# An Efficient Algorithm for Maximizing the Expected Profit from a Serial Production Line with Inspection Stations and Rework

Tal Raviv

*Department of Industrial Engineering, Faculty of Engineering,*

Tel-Aviv University, Tel-Aviv 69978, Israel

E-mail: talraviv@eng.tau.ac.il

Tel +972-3-6406977

May 2012

In this paper, an unreliable serial production line in which nonconforming items are sent back for rework is studied. The line consists of existing machines and optional quality control stations (QCSs). The designer of such a production line needs to decide where to install the QCSs along the line and to determine the production rate, so as to maximize the expected operational profit rate obtained at a steady state. An efficient algorithm for solving this problem is presented; several extensions of the problem are discussed. An extensive simulation study proves the applicability of the model in realistic settings and is used to derive some insights about the nature of optimal solutions.

Keywords: Production, Quality Management, Inspection Allocation, Serial Production Lines

## 1. Introduction

In long production lines, such as those found at semiconductor fabrication facilities (FABs), inspection allocation problems naturally arise and the planner must determine where along the line to locate quality controls stations (QCSs). The location of the QCSs affects both the expected production cost per item and the production rate of the line. Since both measures are of interest to the designer of the quality control system, we advocate using the expected profit per time unit measure that represents both aforementioned measures in terms of monetary units. We refer to this measure as the *profit rate*.

The literature on allocating inspection efforts in multistage production systems dates back to Bishop and Lindsay (1964) who introduced an inspection allocation model to minimize the total production cost per unit. Yum and McDowell (1987) introduced an inspection model that allows rework, off-line repair, and scrapping.  In their model, inspection errors can occur, and variable inspection, production and repair costs are given. They formulated the problem of optimally allocating inspection efforts to minimize total costs per item as a mixed integer linear program. Tayi and Ballou (1988) and Raghavachari and Tayi (1991) developed an integrated framework for manufacturing, inspecting and reprocessing activities in serial production systems operating under a lot-by-lot production mode. Subsequent studies presented heuristic methods to solve similar inspection effort-allocation problems. See, for example, Taneja (1994); Shiau (2002); and Emmons and Rabinowitz (2002). For literature surveys, see Raz (1986) and Mandroli et al (2006).

The inspection effort allocation literature, mentioned above, focuses on production systems that are in *statistical control* where the goal is to detect random defects of the processed items rather than monitoring the state of the production process. From a mathematical point of view the two approaches are very different. The former assumes that the success probabilities of consecutive operations are independent, an assumption which is valid only while the system is under statistical control. The latter approach tries to detect when a machine goes out of control by exploiting the dependence between success probabilities. This dependence is caused by changes in the state of the machine. Clearly, the inspection effort allocated for full inspection at various points along the line is aimed at complementing the *statistical process control* rather than replacing it.

Penn and Raviv (2007) considered inspection effort allocation in serial production lines with scrapping only. In their model, the production rate is a decision variable and the objective is to maximize the profit rate rather than minimize the production cost per item, studied earlier. Higher production rates typically result in higher marginal production and inspection costs per item. Due to this trade-off, the optimal rate is not necessarily the highest possible one. Penn and Raviv (2008) extended this model to incorporate the holding costs of work-in-process and explored the interaction between production rate, configuration of the inspection system, and inventory.

In this study, we consider a serial production line with $N$ machines and an infinite number of identical items to be produced. Processing an item consists of a series of $N$ operations, where each operation is carried out by a distinct machine. The machines are denoted by $M_1, \ldots, M_N$. The cost of performing the $i^{th}$ operation is denoted by $c_i$ and its processing time is assumed to be an *i.i.d* random variable with mean $x_i$. We make no additional assumptions about the distribution of this random variable. Operation $i$, if performed on a conforming item, succeeds with known probability $p_i$ and fails with probability $1 - p_i$. If one or more of the operations of an item fails, it is regarded as nonconforming. Let $p_{u,v} = \prod_{i=u+1}^{v} p_i$ denote the conditional probability that an item is a conforming one upon departure from $M_v$ given that it was conforming upon its previous departure from $M_u$. Note that the assumption of independence among the failure events is unnecessary because $p_i$ is defined as the conditional probability of failure in operation $i$, given that the item was a conforming one before the operation began.

Quality control stations (QCSs) can be installed between any pair of machines and after the last machine in the line. A sequence of machines followed by a QCS is referred to as a *segment*. If a QCS is not installed after the last machine, the sequence of the machines after the last QCS in the line is also considered a segment. An installed QCS detects all the nonconforming items generated by the machines belonging to the segment that contains it. Nonconforming items are sent upstream to the beginning of the segment for rework.

Machines and QCSs are jointly referred to as “stations”. Each item can be processed by a single station at a time and each station can process one item at a time. An unlimited buffer is located in front of each station where all items that have finished their previous operations are waiting to be processed or inspected. A raw material buffer in front of the first machine represents the items that have entered the production line but have yet to begin their first operation. We assume a general stationary arrival process to this buffer, i.e., the inter-arrival times of items at this buffer are *i.i.d*. We denote the expected arrival rate by $a$. If the steady-state departure rate from the system is also $a$, the system is said to be stable and $a$ is said to be the *production rate* of the system.

Note that the assumption of independence between arrivals of items at the first buffer is a simplifying one. Indeed, in actual production systems, the introduction of new items to the system is typically based on the system state and controlled by the operator. Moreover, the assumptions that the buffer space is unlimited and the exclusion of the holding cost of WIP from the profit calculation are useful simplifying assumptions. They allow solving the QCS configuration problem to optimality and the obtained solutions are in many cases nearly optimal even when the amount of WIP is strictly limited by a CONWIP production regime. The above claims are supported by an extensive numerical study presented in Section 6.

A QCS that immediately follows $M_i$, if it is installed, is denoted by $QC_i$. A QCS configuration is denoted by a set $Y \subseteq \{QC_1, QC_2, \dots, QC_n\}$ of its installed QCSs. Figure 1 presents an example of the system introduced above. Machines are depicted as squares and inspection stations as pentagons. The primary stream of items is depicted as a solid arrow, and the reverse stream of nonconforming items that are sent back for rework, as dashed arrows. In this example $Y = \{QC_2, QC_4\}$.

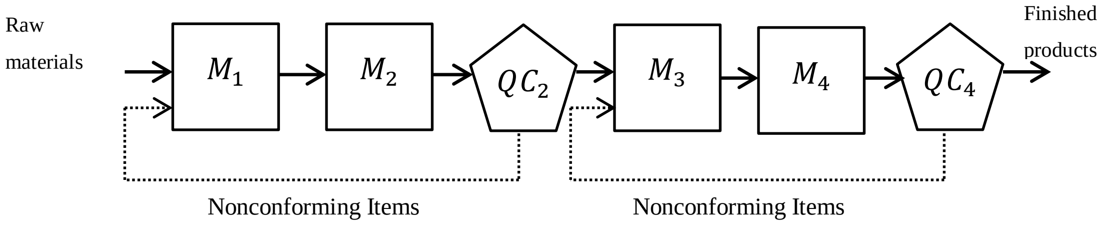

**Figure 1**: An example of a serial production line with two QCSs

For a given QCS configuration *Y*, the location of the last installed QCS before $M_i$ is denoted by $L_i(Y)$, with the conventions that $L_i(Y) = 0$ if no such QCS is installed and that $L_{N+1}(Y)$ is the location of the last QCS in the line. To avoid cumbersome notation, we use $L_i$ rather than $L_i(Y)$, whenever the QCS configuration *Y* is clear from the context. For the example presented in Figure 1, $L_1 = L_2 = 0, L_3 = L_4 = 2$ and $L_5 = 4$.

Since the inspection cost and time are likely to be affected by the number of production steps in the inspected segment we introduce the following notation: The inspection cost per item at $QC_i$ for which $L_i$=$j$ is denoted by $c_i'(j)$. The inspection time is an *i.i.d* random variable with mean $x_i'(j)$. In addition, there is a fixed capital cost per time unit of $f_i'(j)$ associated with each installed QCS regardless of its actual work. Note that the inspection time and costs are likely to be affected by the number and identity of production operations in the segment. The larger this number, the more effort and equipment may be needed to carry out the inspection.

Since capital costs of machines are sunk costs, they are eliminated from the optimization process. Each finished product is sold as a commodity in the market for $r_G$ monetary units. A penalty of $r_B$ monetary units is incurred by each nonconforming product that the system delivers. Note that if $QC_N$ is installed, the system delivers only conforming items.

In our basic setting, nonconforming items detected by $QC_\mathrm{i}$, are sent back for rework on machines $M_{L_i+1}, \ldots, M_i$. The processing times, processing costs, inspection cost, and the success probabilities of reworked items are identical to those of first-time items. An item may be reworked more than once. We introduce and analyze a more general setting in Section 4.

In this study, we present a method for selecting a production rate *a* and a QCS configuration *Y*, so as to maximize *operational profits.* The term "operational profit" refers to the total revenue net of variable production, inspection, and capital costs as well as the penalty cost incurred by nonconforming products delivered by the system. The *net profit* is affected also by the capital cost of machines and by the holding cost of work-in-process (WIP). In Section 6, we demonstrate that the system can typically work at near optimal operational profit with modest levels of WIP.

An intermediate step toward our goal of maximizing operational profit is to solve the problem of minimizing the total costs per item assuming a given production rate. We refer to this problem as the *cost minimization problem* (CMP). The CMP may be of some practical interest intrinsically but it is studied here mainly to serve as a subroutine of the optimization algorithm of the *profit maximization problem* (PMP). The contribution of this study is twofold. First, it introduces a methodology that can efficiently solve a variety of profit maximization problems in the inspection effort allocation domain. This methodology generalizes the method used in Penn and Raviv (2007). Second, it introduces an exact and efficient algorithm to solve an inspection effort allocation problem characterized by rework and capacity constraints. We note that reworking nonconforming items on-line (possibly after performing some offline rehabilitation) is a common practice in some industries.

