



Integrated analysis of quality and production logistics performance in manufacturing lines

Marcello Colledani, Tullio Tolio

► To cite this version:

Marcello Colledani, Tullio Tolio. Integrated analysis of quality and production logistics performance in manufacturing lines. International Journal of Production Research, 2010, pp.1. 10.1080/00207540903443246 . hal-00564786

HAL Id: hal-00564786

<https://hal.science/hal-00564786v1>

Submitted on 10 Feb 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Integrated analysis of quality and production logistics performance in manufacturing lines

Journal:	<i>International Journal of Production Research</i>
Manuscript ID:	TPRS-2008-IJPR-0941.R2
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	30-Sep-2009
Complete List of Authors:	Colledani, Marcello; Politecnico di Milano, Mechanical Engineering Tolio, Tullio; Politecnico di Milano, Mechanical Engineering
Keywords:	PERFORMANCE ANALYSIS, BUFFER SIZING
Keywords (user):	Decomposition Method, Statistical Quality Control



RESEARCH ARTICLE

Integrated analysis of quality and production logistics performance in manufacturing lines

M. Colledani[†], T. Tolio[†]

(Received 00 Month 200x; final version received 00 Month 200x)

This paper presents an analytical method for evaluating the performance of production systems jointly considering in a unique framework quality and production logistics performance measures. In the literature, quality and production logistics have been traditionally investigated with independent approaches. So far, only few recent contributions have focused on their mutual inter-relation, showing that relevant benefits can be achieved from an integrated analysis. This paper contributes to this growing research area by considering, for the first time, production systems in which the behaviour of the machines is monitored by statistical control charts. The control action performed by the quality control system on the machines is directly modeled. Moreover, the impact of the quality control action on the logistic flow of parts in the system is taken into account. In order to estimate the performance of such systems, an approximate analytical method based on the system decomposition technique is developed. The accuracy of the method is assessed by comparing the results with those obtained by simulation. The analysis of the results provides new insights on the relations among the quality and production logistics performance measures and paves the way to the development of integrated design techniques for improved configurations of the system as a whole.

Keywords: Performance Evaluation, Decomposition Method, Statistical Quality Control, Buffer Sizing.

1. Introduction

In the last three decades, the increasing competitiveness of the global market has caused an ever increasing pressure on both quality of products and productivity of the systems producing the products. The progress in technology has provided several possibilities for production managers to perform a better control of production plants performance, both from the point of view of quality and production logistics. Firstly, advances in sensors technology has provided the possibility of rapidly inspect several product characteristics in short time, with high accuracy and on-line [Blais (2003)]. Secondly, production control systems and machine data collection systems have found wide applications in real production plants, for monitoring the states in which the machines produce parts. The results of this technological revolution for manufacturing are that a huge amount of data is day by day collected by production and quality managers, with the aim of having reliable feedbacks from the shop floor.

In spite of these market needs and these technological opportunities, quality and productivity have been traditionally considered by researchers and practitioners as two very separate fields. Considering production logistics performance measures, the Manufacturing Systems Engineering [Gershwin (1994)] research area has been developed for understanding the behaviour of production system and for designing

[†]Politecnico di Milano, Department of Mechanical Engineering - Manufacturing and Production Systems Division. Via La Masa 1, 20156 Milano, Italy.

efficient factories. This research has been focused on productivity-related issues like estimating the production rate and the work in progress of systems and providing a good configuration of buffers. On the other hand, Statistical Process Control [Montgomery (2005)], Total Quality Management [Prajogo and Sohal (2001)] and Six Sigma [Pande and Holpp (2002)] theories have been developed for a better control of manufacturing processes, for meeting higher product quality and for continuous improvement of processes. The link between these two areas have been very rarely considered at a production system level, even if industrial experience has evidenced the need for jointly considering quality and productivity performance measures while designing the manufacturing system [Inman *et al.* (2003)].

There are many aspects that prove that quality and production logistics are mutually related. For instance, the production system architecture affects the performance of the quality control system. It has been shown by Gershwin [Gershwin (2006)] that for a production line with 15 machines the number of bad parts to be scrapped by the system if inspection stations are poorly allocated can be the 15% higher than the number of bad parts produced with a good allocation of the same number of inspection stations.

Moreover, the results coming from researches carried out in Lean Production area [Shingo and Dillon (1989)], [Toyota (1996)] have shown that the reduction of inventory has a positive impact on product quality, since quality problems are identified earlier. However, from the manufacturing system engineering research area, it is known that the production rate of the system is positively affected by the presence of buffers, since they decouple the behaviour of the machines, preventing from the propagation of machine disruptions upstream and downstream the line [Gershwin and Schor (2000)]. Therefore, an interesting trade-off is generated that can be suitably exploited by jointly considering the impact of the inventory configuration on quality and production logistics performance.

In machining and assembly operations, it has been shown that the operating speed is inversely related to the product quality [Blumenfeld and Owen (2008)]. Thus, improving the machine processing rate has a positive impact on the system production rate but may negatively affect the system yield. Finally, frequent machine set-ups among different part types and small production runs may be disruptive for product quality and prevent from process learning through data analysis [Inman *et al.* (2003)]. However, with small batches the quality information feedback can be propagated with short delays, thus enabling a more reactive control of the system.

Not only the system configuration has an impact on the quality control system performance, but also decisions concerning the quality control system configuration have an impact on the system productivity performance. For instance, the corrective actions on the system decided as a consequence of signals generated by the quality control system have an impact on the productivity performance of the manufacturing line. Indeed, stopping a machine for checking its conditions after a signal means interrupting the production of the machine, thus affecting the production rate of the whole system. By considering the quality feedback only at a single resource level in isolation, this global effect cannot be properly estimated. If the machine that has quality problems is the productivity bottleneck of the system, the stop for fixing the problem will have an impact on the production logistics performance of almost the entire line. Again, this trade-off can be only suitably addressed by considering the quality control problem at a system level.

These considerations strongly motivate the need for the proposed research activity and give an idea of its potential impact in terms of knowledge on production systems behaviour as well as in terms of cost reduction for companies. Results in

this area will pave the way to the development of algorithms to jointly address the problems of finding the optimal allocation of inspection devices and buffer storages, and the optimal design of control charts, by performing a complete analysis of the effects of each decision on the overall system performance.

The remaining part of the paper is structured as follows: in the next section, the main contributions in the integrated analysis of quality and productivity performance measures are reported. In Section 3, the features of the studied system are analyzed and the integrated framework is introduced. In Section 4 the approximate analytical method is presented in details. Finally, in Section 5 numerical results are reported, that validate the proposed approach, and the system behaviour is investigated to derive new insights on this topic.

2. Literature Review

Even if the aspects concerning the design of the quality control system and the design of the production system architecture are strongly interrelated, quality and production logistics have almost always been considered in the literature as two separate research areas. However, a growing area of research has focused on this inter-relation, progressively capturing its more interesting aspects. The aim of this section is to provide a detailed analysis of the state of the art in this topic, starting from preliminary studies until more consolidated results. For each work, the main contributions to the problem are highlighted. This section also points out the relevant steps beyond the state of the art of the approach proposed in this paper.

The simulation works proposed in [Bulgak (1992)] and [Li (2005)] address the problem of studying how different system architectures impact on the performance of the quality control system. The main result of these works is the assessment of a fundamental principle: while changing the system architecture both in terms of buffer sizes [Bulgak (1992)] and production strategies [Li (2005)] the fraction of conforming products released in output may drastically change. In other words, there is an impact of the system architecture on the quality of the output production. Unfortunately, the analysis is almost unidirectional, since the impact of the quality control system on the system performance is not investigated. A similar intention is expressed in [Hu *et al.* (1998)] in which different system configurations are compared according to different performance measures, including the capability of the system to produce parts with small variations and the expected productivity. In this contribution, the different layout is intended as different sequence of execution of process operations. Six alternative configurations are considered, having serial lines and parallel lines as extremes. Buffers are not considered in the study. The authors showed that, considering quality as dimensional variation of products, serial lines perform better because there is only one possible path of the part flow and no mixing effect exists. Later [Cheng *et al.* (2000)] showed that U-shaped lines may perform better than serial lines in terms of quality of released output. Even if they provide more qualitative than quantitative analysis of the problem, these contributions highlight a common effort in trying to synthesize general paradigms on the correlation among the production system and the quality of the produced parts. In this sense, their contribution is aligned to the point of view of this paper.

Several contributions in this area are proposed by Tapiero and Hsu. They consider the problem of jointly evaluating quality and productivity in single server queues. They analyzed a M/M/1 queue in which quality monitoring is performed [Hsu and Tapiero (1987)], an M/G/1 queue, both from the quality control [Hsu and Tapiero (1989)], the maintenance point of view [Hsu and Tapiero (1988)] and considering an integrated approach for process monitoring, quality control and maintenance [Hsu

and Tapiero (1992)]. This set of works aims at showing with quantitative models the relation among quality and productivity performance measures in production lines. However, the presence of finite capacity buffers in the analyzed systems is not considered, thus the impact of the system on the quality information feedback is not estimated.

The quality problem is instead addressed at system level in very recent contributions. Gershwin and Kim [Gershwin and Kim (2005)] developed a method for evaluating the effective, the total throughput and the yield of a production line, considering machines subject to "quality failures". When in normal operating conditions the machine does not produce any defective item; after a transition to the quality failure state occurs, the machine produces only defective products. The quality control action is needed for stopping the machine and activate the repairing intervention that enables the machine to go back to the perfect quality state. No possibility of false alarms of the controller is considered. Machines are modeled as Markovian and the quality control action is modeled as a transition that forces the machine to go from the low quality state to an un-operational state for the repairing process. This transition is considered to be fixed and is taken in input, therefore no link to quality control practices is assumed. The presence of buffers with finite capacity is considered and their effect on the quality information delay is investigated. The main result consists in showing a particular behaviour of the throughput of good products as a function of the buffer capacity. Indeed, due to the delay of the quality feedback, generated by the presence of buffers in the case of "ubiquitous inspection", the throughput may also decrease while the buffer capacity increases. In this paper, we will derive a similar result also for the case of systems monitored by statistical control charts. However, due to the fact that the quality control action is directly connected to the control chart parameters, it will be possible to provide an explanation to the phenomenon and to determine the cases in which this phenomenon is visible, also considering the parameters of the quality control system. The authors extended the approach in [Gershwin and Kim (2008)] to longer lines. Furthermore, in [Gershwin and Schick (2007)] a taxonomy for the integrated analysis of quality and productivity performance is proposed. Indeed, dealing with a new area of research which integrates aspects coming from two established areas, it becomes crucial to share and consolidate the terminology, the notation and the basic concepts. Recent contributions are proposed in [Li *et al.* (2007a)] and [Li *et al.* (2007b)] in which the problem of designing specific stations of the system for quality purposes is addressed. Specifically, an application to the painting shop of an automotive plant is considered, in which the developed method is used for increasing the "Quality Buy Rate", i.e., the good job ratio of all jobs. Furthermore, in [Bassetto and Siadat (2009)] the application of a methodology for updating the manufacturing control plan by integrating product, process and tool data within a system level perspective is proposed, related to semiconductor manufacturing. In [Liberopoulos *et al.* (2007)] the production rate of asynchronous unbuffered production lines in which machines are subject to failures are considered. Products under processing are deteriorating items, in the sense that their quality characteristics deteriorate over time. Therefore, if the failure of a machine is long enough, the material under processing in the upstream machines must be scrapped by the system. In this paper, the relation among the system parameters and the quality of the released parts is taken into consideration. However, buffers are not considered in the system analysis. It is worth to make reference to a recent contribution coming from the Hitachi Production Engineering Research Laboratory (PERL) [Nonaka *et al.* (2008)]. The authors analyzed the correlation between the occurrence of machine failures and the quality problems detected on the pro-

duced parts. Following this analysis, the authors came to the conclusion that this correlation is the main cause for quality problems in the analyzed semiconductor manufacturing fab and used this result for bottleneck identification. Although mathematical models for estimating the quality-quantity relation are not provided in this paper, it gives a clear idea of the potential industrial benefits of methodologies addressing this problem.

This paper presents a new approximate analytical method for the integrated analysis of quality and productivity performance of manufacturing lines, where statistical control charts are used to monitor the behaviour of machines in the system. The objective of this work is to model the interactions among quality and productivity parameters and to estimate the effective and the total production rate, the yield and the work in progress of the system. For reviews on early decomposition based approximate analytical methods see [Buzacott and Hanifin (1978)] and [Dallery and Gershwin (1992)]. The decomposition method [Gershwin (1994)] has been developed in the system engineering area with the objective of evaluating the performance of a wide set of manufacturing systems, featuring finite capacity buffers and unreliable machines. Early models dealt with the analysis of serial transfer lines [Gershwin (1987)] and asynchronous flow lines [Burman (1995)], with single failure mode machines. Later, these methods have been extended to deal with more complex machine models, featuring multiple failure modes [Levantesi *et al.* (2003)]. Furthermore, they have been extended to deal with more complex system architectures. Specifically, assembly systems [Gershwin and Burman (2000)], closed loop systems [Gershwin and Werner (2003)], multiple product systems [Colledani *et al.* (2005)] and systems featuring split and merge operations [Helber (1999)] have been analyzed with decomposition based methods. Recently, methods for studying generally complex system layouts [Li (2003)], [Colledani and Tolio (2005)] have been proposed. Approximate analytical methods represent a suitable alternative to simulation when estimating the performance of manufacturing systems. They are accurate in the estimation of the main performance measures. They are extremely fast, allowing to evaluate a large set of system alternatives and parameter settings, in short time. Moreover, they provide parametric equations modeling the dynamics of the material flow in the system. This enables to perform analysis on the propagation of starvation and blocking causes through the line and quality/productivity bottlenecks identification.

