Ann Oper Res (2011) 182: 163–191
DOI 10.1007/s10479-008-0512-1

Probability masses fitting in the analysis of manufacturing flow lines

Jean-Sébastien Tancrez · Philippe Chevalier · Pierre Semal

Published online: 17 February 2009
© Springer Science+Business Media, LLC 2009

Abstract The analysis of manufacturing systems with finite capacity and with general service time distributions is made of two steps: the distributions have first to be transformed into tractable phase-type distributions, and then the modified system can be analytically modelled. In this paper, we propose a new alternative in order to build tractable phase-type distributions, and study its effects on the global modelling process. Called "probability masses fitting" (PMF), the approach is quite simple: the probability masses on regular intervals are computed and aggregated on a single value in the corresponding interval, leading to a discrete distribution. PMF shows some interesting properties: it is bounding, monotonic, refinable, it approximates distributions with finite support and it conserves the shape of the distribution. With the resulting discrete distributions, the evolution of the system is then exactly modelled by a Markov chain. Here, we focus on flow lines and show that the method allows us to compute upper and lower bounds on the throughput as well as good approximations of the cycle time distributions. Finally, the global modelling method is shown, by numerical experiments, to compute accurate estimations of the throughput and of various performance measures, reaching accuracy levels of a few tenths of a percent.

J.-S. Tancrez (✉)
CORE, Université catholique de Louvain, Voie du Roman Pays, 34, 1348 Louvain-la-Neuve, Belgium
e-mail: js.tancrez@uclouvain.be

J.-S. Tancrez
Louvain School of Management, Facultés Universitaires Catholiques de Mons, Chaussée de Binche, 151, 7000 Mons, Belgium

P. Chevalier
Louvain School of Management, CORE, Université catholique de Louvain, Voie du Roman Pays, 34-1348, Louvain-la-Neuve, Belgium
e-mail: philippe.chevalier@uclouvain.be

P. Semal
Louvain School of Management, Université catholique de Louvain, Place des Doyens, 1-1348, Louvain-la-Neuve, Belgium
e-mail: pierre.semal@uclouvain.be

Keywords Manufacturing flow lines · Probability masses fitting · Discretization · Bounds · Performance measures · Distributions

1 Introduction

In the present economic situation, companies are submitted to an ever more competitive environment. Investments have to bring ever higher profit margins. In view of this, inventory has become a crucial question in production, due to the capital it requires and ties up. It is essential for companies to get high productivity, while limiting the capital used by stocks. In this context, the analysis of manufacturing systems with finite storage capacity is of practical interest. Models of such systems need to be realistic, accurate and complete in order to allow to design and manage them efficiently, and, in particular, to find the best balance between productivity benefits and inventory costs.

In the last decades, various analytical models for manufacturing systems with finite buffers have thus been proposed. They can be divided in two categories: exact and approximate result models. In the first category, the so-called state models are of particular interest. They build a continuous (rarely discrete) Markov chain that models the evolution of the system and from which the performance measures can be inferred (see Hillier and Boling 1967 for example). However, these models' applicability is limited to small instances as the state space size of the Markov chain quickly increases with the system size. As a result, approximate models have been proposed, based on the idea of decomposing the system. The decomposition method breaks the system into smaller subsystems, analyze them in isolation and then iteratively adds back the interdependencies between the subsystems (see Dallery and Frein 1993 for a good review). Another approximate approach is the generalized expansion method, proposed by Kerbache and MacGregor Smith (1987). It also decomposes the system, but it adds the concept of an artificial node which registers the blocked jobs. For further details, several comprehensive reviews of analytical models for manufacturing systems are available (Altiok 1996; Buzacott and Shanthikumar 1993; Gershwin 1994; Perros 1994; Dallery and Gershwin 1992; Papadopoulos and Heavey 1996).

An important assumption of these analytical models (exact and approximate) is that they suppose phase-type (PH) distributed processing times, so that the theory of Markov processes can be applied. No analytical model is directly available for generally distributed processing times. Consequently, in practice, in order to analyze real systems with general distributions, one has to build tractable phase-type distributions previously to the application of the analytical models. This preliminary step is part of the modelling process, which should be seen as a whole.

The related literature, concerned with approximating general distributions by phase-type distributions, is quite separated from the literature related to the analytical modelling of manufacturing systems. The former rarely studies the effect of the approximation on the global modelling process (some compute the accuracy reached on single queues but never on networks). The latter takes the PH distributions as given and neglect the effect of the preliminary step, on the approximation error for example. Most of the methods to approximate general distributions by PH distributions fits continuous PH distributions. We first list the main contributions in this literature and then give the few methods fitting discrete PH distributions. There are essentially three kinds of methods.

The first one, historically and in popularity, is moments fitting, i.e. the two (or three) first moments are computed from the original distributions and the phase-type distributions are then built in order to get the same first moments. Various methods have been proposed,

matching two or three moments, with various types of PH distributions, and for original distributions showing various properties. Sauer and Chandy (1975) (using hyperexponential and generalized Erlang distributions), Marie (1980) (using mixed generalized Erlang distributions) and Whitt (1982) (using hyperexponential and mixed Erlang distributions) first proposed closed-form two moments fittings. Altiok (1985) proposed to add the third moment, using mixed Erlang distributions with two stages, and gave the necessary conditions. Botta and Harris (1986) (using generalized hyperexponential distributions) and Johnson and Taaffe (1989) (using mixtures of Erlang distributions), extended the applicability of three moments fitting. More recently, Osogami and Harchol-Balter (2006) proposed a closed-form three moments fitting for any original distributions, using Erlang–Coxian distributions with an almost minimal number of phases. Subsequently, Bobbio et al. (2005) found a fitting with a minimal number of phases (depending on the properties of the distributions), using acyclic PH distributions. Moments fitting is the most popular distribution approximation method. It has the advantage to be simple, computationally cheap, and to lead to a limited number of phases in good cases.[1] However, the extent to which the moments characterize the distributions is not clear (see McCullagh 1994 and Lindsay and Basak 2000). Pearson et al. (1979) showed that distributions with the same four first moments can have very different shapes (percentiles). Concerning the application to queues, Whitt (1984) showed that extremal distributions, with the same two first moments, may lead to very different performance measures on GI/M/1 queues (relative error of 200 percent), and Klincewicz and Whitt (1984) showed that the error can be reduced by imposing shape constraints (8 percent). Very few information is available about the error made when modelling more complex systems. Tancrez et al. (2008) provide a very simple example, two three station lines with triangular distributions showing the same two first moments, where the cycle time differs by 1.62%.

The second kind of method to approximate distributions is based on the computation of the maximum likelihood. In 1992, Bobbio and Cumani (1992) proposed such a method, fitting a Coxian distribution which maximize the likelihood with the original general distribution. Later, Asmussen et al. (1996) made a significant contribution applying the EM (expectation maximization) algorithm for the maximum likelihood estimation. The third kind of method aims at minimizing a given distance measure between the original and the approximated distributions. Examples of such method are the least squares method by Varah (1985) and the method proposed by Schmickler (1992), which mix distance minimization and moments fitting. These two latter methods, likelihood maximization and distance minimization, are quite similar and share the same advantages and weaknesses. The first goal of their authors is to propose more reliable approximations, hoping that more evolved statistical concepts lead to better approximations. Unfortunately, these methods are computationally more expensive (they imply nonlinear optimization) and often lead to a larger number of phases. Lang and Arthur (1997) conducted a vast experimental study of some fitting methods, based on moments fitting as well as on likelihood maximization. They concluded that methods are effective for "phase-type behaved" distributions and quite unsatisfying for others (low variance, steep increase or decrease, multiple modes, finite support). Concerning approximations based on the maximum likelihood principle, Lang and Arthur claim a "moderate order (order 4 or order 8)" is satisfactory for the "phase-type behaved" distributions while an higher order is necessary for other distributions.

[1] The minimal number of phases depends on the moments (see details in Bobbio et al. (2005)). In particular, the following relation, proved by Aldous and Shepp (1987), is well-known: $1/k \leq cv^2 \leq 1/(k-1)$, where k is the number of phases and cv^2 is the squared coefficient of variation of the original distribution.

Finally, the fitting of discrete phase-type distributions has been clearly less investigated. Concerning moments fitting, Adan et al. (1995) proposed a two first moments fitting method and Telek (2000) gave the necessary conditions in terms of the second moment (which are far less restrictive than for continuous distributions). The likelihood maximization was applied by Bobbio et al. (2003) to fit acyclic discrete PH distributions (with infinite support). Bobbio et al. (2004) propose to use the "scale factor" in order to choose between continuous and discrete PH distributions. All in all, moments fitting and likelihood maximization have the same respective advantages and weaknesses than with continuous PH distributions. Bobbio et al. (2004) summarizes the advantages of discrete PH distributions: their ability to approximate original distributions with a low coefficient of variation, with abrupt changes, with finite support, or with deterministic values. Nevertheless, discrete PH distributions have the disadvantage to generate coincident events.

The global modelling process is thus made of two steps: the tractable distributions building and the proper analytical modelling. Our originality mainly lies in the first step. In this paper, we present a new alternative to build tractable distributions, called "probability masses fitting" (PMF). PMF discretizes distributions with finite support. It is quite intuitive: the probability masses on regular intervals are computed and aggregated on a single value in the corresponding interval, leading to a discrete (phase-type) distribution. After PMF, from the discrete distributions, a state model is applied, modelling the evolution of the system by a Markov chain, from which the performance measures can be computed. In a few words, we take the given processing time distributions, in the form of theoretic density functions or histograms, discretize them using PMF, and then exactly model the modified system to finally evaluate the performance measures.

The goal of this paper is to present probability masses fitting as well as to study the global modelling method using PMF jointly with an exact analytical modelling. Our aim is to show the features and advantages of the global modelling approach in the analysis of manufacturing flow lines. In previous articles, we proposed a particular PMF, called "grouping at the end", which aggregates the probability masses at the end of the intervals, and studied its effect on the productivity estimation (Tancrez and Semal 2006; Tancrez et al. 2008). The present paper introduces the general PMF and its properties, and presents the estimation of various performance measures.

The rest of the paper is organized as follows. Probability masses fitting is presented in details in Sect. 2. We show various properties of PMF, such as bounding and monotonicity properties. We present two slightly different alternatives (generating instant jobs or not) and discuss PMF's characteristics. In Sect. 3, the global modelling method (PMF followed by a state model) is presented for manufacturing flow lines. Then, our first point is to translate the PMF properties to the productivity of the modeled system. This is done in Sect. 4. In Sect. 5, we show, by computational results, how our methods behaves. We show that our method leads to accurate estimations of the performance measures and we study how the results change when the system configuration changes. Finally, we conclude in Sect. 6.

2 Probability masses fitting

In this section, we present the proposed method to build tractable discrete phase-type distributions from given general distributions. It is called "probability masses fitting", and denoted PMF. A general definition is first given, followed by some interesting properties. Then, an alternative PMF is presented. Finally, we discuss the advantages and weaknesses of PMF.

