Joint assignment of buffer sizes and inspection points in unreliable transfer lines with scrapping of defective parts

Fatima Zahra Mhada\textsuperscript{a}, Roland P. Malhamé\textsuperscript{b,*} and Robert Pellerin\textsuperscript{c}

\textsuperscript{a}University Mohammed V- Souissi/École Nationale Supérieure d’Informatique et d’Analyse des Systèmes, Rabat, Morocco; \textsuperscript{b}Department of Electrical Engineering, École Polytechnique de Montréal and GERAD, Montréal, Canada; \textsuperscript{c}Department of Mathematics and Industrial Engineering, École Polytechnique de Montréal and CIRRELT, Montréal, Canada

Received 30 May 2013; accepted 1 June 2013

An unreliable single part type transfer line with fixed inter machine buffer sizes is considered. In general, imperfect machines operating on imperfect raw material, or partially processed raw material, will result in the production of a mix of conforming and non conforming parts. The problem of optimal joint assignment of buffer sizes and inspection station positions is considered. The performance measure to be optimized is a combination of work in process storage, finished parts shortage, and parts inspection costs. A simplified quality aware machine Markovian model is proposed, and an approximate decomposition scheme for the analysis of the resulting transfer line model is developed, and then validated against detailed Monte Carlo simulations. The decomposition scheme lends to a dynamic programming based buffer size optimization scheme. The latter is implemented; joint buffer and inspections station positioning optimization results are reported.

Keywords: optimal control; joint buffer sizing inspection station positioning; quality; production lines

1. Introduction

Quality and quantity modeling in manufacturing systems (the case of transfer lines is particularly of interest here) have long been studied separately, even if though they are highly coupled issues. In the literature, it is commonly assumed that product quality is perfect so that the impact of quality failures on the design of production policies is ignored.

Both quantity and quality modeling have the same general objectives: to minimize production cost and to maximize productivity.

Quality problems can in general be classified into two broad categories (Kim & Gershwin, 2008; Montgomery, 1991): (i) common, i.e. individual part related quality failures which can be attributed to an inherent defect in raw material as it interacts with the imperfect tuning of a particular machine and (ii) assignable, or persistent quality failures; they depend on a machine entering a particular absorbing quality failure mode during which, if not attended to, the machine will produce on average a distinctly higher number of defective parts. In practice, a mix of (i) and (ii) quality failure types will occur. Furthermore, one can imagine a class of machine quality states (iii) where quality

*Corresponding author. Email: roland.malhame@polymtl.ca

© 2013 The Author(s). Published by Taylor & Francis.
This is an open-access article distributed under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/3.0/, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the named author(s) have been asserted.
problems occur only intermittently (e.g. a loose machine component, the effect of which on production is recorded only under circumstances of special stress, while the machine appears to behave normally when the stress is removed (Mhada, Malhamé, & Pellerin, 2013). Cases (i) do not require corrective maintenance as there does not appear to be an assignable cause to their quality problems, while cases (iii), despite the fact that they could benefit from corrective maintenance action, will in general escape detection because of the intermittency of the observed quality problems. In general, cases (ii) are the ones which will benefit most from any corrective maintenance action.

An issue specific to the existence of quality problems of type (ii) is that they tend to dictate a preference for low buffer sizes, because this will limit the average time delay between the instant of appearance of an abnormal rate of defects at a given machine, say in a transfer line, and that at which quality problems are first diagnosed by an inspection machine further downstream. Earlier detection of persistent quality problems will in general limit the waste of storage space induced by defective parts in the system, as well as the loss of useful production time of machines downstream of the quality failure, thus improving in part productivity. On the other hand, the lack of reliability of individual machines dictates a priori larger buffer sizes to allow the system to continue to function when a particular machine is down. A first source of quality/quantity considerations conflict is thus identified. While this buffering size conflict is not as obvious for quality cases (i) and (iii), other paramount issues shared by all quality failure cases are as follows: the recognition that the effective cost of buffering is increased by the presence of defective parts in the stored mix and the recognition that inspection units must be placed so as to minimize inspection costs, while scrapping defective parts as soon as possible in order to limit both overall parts storage costs, as well as the loss of useful machine production time. Thus, it becomes clear that globally optimal decisions can only be reached when both quality and quantity issues are captured by adequate mathematical models, and are addressed simultaneously.

Despite the need for a greater integration of quality and production considerations, there are only a limited number of research results in this field: Kim and Gershwin (2005, 2008) study the relationship between quality issues and production policy issues by assuming that machines can enter an/a (unobservable) quality failure mode which is absorbing, until proper maintenance is carried out. Furthermore, their model is hybrid in that while the production process is assumed to be fluid, quality is not a temporary attribute of a flow but rather the property of a discrete part. Although not inclusive of statistical process control, an inspection process is modeled and the analysis succeeds in capturing the previously alluded to quantity/quality dilemma when it comes to buffer sizing in the presence of persistent quality failures.

Colledani and Tolio (2005, 2006, 2009, 2011) consider a production system composed of unreliable manufacturing stations and inspection stations with different failure modes. Statistical quality control charts are introduced at inspection stations and act as noisy measurements on the quality state of the machines. Powerful decomposition methods for the integrated production/quality performance studies of the line are developed. However, the modeling framework remains entirely discrete in the way time flows and in the production process itself. Again, the model permits an analysis of the impact of buffer size on the latency of detection of persistent quality failures, and how that in turn interacts with the statistical control process, and overall performance. An extension of this modeling setup to asynchronous transfer lines is currently under development (Colledani, 2011).
These researches Colledani and Tolio (2005, 2006, 2009, 2011) and Colledani (2011) constitute a very promising framework concentrated on exploring persistent quality failures, although still somewhat incomplete in that it does not appear to address the statistical correlations that will inevitably be present when detecting defective parts at various inspection points, as well as the impact of scrapping (if such is the policy) of defective parts as they are detected. In this paper, our efforts are focused instead on addressing the impact on proper buffer sizing and inspection positioning of cases (i) and (iii) of quality failures, rather than persistent quality failures, under the assumption that scrapping of defective parts upon detection is enforced. Statistical process control is not specifically modeled, and instead 100% inspection stations are considered. The fraction of defective parts in the system is adequately kept track of at all stages. The transfer line model is asynchronous, and an optimization formulation is specifically considered at the outset. Importantly, unlike (Colledani, 2011; Colledani & Tolio, 2005, 2006a, 2009, 2011), our decomposition method for approximate analysis is sequential, and thus lends itself naturally to an effective dynamic programming-based optimization scheme with essentially closed-form individual machining-stage-related cost expressions. Despite large dimensionality issues with dynamic programming in general, it has the definite merit of being an approach for global optimization, as opposed to gradient-based optimization methods which tend to discover local optima.

Meerkov and Zhang (2010) consider serial production lines consisting of producing and inspection machines that follow Bernoulli reliability and quality assumptions to gain insight into the nature of both production and quality bottlenecks. Such production lines are encountered in automotive assembly and painting operations where the downtime is relatively short and the defects are due to uncorrelated random events. Meerkov and Zhang (2011) extend the study to serial production lines with quality–quantity coupling machines, where product quality is interrelated with machine efficiency.

Also recently (Inman, Blumenfeld, Huang, Li, & Li, 2013), a survey of recent advances on the interface between production system design and quality has just been published, it provides evidence of the production system’s impacts on quality and identify opportunities for future research.

In our recent work (Mhada et al., 2013), we have introduced a new stochastic hybrid-state Markovian model with three discrete states. The first two, operational sound and operational defective, are not directly observable, while the third mode, failure, is observable. Production of defective parts is, respectively, initiated and stopped at the random entrance times to and departure times from the defective operational mode. The behavior of the model is numerically investigated, optimized under hedging policies, and subsequently compared to that of a tractable extension of the two-mode Bielecki–Kumar single machine model, where both conforming and defective parts are simultaneously produced in the operational mode, while the ratio of produced non conforming to conforming parts remains fixed (Mhada, Hajji, Malhamé, Gharbi, & Pellerin, 2011).

