Two-stage queueing network models for quality control and testing

Shaul K. Bar-Lev[*], Onno Boxma[†], Wolfgang Stadje[‡],
Frank A. Van der Duyn Schouten[§], Christoph Wiesmeyr[¶]

Abstract

We study sojourn times in a two-node open queueing network with a processor sharing node and a delay node, with Poisson arrivals at the PS node. Motivated by quality control and blood testing applications, we consider a feedback mechanism in which customers may either leave the system after service at the PS node or move to the delay node; from the delay node, they always return to the PS node for new quality controls or blood tests. We propose various approximations for the distribution of the total sojourn time in the network; each of these approximations yields the exact mean sojourn time, and very accurate results for the variance. The best of the three approximations is used to tackle an optimization problem that is mainly inspired by a blood testing application.

1 Introduction

In this paper we study sojourn times in a two-node open queueing network with a processor sharing (PS) node and a delay (D) node. External customers arrive at the PS node and then either leave the system or enter the

[*] Department of Statistics, University of Haifa, Haifa 31905, Israel (bar-lev@stat.haifa.ac.il)

[†] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (boxma@win.tue.nl)

[‡] Department of Mathematics and Computer Science, University of Osnabrück, 49069 Osnabrück, Germany (wolfgang@mathematik.uos.de)

[§] Center for Economic Research, Tilburg University, 5000 LE Tilburg, The Netherlands (f.a.vdrduynschouten@uvt.nl)

[¶] Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (christoph.wiesmeyr@liwest.at)

D node after which they return to the PS node and so on. The special feature of the processor sharing discipline is that all customers are being served simultaneously. When there are k customers in the PS node, they all receive a fraction $1/k$ of the server capacity. In the D node, there is no waiting; a customer who enters the D node leaves it again after some random time, that does not depend on the presence of other customers.

This work has been motivated by some practical aspects which prevail in various industrial (quality control) and medical (blood testing) applications. We briefly discuss an example of each of those.

Consider items which are undergoing a series of quality conformance tests. New items arrive in the system for the first test (PS node), say, a visual inspection for external damages. If an item fails this examination, it is discarded (or forwarded to further activities) and leaves the system. If it passes the first test, it is being forwarded to a processing stage (D node), preparing it for the next planned test in the series. It is then subjected to the second test. If it fails such a test, it leaves the system. Otherwise it is forwarded to the D node preparing it for the third test and so on. In many cases the second and third tests might be parts or modules for examination and testing under operating conditions, respectively. The test node may be represented by a PS node, as all inspected items are immediately taken into consideration, simultaneously receiving the attention of an operator. As we assume that a new test (if required) takes place after some random preparation time, the preparation phase is represented by a D node.

As an example of a medical application, consider the testing for the presence of viruses in blood testing. For example, consider HIV. Until recently, the routine testing was based on the ELISA (Enzyme Linked Immuno-Sorbent Assay) that detects virus-specific antibodies in the blood. This test has high sensitivity and specificity but has a lower analytic detection limit which affects the identification of positive samples very soon after HIV seroconversion, as it takes time to develop a high concentration of antibodies. A new test, PCR (Polymerase Chain Reaction) can detect viral genetic material in the blood and has a much higher sensitivity and specificity. PCR testing is especially advantageous in the window period soon after seroconversion when the virus starts multiplying but antibodies are not yet at high levels. For these samples, the ELISA test will be negative while the PCR test is likely to be positive. However, PCR is very expensive relative to ELISA. Therefore, blood banks in the USA and some countries in Europe established a new protocol whereby all samples are ELISA tested and those which tested negative for ELISA are re-tested either individually or in mini-pools (of 6-12 blood samples) with PCR. A PCR testing requires a special preparation of the sample, a fact which causes a delay (the preparation phase is modeled by the D node). Accordingly, we face here a system with two testing stages: Blood samples arrive at the PS node for an ELISA testing, and if they are found HIV positive they leave the system. Otherwise, they are forwarded to the D node for a preparation to the second stage (a second visit to the PS node) which involves PCR testing. Then, they leave the system either as HIV positive or negative. It should be noted that a similar routine is also applied to detect HCV in blood testing. The rationale behind having processor sharing in these blood testing applications is the following. Because of the limited life times of some substances in the blood, arriving blood samples should immediately be processed. This processing can be done in parallel, for a large collection of samples.

Costs are involved with testing problems like those mentioned above, and minimizing the costs leads to a non-trivial optimization problem. It is very important that the total time for testing a blood sample (the total sojourn time of a customer) is not too long. However, costs are involved in speeding up the procedure, for example by using better equipment or using more (or better trained) personnel, and thus we face a trade-off. To analyze this trade-off, we need to know the probability that the total sojourn time of a customer in the model is less than some threshold value t_0. In the present paper, we therefore study the *sojourn time distribution* of customers in a two-node open queueing network, consisting of a PS node and a D node.

It should be noted that the two-node model under consideration contains a more general feedback mechanism than required for the above-mentioned examples. Furthermore, the method of analysis may be of independent interest. One could, e.g., replace the PS node by a node with another service discipline and still follow the same approach globally. See also the related two-node studies [3, 5, 10], that consider different service disciplines and a less general feedback mechanism.

The paper is organised as follows. In Section 2 we present a detailed description of the two-stage queueing network consisting of a PS node and a D node. In the rest of the paper we focus our attention on the sojourn time of a customer, i.e., the time a customer spends in the system. In view of the fact that it is extremely difficult to obtain exact results for sojourn time distributions if customers can overtake one another (as is the case in processor sharing nodes, while the feedback mechanism in the D node also leads to overtaking), we have to take recourse to approximations. Section 3 discusses three (related) methods to approximate the Laplace-Stieltjes transform (LST) of the joint distribution of the total sojourn time of a customer at the PS node and at the D node. In particular, we obtain expressions for the mean, variance and distribution of the total sojourn time of a customer in the system. We show that each of the three methods yields the exact mean sojourn times at the PS node and at the D node, and hence also the exact mean total sojourn time. In Section 4 we present the results of extensive numerical experiments which test the accuracy of the various approximations. Section 5 is devoted to an optimization problem that is relevant for both the blood testing problem and the quality conformance testing problem. We aim to maximize a certain reward function that involves the probabilities that the sojourn time of an arbitrary customer is below a certain threshold and that it is above another threshold.