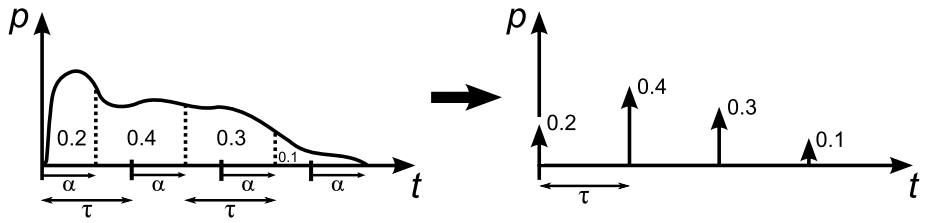


Fig. 1 Discretization by probability masses fitting, with $a = 3$

2.1 Definition

Let us suppose we are given a positive finite processing time distribution and want to transform it into a discrete phase-type distribution by probability masses fitting. To begin, we choose the number of non-zero discrete values, denoted a. Moreover, as we want to infer a discrete Markov chain from the obtained distribution, the size of the interval between two consecutive discrete values, denoted τ, has to be constant. Knowing this, the idea behind probability masses fitting is quite simple: the probability of each discrete value equals the original probability mass (computed as the integral of the density function) around this discrete value. The original distribution is transformed into a discrete one by aggregating the original probability mass distributed in an interval around each discrete value. In order to fix the shift of the intervals, we introduce the parameter α, with $0 \leq \alpha \leq \tau$, which gives the size of the part of the interval located after the discrete value. In summary, PMF transforms a given distribution into a discrete one by aggregating the probability mass distributed in the interval $((j-1)\tau + \alpha, j\tau + \alpha]$ on the point $j\tau$, for each $j = 1, 2, \ldots, a$, and the mass in $[0, \alpha]$ on 0. PMF is illustrated in Fig. 1. Note that τ gives the interval between two discrete values as well as the interval on which the probability mass is computed.

In practice, the parameters a, i.e. the number of non-zero discrete values, and the factor α/τ, i.e. the proportion of the interval which is located after the discrete value, are first chosen. The space τ between two discrete values, or, equivalently, the probability mass interval size, is then deduced from these parameters so that the finite width of the original distribution is covered. In other words, we want the maximum original time *max* to be equal to $a\tau + \alpha$, so that τ is chosen as follows:

$$\tau = \frac{max - \alpha}{a}. \tag{1}$$

Before showing some properties of PMF, let us introduce some notations. The random variable representing the processing time of $w_{i,k}$, the job k on station i, is denoted $l(w_{i,k})$. The realization of $l(w_{i,k})$ in a sample production run r_P (producing P units), i.e. the time the job $w_{i,k}$ takes in this particular run, is denoted $l^r(w_{i,k})$. After PMF with parameter α, we get the corresponding discrete processing times $l_\alpha(w_{i,k})$ and the realizations $l^r_\alpha(w_{i,k})$. Note that, with these notations, the PMF discretization can be formulated as follows:

$$l^r_\alpha(w_{i,k}) \triangleq \left\lceil \frac{l^r(w_{i,k}) - \alpha}{\tau} \right\rceil \tau, \quad \forall r_P, i, k.$$

2.2 Properties

One of the main advantages of probability masses fitting lies in the strong control it offers on the transformation of the individual processing times. The transformation of each original

processing time to a discrete processing time is exactly known. Notably, we know that, when discretized, a processing time will not be shifted by more than α to the left neither by more than $\tau - \alpha$ to the right (see Fig. 1). It leads to the following *bounds* on the processing time.

Proposition 1 *The original processing time of a job, $l^r(w_{i,k})$, can be bounded using its discretized value, $l^r_\alpha(w_{i,k})$.*

$$l^r_\alpha(w_{i,k}) - (\tau - \alpha) \leq l^r(w_{i,k}) \leq l^r_\alpha(w_{i,k}) + \alpha, \quad \forall r_P, i, k. \tag{2}$$

Proof There are two limit cases. The lower bound corresponds to a processing time of length 0 which is increased to τ, with $j = 1, \ldots, a$. The upper bound corresponds to a processing time of length $j\tau + \alpha$ which is decreased to $j\tau$, with $j = 0, \ldots, a$. □

The main interest of these bounds on processing times comes from their extension to bounds on the productivity, which will be shown in Sect. 4.1.

The non-conservation of the distribution expectation, in general, is another characteristic of probability masses fitting. The error made on the expectation is function of the characteristics of the original distribution and of the parameter α, i.e. the shift of the interval on which the probability mass is computed. Consequently, it is quite interesting to study the evolution of the discretized processing time according to the parameter α. It can be checked that this evolution is a *monotonic*, more precisely decreasing, function. Moreover we show that a particular α can always be found so that the expectation of the discretized distribution equals the expectation of the original distribution, if the latter is continuous. In other words, the conservation of the expectation can be obtained by choosing the right parameter α.

Proposition 2 *When discretizing by PMF with two parameters α such that $\alpha_1 \leq \alpha_2$, and with the number of phases a constant, the first discretized processing time will always be larger than the second one.*

$$l^r_{\alpha_1}(w_{i,k}) \geq l^r_{\alpha_2}(w_{i,k}), \quad \forall r_P, i, k.$$

The expectation of a distribution discretized by probability masses fitting is thus decreasing when α is increasing. Moreover, if the original cumulative distribution function $F(t)$ is continuous, the evolution in α of the expectation of the discretized distribution is also continuous. Consequently, a parameter α can always be computed so that the expectation is conserved, i.e. so that the expectation of the discretized distribution is the same as the original one.

Proof To begin, note that $j\tau_1 + \alpha_1 \leq j\tau_2 + \alpha_2, \forall j$, as, by (1), $j\tau + \alpha = j((max - \alpha)/a) + \alpha = (j\, max + \alpha(a - j))/a$, and as $j \leq a$. In order to prove the first part of the proposition, two cases have thus to be considered. In the first case, the original processing time lies between $(j-1)\tau_2 + \alpha_2$ and $j\tau_1 + \alpha_1$. We thus get $l^r_{\alpha_2}(w_{i,k}) = j\tau_2$ and $l^r_{\alpha_1}(w_{i,k}) = j\tau_1$ and, as $\tau_1 \geq \tau_2$ ($\tau = (max - \alpha)/a$), the proposition inequality is valid in this case. In the second case, the original processing time lies between $j\tau_1 + \alpha_1$ and $j\tau_2 + \alpha_2$. We thus get $l^r_{\alpha_2}(w_{i,k}) = j\tau_2$ and $l^r_{\alpha_1}(w_{i,k}) = (j+1)\tau_1$ and the proposition inequality is also valid in this second case.

To prove the continuity in α of $E[t_\alpha]$, the expectation of a discretized distribution, we can simply formulate $E[t_\alpha]$ as a sum of continuous functions. If $f(t)$ and $F(t)$ are the probability density function and the cumulative distribution function of the original distribution, $E[t_\alpha]$

can be written as follows:

$$
\begin{aligned}
E[t_\alpha] &= \sum_{j=0}^{a} j\tau \int_{(j-1)\tau+\alpha}^{j\tau+\alpha} f(t)\,dt = \sum_{j=1}^{a} j\tau \left(F(j\tau+\alpha) - F((j-1)\tau+\alpha)\right) \\
&= \tau\big(F(\tau+\alpha) - F(\alpha)\big) + 2\tau\big(F(2\tau+\alpha) - F(\tau+\alpha)\big) \\
&\quad + \cdots + a\tau\big(F(a\tau+\alpha) - F((a-1)\tau+\alpha)\big) \\
&= -\tau F(\alpha) - \tau F(\tau+\alpha) - \cdots - \tau F((a-1)\tau+\alpha) + a\tau F(max) \\
&= a\tau - \tau \sum_{j=0}^{a-1} F(j\tau+\alpha).
\end{aligned}
$$

This proves the second statement of the proposition. Finally, in order to prove that an α leading to expectation conservation can be found, we just have to show that this expectation is larger than the expectation of the original distribution when α equals zero and smaller when α equals τ. As it is continuous, the expectation of the discretized distribution will thus equal the expectation of the original distribution for a given α. From (2), it can be seen that, when α equals zero, $l^r(w_{i,k}) \leq l_0^r(w_{i,k})$, in other words every processing time is lengthened by the discretization. The expectation of the discretized distribution is thus larger than the original one. Similarly, from (2), when α equals τ, $l_\tau^r(w_{i,k}) \leq l^r(w_{i,k})$, and the expectation of the discretized distribution is thus smaller than the original one. This completes the proof. □

Note that the distribution function $F(t)$ is indeed continuous if the processing times distributions are given in the form of histograms.

2.3 Probability masses fitting without instant job

In the previous section, we introduced the basic probability masses fitting. This discretization potentially generates discrete processing times of length zero with non-zero probability of occurrence (see Fig. 1). In the discretized time, jobs with length zero can thus occur. We call them "instant jobs". It is not difficult to guess that instant jobs complicate the modelling of the system (see Sect. 3). It leads us to propose an alternative to the basic PMF, an alternative which does not generate jobs with length zero. We call it "probability masses fitting without instant job", denoted PMF/nIJ. Moreover, in order to avoid confusion, we will, from now on, call the basic PMF presented previously "probability masses fitting with instant job", and note it PMF/IJ. PMF will be used for the general concept, without differentiation between both alternatives.

In this section, we briefly present PMF/nIJ and its properties. Probability masses fitting without instant job is very similar to its sister with instant jobs. It is just modified in order to remove the possibility of instant jobs. The small difference thus only concerns the first interval (from 0 to α). The original probability mass on this interval $[0, \alpha]$ is aggregated on τ instead of 0. In summary, PMF/nIJ transforms a given distribution into a discrete one by aggregating the probability mass distributed in the interval $((j-1)\tau+\alpha, j\tau+\alpha]$ on the point $j\tau$, for $j = 1, 2, \ldots, a$, and the mass in $[0, \alpha]$ on τ. PMF/nIJ is illustrated in Fig. 2. Compared to Fig. 1, the differences lie in the width of the interval on which the first probability mass is computed and in the absence of discrete probability in zero. Using PMF/IJ or PMF/nIJ, the discretized distributions will have the same values for all but the first interval. Note that PMF/IJ and PMF/nIJ are equivalent when $\alpha = 0$.

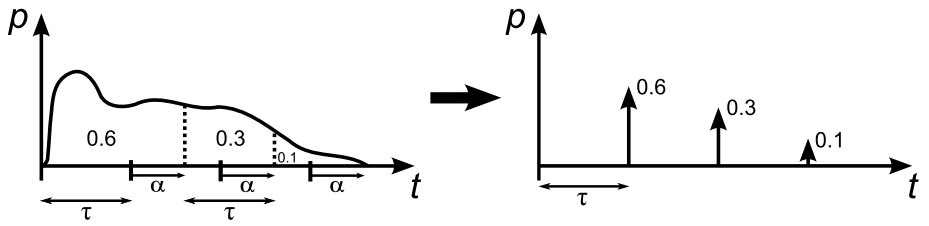


Fig. 2 Discretization by probability masses fitting without instant job, with $a = 3$

The parameters a, α and τ are defined exactly in the same way as for PMF/IJ. The space τ between two discrete values is deduced from a and α/τ using (1). The processing times discretized by PMF/nIJ are denoted $l_{\alpha'}(w_{i,k})$ and the realizations $l^r_{\alpha'}(w_{i,k})$. With these notations, PMF/nIJ can be formulated as follows:

$$\begin{aligned} l^r_{\alpha'}(w_{i,k}) &\triangleq \tau, && \text{if } l^r(w_{i,k}) \leq \tau + \alpha,\ \forall r_P, i, k, \\ &\triangleq \left\lceil \tfrac{l^r(w_{i,k})-\alpha}{\tau} \right\rceil \tau, && \text{if } l^r(w_{i,k}) > \tau + \alpha,\ \forall r_P, i, k. \end{aligned}$$

The properties of probability masses fitting without instant job are similar to the properties of PMF with instant jobs (see Sect. 2.2). We give them shortly. As the first interval is larger using PMF/nIJ (see Fig. 2), the *bounding* property is less good than the one of PMF/IJ. The lower bound is less tight.

