Sojourn time approximations in a two-node queueing network *

O.J. Boxma[a,b], R.D. van der Mei[b,c], J.A.C. Resing[a], K.M.C. van Wingerden[b]

[a]Eindhoven University of Technology, Mathematics and Computer Science
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[b]CWI, Advanced Communication Networks
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[c]Vrije Universiteit, Faculty of Sciences
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

February 28, 2005

Abstract

We develop a method for approximating sojourn time distributions in open queueing networks. The work is motivated by the performance analysis of distributed information systems, where transactions are handled by iterative server and database actions. In this paper we restrict ourselves to a system with one server and a single database, modelled as an open two-node queueing network with a processor sharing node and a first-come first-served node. Extensive numerical results are presented for approximations of the mean sojourn time. The accuracy of the approximations is validated with simulations.

Keywords: two-node open queueing network, sojourn times, approximation.

1 Introduction

The dramatic growth of the Internet and the popularity of PCs have boosted the emergence of so-called Web services technology to compose new and advanced services on top of existing basic services. A typical example of such services that are built on top of existing services is holiday package reservation where the consumer can make a reservation for a hotel, a car and an airline ticket at once. Other examples are PC banking, and on-line services offered by a telephone company that enable the customers to check the status of telephone bills at their home PC with Internet access. A typical feature of this type of distributed applications is that a single transaction initiated by the end user may induce a sequence of server and database transactions. A key factor for the success of this type of services is that the response times observed by the end user are not overly long. This has motivated us to study response times in distributed systems in a queueing-theoretical framework, where the customers represent end-user initiated transactions, the network nodes represent the existing basic services and the response times are modeled as the sojourn times of a customer in the system.

In this paper, we consider a two-node open queueing network with a processor sharing (PS)

*Work carried out within the project EQUANET, funded by the Dutch Ministry of Economic Affairs; part of the research also fits in the European Network of Excellence Euro-NGI. Some preliminary results were presented in [2].

node and a first-come first-served (FCFS) node. External customers arrive at the PS node according to a Poisson process with rate λ. A departing customer subsequently enters the FCFS node with probability p, and leaves the system with probability $1-p$. Upon departure from the FCFS node, a customer always returns to the PS node (see Figure 1). All service times at all visits to both nodes are independent random variables, with distribution $B_{PS}(\cdot)$ and $B_F(\cdot)$ at the PS- and FCFS node, with mean β_{PS} and β_F, respectively. The total load at the PS- and FCFS node is $\rho_{PS} := \lambda\beta_{PS}/(1-p)$ and $\rho_F := p\lambda\beta_F/(1-p)$, respectively, and we assume that both loads are less than one. The PS node may represent a web server, and the FCFS node a database server. Successive visits of a customer represent a sequence of web server and database transactions.

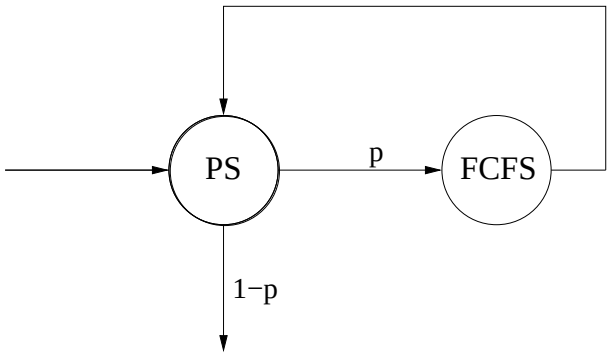

Figure 1: The two-node open queueing network

It is notoriously hard to obtain exact results for sojourn times, especially if some form of overtaking of customers occurs (see [1] for an overview of results on sojourn times in queueing networks). Both processor sharing and feedback induce such overtaking. If the service times at the FCFS node are exponentially distributed, then the joint queue length distribution has a product form, and the mean queue lengths are easily obtained, yielding the mean sojourn times via Little's formula; but otherwise, even the mean sojourn times are not known. The complexity of the problem of obtaining sojourn time results in queues with non-instantaneous feedback was discussed in [5]. Hence in this paper we are looking for approximation methods for the sojourn time distribution in the network. Our study is related to [7]. We consider the same model, extending [7] in three ways: (i) we allow *general* service time distributions in both nodes, (ii) we present a more general approximation method that allows the approximation of sojourn time *distributions* while requiring somewhat less restrictive approximation assumptions, and (iii) we suggest several approximation refinements.

