OR Spectrum manuscript No.
(will be inserted by the editor)

Analytical Evaluation of the Output Variability in Production Systems with General Markovian Structure

Ramiz Assaf · Marcello Colledani · Andrea Matta

Received: date / Accepted: date

Abstract Performance evaluation models are used by companies to design, adapt, manage and control their production systems. In the literature, most of the effort has been dedicated to the development of efficient methodologies to estimate the first moment performance measures of production systems, such as the expected production rate, the buffer levels and the mean completion time. However, there is industrial evidence that the variability of the production output may drastically impact on the capability of managing the system operations, causing the observed system performance to be highly different from what expected. This paper presents a general methodology to analyze the variability of the output of unreliable single machines and small scale multi-stage production systems modeled as General Markovian structure. The generality of the approach allows modeling and studying performance measures such as the variance of the cumulated output and the variance of the inter-departure time under many system configurations within a unique framework. The proposed method is based on the characterization of the autocorrelation structure of the system output. The impact of different system parameters on the output variability is investigated and characterized. Moreover, managerial actions that allow reducing the output variability are identified. The computational complexity of the method is studied on an extensive set of computer experiments. Finally the limits of this approach while studying long multi-stage production lines are highlighted.

Keywords Output Variability · Production Systems · Performance Evaluation · Production variance · Markov chains

1 Introduction

1.1 Motivation

Manufacturing systems engineering methods have been developed in the last decades for investigating the dynamic behavior of manufacturing systems, for estimating their performance and

Ramiz Assaf
Industrial Engineering Department, An-Najah National University, Nablus, Palestine
E-mail: ramizassaf@najah.edu

Marcello Colledani · Andrea Matta
Department of Mechanical Engineering, Politecnico di Milano
via la Masa 1 , 20155 Milano, Italy

for supporting their efficient design, management, improvement and reconfiguration. The most commonly adopted techniques to predict production system performance are simulation and analytical methods. The main advantage of the latter is the ability of rapidly estimating the main performance measures of the system. Moreover, analytical methods allow the user to deeply understand the dynamics of the system behavior, since the relations among the system variables are expressed through equations. During the system configuration/reconfiguration phase [35], these tools are used to select system solutions that profitably exploit the trade-offs between these first order performance measures.

Higher order performance measures are generally difficult to analyze and are rarely considered. However, in the presence of random events and disturbances in the production, higher order performance measures are relevant to correctly predict the system output. Indeed, due to the production variability, the observed performance can be highly different from the average performance. Output variability makes difficult to meet customer orders on time and to ensure the required service level of the system. This problem may directly corrupt the profitability of those systems designed only by considering the mean performance of the system, that are not robust to disturbances. Indeed, low output variance indicates stability of the output of the production line, less unforeseen delays and small fraction of escaped orders, which translates to lower costs. Symmetrically, high variance means instability of the output, i.e. significant differences in production quantity observed on a daily basis. Typical sources of variability in the production system behavior are random failure occurrences and durations. A real case in the automotive sector [8] reports that the weekly output of the production system composed of 22 machines and affected by the occurrences of 144 different failure modes has a coefficient of variation, estimated from the available field data of three months, equal to 0.263. Thus, it is highly probable that the weekly demand will not be met if the system is designed only considering its average performance. Similar data and considerations were given by Gershwin, who showed by simulation that the standard deviation of the weekly production can be over 10% of its mean [12].

1.2 Literature survey

In spite of the relevance of production variability in industry, the number of papers discussing the variability of the output in production lines is fairly limited if compared to the papers on the prediction of first order performance measures of manufacturing systems. Moreover, the underlying assumptions of the available methods are over-simplistic, thus preventing their wide application in industry. Research contributions on production variability deal with both the cumulated production of a transfer line and the interdeparture times of the output process in a time interval.

The output variability of production lines was first studied by Miltenburg [23], who proposed an exact numerical method to calculate the first two moments of the asymptotic measures of the output, i.e. the throughput and the asymptotic variance rate defined as the limiting variance of the output per time unit. The method considered small buffered production lines featuring unreliable machines with geometric/exponential failure and repair times. His approach is based on the state-space representation of the system and the use of the inverse of the fundamental matrix. Since the computational complexity of this method depends on the number of states modeling the system, only simple systems with small number of machines and buffer capacities can be analyzed with success.

Hendricks [16] presented an approach, based on the structural properties of Markov chains, to estimate the asymptotic variance rate of interdeparture times in production lines with exponential processing times, perfectly reliable machines and finite buffer capacities. This work was later

extended [17] to model machines with general processing times. He provided interesting insights on the role of the output autocorrelation structure and the skewness of processing times on the variance of the inter-departure. In particular, it was observed by simulation that, by increasing the skewness of the processing times, the inter-departure variance also increases. The complexity of Hendrick's approach is comparable to Miltenburg's [23], being dependent on the number of states representing the system.

Tan made a series of studies on the output variability of production systems. The works include the analysis and calculation of the output variability for machines in isolation, multi-stage unbuffered lines, production lines with parallel and series machines and small buffered manufacturing systems. He proposed both continuous time models [27, 30, 29] and discrete time models [31–33] for the analysis. In terms of investigated machine models, the studies include reliable machines with exponential processing times [27], unreliable machines with a single failure mode featuring geometrically or exponentially distributed failure times [32] and unreliable machines with Coxian distributed repair time [30]. Performance measures discussed include the asymptotic variance rate of the output and the service level of the system.

Concerning the analysis of multi-stage systems, Tan proposed a matrix-geometric method for the estimation of the asymptotic variance rate in two-machine lines with single failure mode machines and a finite buffer. Compared to Miltenburg's approach, the method proposed by Tan is more efficient in terms of number of executed floating point operations. Tan used the same approach for evaluating the variance of two-stage production lines with single failure mode machines as a function of time, again by exploiting the special structure of the transition matrix [31]. The method uses the Grassman approach [14] to iteratively obtain the performance of interest. The complexity of the adopted procedure depends both on the length of the observed time period and on the size of the Markov chain describing the process, thus on the buffer capacity. Tan [33] increases the computational efficiency of his algorithm allowing evaluating the performance of multi-stage production lines with unreliable machines and finite buffer capacity by using an exact procedure. Moreover, he studies the variability of the output for production lines controlled by different policies such as Kanban, Basestock and CONWIP.

Ou and Gershwin [24] obtained closed form expressions of the variance of the lead time in a two-machine line in which machines may fail in a single mode. Gershwin proposed a method for the calculation of the variance of the output of a single machine with a single failure mode in closed form [12]. His method is based on the solution of the difference equation describing the system dynamics. The developed method is then used to approximate the performance of production lines through a decomposition approach. The effect of previous stages on the last machine in the system is considered by adjusting the failure and repair parameters of the single machine model. However, the method is shown to have large errors in the variance estimation (around 20% compared to simulation results) since the adopted decomposition equations [13] did not capture and propagate the output variability throughout the line. Carrascosa [4] extended the method of Gershwin to the case of the isolated machine with multiple failure modes.

Li and Meerkov [19] studied the variance of the output for production lines composed of unreliable machines and finite buffers. The most limiting assumption to the application of their method is the Bernoulli reliability model, which assumes repair time equal to the cycle time of the machines. The authors focused on the "due time performance" which is an equivalent measure of the service level.

Recently, more complex machine models providing insights about the transient behavior of the system have been studied in depth [20] and [22]. Other works that studied the transient behavior include [11] and [5]. In fact, the work of Dincer and Deler [11] studied both the transient and steady-state variability in the output of small buffered lines with reliable machines featuring exponentially distributed processing times, by adopting n-fold convolution of the inter-arrival

and the processing time distributions. Chen and Yuan [5] focused on the system output mean and variance during the transient period. The approach models long unbuffered production lines with unreliable machines subject to a single failure mode, with exponentially distributed failure and repair times, by using a sample path method.

Ciprut et al [6] used a fluid Markovian model to derive an exact closed-form formula to calculate the first two moments of the asymptotic output for unreliable machines with generally distributed up and down times. An attempt to extend this approach to two-machine one buffer dipoles was made, by approximating the dipole behavior with an equivalent single machine having two switching operational modes. When the second machine is not starved, the equivalent machine is exactly identical to the second machine of the dipole; when the second machine is starved, the equivalent machine behaves as the first machine in the dipole. However, the autocorrelation structure of the starvation times was not considered, thus this approximation may perform poorly in specific configurations. The exact analysis in Ciprut et al [6] was recently extended by Angius et al [1] to handle any system modeled as continuous and discrete time reward models, including machines with general Markovian structure.
Recently, other approximate methods for the analysis of the output variance in long multi-stage lines were introduced. He et al [15] studied serial buffered multi-stage systems, with reliable machines featuring exponential processing times. The approximate method relied on the exact Markovian arrival process analysis of a simplified two-station one buffer sub-system and a compression method that propagates the output variance along subsystems. The difference with decomposition approaches is that the compression (also called aggregation) algorithm does not iterate backwards, form the last subsystem to the first. The system behavior is analyzed and it is shown that the variance of the output always increases with the buffer capacity, for this type of systems. However, no estimation of the method accuracy towards simulation is given in this paper. Another approximate method was proposed by Manitz and Tempelmeier [21], who studied long assembly lines with finite buffers and general service times. Their approach used a two-moment approximation to estimate the output variability in the assembly line, by measuring the coefficient of variation of the inter-departure time.
Other works that studied the inter-departure time variability included [25, 18, 21, 2]. Sabuncuoglu et al [25] studied the effect of different factors of the assembly system on its throughput and inter-departure time variability. Kalir and Sarin [18] proposed a method for reducing inter-departure time in production systems using simulation. Betterton and Silver [2] proposed a method that uses the inter-departure time variance to detect bottlenecks in open asynchronous serial production lines with finite buffers.

Finally, the effect of the autocorrelation structure proposed by Hendricks [16] have been further investigated by Colledani et al [7], for small systems featuring unreliable machines affected by multiple failure modes. They also managed to evaluate approximately the asymptotic variance rate of multi-stage production lines with machines having multiple geometric failure modes [9]. The proposed decomposition method suffers the same limitations of the decomposition method proposed in Gershwin [12]. In summary, the methods available in the literature focus on specific systems and most of them are characterized by exponential or geometrical distributions. Those methods dealing with general systems estimate one single performance measure of the output variability without providing an extensive analysis of the different dimensions with which variability emerges in manufacturing systems. This paper aims at filling this gap by proposing an approach to study the variability of general manufacturing systems in a common and comprehensive framework.

1.3 Contribution

The main contribution of this paper is threefold. Firstly, a general approach to analyze and predict the first two moments of the main output performance measures in manufacturing systems modeled by general Markovian structures is proposed. The approach, based on the autocorrelation structure of the Markovian system, allows to study within a unique framework different performance measures such as the variance of the cumulated output, the variance of the interdeparture time and the transient period of the system, in different system configurations. In the case of unreliable machines with a single reward state, an interesting relationship existing between the variances of the cumulated production and the interdeparture time is derived. This last measure of variability in manufacturing systems with unreliable machines was not investigated in the literature yet.

Secondly, the paper proposes a new approximation to estimate the variance of the cumulated production. Results show the proposed method is accurate enough to estimate other parameters such as, for example, the service level. Thirdly, the impact of the main system and machine parameters on the variability of the output is also investigated, with the objective of deriving insights and new system design and management rules for reducing the variability and meeting the due-time performance of the system. Indeed, very little is known on how to manage production systems to reduce the variability of their output. Important questions like "What is the due date to be quoted for a given order?" and "What is the probability of delivering a given order on time, under a particular system configuration?" still remain unsolved. In this paper, particular attention is given to two issues that do not find clear explanation and, consequently, management guidelines in the available literature. Firstly, the impact of failure and repair times and their distributions on the asymptotic output mean and variance of isolated machines is thoroughly investigated, showing that there are several counterintuitive effects generating chances for production managers to increase the output stability (reduce the output variance) at the cost of slightly reducing the mean production rate. Secondly, the impact of buffers on the output variability of multi-stage systems is investigated in details, providing an explanation to a complex interaction already observed but not clearly explained in the literature [4, 32].

The approach has been applied to single failure and multiple failure isolated machines, with Bernoulli, geometric and generally distributed failure and repair times, as well as applications to buffered two-stage and multi-stage serial systems are discussed in this paper.