In this paper, for the first time, the quality control mechanism is explicitly modeled within the manufacturing system model, creating a complete link among the two interacting systems. It will be shown that this feature allows to understand more deeply the nature of the interaction and to address possible actions to better manage this very common type of production systems. The quality control tools that are modeled in this paper are statistical control charts. They have attracted tremendous attention by researchers and practitioners in the last 40 years. A recent discussion on the application of control charts in the manufacturing sector can be found in [Schipper (1998)]. In this paper, we will consider in details that the monitoring action is performed by Xbar control charts for monitoring product variables. However, the method is easily amenable to extension and application to different types of control charts, such as R, S, EWMA and CUSUM [Montgomery (2005)], [Alwan (2000)].

3. Modeled System

In this section, the characteristics of the considered types of production systems are analyzed and the main assumptions of the system model are detailed. Furthermore,

a new notation is proposed. It is inspired by the two areas of research of interest, but merges both features into a coherent framework. Finally, the main performance measures of interest, which are the output of the method, are presented.

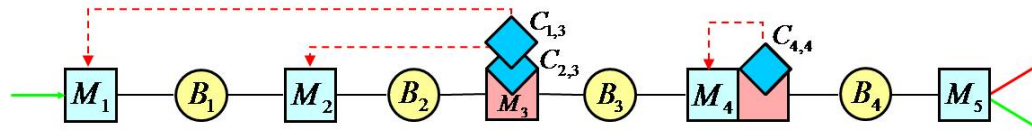


Figure 1. Example of the considered production system layout.

3.1 System Behaviour

The production line monitored by SPC under consideration is represented in Figure 1. The system has a linear layout and is composed of K stations, represented as squares, and $(K - 1)$ buffers, represented as circles. Buffers are frequently present in real production lines, with the function of decoupling the machines. They can be automatic conveyors, AS/RSs, floor space, etc. The considered stations can be *machining stations*, *inspection stations* or *integrated stations*. Machining stations are those realizing machining operations on the parts flowing in the system. Inspection stations are those measuring some quality characteristics of the parts produced at one or more upstream machining stations. Integrated stations are those performing both manufacturing and inspection operations. For instance, in Figure 1, M_1 , M_2 and M_5 are machining stations, M_3 is an inspection station that measures quality characteristics of the parts already processed by the stations M_1 and M_2 and M_4 is an integrated station measuring the quality features of the parts processed at the same stage.

Machining stations take one part from the input buffer, perform a process, and provide in output a finished or semi-finished product. However, the quality of products resulting from the transformation depends on the state of the machine that processed the part. According to the SPC theory, the machining stations can produce being either in control or in the out of control state. The machine is said to process parts in control if each processed quality characteristic is in a statistical control state. If a change in the quality characteristic distribution takes place, the machine is said to be out of control. The in control state is normally characterized by a low fraction of non-conforming parts produced, while the out of control state is normally characterized by a higher fraction of non-conforming parts produced. In real systems, assignable causes for out of controls may be the loss of process settings, the tool wear, the effect of the machine vibrations caused by structural problem occurrence, etc.

In general, several modes of being out of control can be observed for a machining station, each one characterized by its own fraction of non-conformities associated. In this paper, for simplicity, we consider that a unique way of being out of control is possible for each machine. This is a very commonly investigated case in the SPC research area. However, the proposed approach can be easily extended to consider multiple out of control states. Moreover, we consider the case in which product features processed at different production stages are independent. In other words, an out of control of machining station M_i does not affect the quality of the operations performed at the downstream stages M_j , with $j = i + 1, \dots, K$, i.e., the machines add non-conformities of different types to the processed parts. This case is typical of many automated manufacturing systems in which material cutting processes are applied.

Machining stations are also unreliable and subject to *operational failures*. Operational failures are typically random events that instantaneously cause the stop in the production of the machine. They affect the quantity of parts processed by the machine by reducing the fraction of time it is operational, without directly affecting the quality of the produced parts. In real systems, they are typically fuse damages, tool breakages, wrong part positioning in the working area and mechanical jamming. The estimation of the parameters concerning these types of failures is nowadays facilitated, in real systems, by the presence of production monitoring systems. Therefore, machine reliability parameters can be suitably estimated by considering the data collected from the workshop. Multiple modes of operational failures are common in real machines. As an example, the engine blocks machining line recently studied by the authors of this paper [Colledani *et al.* (2009)] was characterized by a total number of failure modes equal to 144, distributed on 22 machines; this means, on average, 7 modes of failures were observed per machine. Even if the operational failures and out of controls are very different random events, that affect respectively the quantity and the quality of the machine output, they may also happen simultaneously. In this case, the repairing intervention due to the operational failure will not reset the machine to the in control condition, since the two causes of problems are of different nature.

Since out of control events are partially observable, control charts have been developed in the SPC theory [Montgomery (2005)], in order to detect them. Control charts are logical devices that perform statistical hypothesis tests on data measured on the produced parts or collected directly from the process. In Figure 1, the control charts are represented as rhombus and named $C_{i,q}$, where i refers to the machining station M_i , that processed the monitored feature, and q is associated to the inspection station M_q that measures the product feature to which the control chart is associated. If $i = q$ then the machining station M_i is said to be *locally monitored* by the control chart $C_{i,i}$, since the inspection of produced parts is carried out just downstream the machining station. Otherwise, if $i < q$, then the machining stations is said to be *remotely monitored* by the control chart $C_{i,q}$, since in this case $(q - i - 1)$ machines and $(q - i)$ buffers are located between the monitored machining station and the inspection device. Remote monitoring is frequently used in manufacturing lines when more then one product feature can be inspected by the same inspection device. Modern inspection systems, indeed, are endowed with proper levels of flexibility and reconfigurability to be able to rapidly inspect many different product characteristics at the same time [Barhak *et al.* (2005)]. Although this practice may be effective in several production lines, in this paper it will be shown that it may negatively impact the overall system performance, due to the delay in the quality feedback it involves.

Basically, a control chart carries out repeated tests of hypothesis on the data coming from the inspected parts. The test is based on two competing hypotheses, namely H_0 , the monitored machine is in control and H_1 , the monitored machine is out of control. This statistical test is subject to two types of errors, named type I error (a false alarm is issued) and type II error (a real out of control is not detected). The first error happens with probability α and the type II error happens with probability β . When, on the basis of the collected data, the control chart identifies an out of control, a signal is sent to the machine that produced the feature monitored by the control chart; the machine receiving the signal is stopped to investigate and fix the problems that kept it out of control.

In order to provide data to be processed by the control charts, inspection plans must be designed. One can design the quality control system to measure all the produced parts, in this case a 100% inspection is performed, or to measure only a

fraction of the produced parts, in this case sampling inspection is adopted. Data collected by the inspection stations are normally used also to decide whether the inspected parts can be considered as conforming or non-conforming. Actions which follow this evaluation generally involve scrapping or reworking of defects. The method allows to model scrap even if, to simplify the analysis, in this paper it will be considered that defective items are scrapped or reworked off-line, downstream the last machine M_K . On-line scrap and rework practices will be investigated in future research activities.

3.2 Assumptions and notation

A detailed list of method assumptions follows, highlighting both the quality and the productivity aspects. The notation is summarized in Table 1.

Discrete Production Line Model. The dynamics of the material flow in the system is modeled through a discrete synchronous line model. A discrete flow of material from outside is supposed to enter the system at the first station M_1 , then moves to the first buffer B_1 , visits the other machines and buffers in sequence until it reaches the last machine M_k , and leaves the system. An infinite supply of raw parts is considered to be available upstream the first machine of the system M_1 . In this paper, we don't consider the presence of defects in raw parts, even if the proposed approach can be extended in order to include the use of acceptance plans. Similarly, we assume that the last machine of the system M_K is never blocked. In discrete models, each machine requires a fixed time to process a part and the number of products stored in each buffer, at time instant t , is an integer number. Stations are characterized by the same deterministic processing time, scaled to time unit. Therefore, at time instant t , each operational machine starts working material simultaneously, if neither starved or blocked. At the end of the time unit t , the buffer levels are changed accordingly. This material flow mechanism is known in literature as *Blocking Before Service* [Dallery and Gershwin (1992)].

Unreliable Machines. The stochastic behaviour of operational failures is modeled by considering geometrically distributed times to failure (TTF) and times to repair (TTR). For the machine M_i , the probabilities of failure and repair in each time unit are respectively $p_{i,f_i} = 1/MTTF_{i,f_i}$ and $r_{i,f_i} = 1/MTTR_{i,f_i}$. The failure modes $f_i = 1, \dots, F_i$ are Operation Dependent (ODF), i.e., the machine can fail in a time unit only if it is operational and neither starved nor blocked. The total number of failure modes of the machine M_i is F_i . The different modes of failures are considered to be mutually exclusive, in the sense that, while being failed in a certain mode, the machine cannot fail in another mode before being repaired. All the previous assumptions are common in multiple failure mode reliability models [Gershwin *et al.* (2002)].

Out of Control State. For the machining station M_i , the transition to the out of control state is assumed to happen with probability $p_i^{quality} = 1/MTOC_i$, with time to out of control (TOC_i) geometrically distributed. The in control conditions are reset with probability $r_i^{quality}$. Not all the machining stations in the system are necessarily subject to out of control. Specifically, the binary indicator variable I_i will assume value 1 if M_i is subject to out of control, 0 otherwise. Also, since samples are taken on the basis of the number of parts processed by the monitored machine, alarms are Operation Dependent and each alarm immediately stops the machine M_i .

Fraction of non conforming parts produced. According to the specification limits, LSL (Lower Specification Limit) and USL (Upper Specification Limit), designed for meeting the product functionality, the fractions of non-conforming parts pro-

duced are γ_i^W and γ_i^O , respectively when the process performed by station M_i is in control (W) and out of control (O).

Control charts. The parameters related to the control chart $C_{i,q}$ are the sample size $m(C_{i,q})$, the number of not measured parts between samples $h(C_{i,q})$, the probability of type I error $\alpha(C_{i,q})$, and the probability of type II error $\beta(C_{i,q})$. These two probabilities are related to the position of the control limits $LCL(C_{i,q})$, (Lower Control Limit) and $UCL(C_{i,q})$ (Upper Control Limit) in the control chart. The performance of the chart in terms of $ARL_0(C_{i,q})$, i.e., the average number of samples to be collected before a false alarm is generated, and $ARL_1(C_{i,q})$, i.e., the average number of samples to be collected before a real out of control is correctly detected, can be estimated as a function of the error probabilities:

$$ARL_0(C_{i,q}) = \frac{1}{\alpha(C_{i,q})} \quad ARL_1(C_{i,q}) = \frac{1}{1 - \beta(C_{i,q})} \quad (1)$$

An example of Xbar control chart is reported in Figure 2. The parameters introduced in this last paragraph are all standard assumptions in Statistical Quality Control theory. Two further comments are required. Firstly, as it can be noticed, the adopted notation is slightly different from the well known SPC notation. However, the two areas that this paper aims at integrating have some overlap in the respective notation. Therefore, a choice has been made about the adopted notation that keeps the content of this manuscript coherent and the parameters definition unique. Secondly, although the SPC theory is widely applied in industry, the problem of estimating in practice the type I and type II error probabilities is still under discussion [Woodall (2000)]. However, our method considers the connection among the production system and the quality control system involving only the sampling parameters, $m(C_{i,q})$ and $h(C_{i,q})$, and the reactivity of the chart, in terms of $ARL_0(C_{i,q})$ and $ARL_1(C_{i,q})$. The first parameters are typically strategic design decisions and the second parameters are local control chart performance measures. Therefore, whatever method is adopted to estimate unknown control chart parameters, as long as these parameters are known, the proposed integrated method can be applied.

