corollary 4.3 *Let $\gamma_{a_0} = K - \zeta_{a_0}$. For any pair of feasible arrival rates a_0 and a_1 such that $a_0 < a_1$,*

$$P^*(a_1) \leq P^*(a_0) + (a_1 - a_0) \cdot \gamma_{a_0}.$$

4.3 A Branch and Bound Algorithm for the Profit Maximization Problem

Based on Corollary 4.3 and on Algorithm 3.1 we present below Algorithm 4.1 which is a branch and bound approximation procedure for solving the profit maximization problem. Let $\mathcal{A} > 0$ and $\mathcal{R} \geq 0$ be the desired absolute and relative optimality errors, respectively. That is, if the value of the optimal solution is OPT, then our algorithm terminates with a solution which is at least $\min\{OPT \cdot (1 - \mathcal{R}), OPT - \mathcal{A}\}$.

The algorithm maintains a list of *active segments* which are continuous subsets of the set of feasible rates. For each segment in the list we store the start point (a_0), the end point (a_1), the optimal profit at a_0 ($P^*(a_0)$) and an upper bound on the expected profit from the system for any arrival rate $a \in [a_0, a_1)$. The list is ordered by the upper bounds.

We start with a single segment that contains all feasible arrival rates. In any iteration, the algorithm removes a segment from the list. The optimal configuration is calculated for the middle point of the segment and the segment is divided into two segments of equal length. If the profit at the middle point is higher than the best known solution, then it is stored as the current best known solution. Next, tighter upper bounds are calculated on the value of the optimal solution within each of the two newly created segments. The segments are returned to the list if their upper bounds are sufficiently larger than the best known solution.

Algorithm 4.1 ***The QCS Branch and Bound Procedure***

Input: a QCS problem $(\mathbf{p}, \mathbf{x}, \mathbf{x}', \mathbf{c}, \mathbf{c}', \mathbf{f}', \mathbf{h}, \mathbf{h}', \mathbf{r_B}, \mathbf{r}(\mathbf{x}))$, optimality errors $\mathcal{A}$ and $\mathcal{R}$.

Initialization: Start with a list of active segments $\mathcal{L}$ that contains a single segment $\left[0, \min_i \left\{ \frac{q_{0,i-1}}{x_i} \right\} \right)$. Set current best known solution $a^* = 0$ (with value $P^*(a^*) = 0$) and set the upper bound relative to this segment to be $\frac{\gamma_0}{x_1}$.

Step1: Remove from $\mathcal{L}$ a segment $[a_0, a_1)$ of maximum upper bound. Let $a_0' = \frac{a_0+a_1}{2}$.

Step2a: Construct a new segment $[a_0, a_0')$. A lower bound on the maximum profit within this segment is given by $P^*(a_0)$. Use Corollary 4.3 to calculate the upper bound $P^*(a_0) + \gamma_{a_0}(a_0' - a_0)$ that associates with the segment. If this upper bound exceeds $\min\{P^*(a^*) \cdot (1 + \mathcal{R}), P^*(a^*) + \mathcal{A}\}$ add the new segment $[a_0, a_0')$ to $\mathcal{L}$.

Step2b: Construct a new segment $[a_0', a_1)$. Calculate $P^*(a_0')$ and its corresponding QCS configuration using Algorithm 3.1. $P^*(a_0')$ is a lower bound on the maximum profit within the segment. If $P^*(a_0') > P^*(a^*)$ then store it as the new best known solution, i.e., set $a^* = a_0'$. Use Corollary 4.3 to calculate the upper bound $P^*(a_0') + \gamma_{a_0'}(a_1 - a_0')$ that associates with the segment. If this upper bound exceeds $\min\{P^*(a^*) \cdot (1 + \mathcal{R}), P^*(a^*) + \mathcal{A}\}$ add the new segment $[a_0', a_1)$ to $\mathcal{L}$.

Step3: If $\mathcal{L}$ is empty then stop and return the current best known solution. Otherwise goto step 1.

Theorem 4.4 *Algorithm 4.1 terminates in a finite number of iterations and achieves an approximate solution of value* $\min\{OPT - \mathcal{A}, OPT \cdot (1 - \mathcal{R})\}$.

Proof. Observe that at any step of the algorithm, a segment is removed from the list and two, one or zero new segments of half length of the removed one, are added to the list. The length of any segment in the list is bounded below by $\frac{\mathcal{A}}{2\gamma_0}$ and hence at some point the list becomes empty and the algorithm stops. Clearly from the algorithm description the value of the solution yielded by the algorithm is at least $\min\{OPT - \mathcal{A}, OPT \cdot (1 - \mathcal{R})\}$. ■

Remark 4.5 If the absolute error is set to $\mathcal{A} = 0$, then Algorithm 4.1 still converges to the optimal solution but the convergence process may be infinite, regardless of the magnitude of $\mathcal{R}$. However, this has no practical implication since there is always some absolute error imposed by the floating point accuracy of the computer.