In Section 2, the cost maximization problem (CMP) is cast as a problem of finding a shortest path on a directed graph. In Section 3, it is shown that for the PMP there is a small finite set of potentially optimal production rates. Once this set is identified, it is possible to solve a CMP for each member of this set and to choose the most profitable one as an optimal solution for the PMP. In Section 4, a richer rework model is introduced and an adaptation of the previously presented algorithms is proposed. In Section 5, we present the results of our numerical study in which we demonstrate the capabilities of the proposed algorithm in dealing with very large instances of the PMP. In Section 6, we study the relationship between the WIP level and the profit rate under a constant WIP (CONWIP) regime using discrete event simulation. We show that the modeling decision to allow all stations to work at their maximal capacity while ignoring the holding costs of WIP does not prevent us from reaching applicable near optimal solutions. In Section 7, we visualized the optimal QCS configuration obtained by our algorithm. We use this visualization to draw insights on the effect of various parameters of the systems on the optimal QCS configuration and the optimal profit under optimal inspection effort allocation. Concluding remarks and directions for future research are presented in Section 8. All the notations in the paper are presented before they are first used but for the convenience of the reader, we also provide a notation summary in the Appendix.

## 2. The Cost Minimization Problem

In this section, we cast the CMP as the shortest path problem. White (1969) introduced this idea for a model with off-line repair and uncapacitated stations, but to the best of our knowledge, it was never adopted for models with rework and capacity constraints.

Consider a system working with arrival rate $a$ under a given QCS configuration. The system is stable only if the arrival rate of first-time items at each of the system segments equals the arrival rate at the first segment. The probability of an item arriving at

$QC_i$ being a conforming one is $p_{L_i,i}$. If the system is stable, then the joint arrival rate of both first-time and reworked items at each station in a segment ending at $QC_i$ is the arrival rate, $a$, of first-time items at the segment multiplied by the mean number of times each item is processed by the stations of the segment, $\sum_{m=0}^{\infty}\left(1-p_{L_i,i}\right)^m = \frac{1}{p_{L_i,i}}$. This is the arrival rate at each station in the segment $M_{L_i+1}, \dots., M_i, QC_i$ is $\frac{a}{p_{L_i,i}}$.

Necessary and sufficient conditions for the stability of the system are that the potential throughput of each station is greater than the arrival rate that it faces. That is

$$\frac{a}{p_{L_i,i}} \leq \frac{1}{x_j} \qquad \forall i \in Y, j = L_i + 1, \dots, i \tag{1}$$

and for QCSs,

$$\frac{a}{p_{L_i,i}} \leq \frac{1}{x_i'(L_i)} \qquad \forall i \in Y. \tag{2}$$

If one or more of the inequalities (1) or (2) holds as equality, then some stations in the system are fully loaded. In this case, the system cannot reach steady-state in terms of the number of items in its buffers, but the production rate converges over time to $a$.

An upper bound on the potential departure rate from a segment, consisting of the stations $M_{u+1}, \dots., M_v, QC_v$, is

$$T(u,v) = \frac{p_{u,v}}{\max(x_{u+1}, \dots, x_v, x_v'(u))}. \tag{3}$$

Note that since no non-conforming items are removed from the system, this is also an upper bound on the arrival rate of *first-time* items at the segment, assuming the system is stable. Now, since the value $T(u,v)$ is determined only by $u$ and $v$, and is not affected by the locations of QCSs outside the segment, it is an upper bound on the production rate in any configuration with $QC_u$, $QC_v$ and no QCS in between. For convenience, let us define

$$T(u, N+1) = \frac{1}{\max(x_{u+1}, \dots, x_N)} \tag{4}$$

for all $u \leq N$, to represent the potential throughput of the last segment if it is not ended by a QCS. Finally, let us define $T(N, N+1) = \infty$ to allow unrestricted flow of items out of $QC_n$, if installed.

Next, let us construct a directed graph $\mathcal{D} = (\mathcal{N}, \mathcal{A})$ with a node set $\mathcal{N} = \{0,1, \dots, N+1\}$, with a node for each machine plus two nodes representing the beginning and the end of the line. The arc set is defined as $\mathcal{A} = \{(u, v): u, v \in \mathcal{N}, u < v, T(u, v) \geq a\}$; with an arc for any segment that corresponds to a feasible segment under arrival rate $a$.

Now let us set the length of each arc $(u, v) \in \mathcal{A}$ as the total cost per item in the segment $M_{u+1}, \dots, M_v, QC_v$,

$$C(u, v) = \frac{f_v'(u)}{a} + \frac{c_v'(u) + \sum_{k=u+1}^{v} c_k}{p_{u,v}}. \tag{5}$$

The first terms of the right-hand side of (5) is the fixed cost divided by the production rate, which represents the contribution of the capital cost to the total production cost of each item. The second term is the variable production and inspection cost per first-time item incurred by all the stations in the segment. In addition, for any arc $(u, N+1)$ let us define

$$C(u, N+1) = \left(1 - p_{u,N}\right) r_B + \sum_{k=u+1}^{N} c_k. \tag{6}$$

Recall that these arcs represent segments at the end of the line that are not ended by a QCS. The first term of (6) is the mean penalty cost for delivered nonconforming items and the second term is the total production cost per item. Note that each item passes exactly once via such segments. Finally let $C(N, N+1) = 0$ represent the fact that no penalty cost is incurred if the last QCS is located immediately after the last machine.

Note that since any arc in the graph represents a feasible segment, any $0 \to (N+1)$ directed path in $\mathcal{D}$, $(0, v_1), (v_1, v_2), \dots., (v_{k-1}, v_k), (v_k, N+1)$ can be associated with a feasible configuration with QCSs installed after stations $v_1, \dots, v_k$. Moreover, since the length of each arc represents the expected cost per item incurred by its corresponding segment, a shortest path represents a configuration that minimizes the total production, inspection, and penalty costs per item subject to the capacity constraints.

The time complexity of a straightforward procedure to construct the graph $\mathcal{D}$ is $O(N^3)$: The number of arcs is $O(N^2)$ and for each arc $(u, v)$, $T(u, v)$ and $C(u, v)$ are calculated using (3)-(6) in $O(N)$. Once the graph is constructed, the shortest path can be found using Dijkstra's algorithm in $O(N^2)$, see Cormen et al (2001). Hence, the complexity is dominated by the graph construction step. Algorithm 1 (see below) is a more cautious procedure for the CMP with overall complexity of $O(N^2)$.

The time complexity of Algorithm 1 is $O(N^2)$ since all the operations inside the internal loop require a constant amount of time. The complexity improvement is obtained by saving on the summation and product operations in (3)-(6) using the recursion relations between $C(u, v)$ and $C(u, v + 1)$ and between $T(u, v)$ and $T(u, v + 1)$.

Clearly, it is impossible to solve the CMP in less than $O(N^2)$ time since this is the order of the input size of the problem. Consequentially, the complexity of the CMP is $O(N^2)$.

```
**Algorithm 1** – Constructing the graph [in $O(N^2)$]
Input:
    Number of machines: $N$
    Machines parameters: $x_v, c_v, p_v$ for all $v = 1, \dots, N$
    QCS parameters: $x'_v(u), c'_v(u), f'_v(u)$ for all $v = 1, \dots, N$ and $u = 0, \dots, v-1$
    Production rate $a$
  Penalty cost $r_B$
Set $\mathcal{N} = \{0,1, \dots, N+1\}$
Set $\mathcal{A} = \emptyset$
for $u = 0$ to $N$
    set $p = 1$
    set $x^* = -\infty$
    set $C = 0$
    for $v = u + 1$ to $N + 1$
        set $p = p \cdot p_v$
        if $x^* < x_v$ then $x^* = x_v$
        if $v < N + 1$ then
            set $T = p/\max\left(x^*, x'_v(u)\right)$
            set $c = c + c_v$
        else
            set $T = 1/x^*$
            set $= c + (1-p)r_B$

        if $T > a$ then
            set $\mathcal{A} = \mathcal{A} \cup \{(\mathrm{u}, \mathrm{v})\}$
            if $\mathrm{v} < \mathrm{N} + 1$ then
                set $C(u, v) = \frac{f'_v(u)}{a} + \frac{c'_v(u) + c}{p}$
            else
                $C(u, v) = c$
```

We conclude this section with a discussion of some extensions to the CMP and describe simple methods for resolving them. In some production lines, it is impossible or undesirable to carry out an inspection of an operation after other operations have already been conducted. For example, inspection of a printed circuit board may be impractical after the board has been installed as a subsystem within a laptop. Expressing such constraints in our model is straightforward. If operation $u$ must be inspected before operation $v$ is carried out, all the arcs $(u', v')$ where $u' < u$ and $v' > v$ are removed from the graph.

Sometimes the rehabilitation process of a nonconforming item consists of an off-line operation that should be carried out prior to sending the item back for rework. We denote the mean cost of the off-line rehabilitation operation for segment $(u, v)$ by $\gamma_v(u)$ and the corresponding processing time of this operation by $\chi_v(u)$. The expected number of times each item passes via this rehabilitation operation is a geometric variable (with a support $\{0,1,2,\dots\}$) and success probability of $p_{u,v}$. For example, integrating this additional rehabilitation cost into our model can be done by adding $\gamma_v(u)\frac{1-p_{u,v}}{p_{u,v}}$ to the length (cost) of the arc$(u, v)$. Similarly, the effect of the processing time for the rehabilitation operation can be introduced into the model by setting the arc throughput to

.

$$T(u,v) = \frac{p_{u,v}}{\max\left(x_{u+1}, \dots, x_v, x'_v(u), \left(1 - p_{u,v}\right)\chi_v(u)\right)}$$

The offline operation costs may include the cost of material handling devices that are used to return the item to the beginning of the segment.

## 3. Algorithm for the Profit Maximization Problem (PMP)

In this section, we consider the problem of maximizing the total revenue rate net of production, inspection, and penalty costs. This problem (PMP) is more complicated than the CMP because, in this case, the production rate and the QCS configuration must be determined simultaneously. Note that there may be a trade-off between the production rate and the total production and inspection cost per item. That is, installing additional QCSs may increase the production rate, but result in higher total cost per item. The algorithm presented below resolves this trade-off.

We now explore some properties of an optimal solution of the PMP. We denote a solution to the PMP by $(a, Y)$ where $a$ is the production rate and $Y$ is the QCS configuration. $(a, Y)$ is feasible if the system with QCS configuration $Y$ is stable under production rate $a$.

