IIE Transactions (2002) **34**, 1–9

An open queueing network for lead time analysis

NICO VANDAELE[1], LIESJE DE BOECK[1] and DOMINIEK CALLEWIER[2]

[1]*Faculty of Applied Economics, Ufsia, Prinsstraat 13, Antwerp, Belgium 2000*
E-mail: nico.vandaele@ua.ac.be or liesje.deboeck@ua.ac.be
[2]*Pharos Management Partners cvba, Kortrijksesteenweg 198, Harelbeke, Belgium 8530*
E-mail: dca@pharos.be

Received April 1999 and accepted November 2000

In this paper we develop an open queueing network for a multi-product multi-machine job-shop in a make-to-order environment. The job-shop produces a variety of products which are characterized by individual general arrival rates and individual general production rates for the machines on their deterministic routings. By incorporating the machines into a general open queueing network, we obtain the average, variance and probability distribution of the individual product lead times. The open queueing network will be illustrated by application to a real-life example existing at Recticel Bedding Hulshout. In addition to using a flexible and computational efficient approach, we methodologically reveal that the approximate queueing network is suitable to rapidly provide an answer to managerial questions.

1. Introduction

In this paper we propose an open queueing network, built on the queueing network of Vandaele (1996) for a multi-product, multi-machine job-shop in a make-to-order environment. Related approaches can be found in Bitran and Tirupati (1988) and Zijm and Buitenhek (1996). The job-shop typically manufactures a wide range of products. All products arrive individually (we assume no batch arrivals), characterized by their own general arrival rates. The products are also processed individually. As such, there are no manufacturing batches and consequently no set-up times. Each product requires several operations performed on different machines. The processing times on the machines depend on the product and follow a general distribution. Each product has a deterministic routing throughout the job-shop.

The key problem is to find satisfactory approximations for the average and the variance of both waiting times and product lead times and for a probability distribution of the product lead times taking into account the above-mentioned job-shop assumptions. Note that the utilization ratios can be obtained exactly. The ultimate goal is the application of the queueing model to real-life problems. Therefore it is necessary to introduce the stochastic nature in all processes (the incorporation of general arrivals and general services): since it enables us to deal with realistic job shops. Moreover, as speed and flexibility have become important competitive factors for almost all manufacturers, a secondary goal consists of allowing the manufacturer to settle this performance measures in a flexible and fast way.

The rest of the paper will be organised as follows. The development of the open queueing network model will be discussed in detail in Section 2. Thereafter, the queueing model will be illustrated by a real-life example in Section 3. The last section, Section 4, incorporates the conclusions of the paper.

2. The open queueing network model

In all queueing networks, there are two fundamental input parameters that underlie all queueing analyses. The first one deals with arrivals, indicating the amount of work the job-shop has to perform. The second parameter relates to processing which denotes the amount of work the job-shop can perform.

Before starting with these two basic characteristics, we introduce some notation which will be used throughout the development process of the open queueing network model. Let p be the product index ($p = 1 \ldots P$, P = number of different products in the job-shop), m the machine index ($m = 1 \ldots M$, M = number of different machines in the job-shop) and o the operation index ($o = 1 \ldots O_p$, O_p = number of operations for product p). We also introduce the binary variable δ_{opm}

where

$$\delta_{opm} = \begin{cases} 1 & \text{if operation } o \text{ of product } p \text{ is} \\ & \text{performed on machine } m, \\ 0 & \text{otherwise.} \end{cases}$$

At this point, we can start defining the input parameters related to the arrival rate and the processing rate. As already stated above, each product p arrives individually at the job-shop and is characterized by an average inter-arrival time denoted by Y_p. The stochastic nature of the inter-arrival time of product p is described by $c^2_{Y_p}$ and $s^2_{Y_p}$, indicating the variance and the squared coefficient of variation of the inter-arrival time of product p respectively. The average arrival rate of product p is denoted by λ_p. Therefore,

$$\lambda_p = \frac{1}{Y_p}. \tag{1}$$

Given the deterministic routing of each product p, the product is processed by a fixed number of operations. The average processing time of operation o on product p is indicated by X_{po}; the variance and the squared coefficient of variation of the processing time of operation o on product p are defined by $c^2_{X_{po}}$ and $s^2_{X_{po}}$ respectively. The average processing rate of operation o on product p is denoted by μ_{po}. Along the same lines as for the arrival process,

$$\mu_{po} = \frac{1}{X_{po}}. \tag{2}$$

The product lead time is approximated by summing the (effective) processing times and the waiting times at each machine where the product is processed. This follows from our initial assumptions of: no batch arrivals; no manufacturing batches; and no (external) set-up times. Since we assume that the variability from breakdowns or internal set ups are reflected in the average and variance of the effective processing times, the processing times can be directly implemented in the model. To approximate the waiting time at machine m, $\mathrm{E}(Wq_m)$, we apply the Kraemer–Lagenbach–Belz formula (Kraemer and Lagenbach-Belz, 1976), which has been extensively used in the literature:

$$\mathrm{E}(Wq_m) = \frac{\rho_m\left(s^2_{Y_m} + s^2_{X_m}\right)X_m}{2(1-\rho_m)} \exp\left\{\frac{-2(1-\rho_m)\left(1-s^2_{Y_m}\right)^2}{3\rho_m\left(s^2_{Y_m} + s^2_{X_m}\right)}\right\} \quad \text{if } s^2_{Y_m} \le 1, \tag{3}$$

or

$$\mathrm{E}(Wq_m) = \frac{\rho_m\left(s^2_{Y_m} + s^2_{X_m}\right)X_m}{2(1-\rho_m)} \quad \text{if } s^2_{Y_m} > 1.$$

In order to use this expression, we have to obtain the following additional parameters: the utilization ratio and the aggregate processing time at each machine m, given by ρ_m and X_m respectively, the squared coefficient of variation of the aggregate processing time and the aggregate inter-arrival time at each machine m, denoted by $s^2_{X_m}$ and $s^2_{Y_m}$ respectively.