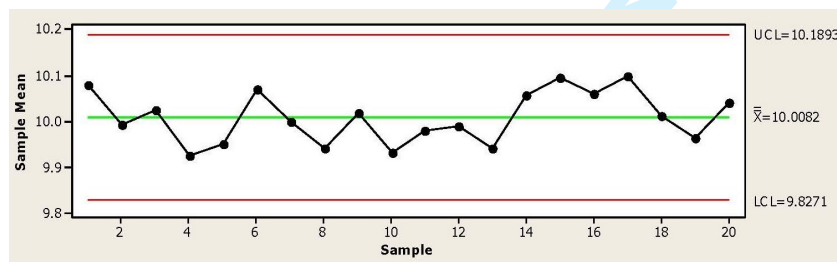


Figure 2. Example of a Xbar control chart monitoring the diameter of ball studs. $LCL(C_{i,q}) = 9.8271mm$, $UCL(C_{i,q}) = 10.1893mm$, $m(C_{i,q}) = 5$, $h(C_{i,q}) = 20$, $ARL_0(C_{i,q}) = 370.3$ and $ARL_1(C_{i,q}) = 2.85$ with a shift of 0.15mm.

False Alarms. The probability that the machine M_i is restarted in a time unit, after a false alarm has occurred, is r_i^{false} . $TTRFA_i$ (Time to Repair False Alarm) are geometrically distributed with mean $1/r_i^{false}$.

Buffer Capacity. The finite capacity of buffer B_i is N_i , which is a positive integer number.

3.3 Performance Measures

The problem addressed in this paper can be formalized as follows: given the system described in Section 3.1 and the modeling assumptions provided in Section 3.2, evaluate the following system level performance measures:

- E^{Tot} : the average total production rate (throughput) of the system, including both conforming and non-conforming parts;
- E^{Eff} : the average effective production rate of the system, including only conforming parts;
- Y^{System} : the system yield, that is the fraction of conforming parts produced in output by the system;
- \bar{n}_i : the average level of buffer B_i , for $i = 1, \dots, K$;
- WIP : the average work in progress in the system.

It is possible to define some structural dependencies among the presented system level performance measures. In particular, the following equation links the effective production rate and the total production rate, to the system yield:

$$Y^{System} = \frac{E^{Eff}}{E^{Tot}} \quad (2)$$

In addition, the *Conservation of Flow Properties* [Gershwin (1994)] can be applied in this type of systems only to the average total production rate E^{Tot} . On the contrary, the average effective production rate E^{Eff} is affected by the quality of the parts processed at each stage, thus it will be a non-increasing function of the system stages.

Table 1. Summary of the Notation.

K	number of machines in the system
M_i	i^{th} machine of the system
B_i	buffer situated downstream M_i
N_i	capacity of B_i
F_i	total number of operational failures for M_i
$p_{i,j}$	failure probability of machine M_i in mode j
$r_{i,j}$	repair probability of machine M_i in mode j
$p_i^{quality}$	probability of out control of M_i
$r_i^{quality}$	probability of repairing the out control of M_i
I_i	binary variable assuming value 1 if M_i is subject to out of control, 0 otherwise.
$C_{i,q}$	control chart located in M_q monitoring M_i
$m(C_{i,q})$	sample size, i.e., number of measured parts in a sample
$h(C_{i,q})$	number of not measured parts between two samples
$UCL(C_{i,q})$	Upper Control Limit of the control chart $C_{i,q}$
$LCL(C_{i,q})$	Lower Control Limit of the control chart $C_{i,q}$
$ARL_0(C_{i,q})$	average number of samples before the chart detects a false out of control
$ARL_1(C_{i,q})$	average number of samples before the chart detects a true out of control
$\alpha(C_{i,q})$	type I error, probability of a false alarm of out of control
$\beta(C_{i,q})$	type II error, probability of not detecting an existing out of control
r_i^{false}	probability that M_i is restarted after a false alarm has occurred
γ_i^W	fraction of non-conforming parts produced by M_i if it is in control
γ_i^O	fraction of non-conforming parts produced by M_i if it is out of control
E^{Tot}	total throughput
E^{Eff}	effective throughput
E^{Def}	defective throughput
Y^{System}	system yield
WIP	work in progress
\bar{n}_i	average number of parts in B_i

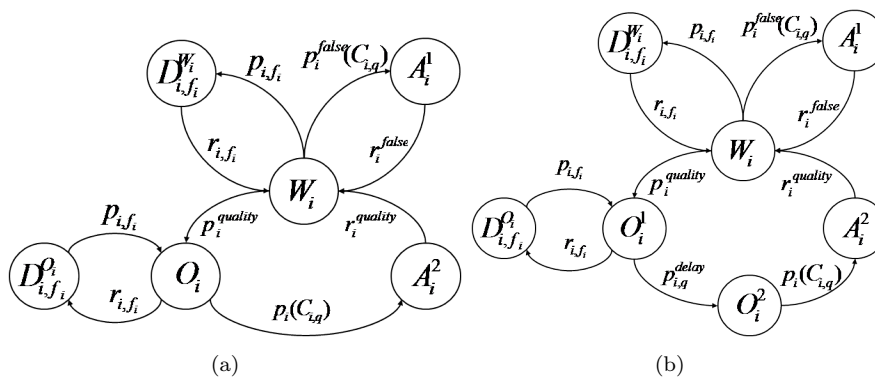


Figure 3. State transition diagrams of the generic machine M_i is isolation, both locally (a) and remotely (b) monitored.

4. The Solution Method

In this section, the decomposition method is presented to evaluate the performance presented in the previous section. The first system under analysis is the single machine monitored by quality control, considered in isolation. Both local and remote monitoring are considered, and the performance measures are derived. Then, by using these results, the method for studying the performance of general multi-stage lines is presented.

4.1 The Isolated Monitored Machine Case

The simplest production model that can be taken into account is the single machine model. A machine M_i is said to work in isolation mode if its behaviour is not influenced by the presence of the other buffers and machines in the line. In the following, the behaviour of manufacturing machines (or integrated machines) subject to out of controls and operational failures is investigated. Since the machine in isolation mode is considered, no information is available regarding the location of the inspection station that feeds the quality control chart monitoring its behaviour. However, it is assumed that the local performance of the quality control chart, in terms of $ARL_0(C_{i,q})$ and $ARL_1(C_{i,q})$, are known. The behaviour of the isolated machine M_i locally monitored by $C_{i,q}$, with $q = i$, can be modeled by the discrete-time discrete-state Markov chain with state transition diagram represented Figure 3(a). The following sequence of events represents the quality control loop, in the case of local monitoring. Suppose that the machine starts its production in the in control state (W_i). It can shift to the out of control state (O_i) with probability $p_i^{quality}$, in a time unit. Since the machine is locally monitored, as soon as it shifts to the out of control state, the inspection device measures parts which have been processed by the manufacturing machine while being out of control. Depending on the reaction time of the control chart in detecting real out of controls, related to the type II error probability of the statistical test, the out of control may or not be identified. The identification of the real out of control happens with unknown probability $p_i(C_{i,q})$, in a time unit. Therefore, the machine is stopped for being reset to the in control conditions, (A_i^2). The repairing intervention resets the machine to the in control condition with probability $r_i^{quality}$, in the time unit. The quality control feedback is only affected by the reactivity of the control chart in identifying out of controls. While being in the in control state, the machine can also be stopped for a false alarm of out of control, generated by the quality control chart.

The probability that this event occurs in a time unit is linked to the type I error probability of the control chart and is equal to $p_i^{false}(C_{i,q})$, still unknown. From this state (A_i^1), the in control conditions are restored with probability $r_i^{false}(C_{i,q})$. Furthermore, operational failures may occur both when producing in control and out of control. These failures immediately turn the machine down. From the down states, respectively ($D_{i,f_i}^{W_i}$) and ($D_{i,f_i}^{O_i}$) for $f_i = 1, \dots, F_i$, the machine is reset to the operational in control or out of control conditions, with probability r_{i,f_i} .

Some transition probabilities among these states are known, since they are system parameters and have been introduced in Section 3.2. The unknown probabilities are determined by using the following equations, that represent the link between the quality control chart $C_{i,q}$ and the monitored machine M_i . Therefore, they have been called *Quality Link Equations*. Consider the transition between state W_i and state A_i^1 ; it is the probability of having a stop for a false alarm, thus it is related to the probability of occurrence of the type I error. From the SPC theory, the ARL_0 is by definition the number of samples to be analyzed before a false alarm is generated. Since the number of parts to be processed by the machine before one sample is observed is equal to $h(C_{i,q}) + m(C_{i,q})$, the first quality link equation can be derived:

$$p_i^{false}(C_{i,q}) = \frac{1}{MTTFA(C_{i,q})} = \frac{1}{ARL_0(C_{i,q}) [h(C_{i,q}) + m(C_{i,q})]} \quad (3)$$

In this equation, an approximation is introduced since the false alarm probability is scaled to the time unit, even if false alarms are possible only when the sample is actually measured. The impact of this approximation on the performance measures will be investigated in Section 5.1.

Similarly, the transition probability from the state O_i to the state A_i^2 depends on the type II error. The more the chart is reactive in identifying out of controls, the higher is this transition probability. It can be evaluated by using the following equation:

$$p_i(C_{i,q}) = \frac{1}{MTTD(C_{i,q})} = \frac{1}{ARL_1(C_{i,q}) [h(C_{i,q}) + m(C_{i,q})]} \quad (4)$$

By using these equations and the input parameters of the model, all the transition probabilities are known, and the Markov chain can be solved. Once the steady-state probability $\pi(X)$ has been calculated, for the generic state X , the performance of the system can be derived. Having named b_i the coefficient:

$$b_i = \frac{p_i^{quality}}{p_i(C_{i,q})} \quad (5)$$

the total average production rate E_i^{TOT} of the isolated machine M_i is given by the sum of the steady-state probabilities of the operational states of M_i , independently from the quality of the parts processed in these states. Thus, E_i^{Tot} is given by:

$$E_i^{Tot} = \pi(W_i) + \pi(O_i) = \frac{1 + b_i}{\left(1 + \sum_{f_i=1}^{F_i} \frac{p_{i,f_i}}{r_{i,f_i}}\right) (b_i + 1) + \frac{p_i^{false}}{r_i^{false}} + b_i \frac{p_i(C_{i,q})}{r_i^{quality}}} \quad (6)$$

This can be considered as the efficiency in isolation mode e_i of the locally monitored machine M_i . The throughput of conforming parts, also named effective throughput E_i^{Eff} of the machine M_i , is obtained by considering the fraction of non-conformities

processed, respectively, in the in control and in the out of control states. It is given by:

$$E_i^{Eff} = \pi(W_i)(1 - \gamma_i^W) + \pi(O_i)(1 - \gamma_i^O) = \frac{(1 - \gamma_i^W) + (1 - \gamma_i^O)b_i}{1 + b_i} E_i^{Tot} \quad (7)$$

Thus the yield Y_i , that is the fraction of the input that is transformed into an output of good quality by the machine M_i , is given by:

$$Y_i = \frac{E_i^{Eff}}{E_i^{Tot}} = 1 - \frac{\gamma_i^W p_i(C_{i,q}) + p_i^{quality} \gamma_i^O}{p_i(C_{i,q}) + p_i^{quality}} \quad (8)$$

In the case of local monitoring, the machine yield only depends on the probability of the machine to go out of control and on the probability of correctly detecting an out of control $p_i(C_{i,q})$.

According to the system assumptions, M_i can be remotely monitored by the control chart $C_{i,q}$. In this case, the Markov chain modeling the behaviour of the remotely monitored machine in isolation becomes more complex, as represented in Figure 3(b). In particular, the following new states substitute the state O in the Markov chain of Figure 3(a):

- Out Of Control State Real But Not Observable (O_i^1): M_i is in this state when an out of control occurred, but the state has not been detected by the control chart $C_{i,q}$, since parts produced in the out of control state have not reached the inspection device yet, but are still crossing the portion of line between the monitored machine and the inspection device. The out of control state is still not observable to the control chart.
- Out Of Control State Real But Not Detected (O_i^2): M_i is in this state when parts produced in the out of control state reach the inspection device M_q , where the control chart $C_{i,q}$ is located. The out of control state is now visible to the control chart.

The only unknown transition in the Markov chain of Figure 3(b) is the transition between the states O_i^1 and O_i^2 . To calculate this transition, an additional *Quality Link Equation* is needed.

$$p_{i,q}^{delay} = \frac{1}{LT(C_{i,q})} \quad (9)$$

Where $LT(C_{i,q})$ is the lead time, i.e., the average number of time units parts spend in the portion of system between the monitored machine M_i and the machine M_q , where the inspection is located. This value is equal to zero if $i = q$, i.e., if the machine is an integrated machine. In this case the Markov chain represented in Figure 3(b) degenerates in the Markov chain of Figure 3(a). The lead time $LT(C_{i,q})$ represents the delay of the quality information, caused by the system architecture and the inspection device location.

Once all the transition probabilities are known in the state transition diagram of Figure 3(b), the steady-state probabilities and the performance measures of interest can be evaluated. For sake of clarity, we just report the expression of the machine yield Y_i , in the case of remotely monitored machine. Since the yield expression gets more complex than in the previous case, we introduce two additional coefficients,

a_i and d_i :

$$a_i = p_i(C_{i,q})p_{i,q}^{delay} \quad d_i = p_i^{quality} \left(p_i(C_{i,q}) + p_{i,q}^{delay} \right) \quad (10)$$

$$Y_i = \frac{\pi(W_i)(1 - \gamma_i^W) + (\pi(O_i^1) + \pi(O_i^2)) (1 - \gamma_i^O)}{\pi(W_i) + \pi(O_i^1) + \pi(O_i^2)} = 1 - \frac{\gamma_i^W a_i + \gamma_i^O d_i}{a_i + d_i} \quad (11)$$

It is interesting to notice that in the case of remote monitoring, also the probability $p_{i,q}^{delay}$ is in the yield formula. This means that the machine yield is not only a function of the performance of the control chart, but strictly depends on the time parts spend in the system among the monitored and the inspection machine.

