Effective process times for multi-server flowlines with finite buffers

Kock, A.A.A.; Etman, L.F.P.; Rooda, J.E.

Published: 01/01/2006

Systems Engineering Group
Department of Mechanical Engineering
Eindhoven University of Technology
PO Box 513
5600 MB Eindhoven
The Netherlands
http://se.wtb.tue.nl/

SE Report: Nr. 2006-08

Effective process times for multi-server flowlines wit finite buffers

A.A.A. Kock, L.F.P. Etman and J.E. Rooda

ISSN: 1872-1567

SE Report: Nr. 2006-08
Eindhoven, February 2006
SE Reports are available via http://se.wtb.tue.nl/sereports

Abstract

An effective process time (EPT) approach is proposed for aggregate model building of multi-server tandem queues with finite buffers. Effective process time distributions of the workstations in the flow line are measured without identifying the contributing factors. A sample path equation is used to compute the EPT realizations from arrival and departure events of lots at the respective workstations. If the amount of blocking in the line is high, the goodness of the EPT distribution fits determines the accuracy of the EPT-based aggregate model. Otherwise, an aggregate model based on just the first two moments of the EPT distributions is sufficient to obtain accurate predictions. The approach is illustrated in an industrial case study using both simulation and analytical queueing approximations as aggregate models.

1 Introduction

Multi-server tandem queues with finite buffers commonly occur in industrial practice. The performance of such lines is typically expressed in terms of throughput and flow time. Irregularities in processing play a key role in the throughput and flow time performance. Due to the limited buffer capacity, blocking of workstations may occur.

For the performance prediction of finitely buffered multi-server tandem queues, typically discrete event simulation models (e.g. [3, 4, 14]) or queueing models (e.g. [9, 6, 15, 24, 22]) are used. Simulation models are usually more accurate than queueing models since they can incorporate more shop-floor realities. On the other hand, queueing models are often computationally far less expensive. Both types of models have to be fed with appropriate data regarding processing, disturbances, and other realities that occur on the shop-floor. Common methods in literature either assume a distribution or measure individual influences on processing [7, 9, 6].

In industrial practice, it is often hard to identify and quantify all relevant shop-floor details that contribute to the flow time performance of the workstations. [12, 13] present an algorithm to obtain effective process time distributions for infinitely buffered workstations from lot arrivals and departures. The advantage of their method is that it does not require the quantification of the individual contributing factors. The motivation of their work is to arrive at a measurable metric for variability at a workstation (variance in processing).

In this paper we generalize this concept to build EPT-based aggregate queueing models of finitely buffered, multi-server tandem flow lines. Using the aggregation based on the effective process time paradigm [11], we aim to arrive at simplified queueing models, either simulation or analytical, for which the aggregate process distribution parameters can be obtained from event data on the shop-floor, such as arrivals and departures.

The contribution of the paper is twofold. First, we show that a sample path equation can be used to compute EPT-realizations in multi-server workstations with blocking. Second, we investigate the effect of the shape of the EPT distribution fit on the accuracy of the EPT-based aggregate queueing model. In particular we consider the offset (i.e., the smallest EPT-realization that was measured) as third distribution parameter in a shifted gamma distribution next to the EPT mean and variance. The accuracy of both the mean flow time and the variance of flow time prediction are considered.

The paper is outlined as follows. First we present our proposed aggregate modeling approach using the effective process time, and give considerations about the applicability of the aggregation. Then the calculation of the effective process time is presented. This is followed by several examples to experimentally investigate the role of the shape of the EPT distribution fit on the accuracy of the aggregate model prediction. Next, in an industrial case problem, the use of EPT-distributions in queueing models and simulation models is illustrated. Finally the main conclusions and some remarks on future work are offered.

2 Aggregate modeling using the effective process time

For the prediction of flow line performance, queueing models are used. Two well-known classes of models are discrete-event simulation models and analytical queueing models.

A simulation model is the imitation of the operation of an actual real-world system [4], in our case a manufacturing flow line. In a simulation model, the various shop-floor realities may be modeled in detail. As an example we cite [3] who included operator behavior in their model. Often it is tried to include the most important details in the model to arrive at an accurate simulation model representation of the factory floor. A drawback is that running a simulation model to obtain statistically relevant outcomes may become computationally expensive. An additional difficulty is to get all required data regarding the shop-floor details in the model. In practice, some of the data may be hard to get.

Analytical queueing models are an interesting alternative to simulation models. One may distinguish exact and approximative analytical models. Examples can be found in [9, 6, 15, 24] and [22]. Such analytical models cannot be as detailed a description as simulation models. These models must adhere to rather restrictive assumptions regarding the inclusion of shop-floor realities. However, if one can keep the number of states in the model limited, analytical queueing models are cheap to evaluate compared to a simulation model. In some cases even exact or explicit approximative expressions can be derived. Even though the number of parameters in analytical models is typically much smaller than in simulation models, feeding the model with appropriate data is nevertheless not trivial.

We aim at an aggregate modeling approach that enables one to obtain its parameters from simple events such as lot arrivals and departures, readily measurable from the shop-floor. For this we start from the effective process time as aggregate process time distribution.