The utilization ratio consists of the arrival rate divided by the processing rate. Since multiple products are processed by the same machine, requiring different operation times, we need the aggregate arrival rate and aggregate processing rate at each machine:

$$\rho_m = \frac{\lambda_m}{\mu_m}. \tag{4}$$

The aggregate arrival rate at machine m (λ_m) is obtained by:

$$\lambda_m = \sum_{p=1}^{P}\sum_{o=1}^{O_p} \lambda_p \delta_{opm}. \tag{5}$$

Note that the same machine can appear multiple times in a routing (cyclic routings). The external aggregate arrival rate (which is part of the aggregate arrival rate λ_m) at machine m equals:

$$\lambda'_m = \sum_{p=1}^{P} \lambda_p \delta_{1pm}. \tag{6}$$

The aggregate processing time (X_m) at machine m is obtained as follows:

$$X_m = \sum_{p=1}^{P}\sum_{o=1}^{O_p} \frac{\lambda_p \delta_{opm}}{\lambda_m} X_{po}, \tag{7}$$

and consequently,

$$\mu_m = \frac{1}{X_m}. \tag{8}$$

Here

$$w_{pom} = \frac{\lambda_p \delta_{opm}}{\lambda_m}, \tag{9}$$

indicates the weight of each product/operation combination in the total arrival rate at machine m. By substituting (7) into (8) and then using (8) and (5) in (4), the utilization ratio at machine m becomes:

$$\rho_m = \sum_{p=1}^{P}\sum_{o=1}^{O_p} \lambda_p \delta_{opm} X_{po}. \tag{10}$$

The squared coefficient of variation of the aggregate processing time at machine m ($s^2_{X_m}$) consists of two effects. The first effect accounts for the variability due to differences in the average processing times, which are caused by the operation/product combinations (heterogeneity variability). The second effect includes the variability inherently present in the individual processing times (natural operation variability). The squared coefficient of variation of the aggregate processing time at machine m is then approximated as:

$$s^2_{X_m} = \frac{\sum_{p=1}^{P}\sum_{o=1}^{O_p} w_{pom}X^2_{po} - \left[\sum_{p=1}^{P}\sum_{o=1}^{O_p} w_{pom}X_{po}\right]^2}{\left[\sum_{p=1}^{P}\sum_{o=1}^{O_p} w_{pom}X_{po}\right]^2} + \sum_{p=1}^{P}\sum_{o=1}^{O_p} w_{pom}s^2_{X_{po}}. \quad (11)$$

The first term of this expression includes the known expression of the squared coefficient of variation. The second term is a weighted average of the squared coefficients of variation of the individual processing times.

To derive an approximation for the squared coefficient of variation of the aggregate inter-arrival time at machine m ($s^2_{Y_m}$), we start by defining the squared coefficient of variation of the aggregate departure time at the preceding machine n, $s^2_{d_n}$. Indeed, $s^2_{d_n}$ plays an important role in resolving the approximation of $s^2_{Y_m}$ through its appearance in the squared coefficient of variation of the aggregate internal departure stream from machine n going to machine m. The squared coefficient of variation of the aggregate departure time at machine n can be obtained using the following expression (Suri *et al.*, 1993; Hopp and Spearman, 1996) which holds for single servers:

$$s^2_{d_n} \approx (1-\rho_n^2)s^2_{Y_n} + \rho_n^2 s^2_{X_n}. \quad (12)$$

Note that we do not intend to focus on parallel machines in this paper. Also note that when the utilization ratio at machine n is close to one, close to zero respectively, we expect $s^2_{d_n}$ to be the same as the squared coefficient of variation of the aggregate processing time at machine n, the squared coefficient of variation of the aggregate inter-arrival time at machine n respectively.

As already stated, the squared coefficient of variation of the aggregate departure time at machine n plays a role in the squared coefficient of variation of the aggregate arrival stream at machine m coming from machine n ($s^2_{Y_{nm}}$). The term $s^2_{Y_{nm}}$ is based on two stochastic processes. The first one is a counting process: the aggregate stream leaving machine n splits into streams to various succeeding machines. The factor t_{nm} which can be proved to be a parameter of a geometric distribution, characterizes the counting process of the number of products between two products leaving machine n for machine m. The second stochastic process is a departure process out of machine n and disturbing the aggregate arrival stream at machine n. This process is characterized by the aggregate arrival rate at machine n and the variance of the aggregate departure rate at machine n. Combining these two processes leads to the calculation of the average and variance of the time between two departures from machine n to machine m. This enables us to deduce the following expression which has been proved in the literature (Shanthikumar and Buzacott, 1993; Suri *et al.*, 1993):

$$s^2_{Y_{nm}} = t_{nm}s^2_{d_n} + (1-t_{nm}), \quad (13)$$

where t_{nm} is the proportion of products leaving machine n and going to machine m (t_{nm}), defined as:

$$t_{nm} = \frac{1}{\lambda_n}\sum_{p=1}^{P}\sum_{o=1}^{O_p-1}\lambda_p\delta_{opn}\delta_{o+1pm}, \quad (14)$$

for $n = 1 \ldots M$ and $m = 1 \ldots M$.
The aggregate arrival rate at machine m can then be formulated as:

$$\lambda_m = \sum_{n=1}^{M}\lambda_n t_{nm} + \lambda'_m, \quad (15)$$

which follows easily from (5), (6) and (14).