**Lemma 1:** If $(a, Y)$ is an optimal solution of a PMP and $a > 0$, then at least one of the stations is operating at full capacity. In other words, there is a $QC_i \in Y$ and $j \in L_i + 1, \dots, j$ such that either

$$\frac{a}{p_{L_i,i}} = \frac{1}{x_j} \tag{7}$$

or

$$\frac{a}{p_{L_i,i}} = \frac{1}{x'_j(L_i)} \tag{8}$$

***Proof:*** The expected profit per item in a given QCS configuration can be expressed as

$$r_G - \mathbb{I}_{\{QC_N \notin Y\}}\left[(1 - p_{L_{N+1},N})r_B + \sum_{k=L_N+1}^{n} c_k\right] - \sum_{v \in Y}\left(\frac{f_v'(L_v)}{a} + \frac{c_v'(L_v) + \sum_{k=L_v+1}^{v} c_k}{p_{L_v,v}}\right).$$

where $\mathbb{I}_{\{QC_N \notin Y\}}$ is an indicator function that equals zero if a QCS is installed after the last machine in the line and zero otherwise. The first term is the revenue while the second is the expected penalty for nonconforming items and the production cost of the segment that is not ended by a QCS, if one exists. The last term is the expected production and inspection costs in all other segments.

Note that the objective is to maximize the profit per time unit. Therefore, if in an optimal solution the production rate is positive ( $a > 0$), then the above sum, which represents the profit per item, must be nonnegative. In such a case, for a fixed configuration, increasing the production rate may only increase the profit rate. Therefore, in an optimal solution, the production rate is the maximal possible rate for the selected configuration. **Q.E.D**

**Corollary 1**: In an optimal solution $(a, Y)$ of the PMP, if $a > 0$, then $a = T(u, v)$ for some $0 \le u < v \le N + 1$.

***Proof:*** The stations that satisfy (7) and/or (8) are those that maximize the production (or inspection) time within the segment. By definition, see (3), these stations determine the value of $T(u, v)$ of the segment. **Q.E.D**

A direct consequence of Corollary 1 is the following corollary:

**Corollary 2:** The value of $a$ in an optimal solution of the PMP is a member of the following finite set $S = \{a: T(u, v) = a, 0 \leq u < v \leq N+1\}$.

By the construction of the set $S$, it is clear that $|S| \leq \binom{N+2}{2}$. Our algorithm for the PMP relies on this observation. Let us denote the objective value (resp., configuration) of an optimal solution of the CMP for arrival rate $a$ by $C^*(a)$ [resp., $Y^*(a)$], then the value of an optimal solution to the PMP is obtained by

$$a^* = \underset{a \in S}{\operatorname{argmax}}\ a[r_G - C^*(a)] \tag{9}$$

The optimization problem (9) can be solved by solving the CMP for each of the members of $S$. The optimal solution of the PMP is $(a^*, Y^*(a^*))$, assuming $r_G - C^*(a) > 0$. Otherwise the optimal solution is $a^* = 0$ and $Y^* = \emptyset$.

Since the computation time of the CMP is $O(N^2)$, the overall time complexity of this algorithm is $O(N^4)$. The space complexity of the procedure is $O(N^2)$, since it is possible to reuse the memory allocated for the representation of the graph.

Next, we present two algorithmic improvements that significantly reduce the computation time of the algorithm in our numerical experiment although they have no implications on the theoretical complexity bound. Both improvement methods are based on identifying members of $S$ that are not candidates for $a^*$ and save the effort to solve their CMP sub-problem.

First, consider $UB = \min_i \left(\frac{1}{x_i}\right)$ as an upper bound on the production rate in any stable solution. Such a rate could be achieved in a configuration without any QCSs at all. We can use this to reduce the set $S$ and remove all the members of $S$ that are greater than $UB$.

Second, observe that $a[r_G - C^*(a)] \leq a[r_G - \sum_{i=1}^{n} c_i]$ because $C^*(a) \geq a \sum_{i=1}^{n} c_i$. The last inequality is true because the total cost per item must include the cost of processing it on each machine along the line at least once (in addition to possible rework and inspection costs). Based on this observation, the algorithm can skip any member $a$ for which $a[r_G - \sum_{i=1}^{n} c_i]$ is smaller than the value of the best solution that was already found. This may be particularly helpful if $S$ is scanned in decreasing order. We note that although the sorting of S requires additional computational effort of $O(N^2 \log^2 N)$, it does not contribute to the theoretical computational complexity of $O(N^4)$ since the sorting is performed only once. The sorting procedure is also instrumental in eliminating non-unique members of $S$ that are clearly unnecessary.

Finally, we note that the number of *unique* members in the set $S$ is affected by the numerical accuracy used to represent its members. If the accuracy is high, it is unlikely to have $T(u_1, v_1) = T(u_2, v_2)$ for different pairs $(u_1, v_1)$ and $(u_2, v_2)$. However, as the numerical accuracy is decreased, the number of pairs that seem identical increases. Therefore, by rounding all the values in $S$ at a certain decimal digit, the size of the set $S$ can be greatly reduced, thus proportionally shortening the computation time. The resulting solution is an approximated one. The absolute approximation error is bounded from above by $10^{-d+1}[r_G - C^*(a)] \leq 10^{-d+1} r_G$ where $d$ represents the numerical accuracy of the calculation, that is, the smallest difference between two numbers that are represented differently by the computer. Since the optimal solution is bounded from above by $a^* r_G$, the relative approximation error of this procedure is bounded from above by $10^{-d+1}$. Clearly, it is possible to increase $d$ artificially in order to save computation time at the expense of accuracy.

## 4. Extensions of the Rework Model

In the previous sections, we considered a production line where all nonconforming items are sent back to the beginning of the segment and reworked by all the stations of the segment in exactly the same way as they were processed in the first place. In this section, we extend the model in the following manner:

1. Allowing rework operations to consume different amounts of resources than first-time operations.
2. Sending items for rework beginning with the operation that created the nonconformity rather than the first operation of the segment.

We denote the ratio between the rework processing time (resp., cost) and the first-time processing time (resp., cost) of an operation on $M_i$ by $\rho_i$ (resp., $\zeta_i$). Note that $\rho_i$ (resp., $\zeta_i$) can be any non-negative number. In particular, $\rho_i > 1$ (resp., $\zeta_i > 1$) indicates that the time (resp., cost) required for a rework operation is greater than the time (resp., cost) required for first-time operations.

The number of times that an item passes through each machine $M_w$ is a geometric random variable with success probability $p_{L_w,w}$, i.e., the average amount of work brought to $M_w$ by a first-time item is $\left(1 + \rho_w \frac{1-P_{L_w,w}}{P_{L_w,w}}\right)$. Therefore, the maximal arrival rate for which $M_w$ is stable is

$$\frac{p_{L_w,w}}{x_w\left(p_{L_w,w}(1-\rho_w) + \rho_w\right)}$$

and $QC_w$, if installed, is stable for an arrival rate not greater than $\frac{p_{u,v}}{x'_v}$. That is, the potential throughput of a segment that consist of the stations $M_{u+1}, \dots, M_v, QC_v$ is given by

$$T(u,v) = min\left(\frac{p_{u,u+1}}{x_{u+1}\left(p_{u,u+1}(1-\rho_{u+1}) + \rho_{u+1}\right)}, \dots, \frac{p_{u,v}}{x_v\left(p_{u,v}(1-\rho_v) + \rho_v\right)}, \frac{p_{u,v}}{x'_v}\right),$$

for all $1 \le u < v < N+1$. The value of $T(u, N+1)$ is calculated exactly as in (4). Similarly, the total inspection and production cost per item incurred by all the stations of the corresponding segment is

$$C(u,v) = \frac{f'_v(u)}{a} + c'_v(u)\left(1 + \frac{1-p_{u,v}}{p_{u,v}}\right) + \sum_{w=u+1}^{v} c_w\left(1 + \zeta_w \frac{1-p_{L_w,w}}{p_{L_w,w}}\right),$$

for all $1 \le u < v < N+1$. The value of $C(u, N+1)$ is calculated exactly as in (6).

Now it is possible to apply the algorithms presented in Sections 2 and 3, using the new values of $C(u,v)$ and $T(u,v)$, to solve the extended version of the CMP and PMP. The complexity of the algorithms is unaffected.

## 5. Benchmarking the Algorithm

In this section, we benchmark the PMP algorithm, described in Section 3, on a set of 44 problem instances, each with 1000 machines. Recall that the algorithm for the PMP calls the CMP algorithm numerous times. We report on the actual number of calls to this algorithm and thus the need to benchmark the CMP algorithm separately is saved.

The algorithm was implemented in Mathwork Matlab v7.11.0 without compilation and ran on an Intel Xenon X3450, 2.66Ghz workstation. The algorithmic improvements presented at the end of Section 3 are included in this implementation and we kept track of the number of calls to the CMP that were eliminated in order to estimate the benefit of these improvements.

As a benchmark, a set of 44 problem instances was constructed based on 22 settings for processing and inspecting times and costs (see Table 1) and two revenue and penalty settings. The first two settings consist of instances with identical characteristics for all the machines and all the QCSs. In these instances, the mean inspection time and cost are proportional to the length of the inspected segment. In the first instance, the fixed capital cost of the QCS is also proportional to the length of the inspected segment while in the second it is fixed for all QCSs regardless of the length of the inspected segment. For instances 3-22, the mean processing time and mean processing cost were selected randomly and independently for each machine from the interval (2,9) and the success probability of each operation was selected from the interval (0.998,1). In instances 3-12, the mean inspection time and costs are proportional to the total inspection time and costs in the corresponding segment. In instances 13-22, the values of these parameters were selected randomly and independently of the corresponding processing operations. Each of the 22 settings was tested with two sets of values for $r_G$ and $r_B$, namely, $r_G = 7000, r_B = 8000$ and $r_G = 20000, r_B = 40000$. The complete dataset and the Matlab code are available from the author upon request.

| ID | Processing times / item | Inspection Time / item | Processing Cost / item | Inspection Cost / item | Success Probability | Capital Cost / time unit |
|---|---|---|---|---|---|---|
| 1 | $x_v = 5$ | $x'_v(u)$<br>$= v - u - 1$ | $c_v = 5$ | $c'_v(u) = v - u$ | $p_v = 0.999$ | $f'_v(u) =$<br>$= 0.1(v - u)$ |
| 2 | $x_v = 5$ | $x'_v(u) = 3$ | $c_v = 5$ | $c'_v(u) = 3$ | $p_v = 0.999$ | $f'_v(u) = 0.3$ |
| 3-12 | $x_v = U(2{,}9)$ | $x'_v(u)$<br>$= 0.2 \sum_{i=u+1}^{v} x_v$ | $c_v = U(2{,}9)$ | $c'_v(u)$<br>$= 0.2 \sum_{i=u+1}^{v} c_v$ | $p_v$<br>$= U(0.998{,}1)$ | $f'_v(u)$<br>$= 0.02 \sum_{i=u+1}^{v} c_v$ |
| 13-22 | | $x'_v(u)$<br>$= U(0.4{,}1.8)$ | | $c'_v(u)$<br>$= U(0.4{,}1.8)$ | | $c'_v(u)$<br>$= U(0.04{,}0.18)$ |

Table 1: Description of the Benchmark Problem Instance

Table 2 presents the performance results of the algorithm. In the first column the IDs of the time/cost settings are presented. In the rest of the table, the first four columns correspond to the $r_G = 7000, r_B = 8000$ setting and the last four to the $r_G = 20000, r_B = 40000$ setting. The first column in each such group presents the objective value of the optimal solution (expected profit per time unit) while the next column presents the computation time in seconds. In the third column, the number of calls to the CMP sub-problems is presented. In the last column of each group, the percentage of these sub-problems out of the total number of members in the set of potentially optimal production rates $S$ is presented.