2.1 Concept

The effective process time aggregates the raw processing time and all the shop-floor realities and disturbances hampering the progress of the processing, into a single process time distribution. Examples of realities and disturbances are machine downs, setup, rework, operator availability, lot size, metrology, tool change, etcetera. The inclusion of multiple phenomena into a single distribution is referred to as aggregation. The phrase effective process time was introduced by [11], although the concept of aggregation is of course not new. Hopp and Spearman defined the effective process time of a lot as 'the time spent by the lot on a workstation from a logistical point of view'. They give explicit expressions to compute the mean EPT and the EPT coefficient of variation from the various outages, either preemptive or non-preemptive. They use the EPT mean and the EPT variance in explicit queueing approximation equations, such as Kingman's equation, to estimate and explain the mean flow time performance.

In many practical cases, the outages may not all be quantifiable. Nevertheless, aggregation such as the EPT is appealing, in particular if the EPT can be measured without identifying the contributing factors. For workstations with infinite buffers, a method to actually do this was first proposed by [12, 13]. From lot arrival and departure events they calculate for each departing lot an EPT realization. By collecting consecutive EPT realizations, a workstation EPT distribution is obtained. All influences on processing at the workstation are then aggregated into the EPT distribution.

This idea may be further generalized into an EPT-based aggregate modeling framework. Then the EPT is not only used as a performance metric quantifying the effective workstation capacity (mean) and variability (variance), but also to build an aggregate simulation or analytical queueing model, so the idea is that the EPT is a measurable quantity on the factory floor while the aggregate queueing model can stay simple and is feeded directly with parameter values obtained from the measured EPT distributions. The basic approach we propose is:

Step 1 Measure arrival and departure events at the workstations in the manufacturing system, and, for multi-server workstations register which lot has been processed on which machine.

Step 2 Translate the events into EPT-realizations, one for each departing lot.

Step 3 From the EPT realizations, compute mean and variance.

Step 4 Build an aggregate queueing model, either simulation or analytical, using the measured EPT means and variances of the workstations.

In this paper we develop the EPT-based aggregate modeling approach for multi-server tandem flow lines subject to blocking. Blocking refers to the situation where a lot cannot be sent away since the receiving buffer of the subsequent station is full. As a consequence, the server cannot commence processing a new lot. Blocking can have a large impact on throughput and flow time performance.

For the aggregate model building of flow lines with blocking we will in particular consider approximative analytical queueing methods such as developed by [22] and [21]. These methods require as input for the workstations the mean and variance of the process time for which we will obviously use the EPT mean and variance. Van Vuuren *et al.*, demonstrated using a range of test problems, the accuracy of their approximation compared to a simulation model representation. A clear advantage of such an analytical approximation is the speed of evaluation compared to running a simulation model.

In the sequel we will use the following notations and definitions. The mean of the EPT distribution is denoted t_e. The ratio of m (number of parallel machines in a workstation) and t_e quantifies the mean effective capacity available at the workstation. The ratio of the raw processing time t_0 and the mean effective process time t_e quantifies capacity loss. The latter ratio relates to the industry metric OEE (see e.g. [18]) and the revision E proposed by [17]. The squared coefficient of variation of the EPT distribution is denoted c_e^2. Following [11] we refer to this as a quantification of variability in processing. We call the model in which certain shop-floor realities are not included explicitly but represented by an aggregate EPT-distribution, an EPT-based aggregate model or simply an EPT-based model. The structure of the EPT-based model (i.e., material flows, number of workstations, number of servers per workstation and number of bufferplaces) is identical to the original system (or detailed model of the original system). Finally, the queuing performance is expressed in throughput (δ [lots/hour]) and flow time (φ [hour]).

2.2 Considerations

For certain cases, shop-floor realities may be aggregated without great loss of accuracy. For an M/G/1 and M/G/n workstation the mean flow time depends solely on the first two moments of the process time distribution. For a multi-server station with generally distributed arrivals (G/G/n) this remarkable property is approximately still valid, provided that service times and arrivals are phase-type distributed [1, 23].

The performance is predicted exact as long as the first two moments of the process time distribution are known, regardless of the shape of the distribution function. This implies that it is sufficient to fit a two-moment distribution (e.g. a Gamma distribution) to the measured EPT realizations.

For finitely buffered flow lines this shape independence property may not hold anymore. As a consequence, the first two moments (mean and variance) may not suffice to obtain accurate predictions from the aggregate queueing model. Then the EPT distribution has to be described more accurately by using a higher order distribution fit. For instance, in most manufacturing lines, processing at the workstations takes at least some minimum time. The shift or offset may be included as third parameter in the distribution fit to account for this, e.g. using a shifted Gamma or other type of distribution. In Section 4 we investigate in further detail the contribution of the offset to the mean flow time for flow lines subject to blocking.

Alternatively, one may decide to include one or more shop-floor realities explicitly in the aggregate queueing model. For instance, if certain lot types give rise to different processing characteristics, one can fit a separate (two-moment) distribution for each lot type. The lot type then becomes an integral part of the aggregate model. For a simulation aggregate model, this poses no additional difficulties. For an analytical aggregate model, new model equations may need to be derived to account for the shop-floor reality that becomes part of the aggregate model (lot type in the example).

One may also want to leave out a certain shop-floor reality from the EPT distribution entirely, and measure and model it separately. This happens when the time scales of events are different. For instance, when the machines are highly reliable, machine downs occur only very infrequently. Then it may happen that for the measurement period under consideration, one may have produced thousands of lots (thus obtained the same amount of EPT realizations) while only a couple of machine downs have occurred. If the downs have a considerable effect on the shape of the EPT distribution, but only few actual down events occur, then no statistically reliable distribution parameter estimates can be obtained. Data on the down behavior should then be collected separately on a different time scale, and be excluded from the EPT. Again, the down then has to be modeled explicitly in the aggregate model. Note in this respect the analytical queueing approximations developed by [19].