Proposition 3 *The original processing time of a job, $l^r(w_{i,k})$, can be bounded using its PMF/nIJ discretized value, $l^r_{\alpha'}(w_{i,k})$.*

$$l^r_{\alpha'}(w_{i,k}) - \tau \leq l^r(w_{i,k}) \leq l^r_{\alpha'}(w_{i,k}) + \alpha, \quad \forall r_P, i, k. \tag{3}$$

Proof The lower bound corresponds to a processing time of length 0 which is increased to τ. The upper bound corresponds to a processing time of length $j\tau + \alpha$ which is decreased to $j\tau, \forall j = 1, \ldots, a$. □

Identically to PMF/IJ, it can be shown that the processing time discretized by PMF/nIJ is a *monotonic*, decreasing, function of α. However, in contrast to PMF/IJ, an α cannot always be found in order to fit the original expectation. The continuity can be proved under the same assumption but an α cannot always be found so that the discretized expectation is smaller than the original one.

Proposition 4 *Discretizing by PMF/nIJ, with two parameters α such that $\alpha_1 \leq \alpha_2$, and a constant, the first discretized processing time will always be larger than the second one.*

$$l^r_{\alpha'_1}(w_{i,k}) \geq l^r_{\alpha'_2}(w_{i,k}), \quad \forall r_P, i, k.$$

Proof As for Proposition 2, $j\tau_1 + \alpha_1 \leq j\tau_2 + \alpha_2, \forall j$, as $j\tau + \alpha = j((max - \alpha)/a) + \alpha = (j\,max + \alpha(a - j))/a$, and as $j \leq a$. In order to prove the proposition, three cases have to be considered. In the first case, the original processing time lies between 0 and α_2. We thus get $l^r_{\alpha'_2}(w_{i,k}) = \tau_2$ and $l^r_{\alpha'_1}(w_{i,k}) = \tau_1$ (since $\tau_1 + \alpha_1 > \alpha_2$) and, as $\tau_1 \geq \tau_2$ ($\tau = (max - \alpha)/a$), the inequality is satisfied. In the second case, the original processing time lies between

$(j-1)\tau_2+\alpha_2$ and $j\tau_1+\alpha_1$ (where $j=1,\ldots,a$). We thus get $l^r_{\alpha'_2}(w_{i,k})=j\tau_2$ and $l^r_{\alpha'_1}(w_{i,k})=j\tau_1$ and, as $\tau_1\geq\tau_2$, the proposition inequality is valid in this case. In the third case, the original processing time lies between $j\tau_1+\alpha_1$ and $j\tau_2+\alpha_2$. We thus get $l^r_{\alpha'_2}(w_{i,k})=j\tau_2$ and $l^r_{\alpha'_1}(w_{i,k})=(j+1)\tau_1$ and this completes the proof. □

2.4 Advantages and weaknesses

We end the presentation of probability masses fitting by a short discussion presenting its main advantages and weaknesses and the differences between PMF with or without instant job. The main *advantages* of probability masses fitting are the following.

- The idea of PMF is *simple* and intuitive. It is natural and easy to understand. The simplicity of the approach is an advantage compared to likelihood maximization and distance minimization methods.
- It preserves the *shape* of the distribution. In fact, it could have been called "shape fitting". Moreover, we show further on that this stays true for the computed throughput distribution. This makes a difference with other fitting methods. Moments fitting, for example, sums up the information in two or three numbers and forgets the shape (see Pearson et al. 1979). This point, the shape conservation, seems quite important. Indeed, in the literature, visual comparisons of the distribution functions is the most common way to evaluate the adequacy of the fitting (see the extensive comparison study by Lang and Arthur (1997) mentioned in the introduction). Moreover, percentiles can be inferred from the computed cycle time distributions. In practice, this allows the manager of the manufacturing system to evaluate the risk and the service level.
- PMF is particularly natural when the data about the processing time distributions is collected in the form of *histograms*, the most common form in practice.
- It is *intelligible*: it offers a strong control on the transformation of the individual processing times. The transformation of each original processing time to a discrete processing time is exactly known. Moreover, each phase of the fitted discrete PH distribution has a clear meaning (for example, it may mean that the original processing time is between 20 and 30 seconds). This is not the case in other fitting methods.
- As a result of the previous advantage, PMF allows to get *bounds* on the productivity of flow lines with general distributions (i.e. through the complete modelling process). This is another characteristic unshared by other fitting methods, and even other modelling methods in general.
- It is *refinable*: the accuracy improves when the number a of discrete values increases. Moreover, at the limit, when the step size decreases, PMF leads to the exact original distribution.
- As will be shown later on, PMF leads to *accurate* estimations of the performance measures of the manufacturing system.

However, probability masses fitting has, of course, some *weaknesses*.

- PMF does not, in general, conserve the *moments* of the distribution. However, we showed that, using PMF with instant jobs, the parameter α can be chosen in order to conserve the expectation, if the original cumulative distribution function is continuous (see Sect. 2.2). Also note that this weakness is shared by likelihood maximization and distance minimization techniques.
- For the PMF to be applied, the distributions have to have a *finite* support (infinite support is impossible in the current setting). However, in the manufacturing application field, it is

always the case in practice. In this context, this property can be seen as an advantage compared to all other proposed distributions fitting methods. Indeed, with other approaches, the support of the fitted distribution is always infinite (a finite support is impossible using continuous phase-type distributions).

- PMF is badly suited for processing time distributions including *rare events*. Indeed, rare events extend the width of the distribution, leading to many insignificant discrete values in the discrete distribution, and increase the complexity without improving the accuracy.

Finally, to compare both PMF alternatives, with or without instant job, we can say that instant jobs have a positive contribution on the accuracy of the distribution fitting, but at the expense of a more complex modelling. This trade-off will be studied in more details by computational experiments (see Sect. 5). PMF with instant jobs also lead to better bounds, in particular lower bounds. Moreover, PMF/IJ has the advantage that it allows, by choosing the adequate shift parameter α, to preserve the expectation of the original distribution.

3 Method

In the introduction, we described the modelling process of a manufacturing system as being composed of two steps: a tractable distributions building preempting the proper analytical modelling. Our originality lies in the first step, where we use probability masses fitting. PMF has now been well defined in the previous section. We are in a position to present the complete modelling method. In very few words, from the discrete processing time distributions built by probability masses fitting, the evolution of the manufacturing system is described by a Markov chain, using a state model, and the performance measures of the system can then be estimated from the chain.

We present the modelling method for *flow lines*. However, note that the approach could be applied to more complex manufacturing systems (assembly/disassembly systems in particular). We are interested in flow lines with asynchronous part transfer and stochastic processing times. They are made of a series of m stations separated by $m-1$ finite buffers. The items are processed sequentially by the stations and stored in the buffers between them when necessary. Such lines experience productivity losses due to blocking and starving. First, a station is said to be blocked when it cannot get rid of an item because the next buffer is full. Second, a station is said to be starved when it cannot begin to work on a new item because the previous buffer is empty. Increasing buffer sizes allows to limit these productivity losses. We make a few classical assumptions on the flow lines we analyze. None of them is restrictive in practice. First, the processing times are generally distributed but finite. Second, the buffer sizes are finite. These finiteness assumptions are not restrictive since they are always satisfied in practice. Third, we suppose that the line operates under saturation, i.e. the first station is never starved and the last station is never blocked. This assumption can be relaxed using an initial and a last station modelling the real arrival process and the real demand. Finally, the blocking policy is chosen to be blocking after service (a blocked job stays in station, after having been processed), the most common in manufacturing. The model could easily be modified in order to follow another policy.

The manufacturing systems it models being well defined, we now present the proposed global *modelling method*. The first step of the method builds discrete distributions by probability masses fitting (see Sect. 2). To begin, the number of discrete values, a, and the parameter α are chosen (the number a is chosen for the longest distribution, it is consistently smaller for shorter distributions). The interval width τ is inferred from (1). Note that τ has to

be the same for each station's distribution. Every processing time distribution is then transformed into a discrete distribution by PMF, i.e. by aggregating the probability mass in the interval $((j-1)\tau+\alpha, j\tau+\alpha]$ on the point $j\tau$, for $j=1,2,\ldots,a$, and the mass in $[0,\alpha]$ on 0 if PMF/IJ is used or τ if PMF/nIJ is used.

The discrete distribution can easily be formulated as a discrete phase-type distribution. An analytical model can then be used. Here, we use a state model, which has the advantage to be exact. Approximate models (decomposition or expansion) could also be applied. The evolution of the production system is described by a Markov chain whose states are the possible combinations of the stages of the various stations and the utilizations of the various buffers. The performance of the production system can then be estimated from the analysis of the Markov chain. Transient performances, like the throughput at some time t, can be derived from the matrix of transition probabilities. The steady-state performances (the cycle time, the work in progress or the flow time for example) can be computed from the steady-state probabilities derived from the Markov chain.

The method has been named "Bounding Discrete Phase-type" (BDPH) since it relies on a discrete phase-type approximation of the various distributions of the system to be studied and since it leads to bounds (Tancrez and Semal 2006). If necessary, we specify if PMF/IJ or PMF/nIJ is used by the notations BDPH/IJ and BDPH/nIJ.

The method can be more clearly understood when explained on a simple *example*. Let us consider a line made of two stations separated by a buffer of size one. Figure 3 shows the application of the BDPH method, with both probability masses fitting alternatives: PMF/IJ and PMF/nIJ. Figure 3a shows the real processing time distributions. Figure 3b shows histograms that could have been collected on the system, as it is often the case in practice. The method can be applied either using the theoretical distribution functions or using the histograms. It works as follows. To begin, PMF transforms the distributions into the discrete distributions shown in Fig. 3c, with two non-zero discrete values ($a=2$) and $\alpha/\tau=1/2$, i.e. aggregating the probability mass in the middle of the interval. The discrete processing time distributions can easily be represented as phase-type distributions (Fig. 3d). Then, the behavior of the system can be modeled by a Markov Chain, i.e. using a state model. The Markov chains given in Fig. 3e list all the possible recurrent states of the system and the transitions between these states. The first symbol of a state refers to the first station, the second to the buffer and the third to the second station. Each station can be starved (S), blocked (B) or in some stage of service (for example, 1 means that the station already spent one time step working on the current job). Each buffer is described by its utilization (0 or 1 for a buffer of size one). For example, state B12 means that the first station is blocked, that the buffer is full and that the second station already worked during two time steps on the current job. From a Markov chain and its stationary probabilities, the performance measures can be computed. For example, the idle (blocked/starved) probability of a machine is easily computed and the productivity easily deduced from it.