4.2 Performance Evaluation of Long Production Lines

In order to evaluate the performance of a general production system of the type of those analyzed in Section 3, an extension of the decomposition technique proposed in [Matta and Tolio (1998)] is used. That method studied the performance of production lines with machines featuring multiple failure modes, from a production logistics point of view, i.e., without considering quality related aspects in the model. The idea of the approach is the following: the general system formed by K stations, is decomposed into $K - 1$ sub-systems each one formed by two pseudo-machines and one buffer, as represented in the right side of Figure 4. The sub-systems, also named building blocks, are analyzed by using the method in [Gershwin *et al.* (2002)], and the results are propagated among sub-systems by using the algorithm in [Dallery *et al.* (1988)]. Given the fact that the behaviour of the machines that are monitored

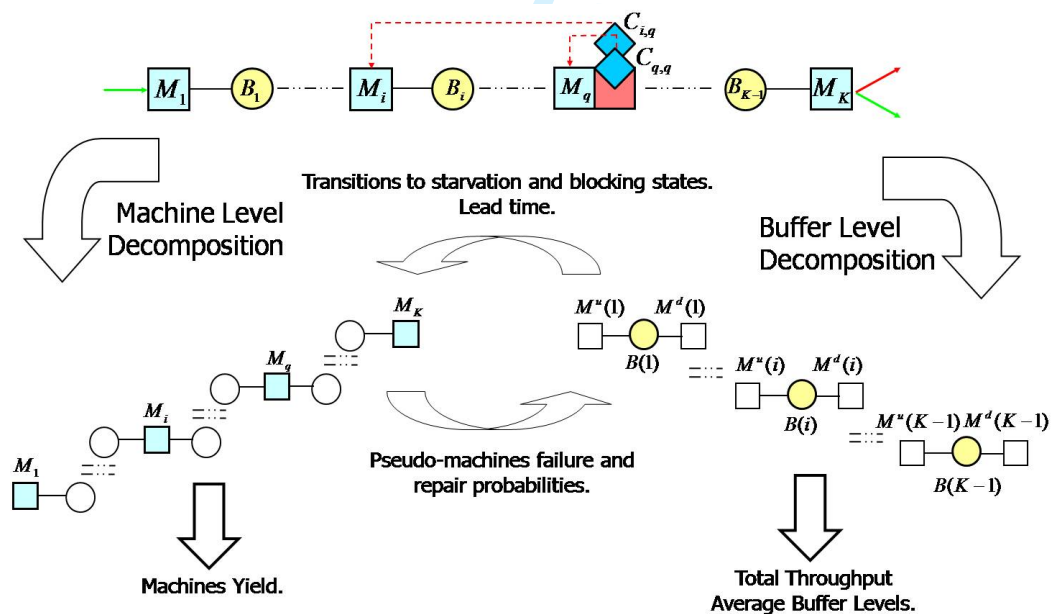


Figure 4. Application of the Two Level Decomposition technique to a general K machines serial line monitored by control charts.

by Statistical Process Control is more complex than the behaviour of multiple failure mode machines, they cannot be directly considered in the decomposition. In other words, there is a gap between the structures of the Markov chains modeling the behaviour of the machines in our system and the structures of the Markov

chains accepted by available decomposition methods. In particular, as we showed in the previous sections, machining stations may have two operational states, i.e., in control and out of control states, with different fraction of non-conformities associated. Currently, decomposition methods able to deal with machines featuring more than one operational state are not available. Thus, a preliminary analysis of these complex machines is needed. In fact, the approach proposed in this paper differs from the classical decomposition. In [Colledani *et al.* (2005)] a similar approach was applied to flexible machine processing multiple part-types and the method was named *Two-level Decomposition*. Also, in [Colledani and Tolio (2005)], the Two-Level Decomposition was successfully applied to the analysis of split and merge lines and assembly/disassembly systems.

More in details, one level of analysis is based on the evaluation of all the state probabilities of each machining station monitored by quality control in the line. These probabilities are obtained by solving discrete time-discrete state Markov chains representing the behaviour of such complex machines, taking also into account the influence of their neighboring buffers. Since the focus is on machines, this level of analysis is named *Machine Level Decomposition* (MLD). This preliminary analysis allows to approximately simplify the behaviour of such complex machines with that of multiple failure modes machines. With these simplified machine models, it is possible to perform the classical multiple failure mode two-machine line analysis, [Gershwin *et al.* (2002)]. It is based on the exact analytical solution of sub-systems formed by two pseudo-machines and one buffer. This second level of analysis is focused on the flow of material crossing the buffer. Indeed, we assign failure and repair parameters to the pseudo-machines of each building block in order to mimic the flow of material through the corresponding buffer of the original line. Therefore, we name this level of analysis *Buffer Level Decomposition* (BLD). The order in which the MLD and the BLD are applied to the machines and buffers in the line is controlled by an algorithm, similar to the DDX [Dallery *et al.* (1988)]. By studying alternately the BLD and the MLD and using the results obtained in one level as input for the other level, it is possible to evaluate the performance of the original complex system, once convergence conditions are met.

The Two-Level Decomposition approach is useful in those cases in which machines with a complex behaviour are included in the system. In this section, we present the analysis performed in both levels and propose the equations for exchanging the parameters from the MLD to the BLD and vice versa.

4.3 Buffer Level Decomposition

According to the decomposition method proposed in [Matta and Tolio (1998)], the generic subsystem $l(i)$ is composed by two pseudo-machines $M^u(i)$ (upstream) and $M^d(i)$ (downstream) and the buffer $B(i)$. The idea is that the flow of material through the buffer $B(i)$ of the sub-system must mimic the flow of material through the correspondent buffer B_i of the original line. In order to achieve this goal, failure modes must be properly assigned to $M^u(i)$ and $M^d(i)$, relating them to each possible cause for interruption of the material flow respectively entering and leaving the buffer. In this sense, machines in the subsystem are called pseudo-machines, because there is no direct matching with the real machines M_i in the original line. To be more precise, each pseudo-machine represents an entire portion of line. For instance, considering the system in Figure 4, the upstream pseudo-machine $M^u(i)$ of the building block $l(i)$ represents the entire portion of line positioned upstream the buffer B_i , i.e., M_1 , B_1 , and so forth until M_i . On the contrary, the direct correspondence exists between the buffers in the original and the decomposed

subsystems.

4.3.1 Structure of the Pseudo-machines

Considering the flow of material entering the original buffer B_i , it can be interrupted for different reasons. For each possible cause of interruption of the material flow entering the buffer B_i , a failure mode is assigned to the pseudo-machine $M^u(i)$.

Firstly, the machine M_i in the original line may fail while producing a part. For this reason, *local failure modes* are assigned to $M^u(i)$. The probabilities of failing and repairing due to these failure modes are equal to those of M_i :

$$p_{i,f_i}^{u(i)} = p_{i,f_i} \quad r_{i,f_i}^{u(i)} = r_{i,f_i} \quad (12)$$

for $f_i = 1, \dots, F_i$. They are data of the problem and are not modified within the algorithm.

Secondly, the machine M_i can be stopped by the control chart controlling it (either for a false alarm or for a correct detection of out of controls). To capture the situation in which the monitored machine M_i is stopped for a false alarm, a failure mode with probability $p_{false}^{u(i)}$ is assigned to the pseudo-machine $M^u(i)$. Similarly, for capturing the situation in which the machine M_i is stopped for a detected out of control, a failure mode with probability $p^{u(i)}(C_{i,q})$ is assigned to $M^u(i)$. These failure probabilities are unknown and will be provided in output by the MLD. However, once M_i is stopped for identified out of control, the probability of being restored to normal in control conditions is by definition $r_i^{quality}$. Moreover, once the machine is stopped given a false alarm of out of control, the probability of being restored to operational conditions is by definition r_i^{false} .

$$r^{u(i)}(C_{i,q}) = r_i^{quality} \quad r_{false}^{u(i)} = r_i^{false} \quad (13)$$

If M_i is not subject to out of control, these two modes of failures are not assigned.

Finally, machine M_j , ($j = 1, \dots, i - 1$), located in the portion of line upstream machine M_i may fail or may be stopped to fix quality problems. If machine M_j stays down for long time, machine M_i will continue the process until the upstream buffer gets empty. In this case, M_i is said to be starved by a specific cause of interruption of flow of M_j . The starvation is modeled through *remote failures* assigned to the pseudo-machine $M^u(i)$. The probability of failing in remote mode is unknown and will be given in output by the MLD. On the contrary, once starved, the probability that M_i is restored to operational is equal to the probability that machine M_j , which caused the starvation, is repaired [Matta and Tolio (1998)].

$$r_{j,f_j}^{u(i)} = r_{j,f_j} \quad (14)$$

for $j = 1, \dots, i - 1$ and $f_j = 1, \dots, F_j + 2I_j$, where I_j is equal to 1 only if machine M_j is subject to out of control. Thus, $2I_j$ are the two additional causes of starvation propagation to M_i that are possible only if the machine M_j is subject to out of control. The total number of failures assigned to the pseudo-machine $M^u(i)$ is therefore equal to:

$$t^u(i) = F_i + 2I_i + \sum_{j=1}^{i-1} [F_j + 2I_j] \quad (15)$$

where the first term refers to local failures, the term $2I_i$ adds two modes of failure

only in case M_i is subject to out of control and the last terms refers to remote starvation failures. Similar considerations can be done for assigning failures of the downstream pseudo-machine $M^d(i)$ of the generic building block $l(i)$, considering the propagation of blocking phenomena (failures of a downstream machine M_k , $k = i + 2, \dots, K$ which causes the upstream buffers to be full) in place of starvation.

4.3.2 Subsystem Performance Measures

Once updated failures probabilities are available from the MLD, the performance measures of the subsystem $l(i)$ can be evaluated by using the exact analytical method proposed in [Gershwin *et al.* (2002)]. In particular, the average throughput $E(i)$ of the subsystem and the average level of buffer $B(i)$, $\bar{n}(i)$, can be estimated. Moreover, the probabilities of the upstream pseudo-machine $M^u(i)$ to be blocked, $Pb_k(i)$, $k = 1, \dots, t^d(i)$, and the probabilities of $M^d(i)$ to be starved, $Ps_j(i)$, $j = 1, \dots, t^u(i)$, are calculated. According to the Little's law, the lead time in the subsystem $l(i)$ can be also estimated:

$$LT(i) = \frac{\bar{n}(i)}{E(i)} + 1 \quad (16)$$

4.3.3 Inputs to the MLD

Once having calculated the performance of the subsystem $l(i)$, the BLD provides in output information that will be taken as input by the MLD analysis. In particular, the transitions to blocking (b) states for the MLD analysis of M_i are calculated by using the decomposition equations provided in [Matta and Tolio (1998)]:

$$r_k^{i,b} = r_k^{d(i)} \quad p_k^{i,b} = \frac{Pb_k(i)}{E(i)} r_k^{i,b} \quad (17)$$

for $k = 1, \dots, t^d(i)$. Similarly, the transitions to starvation (s) states for the MLD analysis of M_{i+1} are given by:

$$r_j^{i+1,s} = r_j^{u(i)} \quad p_j^{i+1,s} = \frac{Ps_j(i)}{E(i)} r_j^{i+1,s} \quad (18)$$

for $j = 1, \dots, t^u(i)$.

4.4 Machine Level Decomposition

4.4.1 State Transition Diagram for M_i

The MLD analysis is explained in details for the remotely monitored machine M_i of the system in Figure 4. The Markov chain representing the behaviour of the complex monitored machine is represented in Figure 5(a). As it can be noticed, this Markov chain is similar to that reported in Figure 3(b), except for the presence of starvation states (blocking states are considered but not represented, for reducing the complexity of the figure). We now investigate the availability of the transition probabilities of the Markov chain in Figure 5(a). In Section 4.1, quality link equations to evaluate the probabilities of entering the states A_i^1 , A_i^2 , and O_i^2 , were provided. In particular, the third quality link equation 9 for the case of machining station remotely monitored expressed $p_{i,q}^{delay}$ as a function of the lead time. It is the time required for parts processed by M_i to reach the inspection point located at

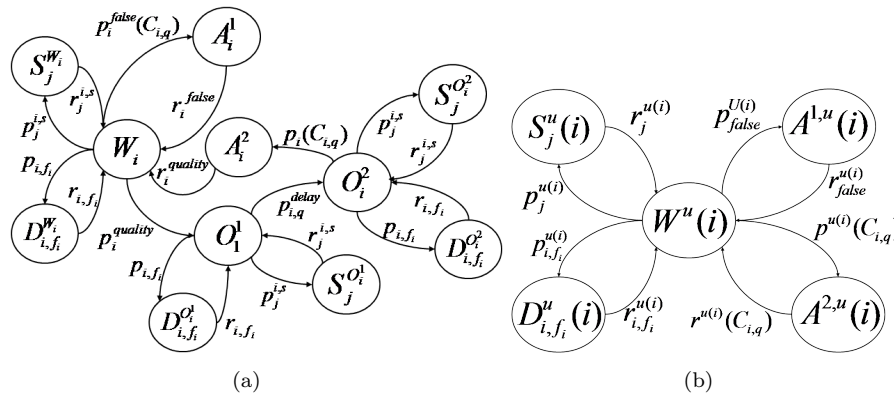


Figure 5. State transition diagram for the remotely monitored machine M_i in the K machines line. State transition diagram for the pseudo-machine $M^u(i)$.