2 Model description

In this section we consider a two-node open queueing network with a processor sharing (PS) node and a delay (D) node, also called infinite server node. The model is depicted in Figure 1. External customers arrive at the PS node according to a Poisson process with rate λ. A departing customer subsequently enters the D node with probability p_1, and leaves the system with probability $1 - p_1$. Upon departure from the D node, a customer always returns to the PS node. After the j-th visit to the PS node, a customer enters the D node with probability p_j and leaves the system with probability $1-p_j$; we assume that $p_K = 0$, implying that no customer visits the PS node more than K times. All service times at all visits to both nodes are independent random variables, with distribution $B_j(\cdot)$ and $D_j(\cdot)$ at the j-th visit to the PS and D node, respectively, and with service time LST (Laplace-Stieltjes transform) $\beta_j(\cdot)$ and $\delta_j(\cdot)$ and mean β_j and δ_j, respectively. In the sequel we shall call a customer a type-j customer when he brings his j-th visit to a queue. Introduce $q_j := \mathbb{P}(\textit{a customer visits the PS node exactly } j \textit{ times})$ $= \prod_{i=1}^{j-1} p_i(1 - p_j)$. The total load of type-j customers at the PS node is $\rho_j := \lambda\beta_j \prod_{i=1}^{j-1} p_i = \lambda\beta_j \sum_{i=j}^{K} q_i$, and the total load at the PS node is $\rho := \sum_{j=1}^{K} \rho_j$. The total load of type-j customers at the D node is $\varphi_j := \lambda\delta_j \prod_{i=1}^{j} p_i = \lambda\delta_j \sum_{i=j+1}^{K} q_i$.

The above model falls into the class of product-form queueing networks as described in [4]. In the following we restrict ourselves to the steady-state behavior of the two-node system. The steady-state joint distribution of the numbers of customers $X_1, \dots, X_K$ of type $1, \dots, K$ at the PS node and the numbers of customers $Y_1, \dots, Y_{K-1}$ of type $1, \dots, K-1$ at the D node is known to exist iff $\rho < 1$, and then to have the following product form:

$$\begin{aligned} &\mathbb{P}(X_1 = i_1, \dots, X_K = i_K, Y_1 = j_1, \dots, Y_{K-1} = j_{K-1}) \\ = \;& (1-\rho)\frac{(i_1 + \cdots + i_K)!}{i_1! \dots i_K!}\rho_1^{i_1} \dots \rho_K^{i_K} \mathrm{e}^{-\sum_{j=1}^{K-1} \varphi_j} \frac{\varphi_1^{j_1}}{j_1!} \dots \frac{\varphi_{K-1}^{j_{K-1}}}{j_{K-1}!}. \qquad (2.1) \end{aligned}$$

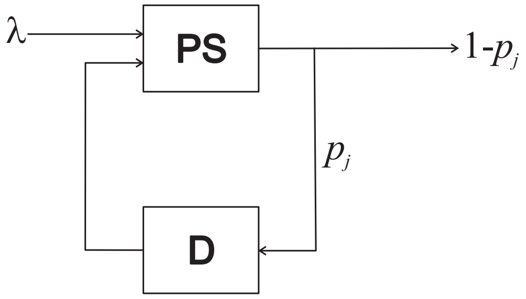

Figure 1: The two-node open queueing network

The mean numbers of customers at both nodes are immediately derived from (2.1), and Little's formula subsequently gives the mean total sojourn times at both nodes:

$$\mathbb{E}S_{PS} = \sum_{j=1}^{K} q_j \sum_{i=1}^{j} \frac{\beta_i}{1-\rho} = \frac{1}{\lambda}\frac{\rho}{1-\rho}, \tag{2.2}$$

$$\mathbb{E}S_{D} = \sum_{l=1}^{K-1} \delta_l \sum_{i=l+1}^{K} q_i = \frac{1}{\lambda}\sum_{l=1}^{K-1} \varphi_l. \tag{2.3}$$

For our purposes we need the full *distribution* of the total sojourn time in the system. Unfortunately, it is virtually impossible to obtain exact results for sojourn time distributions in networks of queues with some form of overtaking of customers (see [2] for an overview of results on sojourn times in queueing networks). Two features of our model, processor sharing and feedback, both entail overtaking. The complexity of the problem of obtaining sojourn time results in queues with non-instantaneous feedback was discussed in [8]. In view of the overtaking problem, we are looking for approximation methods for the sojourn time distribution in the network. Our study is related to [3, 5, 10]. We consider the same model, except that (i) we allow *general* service time distributions in both nodes, (ii) we consider a delay node instead of a FCFS node, and (iii) the feedback probabilities may change in each loop. In the model of [5, 10] there is no product form solution, and even mean sojourn times are not known; those papers concentrate on approximating *mean* sojourn times.

3 Description of the approximation methods

We want to approximate the LST of the joint distribution of the total sojourn time S_{PS} at the PS node and S_D at the D node. Denote by $S_{PS}^{(j)}$ ($S_D^{(j)}$) the total sojourn time at the first j visits to the PS (D) node ($S_D^{(0)} = 0$). We can write, for Re $\omega_1 \geq 0$, Re $\omega_2 \geq 0$:

$$\mathbb{E}[\mathrm{e}^{-\omega_1 S_{PS} - \omega_2 S_D}] = \sum_{j=1}^{K} q_j \mathbb{E}[\mathrm{e}^{-\omega_1 S_{PS}^{(j)} - \omega_2 S_D^{(j-1)}}]. \tag{3.1}$$

Of course, taking $\omega_1 = \omega_2$ yields an expression for the total sojourn time, $S = S_{PS} + S_D$, of a customer in the system. We shall present three different, but related, approximation methods.