In Fig. 3, the difference between probability masses fitting with or without instant job appears, of course, in the processing time distributions (the first step of the PMF/nIJ case gathers two intervals of the PMF/IJ case) and, more interestingly, in the Markov chains that model the evolution of the system. The chain obtained when instant jobs are possible has more states and is clearly more dense. This will affect the speed of the method, in the construction of the chain as well as in its resolution, making the BDPH/IJ method slower (see Sect. 5 for computational times).

The difference comes from the fact that instant jobs make some transitions possible, while they are impossible with PMF/nIJ. For example, using PMF without instant job, four transitions are possible from 101, depending if the stations continue to work on the same

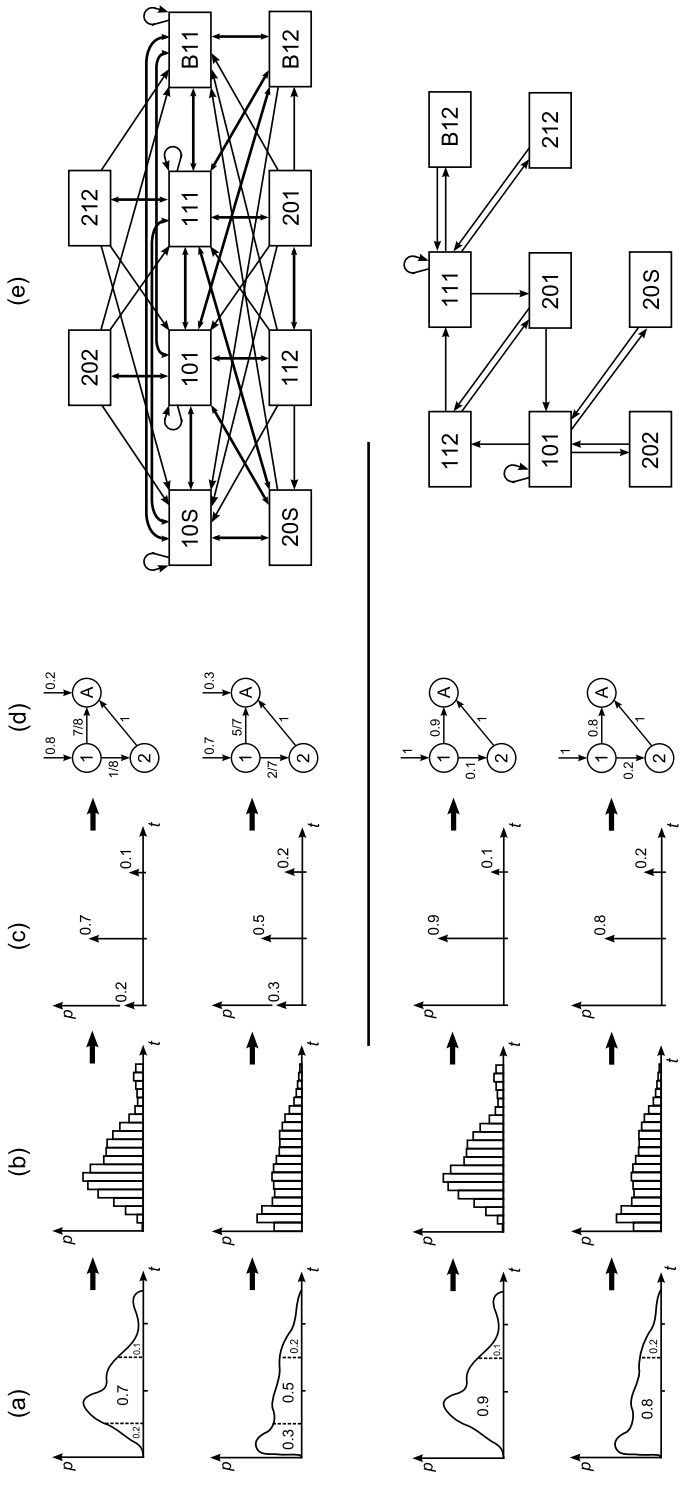


Fig. 3 Steps of the BDPH method (with instant jobs in the *upper part* and without in the *lower part*) applied to a two-station line with processing times (**a**) collected in the form of histograms (**b**): construction of the tractable discrete distributions by PMF (**c**), PH representation (**d**) and Markov chain modelling the evolution of the line (**e**)

job or end. If the first station ends and the second one continues, the new state will be 112 as the first station will put its finished item in the buffer and begin to work on another one, while the second station will work on the same item for a second time step. If both stations finish their jobs, the first station will pass the item to the second station and both of them will begin a new job, leading to the state 101 again. The two other transitions from 101, and, finally, the whole Markov chain can be found following this logic. If PMF with instant jobs is used, more transitions are possible. As can be seen on Fig. 3, eight transitions are possible from state 101, compared to four when PMF/nIJ is used. Let us have look at the transitions from 101, as we did for PMF/nIJ. If the first station ends and the second one continues its job, a possible transition leads to 112 (as for PMF/nIJ). However, if an instant job occurs on the first station, it will block the station and lead to B12. In a second case, if both stations finish their jobs, the transition from 101 to 101 is also possible. It takes place if no instant job occurs, as well as if an equal number of instant jobs occur on each station (an infinite number of them is possible, as they can occur simultaneously on every station). A transition from 101 to B11, for example, is also possible if two more instant jobs occur in the first station, filling the buffer and blocking the station. The Markov chain and its transition probabilities can be found applying this logic for every state.

To conclude this section, we comment on the *complexity* of the proposed method. The size of the Markov chain is, in first approximation, proportional to the number of individual stations and buffer states combinations, i.e. $\prod_{i=1}^{m}(a_i+2)\cdot\prod_{i=2}^{m}(b_i+1)$ with m the number of stations, a_i the number of non-zero discrete values in the discretized processing time distribution of station i, and b_i the size of buffer i. Note that this is a pessimistic estimation as a non-negligible amount of these individual combinations are impossible. It can be seen from the formula that the complexity increases when the number of discrete values a increases, in other words when the discrete time step τ decreases (leading to better accuracy). There is thus a clear trade-off between complexity and accuracy, which can be directly controlled by the parameter a. The size of the Markov chain also quickly increases with the complexity of the system's configuration (number of stations and buffer sizes). The explosion of the state space is the main limitation of the approach, to deal with realistic cases. It is essentially due to the fact that a state model is used in the last step of the method, and is quite independent of the discretization by probability masses fitting. This weakness can be overcome as usual when exact methods become too complex: using approximate methods. Decomposition or expansion models could be applied after probability masses fitting (see the introduction).

4 Properties

In this section, our goal is to translate the properties we observed on probability masses fitting to the global modelling method. First, we reach one of the main results obtained thanks to probability masses fitting: upper and lower bounds on the productivity. Second, we show the monotonicity of the computed productivity in the shift parameter α. Third, we argue that the distribution shape conservation stays true for the computed productivity.

4.1 Bounding methodology

In this subsection, we present the bounding methodology that shows that the BDPH method leads to bounds on the productivity. This methodology relies on two ideas. First, it involves an intelligible transformation, i.e. with good knowledge of its effect on the processing times. Second, the concept of critical path allows to translate the effect of the transformation to the global time.

In order to present the bounding methodology in a general form, we use the following formalism. We consider a probability masses fitting which transforms the original processing time $l^r(w_{i,k})$ to the discrete processing time $\tilde{l}^r(w_{i,k})$, such that $l^r(w_{i,k})$ can be bounded using its discretized value as follows:

$$\tilde{l}^r(w_{i,k}) - \delta_- \leq l^r(w_{i,k}) \leq \tilde{l}^r(w_{i,k}) + \delta_+, \quad \forall r_P, i, k. \tag{4}$$

The terms δ_- and δ_+ stands for the maximum value by which a processing time can be increased and decreased, respectively, when discretized. Using PMF with instant jobs, we have $\delta_- = (\tau - \alpha)$ and $\delta_+ = \alpha$. Using PMF without instant job, we have $\delta_- = \tau$ and $\delta_+ = \alpha$.

4.1.1 Critical path

The notion of critical path is the key idea that allows to translate the properties of probability masses fitting on processing times to properties on the global production time. The notion of critical path has been introduced in a previous communication (Tancrez et al. 2008). We recall its definition and a useful property. The critical path relies on the structural property of the manufacturing system. This property can be found for example in Dallery et al. (1994) for fork-join queueing networks with blocking. We state it for the particular case of flow lines in the following lemma. The moments job $w_{i,k}$ starts and ends are denoted $t_{\text{start}}(w_{i,k})$ and $t_{\text{end}}(w_{i,k})$.

Lemma 1 (Structural property of a flow line) *Given an m-station flow line including a buffer of size b_i before each station i, the moment a job starts is given by the following equation, $\forall i, k$:*

$$\begin{aligned} t_{\text{start}}(w_{i,k}) = \max[\, & t_{\text{end}}(w_{i-1,k}), && (5)\\ & t_{\text{end}}(w_{i,k-1}), && (6)\\ & t_{\text{end}}(w_{i+1,k-b_{i+1}-2}), && (7)\\ & t_{\text{end}}(w_{i+2,k-b_{i+1}-b_{i+2}-3}), && (8)\\ & \ldots, \\ & t_{\text{end}}(w_{m,k-\sum_{j=i+1}^{m} b_j-(m-i+1)})\,]. && (9) \end{aligned}$$

Where, for notation purpose, $t_{\text{end}}(w_{0,k}) = 0, \forall k$.

This equation can easily be understood. In a few words, because of the line structure, a job k can only be started on station i: (5) if its processing in the previous station $i-1$ is ended, (6) if the processing of the previous job $k-1$ in station i is ended and (7-9) if this previous job $k-1$ is not blocked in station i by some unfinished jobs downstream. Furthermore, since there is no reason to wait once all these conditions are satisfied, $t_{\text{start}}(w_{i,k})$ will be given by the maximum of the right hand sides of Lemma 1.

The *critical path* of r_P, denoted $cp(r_P)$, is defined as the sequence of jobs that covers the production run r_P (producing P units) without gap and without overlap. By definition, the length of a run r_P (in other words the time to produce P units in this particular run) can thus be written as a sum of job lengths:

$$l(r_P) = \sum_{w_{i,k} \in cp(r_P)} l^r(w_{i,k}). \tag{10}$$

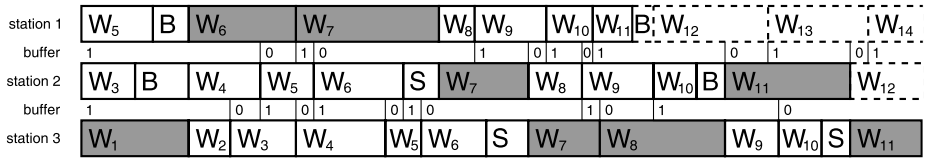


Fig. 4 Gantt chart of a run r_{11} on a three station line with buffers of size one. The critical path is given in *gray*. The time goes from left to right. The state of a station at a given time is represented either by a letter (B for blocked, S for starved) or by the job currently processed. The state of a buffer is represented by its current utilization

The critical path can be built quite easily. Starting with the last job that leaves the system, $w_{m,P}$, we can look which job end, in this precise run, has triggered its start. Repeating this process, we can proceed backwards in time until the start of the run, as every job start is triggered by the end of another job. It follows that every run r_P has at least one critical path.