From this, the following approximation for the squared coefficient of variation of the aggregate inter-arrival time at machine m ($s^2_{Y_m}$) is intuitive:

$$s^2_{Y_m} \approx \sum_{n=1}^{M}\left(\frac{\lambda_n}{\lambda_m}t_{nm}\right)s^2_{Y_{nm}} + \frac{\lambda'_m}{\lambda_m}s'^2_{Y_m}, \quad (16)$$

where t_{nm} and $s^2_{Y_{nm}}$ can be calculated using Equations (14) and (13) respectively and where $s'^2_{Y_m}$ is the squared coefficient of variation of the aggregate external inter-arrival time at machine m. The first term of Equation (16) relates to a weighted average of the squared coefficient of variation of the aggregate internal arrivals at machine m, the second one takes into account the squared coefficient of variation of the aggregate external arrivals at machine m.

The only unknown parameter in the above approximation is $s'^2_{Y_m}$. This approximation is based upon Lambrecht *et al.* (1998):

$$s'^2_{Y_m} \approx \frac{1}{3} + \frac{2}{3}\sum_{p=1}^{P}\frac{\lambda_p\delta_{1pm}s^2_{Y_p}}{\lambda'_m} \quad \text{if } \sum_{p=1}^{P}\delta_{1pm} \geq 2, \quad (17)$$

or

$$s'^2_{Y_m} = s^2_{Y_p} \quad \text{if } \sum_{p=1}^{P}\delta_{1pm} = 1.$$

By using Equations (16), (13) and (12), we obtain for each machine m:

$$-\sum_{n=1}^{M}\lambda_n t^2_{nm}(1-\rho_n^2)s^2_{Y_n} + \lambda_m s^2_{Y_m} = \sum_{n=1}^{M}\lambda_n t_{nm}(t_{nm}\rho_n^2 s^2_{X_n} + 1 - t_{nm}) + \lambda'_m s'^2_{Y_m}, \quad (18)$$

where $s'^2_{Y_m}$ is given in Equation (17) and t_{nm} in Equation (14).

Note that such an equation exists for each machine in the job-shop model, defining a set of M linear simultaneous equations. By solving this set of linear equations, the unknown squared coefficient of variation of the

aggregate inter-arrival time at each machine m ($s^2_{Y_m}$) can be obtained.

At this point all parameters are known to allow the average product lead time to be obtained:

$$\mathrm{E}(LT_p) = \sum_{o=1}^{O_p}\sum_{m=1}^{M}\mathrm{E}(Wq_m)\delta_{opm} + \sum_{o=1}^{O_p} X_{po}. \qquad (19)$$

To find an approximation for the variance of the lead time of product p, the following expression is used:

$$\mathrm{Var}(LT_p) = \sum_{o=1}^{O_p}\sum_{m=1}^{M}\mathrm{Var}(Wq_m)\delta_{opm} + \sum_{o=1}^{O_p} c^2_{X_{po}}. \qquad (20)$$

The second term of the equation is known at this point. For the first term, indicating the variance of the waiting time at each machine included in the routing of product p, the approximation of Whitt (1983) is used. This approximation can only be used when the average waiting time at each machine is available and is defined as:

$$\mathrm{Var}(Wq_m) = [\mathrm{E}(Wq_m)]^2 s^2_{Wq_m}$$

$$s^2_{Wq_m} = \frac{s^2_{D_{q_m}} + 1 - \sigma_{q_m}}{\sigma_{q_m}}$$

$$\sigma_{q_m} = \rho_m + \left(s^2_{Y_m} - 1\right)\rho_m(1-\rho_m)h\left(\rho_m, s^2_{Y_m}, s^2_{X_m}\right)$$

$$h\left(\rho_m, s^2_{Y_m}, s^2_{X_m}\right) = \begin{cases} \dfrac{1+s^2_{Y_m}+\rho_m s^2_{X_m}}{1+\rho_m\left(s^2_{X_m}-1\right)+\rho^2_m\left(4s^2_{Y_m}+s^2_{X_m}\right)} & s^2_{Y_m} \le 1 \\ \dfrac{4\rho_m}{s^2_{Y_m}+\rho^2_m\left(4s^2_{Y_m}+s^2_{X_m}\right)} & s^2_{Y_m} \ge 1 \end{cases} \qquad (21)$$

$$s^2_{D_{q_m}} = 2\rho_m - 1 + \frac{4(1-\rho_m)d^3_{s_m}}{3\left(s^2_{X_m}+1\right)^2}$$

$$d^3_{s_m} = \begin{cases} \dfrac{3}{4}\left[\dfrac{1}{q^2_m} + \dfrac{1}{(1-q_m)^2}\right] & s^2_{X_m} \ge 1 \\ \left(2s^2_{X_m}+1\right)\left(s^2_{X_m}+1\right) & s^2_{X_m} < 1 \end{cases}$$

$$q_m = \frac{1}{2} + \sqrt{\frac{s^2_{X_m}-1}{s^2_{X_m}+1}}$$

where

σ_{q_m} = the probability of delay ($\mathrm{P}(Wq_m > 0)$);
$s^2_{D_{q_m}}$ = the squared coefficient of variation of the conditional waiting time i.e., the waiting time, given that the server is busy;
$d^3_{s_m}$ = $\mathrm{E}[X^3]/\bar{X}^3$.