Remark 4.6 Note that although throughout this paper we assumed the Poison arrival process, the developed methods produce good approximate solution (in the heuristic sense) to other arrival processes. This phenomenon is indicated in the literature and was also

observed by our numerical experiments as described in the next section. It is widely believed that in a tandem of N stations, if the arrival process is stationary and ergodic with a rate of α and the system is stable, then the departure process from the n^{th} station converges to a Poisson process with a rate of α as $n \to \infty$. This conjecture is known as Reiman and Simon conjecture and was partially proved by Mountford and Prabhakar [16] for the case of identical stations. Furthermore, a simulation study conducted by Suresh and Whitt [22] indicates that the convergence rate, in terms of the number of machines in the tandem, is fairly high if the arrival process admits low variability and in particular when the arrival process is deterministic (e.g., the inter-arrival times between any successive arrivals are constant).

5 Computational Results

Our computational study is divided into two parts. The first part is devoted to the computational analysis of our algorithms, while the second part concentrates on the qualitative insights of the obtained solutions.

The first part of our study consists of three sets of experiments described in the subsections below. In 5.1 we test the applicability of Algorithms 3.1 and 4.1 for very large instances of the profit maximization problem. We solved instances with 1000 machines in a very short time. In 5.2 We show that the optimal QCS configuration for a Poisson arrival process remains nearly optimal when the Poisson arrival process is replaced by a deterministic one of the same rate. We compare the results obtained from Algorithm 3.1 with the simulation results of all 2^N possible QCS configurations. For obvious reasons, this experiment is restricted to short production lines. We have tested it on eight machines lines. The created test problem instances differed by the following three criteria:

1. **Success probabilities**: Groups denoted by L possess relatively low success probabilities while H denotes those possess high ones. The success probabilities of the 'H' instances were generated such that $q_{0,N} = \prod_{i=1}^{N} p_i \approx 0.8$ and for the 'L' instances it is $q_{0,N} \approx 0.4$.

2. **Tendency of the processing rates along the line**: For instances denoted by R the expected processing times were sampled from a common distribution (i.i.d) and for those denoted by I, the expected processing times were generated in a way that insures strictly increasing processing times in i, the index of the station.

3. **Tendency of the holding costs along the line**: In problems denoted by R, the holding costs $\mathbf{h}$ and $\mathbf{h}'$ were taken from a common distribution (i.i.d) for all stations and for those denoted by I (resp., D) the holding costs were generated to be monotonously increasing (resp., decreasing) in i, the machine index.

There are 12 combinations of these criteria. A problem instance is denoted by three letters and the number of machines. For example, a problem denoted by $HRD8$ is one with High success probabilities, arbitrary Random processing times, Decreasing holding costs and 8 machines. These 12 combinations represent variety of systems setups. Note that we did not include in our data set problems with decreasing processing times. Such systems are not likely to be used in real life and in many cases they are easier to analyze because the configuration of QCSs located downstream the bottleneck station admits no effect on line throughput, as long as the inspection time is shorter than processing time.

In 5.3 we present the second part of our computational analysis. Here we study the nature of optimal QCS in two types of regular systems and show how the optimal configurations are affected by diverse production rates.

5.1 The Efficiency of the Algorithms

In order to check the applicability of Algorithm 4.1 (the Profit Maximization Algorithm), we defined two revenue functions $r(x) = \alpha \cdot x$ and $r(x) = \beta\sqrt{x}$. The constants α and β were randomly selected in a manner that assures the existence of a profitable solution; this, in order to avoid trivial instances. We coupled this two revenue functions with the 12 above combinations which work out for 24 types of problems. We randomly generated 1200 instances of the problem with 1000 machines each, 50 instances for each problem type.

Our algorithm was applied for these problems; The relative optimality error was set to 0.001 (0.1%) and the absolute optimality gap was set to 0 (which practically means that the absolute error is set to the numerical accuracy of the computer). The running times in seconds and the number of iterations (calls to Algorithm 3.1) were collected. Statistics of this experiment are presented in Table 1.

The algorithm was implemented in Microsoft Visual C++ with LEDA (see [15]) on an Intel Pentium 4, 2Ghz CPU with 512Mb RAM. The source code and data set are available from our site http://www.talraviv.net/ under Publications.

From Table 1, it is apparent that Algorithms 3.1 and 4.1 can be employed to solve efficiently the problems presented in the paper under diverse sets of conditions and for any reasonable size. In particular, we believe that 1000 machines is a reasonable upper bound on the size of serial production lines encountered in real life and the relative optimality guarantee of 0.1% is in most cases more accurate than the problem parameters. Note that Algorithm 3.1 is a subroutine called numerous times in the solution process of Algorithm 4.1. Thus, the problem of determining an optimal QCS configuration for a given arrival rate in a thousand machines line is solved within a fraction of a second.

Model	$r(x) = C \cdot x$			$r(x) = C \cdot \sqrt{(x)}$		
	Average time	Worst time	Average # iterations	Average time	Worst time	Average # iterations
LII1000	0.997	1.673	8.54	3.667	4.567	30.46
LID1000	1.076	1.783	8.70	3.896	4.596	31.14
LIR1000	1.024	1.562	8.44	3.767	4.507	30.52
LRI1000	2.004	2.604	12.00	4.542	6.069	32.60
LRD1000	1.937	2.774	11.44	4.562	5.498	33.06
LRR1000	1.986	2.603	11.82	4.542	5.758	32.52
HII1000	1.716	2.413	11.94	4.640	5.438	35.54
HID1000	1.711	2.494	12.26	4.717	5.978	36.64
HIR1000	1.751	2.473	12.10	4.740	5.899	36.10
HRI1000	1.973	3.265	12.46	4.561	5.508	33.86
HRD1000	1.913	2.864	12.50	4.581	5.739	34.36
HRR1000	1.928	2.834	12.42	4.489	5.398	33.62

Table 1: The average and worst case running times of Algorithm 4.1 in seconds and the average number of calls to Algorithm 3.1 are presented for the two different revenue functions.