Method I: Independence Assumption (IA)
This approximation method is based on the following assumptions.

- *Assumption 1.* $S_{PS}^{(j)}$ and $S_D^{(j-1)}$ are independent, $j = 1, \ldots, K$.
- *Assumption* 2^a. $S_{PS}^{(j)}$ is distributed as the sum of j independent sojourn times $S_1, \ldots, S_j$. The m-th term S_m is distributed as the sojourn time of a special customer with service time distribution $B_m(\cdot)$ in an $M/G/1$ PS queue with arrival rate $\lambda \sum_{i=1}^{K} i q_i$ and service time distribution

$$B(x) := \frac{\sum_{i=1}^{K} (\sum_{j=i}^{K} q_j) B_i(x)}{\sum_{i=1}^{K} i q_i}. \tag{3.2}$$

Assumption 2^a corresponds to having an infinite feedback delay: In effect there are now K independent Poisson arrival streams, the ith one having arrival rate $\lambda \sum_{j=i}^{K} q_j$ and service time distribution $B_i(\cdot)$. This approximation hence is based on several independence assumptions, just like the IA, also called *Independent Flow Time Approximation*, which is a classic approximation method that was proposed for a large class of queueing networks in [9, 13]. Hence we have also called this method IA. From (3.1) and Assumptions 1 and 2^a we obtain:

$$\mathbb{E}[\mathrm{e}^{-\omega_1 S_{PS} - \omega_2 S_D}] \approx \sum_{j=1}^{K} q_j \prod_{i=1}^{j} (\mathbb{E}[\mathrm{e}^{-\omega_1 S_i}]) \prod_{i=1}^{j-1} \delta_i(\omega_2). \tag{3.3}$$

In particular, we find from (3.3) for the mean sojourn times

$$\mathbb{E}S_{PS} = \sum_{j=1}^{K} q_j \mathbb{E}S_{PS}^{(j)} \approx \sum_{j=1}^{K} q_j \sum_{i=1}^{j} \frac{\beta_i}{1-\rho} = \frac{1}{\lambda} \frac{\rho}{1-\rho}, \tag{3.4}$$

$$\mathbb{E}S_D = \sum_{l=1}^{K-1} \delta_l \prod_{i=1}^{l} p_i = \sum_{l=1}^{K-1} \delta_l \sum_{i=l+1}^{K} q_i = \frac{1}{\lambda} \sum_{l=1}^{K-1} \varphi_l. \tag{3.5}$$

These actually agree with the exact mean sojourn times at the PS node and at the D node, as given in (2.2) and (2.3). Hence our approximation also gives the exact mean total sojourn time.

Successive sojourn times at the PS node are nearly independent if the times between successive visits are relatively large; in the latter case, Assumption 2^a is justified. Method II takes the opposite extreme view; there it will be assumed that the times between such successive visits to the PS node are zero. We call this the short-circuit assumption.

Method II: Short-Circuit Assumption (SC)
This approximation method is based on the following assumptions.

- *Assumption 1.* $S_{PS}^{(j)}$ and $S_D^{(j-1)}$ are independent, $j = 1, \dots, K$.
- *Assumption* 2^b. $S_{PS}^{(j)}$ has the same distribution as $\sigma_{PS}^{(j)}$, the total sojourn time after j visits in the PS node short-circuited (i.e., with the D node removed).

From (3.1) and Assumptions 1 and 2^b we obtain:

$$\mathbb{E}[\mathrm{e}^{-\omega_1 S_{PS} - \omega_2 S_D}] \approx \sum_{j=1}^{K} q_j \mathbb{E}[\mathrm{e}^{-\omega_1 \sigma_{PS}^{(j)}}] \prod_{i=1}^{j-1} \delta_i(\omega_2). \tag{3.6}$$

The LST $\mathbb{E}[\mathrm{e}^{-\omega_1 \sigma_{PS}^{(j)}}]$ for an $M/G/1$ PS queue with instantaneous feedback is the same as the LST of the sojourn time of a tagged customer having as service time the sum of the j service times $B_1, \dots, B_j$ in an $M/G/1$ PS queue without feedback, having $\hat{B}(\cdot) = \sum_{j=1}^{K} q_j (B_1(\cdot) * \dots * B_j(\cdot))$ as service time distribution for an arbitrary customer, $*$ denoting a convolution (note that the tagged customer has exactly j passes through the feedback queue, but that an arbitrary customer has a random number of passes). Theorem 2.2 of Ott [11] gives the LST of the sojourn time distribution of a customer with service requirement x in an $M/G/1$ PS queue; integration w.r.t. the density of $B_1 + \dots + B_j$ gives $\mathbb{E}[\mathrm{e}^{-\omega_1 \sigma_{PS}^{(j)}}]$. The LST expression in [11] is quite complicated. If one is satisfied by just obtaining an approximation for the *variance* of $S_{PS}^{(j)}$ (and hence of S_{PS}), then a relatively easy expression for the variance of the former random variable can be taken from [11], see also (3.13) below.

Furthermore, we find from (3.6) for the mean sojourn time

$$\begin{aligned}\mathbb{E}S_{PS} &\approx \sum_{j=1}^{K} q_j \mathbb{E}\sigma_{PS}^{(j)} = \sum_{j=1}^{K} q_j j \frac{1}{1-\rho} \frac{\sum_{i=1}^{K} \beta_i \sum_{j=i}^{K} q_j}{\sum_{j=1}^{K} q_j j} \\ &= \frac{\sum_{i=1}^{K} \beta_i \sum_{j=i}^{K} q_j}{1-\rho} = \frac{1}{\lambda} \frac{\rho}{1-\rho}. \qquad (3.7)\end{aligned}$$

The latter expression agrees with the mean sojourn time in the PS node obtained via the first approximation, and with the exact expression. This makes sense, as the product-form result for numbers of customers reveals an independence of the number of customers in the PS node from the parameters at the delay node. Hence the mean number of customers, and via Little's formula the mean sojourn time, at the PS node is not influenced by the parameters at the delay node. So we might as well take those parameters equal to zero or to infinity – corresponding to the two approximations considered above.