Based on the works (Mhada et al., 2011, in press), we use the two-mode model as a building block in a transfer line approximate decomposition method, aimed at jointly optimizing buffer sizing and inspection station positioning.

The rest of the paper is organized as follows: in Section 2, our quality imperfect, unreliable transfer line model is presented and our optimization problem is stated. In Section 3, we recall optimization results for two basic single machine models which act as building blocks in our decomposition methodology. In Section 4, we extend the Sadr and Malhamé (2004) decomposition/aggregation approach for the approximate
analysis of unreliable transfer lines to the current framework. Section 5 is dedicated to a Monte Carlo simulation-based evaluation of the accuracy of the approximate model of Section 4, while Section 6 extends the Sadr and Malhamé (2004) dynamic programming-based algorithm for buffering optimization, to the current framework where positioning optimization for inspection stations is also required. Numerical results are reported. In Sections 7, we numerically evaluate the added benefits of completely joint vs. “partially joint” optimization of buffer sizing and inspections station positioning in transfer lines, while in Section 8, we study the sensitivity of our optimal solutions to various problem parameters. Section 9 is our conclusion.

2. Modeling framework and statement of the optimization problem

We consider a single part-type transfer line consisting of a series of $n$ unreliable machines $M_i$ separated by $n$ buffers ($i = 1, \ldots, n$), and subjected to a fixed rate of demand for parts. The machines can be in either one of two modes: an operational mode and a failure mode. Failures and repairs occur in continuous time according to a two-mode Markov chain. Both modes are observable. However, while in the operational mode, each machine can impart possible defects on the wip that it processes. For simplicity, we consider that the ratio of wip processed by machine $M_i$ with locally imparted defects to that of processed wip without such locally imparted defects is a constant, machine dependent known fraction $\beta_i$, $i = 1, \ldots, n$.

The production line can contain inspection stations located at the exit of wip $x_i$ and the provisioning point for machine $M_{i+1}$. The presence or absence of these stations is captured by a binary variable $\lambda_i$ with:

$$\lambda_i = \begin{cases} 1 & \text{if there is an inspection station at the exit of the stock } x_i \\ 0 & \text{otherwise} \end{cases}$$

The following notations and assumptions will be used in the development of our models; for machine $M_i$, $i = 1, \ldots, n$:

- $\alpha_i$: the mode, respectively, 1 if machine is operational, and 0 if failed;
- $p_i$: failure rate;
- $r_i$: repair rate;
- $k_i$: machine production capacity. It is assumed that $k_1 \geq k_2 \geq \ldots \geq k_n$. This condition guarantees that when both machines $M_i$ and $M_{i+1}$ are neither starved nor blocked, the rate of production of machine $M_{i+1}$ will never be limited above by that of $M_i$;
- $u_i(t)$: production rate; with $0 \leq u_i(t) \leq k_i$;
- it is assumed that instantaneous wip $x_i(t)$ is a completely homogeneous mixture of good parts denoted $x_{i1}(t)$ and bad parts denoted $x_{i2}(t)$. $q_i(t)$ denotes the ratio at time of bad parts to good parts within wip $x_i(t)$. Note that $q_i(t)$ will in general depend on $\beta_1, \beta_2, \ldots, \beta_i$ and furthermore: $x_i \triangleq x_{i1} \triangleq x_{i2} \triangleq (1 + q_i)x_{i1}$. In Subsection 4.3 below, we show that $q_i$ is a constant which is a function of $\beta_1, \beta_2, \ldots, \beta_i$ and $\lambda_1, \lambda_2, \ldots, \lambda_{i-1}$;
- $z_i$: the inventory level for $x_i$;
- $c_p$: the storage cost per time unit and per part;
- $c_n$: the shortage cost per time unit and per part;
- $c_I$: the inspection cost per pulled part;
- it is assumed that at inspection stations, all parts are inspected and furthermore,
inspection results are fully reliable. Note, however, that the case of partial inspections could in principle be handled. Finally, all finished parts are verified at an inspection station;

- machines $M_i$, $i = 1, \ldots, n - 1$ cannot have any backlog;
- machine $M_n$ is operated under either one of the following two constraints:

1. A finite buffer for finished parts, but unlimited capacity for backlogging demand (this relates to the Bielecki and Kumar (1988) optimization model).
2. Unfulfilled demand cannot be backlogged at $M_n$; however, the long-term probability of availability of conforming finished parts is constrained to be greater than or equal to some fixed value (service level).

- machine $M_1$ is never starved;
- the line must satisfy a demand rate $d$ from the stock of good parts $x_n$, which means that it must satisfy $(1 + q_n)d$ from the total stock $x_n$;
- wip $x_i(t)$ and finished parts inventory $x_n(t)$ evolve, respectively, according to:

$$
\frac{dx_i(t)}{dt} = u_i(t) - u_{i+1}(t) \quad \text{for} \quad i = 1, 2, \ldots, (n - 1) \quad \text{with} \quad x_i(t) \geq 0
$$

$$
\frac{dx_n(t)}{dt} = u_n(t) - (1 + q_n)d
$$

with $u_i(t) = \begin{cases} k_i & \text{if } x_i(t) < z_i \\ (1 + q_i) d & \text{if } x_i(t) = z_i \\ 0 & \text{if } x_i(t) > z_i \end{cases}
$ (2)

- $\tilde{d}_i$, $i = 1, \ldots, n$ is the long-term average number of parts pulled per unit time from the total stock $x_i(t)$. The average unit time inspection cost caused by the existence of defective parts will be proportional to $\tilde{d}_i$. More specifically, it is equal to $c_I \tilde{d}_i \lambda_i$.

As mentioned above, one inspection station is placed at the end of the transfer line, as a result guaranteeing that parts delivered to the customer are all conforming. However, we are considering the problem of adding another inspection station within the transfer line. Thus, the optimization problem studied here is that of the joint placement of an extra inspection station and the sizing of buffer spaces within the transfer line so as to minimize the long-term per unit time average global cost of storage, production shortages, and inspection. More precisely, the cost to be minimized over buffer size parameters and inspection machine position is:

$$
J_{z_i}(x_0^T, z_0^T) = \lim_{T \to \infty} \sup \left( \frac{1}{T} \left( \sum_{i=1}^{n-1} E \left[ \int_0^T (c_p x_i(t) + c_l \tilde{d}_i \lambda_i)dt / (x_i(0), z_i(0)) \right] \right)
\right.

+ E \left[ \int_0^T (c_p x_n^+(t) + c_n x_n^{-1}(t) + c_l (1 + q_n) d \lambda_n)dt / (x_n(0), z_n(0)) \right] \right)
$$

under conditions: $\sum_{i=1}^{n-1} \lambda_i = 1$, $\lambda_n = 1$ and $0 \leq x_i \leq z_i$, $i = 1 \ldots n$
with \( x^+(t) = \max (x(t), 0), x^-(t) = \max (-x(t), 0) \), \( x_0 = [x_1(0), \ldots, x_n(0)]^T \), \( a_0 = [a_1(0), \ldots, a_n(0)]^T \), \( z^T = [z_1, z_2, \ldots, z_n]^T \) and \( \lambda^T = [\lambda_1, \lambda_2, \ldots, \lambda_{n-1}]^T \).