5.2 Deterministic Arrival Process

We believe that optimal QCS configurations for the Poisson arrival process frequently remain optimal or near optimal for non-Poison arrival processes and in particular for the deterministic arrival process. In this section we supply further numerical support for this belief.

Twelve systems of eight machines each based on our 12 categories described at the beginning of this section, were constructed. Each cost minimization problem was solved, using Algorithm 3.1 for three different arrival rates. The arrival rates were selected in order to cover diverse sets of conditions, according to the following method. Observation 2.2 was used to obtain λ_{max}, an upper bound on the feasible arrival rates assuming all eight QCSs are installed. The following arrival rates $\lambda_{low} = 0.5\lambda_{max}$, $\lambda_{med} = 0.8\lambda_{max}$ and $\lambda_{high} = 0.95\lambda_{max}$ were considered.

The procedure recently proposed by Nelson, Swann, Goldsman and Song [17] (NSGS procedure) was used to obtain a near "optimal" configuration for the problem with deterministic arrival process. This procedure finds, with a pre-specified probability $(1 - \alpha)$, a solution which is optimal or within a pre-specified *Indifference Zone* from the optimum. We applied the above procedure with an indifference zone of 2% (of the optimal cost obtained by our algorithm) and $\alpha = 0.05$. It should be noted that NSGS procedure is practical only for very small instances of our problem since the procedure repeatedly runs

numerous simulation sessions for each of the exponentially many possible configurations. In our case, each of the eight machines problem with the above confidence level and indifferent zone, took several minutes to solve.

For the "optimal" solutions obtained by Algorithm 3.1 and by NSGS procedure, further simulations under the deterministic arrival process were conducted, until a relative confidence interval of 0.1% could be obtained.

Table 2 compares the solutions obtained by Algorithm 3.1 with those obtained by the NSGS procedure when both procedures were used for the problem with deterministic arrival process. The values in the "ratio" columns were calculated as follow,

$$100 \times \left(\frac{\text{Cost of the best solution obtained by NSGS procedure}}{\text{Cost of the optimal solution obtained by Algorithm 3.1}} - 1 \right).$$

Problem	Low Rate ($0.5\lambda_{max}$)			Medium Rate ($0.8\lambda_{max}$)			High Rate ($0.95\lambda_{max}$)		
	Ratio	Optimal Configuration		Ratio	Optimal Configuration		Ratio	Optimal Configuration	
		DP 3.1	NSGS		DP 3.1	NSGS		DP 3.1	NSGS
LII8	-	00010001		-	01010101		1.7%	11111101	11110101
LID8	-	00100001		0.4%	01010001	00110001	-	11111001	
LIR8	-1.1%	00100001	00100011	-	00100101		-	10101011	
LRI8	-0.0%	01000001	00100001	-0.4%	01000001	00100001	-0.0%	00100001	00010001
LRD8	-1.2%	00100001	01000001	-1.3%	00100001	01000001	-0.1%	00010001	01010001
LRR8	-	01000001		-	01000001		-	01000001	
HII8	-	00000001		-0.2%	00010001	01000011	-	01010010	
HID8	-	00000000		-1.2%	00100000	0000000	-	00100100	
HIR8	-	00000010		-	00000010		-	00100010	
HRI8	-1.1%	00000001	00000010	-	00000010		-	00000010	
HRD8	-	00000010		-	00000010		-	00000010	
HRR8	-1.6%	00000010	00000001	-2.4%	00000010	00000001	-0.3%	00000010	00000011

Table 2: A comparison between the profit and the QCS configurations for the eight machines system obtained by Algorithm 3.1 and by NSGS procedure.

From Table 2 it is apparent that for our 36 test problems, Algorithm 3.1 returns solutions which are either optimal or very close to optimal. The differences between the solutions obtained by both methods can be partly explained by the estimation error.

Recall that the total costs of a given system under Poisson arrival process and under deterministic one differ only in the holding costs. Thus, for low holding costs it is not surprising that the optimal solutions are similar for both cases. In order to show that the above phenomenon holds also for relatively high holding costs we examined the proportion of holding costs relative to the total expenses. We note that for the instances presented in Table 2 holding costs were a substantial part of the total costs of the optimal solution (26.1% of the total expense on the average with a range of 10.2% to 48.2%).