1.4 Paper Organization

The reminder of the paper is organized as follows: Section 2 defines the different output variability measures of interest for this paper. Section 3 describes the theory and the methodology developed for analyzing the output variability of general Markovian systems. Section 4 presents the application of the general approach to different isolated machine models. Section 5 presents the application of the general approach to multi-stage systems. Section 6 presents numerical results discussing possible variability reduction strategies for different systems. Finally, Section 7 summarizes the main results of this work, draws the conclusions and discusses the future extensions of this work.

2 Measures of output variability

An output variable of interest for this paper is the total amount of parts produced by the system during the time period $[1, t]$. This variable, denoted with Z_t, is random and nonnegative

with expected value $\mathbb{E}[Z_t]$ and variance $var[Z_t]$, both of them dependent on the time period of evaluation as they are increasing functions in t.

For large values of t (i.e., that approaches *infinity*), Z_t is approximately normally distributed with a mean $e \cdot t$ and a variance $v \cdot t$ [23, 30], where e is the mean production rate or the mean throughput of the system, and v is the asymptotic variance rate:

$$e \hat{=} \lim_{t \to \infty} \frac{\mathbb{E}[Z_t]}{t}$$

$$v \hat{=} \lim_{t \to \infty} \frac{var[Z_t]}{t}$$

Performance measures e and v do not depend on the time, so they can be used as output characteristic parameters during the system design phase. The coefficient of variation of Z_t can be approximated using e and v as:

$$cv[Z_t] \approx \frac{\sqrt{v}}{e\sqrt{t}}$$

which approaches *zero* as t approaches infinity, i.e., as the time increases the dispersion of Z_t around the mean decreases. This measure is of practical value for evaluating the uncertainty of the production in a defined time window.

Another measure of the relative output variability is the asymptotic index of dispersion [1]:

$$d = \frac{v}{e}$$

Finally, approximating Z_t with a normal distribution allows calculating the system service level (SL), defined as the probability to meet a certain customer order (composed of x parts) within a certain time t [30]:

$$SL(x,t) \hat{=} P[Z_t \geq x] \approx 1 - \Phi\left(\frac{x - e \cdot t}{\sqrt{v \cdot t}}\right)$$

where $\Phi(\cdot)$ is the cumulative normal distribution function.

Another measure of interest is the interdeparture time, which is defined as the amount of time between two consecutive departures from the system; the interdeparture time is denoted with IDT in this paper.

3 Output variability of general Markovian systems

3.1 Assumptions

In this section, we consider a discrete time system with s different states and an underlying transition probability matrix $\mathbf{P}$. The system has a constant processing cycle time. Time is scaled so that the processing cycle time is one time unit. The system is characterized by up states in the set U, down states in the set D, and transitions among all possible states. Transitions from state j_1 to j_2 occurring with probability (p_{j_1,j_2}) follow the geometric distribution with a mean $(1/p_{j_1,j_2})$. By convention, transitions can happen only at the beginning of a time unit. According to the type of states (operational or down) the elements of the transition probability matrix $\mathbf{P}$ connect, four partitions can be generated, namely $\mathbf{P}_{U,U}$, $\mathbf{P}_{U,D}$, $\mathbf{P}_{D,U}$, and $\mathbf{P}_{D,D}$:

$$P = \begin{bmatrix} \mathbf{P}_{U,U} & \mathbf{P}_{U,D} \\ \mathbf{P}_{D,U} & \mathbf{P}_{D,D} \end{bmatrix}$$

The system has a binary reward column vector $\boldsymbol{\mu}_{s\times 1}$ that governs its output. The reward μ_j assumes the value 1 if j is a productive state, and 0 otherwise, with $j = 1, \ldots, s$.

In the remainder of this section the mathematical derivation of the output variability measures is described. Specifically, in paragraph 3.2 the mathematical analysis points out the autocorrelation structure of the production output and provides an approximate evaluation of the variability measure by series truncation. Exact closed-form formulas in matrix form are instead proposed in paragraph 3.3.

3.2 Analysis of the autocorrelated output process

The system output is a binary random variable Y_i assuming the value *one* if the system produces a piece in period i and *zero* otherwise. The total production output Z_t of the system at time t is defined by the sum of the single outputs:

$$Z_t \hat{=} \sum_{i=1}^{t} Y_i$$

The variance of Z_t is defined as:

$$var\left[Z_t\right] \hat{=} \sum_{i=1}^{t} var\left[Y_i\right] + 2\sum_{i=1}^{t}\sum_{l=i+1}^{t} cov\left[Y_i, Y_l\right] \tag{1}$$

which is the sum of two different components [16]. The first component is related to the variance of the single random variables Y_i, while the second component arises when the series Y_i are not independent but timely autocorrelated. Since we are interested in calculating the steady state performance, we assume that the output process is stationary at the beginning of the analyzed time interval. Thus, equation (1) can be rewritten as:

$$var\left[Z_t\right] = t\sigma_Y^2 + 2\sum_{k=1}^{t-1}(t-k)cov_k\left[Y\right] \tag{2}$$

where Y is the random variable of the stationary output process, and $cov_k[Y]$ is the autocovariance of lag k of the time series Y.

In order to apply equation (2), it is necessary to know the variance and autocovariances of the process output in steady state, i.e. to calculate σ_Y^2 and $cov_k[Y]$. The first is calculated as:

$$\sigma_Y^2 \hat{=} \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = e - e^2 \tag{3}$$

where e is mean production rate calculated as follows:

$$e = \sum_{j=1}^{s} \pi_j \mu_j$$

with π_j the steady state probability of being in state j. Indeed, since Y is binary $\mathbb{E}[Y^2]$ is equal to $\mathbb{E}[Y]$. The interpretation of σ_Y^2 is straightforward: if the system is observed in steady-state n times independently, the variance of the observed Y values tends to σ_Y^2 as $n \to \infty$.

By definition, the autocovariance of Y of lag k is:

$$cov_k[Y] \hat{=} \mathbb{E}[Y_i Y_{i+k}] - \mathbb{E}[Y_i]\mathbb{E}[Y_{i+k}] \tag{4}$$

Again, since Y is binary, $\mathbb{E}[Y_i Y_{i+k}]$ reduces to the probability that the system is operational both at periods i and $i+k$. Therefore, after some manipulations, expression (4) becomes:

$$cov_k[Y] = \sum_{j=1}^{s}\sum_{g=1}^{s} \pi_j \mu_j \mathbf{P}^k_{j,g} \mu_g - e^2 \tag{5}$$

where $\mathbf{P}^k_{j,g}$ is the probability of going from state j to sate g after k steps. It can be noticed that, as the lag k increases, the autocovariance approaches zero [26].

The spectral decomposition of the Perron-Frobenius theorem can be used to formulate the $\mathbf{P}^k$ matrix by means of the eigenvalues and eigenvectors of $\mathbf{P}$:

$$\mathbf{P}^k = \sum_{j=1}^{s} \lambda_j^k \eta_j u_j' \tag{6}$$

where $\lambda_1, ..., \lambda_s$ are the s distinct eigenvalues of $\mathbf{P}$, $u_1, ..., u_s$ and $\eta_1, ..., \eta_s$ the associated sequences of left and right eigenvectors respectively such that $u_r' \eta_j = 0$ if $r \neq j$ and $u_r' \eta_j = 1$ for all $r, j = 1, ..., s$[3]. As a consequence, the autocovariance $cov_k[Y]$ is a function of the eigenvalues of matrix $\mathbf{P}$.

By definition, the autocorrelation function of lag k is:

$$\rho_k[Y] \hat{=} \frac{cov_k[Y]}{\sigma_Y^2} \tag{7}$$

Substituting equations (3), (5) and (7) into (2) the following equation can be derived:

$$var[Z_t] = \sigma_Y^2 \left[t + 2 \sum_{k=1}^{t-1} (t-k) \rho_k[Y] \right] \tag{8}$$

which is an exact formula for calculating the total output variance in a time period $[1, t]$. For small values of t this formula can be directly applied. For large values of t, the above formula becomes not practical; however it is possible to identify a value of k after which ρ_k approaches *zero* and the previous series can be truncated, resulting into a simplified approximate version of the previous equation:

$$var[Z_t] \approx t\sigma_Y^2 + 2\sigma_Y^2 \sum_{k=1}^{k^*} (t-k) \rho_k[Y] \tag{9}$$

where $k^* < t$ is the number of significant lags. The parameter k^* directly depends on the second largest eigenvalue of the matrix $\mathbf{P}$ and it can be estimated by the following equation [26]:

$$k^* = \frac{\log \epsilon}{\log \lambda_2}$$

where ϵ is the desired tolerance in the numerical calculation of ρ_k by using the power method. The higher the value of ϵ, the higher the level of the approximation introduced in the calculation of the output variance. Thus, the parameter k^* represents the number of autocorrelation lags that are considered as significant, at tolerance level ϵ. More formally, k^* is the minimum number of lags such that the sum of the autocorrelation coefficients for $k = k^* + 1, .., +\infty$ is less than ϵ. This increases as the size of $\mathbf{P}$ also increases, indicating that more complex systems will show more complex output autocorrelation structures. As known in literature, the second largest eigenvalue also affects the transient behavior of the system [22, 10]. Therefore, higher values of the second

largest eigenvalue of the transition probability matrix $\mathbf{P}$ result both in longer system settling time in the transient period and more significant output autocorrelation lags.

Equation(9) can be rewritten as:

$$var[Z_t] \approx t\sigma_Y^2 \left(1 + 2\rho_{total}(\epsilon)\right) - 2\sigma_Y^2 \sum_{k=1}^{k^*}(k \cdot \rho_k) \tag{10}$$

where $\rho_{total}(\epsilon) = \sum_{k=1}^{k^*} \rho_k$ is the total significant autocorrelation at tolerance level ϵ. The other performance measures v, $cv[Z_t]$, and d can be calculated as:

$$\begin{aligned} v &\approx \sigma_Y^2(1 + 2\rho_{total}(\epsilon)) \\ cv[Z_t] &\approx \sqrt{\frac{(1-e)}{e \cdot t}(1 + 2\rho_{total}(\epsilon))} \\ d &\approx (1-e)(1 + 2\rho_{total}(\epsilon)) \end{aligned} \tag{11}$$

This result is general and not limited to specific assumptions on the production system. Furthermore, it is also in accordance with the result of Hendricks [16] for a simplified machine model. This analysis can be used to approximate the output variance measures at a given level of tolerance ϵ. However, since the absolute values of $|\rho_k| \leq 1$ for each k, the geometric series is always convergent to a sum. By exploiting the mathematical derivation of the sum of the series ρ_k and $k\rho_k$ from 1 to t, exact closed-form formulas for the variance of the cumulated output and the variance of the inter-departure time can be derived as described in the following section.

3.3 Exact closed-form expressions for the output variance

Starting from the result of Equation (2) a general expression to calculate the variance of Z_t is reported in the following Theorem.

Theorem 1 *Given a production system represented by the transition probability matrix $\boldsymbol{P}$ and reward vector $\boldsymbol{\mu}$, the variance of the cumulated production is:*

$$var\left[Z_t\right] = t\alpha + \beta(t) \tag{12}$$

where:

$$\alpha = e(1-3e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\boldsymbol{P}\boldsymbol{Z}\boldsymbol{\mu} \tag{13}$$

$$\beta(t) = 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\boldsymbol{P}^{t+1} - \boldsymbol{P}\right)\boldsymbol{Z}^2\boldsymbol{\mu} \tag{14}$$

μ_{diag} is a diagonal matrix with the rewards in the diagonal and $\boldsymbol{Z}$ is the Fundamental Matrix:

$$\boldsymbol{Z} = (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{A})^{-1} \tag{15}$$

Proof See Appendix A.

For t large enough (i.e., $t > k^*$) the term $\beta(t)$ tends to β, expressed as:

$$\beta = 2e^2 - 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}\mathbf{Z}^2\boldsymbol{\mu} \tag{16}$$

and the variance becomes a function linearly increasing in time with slope equal to α. Thus, the variance rate is simply:

$$v = \alpha = e(1-3e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}\mathbf{Z}\boldsymbol{\mu} \tag{17}$$

By comparing equation (11) and equation (13), the total autocorrelation term can be expressed in exact terms as:

$$\rho_{total}(\epsilon = 0) = \frac{\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}\mathbf{Z}\boldsymbol{\mu} - e^2}{e(1-e)} \tag{18}$$

while closed-form expressions for $cv[Z_t]^2$, and d are reported on Table 1.