The rest of the paper is organized as follows. In Section 2, we describe several approximation methods, with increasing complexity and accuracy. Section 3 contains an extensive numerical evaluation of the different methods, concentrating on mean sojourn times – discussion of the accuracy of the approximation methods for sojourn time *variances* and *distributions* is the topic of another paper. Finally, in Section 4 we indicate some model extensions which can also be handled by our methods.

2 Description of the approximation methods

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

$$\mathsf{E}[\mathrm{e}^{-\omega_1 S_{PS} - \omega_2 S_F}] = \sum_{k=0}^{\infty} (1-p)p^k \mathsf{E}[\mathrm{e}^{-\omega_1 S_{PS}^{(k+1)} - \omega_2 S_F^{(k)}}]. \tag{2.1}$$

Of course, taking $\omega_1 = \omega_2$ yields an expression for the total sojourn time, S, of a customer in the system. We shall present four different approximation methods.

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

- *Assumption 1.* $S_{PS}^{(k+1)}$ and $S_F^{(k)}$ are independent, $k = 0, 1, \ldots$.
- *Assumption* 2^a. $S_{PS}^{(k+1)}$ is distributed as the sum of $k+1$ independent, identically distributed terms. The individual terms are distributed as σ_{PS}, the stationary sojourn time in an $M/G/1$ PS node with arrival rate $\lambda/(1-p)$ and service time distribution $B_{PS}(\cdot)$. Similarly, $S_F^{(k)}$ is distributed as the sum of k independent, identically distributed terms, where each individual term is distributed as σ_F, the stationary sojourn time in an $M/G/1$ FCFS node with arrival rate $\lambda p/(1-p)$ and service time distribution $B_F(\cdot)$.

It should be noticed that the IA, also called *Independent Flow Time Approximation*, is a classic approximation method, that was proposed for a large class of queueing networks in [6, 10]. From (2.1) and Assumptions 1 and 2^a we obtain:

$$\mathsf{E}[\mathrm{e}^{-\omega_1 S_{PS} - \omega_2 S_F}] \approx \sum_{k=0}^{\infty} (1-p)p^k (\mathsf{E}[\mathrm{e}^{-\omega_1 \sigma_{PS}}])^{k+1} (\mathsf{E}[\mathrm{e}^{-\omega_2 \sigma_F}])^k. \tag{2.2}$$

In particular, we find from (2.2) for the mean sojourn times

$$\mathsf{E}S_{PS} \approx \sum_{k=0}^{\infty} (1-p)p^k (k+1)\mathsf{E}\sigma_{PS} = \frac{\beta_{PS}}{(1-p)(1-\rho_{PS})}, \tag{2.3}$$

$$\begin{aligned} \mathsf{E}S_F \approx \sum_{k=0}^{\infty} (1-p)p^k k \mathsf{E}\sigma_F &= \frac{p}{1-p}\left[\beta_F + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right] \\ &= \frac{p}{1-p}\left[\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}(\beta_F^{(2)} - 2\beta_F^2)\frac{p}{(1-p)(1-\rho_F)}\right], \end{aligned} \tag{2.4}$$

with $\beta_F^{(2)}$ the second moment of the service time distribution in the FCFS node.

Successive sojourn times at the PS (FCFS) 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 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.* $S_{PS}^{(k+1)}$ and $S_F^{(k)}$ are independent, $k = 0, 1, \ldots$.
- *Assumption* 2^b. $S_{PS}^{(k)}$ has the same distribution as $\sigma_{PS}^{(k)}$, the total sojourn time after k visits in the PS node short-circuited (i.e., with the FCFS node removed). Similarly, $S_F^{(k)}$ has the same distribution as $\sigma_F^{(k)}$, the total sojourn time after k visits in the FCFS node short-circuited (i.e., with the PS node removed).