M_q . While considering the machine as integrated in the system, that equation can be further detailed:

$$p_{i,q}^{delay} = \frac{1}{LT(C_{i,q})} = \frac{1}{\sum_{j=i}^{q-1} LT(j)} \quad (19)$$

Where $LT(j)$ is the lead time related to each two-machine one-buffer subsystem included between the monitored station M_i and the control chart located at M_q . This parameter is one of the input taken by the BLD. The only remaining unknown probabilities in this Markov chain are the probabilities of the machine M_i to enter in starvation (blocking) states. These parameters were provided in output by the BLD, specifically in equations 17 and 18. Therefore, all the transition probabilities in this Markov chain are known. It can be noticed that the unknown parameters of the MLD are function of the output parameters of the BLD.

4.4.2 Machine Level Analysis of M_i

The Markov chain in Figure 5(a) can be solved and steady-state probabilities can be calculated. By using the equations provided in Section 4.1, the machine level performance measure of M_i can be estimated. Then, in order to match the simplified multiple failure mode machine structure, the Markov chain of Figure 5(a) is transformed into the Markov chain of Figure 5(b). The transformation is made through the re-distribution of the calculated steady-state probabilities, performed by using the following *State Aggregation Equations*:

$$\pi(W^u(i)) = \pi(W_i) + \pi(O_i^1) + \pi(O_i^2) \quad (20)$$

$$\pi(A^{1,u}(i)) = \pi(A_i^1) \quad (21)$$

$$\pi(A^{2,u}(i)) = \pi(A_i^2) \quad (22)$$

$$\pi(D_{i,f_i}^{u(i)}) = \pi(D_{i,f_i}^{W_i}) + \pi(D_{i,f_i}^{O_i^1}) + \pi(D_{i,f_i}^{O_i^2}) \quad f_i = 1, \dots, F_i \quad (23)$$

$$\pi(S_j^u(i)) = \pi(S_j^{W_i}) + \pi(S_j^{O_i^1}) + \pi(S_j^{O_i^2}) \quad j = 1, \dots, t^u(i-1) \quad (24)$$

These aggregated state probabilities are those of the upstream pseudo-machine $M^u(i)$ of the next two-machine subsystem to be analyzed by the BLD, namely $l(i)$.

4.4.3 Inputs to the BLD

In this paragraph, the unknown failure probability to remote down modes, to the false alarm down mode and to the detected out of control down mode, will be

provided to the BLD analysis of the subsystem $l(i)$. The probability of machine $M^u(i)$ to enter the down state $A^{1,u(i)}$ can be obtained by writing a balance equation at node $A^{1,u(i)}$ of the Markov chain in Figure 5(b) and considering the State Aggregation Equations from 20 to 24:

$$p_{false}^{u(i)} = \frac{\pi(A^{1,u(i)})}{\pi(W^u(i))} r_{false}^{u(i)} = \frac{\pi(A_i^1)}{\pi(W_i) + \pi(O_i^1) + \pi(O_i^2)} r_{false}^{u(i)} \quad (25)$$

Similarly, the probability of machine $M^u(i)$ to enter the down state $A^{2,u(i)}$ is given by:

$$p^{u(i)}(C_{i,q}) = \frac{\pi(A^{2,u(i)})}{\pi(W^u(i))} r^{u(i)}(C_{i,q}) = \frac{\pi(A_i^2)}{\pi(W_i) + \pi(O_i^1) + \pi(O_i^2)} r^{u(i)}(C_{i,q}) \quad (26)$$

Finally, the probability of failing in remote starvation (blocking) mode are updated by using the following equation:

$$p_j^{u(i)} = \frac{\pi(S_j^u(i))}{\pi(W^u(i))} r_j^{u(i)} = \frac{pi(S_j^{W_i}) + \pi(S_j^{O_i^1}) + \pi(S_j^{O_i^2})}{\pi(W_i) + \pi(O_i^1) + \pi(O_i^2)} r_j^{u(i)} \quad (27)$$

for $j = 1, \dots, t^u(i-1)$. Equations 25, 26 and 27 are the output information of the MLD analysis that are used in the BLD analysis. It is worth to highlight that, in case the machine under analysis is not subject to out of control, the machine level analysis highly simplifies and, in principle, can also be skipped. In particular, equations 25, 26 cannot be written and the machine yield must be set to 1 since the machine does not have any effect on the part quality. The Markov chain in Figure 5(a) easily simplifies to the multiple failure modes Markov chain (Figure 5(b)) and equation 27 simply transfer to the next subsystem the transition to starvation states given in 18.

4.5 System Performance Measures

In the Appendix A, the algorithm that controls the sequence of BLD and MLD analysis and enables to solve the provided equations is reported. Once the terminating condition of the algorithm is met, the system level performance measures can be derived. For the conservation of flow, the total production rate E^{Tot} can be expressed as:

$$E^{Tot} = E(1) = E(2) = \dots = E(K-1) \quad (28)$$

The system yield is obtained as the product of the machine Yields Y_i calculated at the MLD.

$$Y^{System} = \prod_{i=1}^K Y_i \quad (29)$$

It must be commented that the machine yield Y_i is an output of the Two-level Decomposition approach. It is calculated by considering in detail the quality information feedback and blocking and starvation phenomena. Thus it will be different from the yield of the machines M_i considered in isolation. Given equation 2 the

effective throughput can be expressed as:

$$E^{Eff} = E^{Tot} Y^{System} \quad (30)$$

Finally, the average buffer levels and the system work in progress can be calculated as:

$$\bar{n}_i = \bar{n}(i) \quad i = 1, \dots, K-1 \quad WIP = \sum_{i=1}^{K-1} \bar{n}_i + K E^{Tot} \quad (31)$$

5. Numerical Results

5.1 Accuracy Testing

The proposed method has been implemented in C_{++} and the results have been compared with simulation. The discrete time simulation model has been built in ARENA, by adopting the same assumptions as those reported in Section 3.2. In order to test the assumption of assimilating a quality monitored machine to a Markov process and the assumption of generating false alarms each time a part instead of a sample is processed, the quality control loop is considered in the simulation model in details. In particular, we assume a certain known distribution for the quality characteristic to be monitored. For each part processed by the machine subject to out of control, we sample the value of the product characteristic from the given statistical distribution. At the inspection station, a sample composed by $m(C_{i,q})$ parts is collected. Furthermore, depending on the quality control chart type adopted, we simulate the real behaviour of the chart by comparing the monitored quality characteristic observed in the sample with the control limits. For instance, if we are simulating a Shewart chart, we calculate the mean of the quality characteristic for the parts in the sample and we compare it with the control limits positioned at 3σ . If the mean of the sample is inside the control limit, no signal is generated, while if it is out of the control limits, the machine is stopped for being repaired. The accuracy of the method has been tested over more then 200 cases. Test systems were characterized by different number and location of the inspection devices, different number of machines varying from 2 to 10, both locally and remotely monitored, different failure parameters varying from 0.0001 and 0.2, different repair parameters varying from 0.001 and 0.8, different quality failure probabilities varying from 0.0001 and 0.1 and different buffer capacities ranging from 4 to 50. The efficiency in isolation of machines varied in the range $[0.8, 0.99]$, values that are typically observed in actual machines. Moreover, both 100% and sampling inspections were considered, with sample size varying from 1 to 25 and number of not measured parts between samples varying from 80 to 800, and equal to zero in the case of 100% inspection. For each case, 10 simulation runs of 5.000.000 time units have been performed. The error in the estimation of the generic performance measure θ is calculated by using the following equation:

$$err\%^\theta = 100 \frac{\theta_{An} - \bar{\theta}_{Sim}}{\bar{\theta}_{Sim}} \quad (32)$$

where $\bar{\theta}_{Sim}$ is the average value of the outputs θ_{Sim}^r of each one of the ten replicates per simulation test. In case θ is the buffer level \bar{n}_i , the denominator in equation 32 is set to the buffer capacity N_i , for making the errors for cases with large and small

buffers comparable. The 95% confidence interval on all the considered performance measures is reported in the following tables.

Tables 2 and 3, and Tables 4 and 5, report the analysis of cases with locally monitored machines, respectively in lines with three different machines and ten equal machines. In Tables 2 and 4, the system data are reported while in Tables 3 and 5, a comparison among the results obtained with the proposed method and simulation is shown. In addition, the CPU time required by simulation and by the proposed analytical method to generate the results is reported in the last column, in seconds. Tables 6 and 7 are related to the system architecture represented in Figure 6, characterized by both locally and remotely monitored machines. Table 6 reports the system data for 3 different cases and Table 7 reports the errors with simulation. Both the cases of sampling ($h > 0, m > 1$) and 100% inspections ($h = 0, m = 1$) are considered in the reported examples.

Table 2. Data for three locally monitored machine cases.

C	N_i	M_i	p_i^{qual}	r_i^{qual}	r_i^{false}	$p_{i,1}$	$r_{i,1}$	γ_i^W	γ_i^O	h	m	ARL_0	ARL_1
1	6	1	0.002	0.6	0.7	0.017	0.102	0.01	0.09	0	1	25	1.11
	12	2	0.04	0.22	0.6	0.09	0.19	0.02	0.17	0	1	25	1.11
	/	3	0.007	0.59	0.98	0.003	0.18	0.02	0.3	0	1	25	1.11
2	4	1	0.006	0.23	0.7	0.016	0.42	0.02	0.22	100	4	370.3	1.087
	4	2	0.02	0.08	0.9	0.21	0.23	0.01	0.11	180	4	370.3	1.064
	/	3	0.041	0.03	0.6	0.12	0.45	0.001	0.01	500	4	370.3	1.137
3	25	1	0.001	0.12	0.34	0.006	0.52	0.001	0.08	400	20	370.3	1.003
	25	2	0.002	0.32	0.78	0.005	0.011	0.01	0.02	100	20	370.3	1.003
	/	3	0.003	0.02	0.33	0.021	0.092	0.001	0.042	200	20	370.3	1.003
4	20	1	0.001	0.012	0.2	0.21	0.006	0.001	0.18	80	10	370.3	1.003
	4	2	0.004	0.12	0.29	0.142	0.11	0.01	0.026	300	6	370.3	1.15
	/	3	0.007	0.09	0.38	0.235	0.092	0.003	0.012	600	4	370.3	1.15

Table 3. Results for three locally monitored machine cases.

Case	Method	E^{Tot}	E^{Eff}	γ^{System}	\bar{n}_1	\bar{n}_2	Time [s]
1	Sim	0.55703	0.52524	0.94291	4.5361	0.84605	532.8
	95% I.C.	± 0.00045	± 0.00038	± 0.00018	± 0.00342	± 0.00286	
	An.	0.55695	0.52515	0.9429	4.5271	0.8522	2.2
	Err%	0.0143	0.016	0.001	0.149	0.051	
2	Sim	0.49427	0.45493	0.9204	3.4094	1.1239	547.8
	95% I.C.	± 0.00034	± 0.00047	± 0.00035	± 0.00152	± 0.00159	
	An.	0.49404	0.45484	0.92066	3.4101	1.1242	2.45
	Err%	0.047	0.0186	0.0285	0.0171	0.0089	
3	Sim	0.62781	0.61158	0.97414	23.846	13.9314	541.2
	95% I.C.	± 0.00209	± 0.00203	± 0.00016	± 0.00778	± 0.0717	
	An.	0.62782	0.61158	0.97413	23.858	13.9167	3.78
	Err%	0.0009	0.0002	0.0007	0.0457	0.059	
4	Sim	0.19714	0.1941	0.98457	8.7769	2.1503	448.8
	95% I.C.	± 0.00028	± 0.00025	± 0.00019	± 0.06194	± 0.00566	
	An.	0.19899	0.19592	0.98456	8.64024	2.14567	2.33
	Err%	0.9415	0.9405	0.001	3.4186	0.0232	

Table 4. Data for ten identical locally monitored machine cases.