The above formulation will be referred to as the \textit{combined storage-shortage cost-minimization problem}. A second optimization problem is one where backlog is not allowed. However, a service-level constraint is imposed dictating that the coefficient of availability of conforming finished products is some fixed desired number \( a_n^{\text{des}} \); this under the additional constraint that the long-term average rate of extraction of conforming finished parts be \( d \). Under these constraints, the cost to be minimized is the same as (3), with \( c_n = 0 \). This second formulation will be referred to as \textit{the storage cost-minimization problem under service-level constraints}.

### 3. Single machine building blocks for transfer line decomposition

In the following, we review analytical and optimization results for two basic single machine models which will serve as building blocks for our transfer line decomposition methodology. In one model (the Bielecki and Kumar (1988) or BK model), backlog is allowed, while in the other (the Hu model (1995)), no backlog is allowed.

Both models involve a single machine producing a single part type and attempting to respond to a constant demand rate \( d \). The machine state changes in continuous time according to a homogeneous Markov process: the state changes from down \((x(t) = 0)\) to up \((x(t) = 1)\) at a rate \( r \) and from up to down at a rate \( p \). When the machine is up, it can produce at any rate \( u(t) \) between zero and a maximum rate \( k \). Storage cost per part per unit time is \( c_p \). The production inventory at time \( x(t) \) is characterized by the following differential equation:

\[
\frac{dx(t)}{dt} = u(t) - d
\]

In what follows, we will recall both the BK model and the Hu model. Thereafter, we present a BK-modified model with the integration of the concept of quality.

#### 3.1. The BK model

In the BK model (Bielecki & Kumar, 1988), under the assumption that backlog is allowed, the objective is to discover a Markovian state feedback policy \( \{\pi\} \) that would minimize the following long-run average cost, where \( c_n \) is the backlog cost per unit part and unit time:

\[
J_{\pi_i}(x(0), z(0)) = \lim_{T \to \infty} \sup \left( \frac{1}{T} E \left[ \int_0^T (c_p x^+(t) + c_n x^-(t)) dt / (x(0), z(0)) \right] \right)
\]

The optimal control policy is characterized by a single critical inventory level called a hedging point \( z^* \), that the production system must maintain as long as possible a kind of insurance policy against the potential cost of shortages following machine failures. More analytical details will be provided in the quality related extension of the model below.

#### 3.2. Hu model

Hu (1995) studies the same production system, under the assumption that backlog is not allowed. The objective is to discover the Markovian state feedback policy \( \{\pi\} \) which
could minimize a long-run average cost consisting of the sum of the inventory cost and the shortage cost. The latter is considered proportional to the time spent at zero inventory. More precisely:

$$J_n(x(0), z(0)) = \lim_{T \to \infty} \sup \left( \frac{1}{T} \mathbb{E} \left[ \int_0^T \left( c_p x(t) + c_n I(x(t) = 0, z(t) = 0) \right) dt / (x_i(0), z_i(0)) \right] \right)$$

(5)

where $c_n$ is the shortage cost per unit time, and $I(.)$ is the indicator function.

Hu shows that the optimal control policy is still characterized by a single critical inventory level called hedging point $z$, and the optimal production rate satisfies:

$$u(t) = \begin{cases} 
  k & \text{if } x(t) < z \\
  d & \text{if } x(t) = z \\
  0 & \text{if } x(t) > z 
\end{cases}$$

Finally, the long-run average cost associated with an arbitrary hedging point $z$ has the form:

$$J(z) = \frac{\rho \left[ \left( c_p \frac{k(1-\exp(-\mu(1-\rho)z))}{1-\rho} - (p + r) z \exp(-\mu(1-\rho)z) \right) + c_n (1-\rho)p \right]}{(p + r)(1-\rho) \exp(-\mu(1-\rho)z)}$$

(6)

with: $\rho = \frac{r(k-d)}{pd}$, $\mu = \frac{p}{k-d}$ and the coefficient of availability of wip, i.e. the steady-state probability that the wip is available is:

$$a = 1 - \frac{p}{(p + r)(1-\rho) \exp(-\mu(1-\rho)z)}$$

(7)

### 3.3. The BK quality modified model

As is the case for the ordinary BK model, machine $M$ has a discrete state $\alpha(t)$ with two possible values: $\alpha(t) = 1$ ($M$ operational) and $\alpha(t) = 0$ ($M$ has failed). The difference here though is that when $M$ is operational, it produces a mixed parts flow including both conforming (good) parts and nonconforming (defective) parts. It is assumed that the ratio between defective parts and good parts is a constant $\beta$. We designate by $x_1(t)$ the inventory of good parts when nonnegative and the backlog of good parts otherwise, while $x_2(t)$ and $x(t)$, respectively, designate the inventory of defective parts, and the total inventory of parts. Note that $x_2(t)$ cannot become negative; also, according to our model, $x_1(t)$ and $x_2(t)$ must reach zero at the same time. Finally, when $x(t)$ is negative, it is equal to $x_1(t)$.

$$x_2(t) = \beta x_1(t) I[x_1(t) \geq 0],$$

(8)

$$x(t) = x_1(t) + x_2(t) = (1 + \beta) x_1(t)$$

(9)

where in the above, $I(.)$ is the indicator function.

The rate of demand for good parts is a constant $d$, because the mixture of conforming and non-conforming parts is assumed to be perfectly homogeneous, the demand rate for total parts is $(1 + \beta) d$.

Parallel to the ordinary BK model analysis, our objective is to determine the optimal feedback production control policy $\{\pi_t\}$ minimizing the following long-term measure of combined storage and backlog costs:
\[ J_{c_n}(x(0), z(0)) = \lim_{T \to \infty} \sup \left( \frac{1}{T} E \left[ \int_{0}^{T} (c_p x^+(t) + c_n x_1^{-}(t)) dt / (x(0), z(0)) \right] \right) \]  

The cost to be minimized in (10) can now be rewritten in terms of the \( x(t) \) variable as:

\[ J_{c_n}(x(0), z(0)) = \lim_{T \to \infty} \sup \left( \frac{1}{T} E \left[ \int_{0}^{T} \left( c_p x^+(t) + \frac{c_n}{1+\beta} x_1^{-}(t) \right) dt / (x(0), z(0)) \right] \right) \]  

with the \( x(t) \) dynamics:

\[ \frac{dx(t)}{dt} = u(t)I[\alpha(t) = 1] - (1+\beta) d \]  

As such, it becomes an ordinary BK problem with a unit backlog storage cost decreased to \( c^*_n = \frac{c_n}{1+\beta} \) and a constant demand rate increased to \( d^* = (1+\beta) d \). This allows us to conclude that the optimal policy is again a hedging policy with values of \( z^* \) and optimal cost \( J^* \) specified below, and obtained based on the expressions in Bielecki and Kumar (1988).

3.3.1. Optimal hedging point (total finished parts) \( z^* \)

\[ z^* = 0 \text{ if } \frac{k p \left( \frac{c_n}{1+\beta} + c_p \right)}{c_p (k - (1+\beta) d) (p+r)} \leq 1 \text{ and } \frac{k - (1+\beta) d}{p} > \frac{(1+\beta) d}{r} \]  

\[ z^* = \infty \text{ if } \frac{k - (1+\beta) d}{p} \leq \frac{(1+\beta) d}{r} \]  

\[ z^* = \frac{1}{r k - r (1+\beta) d - p (1+\beta) d} \ln \left( \frac{k p \left( \frac{c_n}{1+\beta} + c_p \right)}{c_p (k - (1+\beta) d) (p+r)} \right) \text{ otherwise} \]

3.3.2. Optimal cost \( J^* \)

\[ J(z^*) = \frac{c_n p k d}{(p+r) (r k - r (1+\beta) d - p (1+\beta) d)} \text{ if } z^* = 0 \]  

\[ J(z^*) = \frac{c_p (1+\beta) d}{r + p} + \frac{c_p}{r k - r (1+\beta) d - p (1+\beta) d} \ln \left( \frac{k p \left( \frac{c_n}{1+\beta} + c_p \right)}{c_p (k - (1+\beta) d) (p+r)} \right) \text{ if } z^* > 0 \]

For further details and analysis of this model, please refer to Mhada et al. (2011).