It is apparent from the table that the algorithm is capable of solving all the 1000 machine instances in our benchmark set in a reasonable period of time. Even the most difficult instance was solved in slightly more than 31 minutes (instance 18 with $\mathrm{r_G} = 7000, \mathrm{r_B} = 8000$). We also observe that the instances with smaller revenue per conforming item, $r_G = 7000$, are much harder to solve compared to those with the higher $r_G = 20000$. This is because the lower bound on the production rate, discussed at the end of Section 3, grows proportionally with $r_G$. Higher values of this lower bound make it possible to eliminate additional potential production rates from consideration. Indeed, the average computation time of the $r_G = 7000$ instances is about a tenth of the average computation time of those with $r_G = 20000$. It might be the case that instances with smaller $r_G$ could take even a longer period of time to solve because larger number of calls to the CMP would have to be invoked. However, by extrapolating from the share of CMP solved in the hardest instances we can bound the computation time of the 1000 machines PMP instances from above (with the same implementation on the same machine) by about five hours.

Recall that $|S|$ is bounded from above by $\binom{1000+2}{2} = 501{,}501$. Indeed, for instances 3-22, with randomly generated success probabilities and processing times, the actual number of distinct members in $|S|$ was always very close to this upper bound, i.e., at least 99.8% of this number. In fact, if the numerical accuracy of our calculations was unlimited, the number of potential rates should have been, almost surely, exactly 501,501. However, despite the large number of potentially optimal production rates, the CMP is solved for small fractions of these rates. The rest of the candidate rates are eliminated using the bounds constructed for a feasible and optimal production rate.

For instances 1 and 2, where the success probabilities and processing times of all machines is identical, the number of distinct, potentially optimal production rates is 1001 (the number of machines plus one). This is because the production rate dictated by each pair of machines is only determined by the number of machines between the two. In addition, there is one potential rate determined by a configuration with no QCSs at all.

| ID | $r_G = 7000, r_B = 8000$ | | | | $r_G = 20000, r_B = 40000$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Value | Time (sec) | CMP Calls | CMP Share | Value | Time (sec) | CMP Calls | CMP Share |
| 1 | 197.2 | 0.6 | 11 | 1.1% | 2792.0 | 0.4 | 6 | 0.6% |
| 2 | 342.3 | 5.9 | 156 | 15.6% | 2902.1 | 1.4 | 34 | 3.4% |
| 3 | 97.3 | 428.0 | 11927 | 2.4% | 1541.0 | 98.0 | 2730 | 0.5% |
| 4 | 104.4 | 386.5 | 10789 | 2.2% | 1548.2 | 78.1 | 2177 | 0.4% |
| 5 | 98.1 | 402.6 | 11250 | 2.2% | 1541.6 | 70.2 | 1954 | 0.4% |
| 6 | 99.3 | 414.1 | 11543 | 2.3% | 1542.2 | 85.6 | 2385 | 0.5% |
| 7 | 96.4 | 418.4 | 11663 | 2.3% | 1541.6 | 75.3 | 2094 | 0.4% |
| 8 | 111.4 | 356.3 | 9933 | 2.0% | 1554.3 | 74.5 | 2069 | 0.4% |
| 9 | 106.8 | 399.5 | 11138 | 2.2% | 1549.2 | 94.9 | 2640 | 0.5% |
| 10 | 87.7 | 449.0 | 12513 | 2.5% | 1529.6 | 74.4 | 2071 | 0.4% |
| 11 | 102.6 | 371.6 | 10344 | 2.1% | 1548.6 | 61.0 | 1690 | 0.3% |
| 12 | 92.3 | 428.9 | 11956 | 2.4% | 1535.7 | 74.1 | 2063 | 0.4% |
| 13 | 150.6 | 1987.6 | 56992 | 11.4% | 1593.1 | 169.1 | 4828 | 1.0% |
| 14 | 155.2 | 1971.3 | 56576 | 11.3% | 1603.5 | 161.6 | 4608 | 0.9% |
| 15 | 154.5 | 2036.4 | 58420 | 11.7% | 1599.0 | 166.2 | 4737 | 0.9% |
| 16 | 158.4 | 1989.2 | 57023 | 11.4% | 1601.5 | 173.2 | 4940 | 1.0% |
| 17 | 146.8 | 2209.9 | 63321 | 12.7% | 1593.5 | 169.8 | 4830 | 1.0% |
| 18 | 141.8 | 2342.9 | 67318 | 13.4% | 1585.8 | 171.7 | 4884 | 1.0% |
| 19 | 161.2 | 2001.5 | 57364 | 11.5% | 1605.6 | 165.9 | 4729 | 0.9% |
| 20 | 142.0 | 2198.5 | 62995 | 12.6% | 1585.6 | 163.0 | 4645 | 0.9% |
| 21 | 161.0 | 1990.2 | 57007 | 11.4% | 1605.7 | 185.7 | 5297 | 1.1% |
| 22 | 171.7 | 1799.7 | 51466 | 10.3% | 1615.8 | 162.2 | 4611 | 0.9% |
| **Average** | | **1117.7** | | **7%** | | **112.6** | | **0.7%** |

Table 2: Performances of the PMP algorithm

The computation times of the 1000 machines problem instances in our benchmark set seem satisfactory for most applications. However, in order to test the approximation procedure discussed in Section 3, we decided to set the relative error to $10^{-5}$ (0.001%) by rounding the values of the member of $S$ at the sixth digit.

Table 3 below presents the relative and absolute savings in terms of computation times as well as the relative optimality gap for each of our benchmark problem instances.

| ID | $r_G = 7000, r_B = 8000$ | | | $r_G = 20000, r_B = 40000$ | | |
|---|---|---|---|---|---|---|
| | Time Saving (sec.) | Time Saving (%) | Optimality Gap | Time Saving (sec.) | Time Saving (%) | Optimality Gap |
| 1 | 0.1 | 9.6% | 0.00010% | 0.0 | 5.3% | 0.00010% |
| 2 | 0.3 | 5.3% | 0.00039% | 0.0 | 2.3% | 0.00028% |
| 3 | 72.7 | 17.0% | 0.00016% | 38.6 | 39.4% | 0.00037% |
| 4 | 56.2 | 14.5% | 0.00049% | 23.7 | 30.3% | 0.00009% |
| 5 | 55.6 | 13.8% | 0.00084% | 20.4 | 29.1% | 0.00068% |
| 6 | 65.7 | 15.9% | 0.00067% | 28.7 | 33.5% | 0.00047% |
| 7 | 62.4 | 14.9% | 0.00038% | 23.8 | 31.6% | 0.00036% |
| 8 | 52.4 | 14.7% | 0.00060% | 23.2 | 31.2% | 0.00079% |
| 9 | 69.4 | 17.4% | 0.00022% | 35.1 | 36.9% | 0.00075% |
| 10 | 66.2 | 14.7% | 0.00017% | 20.8 | 28.0% | 0.00028% |
| 11 | 47.9 | 12.9% | 0.00051% | 15.2 | 24.9% | 0.00011% |
| 12 | 61.1 | 14.2% | 0.00039% | 20.2 | 27.3% | 0.00068% |
| 13 | 1689.5 | 85.0% | 0.00054% | 133.5 | 78.9% | 0.00072% |
| 14 | 1677.1 | 85.1% | 0.00008% | 126.4 | 78.2% | 0.00075% |
| 15 | 1750.8 | 86.0% | 0.00021% | 133.4 | 80.3% | 0.00059% |
| 16 | 1704.3 | 85.7% | 0.00082% | 136.6 | 78.9% | 0.00086% |
| 17 | 1903.2 | 86.1% | 0.00028% | 133.6 | 78.7% | 0.00020% |
| 18 | 2032.6 | 86.8% | 0.00083% | 137.7 | 80.2% | 0.00028% |
| 19 | 1724.5 | 86.2% | 0.00039% | 132.1 | 79.7% | 0.00055% |
| 20 | 1882.1 | 85.6% | 0.00011% | 130.1 | 79.8% | 0.00061% |
| 21 | 1710.2 | 85.9% | 0.00001% | 148.7 | 80.1% | 0.00016% |
| 22 | 1541.6 | 85.7% | 0.00048% | 130.5 | 80.5% | 0.00007% |
| All | 18225.9 | 74.1% | 0.00039% | 1592.6 | 64.3% | 0.00044% |

Table 3: Performances of the Approximated PMP algorithm

It is apparent from Table 3 that the approximation procedure is capable of saving a significant share of the computation time in exchange for a compromise of less than 0.001% in terms of optimality. The relative savings is particularly large in the hardest instances where it is most needed. The approximation procedure can be used with hard instances if time is a major consideration. Such a case may arise if, for example,  the PMP is embedded as a sub-problem in an algorithm that solves a more general problem or if it is used to guide on-line decisions in a dynamic environment.

It is very likely that the computation time of the PMP algorithm and the approximation procedure are sensitive to parameters of the problem such as the success probabilities of the operations or the ratios between the production costs, inspection costs, revenue, and penalty. However, since we based our numerical experiment on instances that are very large compared to serial production lines encountered in the industry, we believe that our algorithm demonstrates its applicability to a wide range of real-world problem instances.

## 6. Work-In-Process in the Production Line

Until now, we intentionally overlooked the issue of WIP and its interaction with the QCS configuration. However, as indicated by Lemma 1, in an optimal solution prescribed by the PMP model at least one workstation is a bottleneck. That is, its production (or inspection) rate is equal to the arrival rate of the items at its buffer. In such a situation, the WIP queue in the buffers in front of the bottleneck workstations may grow indefinitely. Therefore, even if the holding cost of WIP per item is small, the total holding costs cannot be ignored altogether.