Under the same system assumptions and using the transition probability matrix partitions proposed in section 3, a general expression to calculate the variance of IDT is given in the following Theorem.

Theorem 2 *Given a production system represented by the transition probability matrix* $\boldsymbol{P}$ *and reward vector* $\boldsymbol{\mu}$*, the variance of the inter–departure time is:*

$$var[IDT] = \frac{e-1}{e^2} + \frac{1}{e}\boldsymbol{\pi}_U \boldsymbol{P}_{U,D} 2\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{P}_{D,D})^{-3}\boldsymbol{P}_{D,U}\boldsymbol{\mu}_U \tag{19}$$

Proof See Appendix B.

4 Application to isolated machines

In this section we analyze the output variability measures for some specific isolated machine models, namely, the single failure mode machine model [23, 12], with geometric and generally distributed repair times, the Bernoulli machine [19], and the multiple failure modes geometric machine [9]. For these simple systems, closed form solutions can be derived for the output variability measures by using the proposed approach.

4.1 Single failure machine

A widely analyzed case in the literature is the isolated machine with single failure mode. In the geometric failure and repair time case, this machine can be either up (operational) or down (failed) in a single mode [12]. While operational the machine can fail with probability p at the beginning of the time unit. While failed, it can be restored with probability r. The mean production rate of the machine is:

$$e = \frac{r}{p+r}$$

The expected cumulated production is:

$$\mathbb{E}[Z_t] = \sum_{t=1}^{T} \mathbb{E}[Y_t] = t\frac{r}{r+p}$$

The autocovariance has a special form, as equation (5) simply becomes:

$$cov_k[Y] = e(1-e)(1-p-r)^k \tag{20}$$

Equation (6) becomes:

$$\mathbf{P}^k = \frac{1}{r+p}\begin{bmatrix} r & p \\ r & p \end{bmatrix} + \frac{(1-r-p)^k}{r+p}\begin{bmatrix} p & -p \\ -r & r \end{bmatrix}$$

where $\lambda_2 = 1 - r - p$ is the second largest eigenvalue of $\mathbf{P}$. In this case only, the autocorrelation function of lag k (i.e., $\rho_k[Y]$) coincides with λ_2 to the power k (i.e., λ_2^k), and it is equal to:

$$\rho_k[Y] = \frac{cov_k[Y]}{\sigma_Y^2} = (1 - p - r)^k = \rho^k \tag{21}$$

where $\rho = 1 - p - r$. Substituting equations (3), (20) and (21) into (2) and after some manipulations:

$$var[Z_t] = e(1-e)\left[t + 2\sum_{k=1}^{t-1}(t-k)\rho^k\right] = e(1-e)\left[\frac{t - t\rho^2 - 2\rho + 2\rho^{t+1}}{(1-\rho)^2}\right] \tag{22}$$

This result was first derived in [4] and can alternatively be written as:

$$var[Z_t] = \sigma_Y^2 \cdot \left[\frac{t - t\rho^2 - 2\rho + 2\rho^{t+1}}{(1-\rho)^2}\right] \tag{23}$$

The other output variability measures can be easily calculated as a function of e and ρ, as reported in Table 1.

By using equation (19), the variance of the inter-departure time can be expressed and a simple relation between the variance of the inter-departure time, the asymptotic variance rate and the throughput can be found:

$$var[IDT] = \frac{e-1}{e^2} + \frac{2p}{r^2} = \frac{v}{e^3} \tag{24}$$

The output variability analysis can be reversely used to match the first two moments of the output of a complex manufacturing system with a geometric single failure machine model. Let us assume that a complex manufacturing system produces parts with a throughput e and a cumulative autocorrelation coefficient of the output process ρ_{total}. The parameters p_{eq} and r_{eq} of the geometric single failure mode machine matching the same first two asymptotic moments of the output can be obtained with the following equations:

$$\begin{cases} p_{eq} = \frac{1-e}{\rho_{total}} \\ r_{eq} = \frac{e}{\rho_{total}} \end{cases} \tag{25}$$

Equations (25) can be used to find an equivalent geometric machine on the basis of the estimates for e and ρ from real field data. Alternatively, if v instead of ρ is known or estimated from field data, the parameters p_{eq} and r_{eq} can be obtained with the following equations:

$$\begin{cases} p_{eq} = \frac{2e(1-e)^3}{v} \\ r_{eq} = p_{eq}\frac{e}{1-e} \end{cases} \tag{26}$$

This equivalent machine can be used in a decomposition approach for the approximate evaluation of the performance of multi–stage production lines with general machine behavior. Moreover, this reverse analysis could also be used for propagating both the first and second asymptotic moments of the output between the different subsystems within a new decomposition technique to analyze long production lines. These extensions will be subject of future research activities.

Table 1: Formula for the calculation of output variability measures in the analyzed cases

Variability measure	General structure	Single failure machine	Bernoulli machine	Multiple failure machine
$Var[Z_t]$	$te(1-3e)+2t\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu}+$ $+2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\mathbf{P}^{t+1}-\mathbf{P}\right)\mathbf{Z}^2\boldsymbol{\mu}$	$\sigma_Y^2\cdot\left[\frac{t-t\rho^2-2\rho+2\rho^{t+1}}{(1-\rho)^2}\right]$	$(1-p)p\cdot t$	$te(1-3e)+2t\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu}+$ $+2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\mathbf{P}^{t+1}-\mathbf{P}\right)\mathbf{Z}^2\boldsymbol{\mu}$
v	$e(1-3e)+2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu}$	$e(1-e)\frac{1+\rho}{1-\rho}$	$p(1-p)$	$\left[\sum_{j=1}^{f}I_j\left(\frac{2-r_j}{r_j}\right)-\left(\sum_{j=1}^{f}I_j\right)^2\right]e^3$
d	$(1-3e)+\frac{2}{e}\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu}$	$\frac{(1-e)(1+\rho)}{1-\rho}$	$1-e$	$\frac{v}{e}$
$Var[IDT]$	$\frac{e-1}{e^2}+\frac{1}{e}\boldsymbol{\pi}_U\mathbf{P}_{U,D}\frac{2\mathbf{I}}{\left(\mathbf{I}-\mathbf{P}_{D,D}\right)^3}\mathbf{P}_{D,U}\boldsymbol{\mu}_U$	$\frac{v}{e^3}$	$\frac{v}{e^3}$	$\frac{v}{e^3}$
$cv^2[IDT]$	$e-1+e\boldsymbol{\pi}_U\mathbf{P}_{U,D}\frac{2\mathbf{I}}{\left(\mathbf{I}-\mathbf{P}_{D,D}\right)^3}\mathbf{P}_{D,U}\boldsymbol{\mu}_U$	$\frac{v}{e}$	$1-e$	$d\ \frac{v}{e}$

4.2 Bernoulli Machine

The single failure mode geometric machine can be reduced to the Bernoulli machine when $p + r = 1$ [19]. The mean production rate for the Bernoulli machine is:

$$e = r = 1 - p$$

the autocorrelation among the rewards of the Bernoulli machine does not exist because the random variables Y_k are independent. Therefore, the performance indicators related to the cumulated production can be calculated by setting $\rho = 1 - p - r = 0$. Figure 4 shows the linear behavior of e vs p, in the Bernoulli case. The expected valued of Z_t is simply calculated as:

$$\mathbb{E}[Z_t] = t \cdot e = t \cdot r = t(1-p)$$

and the $var[Z_t]$ becomes:

$$var[Z_t] = r \cdot p \cdot t = (1-p)p \cdot t$$

The other output variability measures are reported on Table 1.

4.3 Multiple failure mode machine

The proposed method can be used for evaluating the output variability of an isolated multiple failure modes machine [9]. In this case, the machine can be up (operational) in one mode or down (failed) in different modes $1, \dots f$ with failure and repair probabilities equal to $p_1, \dots p_f$ and $r_1, \dots r_f$, respectively. When operational, the machine can fail at the beginning of the time unit in one of its failure modes j, with probability p_j. Failure modes are mutually exclusive, in the sense that the machine cannot be down in two different failure modes at the same time. When down in mode j, the machine can be restored into the operational condition at the beginning of the time unit with probability r_j. For each failure mode j, the unavailability factor $I_j = p_j/r_j$ can be defined. The transition probability matrix $\mathbf{P}$ for a multiple failure mode machine with geometric times to repair is:

$$\mathbf{P} = \begin{bmatrix} 1 - \sum_{j=1}^{f} p_j & p_1 & \dots & p_f \\ r_1 & 1 - r_1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_f & 0 & \dots & 1 - r_f \end{bmatrix} \tag{27}$$

The average production rate of the machine is [34]:

$$e = \frac{1}{1 + \sum_{j=1}^{f} I_j} \tag{28}$$

Theorems 1 and 2 hold also for this case, with the only difference that e is specifically calculated by using equation (28) and using a binary reward vector for this machine that is a column vector with element 1 for the operational state and element 0 for each down mode of the machine.

Given the particular structure of the underlying Markov chain, it is possible to obtain rather simple and compact equations for the asymptotic variance rate and the variance of the inter-departure time. The asymptotic variance rate can be expressed as:

$$v = \frac{\sum_{j=1}^{f} I_j \left(\frac{2-r_j}{r_j}\right) - \left(\sum_{j=1}^{f} I_j\right)^2}{(1 + \sum_{j=1}^{f} I_j)^3} = \left[\sum_{j=1}^{f} I_j \left(\frac{2-r_j}{r_j}\right) - \left(\sum_{j=1}^{f} I_j\right)^2 \right] e^3 \tag{29}$$

The variance of the inter-departure time is:

$$var[IDT] = \frac{e-1}{e^2} + 2\sum_{j=1}^{f} \frac{p_j}{r_j^2} \tag{30}$$

By substituting equations (28) and (29) into equation (30) the same simple relation between the variance of the inter-departure time, the asymptotic variance rate and the throughput already found for the single failure case is derived:

$$var[IDT] = -\sum_{j=1}^{f} I_j \left(1 + \sum_{j=1}^{f} I_j\right) + 2\sum_{j=1}^{f} \frac{I_j}{r_j} = \sum_{j=1}^{f} I_j \left(\frac{2}{r_j} - 1\right) - \left(\sum_{j=1}^{f} I_j\right)^2 = \frac{v}{e^3} \tag{31}$$

Finally, the square coefficient of variation of the inter-departure time can be expressed as:

$$cv^2[IDT] = \frac{v}{e} = d \tag{32}$$

Therefore, the coefficient of variation of the inter-departure time for a single machine with multiple failure modes in isolation is equal to the asymptotic index of dispersion of the cumulated production quantity.

5 Application to multi–stage production lines

The method can be used for the analysis of production lines with K unreliable machines and limited buffer capacities. The behavior of each machine is assumed to be described by a discrete time Markov chain of general complexity. In particular, the number of states of machine i (with $i = 1, \ldots, K$) is denoted by S_i and the system state is identified by the vector $\boldsymbol{x} = (n_1, \ldots, n_{K-1}, s_1, \ldots, s_K)$, where n_i indicates the level of buffer B_i. The total number of states is $S = (N_1 + 1) \cdot \ldots \cdot (N_{K-1} + 1) \cdot S_1 \ldots S_K$, where N_i with $i = 1, \ldots, K-1$ is the capacity of buffer B_i.

The approach proposed in this paper also applies in this case. The only issue to take into consideration is the state-explosion phenomenon. Indeed, the size of the transition matrix describing the dynamics of this system depends both on the number of states of the machines composing the system and on the buffer sizes.

Focusing the attention on the last machine of the line, M_K, it is assumed that the random variable Y_i is equal to 1 if the observed machine produces one piece in period i, and 0 otherwise. In particular, Y_i can be zero for different reasons: firstly, the machine M_K can be in a down state; secondly, machine M_K can be starved since one of its upstream machines M_j, with $j = 1, .., K-1$, is in a down state and all the buffers in between are empty. The expected throughput of this system is the sum of all the steady state probabilities in which the last machine is operational and not starved [13]. To be consistent with the notation adopted in the literature, we will use e and v for the mean production rate and asymptotic variance rate of isolated machines, and E and V for the same performance measures calculated for the whole line.