From (2.1) and Assumptions 1 and 2^b we obtain:

$$\mathsf{E}[\mathrm{e}^{-\omega_1 S_{PS} - \omega_2 S_F}] \approx \sum_{k=0}^{\infty} (1-p)p^k \mathsf{E}[\mathrm{e}^{-\omega_1 \sigma_{PS}^{(k+1)}}]\mathsf{E}[\mathrm{e}^{-\omega_2 \sigma_F^{(k)}}]. \tag{2.5}$$

The LST $\mathsf{E}[\mathrm{e}^{-\omega_2 \sigma_F^{(k)}}]$ for an $M/G/1$ FCFS queue with instantaneous feedback follows from Doshi and Kaufman [4]. The LST $\mathsf{E}[\mathrm{e}^{-\omega_1 \sigma_{PS}^{(k+1)}}]$ for an $M/G/1$ PS queue with instantaneous feedback is the same as the LST of the sojourn time of a tagged customer with as service time the sum of the $k+1$ service times $B_1, \ldots, B_{k+1}$ in an $M/G/1$ PS queue without feedback, with

as service time for an arbitrary customer the sum of a geom(p) distributed number of service times B_i (note that the tagged customer has exactly $k+1$ passes through the feedback queue, but that an arbitrary customer has a geom(p) distributed number of passes). Theorem 2.2 of Ott [8] 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 $B_1+\cdots+B_{k+1}$ gives $\mathsf{E}[\mathrm{e}^{-\omega_1\sigma_{PS}^{(k+1)}}]$. These LST expressions in [4, 8] are quite complicated. If one is satisfied by just obtaining an approximation for the mean and variance of $S_{PS}^{(j)}$ and $S_F^{(j)}$ (and hence of S_{PS} and S_F), then relatively easy expressions for the means and variances of the former random variables can be taken from [8] and from either [4] or [11].

In particular, we find from (2.5) for the mean sojourn times

$$\mathsf{E}S_{PS} \approx \sum_{k=0}^{\infty}(1-p)p^k\mathsf{E}\sigma_{PS}^{(k+1)} = \frac{\beta_{PS}}{(1-p)(1-\rho_{PS})}, \tag{2.6}$$

$$\mathsf{E}S_F \approx p\sum_{k=1}^{\infty}(1-p)p^{k-1}\mathsf{E}\sigma_F^{(k)} = \frac{p}{1-p}\left[\frac{\beta_F}{1-\rho_F}+\frac{\lambda}{2}(\beta_F^{(2)}-2\beta_F^2)\frac{p}{1-\rho_F}\right]. \tag{2.7}$$

The latter expression equals p times the mean sojourn time in the $M/G/1$ queue with (instantaneous) Bernoulli feedback (see Formula (35) of Takács [11]); the multiplication by p reflects that the FCFS queue is not visited at all with probability $1-p$.

Method III: Weighted Average Approximation (WA)
Especially when the mean sojourn times at both queues are roughly equal, then the approximations in (2.2) and (2.5) can be improved in the following way. Replace the LST's in the righthand side of (2.1) 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):

$$\begin{aligned}\mathsf{E}[\mathrm{e}^{-\omega_1 S_{PS}-\omega_2 S_F}] \approx & \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathsf{E}[\mathrm{e}^{-\omega_1\sigma_{PS}^{(k+1)}}]+(1-w)(\mathsf{E}[\mathrm{e}^{-\omega_1\sigma_{PS}}])^{k+1}\right)\\ & \left((1-w)\mathsf{E}[\mathrm{e}^{-\omega_2\sigma_F^{(k)}}]+w(\mathsf{E}[\mathrm{e}^{-\omega_2\sigma_F}])^k\right).\end{aligned} \tag{2.8}$$

We choose the weight w as $w = \mathsf{E}\sigma_{PS}^{(1)}/[\mathsf{E}\sigma_{PS}^{(1)}+\mathsf{E}\sigma_F^{(1)}]$ (alternatively, we could have chosen, e.g., $w = \mathsf{E}\sigma_{PS}/[\mathsf{E}\sigma_{PS}+\mathsf{E}\sigma_F]$).

For the mean sojourn times we find in this case

$$\mathsf{E}S_{PS} \approx \frac{\beta_{PS}}{(1-p)(1-\rho_{PS})}, \tag{2.9}$$

$$\mathsf{E}S_F \approx \frac{p}{1-p}\left[\frac{\beta_F}{1-\rho_F}+\frac{\lambda}{2}(\beta_F^{(2)}-2\beta_F^2)\frac{p}{1-\rho_F}\left(1-w+\frac{w}{1-p}\right)\right]. \tag{2.10}$$

Method IV: Weighted Average with Correction (WAC)
In the approximation based on the Independence Assumption (and hence also partly in the Weighted Average Approximation), we approximated individual sojourn times at the PS and FCFS node by stationary sojourn times in $M/G/1$ queues. However, flows in the network are clearly not Poisson flows [3] and hence the approximations probably can be improved by approximating individual sojourn times at the PS and FCFS node by stationary sojourn times in suitably chosen $GI/G/1$ queues.