Taking these considerations into account, the EPT approach may be rephrased as:

Step 0 Define the structure of the model, and define which shop-floor realities or disturbances are modeled explicitly and excluded from aggregation in the EPT.

Step 1 Measure arrival and departure events at the workstations in the manufacturing system; for multi-server workstations register which lot has been processed on which machine; obtain data regarding the explicit realities.

Step 2 Translate the events into EPT-realizations, one for each departing lot.

Step 3 Fit for each workstation a suitable distribution to the measured EPT realizations.

Step 4 Build an aggregate queueing model, either simulation or analytical, using the fitted EPT-distributions.

Step 5 If the EPT model is sufficiently accurate, stop. Otherwise, return to Step 4 to reconsider the distribution fitting or go back to Step 0 to reconsider the aggregation.

Preferably we start with building the simplest possible model, and refine when necessary. The accuracy of an EPT-based model may be validated by comparing the estimated throughput and flow time to the throughput of the actual system and the flow time of the lots in the actual system. We will mainly focus on mean throughput and mean flow time. Higher moments may also be considered but, as we will show, the required quality of the EPT distribution fit regarding the actual shape becomes more pronounced.

2.3 Application

Once a suitable EPT-based model is obtained, it can serve two main purposes.

First, the obtained EPT parameters provide insight in the performance of the flow line. Parameter t_e details the average amount of time claimed by a lot at the workstations. The workstation that has the lowest effective capacity is the actual bottleneck. Parameter c_e^2 quantifies the amount of variability associated with the effective processing of lots. Workstations with a high value for c_e^2 may be a problem since they interrupt the steady flow of lots.

Secondly, the EPT-based model may be used to predict the effect of changes in the line configuration or in numerical optimization procedures. Accurate but quick to evaluate models are then a prerequisite. An analytical model compared to a simulation queueing model has a great advantage here.

3 EPT calculation

[13] (2001,2003) compute EPT-distributions for infinitely buffered multi-server workstations in isolation. They present an EPT algorithm that computes an EPT realization for each departing lot. Their algorithm is based on the observation that as long as there are lots in the workstation capacity is claimed. Each arriving lot starts a new capacity claim if the number of lots in the workstation is less than the number of installed servers. Each departing lot ends a capacity claim. So the number of ongoing capacity claims equals the maximum of the number of lots in the system and the number of servers. The method proposed by Jacobs *et al.* also incorporates time losses due to dispatching issues (assignment of lots to machines) in the EPT, for instance the case that a server should be available for processing but none of the lots waiting in the queue is ever processed on that particular machine. We will refer to this as a violation of the EPT-nonidling assumption as we will explain later in this section.

Workstations subject to blocking cannot be considered in isolation. We therefore follow a different approach to calculate the EPTs. We show that a simple sample path equation can be used to compute the EPT realizations in a flow line subject to blocking. The key observation when blocking is present is that the EPT excludes time losses due to blocking. Blocking is excluded since it is due to the finity of the buffers. The EPT-based model will also have the same finite buffers, which means that the blocking phenomenon is already covered in the structure of the EPT-based aggregate model itself. For similar reasons, starvation of a workstation should not be included in the EPT.

3.1 EPT for finitely buffered, single server workstations

The EPT for a finitely buffered workstation is computed using three events: the possible departure $\mathbf{PD}_{i,j}$ (the time-epoch at which workstation j finishes processing lot i and tries to send it on to the next workstation in the line), the actual departure $\mathbf{AD}_{i,j}$ (the time-epoch at which lot i physically leaves workstation j) and the actual arrival $\mathbf{AA}_{i,j}$ (the time-epoch at which the lot with the ith actual departure enters (the buffer of) workstation j). If no blocking occurs, $\mathbf{PD}_{i,j}$=$\mathbf{AD}_{i,j}$ holds since the receiving workstation has sufficient capacity available to receive the lot. Note that, if transport is instantaneous, $\mathbf{AD}_{i,j}$ equals $\mathbf{AA}_{i,j+1}$.

An EPT-realization ends upon the possible departure of the respective lot. The EPT-realization begins as soon as the workstation could have started processing the lot, that is at the maximum of the moment that the lot arrived in the buffer or the moment that the preceding lot has left. So the EPT-realization begins at $\max\left\{\mathbf{AA}_{i,j}, \mathbf{AD}_{i-1,j}\right\}$ and ends at $\mathbf{PD}_{i,j}$. The EPT-realization can then be computed from:

$$\mathbf{EPT}_{i,j} = \mathbf{PD}_{i,j} - \max\left\{\mathbf{AA}_{i,j}, \mathbf{AD}_{i-1,j}\right\}. \tag{1}$$

which is a reverse use of the sample path equation for finitely buffered, single-server workstations ([6] or [2]); instead of computing departure events, we compute EPT realizations.

3.2 EPT for finitely buffered, multi-server workstations

Calculation of EPTs for multi-server workstations subject to blocking can be done using the same equation: Sort the processed lots by the machine they were processed on; then apply Equation (1) for each machine in the workstation.