Table 2 reports the average computational times for the analysis of 20 systems with different number of machines and buffer capacities. In all cases, the machines are unreliable and can fail in one failure mode, characterized by geometrically distributed times to failure and times to repair. The average value reported for each case is calculated on the basis of 10 different lines, with machine parameters p and r uniformly sampled between 0 and 1. The results were obtained using an Intel Core2 Duo 1.6 *GHz* computer with 3 GB of RAM. The tolerance ϵ was equal to

0.1. It can be seen that the evaluation of production lines can be performed by the proposed method, within reasonable computational times (always lower than 10 minutes). It should also be mentioned that, by using sparse matrix representation of the state space and eliminating transient states, the speed of the method can be drastically improved. Moreover, implementation issues that help increasing the speed of the method are reported in Appendix C.

The estimation of the computational time needed to evaluate a certain system (in seconds) with S states has been fitted (from the experiments related to Table 2) with the following regression model:

$$\hat{T}_{ev} = -1.096 + 0.004358S - 8.6 \times 10^{-7} S^2 + 4.3 \times 10^{-11} S^3 \tag{33}$$

with $R^2 = 98.6\%$, indicating that the regression model well fits the evaluation time T_{ev}.

Table 3 shows the results from the evaluation of thirteen different production lines, with different lengths and machine parameters. For each line, the machines are identical. Each machine is characterized by a single failure mode, with parameters p and r obtained by using the results reported in equation (26), starting from the values of e and v reported in the third column of Table 3. For each line, the asymptotic variance rate is reported, as computed with the proposed approximate method ($\epsilon = 0.01$) and the exact method. The results show that the method is applicable also to production lines of considerable length (a case with 7 machines is reported, where the number of states is 294912). Moreover, it can be noticed that the approximate method always provides the result within the set tolerance level and with lower computational time with respect to the exact method. The difference is particularly significant when the line complexity increases. It is also worth to notice that the approximate method provides the evaluation of the performance measures also in cases where the exact method fails to do so. The last cases are reported to identify the applicability limit of the exact method.

Figure 1 shows the service level estimated using different methods to calculate $var[Z_t]$ for Case 24. The use of the approximation in equation (11) (with $\epsilon = 0.1$) produces accurate estimates compared with the use of the standard approximation $var[Z_t] \approx vt$ adopted in the literature. The accuracy of the method for the same case as a function of the tolerance level is also shown in Figure 2. It can be seen that after a threshold value of ϵ the method provides stable and accurate results.

6 Numerical analysis of the output variability measures

6.1 Single machine

The proposed method is used to derive insights on the behavior of the output variability under changes in the main machine parameters. Firstly, the analysis of the impact of the machine reliability parameters is carried out. Figure 3 shows the behavior of v for the single failure mode geometric machine, with different values of e and ρ. It can be noticed that ρ impacts v more than e does when ρ is higher than 0, i.e. when the machine output is positively autocorrelated, as in most of the real cases.

Figure 4 shows the relationship between v and e and the machine parameters (p, r) in a contour graph. The area in the graph is divided into three regions, namely A, B and C by the curves $p^*(r)$ and $r^*(p)$. The curve $p^*(r)$ is defined as the level set (denoted with $\mathbb{S}_p$) that maximizes the variance rate given a value of r. Similarly, the curve $r^*(p)$ is defined as the level set (denoted with $\mathbb{S}_r$) that maximizes the asymptotic variance rate given a value of p. The curves are calculated by partially differentiating v with respect to r and p, respectively. These level sets are general,

Table 2: Average k^* and computational time for the analysis of multi–stage production lines with single failure geometric machine and equal intermediate buffer capacities, performed by using the proposed approximate method (equation (11)), with $\epsilon = 0.1$.

Case	M	N_i	k^*	S	T_{ev} [s]
1	2	2	75.2	12	0.015
2	2	4	123.4	20	0.010
3	2	6	153.1	28	0.015
4	2	8	192.1	36	0.023
5	2	10	215.5	44	0.038
6	2	12	241.1	52	0.046
7	3	2	92.9	72	0.025
8	3	4	127.3	200	0.062
9	3	6	161.2	392	0.111
10	3	8	199.2	648	0.166
11	3	10	263.0	968	0.227
12	4	2	99.1	432	0.156
13	4	4	141.9	2000	0.596
14	4	6	190.1	5488	1.690
15	4	7	222.3	8192	2.671
16	4	8	235.5	11664	4.026
17	5	2	108.5	2592	3.857
18	5	3	163.1	8192	4.231
19	5	4	155.8	20000	88.209
20	6	2	123.3	15552	14.287

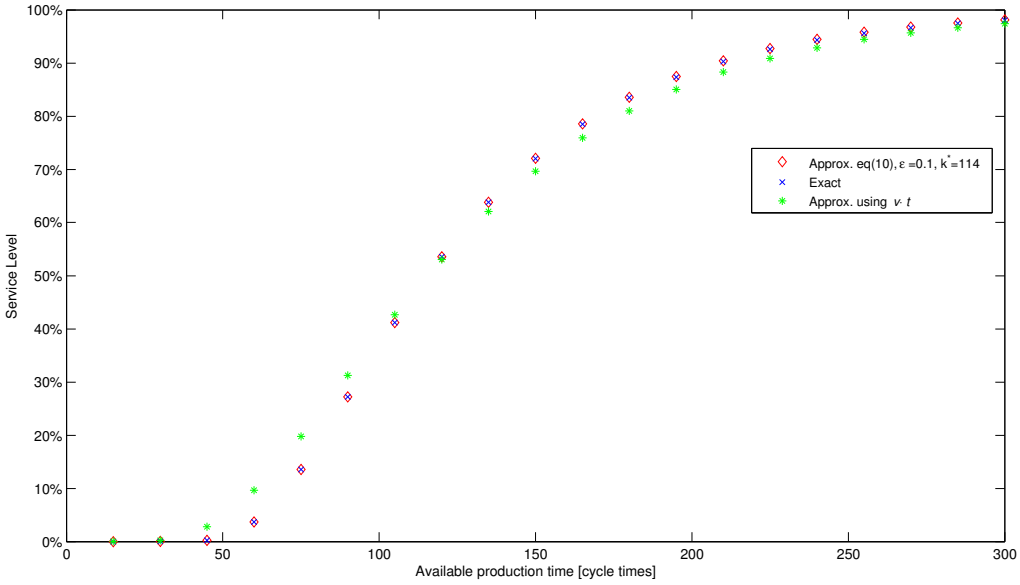

Fig. 1: Exact and approximate $SL(x,t)$ for Case 24, $x = 80$

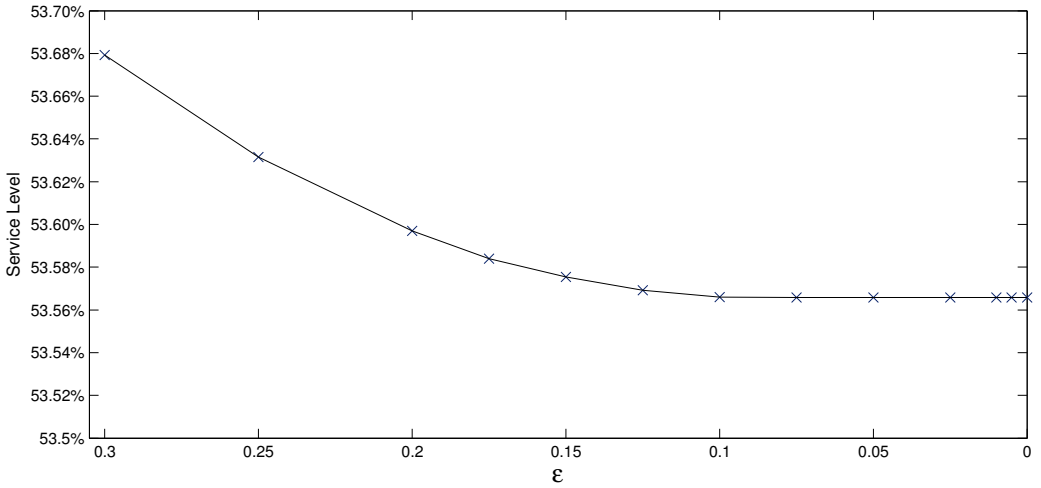

Fig. 2: $SL(x,t)$ as a function of the tolerance level ϵ for Case 24, $x = 80, t = 120$

Table 3: Variability output performances for a set of production lines with different lengths, buffer capacities and machine parameters

Case	M	(e,v)	N_i	$E[IDT]$	$var[IDT]$	E	V_{exact}	V_{approx}	$T_{ev}^{exact}[s]$	$T_{ev}^{approx}[s]$
21	3	(0.9,10)	3-5	1.327	40.387	0.7533	17.183	17.183	0.153	0.125
22	3	(0.95,2)	8-8	1.141	6.224	0.8768	4.114	4.115	0.309	0.309
23	4	(0.7,3)	2-2-2	2.604	32.062	0.3841	1.686	1.687	0.144	0.141
24	4	(0.9,8)	2-2-2	1.443	43.669	0.6931	14.476	14.467	0.184	0.162
25	4	(0.8,4)	4-3-2	1.924	28.648	0.5198	3.812	3.814	0.373	0.309
26	4	(0.9,8)	4-4-4	1.430	42.391	0.6994	14.324	14.324	0.829	0.786
27	4	(0.7,3)	4-4-4	2.364	27.451	0.4230	1.769	1.773	0.673	0.625
28	4	(0.8,4)	10-5-3	1.820	25.451	0.5495	3.815	3.819	1.394	1.322
29	5	(0.8,4)	2-2-2-2	2.195	37.017	0.4556	3.308	3.308	1.340	1.325
30	5	(0.8,4)	4-4-4-4	2.047	39.216	0.4883	3.37	3.374	37.226	11.598
31	6	(0.85,1.3)	2-2-2-2-2	1.943	11.166	0.5145	1.314	1.317	17.539	14.737
32	6	(0.85,1.3)	2-5-2-5-2	1.771	Out of Memory	0.5646	Out of Memory	1.346	-	94.507
33	7	(0.85,1.3)	2-3-3-3-3-2	1.914	Out of Memory	0.5223	Out of Memory	1.228	-	1104.167

they do not depend on the specific machine and are formally defined as:

$$p^*(r) \in \mathbb{S}_p = \{\forall p \in [0,1] : p = 1 - \sqrt{1 - 4r + r^2}\} \ \forall r \in [0,1] \tag{34}$$

$$r^*(p) \in \mathbb{S}_r = \{\forall r \in [0,1] : r = 2 - \sqrt{4 + p^2 - 2p}\} \ \forall p \in [0,1] \tag{35}$$

Let us now to consider a point (p,r) in the graph belonging to region A, or B or C, the following actions are recommended to improve the machine performance:

- *Region A* ($\forall r, p \leq p^*(r)$). Decreasing p or increasing r has a double positive effect, i.e. v decreases and e increases. In other words, actions that increase the machine MTTF or that decrease the MTTR have a positive effect on both e and v.
- *Region B* ($p \geq p^*(r), r \geq r^*(p)$). Increasing r has the same double positive effect on v and e, while decreasing p has the positive effect of increasing e, coupled with the negative effect of increasing v. In this region a rather counterintuitive effect is observed. If the MTTF related of the machine is increased, for example by applying machine improvement plans,

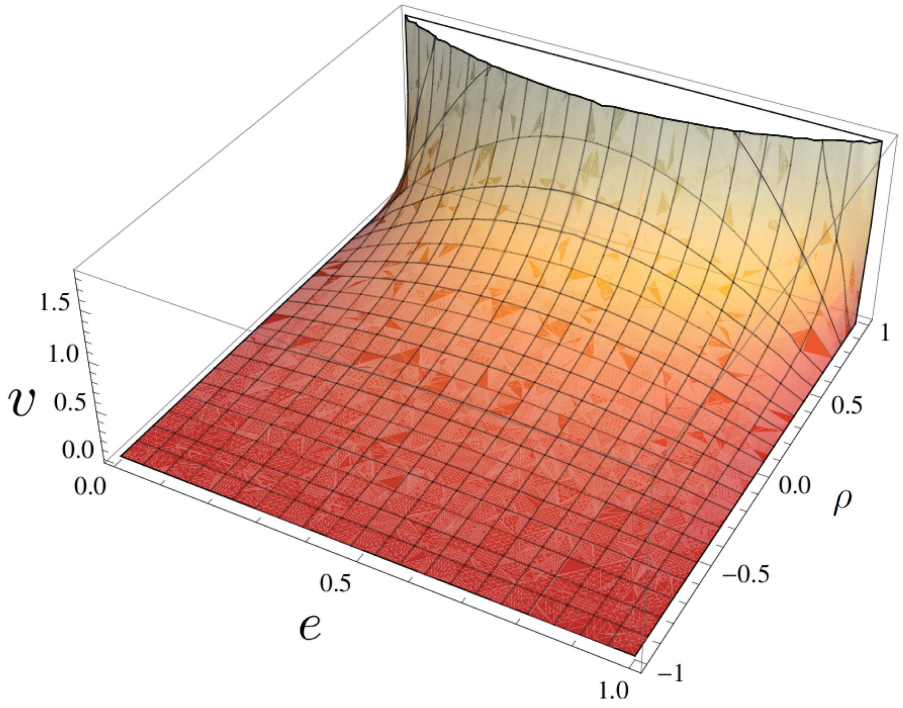

Fig. 3: v vs e and ρ for the single failure geometric machine

the throughput of the system is increased but the output becomes more unstable, since the output variance is higher. On the contrary, up to a certain extend, keeping the machine down for a longer time may be beneficial for v since the loss in the throughput is compensated by a higher stability of the output. The global impact of this action on the machine service level is not shown in the graph and it should be taken into consideration.