The notion of critical path is well illustrated on the Gantt chart associated to a particular run. An example is given in Fig. 4. It can be seen that the critical path covers r_{11} and that each couple of successive jobs corresponds to one of the terms (5–9). The "most perceptive readers" can also observe in Fig. 4 that the predecessor of a job in the critical path can be deduced from the state of the same station just before this job. If the station is previously working, the predecessor is obviously on the same station. If the station is previously starved, the predecessor is on the previous station ($w_{3,11}$ to $w_{2,11}$ for example). If the station is previously blocked, the predecessor is on a later station ($w_{2,11}$ to $w_{3,8}$). In the second case, the same item is twice in the critical path (item 11 in the example). In the third case, the $\sum(b_i + 1)$ items between both stations are "omitted" (items 9 and 10). From these observations, an upper bound can be inferred on $|cp(r_P)|$, the cardinality of the critical path, i.e. the number of jobs in it. Indeed, this upper bound corresponds to the worst case where the critical path goes up from the last station to the first one, without going down.

Lemma 2 *Let r_P be a production run on an m-station flow line, the number of jobs in the critical path satisfies the following upper bound*:

$$|cp(r_P)| \leq P + m - 1.$$

4.1.2 In a particular run

Here, we give the lemma on which the bounding methodology relies. It shows that, if the processing time is bounded, the time needed to achieve a given production in a particular run can also be bounded. The run obtained from the PMF discretization of the original run r_P is denoted $\tilde{r}_P$, and its length is denoted $l(\tilde{r}_P)$.

Lemma 3 *Given a PMF discretization of the processing times $l^r(w_{i,k})$ to $\tilde{l}^r(w_{i,k})$ verifying (4), the time a m-station line takes to produce P units in a production run r_P can be bounded as follows*:

$$l(\tilde{r}_P) - \delta_-(P + m - 1) \leq l(r_P) \leq l(\tilde{r}_P) + \delta_+(P + m - 1). \tag{11}$$

Proof These bounds follow from (4) and (10). However, these equations are not sufficient since the critical path is not the same in continuous and in discretized time. The equation of Lemma 1 is valid for any run: a job $w_{i,k}$ cannot be started before all the jobs on the right

hand side are finished. The absence of overlap in the critical path is thus independent of the considered run. However, which precise job end will trigger the start of job $w_{i,k}$ depends on the processing times and thus on the particular run we consider. In another run, gaps could appear between the job of the sequence $cp(r_P)$. The sequence $cp(r_P)$ is thus just a non-overlapping path (possibly with gaps) in the discretized run $\tilde{r}_p$ and its length is smaller than the length of the critical path $cp(\tilde{r}_p)$ in $\tilde{r}_p$. We get:

$$l(r_P) \overset{(10)}{=} \sum_{w_{i,k}\in cp(r_P)} l^r(w_{i,k}) \overset{(4)}{\leq} \sum_{w_{i,k}\in cp(r_P)} \left(\tilde{l}^r(w_{i,k}) + \delta_+\right)$$

$$\leq \sum_{w_{i,k}\in cp(\tilde{r}_P)} \tilde{l}^r(w_{i,k}) + \delta_+|cp(r_P)| \overset{(10)}{=} l(\tilde{r}_P) + \delta_+|cp(r_P)|.$$

As, by Lemma 2, $|cp(\tilde{r}_P)| \leq P + m - 1$, we get the right inequality of the present lemma. For the left inequality, using the same equations and the fact that $cp(\tilde{r}_p)$ is a non-overlapping path in the original run r_P, we get:

$$l(r_P) \overset{(10)}{=} \sum_{w_{i,k}\in cp(r_P)} l^r(w_{i,k}) \geq \sum_{w_{i,k}\in cp(\tilde{r}_P)} l^r(w_{i,k})$$

$$\overset{(4)}{\geq} \sum_{w_{i,k}\in cp(\tilde{r}_P)} \left(\tilde{l}^r(w_{i,k}) - \delta_-\right) \overset{(10)}{=} l(\tilde{r}_P) - \delta_-|cp(\tilde{r}_P)|.$$

As $|cp(\tilde{r}_P)| \leq P + m - 1$ (Lemma 2), this completes the proof. □

Lemma 3 provides a major result. Considering any sample production run, bounds on the individual processing times can be extended to bounds on the time it takes to produce a given production. Unfortunately, this result cannot yet be directly used since it refers to a sample production run. For the results to be useful, we need to be able to say something about an average production run. This point is tackled in the following.

4.1.3 Transient throughput

We first extend Lemma 3 to the expected time T_P necessary to reach a given production P. By definition, T_P ($\tilde{T}_P$) equals the sum of the lengths of all possible original (discretized) runs r_P ($\tilde{r}_P$), weighted by their probabilities. Consequently, we simply have to check that each of the three terms of (11) is weighted in the same way. That comes from the definition of probability masses fitting. As it aggregates the probability masses, the total probability in continuous time of all the runs r_P which discretize to the same run $\tilde{r}_P$ is equal to the probability of this run $\tilde{r}_P$ in discretized time. As this is true for each discrete run and as an original run has only one discrete correspondent, we can extend Lemma 3 to the following proposition.

Proposition 5 *Given a PMF discretization of the processing times $l^r(w_{i,k})$ to $\tilde{l}^r(w_{i,k})$ verifying (4), the expected time T_P a m-station line takes to produce P units can be bounded using $\tilde{T}_P$, the expected time to produce P units computed in the discretized time. We have*:

$$\tilde{T}_P - \delta_-(P + m - 1) \leq T_P \leq \tilde{T}_P + \delta_+(P + m - 1). \tag{12}$$

Proof By definition, the expected time to produce P, T_P, equals the sum of the lengths of all possible runs r_P, weighted by their probabilities. More formally, we have: $T_P = \int f(r_P) l(r_P) dr_P$, where $f(r_P)$ is the density function of the runs r_P. To get (12) from (11), we have to check that each of the three terms of (11) is weighted in the same way. We thus relate r_P to its discretized correspondent, $\tilde{r}_P$ (which also produces P). Let us note $\gamma(\tilde{r}_P)$ the set of original runs r_P (in infinite number) which have the same discretized correspondent $\tilde{r}_P$. We can decompose the previous integral and use Lemma 3, to get:

$$T_P = \sum_{\tilde{r}_P} \int_{r_P \in \gamma(\tilde{r}_P)} f(r_P) l(r_P) dr_P \leq \sum_{\tilde{r}_P} \big(l(\tilde{r}_P) + \delta_+(P+m-1)\big) \int_{r_P \in \gamma(\tilde{r}_P)} f(r_P) dr_P.$$

As the PMF discretization simply aggregates the probability masses in intervals, $\int_{r_P \in \gamma(\tilde{r}_P)} f(r_P) dr_P$ gives the probability of the run $\tilde{r}_P$, i.e. $P[\tilde{r}_P]$. As $\sum_{\tilde{r}_P} P[\tilde{r}_P] = 1$, we get:

$$T_P \leq \sum_{\tilde{r}_P} l(\tilde{r}_P) P[\tilde{r}_P] + \delta_+(P+m-1) = \tilde{T}_P + \delta_+(P+m-1).$$

The way to the lower bound is very similar. Using Lemma 3, we get:

$$T_P \geq \sum_{\tilde{r}_P} \big(l(\tilde{r}_P) - \delta_-(P+m-1)\big) \int_{r_P \in \gamma(\tilde{r}_P)} f(r_P) dr_P = \tilde{T}_P - \delta_-(P+m-1).$$

□

Note that bounds on the production reached in a fixed time can be derived from the previous proposition.

4.1.4 Steady-state productivity

When focusing on the steady-state productivity, the results get even simpler. The cycle time c is defined as the average time, in steady-state, between the completion of two units by the line, i.e. $c = \lim_{P\to\infty} T_P/P$. The cycle time in the discretized time is denoted $\tilde{c}$, i.e. $\tilde{c} = \lim_{P\to\infty} \tilde{T}_P/P$.

Proposition 6 *Given a PMF discretization of the processing times $l^r(w_{i,k})$ to $\tilde{l}^r(w_{i,k})$ verifying (4), the cycle time c can be bounded using $\tilde{c}$, the cycle time computed in the discretized time*:

$$\tilde{c} - \delta_- \leq c \leq \tilde{c} + \delta_+. \tag{13}$$

Proof These bounds straightforwardly follows from Proposition 5 by dividing (12) by P and making $P \to \infty$. □

Propositions 5 and 6 show that our bounding methodology allows to extend bounds on the individual processing times to prove bounds on the productivity, in transient or in steady-state, from below and from above. The similarity of (4) and (13) is remarkable in the present case of flow lines. We believe that this bounding methodology is an interesting contribution. It is clear that the approach is not restricted to the present application. It could be extended to more complex systems (as a critical path can be defined for them too). Moreover, note that

bounds can be inferred on another performance measure, namely the idle (blocked/starved) time. Indeed, the latter equals the cycle time minus the processing time, on which both we have bounds.

The bounding methodology presented in this subsection can be compared to the bounding methodology proposed in Baccelli and Makowski (1989). The latter also use the structural property of the system, in the form of a recursion equation. It is based on the stochastic ordering of the random variables representing the processing times. However, direct relations between the processing times are required, i.e. the additional terms (δ_- and δ_+) are not allowed.

4.1.5 Bounds by probability masses fitting

The results given earlier can readily be instantiated to probability masses fitting with or without instant job. We just give the bounds on the cycle time. Obviously, the bounds on the transient throughput could also be instantiated.

Corollary 1 *The BDPH/IJ method allows to compute upper and lower bounds on the cycle time. With c_α the cycle time in the time discretized by PMF/IJ, we have*:

$$c_\alpha - (\tau - \alpha) \leq c \leq c_\alpha + \alpha.$$

Proof This corollary is straightforward from propositions 1 and 6 ($\delta_- = (\tau - \alpha)$ and $\delta_+ = \alpha$). □

Corollary 2 *The BDPH/nIJ method allows to compute upper and lower bounds on the cycle time. With $c_{\alpha'}$ the cycle time in the time discretized by PMF/nIJ, we have*:

$$c_{\alpha'} - \tau \leq c \leq c_{\alpha'} + \alpha.$$

Proof This corollary is straightforward from Propositions 3 and 6 ($\delta_- = \tau$ and $\delta_+ = \alpha$). □

These results show that the BDPH method allows to bound the productivity, from above and from below. Furthermore, the accuracy of the bounds is related to the selected step size τ. The bounds thus become tighter, and converge, when the discretization step is decreased, i.e. when the number of discrete values is increased. In other words, the bounds are refinable. Moreover, it allows to a priori choose the desired accuracy of the results. Of course, every accuracy improvement will require additional computational efforts caused by the increase of the state space size (which can be estimated, see Sect. 3).