Despite this shortcoming of the PMP model, we believe that it is useful in designing the QCS configuration and deciding upon production rates in many real-world cases. In particular, we claim that it is typically possible to operate a serial production line with a given QCS configuration very close to its maximum potential production rate (dictated by the bottleneck workstations) while keeping the WIP level relatively low. Hence, if the value of the optimal solution of the PMP is viewed as a theoretical upper bound, it is possible to operate the system with a profit rate that is very close to that bound. This statement is proved empirically by an extensive simulation study reported below.

In self-regulated production systems, the arrival of new items at the system is dynamically controlled by the operator based on the system's state. We show that it is possible to achieve a high production and profit rate with relatively low WIP by using the QCS configuration prescribed by the PMP model and a self-regulating production control. For demonstration purposes, we use a simple pull dispatching rule, namely

CONWIP. Under CONWIP (COnstant WIP) the total WIP level, including the raw material at the first buffer, is kept constant. A new item is dispatched whenever the production of an item is completed. The ability of CONWIP to balance production rate and WIP level in serial production lines is well known, see for example Spearman et.al. (1990) and Masin et. al (1999).

Our simulation study is based on a set of 576 instances with up to 50 stations each. We use twelve settings of processing/inspection times and costs. These settings, presented in Table 4, are equivalent to these of our benchmark set in the previous section but the success probabilities were reduced to create instances with a substantial number of QCSs. Indeed, in the optimal solutions for all the tested instances, the number of installed QCSs was 35-50% of the number of machines. For each time/cost combination, we created instances with ($r_G = 350, r_B = 400$) and ($r_G = 1000, r_B = 2000$) with two different line lengths, namely, $N = 30$ stations and $N = 50$ stations.

| ID | Mean Processing times / item | Mean Inspection Time / item | Mean Processing Cost / item | Mean Inspection Cost / item | Success Probability | Capital Cost / time unit |
|---|---|---|---|---|---|---|
| 1 | $x_v = 5$ | $x'_v(u) = v - u - 1$ | $c_v = 5$ | $c'_v(u) = v - u$ | $p_v = 0.98$ | $f'_v(u) = = 0.1(v - u)$ |
| 2 | $x_v = 5$ | $x'_v(u) = 3$ | $c_v = 5$ | $c'_v(u) = 3$ | $p_v = 0.98$ | $f'_v(u) = 0.3$ |
| 3-7 | $x_v = U(2{,}9)$ | $x'_v(u) = 0.2\sum_{i=u+1}^{v} x_v$ | $c_v = U(2{,}9)$ | $c'_v(u) = 0.2\sum_{i=u+1}^{v} c_v$ | $p_v = U(0.96{,}1)$ | $f'_v(u) = 0.02\sum_{i=u+1}^{v} c_v$ |
| 8-12 |  | $x'_v(u) = U(0.4{,}1.8)$ |  | $c'_v(u) = U(0.4{,}1.8)$ |  | $c'_v(u) = U(0.04{,}0.18)$ |

Table 4: Description of the Simulation Study Instances

The optimal solution of each of the above instances was simulated with six different processing/inspection times as listed below:

- Exponentially distributed times
- Four lognormal distributions with coefficients of variation (CV) equal to $\frac{1}{4}, \frac{1}{2}, 1,$ and 2, i.e., the standard deviation equal to $\frac{1}{4}, \frac{1}{2}, 1,$ and 2 times the mean processing/inspection time.
- Deterministic times

The mean processing/inspection times were the same for all distributions (see Table 4). Note that even with the deterministic times, the arrival process at each machine is stochastic due to the randomness in the success of each operation and the routing of items created by the QCSs.

Let us denote the number of stations (machines and QCSs) in a given configuration by $n$. Each of the combinations above was simulated with two levels of CONWIP, namely, $2n$ and $10n$, in order to explore two extreme candidate values for the CONWIP level. Overall, the full factorial simulation experiment consists of

$$12\ (times/costs) \times 2(N) \times 2(r_G, r_B) \times 6\ (distrbutions) \times 2(CONWIP\ ) = 576$$

configurations.

The period until the first $50n$ items were produced was considered as warm-up time and was eliminated from the estimation of the steady-state performances of each configuration. The estimation was then based on 100 blocks of $1000$ items each. That is, the first block starts when the $50n^{th}$ item is delivered, the second starts when $50n + 1000^{th}$ item is delivered and so on. The simulation was used to estimate the mean profit rate in each configuration. This rate is compared to the upper bound determined by the value of the optimal solution of the PMP. Note that in a given QCS configuration, the production/inspection cost per item is fixed and thus the production rate is proportional to profit rate.

In Tables 5-8, we present the average relative production rates. Each row in these tables corresponds to a group of production/inspection time/cost settings as described in Table 4. Each column corresponds to one of the above six distribution families. In each of the internal cells of these tables, the average profit rate as a fraction of the upper bound, derived from the optimal solution value, is presented together with the average half-width of a 95% confidence interval for these statistics (in parenthesis). The rates in configurations with CONWIP 2 and 10 are separated by a slash. Table 5 and Table 6 correspond to the 30 machine instances while Table 7 and Table 8 correspond to the 50 machine instances. Table 5 and Table 7 correspond to the instances with $r_G = 7000, r_B = 8000$ and Table 6 and Table 8 to the instances with $r_G = 20000$ and $r_G = 40000$.

| ID | Deterministic<br>$CV$=0 | Exponential<br>$CV$=1 | Log Normal<br>$CV = \frac{1}{4}$ | Log Normal<br>$CV = \frac{1}{2}$ | Log Normal<br>$CV = 1$ | Log Normal<br>$CV = 2$ |
|---|---|---|---|---|---|---|
| 1 | 99.75% (0.08%) /<br>99.72% (0.08%) | 97.09% (0.49%) /<br>99.61% (0.61%) | 99.7% (0.19%) /<br>99.77% (0.2%) | 99.19% (0.27%) /<br>99.78% (0.31%) | 96.56% (0.5%) /<br>99.69% (0.61%) | 78.41% (0.72%) /<br>98.82% (1.18%) |
| 2 | 99.77% (0.08%) /<br>99.88% (0.09%) | 91.35% (0.32%) /<br>99.15% (0.69%) | 99.48% (0.17%) /<br>99.81% (0.17%) | 98.28% (0.27%) /<br>99.88% (0.34%) | 91.44% (0.37%) /<br>99.7% (0.64%) | 70.34% (0.63%) /<br>95.79% (1.08%) |
| 3-7 | 99.57% (0.1%) /<br>99.6% (0.1%) | 94.19% (0.37%) /<br>99.65% (0.63%) | 99.63% (0.19%) /<br>99.65% (0.18%) | 99.05% (0.31%) /<br>99.67% (0.34%) | 93.68% (0.4%) /<br>99.51% (0.62%) | 72.29% (0.66%) /<br>96.68% (1.3%) |
| 8-12 | 99.82% (0.11%) /<br>99.85% (0.1%) | 92.5% (0.36%) /<br>99.73% (0.6%) | 99.8% (0.18%) /<br>99.89% (0.2%) | 99.1% (0.3%) /<br>99.88% (0.34%) | 92.09% (0.37%) /<br>99.7% (0.66%) | 70.08% (0.64%) /<br>96% (1.21%) |

Table 5: Estimated average production rate relative to the optimal PMP solution for the 30 machine instances, with $r_G = 7000,\ r_B = 8000$ when using CONWIP=2/CONWIP=10

| ID | Deterministic<br>$CV$=0 | Exponential<br>$CV$=1 | Log Normal<br>$CV = \frac{1}{4}$ | Log Normal<br>$CV = \frac{1}{2}$ | Log Normal<br>$CV = 1$ | Log Normal<br>$CV = 2$ |
|---|---|---|---|---|---|---|
| 1 | 99.89% (0.08%) /<br>99.86% (0.08%) | 97.23% (0.49%) /<br>99.75% (0.61%) | 99.84% (0.19%) /<br>99.91% (0.19%) | 99.33% (0.27%) /<br>99.92% (0.31%) | 96.7% (0.5%) /<br>99.83% (0.61%) | 78.55% (0.72%) /<br>98.95% (1.17%) |
| 2 | 99.84% (0.08%) /<br>99.95% (0.09%) | 91.42% (0.32%) /<br>99.22% (0.69%) | 99.55% (0.16%) /<br>99.88% (0.17%) | 98.35% (0.27%) /<br>99.95% (0.34%) | 91.52% (0.37%) /<br>99.77% (0.64%) | 70.42% (0.63%) /<br>95.86% (1.08%) |
| 3-7 | 99.83% (0.1%) /<br>99.87% (0.1%) | 94.45% (0.37%) /<br>99.91% (0.62%) | 99.9% (0.18%) /<br>99.91% (0.18%) | 99.31% (0.31%) /<br>99.93% (0.33%) | 93.94% (0.4%) /<br>99.77% (0.61%) | 72.56% (0.66%) /<br>96.94% (1.29%) |
| 8-12 | 99.93% (0.11%) /<br>99.96% (0.1%) | 92.61% (0.36%) /<br>99.84% (0.6%) | 99.91% (0.18%) /<br>99.99% (0.2%) | 99.21% (0.3%) /<br>99.98% (0.34%) | 92.2% (0.37%) /<br>99.81% (0.65%) | 70.19% (0.64%) /<br>96.1% (1.2%) |