– *Region C* ($\forall p, r \leq r^*(p)$). Increasing r or decreasing p has the positive effect of increasing e, coupled with the negative effect of increasing v. This is a trade-off region where the only possibility of decreasing the output variability of the machine is to decrease its throughput. Therefore, up to a certain extend, keeping the machine up for a shorter time may be beneficial since the loss in the throughput is compensated by a higher stability of the output. Also in this case, other performance such as the machine service level should also be considered. It should be noticed that machines in this region can drastically affect the system performance of a production system because of the low efficiency. Thus, it would be beneficial to implement improvement actions that upgrade the machine to regions A or B.

This map can be used to select proper machine reconfiguration actions that improve the performance both in terms of asymptotic throughput and variance rate, depending on the position of the machine in the graph. After the estimation of the MTTF and MTTR of the machine, it is possible to identify its relative position respect to the level sets $\mathbb{S}_p$ and $\mathbb{S}_p$. This would help machine designers and production managers identifying the best improvements actions for the analyzed machine.

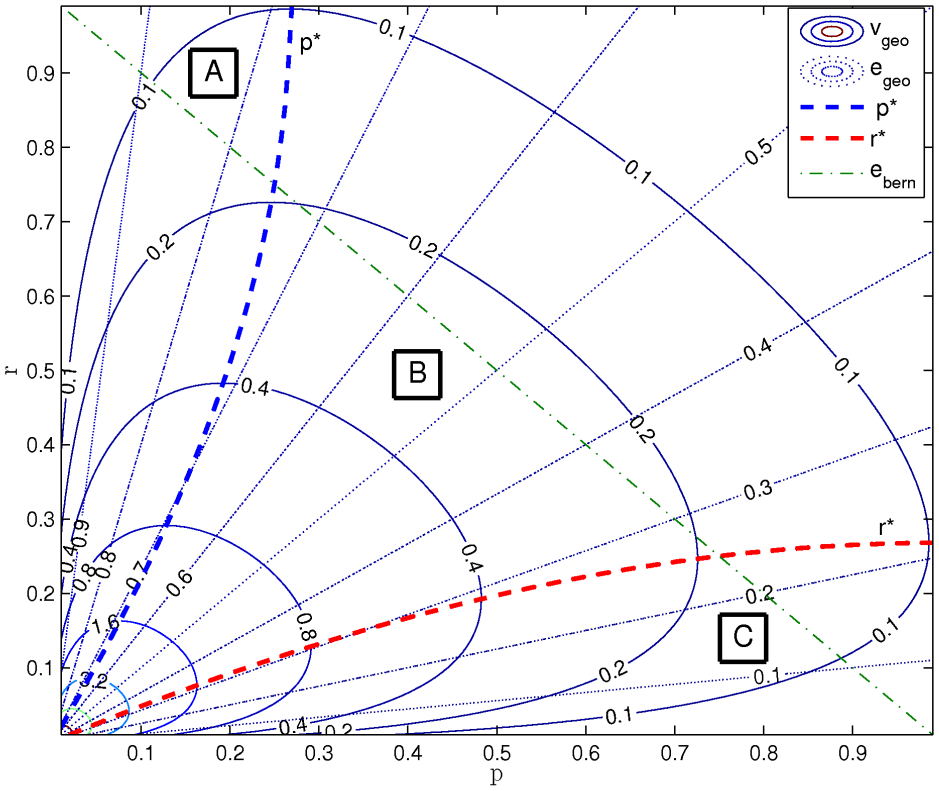

Fig. 4: Contour graph showing efficiency and asymptotic variance rate as a function of p and r for the geometric machine and for the Bernoulli machine (as a function of p).

Figure 5 shows the impact of ρ on the service level for certain values of e and a fixed demand x equal to $0.7143 \cdot t$. When e is equal to x/t, the SL is always 50% regardless of the amount of correlation in the output. When e is greater than x/t increasing ρ will cause SL to decrease, as the probability of obtaining a long consecutive series of no output (i.e. $Y_i = 0$) increases. When e is smaller than x/t, increasing ρ will cause SL to increase, as the probability of having long series of consecutive outputs (i.e. $Y_i = 1$) is higher.

6.2 Two–machine lines

6.2.1 Impact of the buffer size on the output variability

Understanding the impact of the buffer size on the output variance in a two-machine system is a complex task. Carrascosa [4] showed that the shape of the variance rate (V) curve as a function of the buffer capacity N is very sensitive to the machine parameters. In order to understand deeper this behavior, we study a two-machine line characterized by geometric machines subject to a single failure mode. The goal of the analysis is to identify the main factors affecting $V(N)$. To

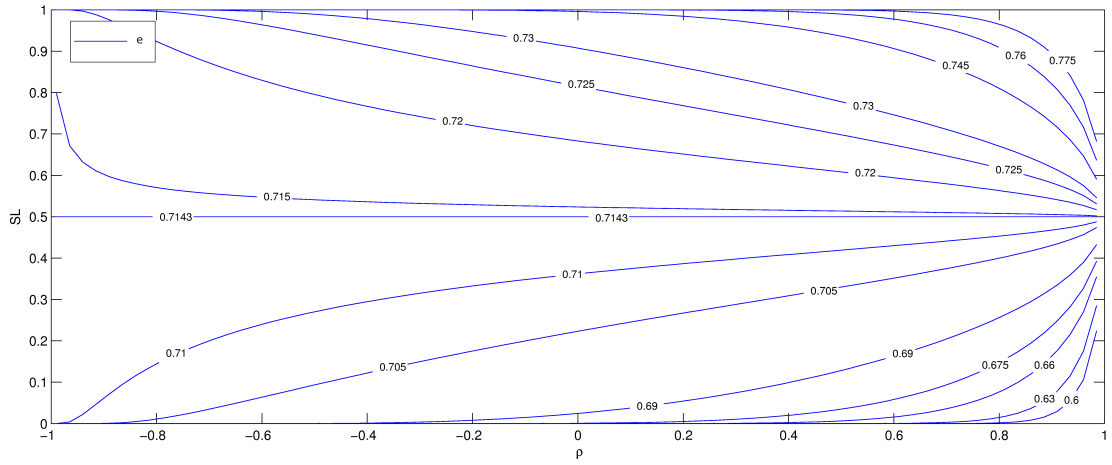

Fig. 5: Service level as a function of ρ and e when the demand x is equal to $10,000$ products within a period of $14,000$ cycle times

this purpose, a design table is constructed considering as factors the efficiency and the asymptotic variance rate of the isolated machines. Specifically, nine set of cases are built combining the cases with $e_1 >, =, < e_2$ and $v_1 >, =, < v_2$. This originates nine possible system classes. In each class, we consider three specific parameter instances respecting the properties of the class and we calculate and plot the variance rate of the two-machine line (V) as a function of the buffer size N, varying from 2 to 200. The results of the experiment are reported in Figure 6 and Table 4.

By analyzing equation (11) applied to the two-machine case:

$$V \approx E(1-E)(1+2\rho_{total}(\epsilon))$$

It is possible to notice that the behavior of the first component of V, i.e. $E(1-E)$ is a concave parabola of the system's throughput E. It is also well known that the throughput E is a monotonic, non-decreasing function of the buffer capacity N [13]. Considering this first term, two different effects of N on V can be observed: if $E(N) < 0.5$ any increase in N will cause E and V to increase, while if $E(N) > 0.5$ any increase in N will cause E to increase, and V to decrease.

The effect of the second term, i.e. $(1+2\rho_{total})$, is much more complex to analyze a priori by only looking at the machine parameters. Indeed, the total autocorrelations ρ_{total} strongly depends on the eigenvalues of the transition matrix of the system, that, unless the system is very simple, is a complex function of the machine parameters. This term has been reported in Table 4. As it can be noticed, there are cases in which ρ_{total} increases with N (for example cases 8, 9, 20 and 21); moreover, there are cases in which it decreases with N (for example cases 2, 4, 15 and 26). More complex cases show a combined effect, i.e. it decreases and then increases with N (for example cases 7, 16, 18 and 19).

The result of the combined effects of these two terms is that V can be a decreasing or an increasing function of N. This result is in accordance with [30] and [4], whereas it contradicts with the findings of Hendricks [16] who noticed that the V always decreases when the buffer capacity increases.

More insights can be obtained from the nine possible combinations of machine parameters presented in Figure 6. When increasing the size of N, V approaches v of the bottleneck machine,

for cases where machines have different efficiencies in isolation (unbalanced lines). The rate of convergence of V when increasing the buffer size is higher for the unbalanced lines.

The counterintuitive mixed effect of the buffer on the variance rate (the buffer size decreases and then increases the variance rate) is particularly visible for cases where the bottleneck machine in terms of isolated throughput is not the bottleneck machine in terms of asymptotic variance rate. For these cases, there is a specific buffer size that minimizes the variance rate. This suggest the development of techniques for optimizing the buffer size to improve the output stability of the system while meeting the target production rate.

In all the analyzed cases, we observed that the asymptotic variance rate of a two-machine line and that of its reversed line, obtained by replacing machine M_1 with machine M_2 and vice versa, is the same.

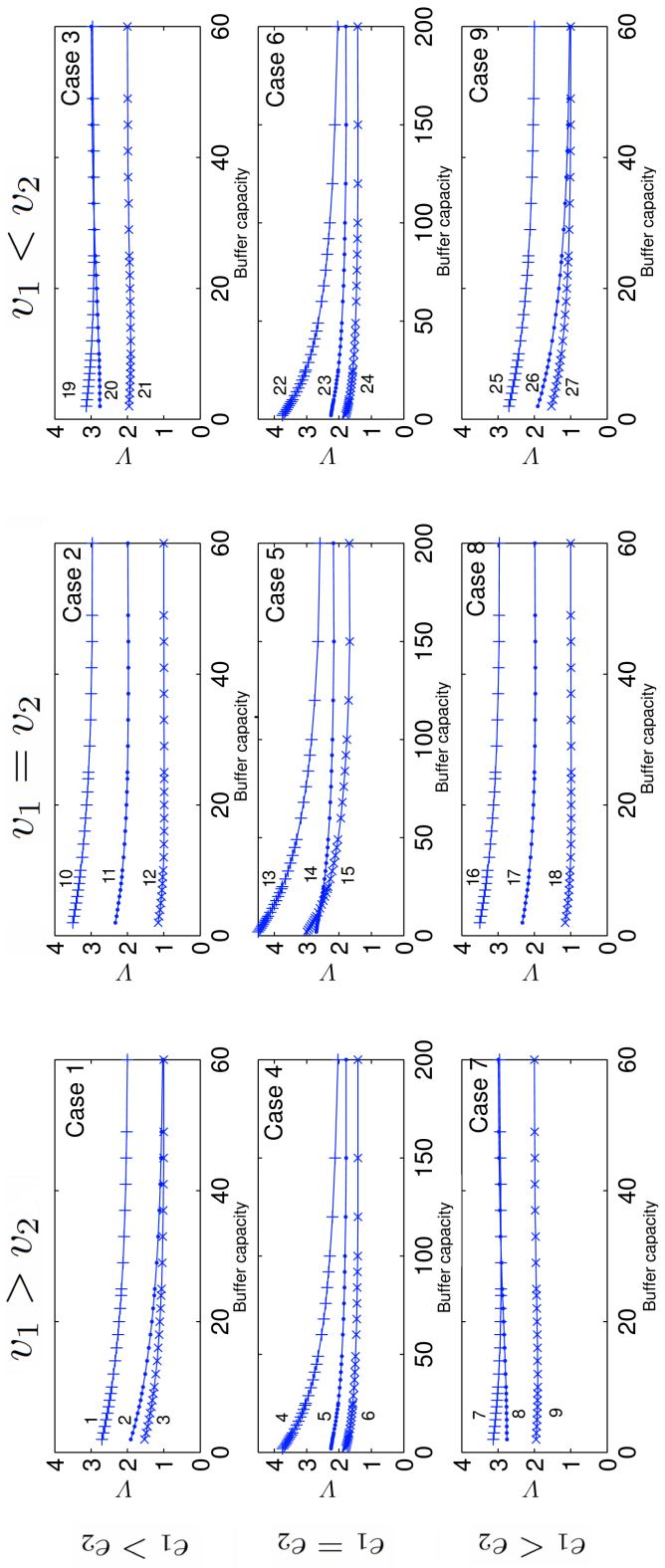

Fig. 6: The asymptotic variance rate V for the nine possible cases of a two-machine line as a function of the buffer capacity. Each output curve represents machine parameters instance that can be found in Table 4
.