It can also be seen from these corollaries that probability masses fitting with instant jobs offers better bounds than PMF without instant job. The gap between both bounds is smaller. In particular, PMF/IJ allows to compute better lower bounds. This is natural since the absence of instant job requires the interval on which the first probability mass is collected to be larger (see Fig. 2), and thus the lower bound on processing times to be worse. On the left-hand side of Fig. 5, we draw the bounds obtained by the method with instant jobs, for a three station line with buffers of size one and processing time distributions beta(2,2), uniform(0,1) and triangular(0,1,0.5). It can be seen that the bounds get sharper when the step size τ decreases. Moreover, the figure reveals that PMF/IJ with the shift parameter α equal to one, i.e. aggregating the probability mass in the beginning of the interval, leads to the best bounds. It comes from the fact that the step size τ, and thus the distance between

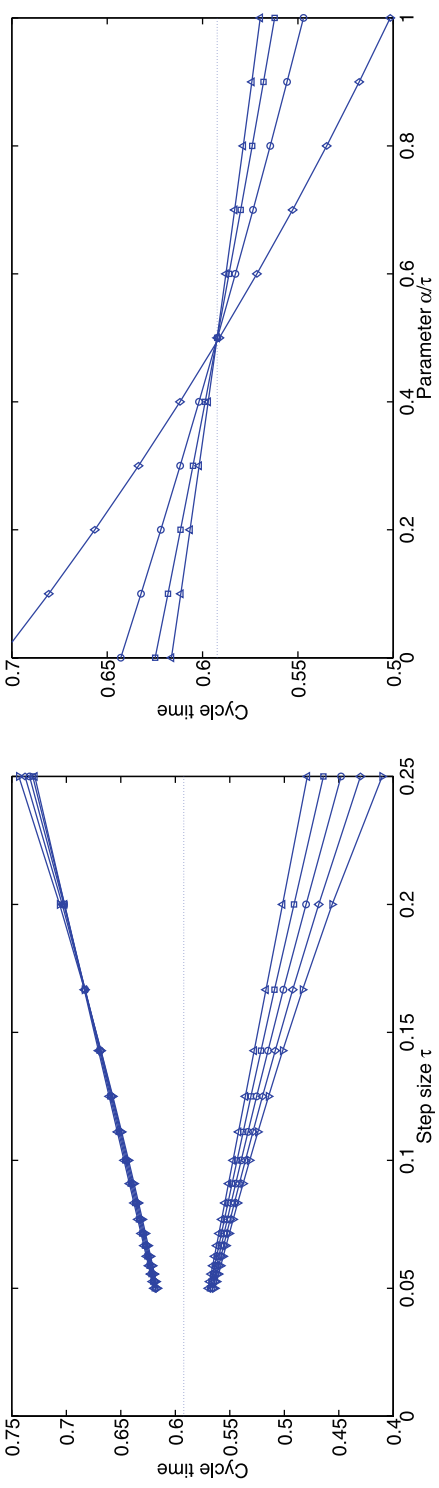


Fig. 5 The BDPH/IJ method is applied an example. On the *left-hand side*, the bounds on the cycle time are computed for various step sizes τ, i.e. various numbers a of steps, with $\alpha = 0$ ($\triangledown$), 0.25 ($\Diamond$), 0.5 ($\odot$), 0.75 ($\square$), 1 ($\triangle$). On the *right-hand side*, the cycle time c_α is computed for various parameters α/τ and with 5 ($\Diamond$), 10 ($\odot$), 15 ($\square$), or 20 ($\triangle$) time steps

both bounds, decreases when α increases ($\tau = (max - \alpha)/a$, see (1)). More precisely, the lower bound shows to be better with $\alpha = 1$, while nothing can be said about the upper bound. As will be shown in Sect. 5 by computational experiments, this observation appears to be not particular to the present example.

4.2 Monotonicity

In the previous section, we extended the bounding property of probability masses fitting to the global modelling method. Similarly, the monotonicity in α of the discretized processing times can also be extended. It can be shown using the critical path as well but it is more easily shown as a corollary of a result given in Baccelli and Liu (1992), which uses the fact that the discretized processing times can be stochastically ordered, and follows the bounding methodology proposed in Baccelli and Makowski (1989). The monotonicity in α of the transient throughput could be shown in the same way.

Proposition 7 *Using the BDPH method with two parameters α such that $\alpha_1 \leq \alpha_2$, the first cycle time will always be larger than the second one.*

$$c_{\alpha_1} \geq c_{\alpha_2} \quad \textit{and} \quad c_{\alpha'_1} \geq c_{\alpha'_2}.$$

Proof The discretized random variables using α_1 or α_2 can be compared stochastically. Proposition 2 (and 4) tells that a realization of the random variable $l^r_{\alpha_1}(w_{i,k})$ (and $l^r_{\alpha'_1}(w_{i,k})$) is always larger than the corresponding realization $l^r_{\alpha_2}(w_{i,k})$ (and $l^r_{\alpha'_2}(w_{i,k})$). We thus have the stochastic ordering inequality $l^r_{\alpha_1}(w_{i,k}) \geq_{st} l^r_{\alpha_2}(w_{i,k})$ (and $l^r_{\alpha'_1}(w_{i,k}) \geq_{st} l^r_{\alpha'_2}(w_{i,k})$). The following increasing convex ordering inequalities are straightforwardly deduced:

$$l^r_{\alpha_1}(w_{i,k}) \geq_{icx} l^r_{\alpha_2}(w_{i,k}) \quad \text{and} \quad l^r_{\alpha'_1}(w_{i,k}) \geq_{icx} l^r_{\alpha'_2}(w_{i,k}).$$

Consequently, as flow lines with finite buffers can be modeled by stochastic decision free Petri nets, Corollary 5.1 of Baccelli and Liu (1992) applies and proves the proposition. □

These results reveal the evolution of the results of the BDPH method in function of the shifting parameter α. When α is increased, the cycle time estimation decreases. Unfortunately, such a monotonic property cannot be proved on the bounds (Corollaries 2 and 1). In the right-hand side of Fig. 5, we illustrate the evolution of the cycle time c_α in α. The decreasing characteristic can be clearly observed.

4.3 Shape conservation

Another interesting property of the probability masses fitting approach lies in the preservation of the shape of the original distribution (see Sect. 2.4). Here, like in previous subsections, we want to extend this PMF property to the global modelling method result. We argue that the shape conservation extends to the cycle time distributions. In other words, we show that we can compute the distribution of the time between two jobs exiting a given station, and that this distribution has a shape close to the real (unknown) one. We do not prove this result but the many example we tested give strong evidence. In the following, we give one example. Others examples show the same behavior.

Let us have a look at the example already analyzed in the previous subsections: a three station line with buffers of size one and processing time distributions beta(2, 2),

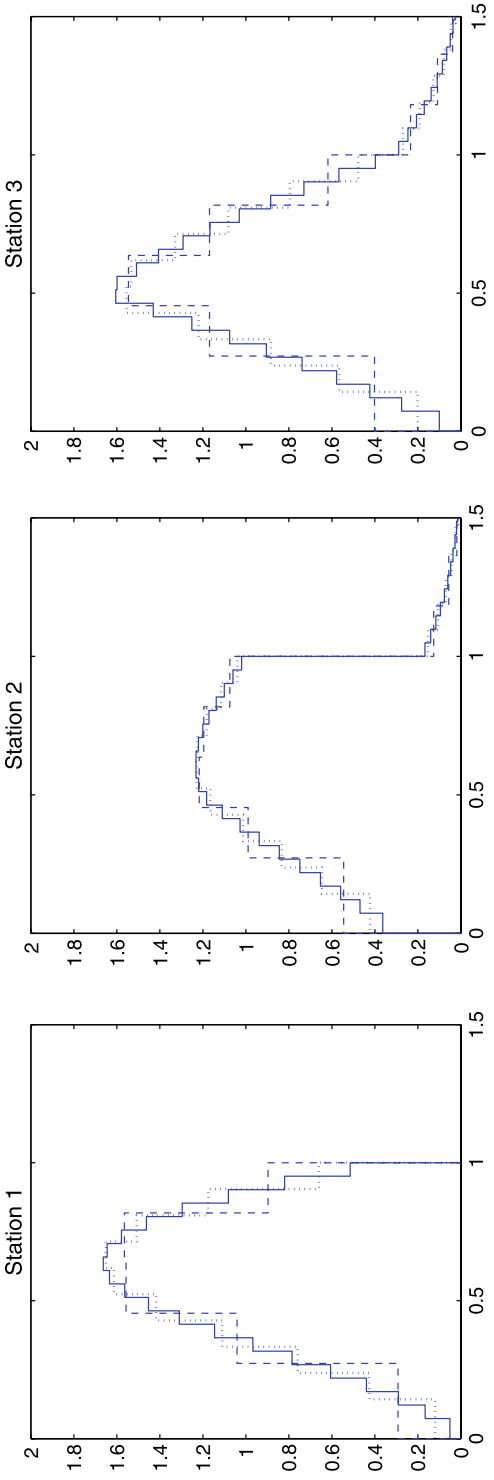


Fig. 6 Cycle time distributions, i.e. the distributions of the time between two items leaving a given station, computed for the three stations of a line, with 5 (*dashed*), 10 (*dotted*) and 20 (*solid*) discretization steps (with PMF/nIJ). It reveals that our modelling method allows to compute a good approximation of the distribution of the flow out of a line

uniform(0, 1) and triangular(0, 1, 0.5). Our method allows to compute discrete estimations of the cycle time distributions. Figure 6 depicts, for each station, the computed discrete cycle time distributions, i.e an estimation of the distribution of the time between two items going out of a station. The three graphs show how starving and blocking impact the cycle time (the distributions are different from the beta, uniform and triangular processing time distributions). On station 2 for example, the uniform distribution is sensibly modified: as the blocking and starving times are added, the cycle time is longer than the processing time, there is a smaller probability for a cycle time to be on the left, and a larger probability to be on the right hand side. The cycle time distributions are computed with 5, 10 and 20 steps. As our method refines and converges when the number of discretization steps increases, it can be supposed that the cycle time distribution computed with 20 steps is accurate. Figure 6 reveals that the shape of a cycle time distribution appears to be independent of the number of discretization steps used. The distribution is of course more detailed with 20 steps but the distributions computed with 5 or 10 steps show the same shape. It tends to show that the cycle time distribution estimations with 5 or 10 steps are already good, that their shape is converging.

In conclusion, we may say that the cycle time distributions computed by our modelling method have a shape close to the real one. Even with 5 or 10 steps, the approximation of the cycle time distribution is already good. The computation of a good cycle time distribution estimation is a significant advantage of our method, compared to other fitting methods which are unable to do so (see Pearson et al. 1979 for example). The distribution offers more detailed information on the behavior of the system, compared to the isolated expectation. It allows to estimate measures such as the variance or the percentiles. Note that other distributions, such as the distribution of the flow time, can also be computed. In practice, the percentiles are crucial information in order to help managers to evaluate the risk and the service level.

5 Computational results

In this section, we show by computational experiments how our modelling method behaves. The impact of the number of discretization steps and of the shift parameter α is shown for the bounds as well as for the estimation of the cycle time. Both probability masses fitting alternatives, with or without instant job, are distinguished. Moreover, the accuracy levels reached for the work-in-progress and for the flow time are given. We then compare these results to the computation time needed to reach them. After that, we study the impact of the configuration of the line, i.e. the number of stations and the buffer sizes, on the accuracy of the performance evaluation.