Table 6: Estimated average production rate relative to the optimal PMP solution for the 30 machine instances, with $r_G = 20000,\ r_B = 40000$ when using CONWIP=2/CONWIP=10

| ID | Deterministic<br>$CV$=0 | Exponential<br>$CV$=1 | Log Normal<br>$CV = \frac{1}{4}$ | Log Normal<br>$CV = \frac{1}{2}$ | Log Normal<br>$CV = 1$ | Log Normal<br>$CV = 2$ |
|---|---|---|---|---|---|---|
| 1 | 99.55% (0.12%) /<br>99.67% (0.12%) | 94.92% (0.46%) /<br>99.02% (0.65%) | 99.6% (0.2%) /<br>99.59% (0.18%) | 99.12% (0.31%) /<br>99.53% (0.36%) | 94.49% (0.45%) /<br>99.29% (0.62%) | 73.48% (0.67%) /<br>96.95% (1.28%) |
| 2 | 99.66% (0.19%) /<br>99.75% (0.17%) | 89.34% (0.38%) /<br>99.17% (0.64%) | 99.19% (0.22%) /<br>99.84% (0.22%) | 97.69% (0.29%) /<br>99.35% (0.31%) | 88.66% (0.38%) /<br>99.1% (0.66%) | 65.52% (0.59%) /<br>93.68% (1.29%) |
| 3-7 | 99.27% (0.11%) /<br>99.28% (0.1%) | 94.36% (0.38%) /<br>99.29% (0.64%) | 99.34% (0.18%) /<br>99.33% (0.18%) | 99.06% (0.32%) /<br>99.28% (0.33%) | 93.89% (0.38%) /<br>99.04% (0.67%) | 71.23% (0.68%) /<br>97.55% (1.42%) |
| 8-12 | 99.78% (0.11%) /<br>99.73% (0.11%) | 94.09% (0.35%) /<br>99.6% (0.64%) | 99.74% (0.19%) /<br>99.77% (0.2%) | 99.52% (0.32%) /<br>99.71% (0.35%) | 93.47% (0.35%) /<br>99.76% (0.7%) | 70.64% (0.66%) /<br>97.14% (1.4%) |

Table 7: Estimated average production rate relative to the optimal PMP solution for the 50 machine instances, with $r_G = 7000,\ r_B = 8000$ when using CONWIP=2/CONWIP=10

| ID | Deterministic<br>$CV$=0 | Exponential<br>$CV$=1 | Log Normal<br>$CV = \frac{1}{4}$ | Log Normal<br>$CV = \frac{1}{2}$ | Log Normal<br>$CV = 1$ | Log Normal<br>$CV = 2$ |
|---|---|---|---|---|---|---|
| 1 | 99.78% (0.12%) /<br>99.89% (0.12%) | 95.14% (0.46%) /<br>99.25% (0.64%) | 99.83% (0.19%) /<br>99.82% (0.18%) | 99.34% (0.31%) /<br>99.76% (0.36%) | 94.72% (0.46%) /<br>99.51% (0.61%) | 73.71% (0.67%) /<br>97.17% (1.27%) |
| 2 | 99.8% (0.19%) /<br>99.9% (0.17%) | 89.48% (0.38%) /<br>99.31% (0.63%) | 99.33% (0.22%) /<br>99.98% (0.22%) | 97.83% (0.29%) /<br>99.49% (0.31%) | 88.8% (0.38%) /<br>99.25% (0.65%) | 65.67% (0.59%) /<br>93.81% (1.28%) |
| 3-7 | 99.74% (0.11%) /<br>99.75% (0.1%) | 94.83% (0.37%) /<br>99.76% (0.63%) | 99.81% (0.18%) /<br>99.8% (0.18%) | 99.53% (0.31%) /<br>99.75% (0.32%) | 94.36% (0.37%) /<br>99.51% (0.66%) | 71.7% (0.67%) /<br>98% (1.39%) |
| 8-12 | 99.94% (0.11%) /<br>99.9% (0.1%) | 94.26% (0.34%) /<br>99.77% (0.63%) | 99.91% (0.19%) /<br>99.94% (0.19%) | 99.68% (0.32%) /<br>99.88% (0.34%) | 93.64% (0.34%) /<br>99.93% (0.69%) | 70.81% (0.66%) /<br>97.3% (1.37%) |

Table 8: Estimated average production rate relative to the optimal PMP solution for the 50 machine instances, with $r_G = 20000,\ r_B = 40000$ when using CONWIP=2/CONWIP=10

It is apparent from the results presented in Tables 5-8 that the profit obtained from both CONWIP levels is typically close to the upper bound obtained by the optimal solutions of the PMP model. As expected, the larger the constant WIP level, the closer the profit rate under CONWIP is to the upper bound. Even for the extremely noisy processing/inspection time ($C.V = 2$), the profit rates were more than $93\%$ of the upper bound in all the 48 instances tested when the CONWIP level was 10.

The values of $(r_G, r_B)$ as well as the number of machines $(N)$ seems to bear no influence on the ratio between the actual profit rate under CONWIP and the theoretical upper bound. To test this hypothesis, we ran a linear regression with this ratio as a dependent variable and with $CV$ (the coefficient of variation), CONWIP (the CONWIP level), and $N$ and $r_G$ as independent variables. Each point in this regression model is the estimated ratio for one of the 576 tested configurations. The results indicated that both CV and CONWIP are statistically significant in this model ($P - value < 10^{-41}$) while $r_G$ and $N$ are not ($P - value >$ 0.58 and $P - value >$ 0.91 respectivly). Therefore, we tend to believe that our finding above holds regardless of the number of stations in the production line and the final product price. As expected, the coefficient of the independent variable $CV$ in the regression model is negative while the coefficient of the CONWIP level is positive.

In conclusion, our simulation study showed that if the system is controlled by a dynamic dispatching policy, such as CONWIP, it yields provably near-optimal profit rates. When the variability of the processing time is not extremely high, such dispatching policies can work with a reasonably low level of WIP.

## 7. Managerial Insights

In this section we employ the devised algorithm to derive some insights about the nature of the optimal inspection effort allocation and its relation with various key parameters of the production line. In particular we examine the effects of the success probabilities, the processing time and cost, and the inspection time and cost (where QCSs are installed) on the QCS configuration and on the expected profit. These insights are derived based on experimentation with four 100-machine instances with characteristics as described in Table 9. We believe that these instances represent a variety of situations. In instances A and C the QCS cost and duration is related to the length of the inspected segment while in B and D the inspection cost is fixed. In instances A and B all the machines and QCSs are identical while in C and D their characteristics are drawn from some probability distribution as described in Table 9. We refer to the values in the table as base values and we will check how the QCS configuration is affected when these values are changed.

| ID | Mean Processing times / item $x_v$ | Mean Inspection Time / item $x'_v(u)$ | Mean Processing Cost / item $c_v$ | Mean Inspection Cost / item $c'_v(u)$ | Success Probability $p_v$ | Capital Cost / time unit $f'_v(u)$ |
|---|---|---|---|---|---|---|
| A | 5 | $1+v-u$ | $c_v = 5$ | $1+v-u$ | $\sqrt[100]{0.7}$<br>$\approx .9964$ | $0.1(v-u)$ |
| B | | $x'_v(u) = 3$ | $c_v = 5$ | 3 | | $.3$ |
| C | $U(2,8)$ | $\mathrm{U}(.5,1.5)$<br>$\times 0.2\sum_{i=u+1}^{v} x_v$ | $c_v = U(2,8)$ | $\mathrm{U}(.5,1.5)$<br>$\times 0.2\sum_{i=u+1}^{v} c_v$ | $U(.995,.999)$ | $0.02\sum_{i=u+1}^{v} c_v$ |
| D | | $U(0.4,1.8)$ | | $U(0.4,1.8)$ | | $U(0.04,0.18)$ |

Table 9: Description of the four 100-machine instances

For each of the four instances we solved the PMP numerous times in order to check the effects of changes to each of the six parameters in the table. We will present here only a sample of the obtained results. A Matlab routine that reproduces the complete set of results is available as an electronic companion to this paper. Our discussion below is based on all of these results.

In order to check the effect of the success probabilities (of the operations) we solved the problems with various probabilities ranging from 0.9 to 0.995. The obtained configurations (optimal solutions) of instance A are depicted in Figure 2 below where each line represents the optimal configuration relative to a given success probability (as shown in the vertical axis of the graph). Each dot represents the location of a QCS (as shown on the horizontal axis).

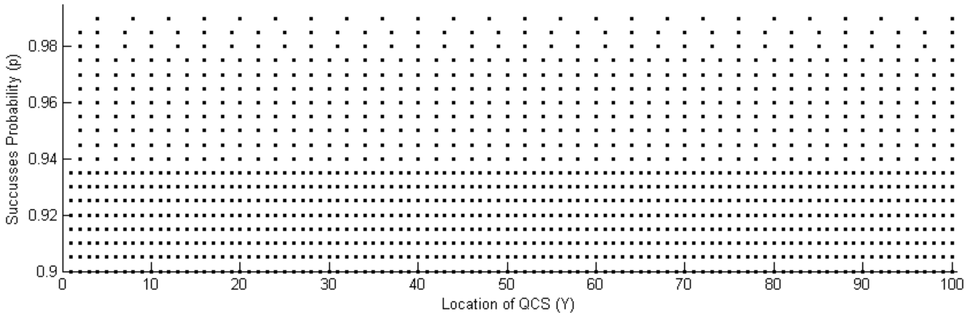

**Figure 2**: The optimal QCS configuration of Instance A with various success probabilities

As expected, one can see that the density of the QCS decreases as the success probability approaches one. The configuration, however, seems robust to slight changes in the success probabilities of the machines, and some locations are attractive enough to host a QCS for the entire range of the checked probabilities.  In Figure 3, we plot the profit rate obtained from the optimal solution against the successes probabilities. It can be observed that the effect is fairly dramatic. Indeed, increasing the reliability of the process not only allows saving on rework and inspection effort, but it also allows increasing the production rate leading to an increase in the profit per time-unit. An optimal configuration of QCS only partially mitigates the negative effects of poor quality.