5.3 Qualitative Insights: Typical QCS Configurations

We turn now to study the effect of production (inspection) rates on the QCS configurations. The executed tests provide us with some insights on the nature of optimal QCS configurations. We consider two systems, each of 100 machines with similar parameters expect for their processing and inspection times. In both systems, and for all the machines, the success probabilities were set to $p_i = 0.99$; the variable processing and inspection costs per unit were set to $c_i = c'_i = 1$; the fixed cost of installing a QCS was set to $f'_i = 0$; the holding cost per unit of time of an item in front of each of the stations was set to $h_i = h'_i = 0.1$; the penalty for delivering a non-conforming item was set to $r_B = 10$. It should be pointed out that in this experiment we are only interested in an optimal QCS configuration for a give production rate. As a result, the price of a conforming product, r_G, has no effect. In System 1, the processing and inspection times are all fixed $x_i = x'_i = 1$ while in System 2 we set the processing times to increase in a rate of 1% per machine. That is $x_i = \frac{x_{i-1}}{0.99}$. In order to make the two systems comparable we normalized the processing times of all machines such that the mean processing time is 1. Also, for each installed QCS, we set its inspection time to equal the processing time of its previous machine on the line. That is, $x'_i = x_i$. Note that, for System 2, maximum production rate can only be achieved by installing a QCS between each pair of machines. On the other hand, in System 1, the bottleneck station is the first machine. Thus, installation of QCS has no effect on the maximum potential production rate of the system. Figures 1 and 2 illustrate optimal QCS configurations in these two systems, each, for 99 possible production rates of 1% to 99% of the maximum possible capacity.

As one would expect, in both systems the number of installed QCSs is non-decreasing with the production rate; optimal solutions seem to be robust to small changes in the rates; optimal solutions are somewhat symmetric in the sense that in each solution, the distances between any pair of consecutive QCSs are approximately the same. System 2 seems to be of a more symmetric pattern. This symmetry is probably due to the symmetric patterns of the production and inspection rates in the systems we studied. Albeit the similarity in the parameters, optimal solutions of System 2 always take more QCSs and this difference increases as the production rate increases. This observation can be explained by the fact that in System 2, if it is possible to remove non-conforming items from the system during the production process, it is reasonable to have the first machines working faster than those down the line. Hence, installation of QCSs helps to reduce costs not only by reducing production costs but also by releasing bottlenecks and reducing holding costs of work in process.

Two conflicting factors affect the decision to install a QCS toward the end of the line. Since the penalty cost imposed on non-conforming items is affected only by the last installed QCS, the values of r_B and the failure probabilities affect the decision to install a QCS toward the end of the line. Indeed, if there is no such penalty, that is $r_B = 0$, then

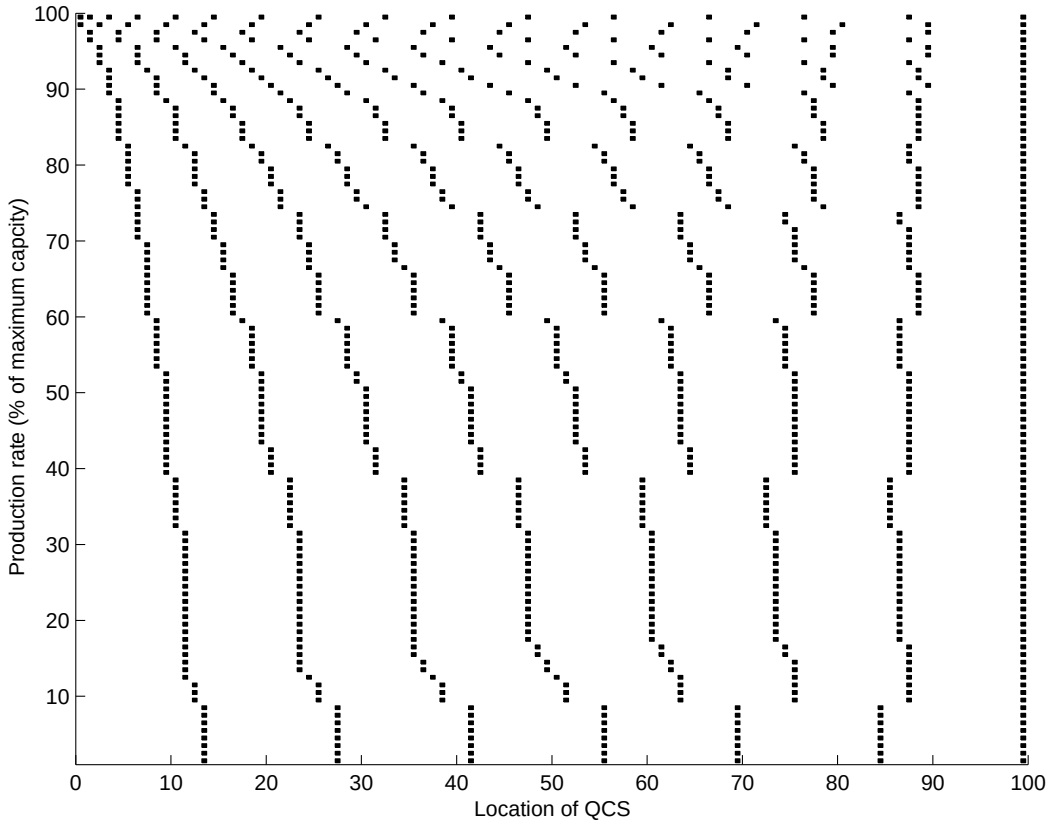

Figure 1: Optimal QCS configurations of System 1 - constant production (inspection) rate. Each row represents an optimal configuration at a given relative rate (production rate/maximum capacity).

optimal solutions tend to be more symmetric as illustrated in Figure 3. However, locating a QCS toward the end of the line has the least effect on the inventory cost and thus makes these locations less attractive for installation of QCSs if holding costs are relatively high.