To begin, we study an assortment of 500 three-station flow lines. The buffer configurations vary from [0 0] up to [2 2]. They are equally shared out according to the global storage space (sum of both buffer sizes) and the space is uniformly divided, i.e. we analyze 100 [0 0] configurations, 100 either [1 0] or [0 1] configurations, 100 [1 1] configurations, 100 either [2 1] or [1 2] configurations, and 100 [2 2] configurations. The processing time distributions are randomly chosen among the 10 following distributions: uniform(0,1), beta(1.3,1), beta(2,2), beta(4,4), beta(5.5,6), beta(8,8), beta(10,9), triangular(0,1,0.5), triangular(0.2,1,0.3) and triangular(0.1,0.9,0.6). These distributions have various expectations and various coefficients of variation. The 500 lines are analyzed using both PMF alternatives (with or without instant job). Five different shift parameters α are used ($\alpha/\tau = 0, 0.25, 0.5, 0.75, 1$) as well as, with PMF/IJ, the parameters α that conserve the expectations of the original distributions (see Proposition 2). The number a of non-zero values, i.e. the number of phases, in the

Table 1 Average relative error, in percent, reached on the lower bound by the BDPH/IJ method. The number of non-zero values in the discrete distribution increases from left to right and the shift parameter α increases from top to bottom. The last row stands for α chosen in order to conserve the expectation of the original distributions

a	5	6	7	8	9	10
$\alpha/\tau = 0$	17.4	14.5	12.4	10.9	9.7	8.7
$\alpha/\tau = 0.25$	16.3	13.7	11.8	10.4	9.3	8.4
$\alpha/\tau = 0.5$	15.3	12.9	11.2	9.9	8.8	8
$\alpha/\tau = 0.75$	14.3	12.2	10.6	9.4	8.4	7.6
$\alpha/\tau = 1$	**13.4**	11.4	10	8.9	8	**7.3**
Exp. cons.	15.5	13.1	11.3	10	8.9	8.1

discretized distributions varies from five up to ten, which seems to be a reasonable choice.[2] We thus made a total of 33000 (500 lines $\cdot$ (5 nIJ + 6 IJ) $\cdot$ 6 a) experiments. The 500 three-station lines were also analyzed by simulation, in order to assess the accuracy of our method. In the following, we give the average relative error, in percent, between the results of our method and the results of the simulation (i.e. $|res_{bdph} - res_{simu}|/res_{simu}$). It is given for various parameter configurations and for various performance measures.

To begin, we are interested in the *bounds* on the cycle time. In Sect. 4.1.5, we argued that probability masses fitting with instant jobs offers better bounds. Moreover, we showed on a simple example that the lower bound appears to be better with $\alpha/\tau = 1$, i.e. the probability mass is aggregated in the beginning of the interval. This observation stays true on the computational experiments, as illustrated in Table 1. This table shows the accuracy of the lower bound computed with PMF/IJ as a function of the shift parameter α and the number a of non-zero values in the discretized distribution. It can be seen that the bound improves when a or α increases. However, as expected, the bounds are not accurate (it is known that the gap between both bounds equals the step size τ).

Of course, estimations of the *cycle time* can also be computed by the BDPH method. They should lead to better accuracy. The cycle time computed directly in the discretized time is not the best possible estimation (except whit shift parameters α that conserve the expectation). Indeed, as previously explained, in general, the distribution discretized by PMF does not have the same expectation as the original processing time distribution. But the error on the average processing time is known, as the expectations of both original and discretized distributions are known. It can thus be subtracted in order to get a better estimation of the cycle time, which is composed of the average processing time plus the average idle time. This estimation thus only includes the error on the idle (blocked/starved) time. Table 2 shows the accuracy reached with various parameters α and various number a of non-zero values in the discretized distributions. First of all, this table shows that our method leads to accurate estimations of the cycle time. The error made by the method amounts to a few tenths of a percent. Of course, the accuracy improves when the number of discretization steps increases. A 0.1 percent accuracy is reached with $a = 10$, while a 0.3 percent accuracy

[2]Lang and Arthur, in their extensive study (1997), mention 4 to 8 phases for "phase-type behaved" distributions and sensibly more (16) for other distributions. Moreover, concerning moments fitting, if we simply take the classical formula on the second moment into account, i.e. $1/k \leq cv^2 \leq 1/(k-1)$ (see Aldous and Shepp 1987), the 10 distributions used in our test would need 3, 5, 5, 9, 12, 17, 23, 6, 8 and 11 phases, in the same listing order.

Table 2 Average relative error, in percent, reached on the cycle time estimation by the BDPH/IJ method (upper table) and by the BDPH/nIJ (lower part). The number of non-zero values in the discrete distributions increases from left to right and the shift parameter α increases from top to bottom (in each part). The row "Exp. cons." stands for α that conserve the expectation. The computational times (in seconds) for the [2 2] buffer configuration are given in italic

a	5	6	7	8	9	10
$\alpha/\tau = 0$	0.63	0.53	0.46	0.42	0.37	0.33
$\alpha/\tau = 0.25$	0.35	0.26	0.22	0.20	0.18	0.16
$\alpha/\tau = 0.5$	**0.31**	0.22	0.18	0.14	0.12	**0.1**
$\alpha/\tau = 0.75$	0.51	0.39	0.33	0.28	0.25	0.22
$\alpha/\tau = 1$	0.78	0.63	0.53	0.46	0.41	0.37
Exp. cons.	**0.21**	0.15	0.13	0.1	0.09	**0.08**
Comp. time	*1.4* s	*2.1* s	*3.5* s	*5.7* s	*10* s	*18* s

a	5	6	7	8	9	10
$\alpha/\tau = 0$	0.63	0.53	0.46	0.42	0.37	0.33
$\alpha/\tau = 0.25$	0.52	0.38	0.31	0.27	0.23	0.21
$\alpha/\tau = 0.5$	**0.31**	0.21	0.16	0.13	0.11	**0.09**
$\alpha/\tau = 0.75$	0.86	0.6	0.47	0.38	0.31	0.27
$\alpha/\tau = 1$	1.21	0.87	0.68	0.56	0.48	0.41
Comp. time	*0.2* s	*0.3* s	*0.7* s	*1.7* s	*2.2* s	*5.8* s

is already obtained with five non-zero values in the discretized distribution. Note that these accuracy levels essentially reflect the relative error brought by the distribution fitting (first step), i.e. PMF, as the analytical modelling (second step of the global modelling) is exact. No results are available in the literature to directly compare these accuracy levels (see the introduction). We recall that a very simple example can be found where two moments fitting methods could not make the difference between two lines' cycle times while the real difference equals 1.62%. In this case, the PH distributions given by the moments fitting would include at least 8 phases. Furthermore, it can be seen that the best results are obtained with $\alpha/\tau = 0.5$ and with α which conserves the expectation. This is quite natural: aggregating the probability masses in the middle of the interval offers a better approximation of the distribution. This choice leads to a better estimation of the cycle time, which means a better estimation of the idle time. It can also be seen in Table 2 that the best absolute results are obtained choosing the shift parameter α in order to conserve the expectation of the original distribution, using PMF/IJ. However, when the computational time[3] is also taken into account, the best option shows to be PMF/nIJ with $\alpha/\tau = 0.5$, i.e. probability masses fitting aggregating in the middle of the interval, without instant job. Indeed, for example, an average error of 0.21% is reached in 0.3 seconds with PMF/nIJ while PMF/IJ needs 1.4 seconds. Similarly, PMF/nIJ leads to a relative error of 0.13% in 1.7 seconds while PMF/IJ takes 3.5 seconds.

Table 3 show the relative error obtained on two other performance measures: the *work-in-progress* (WIP) and the *flow time*. The WIP is easily computed from the occupations of the buffers and of the stations while the flow time is deduced from the Little's law. The

[3]Using the Gaussian elimination implemented in MATLAB® on a 2.16 GHz usual PC, 2 GB RAM.

Table 3 Average relative error, in percent, reached on the work-in-progress estimation (upper part) and on the flow time estimation (lower part), using PMF/IJ as well as PMF/nIJ. The number of non-zero values in the discrete distribution increases from left to right

a	5	6	7	8	9	10
PMF/IJ, $\alpha/\tau = 0.5$	0.38	0.27	0.21	0.17	0.15	0.13
PMF/IJ, Exp. cons.	0.2	0.16	0.13	0.12	0.11	0.1
PMF/nIJ, $\alpha/\tau = 0.5$	**0.36**	0.23	0.18	0.15	0.12	**0.11**
a	5	6	7	8	9	10
PMF/IJ, 0.5	0.48	0.34	0.26	0.21	0.18	0.16
PMF/IJ, Exp. cons.	0.26	0.2	0.17	0.14	0.13	0.12
PMF/nIJ, 0.5	**0.41**	0.27	0.21	0.17	0.14	**0.13**

table shows that our method leads to accurate estimations of these performances measures too. As for the cycle time, the level of accuracy is in the order of a few tenths of a percent. Again, the PMF/IJ with α that conserve the expectation leads to the better absolute results. When the computation times given in Table 2 are taken into account, it can be seen that, like for the cycle time, PMF/nIJ aggregating in the middle of the interval reveals to be the best choice. In the computation of the work-in-progress, for example, a relative error of 0.12% is reached in 2.2 seconds, compared to 5.7 seconds with PMF/IJ. Concerning the flow time, a error of 0.17% is obtained in 1.7 seconds, compared to 3.5 seconds.

At this point, it is interesting to study how the method behaves when the *configuration* of the line changes, i.e. when the number of stations and the buffer sizes vary. For this, we run a new set of experiments: 500 other flow lines are analyzed. Half of them is made of 2 stations and the other half is made of 4 stations. As for the three station lines previously studied, the total storage space (sum of every buffer sizes) vary from zero to four, and is uniformly shared out. The processing time distributions are randomly chosen among the 10 distributions listed previously. For this experiment, we focus on the cycle time estimation computed by the BDPH/nIJ method with $\alpha/\tau = 0.5$, i.e. aggregating the probability mass in the middle of the interval, which showed to be the best option in the previous experiments. Moreover, we chose a, the number of non-zero values in the discretized distributions, to be equal to eight. This choice leads to excellent accuracy in previous experiments (0.13%, see Table 2), while still increasing a only offers marginal benefit. An other choice of a would lead to the same conclusions. Again, the 500 new flow lines are also analyzed by simulation, in order to assess the accuracy of our method. We give the average relative error, in percent, between the results of our method and the results of the simulation.

From Table 4, the first conclusion is obvious: the accuracy of the estimations reveals to be remarkably stable. Whatever the dimension which varies, the number of stations or the buffer sizes, the relative error observed reveals to be nearby constant ($0.13 \pm 0.02\%$). In other words, the accuracy of the results does not deteriorate when the configuration gets more complex. This is another advantage of our method. Table 4 also shows the computational time needed. It illustrates what we already revealed previously: the complexity, due to the state model, is the main weakness of the global modelling method. The computational time increases very quickly when the system gets more complex.