This approach of calculating the EPT realizations per machine assumes that waiting lots will be processed on the next available machine. This is often referred to as the non-idling assumption. Note that in our case the non-idling assumption has to be interpreted from the EPT point of view. From an EPT point of view the state of a machine that finishes processing a lot changes from busy to available. The machine is from an EPT point of view busy again when the next lot to be processed is present in the queue. Actual loading of the lot on the machine may be delayed for whatever reason.

The EPT-nonidling assumption is violated when a machine comes available and lots are present in the buffer but none of these will be processed on the respective machine. By applying Equation (1) for each machine separately this particular loss of capacity is not accounted for in the EPT and has to be accounted for separately. This case will not be considered further in this paper.

Finally, if we have an infinitely buffered workstation instead of a finitely buffered one, $\mathbf{PD}_{i,j}$ may be replaced by $\mathbf{AD}_{i,j}$ in Equation (1). When the EPT-nonidling assumption is satisfied, then it can be shown that using Equation (1) is equivalent to the algorithm proposed by [13].

4 Examples

In this section, the applicability of the EPT-method for finitely buffered, multiple server flowlines is evaluated using several examples. First, we briefly illustrate that Equation (1) provides the correct EPT-parameters. Next, we show that EPT-based models for finitely buffered flow lines may require more input than just the first two moments of the EPT distribution. We study this more extensively for the 'offset' as third distribution parameter. Finally, we show that the variance of the flow time distribution may also be approximated using the EPT approach.

4.1 Validation of Equation (1)

Consider a two-workstation flow line. The first workstation, which consists of a single server, is never starved. The service time at the first workstation is exponentially distributed with mean process time $\lambda^{-1} = 1.00$ [hr/lot]. The second workstation, which is never blocked, consists of two (identical) parallel servers and a single buffer space. The process times are again exponentially distributed, with mean process time $\mu^{-1} = 2.05$ [hr/lot].

Following the EPT approach, events are measured per lot per workstation. These events are the actual and possible departures, and the arrivals. The collected events are used as input for Equation (1), with which EPT realizations are computed. The gathered EPT realizations are represented as gamma distributions. For the first workstation, the mean effective process time we measure is $t_{e,0} = 1.0$ [hr]; whereas the squared coefficient of variation is $c^2_{e,0} = 1.0$. For the second workstation, parameters $t_{e,1} = 2.05$ [hr] and $c^2_{e,1} = 1.0$ are measured. These values correspond with the input given above.

4.2 Influence of the EPT-distribution shape

Consider a line consisting of three unbuffered workstations. The first workstation is never starved, the third workstation is never blocked. The first workstation contains one machine, the second and third workstation each contain two machines. The clean process time on the first workstation is triangularly distributed with minimum 0.9, maximum 1.1 and modus 1.0. On the second and third workstation, the process time is also triangularly distributed, but now with minimum 1.8, maximum 2.2 and modus 2.0.

On all machines, a setup is required after every 10th lot that has been processed. A setup is triangularly distributed with minimum 0.5, maximum 1.5 and mean 1.0. Machines are prone to failure. The busy time between failures is exponentially distributed on each machine with mean $t_f = 15.0$. After a failure, the machine should be repaired. The repair time is exponentially distributed with mean $t_r = 3.0$. After a repair, processing of the lot is resumed where it was left. For this system, the simulated mean flow time is $\varphi = 7.111$. The 95% confidence interval of the simulation results presented in this section is less than 1% of the corresponding parameter.

From this system, EPT-realizations were obtained using Equation (1). The mean and variance of the distributions were $t_{e,0} = 1.292$, $c^2_{e,0} = 0.777$, $t_{e,1} = 2.490$, $c^2_{e,1} = 0.400$ and $t_{e,2} = 2.492$, $c^2_{e,2} = 0.405$ for the three workstations respectively. These values were inserted in an EPT-based model. The model approximates $\tilde{\varphi} = 7.563$. Hence, it overestimates the flow time by 6.4%.

From our measurements, we know that in the real system, the smallest EPTs measured at the workstations (referred to as offset) were respectively $\Delta_0 = 0.9$, $\Delta_1 = 1.8$ and $\Delta_2 = 1.8$. However, this knowledge is not used in the EPT-based model. By fitting a shifted gamma distribution ([8]), this offset can be included in the EPT model. The estimated parameters of the shifted gamma distribution are $\Delta_{e,0} = 0.9$, $t_{e,0} = 1.292$, $c^2_{e,0} = 0.777$, $\Delta_{e,1} = 1.8$, $t_{e,1} = 2.490$, $c^2_{e,1} = 0.400$ and $\Delta_{e,2} = 1.8$,$t_{e,2} = 2.492$, $c^2_{e,2} = 0.405$. Then, the EPT-based model approximates $\tilde{\varphi} = 7.223$. Now, the mean flow time is only overestimated by 1.6%. Inclusion of the offset here improves the accuracy of the EPT model.

4.3 Relevance of the offset

In many practical cases, a minimum (positive) value for the process time distribution is present (processing requires at least a fixed minimum amount of time). As the previous example illustrates, for flow lines subject to blocking the shape of the EPT distribution may need to be represented in more detail than just using the first two moments to obtain a sufficient prediction accuracy of the EPT-based model. In this subsection, we experimentally investigate the contribution of the offset. Our hypothesis is that the shape of the process time distribution (i.e. inclusion of the offset in this example) becomes increasingly important when flow times on one workstation heavily affect flow times on other workstations, i.e. when blocking occurs. The stronger the effect of blocking is, the stronger we expect the shape of the EPT distribution fit to impact the accuracy of the EPT-based model.