Method III: Weighted Average Approximation (WA)

Method I should work well if $\mathbb{E}S_D >> \mathbb{E}S_{PS}$, and Method II should work well if $\mathbb{E}S_D << \mathbb{E}S_{PS}$. If the mean sojourn times at both queues are roughly of the same size, then the approximations in (3.3) and (3.6) can be improved in the following way. Replace the LST's in the righthand side of (3.1) by weighted sums of LST's that correspond to the two extremes of short-circuiting (i.e., immediate feedback to the same queue) and independence of successive sojourn times of a customer at the same queue (i.e., feedback after an infinite amount of time):

$$\mathbb{E}[\mathrm{e}^{-\omega_1 S_{PS} - \omega_2 S_D}] \approx \sum_{j=1}^{K} q_j \left(w\mathbb{E}[\mathrm{e}^{-\omega_1 \sigma_{PS}^{(j)}}] + (1-w) \prod_{i=1}^{j} (\mathbb{E}[\mathrm{e}^{-\omega_1 S_i}]) \right) \prod_{i=1}^{j-1} \delta_i(\omega_2). \qquad (3.8)$$

We propose to choose the weight w as $w = \mathbb{E}S_{PS}/[\mathbb{E}S_{PS} + \mathbb{E}S_D]$. For the *mean* sojourn times we of course again find the exact result.

Sojourn time variances for Methods I-III

The starting point is:

$$\begin{aligned}\mathrm{Var}(S) &= \mathrm{Var}(S_{PS}) + \mathrm{Var}(S_D) + 2\mathrm{cov}(S_{PS}, S_D) \\ &= \mathrm{Var}(S_{PS}) + \mathrm{Var}(S_D) + 2\mathbb{E}S_{PS}S_D - 2\mathbb{E}S_{PS}\mathbb{E}S_D. \qquad (3.9)\end{aligned}$$

$\mathbb{E}S_{PS}$ and $\mathbb{E}S_D$ are known exactly, see above, and so is $\mathrm{Var}(S_D)$: Obviously, $\mathrm{Var}(S_D) = \sum_{j=1}^{K} q_j \mathbb{E}((S_D^{(j)})^2) - \left(\sum_{j=1}^{K} q_j \mathbb{E}(S_D^{(j)})\right)^2$.

In Method I we approximate, via (3.3),

$$\mathbb{E}(S_{PS}S_D) \approx \sum_{j=1}^{K} q_j \mathbb{E}S_{PS}^{(j)}[\delta_1 + \cdots + \delta_{j-1}] = \sum_{j=1}^{K} q_j \sum_{i=1}^{j} \frac{\beta_i}{1-\rho}(\delta_1 + \cdots + \delta_{j-1}). \tag{3.10}$$

Alternatively, in Method II, we find the same result, via (3.6),

$$\mathbb{E}(S_{PS}S_D) \approx \sum_{j=1}^{K} q_j \mathbb{E}\sigma_{PS}^{(j)}[\delta_1 + \cdots + \delta_{j-1}], \tag{3.11}$$

as $\mathbb{E}\sigma_{PS}^{(j)} = \sum_{i=1}^{j} \beta_i/(1-\rho)$.

The main problem is to approximate $\mathrm{Var}(S_{PS})$. We discuss how this is done in each of the three methods.
Method I (Independence Assumption): $\mathrm{Var}(S_{PS}) = \mathbb{E}(S_{PS}^2) - (\mathbb{E}S_{PS})^2$, and we approximate $\mathbb{E}(S_{PS}^2) \approx \sum_{j=1}^{K} q_j \sum_{i=1}^{j} \mathbb{E}(S_i^2)$. We shall determine $\mathbb{E}(S_i^2)$, S_i being the sojourn time of a special customer with service time distribution $B_i(\cdot)$ in an $M/G/1$-PS queue with general service time distribution as given in (3.2). Let ρ denote the total traffic load, and $F(\cdot)$ the waiting time distribution in the corresponding $M/G/1$ queue with FCFS instead of PS service discipline. Formula (2.33) of Ott [11] gives the variance of the sojourn time in this $M/G/1$-PS queue of a customer with service requirement x. Adding $(\mathbb{E}(S_i|x))^2 = x^2/(1-\rho)^2$ gives

$$\mathbb{E}(S_i^2|x) = \frac{2x^2}{(1-\rho)^2} - \frac{2}{(1-\rho)^2}\int_{y=0}^{x}(x-y)F(y)\mathrm{d}y. \tag{3.12}$$

Integrating over the distribution $B_i(\cdot)$ then yields $\mathbb{E}(S_i^2)$.

Method II (Short-Circuit Assumption): Take the variance of the total sojourn time in the $M/G/1$-PS queue with instantaneous Bernoulli feedback, which is also the total sojourn time in the $M/G/1$-PS queue without feedback but with service time distribution $\hat{B}(x) = \sum_{j=1}^{K} q_j (B_1 * \cdots * B_j(x))$, $*$ again denoting convolution. With $\hat{\beta}$ and $\hat{\beta}_2$ the first two moments of $\hat{B}(\cdot)$, $\hat{\rho}$ the traffic load, $\hat{S}$ the sojourn time and $\hat{F}(\cdot)$ the waiting time distribution in the corresponding $M/G/1$ queue with FCFS instead of PS service discipline, Formula (2.36) of Ott [11] gives:

$$\mathrm{Var}(\hat{S}) = \frac{2\hat{\beta}_2 - \hat{\beta}^2}{(1-\hat{\rho})^2} - \frac{2}{(1-\hat{\rho})^2}\int_{x=0}^{\infty}\int_{y=0}^{x}(x-y)\hat{F}(y)\mathrm{d}y\mathrm{d}\hat{B}(x). \tag{3.13}$$

Method III: Take a weighted average, starting from (3.8).