Note that we assume that there is independency between the waiting time and the processing time at each machine as well as between the waiting times of the different machines. We are aware that this independency assumption only holds for deterministic routings. When dealing with stochastic routings, we refer to Shantikumar and Sumita (1988) who determine the variance of the lead time for a strictly symmetric job shop.

Given the average ($\mathrm{E}(LT_p)$) and variance ($\mathrm{Var}(LT_p)$) of the individual product lead time and assuming a probability distribution for the product lead time, the product lead time for a customer service level of P% can be obtained. We postulate a lognormal distribution for the product lead time as proposed by Vandaele (1996). The parameters for this distribution (μ_{LN} and σ^2_{LN}) become:

$$\mu_{\mathrm{LN}} = \ln\left(\frac{\mathrm{E}(LT_p)}{\sqrt{\left(\mathrm{Var}(LT_p)/\mathrm{E}(LT_p)^2\right)+1}}\right) \qquad (22)$$

$$\sigma^2_{\mathrm{LN}} = \ln\left(\frac{\mathrm{Var}(LT_p)}{\mathrm{E}(LT_p)^2}+1\right).$$

The lead time guaranteeing a service level of P% for product p is:

$$LT_{p,P} = \exp\{\mu_{\mathrm{LN}} + z_P\sigma_{\mathrm{LN}}\}, \qquad (23)$$

where z_P is the tabulated standard normal variable yielding a cumulative percentage P.

As all expressions to obtain the performance measures of the open queueing network model are developed, we immediately apply them to a real-life example discussed in Section 3.

3. Real-life application

One of the ways to check the usability of an open queueing network model is its application to a real-life example. The example we study is from Recticel Bedding Hulshout, a Belgian mattress manufacturer with a strong European dimension. Within their production lay-out redesign project, they aimed at testing the feasibility of a new packing unit in different scenarios. The packing unit must be capable of processing the volume of mattresses out of the new production lines. The mix of these mattresses is dependent on different wrapping procedures: not wrapped (W_{n}), once wrapped (W_{o}) and double wrapped (W_{d}) mattresses. In this packing environment, products are units of different modes of mattresses, a mode being a typical wrapping form, coming from a specific feeding transfer line. There are 16 different product types in our model: $\mathrm{P}_1 \ldots \mathrm{P}_{16}$.

The new lay-out of the packing unit, provided by the company, can typically be modeled as a multi-product multi-machine queueing network. The products are characterized by general individual inter-arrival times and general individual processing times and follow a deterministic routing throughout the layout of the packing

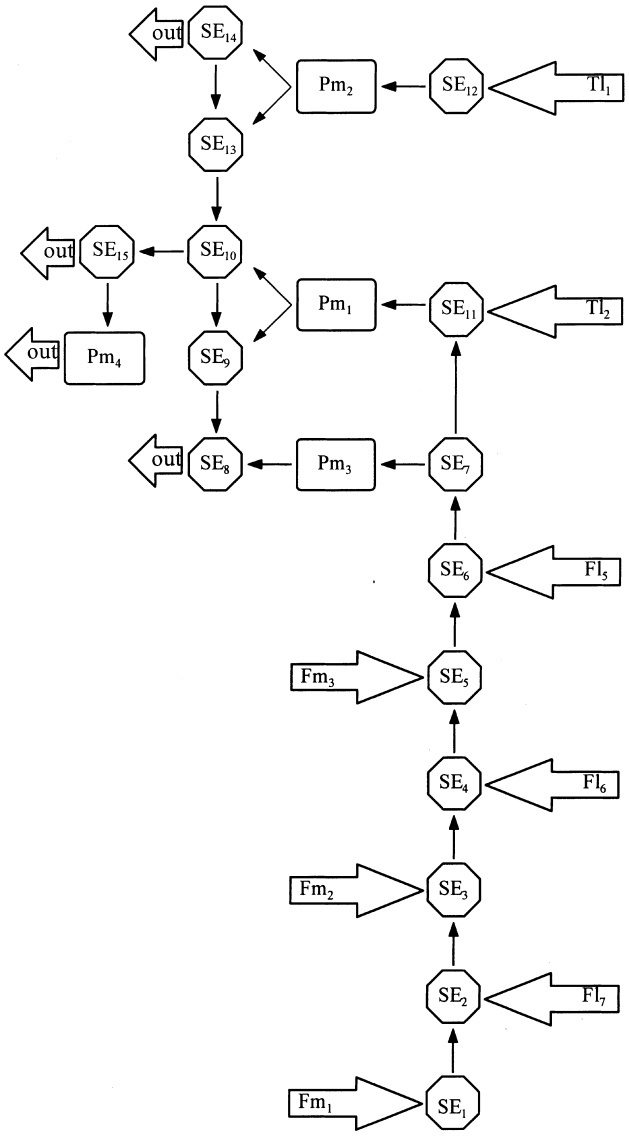

Fig. 1. The lay-out of the packing unit.

unit. Within the different scenarios, we focused on the upgrade of one packing machine, increased demand, demand mix changes and a processing time increase due to a new packing procedure. Therefore, we used the open queueing network model developed in Section 2.