To summarize, interesting observations can be drawn from the computational experiments. First, the best lower bound is offered by PMF/IJ aggregating in the beginning of the interval ($\alpha = \tau$). Second, and most importantly, our method computes accurate estimations

Table 4 Average relative error, in percent, reached on the cycle time estimation with PMF/nIJ aggregating in the middle of the interval. The results are given for lines with various number of stations and various total storage space (sum of every buffer size). The computational times are given in italic

$\sum b_i$	0	1	2	3	4	Total
2 stations	0.11	0.11	0.13	0.14	0.14	0.13
	0.1 s	*0.1* s	*0.1* s	*0.1* s	*0.1* s	
3 stations	0.15	0.12	0.13	0.12	0.11	0.13
		0.1 s	*0.2* s	*0.3* s	*0.4* s	*1.7* s
4 stations	0.15	0.12	0.12	0.13	0.12	0.13
	0.9 s	*1.6* s	*4.7* s	*54* s	*200* s	
Total	0.14	0.12	0.13	0.13	0.12	0.13

of the performance measures (cycle time, work-in-progress and flow time). Average relative errors are in the order of a few tenths of a percent. Even with five non-zero values in the discretized distributions, i.e. with five phases in the phase-type distributions, excellent accuracy levels are reached (0.3%), which reflect the quality of PMF (the analytical modelling is exact). Third, the accuracy level reached by the estimations shows to be remarkably stable when the system's configuration changes.

6 Conclusion

We introduce a new approach in order to build tractable phase-type distributions which are required by analytical models of stochastic manufacturing systems, and we study the global modelling method using it jointly with an exact analytical modelling. Probability masses fitting is simple: the probability masses on regular intervals are computed and aggregated on a single value in the corresponding interval, leading to a discrete distribution. The place where the value lies in the interval is chosen by the shift parameter α. Moreover, two alternative PMF are proposed, generating potential instant jobs, i.e. processing times of length zero, or not. Two characteristics follow directly from the definition. First, PMF is refinable: the approximation becomes more and more accurate when the interval size is decreased or, equivalently, when the number of intervals is increased. Second, PMF conserves the shape of the original distribution. Also note that PMF requires the processing time distributions to have finite support. As this is always the case in manufacturing systems, it can be seen as an advantage of PMF compared to other approaches. Two other properties of probability masses fitting were shown in the text. First, the discretized times allow to bound the original times. Second, the evolution of the discretized times in the shift parameter α is a monotonic, decreasing, function. Moreover, we showed that, using probability masses fitting with instant jobs, the shift parameter α can be chosen in order to conserve the mean of the original distribution.

After probability masses fitting, various state-of-the-art analytical models can be applied. In this paper, we chose a state model. In other words, from the discrete phase-type distributions built by PMF, the evolution of the system is modeled by a Markov chain and the performances are evaluated from it. A state model has the advantage to be exact but has a high complexity. The Markov chain size increases quickly when the system size increases. Here, the method is applied to manufacturing flow lines. The properties shown on probability

masses fitting can be extended to the global modelling method. The first remarkable result concerns bounds on the productivity. Using the concept of critical path, we prove refinable upper and lower bounds on the transient throughput and on the cycle time. Such bounds, for general processing times, have not been found by other methods. PMF with instant jobs leads to tighter bounds and, in particular, a better lower bound is obtained when the probability masses are aggregated in the beginning of the interval. Moreover, the monotonicity of PMF can be extended to the cycle time computed by the global method. Finally, interestingly, the proposed method allows to compute a good approximation of the cycle time distribution, what is a far more detailed information than the isolated expectation and cannot be computed by other methods. In practice, this feature allows to compute percentiles and helps the manager to evaluate the risk and the service level.

On computational experiments, our modelling method proves to offer accurate estimations of various performance measures, namely the cycle time, the work-in-progress and the flow time. Using only five discretization steps, an accuracy level of 0.3% is reached. This reflects the quality of PMF, even with a limited number of phases, as the subsequent analytical modelling is exact. Comparing various parameters options, PMF without instant job aggregating the probability mass in the middle of the interval appears to be the best choice, in the sense of the trade-off between accuracy and computational time. When the configuration of the line (number of stations and buffer sizes) is changed, the results' accuracy reveals to be remarkably stable.

Consequently, we believe that probability masses fitting can be considered as a valuable alternative in order to build tractable distributions for the analytical modelling of manufacturing systems. Moreover, it has interesting consequences on the global modelling method. Accordingly, the application of probability masses fitting could bring about much interesting *future work*. In order to improve its applicability, and avoid the complexity of state models, approximate analytical models could be applied after PMF. In particular, decomposition methods could benefit from the computation of detailed approximations of the cycle time distributions for each station. Moreover, the probability masses fitting approach could of course be applied to more complex manufacturing systems, such as assembly/disassembly systems. Finally, we believe that the ability to compute a good approximation of the distribution of the cycle time is an important advantage that could be further investigated and exploited. In particular, one could try to prove convergence results for the shape of the cycle time distribution.

Acknowledgements The work of J.-S. Tancrez has been funded by the European Social Fund and the Walloon Region of Belgium. It is made in collaboration with ArcelorMittal Fontaine. J.-S. Tancrez is also grateful to Prof. F. Riane for his unfailing support. The authors are also grateful to the anonymous referees for their helpful remarks.

References

Adan, I. J., van Eenige, M. J., & Resing, J. A. (1995). Fitting discrete distributions on the first two moments. *Probability in the Engineering and Informational Sciences*, *9*, 623–632.

Aldous, D., & Shepp, L. (1987). The least variable phase type distribution is Erlang. *Stochastic Models*, *3*, 467–473.

Altiok, T. (1985). On the phase-type approximations of general distributions. *IIE Transactions*, *17*(2), 110–116.

Altiok, T. (1996). *Performance analysis of manufacturing systems*. New York: Springer.

Asmussen, S., Nerman, A., & Olsson, M. (1996). Fitting phase-type distributions via the em algorithm. *Scandinavian Journal of Statistics*, *23*, 419–441.

Baccelli, F., & Liu, Z. (1992). Comparison properties of stochastic decision free Petri nets. *IEEE Transactions on Automatic Control*, *37*, 1905–1920.
Baccelli, F., & Makowski, A. M. (1989). Queueing models for systems with synchronization constraints. *Proceedings of the IEEE*, *77*, 138–161.
Bobbio, A., & Cumani, A. (1992). Ml estimation of the parameters of a ph distribution in triangular canonical form. In G. Balbo & G. Serazzi (Eds.), *Computer performance evaluation* (pp. 33–46). Amsterdam: Elsevier.
Bobbio, A., Horváth, A., & Telek, M. (2004). The scale factor: a new degree of freedom in phase-type approximation. *Performance Evaluation*, *56*(1–4), 121–144.
Bobbio, A., Horváth, A., Scarpa, M., & Telek, M. (2003). Acyclic discrete phase type distributions: properties and a parameter estimation algorithm. *Performance Evaluation*, *54*, 1–32.
Bobbio, A., Horváth, A., & Telek, M. (2005). Matching three moments with minimal acyclic phase type distributions. *Stochastic Models*, *21*, 303–326.
Botta, R. F., & Harris, C. M. (1986). Approximation with generalized hyperexponential distributions: weak convergence results. *Queueing Systems*, *1*(2), 169–190.
Buzacott, J. A., & Shanthikumar, J. G. (1993). *Stochastic models of manufacturing systems*. Englewood Cliffs: Prentice-Hall.
Dallery, Y., & Frein, Y. (1993). On decomposition methods for tandem queueing networks with blocking. *Operations Research*, *41*(2), 386–399.
Dallery, Y., & Gershwin, S. B. (1992). Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems*, *12*, 3–94.
Dallery, Y., Liu, Z., & Towsley, D. (1994). Equivalence, reversibility, symmetry and concavity properties in fork/join queueing networks with blocking. *Journal of the ACM*, *41*, 903–942.
Gershwin, S. B. (1994). *Manufacturing systems engineering*. Englewood Cliffs: Prentice-Hall.
Hillier, F. S., & Boling, R. W. (1967). Finite queues in series with exponential or Erlang service times—a numerical approach. *Operations Research*, *15*(2), 286–303.
Johnson, M. A., & Taaffe, M. R. (1989). Matching moments to phase distributions: Mixtures of Erlang distributions of common order. *Stochastic Models*, *5*(4), 711–743.
Kerbache, L., & MacGregor Smith, J. (1987). The generalized expansion method for open finite queueing networks. *European Journal of Operational Research*, *32*, 448–461.
Klincewicz, J. G., & Whitt, W. (1984). On approximations for queues, ii: Shape constraints. *AT&T Bell Laboratories Technical Journal*, *63*(1), 139–161.
Lang, A., & Arthur, J. L. (1997). Parameter approximation for phase-type distributions. In S. R. Chakravarthy & A. S. Alfa (Eds.), *Matrix analytical methods in stochastic models* (pp. 151–206). New York: Dekker.
Lindsay, B. G., & Basak, P. (2000). Moments determine the tail of a distribution (but not much else). *The American Statistician*, *54*(4), 248–251.
Marie, R. (1980). Calculating equilibrium probabilities for λ(n)/ck/1/n queues. In *PERFORMANCE '80: Proceedings of the 1980 international symposium on computer performance modelling, measurement and evaluation* (pp. 117–125). New York: ACM.
McCullagh, P. (1994). Does the moment-generating function characterize a distribution? *The American Statistician*, *48*(3), 208.
Osogami, T., & Harchol-Balter, M. (2006). Closed form solutions for mapping general distributions to quasi-minimal ph distributions. *Performance Evaluation*, *63*(6), 524–552.
Papadopoulos, H. T., & Heavey, C. (1996). Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. *European Journal of Operational Research*, *92*, 1–27.
Pearson, E. S., Johnson, N. L., & Burr, I. W. (1979). Comparisons of the percentage points of distributions with the same first four moments, chosen from eight different systems of frequency curves. *Communications in Statistics–Simulation and Computation*, *8*(3), 191–229.
Perros, H. G. (1994). *Queueing networks with blocking: exact and approximate solutions*. Oxford: Oxford University Press.
Sauer, C. H., & Chandy, K. M. (1975). Approximate analysis of central server models. *IBM Journal of Research and Development*, *19*(3), 301–313.
Schmickler, L. (1992). Meda: mixed Erlang distributions as phase-type representations of empirical distribution functions. *Stochastic Models*, *8*(1), 131–156.
Tancrez, J.-S., & Semal, P. (2006). The bounding discrete phase-type method. In A. N. Langville & W. J. Stewart (Eds.), *MAM 2006: Markov anniversary meeting* (pp. 257–269). Raleigh: Boson.
Tancrez, J.-S., Semal, P., & Chevalier, P. (2008, in press). Histogram based bounds and approximations for production lines. *European Journal of Operational Research*. doi:10.1016/j.ejor.2008.03.032.
Telek, M. (2000). Minimal coefficient of variation of discrete phase type distributions. In *3rd international conference on matrix-analytic methods in stochastic models, MAM3* (pp. 391–400). Leuven: Notable Publications.

Varah, J. M. (1985). On fitting exponentials by nonlinear least squares. *SIAM Journal on Scientific and Statistical Computing*, *6*(1), 30–44.
Whitt, W. (1982). Approximating a point process by a renewal process, i: Two basic methods. *Operations Research*, *30*(1), 125–147.
Whitt, W. (1984). On approximations for queues, i: Extremal distributions. *AT & T Bell Laboratories Technical Journal*, *63*(1), 115–138.