3.1 Discussion

In this subsection we discuss the assumptions. Assumption 1 was already introduced and motivated in [10], for the two-node model with a PS node and an FCFS node. Simulation experiments in [10] indicate that Assumption 1 is in most cases quite close to reality in that PS/FCFS model. In the present model, we have a product-form network, and hence the queue lengths at jump epochs of customers are independent. Moreover, the times at the D node are independent from those at the PS node. All this forms additional motivation for the Independence Assumption 1. One may expect that Assumption 2^b (and hence Method II) works well in most cases. The reason for this is the following. Short-circuit assumption 2^b should be quite accurate if $\mathbb{E}S_D << \mathbb{E}S_{PS}$. It will not be accurate if $\mathbb{E}S_{PS} << \mathbb{E}S_D$, but in the latter case the contribution of the PS node to the total sojourn time will be much smaller than the contribution of the D node – and the latter contribution is exact.

3.2 The sojourn time distribution

In Section 5 we need an (approximate) expression for the *distribution* of the total sojourn time of a customer. Earlier in this section we have obtained an exact expression for its mean, and approximations for its variance and its LST (for the latter, take $\omega_1 = \omega_2 = \omega$ in the expression for the joint transform of S_{PS} and S_D). One could now invert this LST numerically (using, e.g., the well-known procedure expounded by Abate and Whitt in several papers; see, e.g., [1]). However, we would like to have relatively simple, explicit expressions for optimization purposes. Hence we follow another road: We approximate $\text{Var}(S)$ as above (using Method III in Section 5 on optimization), and subsequently approximate $\mathbb{P}(S > x)$ by using a two-moment approximation for this distribution. The following method is suggested for this purpose, cf. [15]. Consider the squared coefficient of variation c_S^2 of S. We distinguish between two cases, viz. (i) $c_S^2 \leq 1$ and (ii) $c_S^2 > 1$. In case (i), we proceed as follows. If

$$\frac{1}{k} \leq c_S^2 \leq \frac{1}{k-1},$$

for some $k \in \{2, 3, \dots\}$, then the approximating sojourn time is with probability p a sum of $k-1$ independent exponentials with common mean $1/\mu$ (hence an Erlang-$(k-1)$ distributed random variable), and with probability $1-p$ a sum of k independent exponentials with common mean $1/\mu$. By choosing (cf. [14])

$$p = \frac{1}{1+c_S^2}[kc_S^2 - (k(1+c_S^2) - k^2c_S^2)^{1/2}], \quad \mu = \frac{k-p}{\mathbb{E}S},$$

this so-called $E_{k-1,k}$ distribution matches $\mathbb{E}S$ and c_S^2.
In case (ii), we approximate $\mathbb{P}(S < x)$ by a hyperexponential distribution with balanced means, cf. p. 359 of [14]. Its density is given by

$$f(t) = r_1\mu_1 \mathrm{e}^{-\mu_1 t} + r_2\mu_2 \mathrm{e}^{-\mu_2 t}, \quad t > 0,$$

where $r_1 + r_2 = 1$ and $r_1, r_2 \geq 0$. We take $r_1 = \frac{1}{2}(1 + \sqrt{\frac{c_S^2-1}{c_S^2+1}})$, and choose μ_1 and μ_2 such that $r_1/\mu_1 = r_2/\mu_2$ (balanced means), hence $1/\mu_1 = \mathbb{E}S/(2r_1)$.

4 Numerical examples

This section contains a number of numerical examples in order to test the various approximations of the previous section. We restrict ourselves to the case in which customers visit the PS node at most twice: $p_2 = 0$, as this is the most relevant case for the blood testing example that occupies a central place in our study. For computing the approximation of $\mathbb{P}(S < t)$ we need $\mathbb{E}(S)$ and $\mathrm{Var}(S)$. Since the first moment can be computed exactly from (2.2) and (2.3), the variance is the crucial ingredient for approximating $\mathbb{P}(S < t)$. Therefore we will give some examples of computing $\mathrm{Var}(S)$ and compare it to values obtained by simulating the system. We let the service times $B_j \sim exp(1/\beta) = exp(\mu)$ for $j = 1, 2$.

4.1 Method I (IA)

In Method I, the first and second sojourn times in the PS node are assumed to be independent (an accurate approximation if $\beta \ll \delta$). Hence

$$S_{PS} = \begin{cases} S_1 & \text{w.p. } 1-p_1, \\ S_1 + S_2 & \text{w.p. } p_1, \end{cases} \tag{4.1}$$

and

$$\mathrm{Var}(S_{PS}) = \mathrm{Var}(S_1) + p_1\mathrm{Var}(S_2) + (p_1 - p_1^2)(\mathbb{E}S_2)^2.$$

Here $\mathrm{Var}(S_i)$, for both $i = 1, 2$, is the variance of the sojourn time in an $M/G/1$ queue with arrival rate $\lambda(1+p_1)$ and with service time distribution $\frac{1}{1+p_1}B_1(\cdot) + \frac{p_1}{1+p_1}B_2(\cdot)$. This service time distribution is again $exp(\mu)$, hence the calculations are somewhat simpler than those described in subsection 3.2 (compare in particular 3.12). This leads to

$$\mathrm{Var}(S_{PS}) = (1+p_1)\mathrm{Var}(S_1) + (p_1 - p_1^2)(\mathbb{E}(S_1))^2, \tag{4.2}$$

where S_1 is the sojourn time in an M/M/1-PS queue with arrival intensity $\lambda(1+p_1)$. By [6] we have

$$\mathrm{Var}(S_1) = \frac{1}{\mu^2(1-\rho)^2}\,\frac{2+\rho}{2-\rho}, \tag{4.3}$$

which leads to

$$\mathrm{Var}(S_{PS}) = \frac{1}{\mu^2(1-\rho)^2}\left(\frac{(1+p_1)(2+\rho)}{2-\rho} + p_1(1-p_1)\right). \tag{4.4}$$