First, consider a three-workstation flow line with one server per workstation. Process times are distributed with a shifted gamma distribution with mean 1.0 and squared coefficient of variation of 1.0. The offset (or shift) is taken at 0.0 and 0.9. In Figure 1, we see that the influence of the offset is reduced if the buffersize is increased for both throughput and flow time. Increasing the buffer level corresponds to decreasing the amount of blocking. Hence, this observation confirms our hypothesis.

Insert Figure 1 about here

Next, consider a ten-workstation flow line with $n \in \{1..10\}$ servers per workstation. Each workstation has one bufferplace. Process times are distributed according to a shifted gamma distribution with mean 1.0 and a squared coefficient of variation of $c_{\mathrm{e}}^2 \in \{0.5, 1.0, 2.0\}$ and offsets (shifts) of 0.0 and 0.9 respectively. The results are displayed in Figure 2. Herein, $d_\delta = \frac{\delta_{\Delta=0.0}-\delta_{\Delta=0.9}}{\delta_{\Delta=0.0}}$ and $d_\phi = \frac{\varphi_{\Delta=0.0}-\varphi_{\Delta=0.9}}{\varphi_{\Delta=0.0}}$. From this figure, we see that the influence of the offset becomes smaller as there are more parallel servers in the system. Including extra parallel servers leads to a reduction of blocking. Again, this observation confirms our hypothesis. The second observation from Figure 2 is that, if the level of variability in the line (i.e. c_{e}^2) is reduced, the relevance of the offset also becomes smaller. Reducing the variability implies that the level of blocking is also reduced. Hence, again our hypothesis is confirmed.

Insert Figure 2 about here

From these experiments, we conclude that the offset only needs to be included in the EPT distribution fit if the amount of blocking is high, that is, for few parallel servers, small buffer sizes, and high levels of variability. Otherwise, an EPT distribution fit with just the mean and variance is sufficient. This does not only hold for the offset but for the distribution shape in general. The advantage is then that analytical queueing models based on the first two moments of the process time distribution, such as proposed by [22] and [21], can be used.

4.4 Estimation of the variance of the flow time

Estimation of the variance of the flow time is relevant for instance in the context of customer reliability. In this example, we experimentally investigate the possibility to estimate the second moment of the flow time. Reconsider the three workstation example of Section 4.2, where the first workstation consisted of one server, while the second and third workstation both had two servers. All three workstations are unbuffered. For that system, we obtained $\varphi = 7.111$. The variance of the flow time can also be measured: $S_\varphi^2 = 8.611$.

If we build an EPT-based model using solely t_{e} and c_{e}^2, then we approximate $\tilde{\varphi} = 7.563$ and $\tilde{S}^2{}_\varphi = 6.457$, which are respectively an overestimation of 6.4% and an underestimation of 25 %. By explicitly including the offset in the EPT-based model using a shifted gamma distribution, we approximate $\tilde{\varphi} = 7.223$ and $\tilde{S}_\varphi^2 = 8.120$, an overestimation of 1.6% and an underestimation of 5.7% respectively.

Including more detail in the distribution fit further enhances the accuracy of the EPT-based model. Therefore, using the work of [16], we fit a shifted Erlang-Coxian distribution to the EPT of a machine. Then, we obtain $\tilde{\varphi} = 7.118$ and $\tilde{S}_\varphi^2 = 8.434$, an overestimation of 0.1% and an underestimation of 2.1% respectively. We see that describing the EPT distribution in greater detail, the prediction accuracy of the EPT model increases. To accurately predict the variance in the flow time a more detailed distribution fit is required compared to predicting just the mean flow time.

5 Industrial case

The proposed method is tested on a case inspired by industry practice. The industrial case considers a manufacturing line for lamp sockets, see [20]. The layout of the case is shown in Figure 3.

Insert Figure 3 about here

In supply station S_0, sheets of aluminum are die-cut into small cilinders. The rolls of aluminum-sheet arriving at S_0 are large enough to safely assume that S_0 is never starving. The lots of cylinders are transported to W_0, where screw thread is cut in the cilinders. Next, the lamp sockets are placed inside a glass-oven (W_1), where a small amount of liquid glass is poured into the sockets. In W_2, the finishing bath, the socket is bathed in a solvent of nickel or stain. Finally, in W_3, lots are packed into carton boxes and cleared away for shipping. It is assumed that W_3 is never blocked.

Workstation S_0 has two parallel servers. In W_0, lots can be placed in a finite buffer of capacity two; the workstation has four parallel machines. W_1 has a finite buffer of capacity four, and one server. W_2 has a single server and a single bufferspace; finally W_3 has a single server and four buffer spaces. Note that each single lot in this case corresponds to 6000 bulbs. The process times are approximately constant on the workstations, aside from the failure behavior. The time consumed by a lot on the workstation is thus accurately captured by the clean process time, the busy time between failures (exponentially distributed) and a description of the failure behavior.

In this paper, failure behavior is assumed that consists of up to two exponentially distributed stages. First, when a machine breaks down an operator will check whether he can make an emergency repair, with rate λ_0. With probability p, the emergency repair suffices and the machine is fixed. With probability $1 - p$, the repair is not sufficient and a professional mechanic has to be notified. This mechanic repairs the machine in the second stage with rate λ_1, and repairs the machine with probability 1. The respective parameters for all workstations are presented in Table 1. In the table, b refers to the number of bufferspaces per workstation, m refers to the number of parallel machines, μ_0 is the inverse of the clean process time and μ_b is the inverse of the mean time to failure.