C	N_i	M_i	p_i^{qual}	r_i^{qual}	r_i^{false}	$p_{i,1}$	$r_{i,1}$	γ_i^W	γ_i^O	h	m	ARL_0	ARL_1
1	4	$\forall i$	0.006	0.42	0.65	0.01	0.1	0.02	0.25	150	4	370.3	1.018
2	6	$\forall i$	0.101	0.57	0.8	0.159	0.6	0.05	0.32	0	1	370.3	1.1

Table 5. Results for ten identical locally monitored machine cases.

Case	Method	E^{Tot}	E^{Eff}	γ^{System}	WIP	Time [s]
1	Sim	0.60257	0.25686	0.42628	18.015	1872
	95% I.C.	± 0.00053	± 0.00072	± 0.00113	± 0.02715	
	An.	0.60261	0.25705	0.42656	18	13.86
	Err%	0.0075	0.0723	0.0642	0.0853	
2	Sim	0.73142	0.39664	0.54231	26.9855	1882.2
	95% I.C.	± 0.00017	± 0.00016	± 0.002	± 0.03047	
	An.	0.73921	0.40074	0.54211	27	11.2
	Err%	1.0652	1.0333	0.0366	0.0534	

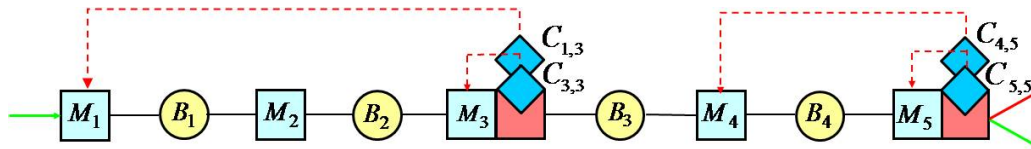


Figure 6. Example of the considered production system layout.

Table 6. Data for five machine systems with remote monitoring of Figure 6.

C	N_i	M_i	p_i^{qual}	r_i^{qual}	r_i^{false}	$p_{i,1}$	$r_{i,1}$	γ_i^W	γ_i^O	h	m	ARL_0	ARL_1
1	8	1	0.002	0.51	0.4	0.023	0.203	0.001	0.1	150	4	370.3	1.188
	4	2	/	/	/	0.089	0.319	/	/	/	/	/	/
	30	3	0.009	0.32	0.9	0.102	0.412	0.002	0.13	150	4	370.3	1.188
	32	4	0.007	0.12	0.4	0.076	0.098	0.002	0.02	200	8	370.3	1.004
	/	5	.0006	0.103	0.7	.0012	0.009	0.001	0.11	200	8	370.3	1.004
2	12	1	.0012	0.098	0.24	0.008	0.102	0.001	0.12	400	4	370.3	1.023
	12	2	/	/	/	0.002	0.089	/	/	/	/	/	/
	38	3	.0002	0.201	0.65	.0009	.0087	0.001	0.19	400	4	370.3	1.13
	28	4	.0006	.2901	0.19	.0102	0.023	0.001	0.22	100	4	370.3	1.268
	/	5	.0009	0.146	0.25	.0068	0.013	0.001	0.5	100	4	370.3	3.645
3	4	1	0.004	0.12	0.29	0.142	0.221	0.01	0.026	250	6	370.3	1.016
	31	2	/	/	/	0.104	0.098	/	/	/	/	/	/
	3	3	.0001	0.7	0.67	0.03	0.12	0.005	0.42	250	6	370.3	1.086
	25	4	0.091	0.24	0.92	0.174	0.03	0.003	0.09	500	10	370.3	1.271
	/	5	.0002	0.012	0.32	0.005	0.42	0.002	0.118	500	10	370.3	1.083

Table 7. Results for five machine systems with remote monitoring of Figure 6.

Case	Method	E^{Tot}	E^{Eff}	γ^{System}	\bar{n}_1	\bar{n}_2	\bar{n}_3	\bar{n}_4	Time [s]
1	Sim	0.53367	0.51747	0.96964	6.883	2.546	23.394	3.075	652.2
	95% I.C.	± 0.001	± 0.0009	± 0.0001	± 0.0041	± 0.0034	± 0.0474	± 0.0527	
	An.	0.53361	0.51579	0.96653	6.9208	2.5725	23.6546	3.0581	4.8
	Err%	0.01	0.3305	0.3205	0.4801	0.6595	0.88	0.0526	
2	Sim	0.52116	0.48178	0.92443	9.7345	10.1829	33.4719	13.0687	708
	95% I.C.	± 0.002	± 0.0024	± 0.0026	± 0.0149	± 0.0196	± 0.1257	± 0.1601	
	An.	0.52371	0.48328	0.92279	9.8087	10.1981	33.6637	12.987	6.21
	Err%	0.4898	0.3115	0.1773	0.6183	0.1266	0.5045	0.324	
3	Sim	0.14317	0.13552	0.94657	3.5323	29.337	2.7504	0.1615	508.2
	95% I.C.	± 0.0006	± 0.0005	± 0.001	± 0.0023	± 0.0205	± 0.0027	± 0.003	
	An.	0.14345	0.13574	0.94619	3.5445	29.3515	2.7486	0.1617	5.12
	Err%	0.1571	0.198	0.0407	0.303	0.0466	0.0614	0.0006	

As it can be noticed, the proposed method is very accurate, both in terms of the total and the effective throughput, of the system yield and of the average buffer level. As a summary, the maximum error on all the studied cases remains below 2% on the effective and the total throughput, with 90% of the cases below 1.5%. Regarding the average buffer levels, the maximum error is 4.2% and in the 90% of the cases it is below 2.5%. Moreover, even if the convergence of the algorithm cannot be formally proved (like usual for decomposition methods Dallery *et al.* (1988)), in

all the 200 analyzed cases, the method showed to converge rapidly, always with less than 15 complete iterations. This resulted in extremely high speed of convergence, comparable to classical decomposition based methods: as evidenced in the tables, the method provided the performance measures in few seconds. Thus, if compared to discrete time simulation in terms of speed, our approximate analytical method is typically faster of two orders of magnitude. It must be commented that, in terms of algorithmic complexity, our Two-Level Decomposition only differs from classical decomposition approaches for the application of the MLD. However, this additional step only entails the analysis of a discrete time-discrete state Markov chain, for each machine affected by quality problems, and the application of the state aggregation equations. Thus, the additional complexity of our approach if compared to existing methods is very limited.

5.2 Operating Curves and Properties of the System

The experiments reported in the previous section show the good accuracy and the high convergence speed of the proposed method. Therefore, we are allowed to make use of this method to investigate some properties of the type of systems under analysis.

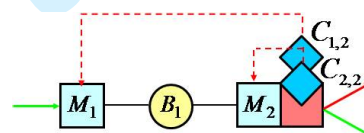


Figure 7. Two-machine line considered in the analysis. The only inspection station is located at stage 2. M_1 is remotely monitored by $C_{1,2}$ and M_2 is locally monitored by $C_{2,2}$.

Table 8. Parameters of the two-machine line under analysis.

	p_i	r_i	$p_i^{quality}$	$r_i^{quality}$	r_i^{false}	γ_i^W	γ_i^O
M_1	0.007	0.194	0.007	0.102	0.9	$1.58 \cdot 10^{-5}$	$2.22 \cdot 10^{-1}$
M_2	0.012	0.1	0.005	0.1	0.95	$2 \cdot 10^{-9}$	$2.5 \cdot 10^{-3}$
	$m(C_{i,q})$	$h(C_{i,q})$	$ARL_0(C_{i,q})$	$ARL_1(C_{i,q})$			
$C_{1,2}$	1	0	370	1.18			
$C_{2,2}$	1	0	144	1.446			

The first experiment considers a simple two-machine line. M_1 is a machining station subject to out of controls and it is remotely monitored by the control chart $C_{1,2}$. This control chart uses the information measured by the inspection station that is integrated just downstream M_2 . The machine M_2 is also subject to out of controls and is locally monitored by the control chart $C_{2,2}$. The layout of the system under analysis is reported in Figure 7. It is worth to underline that a similar system cannot be analyzed in an integrated way by any method existing in the literature. The parameters for this system are reported in Table 8. As it can be noticed, 100% inspection is considered ($h = 0, m = 1$). The efficiency in isolation for the two machines, calculated by using equation 6 since no information is available a priori on the transition $p_{1,2}^{delay}$, is respectively equal to 0.903 and 0.85. Thus, in this test case, $e_1 > e_2$. In this experiment, we aim at investigating the behaviour of the total throughput, the effective throughput, the work in progress and the system yield as a function of the buffer capacity N_1 . For this reason, we varied N_1 in the range $[3, 90]$ and we run the proposed algorithm for all the different resulting systems. The plots of these four performance measures are reported in Figure 8. In the literature it

is well known ([Dallery and Gershwin (1992)], [Meester and Shantikumar (1990)]) that the increasing capacity of the buffer has a positive impact on the system throughput. This is perfectly in line with the results reported in Figure 8 concerning the total production rate of the system (bottom right box). E^{Tot} is shown to be a monotonic increasing function of N_1 , with asymptotic behaviour as N assumes large values. However, by observing the effective throughput E^{Eff} (bottom left box) a different behaviour can be observed. Indeed, E^{Eff} is shown to be a concave function of the buffer capacity N_1 , presenting a maximum for $N_1^{max} = 24$.

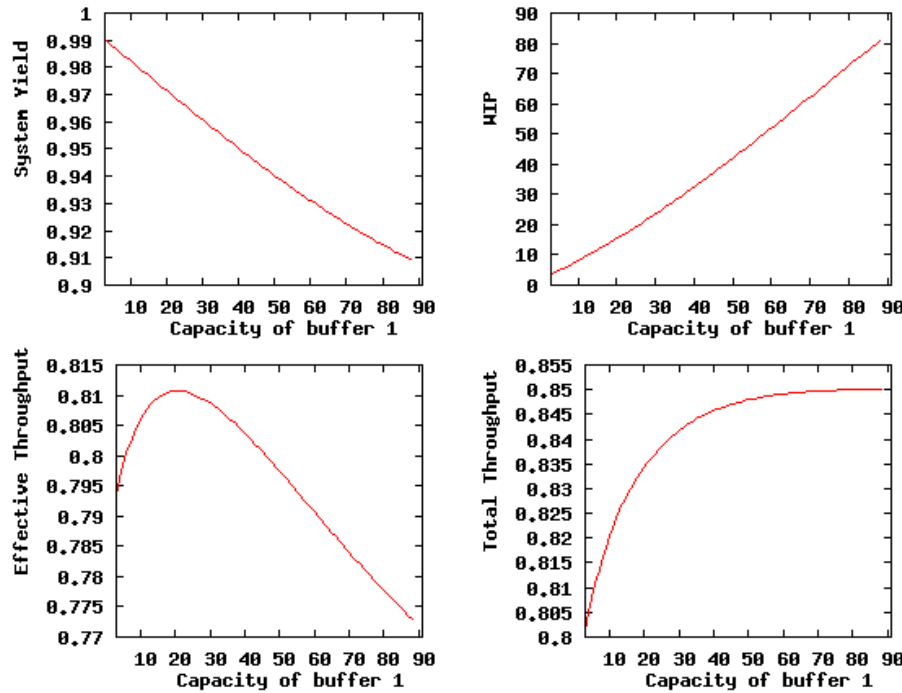


Figure 8. System performance measures under variation of the capacity of the buffer B_1 .

The explanation of this behaviour is the following. By increasing the buffer capacity two opposing effects are observable in the production line. On the one hand, the total throughput of the system increases, with higher impact when the buffer is small (the total throughput curve has higher derivative). From the other hand, the delay of the quality information feedback increases. This delay is due to the fact that when the machine M_1 goes out of control, even if 100% inspection is performed, the parts do not instantaneously reach the inspection point, but are stored in the buffer B_1 before being processed by M_2 . Therefore, the longer is the queue in B_1 , the longer is the time required by the control chart to generate the quality control feedback (alarm). Since, in this case, the average buffer level increases quasi-linearly with the buffer capacity ($e_1 > e_2$), the delay in the quality information feedback has higher impact for higher buffer sizes. Therefore, two effects are found to be in trade-off: when the positive impact of the buffer on the total throughput is more important (low buffer sizes), then the effective throughput increases; when the delay in the quality information feedback is more important (high buffer sizes), then the relative weights of these two phenomena change and the effective throughput starts decreasing. This analysis demonstrates that the buffers in the system should be designed to increase the total throughput, with particular attention to the effects they have on the quality control system reactivity. In other

terms, they should be designed by using an integrated quality-quantity framework. Furthermore, this analysis shows that having low work in progress in the system may increase the quality of the parts released in output by the system, like Just in Time (JIT) advocates suggest. However, if the work in progress gets too small, the production rate of the system decreases and the effective throughput may decrease accordingly. Thus, this trade-off cannot be addressed unilaterally, but the mutual relation among quality and productivity measures provides reasonable answers to these issues.