4. Flow line decomposition

Decomposition techniques play an important role in the analysis of performance for a given choice of buffer sizes in a transfer line, thus paving the way for parameter
optimization. Thanks to such decomposition techniques, a line of $n$ machines is approximately decomposed into $n$ separate machines. To do this, the influence of the universe upstream and downstream of each machine shall be approximately represented. The decomposition technique used in this work is an extension of the approach in Sadr and Malhamé (2004) to the case of unreliable and imperfect machines transfer lines, with scrapping of defective parts. We recall that Sadr and Malhamé (2004) is based on two strategies of approximation:

- The hypothesis of decoupling of any given machine mode and the binary activity state of the buffer supplying that machine. This assumption is referred to as the machine decoupling approximation.
- The demand averaging principle (DAP).

4.1. The combined wip supply/machine building blocks

The decomposition/aggregation methodology of Sadr and Malhamé (2004) provides a tractable approximation for performance evaluation of failure prone transfer lines (with perfect quality) under the class of Kanban policies. Furthermore, because of the peculiar resulting unidirectional causality propagation (from upstream to downstream), it allows one to define sequential decision stages (sequential buffer sizing). As a result, dynamic programming becomes an easily implemented optimization tool. The approximation method is based on building blocks involving the aggregation of an on–off process of wip availability, say the indicator function of active $x_i$, and the on–off machine it supplies, say $M_{i+1}$, into what is called a pseudo-machine, say $\tilde{M}_{i+1}$ (see Figure 1). This virtual machine is never starved, but is more likely to fail than the original one because its failures must account for starvation phenomena. In effect, $\tilde{M}_{i+1}$ is an aggregate representation of the complete transfer line up to machine $M_{i+1}$, as it appears viewed from the rest of the transfer line downstream.

In Figure 1, $M_{i+1}$ is a machine with state $z_{i+1}(t)$ which can be zero (failed) or 1 (operational) with, respectively, repair rate $\tilde{r}_{i+1}$ and failure rate $\tilde{p}_{i+1}$. Using the machine decoupling approximation, it is obtained as the Cartesian product of the Markov chain associated with machine $M_{i+1}$ and that associated with the $x_i$ supply availability indicator $I_i(t)$ ($I_i(t) = 1$ indicates that $x_i(t)$ is active). This four-state machine is subsequently collapsed into a further simplified two-state machine as shown to the right of Figure 1.

![Figure 1. The pseudo-machine $\tilde{M}_{i+1}$, $i = 1 \ldots n - 1$.](image-url)
\( \tilde{M}_{i+1} \) represents the machines upstream of given buffer \((i + 1)\). Following Sadr and Malhamé (2004) which was initially developed for transfer lines with only conforming parts, \( p_{si}, \tilde{r}_i \) and \( \hat{p}_i \) are given by the following expressions for \( i = 1, \ldots, n \):

\[
\begin{align*}
ps_i &= \tilde{r}_i \frac{1 - a_i}{a_i} \\
\tilde{r}_i &= \frac{(ps_{i-1} + p_i) \tilde{r}_{i-1} r_i}{p_i \tilde{r}_{i-1} + ps_{i-1} r_i} \\
\hat{p}_i &= \left( \frac{(ps_{i-1} + \tilde{r}_{i-1})(p_i + r_i)}{\tilde{r}_{i-1} r_i} - 1 \right) \tilde{r}_i
\end{align*}
\]

with: \( \tilde{r}_1 = r_1, \hat{p}_1 = p_1 \) and where \( a_i \) is the coefficient of availability of wip at buffer \( i \), i.e. the steady-state probability that wip is available at buffer \( i \).

The calculation of availability coefficient \( a_i \) is based on the so-called DAP, which can be applied only if the transfer line demand rate is a known constant (under our assumptions, it is the case). However, in the current quality aware context, the above calculation must be adjusted so as to account for the presence of both conforming and nonconforming parts, as well as that of inspection stations internal to the line. Details will be given in the next subsection as we discuss the all important DAP.

The condition of feasibility of demand must be satisfied by each pseudo-machine \( \tilde{M}_{i+1} \): if \( \tilde{M}_{i+1} \) is able to meet demand in the long run, then it is necessary that its long-term average production rate when always operated at its current full capacity exceeds the long-term average demand i.e.:

\[
\frac{\tilde{r}_i k_i}{\tilde{r}_i + \hat{p}_i} > \hat{d}_i
\]

### 4.2. The demand averaging principle

The DAP (Sadr & Malhamé, 2004) is used to approximate the effect of the machines downstream of a given buffer \( i \). It is based on recognizing that for a transfer line with perfect quality, if finished parts are being pulled at long-term rate \( d \), then all buffers in the transfer line will be subjected to that same long-term rate of extraction \( d \). Thus, rate \( d \) is an invariant across the transfer line. This observation Sadr and Malhamé (2004) must, however, be reassessed in the current context of mixed conforming and nonconforming parts. Indeed, if there are no inspection stations internal to the line, then while \( d \) is the rate at which conforming parts are being extracted, it is actually the total rate, both conforming and nonconforming parts, at which parts are being extracted from the finished parts buffer that becomes the line invariant. This rate is in fact \((1 + q_\theta)d\). Given this rate invariance property, the demand averaging principle is the approximation by virtue of which the actual complex stochastic parts extraction process from any buffer in the line is replaced by the simplest process, namely a constant consistent with that rate constraint. As a result, and according to DAP, buffer \( i \) is considered to be subjected to a constant rate of total parts extraction of value \( \frac{d_i}{a_i} = \frac{(1+q_\theta)d}{a_i} \) while it is active and where it is recalled that \( a_i \) is the a priori unknown total parts wip availability coefficient at buffer \( i \).

However, the presence of an inspection station (see Figure 2) will complicate things since an inspection station at the position \( j \) (\( \lambda_j = 1 \)) divides the line in two parts: the part of the line upstream of an inspection station and that downstream of it.
While it is straightforward to include more than one inspection station within the line, the focus here is on the case of a single internal inspection station placement.

Downstream of the inspection station, buffers when active must meet the demand rate:

\[ a_i \frac{d}{d_a} = \frac{1}{1 + q_n} \left( 1 + q_n \right) a_i d_a, \quad i > j, \]

while the upstream part of the line should provide a rate \((1 + q_n) d\) of good parts to the downstream part i.e. a rate \((1 + q_n) (1 + q)\) of total parts. Thus, buffers upstream of the inspection station, when active, must meet the long-term average demand rate:

\[ a_i \frac{d}{d_a} = \frac{1}{1 + q_n} \left( 1 + q_n \right) \left( 1 + q_n \right) d, \quad i \leq j. \]

We note that whatever the position of the inspection station in the line \((\forall j = 1, \ldots, n - 1)\), the demand rate \(d_i = (1 + q_n) (1 + q)\) \(d\), \(i \leq j\) is constant and greater than the demand rate \(d_i = (1 + q_n) d\) \(i > j\), itself also a constant.

Again summarizing, of all stochastic processes consistent with the above buffer rates of parts long-term extraction rate constraints, DAP is the approximation whereby the actual rates are taken to be constant while a buffer is active.