Clearly the arrival rates λ_{PS} and λ_F at the PS and the FCFS node are given by $\lambda_{PS} = \lambda/(1-p)$ and $\lambda_F = \lambda p/(1-p)$, respectively. In order to approximate the squared coefficients of variation $c_{A_{PS}}^2$, $c_{D_{PS}}^2$, $c_{A_F}^2$, and $c_{D_F}^2$ of the interarrival and interdeparture times at the PS and the FCFS node, we used the following four approximate equations:

$$c^2_{D_{PS}} = \rho^2_{PS} + (1-\rho^2_{PS})c^2_{A_{PS}}, \tag{2.11}$$

$$c^2_{A_F} = pc^2_{D_{PS}} + 1 - p, \tag{2.12}$$

$$c^2_{D_F} = \rho^2_F c^2_{B_F} + (1-\rho^2_F)c^2_{A_F}, \tag{2.13}$$

$$c^2_{A_{PS}} = 1 - vp + vpc^2_{D_F}, \tag{2.14}$$

where

$$v = \left[1 + 4(1-\rho_{PS})^2\left(\frac{1}{(1-p)^2+p^2} - 1\right)\right]^{-1}.$$

Here, (2.11) is based on [14], where it is backed by extensive simulation results. Noticably, this approximate relation was found to be remarkably accurate regardless of the service time distribution. Furthermore, (2.12),(2.13) and (2.14) are based on Whitt's QNA paper [12]. Remark that the approximate formulas for $c^2_{D_{PS}}$ and $c^2_{D_F}$ are similar, except for one major difference: The service time variability does not appear in the formula for $c^2_{D_{PS}}$, reflecting near-insensitivity.

From (2.11-2.14) we get the following expressions for $c^2_{A_{PS}}$ and $c^2_{A_F}$:

$$c^2_{A_{PS}} = 1 - \frac{pv\rho^2_F(1-c^2_{B_F})}{1-p^2v(1-\rho^2_{PS})(1-\rho^2_F)},$$

$$c^2_{A_F} = \frac{1-p^2v(1-\rho^2_{PS})(1-\rho^2_F c^2_{B_F})}{1-p^2v(1-\rho^2_{PS})(1-\rho^2_F)}.$$

Once we have approximated these squared coefficients of variation of the interarrival times at the two nodes, we can approximate the sojourn times of the individual visits at the nodes by using results of Sengupta [9] for the $GI/G/1$ PS node and by using results of Whitt [12] for the $GI/G/1$ FCFS node.

The approximation for the mean sojourn time in the $GI/G/1$ PS queue in [9] is given by $\mathsf{E}\sigma_{PS} \approx \beta_{PS}/(1-\eta)$, where η is the smallest positive root of the equation $\eta = \alpha((1-\eta)/\beta_{PS})$ with $\alpha(\cdot)$ the Laplace-Stieltjes transform of the interarrival times. For the interarrival time distribution we choose a hyperexponential distribution of order two with balanced means, with mean $1/\lambda_{PS}$ and squared coefficient of variation $c^2_{A_{PS}}$. Note that Sengupta's approximation for the mean sojourn time is only based on the mean service time and not on the coefficient of variation $c^2_{B_{PS}}$.

The approximation in [12] for the mean sojourn time in a $G/G/1$ FCFS queue with mean service time β, load ρ and squared coefficients of variation c^2_A, c^2_B reads:

$$\mathsf{E}\sigma \approx \beta + \frac{\beta\rho\left(c^2_A + c^2_B\right)g}{2(1-\rho)},$$

where

$$g = \begin{cases} exp\left[-\frac{2(1-\rho)}{3\rho}\frac{(1-c^2_A)^2}{c^2_A+c^2_B}\right], & \text{if } c^2_A < 1, \\ 1, & \text{if } c^2_A \geq 1. \end{cases}$$