One should observe that deriving a rule of the thumb for obtaining optimal QCS configurations is hard to achieve. This is because optimal configurations, in real production lines, are very sensitive to many parameters that may vary significantly from station to station along the line. However, an important observation from our experiment above is that optimal QCS configurations are relatively robust to moderate changes in the production rate, especially if the system is not working near its maximum possible capacity. Clearly, such changes are frequently required due to changes in the market.

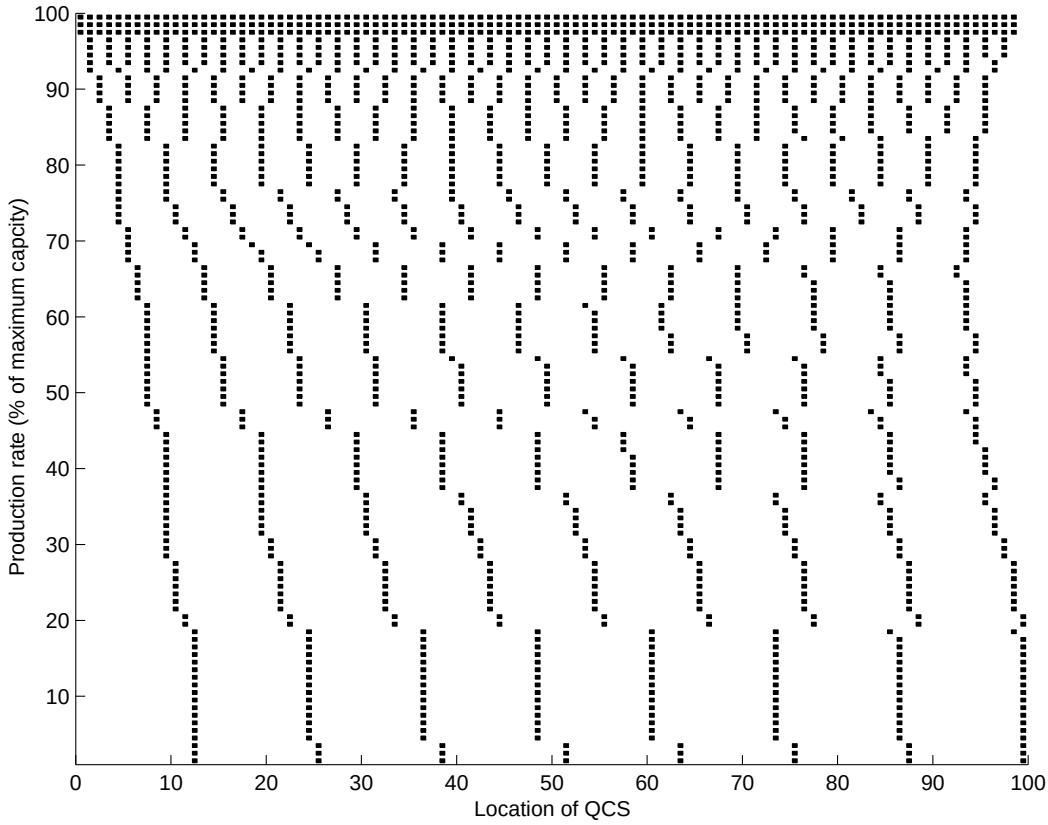

Figure 2: Optimal QCS configurations of System 2 - decreasing production (inspection) rate. Each row represents the optimal configuration at a given relative rate (production rate/maximum capacity).

6 Discussion

In this paper we presented a dynamic programming algorithm and a branch and bound strategy to solve the problem of determining an optimal QCS configuration along a serial production line. Two versions of this problem were considered: minimization of the cost per time unit under a given production rate and maximization of the profit where the QCS configuration and the production rate are to be selected simultaneously. As was pointed out throughout the paper, the latter problem is much harder than the former one.

We point out that the model discussed in this paper, as oppose to previous studies in the literature, captures the effect of the inspection process on the line throughput and on the level of work in process. Clearly there is a tradeoff related to installation of QCSs. On the pros side, QCSs save resources otherwise spent on non-conforming products, allow to increase the line throughput by reducing the load on the bottleneck stations and reduce the work in process on the stations that follow them. On the cons side, QCSs incur their