### 4.3. Recursive calculation of the \(q_i\)'s

Recall that \(x_{i_2}\) is the part of the total stock \(x_i\) consisting of non-conforming parts and \(x_{i_1}\) corresponds to the conforming parts. Thus, according to the assumptions of our model that means:

\[ x_{i_2} = q_i x_{i_1} \]  \hspace{1cm} (19)

\[ x_i = x_{i_1} + x_{i_2} = (1 + q_i) x_{i_1} \]  \hspace{1cm} (20)

\[ x_i = \frac{(1 + q_i)}{q_i} x_{i_2} \]  \hspace{1cm} (21)

Our aim is to calculate \(q_{i+1}\), the ratio of nonconforming to conforming parts in \(x_{i+1}\) given the parameters of the mix at buffer \(i\). More specifically, \(x_{i+1} = q_{i+1} x_{i+1}\). Indeed, only a fraction of the stock of good parts that will be processed by machine \(M_{i+1}\) remains conforming. This fraction is \(\frac{1}{1 + p_{i+1}}\); so for a small time increment \(\delta t\), the incoming quantity of good parts stored in buffer \(i + 1\) is equal to \(\frac{1}{1 + p_{i+1}}\) times the quantity of the stock \(x_{i_1}\) processed during \(\delta t\), say \(\delta x_{i_1}\). The rest of \(\delta x_{i_1}\), \(i.e.\) \(\frac{p_{i+1} \delta x_{i_1}}{1 + p_{i+1}}\) will be stored in buffer \(i + 1\) as nonconforming parts. Also, during the same time interval \(\delta t\), machine \(M_{i+1}\) will process a quantity \(\delta x_{i_2}\) of already nonconforming parts within wip \(i\), associated with \(\delta x_{i_1}\) and given as \(q_i \delta x_{i_1}\), unless an inspection station is present at buffer \(i\) \((\lambda_i = 1)\), in which case \(\delta x_{i_2}\) is rejected, and only good parts are processed by \(M_{i+1}\). If no inspection station is present at buffer \(i\), \(\delta x_{i_2}\) persists as noncon-
forming in wip \( i + 1 \). The resulting net ratio of nonconforming to conforming parts within wip \( i + 1 \) is then given by the following recursive calculation:

\[
q_{i+1} = \frac{(1 - \lambda_i) \delta x_{i+1} + \frac{\beta_{i+1}}{1 + \beta_{i+1}} \delta x_i}{\left( \frac{1}{1 + \beta_{i+1}} \right) \delta x_i}
\]

\[
= \frac{((1 - \lambda_i) q_i) \delta x_i + \frac{\beta_{i+1}}{1 + \beta_{i+1}} \delta x_i}{\left( \frac{1}{1 + \beta_{i+1}} \right) \delta x_i}
\]

\[
= (1 - \lambda_i) q_i (1 + \beta_{i+1}) + \beta_{i+1}
\]

It is initialized by \( q_1 = \beta_1 \).

4.4. Evaluation of internal buffer \( i \) induced costs

The cost of storage in buffer \( i \) is given by:

\[
J_i(z_i, \lambda_i) = \lim_{t \to \infty} E[c_p x_i(t)]
\]

Furthermore, under DAP, when active, buffer \( i \) is subjected to a constant, coefficient of availability dependent demand, so that wip evolves according to:

\[
\frac{dx_i(t)}{dt} = u_i(t) - \frac{d_i}{a_i}
\]

with \( u_i(t) = \begin{cases} 
  k_i & \text{if } x_i(t) < z_i \\
  \frac{d_i}{a_i} & \text{if } x_i(t) = z_i \\
  0 & \text{if } x_i(t) > z_i
\end{cases} \)

The cost and dynamics under (24) characterize a Hu (1995) machine subject to a constant, a priori unknown, demand \( \frac{d_i}{a_i} \) (recall the discussion in Subsection 4.2). Based on (7), and the pseudo-machine \( \tilde{M}_{i+1} \) obtained in Subsection 4.1, one can write:

\[
a_i = 1 - \frac{\tilde{p}_i}{\tilde{p}_i + \tilde{r}_i} \frac{(1 - \rho_i) (1 - \rho_i)}{1 - \rho_i \exp(-\mu_i(1 - \rho_i)z_i)}
\]

with: \( \rho_i = \frac{\tilde{r}_i(k_i - \mu_i)}{\tilde{p}_i + \tilde{r}_i} \mu_i = \frac{\tilde{p}_i}{k_i - \mu_i} \) and

\[
J_i(z_i, \lambda_i) = \frac{c_p \rho_i}{(\tilde{p}_i + \tilde{r}_i)(1 - \rho_i \exp(-\mu_i(1 - \rho_i)z_i))}
\]

under constraints: \( \sum_{i=1}^{n-1} \lambda_i = 1 \) and \( \lambda_n = 1 \).

Note that for given \( \tilde{p}_i, \tilde{r}_i \) and \( z_i \), (25) constitutes an implicit equation for unknown coefficient of availability \( a_i \) (a fixed point is obtained by using iterated substitutions starting from initial guess \( a_i^{(0)} = 1 \) much like in Sadr and Malhamé (2004)).

4.5. Evaluation of buffer \( n \) induced costs

Two distinct models will be used for the analysis of buffer \( n \) induced costs: the first model corresponds to a situation where backlog is allowed, but negative excursions are
penalized at a cost of \( c_n \) per part and unit time. This is the so-called combined storage-backlog cost-minimization problem. Calculations are carried out using a version of the BK theory (Sadr & Malhamé, 2004), modified to account for the presence of both conforming and nonconforming parts; the second model corresponds to a situation where backlog is not allowed. This is the so-called storage cost-minimization problem under service-level constraints. Calculations are carried out using the Hu theory with parameters dependent on the imposed service-level constraint.

4.5.1. Combined storage-shortage cost-minimization problem

Based on the BK theory Bielecki and Kumar (1988) and our previous work Mhada et al. (2011), the following expressions can be written for pseudo-machine \( M_{i+1} \) associated with storage as well as backlog costs:

\[
J_n(z_n, \lambda_n) = \lim_{t \to -\infty} E \left[ c_p x_n^+ (t) + c_n x_n^- (t) \right] = c_p z_n + \frac{\hat{p}_n k_n}{\delta (\hat{p}_n + \hat{r}_n) (k_n - \hat{d}_n)} \left( \frac{c_n}{1 + q_n} \exp(-\delta z_n) - c_p (1 - \exp(-\delta z_n)) \right)
\]

with \( \delta = \frac{\hat{r}_n}{\hat{d}_n} - \frac{\hat{p}_n}{k_n - \hat{d}_n} > 0 \) and \( \tilde{d}_n = (1 + q_n) d \).

The optimal hedging point (total finished parts) \( z^* \) is:

\[
\begin{align*}
  &z^* = 0 \quad \text{if} \quad \frac{k_n\hat{p}_n(\frac{c_n}{1+q_n} + c_p)}{c_p}(k_n - d_n)(\hat{p}_n + \hat{r}_n) \leq 1 \quad \text{and} \quad \frac{k_n - \tilde{d}_n}{\hat{p}_n} > \frac{\tilde{d}_n}{\hat{r}_n} \\
  &z^* = \infty \quad \text{if} \quad \frac{k_n - \tilde{d}_n}{\hat{p}_n} < \frac{\tilde{d}_n}{\hat{r}_n} \\
  &z^* = \frac{1}{\frac{\tilde{d}_n}{\hat{r}_n} - \frac{\hat{p}_n}{k_n - \tilde{d}_n}} \ln \left( \frac{k_n\hat{p}_n(\frac{c_n}{1+q_n} + c_p)}{c_p}(k_n - d_n)(\hat{p}_n + \hat{r}_n) \right) \quad \text{otherwise}
\end{align*}
\]

and the optimal cost \( J^* \) is:

\[
\begin{align*}
  &J(z^*) = \frac{c_n \hat{p}_n k_n \tilde{d}_n}{(\hat{p}_n + \hat{r}_n)(\hat{r}_n k_n - \hat{r}_n \hat{d}_n - \hat{p}_n \tilde{d}_n)} \quad \text{if} \quad z^* = 0 \\
  &J(z^*) = \frac{c_p \tilde{d}_n}{\hat{r}_n + \hat{p}_n} + \frac{c_p}{\frac{\tilde{d}_n}{\hat{r}_n} - \frac{\hat{p}_n}{k_n - \tilde{d}_n}} \ln \left( \frac{k_n\hat{p}_n(\frac{c_n}{1+q_n} + c_p)}{c_p}(k_n - d_n)(\hat{p}_n + \hat{r}_n) \right) \quad \text{if} \quad z^* > 0
\end{align*}
\]