4.2 Method II (SC)

This is the short-circuit assumption (valid if $\beta \gg \delta$), which is equivalent to having no delay node. Thus the service time distribution is given by

$$\begin{cases} B(x) = (1-p_1)B_1(x) + p_1 B_2(x), \text{ with} \\ B_1 \sim exp(\mu) \\ B_2 \sim Erl(2,\mu). \end{cases} \tag{4.5}$$

Therefore the system behaves like an M/G/1-PS queue, with service time distribution $B(x)$, for which the variance of the sojourn time is given by (3.13) (ignoring the hats that there referred to $\hat{B}(\cdot)$). By changing the integration order one can see that

$$\begin{aligned} &\int_0^\infty \int_0^x (x-y)F(y)\,\mathrm{d}y\mathrm{d}B(x) = \\ &\quad (1-p_1)\int_0^\infty F(y)h_1(y)\,\mathrm{d}y + p_1\int_0^\infty F(y)h_2(y)\,\mathrm{d}y = \\ &\quad (1-p_1)I_1 + p_1 I_2, \end{aligned} \tag{4.6}$$

where

$$\begin{cases} h_1(y) = \frac{1}{\mu}e^{-\mu y}, \\ h_2(y) = \frac{2+\mu y}{\mu}e^{-\mu y}. \end{cases} \tag{4.7}$$

By computing the first and second moment of the service time distribution (3.13) now gives

$$\mathrm{Var}(S_{PS}) = \frac{3+6p_1-p_1^2}{\mu^2(1-\rho)^2} - \frac{2}{(1-\rho)^2}((1-p_1)I_1 + p_1 I_2). \tag{4.8}$$

The function $F(\cdot)$ is the waiting time distribution in the corresponding M/G/1-FCFS queue, of which the LST is given by

$$\varphi(s) = \int_0^\infty e^{-sx}\,\mathrm{d}F(x) = \frac{(1-\rho)s}{s-\lambda(1-\beta(s))}, \tag{4.9}$$

where $\beta(s)$ is the LST of the service time distribution $B(\cdot)$. By integration by parts we can express the integral I_1 as $\varphi(\mu)$, yielding

$$I_1 = \frac{1}{\mu^2}\,\frac{4\mu(\rho-1)}{\lambda(2+p_1)-4\mu}. \tag{4.10}$$

For I_2 we get

$$\begin{aligned} I_2 &= 2I_1 - \left[\frac{d}{da}(\frac{1}{a}\varphi(a))\right]_{a=\mu} \\ &= 2I_1 - \frac{2(8\mu^3-2\lambda\mu^2)(\rho-1)}{\mu(4\mu^2-\lambda(2\mu+p_1\mu))^2}. \end{aligned} \tag{4.11}$$

It should be noted that in both Method I and II we subsequently use (3.9) to approximate the variance of the total sojourn time, and finally we use the method outlined in Subsection 3.2 to approximate the full sojourn time distribution $\mathbb{P}(S<t)$. The latter distribution is needed in the optimization problem that will be tackled in Section 5.

4.3 Examples

We give some numerical examples for calculating the variance of the total sojourn time based on the above discussed approximations. The Poisson arrival process has intensity $\lambda = \frac{1}{100}$.

Case $p_1 = 0.5$**:**

$PS \sim exp(0.1)$, $D \sim det(1)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	282.8	0.25	289.0
Method I (IA)	275.8	0.25	282.0
relative error (in %)	2.5	0	2.4
Method II (SC)	288.8	0.25	295.0
relative error (in %)	-2.1	0	-2.1
Method III (WA)	288.1	0.25	294.3
relative error (in %)	-1.8	0	-1.8

$PS \sim exp(1)$, $D \sim det(1)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	1.833	0.25	2.589
Method I (IA)	1.827	0.25	2.585
relative error (in %)	0.3	0	0.2
Method II (SC)	1.836	0.25	2.594
relative error (in %)	-0.2	0	-0.2
Method III (WA)	1.832	0.25	2.589
relative error (in %)	0.1	0	0

$PS \sim exp(0.1)$, $D \sim det(10)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	283.2	25	367.7
Method I (IA)	275.9	25	359.7
relative error (in %)	2.6	0	1.5
Method II (SC)	288.9	25	372.7
relative error (in %)	-2.0	0	-2.0
Method III (WA)	286.0	25	369.8
relative error (in %)	-1.0	0	-1.2

$PS \sim exp(1)$, $D \sim det(10)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	1.816	25	31.9
Method I (IA)	1.827	25	31.9
relative error (in %)	-0.6	0	0
Method II (SC)	1.836	25	31.9
relative error (in %)	-1.1	0	0
Method III (WA)	1.828	25	31.9
relative error (in %)	-0.7	0	0

The approximation of Var(S_{PS}) does not depend on the sojourn time distribution in the D node, except for its mean (as $\mathbb{E}S_D$ is used in the weighted average approximation method III, see below (3.8)). Therefore the above shown tables can easily be extended to arbitrary service (= sojourn) time distributions in the D node.

Case $p_1 = 1$: Also in this case we look at a deterministic service time distribution in the D node. Since we always have exactly two loops, Var$(S_D) = 0$. Furthermore S_{PS} and S_D are independent, which yields Var$(S) =$ Var$(S_{PS})+$ Var(S_D).