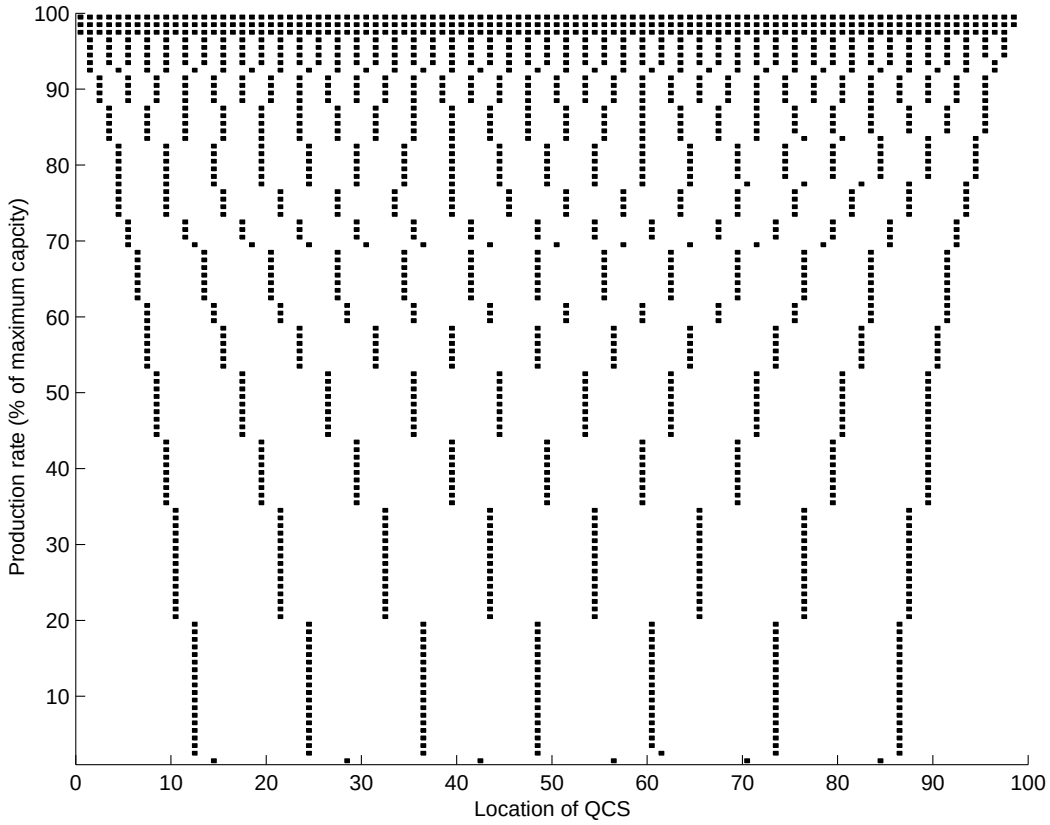

Figure 3: Optimal QCS configurations - decreasing production (inspection) rate with no penalty. Each row represents an optimal configuration at a given relative rate (production rate/maximum capacity).

own costs, and create new queues in the system that might increase WIP and flow time.

Throughout the analytical part of the paper we assumed that the arrival process of jobs into the system is Poisson. This assumption is not suitable for most real life production environments where the jobs are dispatched into the system by a decision of the system operator. We used numerical experiments to demonstrate that the Poisson arrival assumption leads to near optimal solutions also for deterministic arrival processes. This is true at least for small instances of the problem, for which we were able to estimate the optimal solutions by an enumerative method. However, we expect the method to be even more accurate for larger instances since the arrival process into machine i stochastically approaches Poisson process as $i \to \infty$. See the discussion on Reiman and Simon conjecture in Remark 4.6.

Note that although our aim in this study was to optimize the steady state performances of the systems, we believe that the method is also well suited for high multiplicity problems where a large but finite number of identical or similar products are to be produced, and

the goal is to minimize the total cost or maximize the total profit.

We propose further research to focus on the following directions: (1) To perform some sensitivity analysis to study the affect of different failure probabilities, arrival rates, processing times etc.. (2). To consider some dependence in the operations' failures probabilities. For example, to allow the failure probabilities to depend on the state of the machine. (3). The policy of 100% inspection in each installed QCS may be sub-optimal even under the assumption of independent failures on any machine. In particular, for slow QCSs, it might be better to inspect subsets of the jobs, so part of the benefit from inspecting is gained without creating a new bottleneck in the system. In general, the decision whether to check a job or not should be made on-line, based on the state along the line. (4) To allow Type I and Type II inspections errors in our model. (5) To model the case where the inspection operation itself may damage the product with some probability.

The production model presented in this paper can be further extended to capture a variety of manufacturing environments such as allowing repairs, reworks and machine breakdowns. In addition, other manufacturing environments such as job shop, assembly lines and multi-stage shop should be considered. Also, the ideas presented here can be adopted to some problems in other areas, such as determining optimal integrity check points in communication networks or during long service processes.

Acknowledgment: Partial support was received from the fund for the promotion of research at the Technion. We are grateful to the AE and the referees for their valuable comments that lead to an improvement of the paper.

References

[1] D. S. Bai and H. J. Yun. Optimally allocation of inspection effort in a serial multi-stage production system. *Computers and Industrial Engineering*, 30(3):387–396, 1996.

[2] M. Ben-Daya and A. Rahim. Optimal lot-sizing, quality improvement and inspection errors for multi-stage production system. *Computers and Industrial Engineering*, 41(1):65–79, 2003.

[3] S. R. Bowling, S. Kaewkuekool M. T. Khsawneh, and B. R. Cho. A markovian approach to determining optimum process target level for a multi-stage serial production system. *European Journal of Operational Research*, 159:636–650, 2004.

[4] A. K. Chakravarty and A. Shtub. Strategic allocation of inspection effort in a serial, multi-product production system. *IIE Transactions*, 19(1):13–22, 1987.

[5] T. J. Chen and A. C. Thornton. Quantitative selection of inspection plans. *Proceeding of the 1999 ASME Design Engineering Technical Conferences, Las Vegas, Nevada*, pages 1–11, 1999.