Insert Table 1 about here

A detailed simulation model is built using the simulation modeling language χ-0.8 [10, 5]. In the detailed model, workstations have clean process times modified by failures and repairs as quantified in Table 1. In the case, the detailed simulation model was treated as the real life situation, from which the **AA**, **PD**, and **AD** events were measured for each workstation. Using the EPT-algorithms presented in Section 3, the EPT-realizations for all workstations were gathered. These EPT realizations were fitted into (shifted) gamma distributions. The obtained EPT-parameters are reported in Table 2. The following EPT-based aggregate models were built: a simulation model in which the offset is incorporated in the EPT-distribution fits (this model is referred to as EA-1), a simulation model in which the offset is included in the EPT-distribution fit at W_1, W_2, W_3 (referred to as EA-2), a simulation in which the EPT-distribution fits have no offsets (i.e., all shifts in the shifted gamma distribution are set to zero) (called EA-3) and a queueing approximation model using the approach of [22] (labeled EA-4).

Insert Tables 2 and 3 about here

Simulation results comparing the three EPT-based models to the detailed model are presented in Table 3. These results show that all models are very close to each other, since the amount of blocking and starvation of the bottleneck workstation (W_1) is low. The low level of blocking and starvation is reflected by the obtained throughput ($\delta = 3.460$), which is nearly equal to the theoretical upperbound for the bottleneck ($\delta_{\max} = t_e^{-1} = 0.2888^{-1} = 3.462$). This illustrates that in a (highly) unbalanced line, the level of blocking and starvation at the bottleneck workstation is decisive for the relevance of the offset.

This assertion is tested by changing the configuration of the line. First, the clean process times are changed to make the line more evenly balanced. Furthermore, in order to increase the variance in the line, the mean times between failure are decreased. The changes are given in Table 4, along with the new EPT-parameters. The new results of the four EPT-models, compared to the original model, are presented in Table 5. The relevance of the offset has indeed increased. However, the influence is still reasonably small, for EA-3 the approximation error has grown to 14% for flow time and 4% for throughput. The queueing model (EA-4) tries to approximate the behavior of EA-3. The error present in the queueing approximation happens to cancel out the error induced by neglecting the offset. In other cases, the two errors may add up. Summarizing, the case study illustrates that, for moderate levels of variability and moderate levels of buffering, the shape of the distribution fit (in this case represented by the offset) is not very influential on the prediction of the flow line performance. The EPT-based aggregate models stille provide accurate approximations.

Insert Tables 4 and 5 about here

The EPT-parameters of Table 2 can be used to perform a bottleneck analysis. Workstations with low effective capacity $r_{e_j} = m_j/t_{e_j}$ (with m_j the number of servers on workstation j) or high c_e^2 are potential bottlenecks. A closer look at these bottleneck stations may reveal options for improvement. Before they are implemented on the shop-floor, the effects of changes in t_e and c_e^2 can be predicted using the EPT-based aggregate model.

6 Conclusion and future work

The process time distributions play a key role in the throughput and flow time performance of a multi-server tandem queue subject to blocking. In industry practice, often only average production losses are quantified. In this paper, an effective process time (EPT) approach is proposed that enables one to measure aggregate process time distributions of workstations which incorporate outages that delay the processing without the need to quantify each of the contributing factors. The mean and variance of a measured EPT distribution quantify the effective workstation capacity and variability, respectively, which can be used for bottleneck analysis. The measured EPT distributions may also be fitted using a suitable distribution function for EPT-based aggregate model building. The EPT-based aggregate model can either be a simulation or an analytical queueing model with the advantage that it does not require the explicit modeling of the shopfloor details that are covered by the EPT distributions.

The EPT distribution of a finitely buffered, multi-server workstation can be determined using three manufacturing events: (1) the arrival of a lot in the (buffer) of the workstation, (2) the moment in time at which processing of the lot is finished and (3) the departure of the lot from the workstation. Using a simple sample path equation, these events can be translated into EPT-realizations.

For performance prediction using the EPT-based queueing model, often just the first two moments of the EPT workstation distributions suffice. Then computationally cheap queueing models, such as proposed by [22] and [21], can be used with the measured EPT mean and variance as input. However, if blocking plays a major role in the system, then the shape of the EPT distribution needs to be represented more accurately. This happens when buffer sizes are small or zero, variability is high, and only few (or just one) parallel servers are present in a workstation. We have illustrated this in examples using the offset as 'third' distribution parameter, representing a minimum positive process time. We also showed that the EPT distribution shape needs to be represented in greater detail if an accurate prediction of for instance the variance of the flow time is desired.

The EPT-based models presented in this paper assume that the EPT non-idling assumption holds. This implies that, from an EPT point of view, a server is not idle if an unprocessed lot is in the buffer. This assumption may be violated when one machine has a long down and the other machine(s) in the workstation take over. [13] proposed a method to cope with such a situation for infinitely buffered multi-server workstations. In future work, addressing violation of the non-idling assumption will be further investigated, also for the finitely buffered case.

The method developed in this paper is potentially very interesting for performance analysis of asynchronous assembly lines, as for instance encountered in automotive industry. Assembly of various components into an assembled part occurs at various stages of production. We are currently investigating the EPT of an assembly machine, and the role of transport therein.