$PS \sim exp(0.1)$, $D \sim det(1)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	429.8	0	429.8
Method I (IA)	381.9	0	381.9
relative error (in %)	11.1	0	11.1
Method II (SC)	424.6	0	424.6
relative error (in %)	1.2	0	1.2
Method III (WA)	422.9	0	422.9
relative error (in %)	1.6	0	1.6

$PS \sim exp(1)$, $D \sim det(1)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	2.136	0	2.136
Method I (IA)	2.124	0	2.124
relative error (in %)	0.6	0	0.6
Method II (SC)	2.15	0	2.15
relative error (in %)	-0.7	0	0.7
Method III (WA)	2.142	0	2.142
relative error (in %)	-0.2	0	0.2

$PS \sim exp(0.1)$, $D \sim det(10)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	405	0	405
Method I (IA)	381.9	0	381.9
relative error (in %)	5.7	0	5.7
Method II (SC)	424.6	0	424.6
relative error (in %)	-4.8	0	4.8
Method III (WA)	412.4	0	412.4
relative error (in %)	-1.8	0	1.8

$PS \sim exp(1)$, $D \sim det(10)$

	Var(S_{PS})	Var(S_D)	Var(S)
Simulated Value	2.156	0	2.156
Method I (IA)	2.125	0	2.125
relative error (in %)	1.4	0	1.4
Method II (SC)	2.151	0	2.151
relative error (in %)	0.2	0	0.2
Method III (WA)	2.129	0	2.129
relative error (in %)	1.2	0	1.2

Conclusion: In all test cases, the three approximation methods of Section 3 yield very good results for the sojourn time variances. Method I usually yields the least good approximation, and Method III is slightly better than Method II; both latter methods result in errors that are typically below 2%.

5 Optimization

In this section we consider an optimization problem that is inspired by the blood testing example of Section 1, as well as by the quality conformance testing example of that section. In both examples, it is important to finish the whole testing procedure within a reasonable amount of time, say before a predetermined time t_0. In quality conformance it is important to have an upper bound on the quality control time of the items, because one wants to either deliver them or process them quickly without unnecessary delay. Therefore we assign a reward C_0 to each item that is processed in less than t_0 time units (or equivalently, an item has a sojourn time less than t_0). In blood testing, patients are entitled to have their results within a short period of time (i.e., before t_0) so that treatment can start immediately and the risk of infecting other individuals is diminished. Hence we assign a reward to having a sojourn time less than some value t_0. For the same reasons, in both applications excessively long sojourn times of items in the testing network have to be avoided so that we assign a penalty C_1 on each item whose sojourn time is greater than some prespecified time $t_1 > t_0$. The sojourn time may be reduced by speeding up the service in the PS node and/or the D node. This might be done by using better equipment and/or more (or better trained) personnel, but that involves costs. We study this trade-off problem, aiming at optimizing a 'reward'. Our decision variables are the mean service times at the PS and D nodes: the β_i and δ_i. We assume that there are costs $f(\beta_j)$ involved with having certain mean service times β_j, and similarly costs $g(\delta_j)$ with having mean service times δ_j. $f(\cdot)$ and $g(\cdot)$ are taken to be non-increasing functions. So the following reward per time unit must be maximized (notice that the multiplication by λ in the first terms results in rewards per customer times the number of customers per time unit):

$$C := \lambda C_0 \mathbb{P}(S < t_0) - \lambda C_1 \mathbb{P}(S > t_1) - \sum_{j=1}^{K} f(\beta_j) - \sum_{j=1}^{K-1} g(\delta_j). \tag{5.1}$$

Alternatively, we can put a constraint on large sojourn times by requiring that $\mathbb{P}(S > t_1) \leq \alpha$, for some small $\alpha > 0$, in which case the term $\lambda C_1 \mathbb{P}(S > t_1)$ has to be omitted from (5.1).

In our numerical example we treat the case $K = 2$, i.e.,

$$C := \lambda C_0 \mathbb{P}(S < t_0) - \lambda C_1 \mathbb{P}(S > t_1) - \sum_{j=1}^{2} f(\beta_j) - g(\delta_1), \tag{5.2}$$

and two choices of cost functions:

A. $f(\beta) = a_0 + a_1/\beta$ and $g(\delta) = b_0 + b_1/\delta$, where $\beta > \beta_0$ and $\delta > \delta_0$, with β_0 and δ_0 being the minimum possible mean service times, and a_0 and b_0 being the fixed minimum costs.

B. $f(\beta) = a_0 + a_1 e^{-\beta}$ and $g(\delta) = b_0 + b_1 e^{-\delta}$.

Note that the system parameters are $C_0, C_1, \lambda, a_0, b_0, a_1, b_1, \beta_0$ and δ_0. We try to choose β_1, β_2 and δ_1 such that the reward C is maximized. Since the optimization problem (5.2) is analytically intractable we deal with it by numerical methods.

First we note that the parameters a_0 and b_0 only cause a vertical displacement and can therefore be omitted. We also assume that the service time in the PS node is exponentially distributed with parameter $1/\beta$, where β stays the same for every loop. Further let the D node have a deterministic distribution with parameter δ. We also note that w.l.o.g. we can choose $C_0 = 1$ since other values would only result in another scaling of the whole objective function. The arrival process parameter λ can be chosen to be 1, since changing it would only result in a different time scaling. We have chosen $t_0 = 0.5$ and $t_1 = 3$. For the functions f and g we choose the ones given in A above.

The constrained problem: In this case the objective function reads as

$$\begin{cases} C = C_0 \mathbb{P}(S < t_0) - 2f(\beta) - g(\delta), \\ \mathbb{P}(S > t_1) \leq \alpha, \end{cases} \tag{5.3}$$

where we fix $\alpha = 0.1$. The parameters left are a_1 and b_1, so we will discuss different choices and the corresponding solutions β and δ:

a_1	b_1	β	δ
1/20	1/20	0.34	0.49
1/20	1/4	0.30	1.10
1/4	1/20	0.35	0.34
1/4	1/4	0.32	0.78

In all these examples the maximum is attained at the boundary imposed by the constraint.