This formula for the $GI/G/1$ queue does not take into account feedback. Following an idea of Whitt [13], we use the above $GI/G/1$ formula, eliminating feedback by taking instead load $\rho_F = \lambda\beta_F p/(1-p)$ (as before) and replacing c^2_B by $p + (1-p)c^2_{B_F}$. As a consequence, the total mean sojourn time in the PS node and the FCFS node in the WAC method can be approximated by the following expressions:

$$\mathsf{E}S_{PS} \approx \frac{1}{1-p}\left(w\frac{\beta_{PS}}{1-\rho_{PS}} + (1-w)\frac{\beta_{PS}}{1-\eta}\right), \tag{2.15}$$

and

$$\begin{aligned} \mathsf{E}S_F \approx\ & (1-w)\frac{p}{1-p}\left[\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}(\beta^{(2)}_F - 2\beta^2_F)\frac{p}{1-\rho_F}\right] \\ +\ & w\frac{p}{1-p}\left[\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}\left(\beta^{(2)}_F - 2\beta^2_F + \frac{\beta^2_F}{1-p}\left(c^2_{A_F} + 1 - \frac{2}{g}\right)\right)\frac{pg}{1-\rho_F}\right]. \end{aligned} \tag{2.16}$$

3 Numerical results

To assess the accuracy of the approximations discussed in Section 2, we have performed numerical experiments, comparing the approximations with simulations. We have checked the accuracy of the approximations for many parameter combinations, by varying the arrival rate (λ), the loads per queue (ρ_{PS}, ρ_F), the variability in the service-time distributions (c^2_{PS}, c^2_F) and the feedback probability (p). From the simulations we have calculated point estimations for the mean sojourn times, and 95% confidence intervals. Denoting the point estimations from the simulations by z_{sim} and the approximations by z_{app}, the relative error of the approximations is defined as

$$\Delta\% = \frac{z_{app} - z_{sim}}{z_{sim}} \times 100\%. \tag{3.1}$$

The results of the experiments are outlined below.

Table 1 below shows the results for highly asymmetrical systems in which one node is heavily loaded (i.e., 90%) while the other one is lightly loaded (i.e., 10%). The service times are H_2-distributed (with balanced means), and $p = 0.8$. For any model instance, the mean service times (β_{PS} and β_F) were chosen such that the listed load-values were realized. Table 1 shows the point estimations and approximations for the total mean sojourn times, and the corresponding relative errors. Confidence intervals are not shown for compactness of the presentation. The results presented in Table 1 show that the simple "naive" approximation, IA,

λ	ρ_{PS}	ρ_F	c^2_{PS}	c^2_F	sim	IA	$\Delta\%$	SC	$\Delta\%$
0.36	0.9	0.1	1.67	1.67	25.2	25.3	0.6	25.3	0.6
0.09	0.9	0.1	1.67	4.56	100.8	101.5	0.6	101.3	0.5
0.36	0.9	0.1	4.56	4.56	25.6	25.4	-0.9	25.3	-1.1
0.09	0.9	0.1	4.56	1.67	103.2	101.3	-1.9	101.2	-1.9
0.01	0.1	0.9	1.67	1.67	958.5	1181.1	23.2	965.1	0.7
0.04	0.1	0.9	1.67	4.56	302.0	587.8	94.6	299.8	-0.7
0.04	0.1	0.9	4.56	4.56	299.4	587.8	96.3	299.8	0.1
0.01	0.1	0.9	4.56	1.67	951.2	1181.1	24.2	965.1	1.5

Table 1: $E[S]$ for highly asymmetically loaded systems with varying variability of the service-time distributions: approximations versus simulations.

that completely ignores any correlations between successive sojourn times may lead to large relative errors, particularly in the cases where the FCFS node is dominant. Note that in case the PS node is dominant the correlation between successive sojourn times does not play a key role for the mean sojourn times, which explains why IA performs quite well in the first four models. Moreover, the results show that the SC approximation, which does take into account correlations between successive sojourn times, here leads to highly accurate results.