To transform the packing unit into an open queueing network model, we link the building blocks of the layout with the elements of the queueing model, as can be seen in Fig. 1.

- *Feeding transfer lines:*

These are conveyors that transport mattresses from the production lines to the packing unit. As transfer lines, they indicate the arrival point of products at the packing unit. There are eight feeding transfer lines: Filling machine 1 (Fm_1), Filling machine 2 (Fm_2), Filling machine 3 (Fm_3), Filling line 5 (Fl_5), Filling line 6 (Fl_6), Filling line 7 (Fl_7), Taping line 1 (Tl_1) and Taping line 2 (Tl_2).

- *Transport change-over points:*

These serve as inter-connection between transport lines to make mattresses change position. They constitute the first type of service elements and are 15 in number: $SE_1 \ldots SE_{15}$.

- *Packing machines:*

These are the machines to accomplish the wrapping procedures on the mattresses and make-up the second type of service elements. They are four in number: Packing machine 1 (Pm_1), Packing machine 2 (Pm_2), Packing machine 3 (Pm_3) and Packing machine 4 (Pm_4). The first three packing machines handle mattresses that do not need wrapping or that only need to be wrapped once. As such, they also wrap mattresses for the first time. The last packing machine only processes double wrapped mattresses and wraps them for a second time.

- *Transport lines:*

Besides moving the mattresses, they serve as buffers in front of each service element. There are 19 buffers.

The scenarios are based on current and future arrival processes and current and future service processes respectively. The arrival process is characterized on the one hand by the product mix (relative share of each mode) coming out of each feeding transfer line and on the other hand by an aggregate inter-arrival time of the product stream out of each feeding transfer line and its variance (for more details on how to obtain the individual inter-arrival times, we refer to Vandaele *et al.* (1999)). The difference between the current and the future arrival process consists of changes in these two characteristics. The service process consists of the average processing times of each service element and its variance. The future service processes apply only to the packing machines. Their processing times are increased by 25% (except for mattresses of type W_n) compared to the current service processes due to a new packing procedure (thicker plastic). The processing times of the transport changeover points do not change, they remain the same for both current and future service processes. The five scenarios are:

- *Scenario 1* ($Scen_1$):

This scenario copes with current arrival processes and current service processes.

- *Scenario 2* ($Scen_2$):

This scenario copes with future arrival processes and current service processes.

- *Scenario 3* ($Scen_3$):

This scenario copes with current arrival processes and future service processes.

- *Scenario 4* ($Scen_4$):

This scenario copes with future arrival processes and future service processes.

- *Scenario 5* ($Scen_5$):

This scenario is the same as Scenario 1 except for Packing machine 3 not being upgraded and thus having a higher processing time.

Some additional assumptions of the queueing model are: a buffer of infinite capacity is placed in front of each service element and the model obeys the traditional FCFS-discipline. Also at the transport changeover points, even if two or more input streams are present, the product types continue to obey the joint FCFS-discipline (a rule like 'main road first' (long line) does not apply here). As mattresses of mode W_n only briefly occupy the packing machines, they are given a small processing time. All service elements have a spatial capacity of one product. As already stated, the model copes with general arrival processes and general service processes and the products have a deterministic routing.

For more details on the description of the lay-out of the new packing unit and its parameters, we refer to Vandaele *et al.* (1999).

Given all the input parameters for the open queueing network model, we obtain the performance measures developed in Section 2: the utilization ratio of each service element (ρ_m), the average waiting time at each service element ($E(Wq_m)$) and the variance of the waiting time ($Var(Wq_m)$), the average product lead time ($E(LT_p)$) and the standard deviation of the product lead time ($STDV(LT_p)$). Because the waiting queue at each service element is of major importance for the company, we like to have an idea about the acceptable queue length (finite buffer size) at each service element. This is done by dividing the 95 percentiles of the waiting time ($W_{m,95}$) by the average processing time at each service element. Assuming a log-normal distribution for the waiting time, the 95 percentile of the waiting time is computed using Equations (22) and (23). To get an idea of the significance of the acceptable queue length ($Q_{m,95}$), it is compared with the average queue at each service element ($E(Q_m)$) (the latter is obtained by multiplying the average waiting time at each service element by its aggregate arrival rate (Little's Law)).

The results are given in Tables 1, 2, 3 and 4 in which NA stands for not applicable. More specifically, this means that in Scenario's 2 and 4 Service element 13 (SE_{13}) does not appear as well as Product types 14 (P_{14}) and 16 (P_{16}). In Scenario's 1, 3 and 5 Product type 4 (P_4) is not processed anymore. For more details, we refer to Vandaele *et al.* (1999).