4.5.2. Storage cost-minimization under service-level constraints

Based on (7), and for a given desired coefficient of availability of conforming finished parts \( \alpha_n^{\text{des}} \), one can write the constraint:

\[
\alpha_n^{\text{des}} = 1 - \frac{\hat{p}_n}{(\hat{p}_n + \hat{r}_n)} \left( 1 - \rho_n \exp(-\mu_n (1 - \rho_n) z_n) \right)
\]

The required hedging point \( z_n(\alpha_n^{\text{des}}) \) can be readily calculated using (30), in terms of (yet to be designed) variables \( \hat{r}_n \) and \( \hat{p}_n \) as:

\[
\begin{align*}
  z_n(\alpha_n^{\text{des}}) = \frac{1}{-\mu_n (1 - \rho_n)} \ln \left[ \frac{1}{\rho_n} \left( 1 - \frac{(1 - \rho_n)}{(1 - \alpha_n^{\text{des}})(\frac{\hat{p}_n + \hat{r}_n}{\hat{p}_n})} \right) \right]
\end{align*}
\]
with: \( \rho_n = \frac{\hat{r}_n(k_n - \frac{\hat{\alpha}_n}{\hat{\alpha}_n^2})}{\frac{\hat{r}_n \hat{\alpha}_n}{\hat{\alpha}_n^2}} \), \( \mu_n = \frac{\hat{p}_n}{(k_n - \frac{\hat{\alpha}_n}{\hat{\alpha}_n^2})} \) and the corresponding cost is:

\[
J_n(a_{\text{des}}^n, \hat{p}_n, \hat{r}_n) = \rho_n e^p \frac{k_n (1 - \exp(-\mu_n(1 - \rho_n) z_n(a_{\text{des}}^n)))}{(\hat{p}_n + \hat{r}_n)} \left( \frac{1}{1 - \rho_n} z_n(a_{\text{des}}^n) \exp(-\mu_n(1 - \rho_n) z_n(a_{\text{des}}^n)) \right)
\]

(32)

5. Approximate model validation

In a preliminary verification step, the approximate theoretical expressions derived in Section 4 have been validated against the results of Monte Carlo simulations for a large sample of four machine lines, and the results differed by at most 4%.

With this validation behind us, our objective now is to determine buffer sizes and the one extra inspection station location which solve the combine storage-shortage cost-minimization problem. So, in the next section we present a dynamic programming-based optimization method inspired by the work of Sadr and Malhamé (2004) as adapted it to the current problem at hand.

6. Optimization

The objective is to find all \( z_i \) and the only \( \lambda_i \neq 1 \), for \( i = 1, \ldots, n - 1 \) with \( \lambda_n = 1 \), values that minimize either the combined storage-shortage cost, or the storage cost under service-level constraints. It is possible to rewrite the optimization problem as a dynamic programming problem for a fixed choice of inspection station positions (\( \lambda \) vector fixed):

\[
J^*(\lambda) = \inf_{a_i \in A(a_{\text{des}}^i, \lambda), i = 1, \ldots, n} \left( \sum_{i=1}^{n-1} T^{(i)}(\hat{r}_i, \hat{p}_i, a_i, q_i) + T_F(\hat{r}_n, \hat{p}_n, q_n) + c_I \sum_{i=1}^{n} \lambda_i \hat{d}_i \right)
\]

(33)

\[
T^{(i)}(\hat{r}_i, \hat{p}_i, a_i, q_i) = c_p \left( \frac{k_i \hat{p}_i}{\sigma_i (k_i - \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2}) (\hat{r}_i + \hat{p}_i)} - \frac{k_i (1 - a_i)}{\sigma_i (k_i - \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2})} - \frac{1}{\sigma_i} \frac{(1 - a_i) (\hat{r}_i + \hat{p}_i)}{\sigma_i^2 (k_i - \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2}) \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2}} \right) - \ln \left( \frac{\hat{p}_i \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2}}{(\hat{r}_i (k_i - \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2}) \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2})} \right), \quad i = 1, \ldots, n - 1
\]

with: \( \sigma_i = (\hat{p}_i + \hat{r}_i) \frac{\hat{\alpha}_i}{\hat{\alpha}_i^2} - k_i \hat{r}_i \).

For the combined storage-shortage cost-minimization problem:

\[
T_F(\hat{r}_n, \hat{p}_n, q_n) = \left( c_p z_n^* + \frac{\hat{p}_n k_n}{\delta (\hat{p}_n + \hat{r}_n) (k_n - \frac{\hat{\alpha}_n}{\hat{\alpha}_n^2})} \left( \frac{c_n}{1 + q_n} - c_p \frac{1 - \exp(-\delta z_n^*)}{-\exp(-\delta z_n^*)} \right) \right)
\]

(34)

where \( \delta = \frac{\hat{\alpha}_n}{\hat{\alpha}_n^2} - \frac{\hat{p}_n \hat{\alpha}_n}{\hat{\alpha}_n^2} \) > 0 and the expression of \( z_n^* \) is given in (28), thus leading to the alternate cost expressions in (29).

For storage cost-minimization under service-level constraints problem:

\[
T_F(\hat{r}_n, \hat{p}_n, q_n) = \frac{\rho_n e^p k_n \left( \hat{p}_n \frac{(1 - \exp(-\mu_n(1 - \rho_n) z_n(a_{\text{des}}^n)))}{(\hat{p}_n + \hat{r}_n)} \frac{(1 - \rho_n) z_n(a_{\text{des}}^n)}{z_n(a_{\text{des}}^n)} \right)}{(\hat{p}_n + \hat{r}_n) (1 - \rho_n \exp(-\mu_n(1 - \rho_n) z_n(a_{\text{des}}^n)))}
\]

(35)
where \( q_n = \frac{\tilde{r}_n(k_n - \frac{1}{2})}{\tilde{p}_n} \), \( \mu_n = \frac{\tilde{p}_n}{(k_n - \frac{1}{2})} \) and the expression of \( z_n(a_n^{\text{des}}) \) is given in (31); if \( j \) is the inspection station location then for \( i = 1, \ldots, n \),

\[
\begin{cases}
    \bar{d}_i = (1 + q_n) d & \text{if } i > j \\
    d_i = (1 + q_n)(1 + q_j) d & \text{if } i \leq j
\end{cases}
\]

(36)

Note that, in (33), the decision variables are the coefficients of availability of parts (both conforming and non-conforming), instead of the buffer sizes, because their range is bounded; the state space itself is two-dimensional at each stage \( (\tilde{r}_i, \tilde{p}_i) \). Also note that buffer sizes \( z_i \) can be immediately calculated from the \( a_i \)'s, once the latter have been obtained. Finally, note that this is a constrained dynamic programming problem in the sense that, at each stage, the range of permissible \( a_i \)'s as defined by the sets \( A_i(a_{i+1}, \lambda) \) is dependent on the state at the next future stage \( (\tilde{r}_{i+1}, \tilde{p}_{i+1}) \). See Sadr and Malhamé (2004) for further details on these constraint sets. In the following, we will detail our numerical algorithm as applied to a homogeneous transfer line.