6.3 Multi-stage lines

6.3.1 Conservation of the output variability in the system

In this section, a system composed of three machines with geometrical up and down times and single failure mode is considered. Machines are identical with parameters $p = 0.01$ and $r = 0.2$. The buffers are identical with capacity equal to 5. The goal of this experiment is to investigate the conservation of variability throughout the machine stages. Indeed, it is well known that the mean production rate is conserved in production lines, for the so-called "conservation of flow property". This is a very important property that has been exploited by all the developed approximate methods based on the system decomposition. To show this effect, we considered the time dependent variance rate of the system calculated by focusing respectively on the first and the second machine in the system. This implies considering two different reward vectors in the two cases. Then, we plot the difference between the variance rate computed on the first machine and the variance rate calculated on the second machine, as a function of time. The result is shown in figure 7.

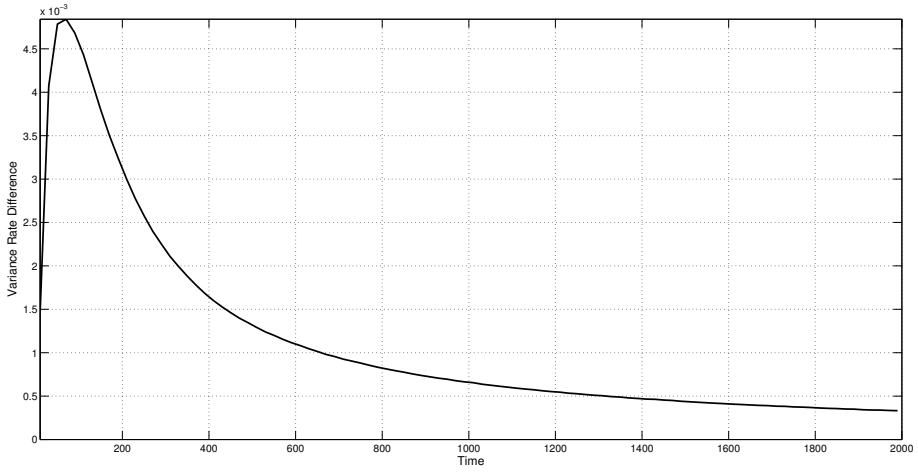

Fig. 7: Difference in the variance rate calculated on the first and the second machine, for a buffered three machine line, as a function of time.

As it can be noticed, the difference increases and then decreases, approaching zero for large times. The absolute value is always relatively small (order of 10^{-3}). The asymptotic variance rate in this case is $V = 0.7088$, and is the same for the two machines. As a second experiment, we modified the parameters of the second machine by reducing the previous values by a factor 100. This results in a second machine that has the same efficiency in isolation but much higher output variability. Then, we repeated the previous analysis. The results for this case are reported in Figure 8.

In this case, the difference decreases and then increases, still approaching zero for large times. Moreover, the time required to approach zero is much higher than the previous case. The asymptotic variance rate in this case is $V = 35.0574$, and is the same for the two machines.

Table 4: The values of V, E and ρ_{total} for the studied 27 experiments at buffer levels $N = 5$, 20, 45, 100 and 200.

Case ID	Instance	e_1	v_1	e_2	v_2	N	5	20	45	100	200
1 $e_1 > e_2$ $v_1 > v_2$	1	0.9	3	0.7	2	V	2.5861	2.2286	2.0298	1.9935	1.9998
						E	0.6585	0.6821	0.6951	0.6997	0.7000
						ρ_{total}	5.2500	4.6385	4.2882	4.2435	4.2614
	2	0.9	3	0.7	1	V	1.7589	1.3268	1.0750	1.0010	1.0000
						E	0.6591	0.6833	0.6958	0.6998	0.7000
						ρ_{total}	3.4143	2.5656	2.0395	1.8824	1.8809
	3	0.9	2	0.7	1	V	1.4056	1.1052	1.0049	0.9996	0.9936
						E	0.6627	0.6894	0.6985	0.7000	0.7000
						ρ_{total}	2.6441	2.0809	1.8857	1.8798	1.8659
2 $e_1 = e_2$ $v_1 > v_2$	4	0.9	3	0.9	2	V	3.6200	3.1410	2.6952	2.2808	2.0313
						E	0.8257	0.8486	0.8661	0.8806	0.8891
						ρ_{total}	12.0775	11.7250	11.1170	10.3436	9.8006
	5	0.7	3	0.7	2	V	2.2200	2.0653	1.9250	1.8124	1.7458
						E	0.5669	0.6260	0.6575	0.6780	0.6883
						ρ_{total}	4.0209	3.9107	3.7740	3.6509	3.5686
	6	0.7	3	0.7	1	V	1.7489	1.6014	1.4979	1.4250	1.3812
						E	0.5725	0.6348	0.6641	0.6819	0.6905
						ρ_{total}	3.0729	2.9538	2.8573	2.7849	2.7315
3 $e_1 < e_2$ $v_1 > v_2$	7	0.7	3	0.9	2	V	3.0712	2.9298	2.9356	2.9941	3.0000
						E	0.6607	0.6865	0.6974	0.6999	0.7000
						ρ_{total}	6.3497	6.3061	6.4549	6.6277	6.6428
	8	0.7	3	0.9	1	V	2.7620	2.8478	2.9705	3.0001	3.0000
						E	0.6661	0.6937	0.6995	0.7000	0.7000
						ρ_{total}	5.7090	6.2009	6.5663	6.6432	6.6424
	9	0.7	2	0.9	1	V	1.9287	1.9294	1.9889	2.0000	2.0000
						E	0.6679	0.6952	0.6998	0.7000	0.7000
						ρ_{total}	3.8479	4.0525	4.2332	4.2619	4.2503
4 $e_1 > e_2$ $v_1 = v_2$	10	0.9	3	0.7	3	V	3.4105	3.1222	2.9749	2.9820	2.9995
						E	0.6580	0.6809	0.6943	0.6996	0.7000
						ρ_{total}	7.0773	6.6852	6.5086	6.5941	6.6415
	11	0.9	2	0.7	2	V	2.2414	2.0244	1.9759	1.9976	2.0000
						E	0.6616	0.6879	0.6979	0.7000	0.7000
						ρ_{total}	4.5053	4.2143	4.1864	4.2558	4.2619
	12	0.9	1	0.7	1	V	1.0807	0.9873	0.9978	1.0000	1.0000
						E	0.6704	0.6967	0.6999	0.7000	0.7000
						ρ_{total}	1.9455	1.8362	1.8752	1.8807	1.8809
5 $e_1 = e_2$ $v_1 = v_2$	13	0.9	3	0.9	3	V	4.3758	3.8805	3.3690	2.8487	2.5475
						E	0.8246	0.8452	0.8624	0.8777	0.8872
						ρ_{total}	14.6243	14.3311	13.6944	12.7728	12.2299
	14	0.7	3	0.7	3	V	2.6787	2.5191	2.3515	2.2039	2.1329
						E	0.5628	0.6186	0.6515	0.6743	0.6862
						ρ_{total}	4.9434	4.8387	4.6785	4.5177	4.4523
	15	0.9	2	0.9	2	V	2.8753	2.4384	2.0687	1.7566	1.6186
						E	0.8274	0.8530	0.8704	0.8837	0.8910
						ρ_{total}	9.5664	9.2227	8.6688	8.0448	7.8340
6 $e_1 < e_2$ $v_1 = v_2$	16	0.7	3	0.9	3	V	3.4105	3.1222	2.9751	2.9823	3.0005
						E	0.6580	0.6809	0.6943	0.6996	0.7000
						ρ_{total}	7.0773	6.6853	6.5091	6.5948	6.6440
	17	0.7	2	0.9	2	V	2.2414	2.0244	1.9760	1.9978	2.0000
						E	0.6616	0.6879	0.6979	0.7000	0.7000
						ρ_{total}	4.5053	4.2143	4.1866	4.2562	4.2618
	18	0.7	1	0.9	1	V	1.0807	0.9873	0.9978	1.0001	1.0001
						E	0.6704	0.6967	0.6999	0.7000	0.7000
						ρ_{total}	1.9455	1.8362	1.8752	1.8811	1.8812
7 $e_1 > e_2$ $v_1 < v_2$	19	0.9	2	0.7	3	V	3.0712	2.9298	2.9354	2.9937	3.0000
						E	0.6607	0.6865	0.6974	0.6999	0.7000
						ρ_{total}	6.3497	6.3060	6.4545	6.6267	6.6427
	20	0.9	1	0.7	3	V	2.7620	2.8479	2.9703	2.9997	3.0000
						E	0.6661	0.6937	0.6995	0.7000	0.7000
						ρ_{total}	5.7090	6.2010	6.5659	6.6422	6.6428
	21	0.9	1	0.7	2	V	1.9287	1.9294	1.9889	1.9999	2.0000
						E	0.6679	0.6952	0.6998	0.7000	0.7000
						ρ_{total}	3.8479	4.0525	4.2331	4.2618	4.2619
8 $e_1 = e_2$ $v_1 < v_2$	22	0.9	2	0.9	3	V	3.6200	3.1410	2.6954	2.2817	2.0389
						E	0.8257	0.8486	0.8661	0.8806	0.8891
						ρ_{total}	12.0775	11.7252	11.1178	10.3480	9.8362
	23	0.7	2	0.7	3	V	2.2200	2.0653	1.9252	1.8122	1.7457
						E	0.5669	0.6260	0.6575	0.6780	0.6883
						ρ_{total}	4.0209	3.9107	3.7742	3.6506	3.5685
	24	0.7	1	0.7	3	V	1.7489	1.6014	1.4981	1.4249	1.3812
						E	0.5725	0.6348	0.6641	0.6819	0.6905
						ρ_{total}	3.0729	2.9539	2.8579	2.7848	2.7315
9 $e_1 < e_2$ $v_1 < v_2$	25	0.7	2	0.9	3	V	2.5861	2.2286	2.0299	1.9941	1.9998
						E	0.6585	0.6821	0.6951	0.6997	0.7000
						ρ_{total}	5.2500	4.6385	4.2885	4.2449	4.2614
	26	0.7	1	0.9	3	V	1.7590	1.3268	1.0752	1.0014	1.0000
						E	0.6591	0.6833	0.6958	0.6998	0.7000
						ρ_{total}	3.4144	2.5657	2.0398	1.8832	1.8808
	27	0.7	1	0.9	2	V	1.4056	1.1052	1.0049	0.9999	0.9938
						E	0.6627	0.6894	0.6985	0.7000	0.7000
						ρ_{total}	2.6441	2.0810	1.8859	1.8807	1.8662

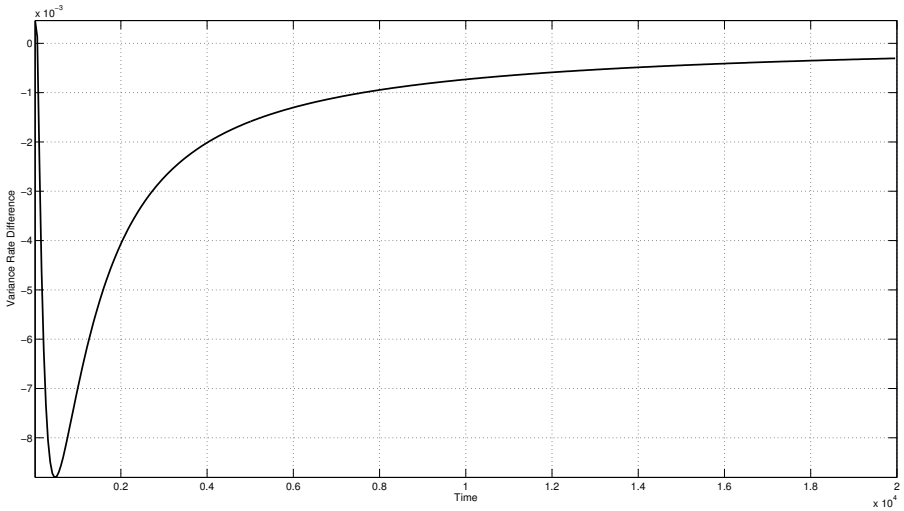

Fig. 8: Difference in the variance rate calculated on the first and the second machine, for a buffered three machine line with modified parameters, as a function of time.