In Table 1 we observe that Pm_3 at Scenario 5 is nearly completely utilized and therefore has an extremely high waiting time. Consequently, this scenario must be avoided. Pm_3 must be upgraded into a faster machine because a small disruption on this packing machine would cause considerable congestion problems. As the utilization ratios in all the other scenarios never rise above 65%, they are all

Table 1. The utilization ratio and the average waiting time at each service element

	ρ_m					$E(Wq_m)$ (seconds)				
	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$
SE_1	1.6	1.6	1.6	1.6	1.6	0	0	0	0	0
SE_2	4.2	4.2	4.2	4.2	4.2	0	0	0	0	0
SE_3	5.7	5.7	5.7	5.7	5.7	0	0	0	0	0
SE_4	7.4	7.8	7.4	7.8	7.4	0	0	0	0	0
SE_5	9	9.4	9	9.4	9	0	0	0	0	0
SE_6	11.6	12	11.6	12	11.6	0	0	0	0	0
SE_7	34.7	31.4	34.7	31.5	34.7	0.2	0.5	0.2	0.5	0.2
SE_8	19.9	19.7	19.9	19.7	19.9	0.5	0.8	0.4	0.7	0.2
SE_9	8.4	11.4	8.4	11.4	8.4	0.5	0.8	0.5	0.8	0.5
SE_{10}	6.7	8.7	6.7	8.7	6.7	0.2	0.3	0.2	0.2	0.2
SE_{11}	12	20.6	12	20.6	12	1.1	2.6	1.1	2.6	1.1
SE_{12}	21.6	34.7	21.6	34.7	21.6	5.2	14.2	5.2	14.2	5.2
SE_{13}	2.3	NA	2.3	NA	2.3	0.1	NA	0.1	NA	0.1
SE_{14}	6	10.4	6	10.4	6	0	0.5	0	0.4	0
SE_{15}	6.2	14.9	6.2	14.9	6.2	0.2	1.2	0.2	1.1	0.2
Pm_1	31.3	46.9	38.5	57.8	31.3	9.8	22	17.8	45.5	9.8
Pm_2	29.6	52	36.9	65	29.6	11.5	44.4	21.2	104	11.5
Pm_3	49.5	41.7	61.9	52.1	99	1.4	2.4	3.9	5	559.1
Pm_4	1.4	26.2	1.7	32.8	1.4	0.2	7.1	0.3	11.9	0.2

Table 2. The variance and the 95 percentile of the waiting time at each service element

	$Var(Wq_m)$ $(seconds^2)$					$W_{m,95}$ $(seconds)$				
	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$
SE_1	0	0	0	0	0	0	0	0	0	0
SE_2	0	0	0	0	0	0	0	0	0	0
SE_3	0	0	0	0	0	0	0	0	0	0
SE_4	0	0	0	0	0	0	0	0	0	0
SE_5	0	0	0	0	0	0	0	0	0	0
SE_6	0	0	0	0	0	0	0	0	0	0
SE_7	0.6	3.1	0.6	3.1	0.6	0.6	1.5	0.6	1.5	0.6
SE_8	3.3	5.9	2.7	5.4	1	1.5	2.4	1.2	7.5	0.6
SE_9	4.3	6.7	4.2	6.2	4.3	1.7	2.6	1.7	2.4	1.7
SE_{10}	0.8	0.9	0.7	0.7	0.8	0.7	0.8	0.6	0.8	0.7
SE_{11}	11.2	33.3	11.2	33.3	11.2	3.2	7.3	3.2	7.3	3.2
SE_{12}	96.5	391.7	96.5	391.7	96.5	13.6	32.8	13.6	32.8	13.6
SE_{13}	0.9	NA	0.9	NA	0.9	0.4	NA	0.4	NA	0.4
SE_{14}	0	2.1	0	1.5	0	0	1.5	0	1.2	0
SE_{15}	0.7	10.8	0.7	9.3	0.7	0.6	3.5	0.6	3.2	0.6
Pm_1	235.1	795.9	661	2828.4	235.6	24.2	24.2	41.8	96.6	23.8
Pm_2	335.8	2543.9	884.3	11 965.9	335.8	28.2	93	49.1	209.7	28.2
Pm_3	11.6	29.9	49.1	86	320 436.6	4.1	6.7	10	13	1106.3
Pm_4	4.2	181.1	8.2	407.2	0	0.6	18.6	1	29.9	0.3

considered feasible. Moreover, knowing that the wrapping process is a standardized automated process, utilization ratios of 80% can still be afforded so there is some room for further demand changes (volume, mix,...). Note that those machines with the highest utilization ratios also have the highest waiting times.

In Table 2 we observe high variances for the waiting times, and consequently high 95 percentiles of the waiting times.

Table 3 indicates that the average queues are all smaller than unity except for Pm_2 in Scenario's 2 and 4 and Pm_3 in Scenario 5. Therefore the acceptable queues

Table 3. The acceptable queue length and the average queue at each service element