6.1. Numerical procedure

The numerical algorithm proceeds as follows:

- For each choice of the positioning of the internal inspection station \( \lambda_j = 1; j = 1, \ldots, n - 1 \) and \( \lambda_i = 0, i = 1, \ldots, n - 1, i \neq j \), we first calculate the different values of \( q_i \) and \( \bar{d}_i \) based on (22) and (36).
- For each fixed value of the \( \lambda \) vector, we solve a dynamic programming problem to determine the associated optimal buffer sizes, and minimal cost. This algorithm is deployed in two phases: (1) State space generation and (2) Application of the dynamic programming algorithm.

(1) State space generation: here, the decision variable \( a_i, i = 1, \ldots, n - 1 \) is discretized between its lower bound defined by feasibility condition (15) and its upper bound of 1. This discretization is then used to generate the sequence of discrete grid points in the two-dimensional \( [\tilde{r}_i, \tilde{p}_i] \) space, starting from the single point \( \tilde{r}_1 = r_1, \tilde{p}_1 = p_1 \), and using ((15)–(17)).

(2) Application of the dynamic programming algorithm: this yields the optimal sequence of \( a_i \) decisions, the optimal trajectory on the sequence of \( [\tilde{r}_i, \tilde{p}_i] \) planes, and the corresponding optimal buffer sizes \( z_i \) (based on (7)), for the given choice of the \( \lambda \) vector.

We add to the total cost of storage and possibly shortage (if the combined storage-shortage cost-minimization version of the problem is considered), as calculated by dynamic programming, the inspection cost corresponding to the fixed value of the \( \lambda \) vector.

When all permissible configurations of the \( \lambda \) vector have been explored, we compare the resulting associated optimal costs to determine the position \( \lambda_j = 1, j = 1, \ldots, n - 1 \), for which the cost is lowest. This position will represent the solution to our joint optimal buffer sizing and inspection station positioning problem.

6.2. A homogeneous machines line

We show an example of a dynamic programming problem solution. The line to be optimized is a 10 homogeneous machines line with \( \beta_i = 0.1, p_i = 0.2, r_i = 0.9 \) for
i = 1, . . . , n, with machine production capacity \( k_i = 4 \), demand rate \( d = 1 \), and storage, inspection and shortage costs given, respectively, as \( c_p = 1 \), \( c_i = 2 \) and \( c_n = 10 \). We study both the combined storage-shortage cost-minimization version of the problem, and that of storage cost-minimization under service-level constraints.

6.2.1. Combined storage-shortage cost-minimization problem

Figure 3 presents the optimal solution: the first picture on the left displays the optimal cost as a function of the location of the extra inspection station. The other pictures on the right display the different values of the optimal buffer sizes \( z_i \) and coefficient of availability of wip or inventory \( a_i \) for the case of an optimally located internal inspection station (following buffer 5 in this case).

We notice that for a homogeneous unreliable transfer line, the total minimal cost is a convex function of the position where \( \lambda_i \) is equal to 1, with \( \lambda_5 = 1 \) identifying the center of the line as the optimal inspection station position. This result is rather plausible for a homogeneous line in that the inspection station must not be placed too early as its usefulness would be limited as an instrument of rejection of non-conforming parts, and not too late, in that, the more non-conforming parts in the system, the higher the storage costs are and the less efficient the transfer line becomes in terms of productivity. Note that in the section of the line upstream of the inspection station, the optimal buffer sizes are greater than those downstream of the inspection station; this is because for the chosen \( \beta = 0.1 \) (1 non-conforming part for 10 conforming parts), the upstream section must satisfy a demand rate \textit{twice as large} as that of the downstream section.

The first machine in the line and the one just after the inspection station (\( M_6 \)) have almost the same optimal buffer size (since the inspection station resets to zero the \( q_5 \) parameter), yet their availability coefficients are very different (since the first machine is never starved while the second one is). The decrease in required buffer size from position
6.2.2. Storage cost minimization under service-level constraints

Figures 4, presents the optimal solution with a required conforming finished parts availability rate $a_n^{\text{des}}$ equal to 0.95. The first picture on the left displays the optimal cost as a function of the location of the extra inspection station. The other pictures on the right present the different values of the optimal buffer sizes $z_i$ and coefficient of availability of wip or inventory $a_i$ for the case of an optimally located internal inspection station ($\lambda_4 = 1$).

We notice a look almost similar to the case presented in Section 6.2.1 (for $a_i$ or $z_i$). The difference is noted at the optimal position of the inspection station (after $M_4$ in this case) and the $z_n(a_n^{\text{des}})$ necessary to meet the desired service-level constraint.

7. Added value of joint buffer optimization and inspection station positioning

In the following, we shall study the possible gains from a fully joint consideration of buffer sizing and inspection station positioning by comparing the results based on our current modeling and optimization methodology, with those obtained when following only partially joint numerical optimization approaches such as found in Schick, Gershwin, and Kim (2005) and Colledani and Tolio (2011). Indeed, in both of these works, an attempt is made at beating combinatorial complexity of the joint optimization problem at hand, by first picking a uniform buffer size allocation in the transfer line,
based on which a best positioning of inspection stations is sought in terms of overall productivity.

Thus, Schick et al. (2005) simulated an unreliable production line of 15 machines and 14 buffers where the machines are all identical as well as buffer sizes; they subsequently tested the performance of all possible combinations of location of inspection stations, i.e. $2^{14}$ possible cases (in all the cases considered, there is always an inspection station after the finished goods inventory). Following this study, Gershwin (2006) showed that for a production line with 15 machines (same model as in Schick et al. (2005)), the number of defective parts to be rejected from the system, if the inspection stations are poorly allocated, may be as much as 15% higher than that produced under a more adequate distribution of the same number of inspection stations.

Colledani and Tolio (2011) studied production lines in which the positioning of the inspection stations is arbitrary, and then analytically evaluated line performance accounting for statistical control charts parameters. However, when carrying out comparative studies, they fixed the buffering within the transfer lines to a constant.

Following the partial test approach of the above researchers, we will consider a homogeneous production line of 10 machines (same parameters as in Section 6.2) where buffers are identically sized and subsequently determine the best and the worst location for a single internal inspection station. For this example, and a uniform buffer size of 5, the results are as follows for the combined storage-shortage cost-minimization problem:

- The worst inspection station location is when $\lambda_1 = 1$ i.e. after the first machine
- The best inspection station location is when $\lambda_9 = 1$ i.e. before the last machine

Figure 5. The optimal stock sizes.
In the next step, and by choosing as positions of inspection stations those identified in the previous step (i.e. the best and worst location), we calculate the optimal buffer sizes and optimal cost.

Figure 5 presents the optimal buffering size distribution, respectively, associated with (i) the case where the inspection station is located after the first machine and (ii) before the last machine. The worst location associated optimal cost is 32.12 and the best location associated optimal cost is 27.26. When comparing these two optimal costs (32.12, 27.26) with that obtained by our method (24.31), we notice that in the partial approaches of Schick et al. (2005) and Colledani and Tolio (2011), the cost differential between best and worst is about 15% while based on our fully joint buffering and inspection station positioning methodology, the gain over the worst case configuration is on the order of 25% (based on the results in Subsection 6.2). This is an indication of the kind of improvement margins that may be achievable when buffer sizing and inspection station positioning are considered as a single global optimization problem.

8. Further numerical results of interest

8.1. Influence of inspection costs per part on the optimal positioning/necessity of the internal inspection station

In this section, we will show the effect that an increase in inspection cost $c_I$ has on the optimal positioning of the inspection stations. Thus, keeping $c_p$ and $c_n$ costs parameters constant, we vary the $c_I$ cost and note every time the optimal internal inspection station position for the homogeneous line described in Subsection 6.2.1.