[6] E. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[7] S. E. Elmaghraby. Comments on a dp model for the optimal inspection strategy. *IIE Transactions*, 18(1):104–108, 1986.

[8] E. A. Elsayed. Invited paper: Perspective and challenges for research in quality and reliability engineering. *International Journal of Production Research*, 38(9):1953–1976, 2000.

[9] Z. Drezner H. Gurnani and R. Akella. Capacity planning under different inspection strategies. *European Journal of Operational Research*, 89:302–312, 1996.

[10] R. R. Inman, D. E. Blumenfeld, N. Huang, and J. Li. Designing production systems for quality: Research opportunities from an automotive industry perspective. *International Journal of Production Research*, 41(9):1953–1971, 2003.

[11] V. Kakade, J. F. Valenzuela, and J. S. Smith. An optimization model for selective inspection in serial manufacturing systems. *International Journal of Production Research*, 42(18):3891–3909, 2004.

[12] L. Kleinrock. *Queueing systems*. Wiley, New York, 1975.

[13] K. Kogan and T. Raz. Optimal allocation of inspection effort over a finite planning horizon. *IIE Transactions*, 34:515 527, 2002.

[14] G.F. Lindsay and A.B. Bishop. Allocation of screening inspection effort: a dynamic programming approach. *Management Science*, 10:342–352, 1965.

[15] K. Mehlhorn and S. Näher. *Leda - a platform for combinatorial optimization and geometric computing*. Cambridge University Press, 1999.

[16] T. S. Mountford and B. Prabhakar. On the weak convergence of departures from an infinite series of $\cdot/m/1$ queues. *Annals of Applied Probability*, 5:121–127, 1995.

[17] B. L. Nelson, J. Swann, D. Goldsman, and W. Song. Simple procedures for selecting the best simulated system when the number of alternatives is large. *Operations Research*, 49:950–963, 2002.

[18] M. Penn and T. Raviv. A polynomial time algorithm for solving a quality control station configuration problem. *Discrete Applied Mathematics, to appear*, 2006.

[19] A. Agnetis R. Rebello and P. B. Mirchandani. Specialized inspection problems in serial production systems. *European Journal of Operational Research*, 80, 1995.

[20] T. Raz. A survey of models for allocating inspections effort in multistage production system. *Journal of Quality Technology*, 18:239–247, 1986.

[21] Y. Shiau. Inspection allocation planning for a multiple quality characteristic advanced manufacturing system. *The International Journal of Advanced Manufacturing Technology*, 21, 2003.

[22] S. Suresh and W. Whitt. The heavy-trafic bottleneck phenomenon in open queueing networks. *Operations Research Letters*, 9:355–362, 1990.

[23] B.J. Yum and E.D. McDowell. Optimal inspection policies in a serial production system including scarp, rework and repair: an MILP approach. *International Journal of Production Research*, 25:1451–1464, 1987.

Appendix

Proof of Lemma 4.1. Recall that the function $C^*(a)$ is obtained as a minimization over all possible configurations of $C(Y,a)$. Thus, it is suffice to show that $C(Y,a)$ is convex, continuous and differentiable with respect to a. Let us write this function explicitly

$$\begin{aligned} C(Y,a) \equiv \;& a \cdot \textstyle\sum_{i=1}^{N} q_{0,L_i(Y)} \cdot [\mathcal{C}_i(Y,a) + \mathcal{C}'_i(Y,a)] + \sum_{i=1}^{N} Y_i f'_i \\ & + (1 - Y_N) \cdot a \cdot q_{0,L_Y(N)} \cdot (1 - q_{L_Y(N),N}) \cdot r_B. \end{aligned}$$

Thus, $C(Y,a)$ is a sum of linear functions and the functions

$$\mathcal{H}_i(Y,a) = \frac{h_i \cdot q_{0,L_i(Y)} \cdot a}{q_{0,L_i(Y)} \cdot a - \frac{1}{x_i}} \qquad \textit{and} \qquad \mathcal{H}'_i(Y,a) = \frac{h'_i \cdot q_{0,L_i(Y)} \cdot a}{q_{0,L_i(Y)} \cdot a - \frac{1}{x'_i}}.$$

Deriving $\mathcal{H}_i(Y,a)$ twice we obtain,

$$\frac{\partial^2 \mathcal{H}_i(Y,a)}{\partial a} = \frac{2\, q_{0,L_i(Y)}\, h_i}{\left(a\, q_{0,L_i(Y)} - \frac{1}{x_i}\right)^2} - \frac{2\, a\, q_{0,L_i(Y)}\, h_i}{\left(a\, q_{0,L_i(Y)} - \frac{1}{x_i}\right)^3} = -\frac{2\, q_{0,L_i(Y)}\, h_i\, x_i^2}{\left(a\, x_i\, q_{0,L_i(Y)} - 1\right)^3}.$$

Since we are interested in stable systems, it follows from Observation 2.2 that the relevant arrival rates are those for which $a \cdot q_{0,L_i(Y)} < \frac{1}{x_i}$ and $a \cdot q_{0,L_i(Y)} < \frac{1}{x'_i}$. Thus, it is easy to see that the second derivative is positive for all a in the relevant domain and the convexity of $\mathcal{H}$ is established. Now, since $C(Y,a)$ is obtained as sum of continuous differentiable and convex functions it follows that it is continuous, differentiable and convex. ■