Acknowledgments

This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Dutch Ministry of Economic Affairs. The authors would like to thank the anonymous reviewers for their comments that helped to improve the paper, and Marcel van Vuuren of the Eindhoven University of Technology and Simone Resing of CQM for their help with the case.

Bibliography

[1] I. Adan. *College sheets SPD*. Eindhoven University of Technology, 2001.

[2] I.J.B.F. Adan and J. van der Wal. Monotonicity of the throughput in single server production and assembly networks with respect to the buffer sizes. In H.G. Perros and T. Altiok, editors, *Queueing networks with blocking*, pages 345–356, 1989.

[3] T. Baines, S. Mason, P.O. Siebers, and J. Ladbrook. Humans: the missing link in manufacturing simulation? *Simulation Modelling: practice and theory*, 12(7-8):515–526, 2004.

[4] J. Banks. Introduction to simulation. In P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 7–13, 1989.

[5] D.A. van Beek, K.L. Man, M.A. Reniers, J.E. Rooda, and R.R.H. Schiffelers. Syntax and consistent equation semantics of hybrid chi. *Journal of Logic and Algebraic Programming*, 68:129–210, 2006.

[6] J.A. Buzacott and J.G. Shanthikumar. *Stochastic models of manufacturing systems*. Prentice Hall, Englewood Cliffs, New Jersey, 1 edition, 1993.

[7] L. Chen and C-L. Chen. A fast simulation approach for tandem queueing series. In O. Balci, R.P. Sadowski, and R.E. Nance, editors, *Proceedings of the 1990 Winter Simulation Conference*, pages 539–546, 1990.

[8] R. Christensen. *Data Distributions: A Statistical Handbook*. Entropy Limited, Lincoln, Massachussetts, second edition, 1989.

[9] Y. Dallery and S.B. Gershwin. Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems: Theory and Applications*, 12:3–94, 1992.

[10] A.T. Hofkamp and J.E. Rooda. *χ Reference manual*. Systems Engineering Group, TU/E, 11 2002. URL:http://se.wtb.tue.nl/.

[11] W.J. Hopp and M.L. Spearman. *Factory physics: foundations of manufacturing management*. London: Irwin McGraw-Hill, 2nd edition, 2001.

[12] J.H. Jacobs, L.F.P. Etman, E.J.J. van Campen, and J.E. Rooda. Quantifying operational time variability: the missing parameter for cycle time reduction. In *2001 IEEE/SEMI Advanced semiconductor manufacturing conference*, pages 1–10, 2001.

[13] J.H. Jacobs, L.F.P. Etman, E.J.J. van Campen, and J.E. Rooda. Characterization of operational time variability using effective process time. *IEEE Transactions on Semiconductor Manufacturing*, 16:511–520, 2003.

[14] A.M. Law and W.D. Kelton. *Simulation modeling and analysis*. McGraw-Hill Higher Education, Boston, 3 edition, 2000.

[15] J. MacGregor Smith. M/g/c/k performance models. In *Fifth International Conference on "Analysis of Manufacturing Systems-Production Management"*, pages 177–184, 2005.

[16] T. Osogami and M. Harchol-Balter. Closed form solutions for mapping general distributions to quasi-minimal ph distributions. *Performance Evaluation*, 63:524–552, 2006.

[17] A.J. de Ron and J.E. Rooda. Equipment effectiveness: Oee revisited. *IEEE Transactions on Semiconductor Manufacturing*, 18(1):190–196, 2005.

[18] SEMI. Standard for definition and measurement of equipment productivity. Technical Report SEMI E79-0200, Sematech, 2000. Originally published in 1999.

[19] T. Tolio, S.B. Gershwin, and A. Matta. Analysis of two-machine lines with multiple failure modes. *IIE Transactions*, 35:51–62, 2002.

[20] M. van Vuuren. Performance analysis of multi-server tandem queues with finite buffers. Master's thesis, Eindhoven University of Technology, Department of Mathematics, June 2003.

[21] M. van Vuuren. *Performance analysis of manufacturing systems: queueing approximations and algorithms*. PhD thesis, Department of Mathematics and Computer Science, Eindhoven University of Technology, 5 2007.

[22] M. van Vuuren and I.J.B.F. Adan. Performance analysis of tandem queues with small buffers. In *Fifth International Conference on "Analysis of Manufacturing Systems-Production Management"*, pages 127–135, 2005.

[23] M. van Vuuren, I.J.B.F. Adan, and S.A.E. Resing-Sassen. Approximation of the σ gi/g/s queue by using aggregation and matrix analytic methods. *Stochastic Models*, 21:767–784, 2005.

[24] M. van Vuuren, I.J.B.F. Adan, and S.A.E. Resing-Sassen. Performance analysis of multi-server tandem queues with finite buffers and blocking. *OR Spectrum*, 27:315–339, 2005.

Author biographies

A.A.A. Kock is a Ph.D. student in the Systems Engineering group of the department of Mechanical Engineering at the Eindhoven University of Technology. His research work is on the effective process time for performance analysis of complex manufacturing machines and systems. Ad Kock finished his M.Sc. in Mechanical Engineering at the Eindhoven University of Technology in 2003.

L.F.P. Etman is an assistant professor in the Systems Engineering group of the department of Mechanical Engineering at the Eindhoven University of Technology. His research interests include simulation-based optimization, multi-disciplinary design optimization, and the EPT methodology for performance analysis of manufacturing systems. Pascal Etman received his M.Sc. and Ph.D. in mechanical engineering at the Eindhoven University of Technology in 1992 and 1997.