In Figure 8 the effective throughput, the total throughput, the system yield and the work in progress for five different values of the failure probability $p_1 = [0.007, 0.013, 0.019, 0.025, 0.031]$ are reported. These values generate the following efficiencies in isolation of machine M_1 , $e_1 = [0.903, 0.878, 0.855, 0.833, 0.812]$. The system is the same as that analyzed in the previous example and the efficiency in isolation of the second machine is kept constant ($e_2 = 0.85$). While decreasing the efficiency in isolation of the first machine, the value of the buffer capacity that maximizes the effective throughput shifts toward the right side of the graph (bottom left box). Numerical values for the optimal buffer sizes are $N_1^{max} = [24, 31, 39, 49, 61]$, for each one of the five test cases. As a system design rule, it can be noticed that while increasing the value of p_1 , the optimal buffer capacity increases.

The motivation for this behaviour is the following. When the efficiency in isolation of the first machine decreases, then the slope of the work in progress curve as a function of the buffer capacity also decreases (top right box). In other words, less material will be accumulated in the buffer, since the first machine is more frequently failed due to operational failures. This effect directly impacts on the delay of the quality information feedback. If less material accumulates in the buffer, the lead time is shorter and the delay of the quality information feedback reduces. Thus, the control chart identifies the out of control with higher reactivity. This is visible also by looking at the System Yield graph (top left box). Indeed, for higher values of p_1 the yield decreases quasi-linearly but with lower slope.

We now focus on the trade-off already analyzed between the positive impact of buffers on the total throughput and their negative impact on the quality information feedback delay. As the efficiency of machine M_1 reduces (higher p_1), the second effect loses importance. Therefore, the positive impact of buffers on the total throughput will dominate the negative impact of buffers on the delay of the quality information feedback for a larger set of buffer capacities. As a consequence, the effective throughput curve increases for a larger set of buffer sizes. If the efficiency in isolation of the first machine is further reduced ($e_1 < e_2$), then the work in progress curve becomes an asymptotic function of the buffer capacity. This means that the delay in the quality information feedback has an upper bound. Therefore, for values of p_1 such that ($e_1 < e_2$) the effective throughput curve may become a monotonic increasing function of the buffer capacity, exactly like the total throughput curve. These considerations lead to important consequences on the design of production lines. Depending on the machine reliability parameters, the impact of the buffer size on the production line performance may be positive (when $e_1 < e_2$) but also negative (when $e_1 \geq e_2$). Therefore the properties of the effective throughput curve for this type of systems have been characterized and motivated.

There is another relevant information that can be derived from the analysis of these graphs. Indeed, a counterintuitive rather than surprising behaviour can be observed. Consider the two graphs reported in the bottom left box of Figure 9, for two values of $p_1 = [0.007, 0.019]$. It can be observed that the two curves have an intersection point, approximately corresponding to a value of the buffer capacity

equal to $N_1 = 56$. For $N_1 < 56$ an effort in reducing the probability of failure of machine M_1 positively impact on the effective production rate of the system. This is perfectly in line with the motivation behind the application of common machine efficiency improvement practices. However, for values of the buffer capacity $N_1 > 56$ an effort in reducing the probability of failure of machine M_1 have a negative impact on the effective production rate of the system, causing the system to deliver less conforming parts in output. Intuitively, in this region of the graph, if the first machine efficiency is increased, it stays down for a smaller fraction of time and it is enabled to produce more parts. However, since the control chart has high delay in the detection of out of controls, due to the time consuming storage of parts in the buffer, most of the parts the machine produces will be non-conforming. Thus, the effective production rate decreases. This analysis demonstrates that by applying local improvement actions (reduction of the machine failure probability), without jointly considering the impact on the entire system both in terms of quality and productivity, the action may result in a global degradation of the performance at a system level.

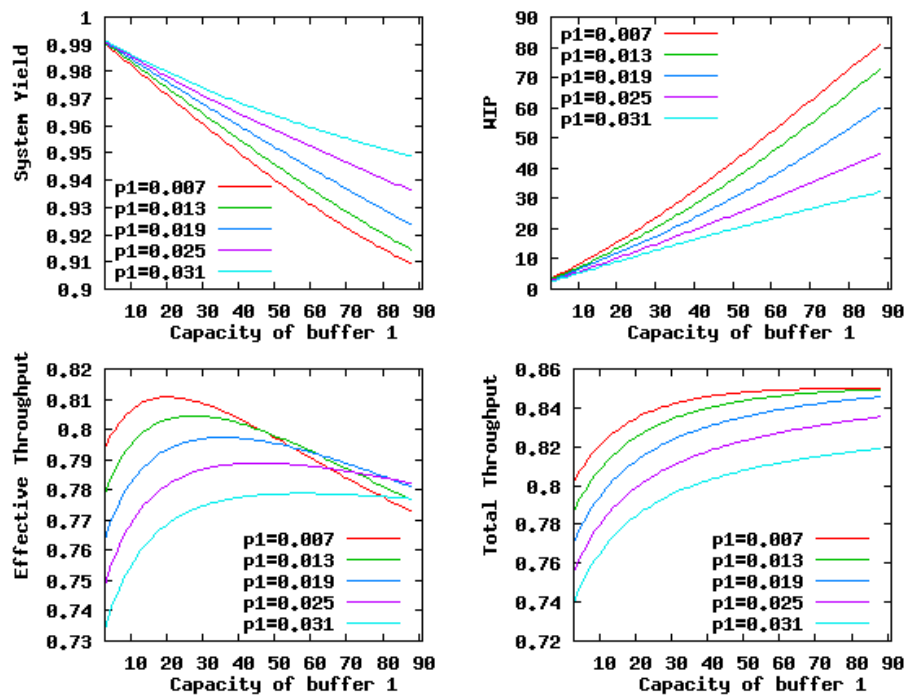


Figure 9. System performance measures under variation of the capacity of buffer B_1 for 5 different values of $p_1 = [0.007, 0.013, 0.019, 0.025, 0.031]$. Resulting efficiency in isolation of M_1 varies with values $e_1 = [0.903, 0.878, 0.855, 0.833, 0.812]$.

The dependency of the effective production rate on the sampling frequency is addressed in the following. For investigating this aspect, the effective throughput curve is charted for five different values of $h(C_{1,2}) = [0, 30, 60, 90, 120]$. Indeed, the inverse of the number of not measured parts between samples $h(C_{1,2})$ is a measure of the sampling frequency (in case $h(C_{1,2}) > 0$). The graphs are reported in Figure 10. Lower monitoring effort of the quality control system leads to lower effective throughput. This is reasonable because in case of out of control, more time is needed by the control chart to generate the alarm. However, another interesting consideration can be done. When performing 100% monitoring the quality control

system is itself very reactive (ARL_1 is small). However, the effect of the delay due to the storage time of parts in the buffer is predominant and causes the delay in the quality information feedback. On the contrary, when the responsiveness which is proper of the control system decreases (sampling inspection case), the impact of the delay due to the buffer progressively smoothes, until it is negligible if compared to the delay due to the chart parameters (case $h(C_{1,2}) = 120$). Therefore, by observing the system behaviour, it is possible to distinguish the effect of the two root causes for the delay in the quality feedback. It is relevant to identify which of the two causes is more influential, in order to select the right actions to improve the system performance, depending on the operating conditions of the system. Furthermore, depending on the sampling policy, the maximum in the system effective throughput can be found for different values of N_1 . In particular:

- If the control chart monitoring effort is high (low values of $h(C_{1,2})$), then the maximum will be visible for lower values of N_1 .
- If the control chart monitoring effort is low (high values of $h(C_{1,2})$), then the maximum will be visible for higher values of N_1 .

This analysis underlines the need for jointly taking into consideration the design of the system layout parameters, such as the capacity of the buffers, and the quality control parameters, such that the sampling inspection frequency, in order to select a globally optimized solution.

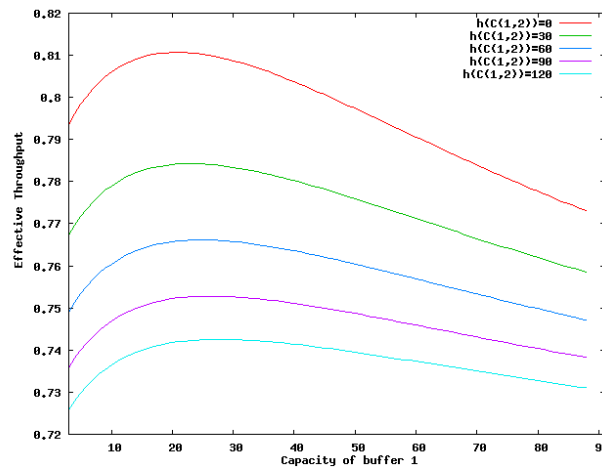


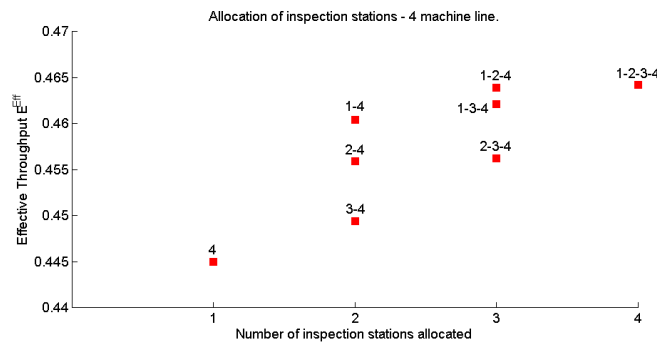
Figure 10. Effective throughput under variation of the capacity of buffer B_1 for 5 different values of the number of parts between samples $h(C_{1,2}) = 0, 30, 60, 90, 120$. By decreasing the monitoring effort the maximum effective throughput is found for higher buffer capacity.

The last experiment aims at testing the effect of different locations of the inspection stations in the system. For this purpose, a system formed by 4 stations, with different reliability parameters, has been considered. Data of the system are reported in Table 9. Each machine is subject to out of control and may be monitored locally or remotely. Considering all the possible locations of inspection points in the system, eight different configurations are possible, since the last station must be always monitored, being subject to out of control. We refer to a specific allocation of inspection station by using this notation: configuration 1 – 2 – 4 has one inspection station integrated in M_1 , thus this machine is locally monitored, the same is true for M_2 , while the last inspection point is located downstream M_4 and monitors M_3 in a remote way and M_4 in a local way. In Figure 11 the effective throughput as a function of the number of inspection stations allocated in the system is reported for all the eight possible configurations.

Three relevant considerations can be done:

Table 9. Data for the four machines in the test line for the inspection location experiment.

N_i	M_i	p_i^{qual}	r_i^{qual}	r_i^{false}	$p_{i,1}$	$r_{i,1}$	γ_i^W	γ_i^O	h	m	ARL_0	ARL_1
3	1	0.08	0.52	0.94	0.1	0.4	0.016	0.42	0	1	370.3	4.72
3	2	0.06	0.54	0.9	0.08	0.5	0.016	0.42	0	1	370.3	8.69
3	3	0.06	0.56	0.88	0.1	0.49	0.016	0.184	0	1	370.3	14.967
/	4	0.05	0.58	0.78	0.1	0.48	0.016	0.115	0	1	370.3	27.831

Figure 11. System effective throughput E^{Eff} as a function of the number of inspection stations allocated in the line. 8 possible allocations exist for the 4 machine line under analysis.

- The inspection station allocation strongly impacts on the production rate of the system (max difference = 4.5%).
- The big improvement in the effective throughput is given by passing from one inspection device to two inspection devices, correctly allocated (+3.46%). The addition of new inspection devices only partially increases the effective throughput of the system.
- A lower number of inspection stations coherently positioned performs better than a higher number of stations poorly deployed. For instance, in the case of solution 1 – 4 and 2 – 3 – 4, the money investment for the additional inspection device is not correctly exploited.

In Table 10, also the average work in progress for each allocation is provided. It can be seen that, for each number of inspection stations allocated, the best solutions are also those characterized by lower work in progress. This means lower response time of the quality information feedback.

Table 10. Results for the 8 possible allocations of inspection stations for the analyzed 4 machines system.

Allocation	E^{Eff}	E^{Tot}	Y^{System}	WIP
4	0.444992	0.629997	0.70634	4.28407
1-4	0.46042	0.623776	0.738118	4.16414
2-4	0.455928	0.624705	0.72983	4.21379
3-4	0.449405	0.627483	0.716203	4.26193
1-2-4	0.463879	0.621502	0.746382	4.15312
1-3-4	0.462095	0.62243	0.742427	4.1644
2-3-4	0.456245	0.624201	0.730927	4.21827
1-2-3-4	0.464226	0.621014	0.747529	4.1576

From this experiment, it appears clear that neglecting the impact of the system architecture on the propagation of the quality information can lead to wrong money investment, also of consistent entity. Moreover, also operating costs may be increased, by taking wrong decisions caused by a unidirectional vision of the problem. Indeed solutions characterized by higher inventory may be selected. The need of considering the quality problem at a system level is therefore evidenced by the proposed results.

6. Conclusions

An approximate analytical method for evaluating the performance of a production system where machines are monitored by using Statistical Process Control has been developed. In particular, the method considers the presence of inspection and integrated stations in the line that are subject to operational failures and to out of controls. While operational failures are easy to detect, out of controls need quality control tools to be detected. For this reason, in the system, inspection points which collect data to compile control charts are included. Numerical results show that the method is accurate both in evaluating the system throughput and the system yield. Moreover, the proposed method has been used for deriving new insights concerning the behaviour of the system. The main results are summarized in the following, in terms of practical rules for operating and designing similar systems.