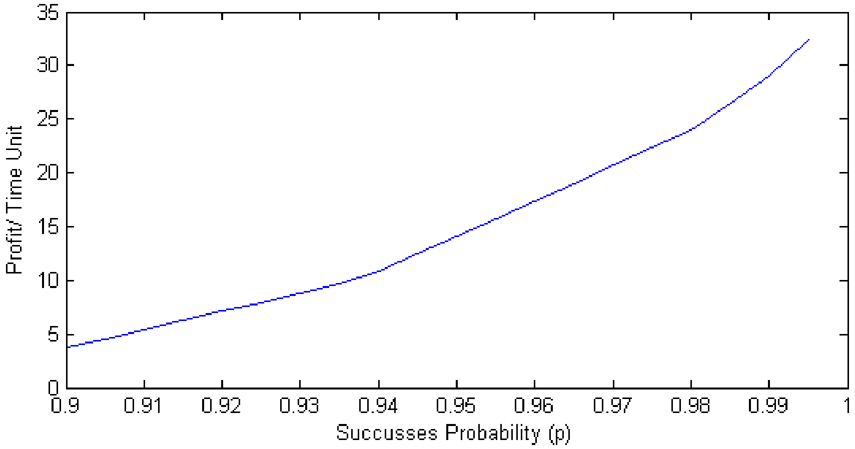

**Figure 3**: The optimal profit obtained from Instance A with various success probabilities

Next we examine the effect that the machines' processing times have on the optimal QCS configuration. In Figure 4, each line represents the optimal QCS configuration that corresponds to the base processing time multiplied by a factor as given in the vertical axis. We observe that the number of QCSs in the optimal configuration is decreased as the processing time increases. This phenomenon can be attributed to the fact that the reduction in production rate imposed by the slower machines increases the relative weight of the fixed inspection cost $f'$. Indeed, since the utilization of the QCSs decreases, the mean inspection cost per item increases. When the processing time is long enough, it becomes impossible to make a positive profit from the system and the optimal solution is not to produce at all; therefore, no QCSs are installed. Nevertheless, as with the success probability, it seems that the optimal QCS configuration is fairly robust to small changes in the processing times and many QCS locations remained unchanged or changed slightly when this parameter was increased. The expected profit is plotted versus the processing time on the right hand side of the figure. It appears that the profit rate decreases super-linearly since it is affected by both the reduction in the production rate and the higher average inspection cost per item.

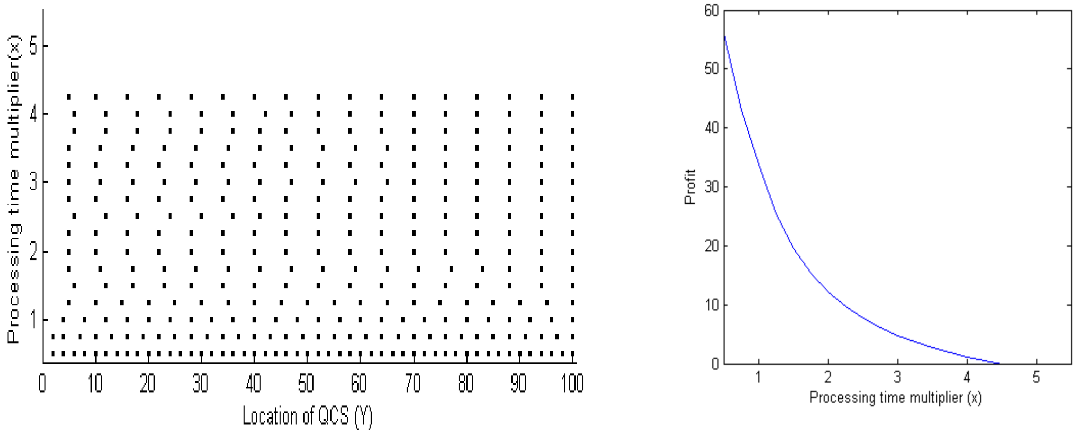

**Figure 4**: The optimal QCS configuration and profit with various processing times (Instance A)

In Figure 5, the effect of the inspection time on the QCS configuration and profits is examined. Here the results are somewhat intriguing: in instance A, increasing all the inspection times by some common factor results in an increase in the density of QCSs along the line. This is because in this instance, the inspection time is proportional to the length of the inspected segment. Thus, in order to avoid situations where the QCSs become bottlenecks and slow down the line, it is better to use more QCSs that cover shorter segments, even at the expense of a higher inspection effort. At some point, the optimal solution is to install a QCS after each machine; from this point on the production rate is dictated by the inspection rate. With the inspection cost structure of Instance A, it is optimal to stick with this configuration throughout the rest of the test range. A similar phenomenon is observed in Instance C where the inspection times are also related to the length of the segment. However, in instances B and D, the effect of increasing the inspection time is the opposite - once the QCSs become bottlenecks, the optimal number of QCSs begins to decrease.

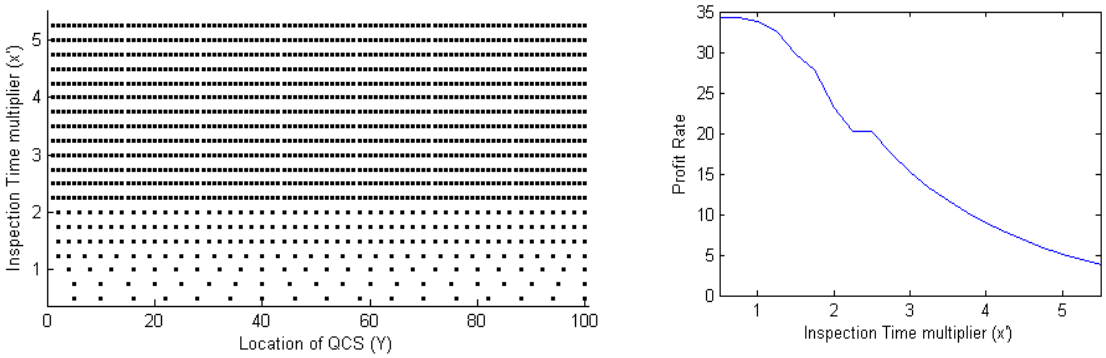

**Figure 5**: The optimal QCS configuration and profit with various inspection times (Instance A)

In figures Figure **6**, Figure **7**, Figure 8 we see that increasing the processing costs or inspection costs proportionally does not affect the optimal QCS configuration as long as it still possible to operate the system profitably. We note that this phenomenon is directly related to the cost structure of Instance A where the inspection costs are proportional to the length of the segment. In this case, no savings can be gained from switching to a sparser QCS configuration. In Figure 9, we show the same results for instance B where the inspection costs are fixed regardless of the length of the inspected segment. Here, an increase in the inspection cost makes inspection effort less attractive. As a result, the number of installed QCS is gradually decreased as the inspection costs increase. However, even in this case, the processing costs still do not affect the optimal QCS configuration significantly as long as it is still profitable to produce (see first row of Figure 9). This can be explained by the fact that the savings that could be obtained by reducing the rework and processing cost by installing more QCSs is canceled by the additional inspection cost per item due to the reduction in the production rate. We note that a similar phenomenon is observed in instances C and D (with random parameters). In these instances the number of QCSs remains more or less constant as the processing cost is increased, however their location changes. This is visualized in Figure 10.