Figure 6 shows that the more the higher inspection cost, further off center and downstream the optimal inspection station position is. An intuitive explanation of this phenomenon may be as follows: machines up to the first inspection station have to work at a rate sufficient to make up for all the non-conforming parts in the system, both as produced by themselves up to the inspection station in question and as produced in the

Figure 6. Influence of $c_I$ on optimal inspection station position.
rest of the line; this is unlike the machines past the inspection station which have to work at a common reduced rate. This means that, although the internal inspection station works as hard no matter where it is located (and thus accrues the same inspection costs per unit time) except that the more upstream it is located, the less parts it will reject, and thus the harder the second inspection station at the end of line will have to work. On the other hand, early inspection will remove defective parts sooner, thus reducing internal storage costs and gaining productive machine time. However, as inspection costs rise, the additional inspection costs at the end inspection station become the dominant factor and thus it pays to push the first inspection station further downstream so as to achieve overall inspection costs reduction. In effect, this suggests that as inspection costs increase sufficiently, it may be overall more economical to keep only a single inspection station at the end of the line: Figure 7 shows that for $c_I$ above a certain threshold, a line without an internal inspection station is more economical than a line that has one.

8.2. Assessing the influence of an individual poor quality machine

Figure 8 displays the optimal position of the inspection station for the homogeneous line of Subsection 6.2.1, when the $\beta_j$ coefficient of a particular machine $M_j$ is increased while all other $\beta_i$’s $i \neq j$ remain identical for $j = 1\ldots10$. For example, for $\beta_j = 0.15$, notice that, relative to the unperturbed case, the optimal inspection station position moves from $\lambda_4 = 1$ to $\lambda_5 = 1$, as long as the poor quality machine is located at position 5 or higher. The optimal inspection station position reverts back to 4 if the poor quality machine is placed at positions 1–4. As one further reduces the quality of the machine to $\beta_j = 0.2$, the optimal inspection station position is further pulled down to 3 as long as

Figure 7. Relationship between $c_I$ and the need for an internal inspection station.
Note: We conducted a similar analysis for $c_n$ and $c_p$, and we noticed no change in the optimal position of the internal inspection station (center of the line).
the index of the poor machine is 3 or less; it moves and stays at 4 if the poor machine index is 4 or greater. Finally, a further decrease in quality to $\beta_j = 0.25$ induces a motion of the optimal inspection station position at 2, if the poor quality machine index is 1 or 2, while it moves permanently to 3 once the poor quality machine index is 3 or higher.

In conclusion, the introduction of a particularly poor single machine in an otherwise homogeneous line appears to favor configurations with earlier inspections, irrespective of the position of that machine; on the other hand, this influence appears to be most significant when the perturbed machine position is to the left of the optimal inspection station position that would prevail in the unperturbed line. This point to the fundamental asymmetry, or directionality, of quality effects in a transfer line.

9. Conclusions and future work

This paper develops an approximate optimization formalism when joint decisions of buffer sizing and inspection station placement have to be made in unreliable transfer lines with imperfect machines producing as part of their normal operation both conforming and nonconforming parts. In effect, the focus has been on nonassignable or random quality failures, with scrapping of defective parts upon detection by an inspection station.

It is assumed that the proportion between parts correctly machined and those improperly machined remains constant for a given machine. In addition, the line includes two inspection stations; the location of one station is fixed (dedicated to the inspection of finished parts), while the location of the other station is chosen so as optimize the total per unit time average cost (storage cost, possibly shortage, and inspection cost). For a constant rate of demand for finished conforming parts, two versions of
the cost-minimization problem have been considered: one in which backlog is allowed at a cost and the other one where delays in delivery are tolerated as long as their probability of occurrence is less than some number (service-level constraint). A dynamic programming-based optimization framework has been presented. The importance of the joint consideration of buffer sizing and inspection station positioning is confirmed. Also, inclusion of inspection costs appears to be a useful device for excluding configurations with an excessive number of inspection stations.

In future work, based on our quality aware production model, we shall consider the complex combinatorics of multiple inspection stations in transfer lines, and develop for long homogeneous transfer lines criteria for specifying the optimal frequency of inspection stations. In addition, we believe that the rich modeling framework of Colledani and Tolio (2005, 2006, 2009, 2011) and Colledani (2011) could be eventually integrated into our current setup to produce an approximation framework whereby quantity/quality design decisions in the presence of statistical process control and scrapping of detected defective parts, which could be jointly optimized via dynamic programming.

References
Bielecki, T., & Kumar, P. R. (1988). Optimality of zero inventory policies for unreliable manufacturing systems. *Operations Research, 36*, 532–541.

Colledani, M. (2011, August). Joint design of quality and production control in multi-stage asynchronous manufacturing systems. *18th IFAC World Congress, 18*, 10408–10415.

Colledani, M., & Tolio, T. (2005). Impact of statistical process control (SPC) on the performance of production systems-Part 2 (large systems). The international conference on analysis of manufacturing systems-production management, Zakynthos Island, Greece.

Colledani, M., & Tolio, T. (2006). Impact of quality control on production system performance. *Annals of the CIRP, 55*, 453–456.

Colledani, M., & Tolio, T. (2009). Performance evaluation of production systems monitored by statistical process control and off-line inspections. *International Journal of Production Economics, 120*, 348–367.

Colledani, M., & Tolio, T. (2011). Integrated analysis of quality and production logistics performance in manufacturing lines. *International Journal of Production Research, 49*, 485–518.

Gershwin, S. B. (2006). How do quantity and quality really interact? Precise models Instead of strong opinions. *Information Control Problems in Manufacturing, 1*, 33–40.

Hu, J. Q. (1995). Production rate control for failure-prone production systems with no backlog permitted. *IEEE Transactions on Automatic Control, 40*, 291–295.

Inman, R., Blumenfeld, D. E., Huang, N., Li, J., & Li, J. (2013). Survey of recent advances on the interface between production system design and quality. *IEEE Transactions, 45*, 557–574.

Kim, J., & Gershwin, S. B. (2005). Integrated quality and quantity modeling of a production line. *Spectrum, 27*, 287–314.

Kim, J., & Gershwin, S. B. (2008). Analysis of long flow lines with quality and operational failures. *IEE Transactions, 40*, 284–296.

Meerkov, S. M., & Zhang, L. (2010). Product quality inspection in Bernoulli lines: Analysis, bottlenecks, and design. *International Journal of Production Research, 48*, 4745–4766.

Meerkov, S. M., & Zhang, L. (2011). Bernoulli production lines with quality-quantity coupling machines: Monotonicity properties and bottlenecks. *Annals of Operations Research, 182*, 119–131.

Mhada, F., Hajji, A., Malhamé, R., Gharbi, A., & Pellerin, R. (2011). Production control of unreliable manufacturing systems producing defective items. *Journal of Quality in Maintenance Engineering, 17*, 238–253.

Mhada, F., Malhamé, R., & Pellerin, R. (2013). A stochastic hybrid state model for optimizing hedging policies in manufacturing systems with randomly occurring defects. *Discrete Event Dynamic Systems: Theory and Applications*. Retrieved from [http://dx.doi.org/10.1007/s10626-013-0160-8](http://dx.doi.org/10.1007/s10626-013-0160-8), in press.

Montgomery, D. C. (1991). *Introduction to statistical quality control*. New York, NY: John Wiley.
Sadr, J., & Malhamé, R. (2004). Unreliable transfer lines: Decomposition/aggregation and optimization. *Annals of Operations Research, 125*, 167–190.

Schick, I. C., Gershwin, S. B., & Kim, J. (2005) *New results on the integrated analysis of quality and quantity in production lines*. Proceedings of the 5th international conference on Analysis of Manufacturing Systems-Production Management, Zakynthos Island, Greece.