- Lower work in progress does not always mean higher conforming parts production rate, since the behaviour strictly depends on the buffer capacity.
- The design of buffers should be performed by jointly considering their impact on quality and productivity.
- Local improvements (increasing the machine reliability) can have negative impact on the system level performance.
- Allocating higher number of inspection devices does not always mean higher conforming parts production rate.

Concerning the method proposed in this paper, future investigations will regard the analysis of the robustness of our approach to changes in the system assumptions. Different types of control charts, different distributions of random events and different processing times at different stages will be considered. Furthermore, the area of research that jointly addresses quality and productivity aspects is almost new and rich of topics which need further investigation. Future developments include the generalization of the method to deal with:

- unreliable buffers and buffer delays due to non zero traveling times of parts;
- scrapping of non-conforming parts and on-line rework of repairable products;
- multiple modes of out of controls of machines and dependency of features between different production stages;
- general production stopping rules, based on the number of non-conformities produced in sequence or other run-rules;
- degrading processes with continuous deterioration of part quality;
- quality control in complex systems, like multiple products and loop systems;
- impact of production control rules (Kanban, Basestock and Conwip) on quality.

In addition, the proposed performance evaluation method will be used for developing models for the optimal joint design of control charts, inspection point locations and buffer sizes.

Acknowledgements

The authors would like to thank prof. B.M. Colosimo and Prof. Q. Semeraro at the Department of Mechanical Engineering of Politecnico di Milano for their interesting comments and suggestions. The research is partially funded by the Roberto Rocca Project "*Interactions Among Quality and Productivity Performance Measures in Production Systems*" in collaboration with Dr. Stanley Gershwin and Dr. Irvin Schick from MIT Laboratory for Manufacturing and Productivity.

References

- Alwan, L.C., 2000. *Statistical Process Analysis*. McGraw Hill.
- Barhak, J., Djurdjanovic, D., Spicer, P. and Katz, R., 2005. Integration of Reconfigurable Inspection with Stream of Variations Methodology. *International Journal of Machine Tools and Manufacture*, 45, 407–419.
- Bassetto, S. and Siadat, A., 2009. Operational Methods for Improving Manufacturing Control Plans: Case Study in a Semiconductor Industry. *Journal of Intelligent Manufacturing*, 20, 55–65.
- Blais, F., 2003. A Review of 20 Years of Range Sensor Development. *Videometrics*, VII, SPIE 5013, 62–76.
- Blumenfeld, D.E. and Owen, J.H., 2008. Effects of Operating Speed on Production Quality and Throughput. *International Journal of Production Research*, 46, 7039–7056.
- Bulgak, A.A., 1992. Impact of Quality Improvement on Optimal Buffer Designs and Productivity in Automatic Assembly Systems. *Journal of Manufacturing Systems*, 11 (2), 124–136.
- Burman, M.H., 1995. New Results in Flow Lines Analysis. Thesis (PhD). Massachusetts Institute of Technology.
- Buzacott, J.A. and Hanifin, L.E., 1978. Models of Automatic Transfer Lines With Inventory Banks: a Review and Comparison. *IIE Transactions*, 10, 197–207.
- Cheng, C.H., Miltenburg, J. and Motwani, J., 2000. The Effect of Straight and U-shaped Lines on Quality. *IEEE Transactions On Engineering Management*, 47 (3), 321–334.
- Colledani, M., Ekvall, M., Moriggi, P., Polato, A., Lundholm, T. and Tolio, T., 2009. Analytical Methods to Support Continuous Improvements at Scania. *International Journal of Production Research*, on-line first doi: 10.1080/00207540802538039.
- Colledani, M., Matta, A. and Tolio, T., 2005. Performance Evaluation of Production Lines with Finite Buffer Capacity Producing Two Different Products. *OR Spectrum*, 27, 243–263.
- Colledani, M. and Tolio, T., 2005. A Decomposition Method to Support the Configuration/Reconfiguration of Production Systems. *CIRP Annals*, 54 (1), 441–444.
- Dallery, Y., David, R. and Xie, X.L., 1988. An Efficient Algorithm for Analysis of Transfer Lines with Unreliable Machines and Finite Buffers. *IEEE Transactions*, 20(3), 280–283.
- Dallery, Y. and Gershwin, S.B., 1992. Manufacturing Flow Line Systems: A Review of Models and Analytical Results. *Queueing Systems Theory and Applications*, 12, 3–94.
- Gershwin, S.B., 1987. An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking. *Operations Research*, 35, 291–305.
- Gershwin, S.B., 1994. *Manufacturing Systems Engineering*. Englewood Cliffs, New Jersey: Prentice Hall.
- 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.
- Gershwin, S.B. and Kim, J., 2005. Integrated Quality and Quantity Modeling of a Production Line. *OR Spectrum*, 287–314.
- Gershwin, S.B. and Kim, J., 2008. Analysis of Long Flow Lines With Quality and Operational Failures. *IIE Transactions*, 40 (3), 284–296.

- 1 Gershwin, S.B., Matta, A. and Tolio, T., 2002. Analysis of Two-Machine Lines
2 With Multiple Failure Modes. *IIE Transactions*, 34, 51–62.
- 3 Gershwin, S.B. and Schick, I.C., 2007. Taxonomy of Quality/Quantity Issues in
4 Manufacturing Systems. May 11–16., 6th Conference on the Analysis of Man-
5 ufacturing Systems, Lunteren, The Netherlands.
- 6 Gershwin, S.B. and Schor, J.E., 2000. Efficient Algorithms for Buffer Space Allo-
7 cation. *Annals of Operations Research*, 93, 117–144.
- 8 Gershwin, S.B. and Werner, L.M., 2003. An Approximate Analytical Method for
9 Evaluating the Performance of Closed Loop Flow Systems with Unreliable
10 Machines and Finite Buffers. *International Journal of Production Research*,
11 45 (14), 3085–3111.
- 12 Gershwin, S. and Burman, M., 2000. A Decomposition Method for Analysing In-
13 homogeneous Assembly/Disassembly Systems. *Annals of Operations Research*,
14 93, 91–116.
- 15 Helber, S., 1999. *Performance Analysis of Flow Lines with Non-Linear Flow of*
16 *Material*. Springer.
- 17 Hsu, L.F. and Tapiero, C.S., 1987. Quality Control of the M/M/1 Queue. *Interna-*
18 *tional Journal of Production Research*, 25 (3), 447–455.
- 19 Hsu, L.F. and Tapiero, C.S., 1987. Maintenance of an Unreliable M/G/1 Queue-
20 like Job Shop. *Queueing Systems: Theory and Applications*, 2, 333–350.
- 21 Hsu, L.F. and Tapiero, C.S., 1989. Quality Control of the M/G/1 Queue. *European*
22 *Journal of Operational Research*, 42 (1), 88–100.
- 23 Hsu, L.F. and Tapiero, C.S., 1992. Integration of Process Monitoring, Quality Con-
24 trol and Maintenance in an M/G/1 Queue-like Production System. *Interna-*
25 *tional Journal of Production Research*, 30 (10), 2363–2379.
- 26 Hu, S.J., Koren, Y. and Weber, T.W., 1998. Impact of Manufacturing System
27 Configuration on Performance. *CIRP Annals*, 47 (1), 369–372.
- 28 Inman, R., Blumenfeld, R., Huang, R. and Li, J., 2003. Designing Production
29 Systems for Quality: Research Opportunities from an Automotive Perspective.
30 *International Journal of Production Research*, 41 (9), 1953–1971.
- 31 Levantesi, R., Matta, A. and Tolio, T., 2003. Performance Evaluation of Continu-
32 ous Production Lines with Machines Having Different Processing Times and
33 Multiple Failure Modes. *Performance Evaluation*, 51 (2–4), 247–268.
- 34 Li, J., 2003. Modelling and Analysis of Complex Production Systems. Fifth Aegean
35 International Conference on Analysis of Manufacturing Systems, 203–212.
- 36 Li, J., 2005. Comparative Study of Push and Pull Systems Considering Quality
37 Performance in a Cell-based Job Shop Environment. *International Journal of*
38 *Production Research*, 43 (23), 4999–5017.
- 39 Li, J., Blumenfeld, D.E. and Marin, S.P., 2007a. Manufacturing System Design
40 to Improve Quality Buy Rate: an Automotive Paint Shop Application Study.
41 *IEEE Transactions on Automation Science and Engineering*, 4 (1), 75–79.
- 42 Li, J., Blumenfeld, D.E. and Marin, S.P., 2007b. Production System Design for
43 Quality Robustness: Theory and Application in Automotive Paint Shops. *IIE*
44 *Transactions*, 39.
- 45 Liberopoulos, G., Kozanidis, G. and Tsarouhas, P., 2007. Performance Evaluation
46 of an Automatic Transfer Line with WIP Scrapping During Long Failures.
47 *Manufacturing and Service Operations Management*, 9 (1), 62–83.
- 48 Matta, A. and Tolio, T., 1998. A Method for Performance Evaluation of Automated
49 Flow lines. *CIRP Annals*, 47 (1), 373–376.
- 50 Meester, L.E. and Shantikumar, J.G., 1990. Concavity of Throughput of Tandem
51 Queueing Systems with Finite Buffer Storage Capacity. *Advances in Applied*
52 *Probability*, 22 (1), 764–767.

- Montgomery, D., 2005. *Introduction to Quality Control*. 5th John Wiley & Sons Inc.
- Nonaka, Y., Lengyel, A., Ono, M. and Sugimoto, K., 2008. Correlation Analysis of TSUNAMI Phenomena and Failure Rate Fluctuation in Manufacturing System. Proceedings of the 17th IFAC World Congress, Seoul, Korea, July 6-11.
- Pande, P. and Holpp, L., 2002. *What is six sigma?*. McGraw-Hill.
- Prajogo, F.I. and Sohal, A.S., 2001. TQM and Innovation: a Literature Review and Research Framework. *Technovation*, 21(9), 539–558.
- Schippers, W.A.J., 1998. Applicability of Statistical Process Control Techniques. *International Journal of Production Economics*, 56 (7), 525–535.
- Shingo, S. and Dillon, A.P., 1989. *A Study of the Toyota Production System from an Industrial Engineering Viewpoint*. Productivity Press.
- Toyota, 1996. *The Toyota Production System*. .
- Woodall, W.H., 2000. Controversies and Contradictions in Statistical Process Control. *Journal of Quality Technology*, 32, 341–378.

Appendix A - The Algorithm

Phase 1 - Initialization of subsystems. Creation of the pseudo-machines structures and assignment of known probabilities.

For $i = 1, \dots, K - 1$, consider the subsystem $l(i)$. Set $N(i) = N_i$. Set all the failure and repair probabilities of the local failure modes of the pseudo-machine $M^u(i)$ using equation 12. Set all the repair probabilities of the remote failures modes of the pseudo-machine $M^u(i)$ using equations 14. If M_i is a machining station subject to out of controls then set the repair probabilities of the failures modes related to quality control of the pseudo-machine $M^u(i)$ using equations 13. Set the unknown failure probabilities for these modes to a given small value $\epsilon = 0.05$.

For $i = K - 1, \dots, 1$, repeat the same procedure for initializing the parameters of the downstream pseudo-machines $M^d(i)$.

For $i = 1, \dots, K - 1$, calculate all the performance measures of subsystem $l(i)$ by using the exact method in [Gershwin *et al.* (2002)] and calculate the transition to starvation and blocking states by using equations 17 and 18.

Phase 2 - Upstream pseudo-machines analysis. For $i = 1, \dots, K - 1$:

Step 1 - MLD

- Calculate quality related transitions of the Markov chain in Figure 5(a) for the machine M_i , by using quality link equations 3, 4 and 19;
- Solve the Markov chain in Figure 5(a) and calculate the state probabilities for the Markov chain in Figure 5(b) by using state aggregation equations from 20 to 24.
- Compute the machine yield of M_i by using equation 8 if M_i is locally monitored or 11 in case M_i is remotely monitored.
- Calculate the unknown failure probabilities to be transferred to the BLD by using equations 25, 26 and 27.

Step 2 - BLD

- Compute the performance of the Building Block $l(i)$. Calculate the throughput and the average buffer levels by using the method [Gershwin *et al.* (2002)]. Calculate the subsystem Lead Time by using 16.
- Update the transition to starvation and blocking states to be transferred to the MLD by using equations 17 and 18.

Phase 3 - Downstream pseudo-machines analysis For $i = K - 1, \dots, 1$:
Perform Step 1 and 2, considering the machine M_{i+1} in the MLD and the pseudo-machine $M^d(i)$ in the BLD.
Applying alternately Phase 2 and Phase 3 of the algorithm the performance of the line can be obtained. The algorithm stops when no significant changes in the unknown parameters are observed.

For Peer Review Only