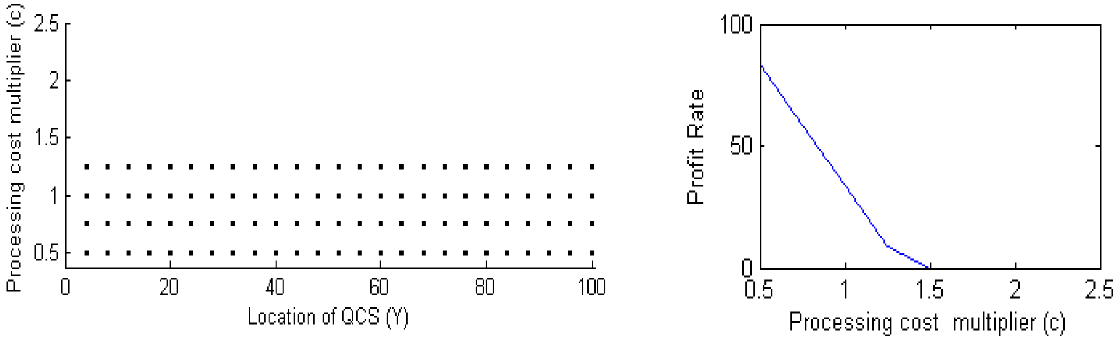

**Figure 6**: The optimal QCS configuration and profit with various processing costs (Instance A)

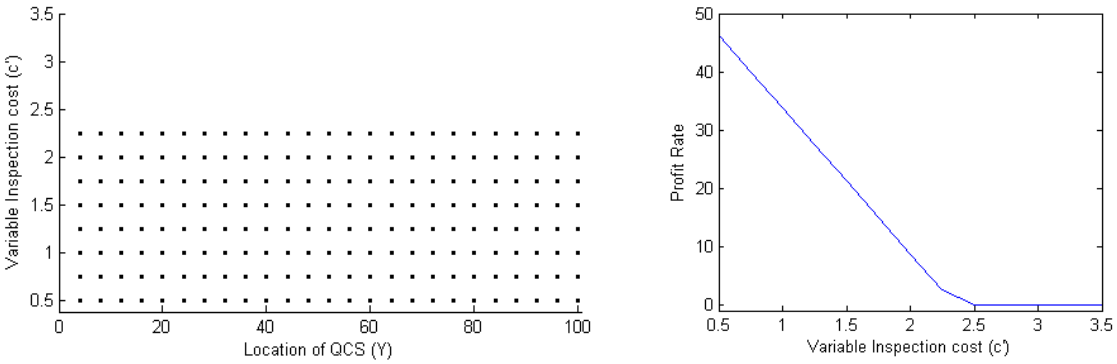

**Figure 7**: The optimal QCS configuration and profit with various variable inspection costs (Instance A)

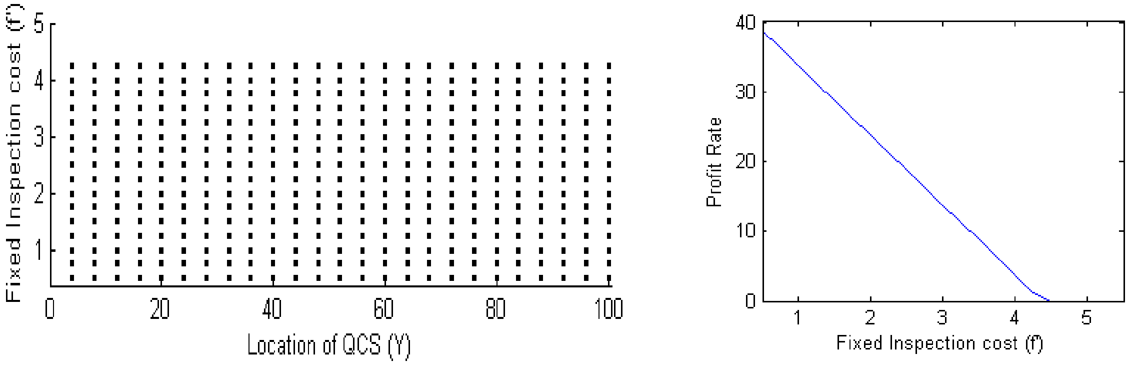

**Figure 8**: The optimal QCS configuration and profit with various fixed inspection costs (Instance A)

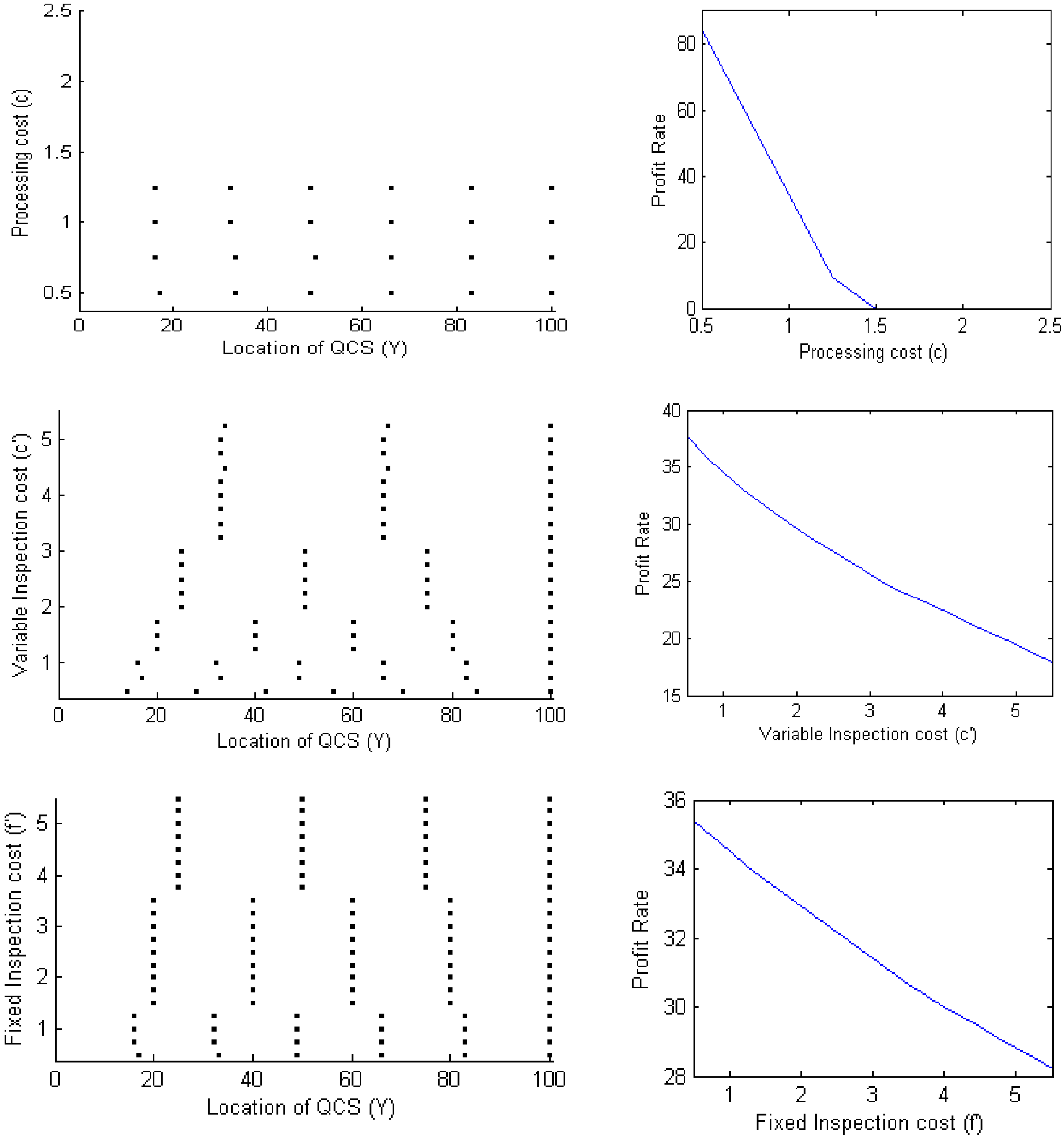

**Figure 9**: The optimal QCS configuration and profit with various processing and inspection costs (Instance B)

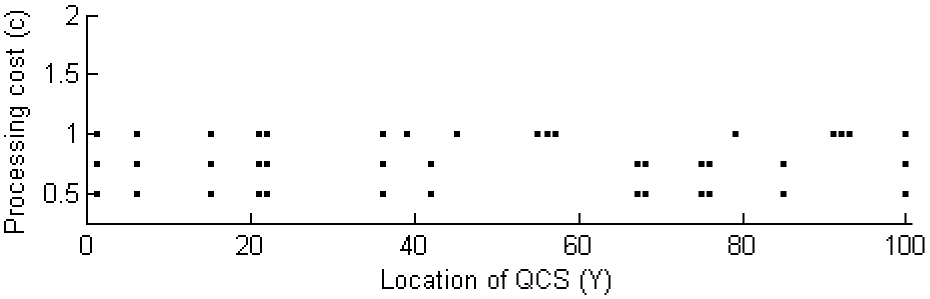

**Figure 10**: The optimal QCS configuration with various processing costs (Instance C)

## 8. Conclusions and further research

In this study, we formulated a profit maximization problem of a manufacturing system as a series of shortest path problems. We believe that this methodology may be useful for other problems that can be decomposed in a similar way. Specifically, good candidates for this approach are problems that involve allocating resources along a production line, where the allocation affects the potential rate of the line in a complex way.

Quality control systems in production systems, such as those we find at semiconductors fabrication facilities, may be much more complex than the simple serial line model presented in this paper and indeed further study is needed to provide a complete solution for their optimal design problem. However, we believe that the optimal solution of our simplistic model can provide initial guidelines for the designers of these lines especially where the number of potential locations for QCSs is large and it is impractical to test or simulate significant subsets of the feasible solutions.

Next, we discuss several problems for further research. Penn and Raviv (2008) incorporated holding costs into a QCS configuration model, where nonconforming items are scrapped. Further study should explore a similar model in the online rework setting studied here.

In practice, many production lines are re-entrant, i.e., jobs may visit some machines several times. In this setting, the workload on each station may be affected by several operations that may belong to several different segments (inspection-wise). Hence, the decomposition used in this study cannot be directly applied and a different approach is called for.

Many authors, e.g., Yum and McDowell (1987), considered a model where inspection errors may occur. However, no efficient solution approach was introduced for the inspection allocation problem in this setting. We believe that the methodology presented here may serve as a vehicle for finding efficient approximation algorithms for such problems.

## Appendix – List of Notation

$a$ Arrival rate of item to the system / production rate in a stable system

$c_i$ Variable production cost per item on $M_i$

$c'_v(u)$ Variable inspection cost per item on $QC_v$ assuming the previous installed QCS is $QC_u$. $c'_v(0)$ is the inspection cost on $QC_v$ assuming it is the first QCS in the line.

$C(u,v)$ Total production, inspection and penalty cost per item incurred by the segment $M_{u+1},\dots,M_v,QC_v$

$f'_v(u)$ Fixed cost per unit of time for installation of $QC_v$ assuming the previous installed QCS is $QC_u$

$L_i$ The location of the QCS that precedes $M_i$ in a given configuration (or 0 there is no such QCS). $L_{N+1}$ is the location of the last QCS in the line.

$M_i$ Machine $i$

$N$ Number of machines in the line

$p_i$ Success probability of the operation on machine $M_i$

$p_{u,v}$ Success probability on all the machines $M_{u+1},\dots,M_v$

$QC_i$ Quality control station located after $M_i$

$r_B$ Penalty incurred by a nonconforming item delivered by the system

$r_G$ Market price of a finished product

$T(u, v)$ Maximum throughput of the segment $M_{u+1}, \dots, M_v, QC_v$

$x_i$ Mean processing time of operations on machine $M_i$

$x'_v(u)$ Mean processing time of inspection task on $QC_v$ assuming the previous installed QCS is $QC_u$

$Y$ The set of installed QCSs in a particular solution

$Y^*(a)$ The optimal QCS configuration for production rate $a$

$\rho_i$ The ratio between the rework processing time and the “first time” processing time of an operation on $M_i$

$\zeta_i$ The ratio between the rework cost and the “first time” cost of an operation on $M_i$

## References

Cormen H.T., C.E. Leiserson, R.L. Rivest, and C. Stein, 2001, Introduction to Algorithms, Second Edition, MIT Press and McGraw-Hill

Emmons H. and G. Rabinowitz, 2002, Inspection allocation for multistage deteriorating production systems, IIE Transactions, 34, 1031–1041

Lindsay G.F. and A.B. Bishop, 1964, Allocation of screening inspection effort: a dynamic programming approach, Management Science, 10:342-352

Mandroli S.S., A.K. Shrivastava, and Y. Ding, 2006, A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes, IIE Transactions 38(4), 309-328

Masin M., Y.T.Herer, and E.M. Dar-El, 1999, CONWIP-based production lines with multiple bottlenecks: performance and design implications, IIE Transactions, 31, 99-111

Penn M. and T. Raviv, 2007, Optimizing the Quality Control Station Configuration, Naval Research Logistics, 54, 301-314

Penn M. and T. Raviv, 2008, A Polynomial Time Algorithm for Solving a Quality Control Station Configuration Problem, Discrete Applied Mathematics, 156, 412-419

Raghavachari M. and G.K. Tayi, 1991, Inspection Configuration and Reprocessing Decisions in Serial Production Systems, International Journal of Production Research, 29:5, 897-911

Raz T., 1986, A Survey of Models for Allocating Inspections Effort in Multistage Production System. Journal of Quality Technology, 18, 239-247.

Shiau Y-R., 2002, Inspection resource assignment in a multistage manufacturing system with an inspection error model, International Journal of Production Research, 40:8, 1787-1806

Spearman, M., D. Woodruff, and W. Hopp, 1990, CONWIP: a pull alternative to kanban, International Journal of Production Research 28, 879-894

Taneja M., S. M. Sharma and N. Viswanadham, 1994, Location of Quality-Control Stations in Manufacturing Systems: A Simulated Annealing Approach, Systems Practice, 7:4, 367-380

Tayi G. K. and D. P. Ballou, 1988, An Integrated Production-Inventory Model with Reprocessing and Inspection, International Journal of Production Research, Vol. 26:8, 1299-1315

White L.S., 1969, Shortest route models for the allocation of inspection effort on a production line, Management Science, 15, 249-259

Yum B.J. and E.D. McDowell, 1987, Optimal Inspection Policies in a Serial Production System Including Scrap, Rework and Repair: an MILP approach, International Journal of Production Research, 25:1451-1464