When the loads of the queues are roughly the same, the SC approximation becomes inaccurate (see also Tables 3 to 5 below), which has motivated us to develop the refined approximations WA and WAC. To assess their accuracy, we consider the set of model instances listed in Table 2 below. The parameter sets have been constructed such that the loads at both queues are the same ($\rho_{PS} = \rho_F =: \rho$) and significant ($\rho = 0.5$ or 0.8) while the squared coefficient of variation of the service time distributions is varied (only results for values larger than 1 are presented). Table 3 shows the results for the total mean sojourn times $E[S]$ for each of the twelve models specified in Table 2. The results in Table 3 show that indeed the IA and SC approximations tend to become less accurate when the loads of the two servers are roughly equal. Moreover, we observe that the WA and WAC approximations, which *do* take into account the impact of "delayed feedback", indeed lead to much more accurate performance predictions.

Number	p	λ	ρ	c^2_{PS}	c^2_F
1	0.5	0.20	0.8	1.67	1.67
2	0.5	0.80	0.8	1.67	1.67
3	0.8	0.08	0.8	1.00	4.56
4	0.8	0.10	0.5	1.00	4.56
5	0.8	0.10	0.5	4.56	4.56
6	0.8	0.10	0.5	1.67	4.56
7	0.8	0.08	0.8	1.67	1.67
8	0.8	0.32	0.8	1.67	1.67
9	0.8	0.08	0.8	1.67	4.56
10	0.8	0.08	0.8	4.56	4.56
11	0.8	0.32	0.8	4.56	1.67
12	0.8	0.08	0.8	4.56	1.67

Table 2: Parameters for the model instances to be tested when the loads at both nodes are equal.

Number	sim	IA	$\Delta\%$	SC	$\Delta\%$	WA	$\Delta\%$	WAC	$\Delta\%$
1	44.1	45.3	2.8	42.7	-3.2	43.5	-1.4	43.8	-0.7
2	11.1	11.3	2.5	10.7	-3.5	10.9	-1.7	11.0	-0.9
3	154.1	171.1	11.0	114.2	-25.9	133.9	-13.1	141.3	-8.4
4	24.6	28.9	17.7	21.8	-11.3	24.2	-1.4	23.1	-6.0
5	23.5	28.9	23.0	21.8	-7.3	24.2	3.1	23.1	-1.7
6	24.2	28.9.	19.3	21.8	-10.0	24.2	0.0	23.1	-4.6
7	109.6	113.3	3.5	102.7	-6.3	107.2	-2.2	107.5	-1.9
8	27.4	28.33	3.3	25.7	-6.5	26.8	-2.4	26.9	-2.1
9	145.2	171.1	17.9	114.2	-21.3	133.9	-7.8	141.3	-2.7
10	133.1	171.11	28.6	114.2	-14.2	133.9	0.6	141.3	6.2
11	26.8	28.33	5.7	25.7	-4.3	26.89	-0.1	26.9	0.2
12	107.9	113.3	5.0	102.7	-4.9	107.2	-0.6	107.5	-0.5

Table 3: $E[S]$ for symmetrically loaded systems with varying variability of the service-time distributions: approximations versus simulations.

To analyze the accuracy of the approximations in more detail, Tables 4 and 5 show the results for the PS node and the FCFS node, respectively. The results in Table 4 show that

Number	sim	IA/SC/WA	Δ%	WAC	Δ%
1	21.12	20.00	-5.3	21.05	-0.3
2	5.27	5.00	-5.2	5.26	-0.2
3	76.22	50.00	-34.4	73.45	-3.6
4	11.97	10.00	-16.4	10.98	-8.3
5	11.10	10.00	-9.9	10.98	-1.1
6	11.71	10.00	-14.6	10.98	-6.2
7	54.25	50.00	-7.8	53.95	-0.6
8	13.59	12.50	-8.0	13.49	-0.7
9	70.46	50.00	-29.0	73.45	4.2
10	61.68	50.00	-18.9	73.45	19.1
11	13.12	12.50	-4.75	13.49	2.8
12	52.53	50.00	-4.81	53.95	2.7

Table 4: $E[S_{PS}]$ for symmetrically loaded systems with varying variability of the service-time distributions: approximations versus simulations.