	$Q_{m,95}$					$E(Q_m)$				
	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$
SE_1	0	0	0	0	0	0	0	0	0	0
SE_2	0	1	0	1	0	0	0	0	0	0
SE_3	0	1	0	1	0	0	0	0	0	0
SE_4	0	1	0	1	0	0	0	0	0	0
SE_5	0	1	0	1	0	0	0	0	0	0
SE_6	0	1	0	1	0	0	0	0	0	0
SE_7	1	1	1	1	1	0.01	0.02	0.01	0.02	0.01
SE_8	1	1	1	1	1	0.02	0.03	0.02	0.03	0.01
SE_9	1	1	1	1	1	0.01	0.01	0.01	0.01	0.01
SE_{10}	1	1	1	1	1	0	0.01	0	0.01	0
SE_{11}	1	1	1	1	1	0.03	0.11	0.03	0.11	0.03
SE_{12}	1	3	1	3	1	0.11	0.7	0.11	0.07	0.11
SE_{13}	1	NA	1	NA	1	0	NA	0	NA	0
SE_{14}	0	1	0	1	0	0	0.02	0	0.02	0
SE_{15}	1	1	1	1	1	0	0.03	0	0.03	0
Pm_1	2	3	4	7	2	0.27	0.89	0.48	1.84	0.27
Pm_2	2	5	4	11	2	0.25	2.2	0.46	5.14	0.25
Pm_3	1	1	1	1	37	0.05	0.07	0.13	0.07	18.49
Pm_4	1	1	1	1	1	0	0.06	0	0.1	0

Table 4. The average and the standard deviation of the product lead time

	$E(LT_p)$ (seconds)					$STDV(LT_p)$ (seconds)				
	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$	$Scen_1$	$Scen_2$	$Scen_3$	$Scen_4$	$Scen_5$
P_1	66	80.7	74	104.1	65.8	16.5	29.1	26.1	53.6	16.4
P_2	48.1	49.6	54.2	55.8	620.4	4	6.3	7.3	9.8	566.1
P_3	45.1	46.6	51.2	52.8	617.4	4	6.3	7.3	9.8	566.1
P_4	NA	119.5	NA	158.9	NA	NA	32.1	NA	57.3	NA
P_5	60	74.7	68	98.1	59.8	16.5	29.1	26.1	53.6	16.4
P_6	42.1	43.6	48.2	49.8	614.4	4	6.3	7.3	9.8	566.1
P_7	39.1	40.6	45.2	46.8	611.4	4	6.3	7.3	9.8	566.1
P_8	54	68.7	62	92.1	53.8	16.5	29.1	26.1	53.6	16.4
P_9	36.1	37.6	42.2	43.8	608.4	4	6.3	7.3	9.8	566.1
P_{10}	33.1	34.6	39.2	40.8	605.4	4	6.3	7.3	9.8	566.1
P_{11}	37.8	52.3	45.8	75.6	37.6	16.5	29	26.1	53.6	16.4
P_{12}	35.2	50	47	77.1	35.2	16.3	29	26	53.6	16.3
P_{13}	72.3	94	91.7	133.4	72.3	16.5	32	26.2	57.3	16.5
P_{14}	57	NA	66.6	NA	56.7	21	NA	31.5	NA	21
P_{15}	44.6	87.1	58.1	150.3	44.6	20.8	54.2	31.3	111.2	21
P_{16}	105.3	NA	126.4	NA	105.3	21	NA	31.5	NA	21

are not significant except for the three cases just mentioned. Given that Scenario 5 must be avoided due to its very high utilization ratio at Pm_3 and that Pm_2 has a buffer capacity of three mattresses, we can conclude that the buffer capacities are acceptable. At this point we have to clarify how blocking is approached by the line supervisors. As can be seen from Table 3, blocking is very unlikely to occur. If eventually the packing line is prone to blocking, then an operator removes the blocking mattress from the line and puts it back when the buffer allows space at a later moment. Note that this is possible since the packing line is not fully automated and the mattresses are not continuously traced. Of course the new design of the packing line should ensure that this occasional procedure is limited to an acceptable minimum. Avoiding blocking completely would be too expensive in terms of automated roller banes.

In Table 4 we observe that the only significant product lead times relate to Scenario 5. The higher product lead times in the other scenarios apply to those products that cover the longest path throughout the packing unit. Note also that the squared coefficients of variation of the product lead times are relatively low: they never rise above 0.85 for Scenario 5 and they do not exceed 0.55 for the other scenarios.

To test the accuracy of our open queueing network results, we also performed a simulation study. The simulation model is built identically to the lay-out as given in Fig. 1. The same scenarios are simulated, both with infinite and finite buffers. For more information on the simulation model and the simualtion results, we refer the interested reader to Vandaele *et al.* (1999). (For the 95 percentiles of the average waiting times and the corresponding waiting queues and for the variances of the waiting times and the lead times, we have no simulation results.) If we compare the simulation results (which can be found in Vandaele *et al.* (1999)) and the queueing results, we can draw two conclusions. The expressions for the utilization ratios and the average product lead times perform very well. The utilization ratios of the queueing model are almost equal to the results of the simulation model. Also the product lead times of the queueing model and the simulation model are close. Secondly, although the absolute values for the average waiting times, and consequently the average waiting queueus of both models do not quite match, we can clearly observe sufficient similarity between both results. More specifically, if we compare the results of waiting times and waiting queues for the different machines, both models lead to the same managerial conclusions.

4. Conclusions

An open queueing network model for job shops is developed in which products have general individual arrival processes, general individual service processes and deterministic routings. Once the input parameters (the average inter-arrival time of each product and its variance and the average operation processing time of each product on each machine of its deterministic routing and its variance) are known, the performance measures can be easily obtained. In terms of different what-if scenarios, the open queueing network model is a very efficient approach, computationally fast and leads quickly to effective answers to the managerial questions. Besides this, the methodology outlined for developing the queueing model

is highly suited for real-life applications. In particular, these applications fit in feasibility analysis of job shops and production cells. Moreover, the queueing model can provide evidence of potential improvements in job shops as congestion points can be detected, having high utilization ratios and huge waiting times.