J.E. Rooda is professor of Systems Engineering in the department of Mechanical Engineering at the Eindhoven University of Technology. His research interests include design and analysis of manufacturing systems, manufacturing control, and supervisory machine control. Koos Rooda received his M.Sc. in Food Technology from Wageningen University in 1971, and his Ph.D. from Twente University in 1978.

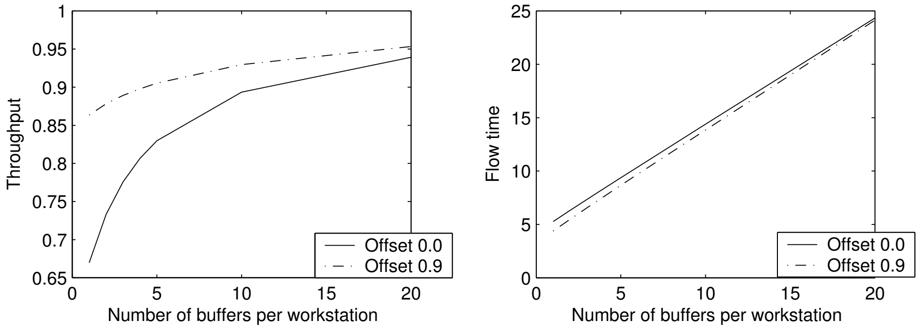

Figure 1: Influence of buffer size on throughput δ and flow time φ for a three workstation flow line

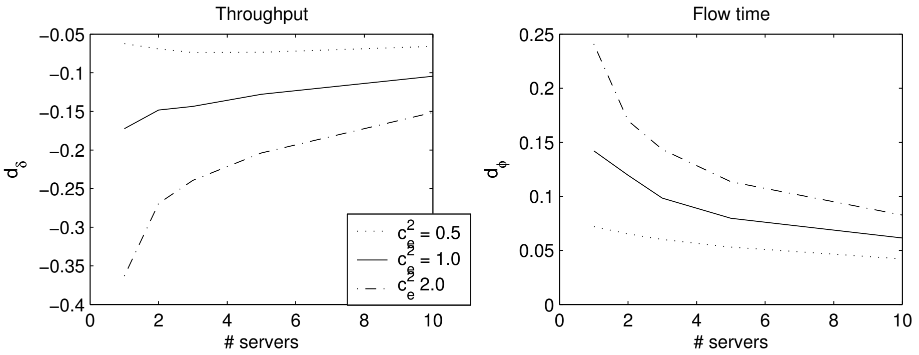

Figure 2: Relative difference between a 10-station flow line with and without inclusion of an offset of 0.9 for various levels of variability. The workstation parameters are $t_e = 1.0$ and cap = 1.

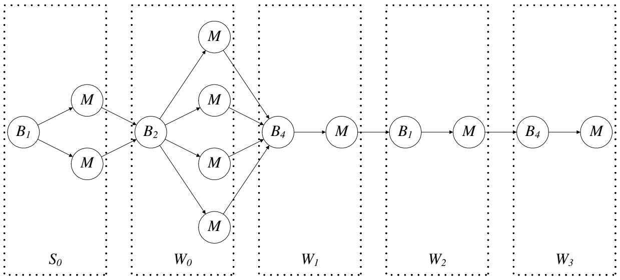

Figure 3: Layout industrial case

Station	m	b	μ_0 [lot/hr]	μ_b [f/hr]	λ_0 [r/hr]	p	λ_1 [r/hr]
S_0	2	1	5.89	0.016	2	0.2	0.8
W_0	4	2	1.54	0.003	2	0.4	0.8
W_1	1	4	3.56	0.040	2	0.8	1
W_2	1	1	32.67	0.020	12	0.5	1
W_3	1	4	16.44	0.040	12	0.5	3

Table 1: Parameters of the workstations

Workstation	t_e [hr]	c_e^2 [-]	Δ_e [hr]
S_0	0.1738	0.3572	0.1698
W_0	0.6518	0.0143	0.6494
W_1	0.2888	0.1497	0.2809
W_2	0.0310	0.7141	0.0306
W_3	0.0614	0.0988	0.0608

Table 2: EPT-parameters of the workstations

Parameter	Original	EA-1	EA-2	EA-3	EA-4
δ [lots/hr]	3.460	3.460	3.460	3.462	3.453
φ [hr]	4.138	4.138	4.139	4.136	4.04

Table 3: Estimated throughput and flow time

Station	μ_0 [lot/hr]	μ_b [f/hr]	t_e [hr]	c_e^2 [-]	Δ_e [hr]
S_0	1.78	0.50	0.9836	1.1637	0.5618
W_0	0.89	0.10	1.2639	0.2196	1.1236
W_1	3.56	0.60	0.3990	1.1670	0.2809
W_2	3.56	0.30	0.3301	0.8500	0.2809
W_3	3.56	0.60	0.3230	0.2468	0.2809

Table 4: Changed parameters of the workstations and resulting EPT-parameters

Parameter	Original	EA-1	EA-2	EA-3	EA-4
δ [lots/hr]	1.925	1.931	1.899	1.860	1.933
φ [hr]	5.586	5.467	5.503	6.396	6.09

Table 5: Estimated throughput and flow time after changes