Number	sim	IA	Δ%	SC	Δ%	WA	Δ%	WAC	Δ%
1	22.97	25.3	10.3	22.67	-1.3	23.47	2.2	22.75	-1.0
2	5.78	6.3	9.6	5.67	-2.0	5.87	1.5	5.69	-1.6
3	77.86	121.11	55.5	64.22	-17.5	83.91	7.8	67.88	-12.8
4	12.58	18.9	50.2	11.78	-6.4	14.22	13.0	12.11	-3.7
5	12.38	18.9	52.6	11.78	-4.8	14.22	14.9	12.11	-2.2
6	12.50	18.9	51.1	11.78	-5.8	14.22	13.7	12.11	-3.2
7	55.30	63.3	14.5	50.00	-4.8	57.17	3.4	53.50	-3.3
8	13.86	15.8	14.3	13.17	-5.0	14.29	3.1	13.38	-3.5
9	74.72	121.1	62.1	64.22	-14.1	83.91	12.3	67.88	-9.2
10	71.42	121.1	69.6	64.22	-10.1	83.91	17.5	67.88	-5.0
11	13.69	15.8	15.7	13.17	-3.8	14.29	4.4	13.38	-2.3
12	55.43	63.3	14.3	52.67	-5.0	57.17	3.1	53.50	-3.5

Table 5: $E[S_F]$ for symmetrically loaded systems with varying variability of the service-time distributions: approximations versus simulations.

the total mean sojourn time at the PS node may deviate quite strongly from the real (i.e., simulated) value for the IA, SC and WA approximation. Note that these approximations are the same (see also the respective relations in Section 2), and do not take into account the second moment of the service time distributions at the PS node. Moreover, we observe that these approximations consistently underestimate the mean sojourn time at the PS node. This can be explained by the fact that the approximations assume Poisson arrivals at the PS node, while the simulation results show that the arrival process at the PS node is "burstier than Poisson" (in terms of the squared coefficient of variation of the interarrival times, exceeding one) for all the model instances listed in Table 2. Note that this observation is not generally true for service-time distributions with squared coefficient of variation smaller than one. We also observe that the WAC-approximation for the mean sojourn time at the PS node, that explicitly takes into account the fact that the arrival process at the PS node is not Poisson (although correlations between the interarrival times are not taken into account), performs significantly better with relative errors of only a few percent (except for an outlier in model 10).

The results in Table 5 for the FCFS queue show similar results. The IA approximation performs badly in all cases, and the SC and WA approximation perform better, but still with errors that sometimes exceed 10%. Again, the WAC approximation improves the accuracy

by an order of magnitude (except for an outlier in model 3). We emphasize that the models listed in Table 2 are worst-case models, and the approximations should be judged from that perspective.

In summary, the WAC approximation outperforms the other approximations. Apparently, the inclusion of (1) a weighing between the extreme cases of immediate feedback and feedback after an infinite amount of time to the same queue and (2) non-Poisson arrival processes in the approximation captures the dominant factors that determine the sojourn times.

4 Extensions

The results from this paper can be extended in various ways. First, although the focus in this paper has been on the mean sojourn times, the methods are also applicable for approximating higher moments and even distributions of the sojourn times. Working out the details and validating the quality of the resulting approximations is an interesting topic for further research. Second, the relatively good accuracy of the WAC approximation that captures non-Poisson arrivals suggests that further refinements can be obtained if the correlations in the arrival processes at the nodes are captured. This requires in-depth characterization of the correlation structures within the arrival processes, and the development of approximations for the sojourn times that take into account these correlation structures. Finally, from an application point of view extension of the results to queueing networks with more than two nodes and to queueing networks with deterministic routing is of key interest.

References

[1] 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.

[2] 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, HET-NETs (Ilkley, UK, July 2004).

[3] R.L. Disney and P.C. Kiessler. Traffic processes in queueing networks: a Markov renewal approach (Johns Hopkins University Press, Baltimore, 1987).

[4] B.T. Doshi and J.S. Kaufman. Sojourn time in an M/G/1 queue with Bernoulli feedback. In: O.J. Boxma, R. Syski (eds.). Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen (North-Holland Publ. Cy., Amsterdam, 1988), pp. 207-233.

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

[6] 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.

[7] 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.

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

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

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

[11] L. Takács. A single-server queue with feedback. *The Bell System Technical Journal* 42 (1963) 505-519.

[12] W. WHITT. The Queueing Network Analyzer. *The Bell System Technical Journal* 62 (1983) 2779-2815.

[13] W. WHITT. Performance of the Queueing Network Analyzer. *The Bell System Technical Journal* 62 (1983) 2817-2843.

[14] K.M.C. VAN WINGERDEN. Approximations for Sojourn Times in Queueing Networks. M.Sc. thesis, Tilburg University, Department of Econometrics, February 2005.