We repeated this analysis for 100 cases and similar results were always observed. Therefore, by numerical facts, we can generalize this result within the following statement:

Numerical Fact 3 *The asymptotic variance rate is conserved throughout the stations of a multi-stage production line.*

This is a very important property that will be useful for the approximate evaluation of the asymptotic variance rate of the output by applying system decomposition. Indeed, this means that both the throughput and the asymptotic variance rate should be conserved in the different building blocks obtained by decomposition. Further investigation is needed for a formal proof of this statement.

7 Conclusions

This paper proposes a methodology to calculate several output variability indicators for single and small multi-stage manufacturing systems modeled as general Markovian structure and binary reward, including the variance of the cumulated production and the inter departure time. The proposed method exploits the special autocorrelation structure of the output of markov-reward systems to compute the variability measures in an approximate way. The approach is general and it can be applied to several different system architectures.

Results show relevant relations between the output variance and the machine reliability parameters and the buffer sizes. In particular, depending on the machine parameters, reducing the MTTR or increasing the MTTF of the machine may have a positive or negative impact on the output variability. This counter intuitive result is important while choosing improvement options that will have positive effect on both e and v. Moreover, the paper shows that increasing the buffer size may reduce or increase the output variability, and an explanation for this behavior is drawn. Furthermore, the paper shows by numerical experiments that the time dependent

variance of the production is not conserved throughout the stages of a production line but the asymptotic variance rate is conserved. This paves the way to the development of approximate analytical methods based on system decomposition to propagate both the asymptotic first and second moment of the cumulated output. Obviously, the problem of the complexity that increases with the buffer sizes and line length should be taken in serious consideration. Moreover, future research will be focused on the formulation and solution of new buffer allocation problems to jointly meet the desired target production rate and the target service level of the system.

A Proof of Theorem 1

According to equation (2) and equation (3):

$$var\left[Z_t\right] = te(1-e) + 2\sum_{k=1}^{t-1}(t-k)cov_k\left[Y\right] \tag{36}$$

Rearranging the previous expression:

$$var\left[Z_t\right] = te(1-e) + 2t\sum_{k=1}^{t} cov_k\left[Y\right] - 2\sum_{k=1}^{t} kcov_k\left[Y\right] \tag{37}$$

where:

$$cov_k[Y] = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}^k\boldsymbol{\mu} - (\boldsymbol{\pi}\boldsymbol{\mu})^2 = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}^k\boldsymbol{\mu} - e^2 \tag{38}$$

The sums in the second and the third terms of equation 37 can be expressed as known geometric sums:

$$\sum_{k=1}^{t} cov_k = \sum_{k=1}^{t}\left(\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}^k\boldsymbol{\mu} - e^2\right) = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t}\mathbf{P}^k\boldsymbol{\mu} - te^2 \tag{39}$$

By adding and removing the term $\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\sum_{k=1}^{t}\mathbf{A}^k\right)\boldsymbol{\mu}$, where A is a square $(s\ x\ s)$ matrix with identical rows formed by the transpose of the steady state probability vector $\boldsymbol{\pi}$, the following can be obtained:

$$\begin{aligned}\sum_{k=1}^{t} cov_k &= \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t}(\mathbf{P}-\mathbf{A})^k\boldsymbol{\mu} + \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t}\mathbf{A}^k\boldsymbol{\mu} - te^2 = \\ &= \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t}(\mathbf{P}-\mathbf{A})^k\boldsymbol{\mu} + t\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{A}\boldsymbol{\mu} - te^2 = \\ &= \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t}(\mathbf{P}-\mathbf{A})^k\boldsymbol{\mu} + te^2 - te^2 = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t}(\mathbf{P}-\mathbf{A})^k\boldsymbol{\mu}\end{aligned} \tag{40}$$

It is worth to recall that, due to the general solution of discrete time Markov chains, $\mathbf{PA} = \mathbf{A}$ and $\mathbf{A}^k = \mathbf{A}$ for each value of $k > 0$. By using the known sum results it is possible to write:

$$\sum_{k=1}^{t} cov_k = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}(\mathbf{P}-\mathbf{A})(\mathbf{I}-(\mathbf{P}-\mathbf{A})^t)(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-1}\boldsymbol{\mu} \tag{41}$$

Similarly:

$$\sum_{k=1}^{t} k\cdot cov_k = \sum_{k=1}^{t} k\left(\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}^k\boldsymbol{\mu} - e^2\right) = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t} k\mathbf{P}^k\boldsymbol{\mu} - \left(\frac{t^2}{2}+\frac{t}{2}\right)e^2 \tag{42}$$

Moreover, by adding and removing the term $\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\sum_{k=1}^{t}\mathbf{A}^k\right)\boldsymbol{\mu}$, the first term can be expressed as:

$$\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t} k\mathbf{P}^k\boldsymbol{\mu} = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t} k(\mathbf{P}-\mathbf{A})^k\boldsymbol{\mu} - \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t} k\mathbf{A}^k\boldsymbol{\mu} \tag{43}$$

therefore, by using the known geometric sum:

$$\sum_{k=1}^{t} k(\mathbf{P}-\mathbf{A})^k = (\mathbf{P}-\mathbf{A})(\mathbf{I}-(\mathbf{P}-\mathbf{A})^t)(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-2} - t(\mathbf{P}-\mathbf{A})^{t+1}(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-1} \tag{44}$$

and, for the second term:

$$\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\sum_{k=1}^{t} k\mathbf{A}^k\boldsymbol{\mu} = \left(\frac{t^2}{2}+\frac{t}{2}\right)\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{A}\boldsymbol{\mu} = \left(\frac{t^2}{2}+\frac{t}{2}\right)e^2 \tag{45}$$

by substituting equations (44), (45) and (43) into equation (42) the following can be obtained:

$$\sum_{k=1}^{t} kcov_k = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left[(\mathbf{P}-\mathbf{A})(\mathbf{I}-(\mathbf{P}-\mathbf{A})^t)(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-2} - t(\mathbf{P}-\mathbf{A})^{t+1}(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-1}\right]\boldsymbol{\mu} \tag{46}$$

Finally, by substituting equations (41) and (46) into equation (37) we obtain:

$$var\left[Z_t\right] = te(1-e) + 2t\boldsymbol{\pi}\boldsymbol{\mu}_{diag}(\mathbf{P}-\mathbf{A})(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-1}\boldsymbol{\mu} - 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}(\mathbf{P}-\mathbf{A})(\mathbf{I}-(\mathbf{P}-\mathbf{A})^t)(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-2}\boldsymbol{\mu} \tag{47}$$

This is a close form expression of the variance of the cumulated number of parts produced at time t by a general Markovian system with transition probability matrix $\mathbf{P}$ and reward vector $\boldsymbol{\mu}$. It can be rewritten in the following form:

$$var\left[Z_t\right] = t\alpha + \beta(t)$$

where:

$$\alpha = e(1-e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}(\mathbf{P}-\mathbf{A})(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-1}\boldsymbol{\mu} \tag{48}$$

and:

$$\beta(t) = -2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}(\mathbf{P}-\mathbf{A})(\mathbf{I}-(\mathbf{P}-\mathbf{A})^t)(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-2}\boldsymbol{\mu} \tag{49}$$

It can be easily shown that, since $\mathbf{P}^t$ approaches $\mathbf{A}$ as t tends to infinity, the term $\frac{\beta(t)}{t}$ tails off, and the asymptotic variance rate expression becomes:

$$v = \alpha = e(1-e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}(\mathbf{P}-\mathbf{A})(\mathbf{I}-\mathbf{P}+\mathbf{A})^{-1}\boldsymbol{\mu} \tag{50}$$

The Fundamental Matrix $\mathbf{Z}$ of a discrete time Markov chain with transition probability matrix P can be expressed as:

$$\mathbf{Z} = (\mathbf{I}-\mathbf{P}+\mathbf{A})^{-1} \tag{51}$$

The properties of the fundamental matrix are such that:

$$\mathbf{AZ} = \mathbf{AZ}^2 = \mathbf{A} \tag{52}$$

Therefore, more compact expressions of α and β can be obtained:

$$\begin{aligned}\alpha &= e(1-e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}(\mathbf{P}-\mathbf{A})\mathbf{Z}\boldsymbol{\mu} = e(1-e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu} - 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{AZ}\boldsymbol{\mu} \\ &= e(1-e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu} - 2e^2 = e(1-3e) + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu}\end{aligned} \tag{53}$$

and:

$$\beta(t) = 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\mathbf{P}^{t+1}-\mathbf{P}\right)\mathbf{Z}^2\boldsymbol{\mu} \tag{54}$$

Therefore, the final expression of the variance is:

$$var\left[Z_t\right] = te(1-3e) + 2t\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{PZ}\boldsymbol{\mu} + 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\mathbf{P}^{t+1}-\mathbf{P}\right)\mathbf{Z}^2\boldsymbol{\mu} \tag{55}$$

B Proof of Theorem 2

By referring to the partitions of the transition probability matrix $\mathbf{P}$ defined in section 3, the mean inter departure time can be obtained with the following equation:

$$\mathbb{E}[IDT] = \frac{1}{e}\boldsymbol{\pi}_U \left(\mathbf{P}_{U,U} + \mathbf{P}_{U,D} \sum_{k=2}^{\infty} k \mathbf{P}_{D,D}^{k-2} \mathbf{P}_{D,U} \right) \boldsymbol{\mu}_U \tag{56}$$

The first term in brackets reflects the situation in which the inter-departure time assumes value 1 since the system makes a transition from one operational state to another operational state or it stays in the same operational state. The second term in brackets reflects the situation in which the system makes a transition to a non-operational state and the inter-departure time increases of one unit for any time step it remains in the down state, until the system goes back to an operational state. The multiplying factor is the conditional probability of the system being in any specific operational state.