As the queueing model explicitly modeled complexity and stochasticity, we were able to make the design and operation of an automated packing unit a smooth, reliable and efficient unit in the entire production facility. We modeled it as a part of a larger manufacturing system thereby focusing on short-term dynamics, in terms of the stochastic network, and long-term dynamics, in terms of buffering the system with safety processing and buffer capacity (exposed in the different scenarios studied). The links with the feeding transfer lines (production lines) were crucial and reflected the short and longer term dynamics expressed by product volumes and product mix. Methodologically, we revealed that the decentralized decisions (individual product performance measures) can be taken by first conducting an aggregate yet centralized analysis (queueing model) evaluating the overall efficiencies of the packing unit.

Nevertheless, it must be kept in mind that when there is need for very detailed and precise information, a complementary simulation study can be performed.

Acknowledgements

This research was supported by the National Science Foundation (FWO), project G.0063.98 and the BOF fund of the University of Antwerp.

References

Bitran, G.R. and Tirupati, D. (1988) Multiproduct queueing networks with deterministic routing: decomposition approach and the notion of interference. *Management Science*, **34**(1), 75–100.

Hopp, W.J. and Spearman, M.L. (1996) *Factory Physics*, McGraw-Hill, Chicago, Illinois.

Kraemer, W. and Lagenbach-Belz, M. (1976) Approximate formulae for the delay in the queueing system GI/GI/1, in *Congressbook of the Eight International Teletraffic Congress*, pp. 235-1/8.

Lambrecht, M.R., Ivens, P.L. and Vandaele, N.J. (1998) ACLIPS: a capacity and lead time integrated procedure for scheduling. *Management Science*, **44**(11), 1548–1561.

Shanthikumar, J.G. and Buzacott, J.A. (1993) *Stochastic Models of Manufacturing Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Shantikumar, J.G. and Sumita, U. (1988) Approximations for the time spent in a dynamic job-shop with applications to due-date assignment. *International Journal of Production Research*, **26**, 1329–1352.

Suri, R., Sanders, J.L. and Kamath, M. (1993) Performance evaluation of production networks, in *Logistics of Production and Inventory, Handbooks in Operations Research and Management Science*, Vol. 4, S.C. Graves, A.H.G. Rinnooy Kan and P.H. Zipkin (eds.), North Holland, Amsterdam, The Netherlands, pp. 199–286.

Vandaele, N.J. (1996) The Impact of Lot Sizing on Queueing Delays: Multi Product, Multi Machine Models. Ph.D. thesis, Department of Applied Economics, Katholieke Universiteit Leuven, Belgium.

Vandaele, N., De Boeck, L. and Callewier, D. (1999) A queueing based analysis of a packing line, in *Proceedings of the Ninth International Conference on Flexible Automation and Intelligent Manufacturing*, Begell House, New York, NY.

Whitt, W. (1983) The queueing network analyzer. *The Bell System Technical Journal*, **62**(9), 2779–2815.

Zijm, W.H.M. and Buitenhek, R. (1996) Capacity planning and lead time management. *International Journal of Production Economics*, **46**, 165–179.

Biographies

Nico Vandaele is Professor of Operations Management and Operations Research at UFSIA, the University of Antwerp and at the Catholic University of Leuven, Belgium. He received his Ph.D. in Operations Management from the Catholic University of Leuven and holds a Commercial Engineering Degree. He currently teaches courses in Operations Management and Operations Research and has published in *Management Science*, *Interfaces*, *European Journal of Operations Research*, *Transportation Research*, *IIE Transactions*, *Production and Operations Management*. He is a member of INFORMS, EURO, APICS, POMS and EUROMA. He is also a member of the Board of PICS Belgium. His research interests are queueing networks and their applications to production planning, scheduling and traffic management.

Liesje De Boeck is a doctoral research assistant at UFSIA, the University of Antwerp, Belgium. In 1997 she graduated as a Commercial Engineer at the Catholic University of Leuven, Belgium. Currently, she is working on a 4-year project, supported by the Flemish Science Foundation (FWO) at UFSIA. The project extends queueing networks to practical applications for production planning. She is a member of INFORMS. Her key research topics are queueing approximations, lead time analysis and performance evaluation of production systems.

Dominiek Callewier is founder and managing director of Pharos Management Partners, a company that specializes in interim and project management in industrial companies. In this function, he is responsible for several assignments – Sadef, Vyncke, Sabena, All Crump, Stas – in the field of operations management and improvement management. Previously, he spent 4 years as a manager at Ernst and Young Consulting on assignments such as Recticel and Sabena and 3 years as an assistant in the Operations Management Group at the Catholic University of Leuven, Belgium. He holds a Masters degree in Management (K.U.Leuven), a degree in engineering (Groep T) and is Certified in Production and Inventory Management (CPIM) and Certified in Integrated Resources Management (CIRM). He has lectured at universities such as UFSIA, the University of Antwerp, the Catholic University of Leuven and Leti-Lovanium (Saint-Petersburg) and in programmes for PICS Belgium, CIM_CIL and SBM. He has published numerous articles on operations management and a book on Total Quality Management.

Contributed by the Decentralized Control of Manufacturing Systems Department