The unconstrained problem: In this case the objective function is given by

$$C = C_0\mathbb{P}(S < t_0) - C_1\mathbb{P}(S > t_1) - 2f(\beta) - g(\delta). \tag{5.4}$$

In this case we have to choose another parameter, namely C_1 and this leads to the following solutions:

$C_1 = 1$:

a_1	b_1	β	δ
1/20	1/20	0.23	0.38
1/20	1/4	0.26	0.97
1/4	1/20	0.47	0.51
1/4	1/4	0.51	1.39

$C_1 = 5$:

a_1	b_1	β	δ
1/20	1/20	0.20	0.34
1/20	1/4	0.20	0.76
1/4	1/20	0.30	0.33
1/4	1/4	0.30	0.71

To illustrate the problem better we also include a plot of the objective function. The chosen parameters are $C_1 = 5$, $a_1 = b_1 = 1/4$. The plot can be found in Figure 2.

Furthermore we give some results for the set B of functions f and g in the unconstrained case, i.e.

$$\begin{cases} f(\beta) = a_1 e^{-\beta}, \\ g(\delta) = b_1 e^{-\delta}. \end{cases} \tag{5.5}$$

We choose $\delta \geq \delta_0 = 0.1$ and $\beta \geq \beta_0 = 0.1$ as minimal service time requirements. As above, $C_1 = 5$.

a_1	b_1	β	δ
1	1	0.10	0.79
1	2.5	0.10	1.75
2.5	1	0.25	0.77
2.5	2.5	0.23	1.60

We note that in the first case the optimal points are at the boundary imposed by the constraints $\delta \geq 0.1$ and $\beta \geq 0.1$. In the second line appears $\beta = 0.10$

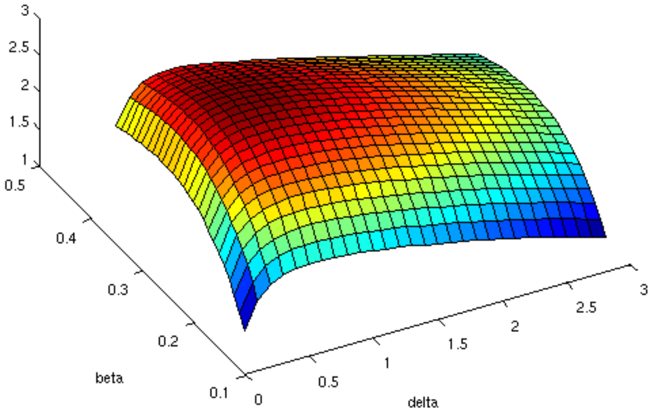

Figure 2: The objective function for the unconstrained case

while with more accuracy this should be $\beta = 0.1021$. This means that this point is not at the boundary.

Acknowledgment. The research of Onno Boxma was supported by the BSIK/BRICKS project.

References

[1] J. Abate and W. Whitt. The Fourier-series method for inverting transforms of probability distributions. Queueing Systems 10 (1992) 5-88.

[2] O.J. Boxma and H. Daduna. Sojourn times in queueing networks. In: H. Takagi (ed.). Stochastic Analysis of Computer and Communication Systems (North-Holland Publ. Cy., Amsterdam, 1990), pp. 401-450.

[3] O.J. Boxma, B.M.M. Gijsen, R.D. van der Mei and J.A.C. Resing. Sojourn-time approximations in two-node queueing networks. Proceedings 2nd international working conference on Performance Modelling and Evaluation of Heterogeneous Networks, HETNETs (Ilkley, UK, July 2004).

[4] F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios. Open, closed and mixed networks of queues with different classes of customers. J. ACM 22 (1975) 248-260.

[5] O.J. Boxma, R.D. van der Mei, J.A.C. Resing and K.M.C. van Wingerden. Sojourn time approximations in a two-node queueing network. In: Proc. ITC-19, Beijing, 2005, pp. 1121-1133.

[6] E.G. Coffman, R.R. Muntz and H. Trotter. Waiting time distributions for processor-sharing systems. J. ACM 17 (1970) 123-130.

[7] R.L. Disney and P.C. Kiessler. *Traffic Processes in Queueing Networks: A Markov Renewal Approach* (Johns Hopkins University Press, Baltimore, 1987).

[8] R.D. Foley and R.L. Disney. Queues with delayed feedback. Advances in Applied Probability 15 (1988) 162-182.

[9] S.D. Hohl and P.J. Kuehn. Approximate analysis of flow and cycle times in queuing networks. In: L.F.M. de Moraes, E. de Souza e Silva and L.F.G. Soares (eds.). Proc. 3rd Int. Conf. on Data Communication Systems and their Performance (North-Holland Publ. Cy., Amsterdam, 1987), pp. 471-485.

[10] R.D. van der Mei, B.M.M. Gijsen, N. in 't Veld and J.L. van den Berg. Response times in a two-node queueing network with feedback. Performance Evaluation 49 (2002) 99-110.

[11] T.J. Ott. The sojourn time distribution in the M/G/1 queue with processor sharing. Journal of Applied Probability 21 (1984) 360-378.

[12] B. Sengupta. An approximation for the sojourn-time distribution for the $GI/G/1$ processor-sharing queue. Stochastic Models 8 (1992) 35-57.

[13] J.G. Shanthikumar and J.A. Buzacott. The time spent in a dynamic job shop. European Journal of Operational Research 17 (1984) 215-226.

[14] H.C. Tijms. *Stochastic Models: An Algorithmic Approach.* Wiley, Chichester, 1994.

[15] M. van Vuuren. *Performance Analysis of Manufacturing Systems.* PhD Thesis, Eindhoven University of Technology, 2007.