By using the results for known sums of geometric series, the following can be written:

$$\sum_{k=2}^{\infty} k \mathbf{P}_{D,D}^{k-2} = \sum_{q=0}^{\infty} (q+2) \mathbf{P}_{D,D}^{q} = \left(2\mathbf{I} - \mathbf{P}_{D,D}\right) \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-2} \tag{57}$$

Therefore:

$$\mathbb{E}[IDT] = \frac{1}{e}\boldsymbol{\pi}_U \left(\mathbf{P}_{U,U} + \mathbf{P}_{U,D} \left(2\mathbf{I} - \mathbf{P}_{D,D}\right) \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-2} \mathbf{P}_{D,U} \right) \boldsymbol{\mu}_U \tag{58}$$

It can be easily proved that, for any system, the mean inter-departure time is the inverse of the throughput, i.e:

$$\mathbb{E}[IDT] = \frac{1}{e} \tag{59}$$

The variance of the inter-departure time can be expressed as a function of the second and the first moments:

$$var[IDT] = \mathbb{E}[IDT^2] - \mathbb{E}[IDT]^2 \tag{60}$$

The second moment of the inter-departure time can be expressed as follows:

$$\mathbb{E}[IDT^2] = \frac{1}{e}\boldsymbol{\pi}_U \left(\mathbf{P}_{U,U} + \mathbf{P}_{U,D} \sum_{k=2}^{\infty} k^2 \mathbf{P}_{D,D}^{k-2} \mathbf{P}_{D,U} \right) \boldsymbol{\mu}_U \tag{61}$$

By using the results for known sums of geometric series, the following can be written:

$$\sum_{k=2}^{\infty} k^2 \mathbf{P}_{D,D}^{k-2} = \sum_{q=0}^{\infty} (q^2 + 4q + 4) \mathbf{P}_{D,D}^{q} = \left(\mathbf{P}_{D,D}^2 - 3\mathbf{P}_{D,D} + 4\mathbf{I}\right) \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-3} \tag{62}$$

Therefore, the second moment of the inter-departure time can be expressed as:

$$\mathbb{E}[IDT^2] = \frac{1}{e}\boldsymbol{\pi}_U \left[\mathbf{P}_{U,U} + \mathbf{P}_{U,D} \left(\mathbf{P}_{D,D}^2 - 3\mathbf{P}_{D,D} + 4\mathbf{I}\right) \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-3} \mathbf{P}_{D,U} \right] \boldsymbol{\mu}_U \tag{63}$$

By rearranging the previous equation we obtain:

$$\mathbb{E}[IDT^2] = \frac{1}{e}\boldsymbol{\pi}_U \left[\mathbf{P}_{U,U} + \mathbf{P}_{U,D} \left(\left(2\mathbf{I} - \mathbf{P}_{D,D}\right) \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-2} + 2\mathbf{I} \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-3} \right) \mathbf{P}_{D,U} \right] \boldsymbol{\mu}_U \tag{64}$$

Therefore, it is possible to express the second moment as a function of the first moment:

$$\mathbb{E}[IDT^2] = \mathbb{E}[IDT] + \frac{1}{e}\boldsymbol{\pi}_U \mathbf{P}_{U,D} 2\mathbf{I} \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-3} \mathbf{P}_{D,U} \boldsymbol{\mu}_U \tag{65}$$

By substituting equations (59) and (65) into equation (60) the closed form expression for the variance of the inter-departure time for any general system can be obtained as follows:

$$var[IDT] = \frac{e-1}{e^2} + \frac{1}{e}\boldsymbol{\pi}_U \mathbf{P}_{U,D} 2\mathbf{I} \left(\mathbf{I} - \mathbf{P}_{D,D}\right)^{-3} \mathbf{P}_{D,U} \boldsymbol{\mu}_U \tag{66}$$

C Implementation Issues

In this section, useful rearrangements of the proposed equations that positively contribute to reduce the computational time of the proposed approaches are proposed, both for the approximate formula and the exact formulas proposed in this paper.

C.1 Approximate formula for the variance of the cumulated production

Equation (5) can be rearranged in order to obtain a recursive function to be evaluated for $k = 1, .., K^*$.

$$cov_k[Y] = \sum_{g=1}^{s} C_{k,g}\mu_g - e^2 \tag{67}$$

where

$$C_{k,g} = \sum_{j=1}^{s} C_{k-1,j}\mathbf{P}_{j,g} \quad k = 2, .., k^*, g = 1, .., s$$

$$C_{1,g} = \sum_{j=1}^{s} \pi_j \mu_j \mathbf{P}_{j,g} \tag{68}$$

In this way, C_k is a row vector that is computed by recursion at each significant time step $k = 1, .., K^*$, thus avoiding several computations of the power matrix. For a similar approach see [33]. $\rho_{total}(\epsilon)$ then becomes:

$$\rho_{total}(\epsilon) = \frac{\sum_{k=1}^{k^*}\sum_{g=1}^{s} C_{k,g}\mu_g - k^* e^2}{e(1-e)} \tag{69}$$

C.2 Exact formula for the variance of the cumulated production

From a computational point of view, the most complex aspect of equations (13) and (14) is the calculation of the elements of the fundamental matrix Z. A computationally efficient method to address this problem is described in the following. Equation (13) is rewritten in the following terms:

$$\alpha = e(1-3e) + 2Q\boldsymbol{\mu} \tag{70}$$

where Q is a row vector of the form:

$$Q = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P}\mathbf{Z} = q(I - P + A)^{-1} \tag{71}$$

where q is a row vector of s elements, easily obtained as:

$$q = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\mathbf{P} \tag{72}$$

The vector Q can be thus obtained by solving the following system of equations:

$$\begin{aligned} Q(I - P + A) &= q \\ Qu &= 0 \end{aligned} \tag{73}$$

where u is a column vector with s elements equal to 1. Similarly, for $\beta(t)$:

$$\beta(t) = 2\boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\mathbf{P}^{t+1} - \mathbf{P}\right)\mathbf{Z}^2\boldsymbol{\mu} = 2W\boldsymbol{\mu} \tag{74}$$

where W is a row vector of the form:

$$W = W_1\mathbf{Z} \tag{75}$$

and:

$$W_1 = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\mathbf{P}^{t+1} - \mathbf{P}\right)\mathbf{Z} = w(I - P + A)^{-1} \tag{76}$$

where w is a row vector of s elements, easily obtained as:

$$w = \boldsymbol{\pi}\boldsymbol{\mu}_{diag}\left(\mathbf{P}^{t+1} - \mathbf{P}\right) \tag{77}$$

The vector W_1 can be thus obtained by solving the following system of equations:

$$\begin{aligned} W_1(I - P + A) &= w \\ W_1 u &= 0 \end{aligned} \tag{78}$$

Finally, the vector W can be obtained by solving this second system of equations:

$$\begin{aligned} W(I - P + A) &= W_1 \\ W u &= 0 \end{aligned} \tag{79}$$

Therefore, the variance formula reduces to:

$$var\left[Z_t\right] = te(1 - 3e) + 2tQ\boldsymbol{\mu} + 2W\boldsymbol{\mu} \tag{80}$$

C.3 Exact formula for the variance of the inter-departure time

The same procedure can be adopted for increasing the computational efficiency in the calculation of the term $(\mathbf{I} - \mathbf{P}_{D,D})^{-3}$ in equation 66.

References

1. Angius A, Horvath A, Colledani M (2011) Moments of cumulated output and completion time of unreliable general markovian machines. In: Proceeding of the 18th World Congress of the International Federation of Automatic Control (IFAC)
2. Betterton C, Silver S (2012) Detecting bottlenecks in serial production lines a focus on interdeparture time variance. International Journal of Production Research 50(15):4158–4174, DOI 10.1080/00207543.2011.596847
3. Bremaud P (1999) Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues. Springer, USA
4. Carrascosa M (1995) Variance of the output in a deterministic two-machine line. Master's thesis, Massachusetts Institute of Technology, Cambridge MA 02139
5. Chen CT, Yuan J (2004) Transient throughput analysis for a series type system of machines in terms of alternating renewal processes. European Journal of Operational Research 155(1):178 – 197, DOI 10.1016/S0377-2217(02)00838-X
6. Ciprut P, Hongler MO, Salama Y (2000) Fluctuations of the production output of transfer lines. Journal of Intelligent Manufacturing 11:183–189, 10.1023/A:1008942917166
7. Colledani M, Matta A, Tolio T (2008) Analysis of the production variability in manufacturing lines. ASME Conference Proceedings 2008(48357):381–390, DOI 10.1115/ESDA2008-59408
8. Colledani M, Ekvall M, Lundholm T, Moriggi P, Polato A, Tolio T (2010) Analytical methods to support continuous improvements at scania. International Journal of Production Research 48(7):1913–1945, DOI 10.1080/00207540802538039
9. Colledani M, Matta A, Tolio T (2010) Analysis of the production variability in multi-stage manufacturing systems. CIRP Annals - Manufacturing Technology 59(1):449 – 452, DOI 10.1016/j.cirp.2010.03.142
10. Cox D, Miller D (1977) The Teory of Stochastic Processes. Science Paperbacks, Wiley
11. Dincer C, Deler B (2000) On the distribution of throughput of transfer lines. The Journal of the Operational Research Society 51(10):1170–1178
12. Gershwin SB (1993) Variance of the output of a tandem production system. In: Onvural RD, Akyildiz IF (eds) Proceedings of the Second International Conference on Queuing Networks with Finite Capacity
13. Gershwin SB (1994) Manufacturing Systems Engineering. PTR Prentice Hall
14. Grassman WK (1993) Means and variances in markov reward systems. In: Meyer CD, Plemmons RJ (eds) Linear Algebra, Markov Chains and Queueing Models, The IMA Volumes in Mathematics and Its Applications, vol 48, Springer-Verlag, New York, pp 193–204
15. He XF, Wu S, Li QL (2007) Production variability of production lines. International Journal of Production Economics 107(1):78–87, DOI 10.1016/j.ijpe.2006.05.014, special Section on Building Core-Competence through Operational Excellence
16. Hendricks KB (1992) The output processes of serial production lines of exponential machines with finite buffers. Operations Research 40(6):1139–1147
17. Hendricks KB, McClain JO (1993) The output processes of serial production lines of general machines with finite buffers. Management Science 39(10):1194–1201
18. Kalir A, Sarin S (2009) A method for reducing inter-departure time variability in serial production lines. International Journal of Production Economics 120(2):340–347

19. Li J, Meerkov S (2000) Production variability in manufacturing systems: Bernoulli reliability case. Annals of Operations Research 93(1):299–324, DOI 10.1023/A:1018928007956
20. Li J, Meerkov S (2009) Production Systems Engineering. Springer
21. Manitz M, Tempelmeier H (2010) The variance of inter-departure times of the output of an assembly line with finite buffers, converging flow of material, and general service times. OR Spectrum pp 1–19, 10.1007/s00291-010-0216-1
22. Meerkov S, Shimkin N, Zhang L (2010) Transient behavior of two-machine geometric production lines. Automatic Control, IEEE Transactions on 55(2):453 –458, DOI 10.1109/TAC.2009.2036328
23. Miltenburg GJ (1987) Variance of the number of units produced on a transfer line with buffer inventories during a period of length t. Naval Research Logistics 34(6):811–822, DOI 10.1002/1520-6750(198712)34:6¡811::AID-NAV3220340606¿3.0.CO;2-Z
24. Ou J, Gershwin SB (1989) The variance of the lead time of a two machine transfer line with a finite buffer. Techical report LMP-90-028, Laboratory for Manufacturing and Productivity, MIT
25. Sabuncuoglu I, Erel E, Gurhan Kok A (2002) Analysis of assembly systems for interdeparture time variability and throughput. IIE Transactions 34(1):23–40, DOI 10.1080/07408170208928847
26. Stewart WJ (2009) Probability, Markov chains, queues and simulation: the mathematical basis of performance modeling, 1st edn. Princeton University Press
27. Tan B (1997) Variance of the throughput of an n-station production line with no intermediate buffers and time dependent failures. European Journal of Operational Research 101(3):560–576, DOI 10.1016/S0377-2217(96)00191-9
28. Tan B (1998) Agile manufacturing and management of variability. International Transactions in Operational Research 5:375–388, DOI 10.1016/S0969-6016(98)00024-0
29. Tan B (1998) An analytical formula for variance of output from a series-parallel production system with no interstation buffers and time-dependent failures. Mathematical and Computer Modelling 27(6):95–112, DOI 10.1016/S0895-7177(98)00031-4
30. Tan B (1998) Effects of variability on the due-time performance of a continuous materials flow production system in series. International Journal of Production Economics 54(1):87–100, DOI 10.1016/S0925-5273(97)00132-1
31. Tan B (1999) Variance of the output as a function of time: Production line dynamics. European Journal of Operational Research 117(3):470–484, DOI 10.1016/S0377-2217(98)00266-5
32. Tan B (2000) Asymptotic variance rate of the output in production lines with finite buffers. Annals of Operations Research 93(1):385–403, DOI 10.1023/A:1018992327521
33. Tan B (2002) State-space modeling and analysis of pull-controlled production systems. In: Gershwin S, Dallery Y, Papadopoulos C, Smith JM (eds) Analysis and Modeling of Manufacturing Systems, Operations Research Management Science, KAP, pp 363–398
34. Tolio T, Matta A, Gershwin S (2002) Analysis of two-machine lines with multiple failure modes. IIE Transactions 34:51–62, 10.1023/A:1019293215459
35. Tolio T, Ceglarek D, ElMaraghy H, Fischer A, Hu S, Laperrire L, Newman S, Vncza J (2010) Speciesco-evolution of products, processes and production systems. CIRP Annals - Manufacturing Technology 59(2):672 – 693, DOI 10.1016/j.cirp.2010.05.008