Proof of Lemma 4.2. Let us write $C^*(a_1)$ explicitly in terms of the optimal configuration $Y^*(a_1)$ at a_1,

$$\begin{array}{rl} C^*(a_1) = & a_1 \cdot q_{0,L_N(Y^*(a_1))} \cdot (1 - q_{L_N(Y^*(a_1)),N}) \cdot r_B + \\ & a_1 \cdot \sum_{i=1}^N q_{0,L_i(Y^*(a_1))} \cdot [\mathcal{C}_i(Y^*(a_1), a_1) + \mathcal{C}_i'(Y^*(a_1), a_1)] + \\ & \sum_{i=1}^N f_i' Y_i^*(a_1). \end{array}$$

Let us denote $\mathcal{Y} \equiv Y^*(a_1)$. Note that if $C^*(a)$ is differentiable at a_1, then there is a neighborhood of a_1 for which $\mathcal{Y}$ remains an optimal configuration. Now,

$$\begin{array}{rl} \frac{\partial C^*(a)}{\partial a}(a_1) = & q_{0,L_N(\mathcal{Y})} \cdot (1 - q_{L_N(\mathcal{Y}),N}) \cdot r_B + \\ & \sum_{i=1}^N q_{0,L_i(\mathcal{Y})} \left\{ \mathcal{C}_i(\mathcal{Y}, a_1) + a_1 \frac{\partial \mathcal{C}_i(\mathcal{Y},a)}{\partial a}(a_1) \right\} + \\ & \sum_{i=1}^N q_{0,L_i(\mathcal{Y})} \left\{ \mathcal{C}'_i(\mathcal{Y}, a_1) + a_1 \frac{\partial \mathcal{C}'_i(\mathcal{Y},a)}{\partial a}(a_1) \right\} \\ \geq & \sum_{i=1}^N q_{0,L_i(\mathcal{Y})} \left\{ \mathcal{C}_i(\mathcal{Y}, a_1) + a_1 \frac{\partial \mathcal{C}_i(\mathcal{Y},a)}{\partial a}(a_1) \right\} \\ = & \sum_{i=1}^N \left\{ q_{0,L_i(\mathcal{Y})} \left(c_i + \frac{h_i}{\frac{1}{x_i} - a_1 \cdot q_{0,L_i(\mathcal{Y})}} \right) + \frac{a_1 \cdot h_i \cdot q_{0,L_i(\mathcal{Y})}^2}{\left(\frac{1}{x_i} - a_1 \cdot q_{0,L_i(\mathcal{Y})} \right)^2} \right\} \\ \geq & \sum_{i=1}^N \left\{ q_{0,i-1} \left(c_i + \frac{h_i}{\frac{1}{x_i} - a_1 \cdot q_{0,i-1}} \right) + \frac{a_1 \cdot h_i \cdot q_{0,i-1}^2}{\left(\frac{1}{x_i} - a_1 \cdot q_{0,i-1} \right)^2} \right\} \\ \geq & \sum_{i=1}^N \left\{ q_{0,i-1} \left(c_i + \frac{h_i}{\frac{1}{x_i} - a_0 \cdot q_{0,i-1}} \right) + \frac{a_0 \cdot h_i \cdot q_{0,i-1}^2}{\left(\frac{1}{x_i} - a_0 \cdot q_{0,i-1} \right)^2} \right\}. \end{array} \tag{13}$$

The first inequality is due to the facts that

$$q_{0,L_N(\mathcal{Y})} \cdot (1 - q_{L_N(\mathcal{Y}),N}) \cdot r_B \geq 0$$

and

$$\left\{ \mathcal{C}'_i(\mathcal{Y}, a_1) + a_1 \frac{\partial \mathcal{C}'_i(\mathcal{Y}, a)}{\partial a}(a_1) \right\} \geq 0$$

for all i. This is because $\mathcal{C}'_i(\mathcal{Y}, a)$ is a non-negative and increasing function of a. The second inequality in (13) is due to the fact that $q_{0,L_i(Y)} \geq q_{0,i-1}$, since $L_i(Y) \leq i - 1$ for any configuration Y. Now it is apparent that the expression

$$q \left(c_i + \frac{h_i}{\frac{1}{x_i} - a_1 \cdot q} \right) \tag{14}$$

is non-decreasing in q within the relevant domain. To see why the expression

$$\frac{a_1 \cdot h_i \cdot q_{0,L_{i-1}}^2}{\left(\frac{1}{x_i} - a_1 \cdot q_{0,L_{i-1}} \right)^2} \tag{15}$$

is also non-decreasing, we derive it with respect to q and obtain

$$\frac{2\,a_1^2\,h_i\,q^2}{\left(\frac{1}{x_i}-a_1\,q\right)^3}+\frac{2\,a_1\,h_i\,q}{\left(\frac{1}{x_i}-a_1\,q\right)^2}$$

which is also non-negative in the relevant domain of a and for all positive x and non-negative h and q. The last inequality of (13) follows from the fact that expressions (14) and (15) are also non-decreasing in a in the relevant domain and the fact that $a_0 < a_1$. ■