Sojourn Time Approximations in Queueing Networks with Feedback[1]

B.M.M. Gijsen[a], [2]R.D. van der Mei[b,c], P. Engelberts[a], J.L. van den Berg[a,d] and K.M.C. van Wingerden[b]

[a]TNO Telecom, Quality of Service, P.O. Box 5050, 2600 GB Delft, The Netherlands
[b]CWI, Advanced Communication Networks, P.O. Box 94079, 1098 SJ Amsterdam, The Netherlands
[c]Vrije Universiteit, Faculty of Sciences, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[d]University of Twente, Faculty of Sciences, P.O. Box 217, 7500 AE Enschede, The Netherlands

This paper is motivated by the response-time analysis of distributed information systems, where transactions are handled by a sequence of front-end server and back-end server actions. We study sojourn times in an open queueing-network with a single Processor Sharing (PS) node and an arbitrary number of M multi-server First-Come-First-Served (FCFS) nodes. Customers arrive at the PS according to a Poisson process. After departing from the PS node a customer jumps to FCFS node k with probability p_k, and departs from the system with probability $1-p$, where $p = \sum_{k-1}^{M} p_k$ $(0 < p < 1)$. After receiving service at a FCFS node, a customer jumps back to the PS node. For this model, we focus on the mean and the variability of the sojourn time of an arbitrary customer in the system. The model is a product-form network, which immediately leads to a closed-form expression for the mean sojourn times. The variance of the sojourn times, however, does not admit an exact expression; the complexity is caused by the possibility of overtaking. To this end, we propose a new methodology for deriving closed-form approximations for the variance of sojourn times in queueing networks with feedback. Numerical results from extensive experimentation with simulations demonstrates that the approximations are highly accurate for a wide range of parameter values.

Keywords: queueing networks, sojourn time, response time, feedback, approximation

1 Motivation and Background

The recent emergence of Web technology has boosted to the development of applications running in a distributed computing environment where data is collected from diverse and remote information systems, and processed before a response is returned to the end user. A typical feature of such applications is that a single transaction initiated by the end user may initiate a cascade of sub-transactions to be performed on the different information system, each of which can handle a number of sub-transactions in parallel. Examples of such distributed applications are on-line ticketing, electronic banking, on-line shopping, location-based services and the like. A key factor for the success of this type of distributed applications is the ability to predict and control the performance in terms of the end-to-end response times, i.e. the response times experienced by the end user. For the user-perceived quality, both the *mean* and the *variability* of the response times are of key metrics.

Motivated by this, we model the end-to-end response time as the sojourn time of a customer in an open queueing network, where the customers represent transactions, the nodes represent the different application servers and information systems. In this context, we focus on the mean and the variance of the sojourn times of customers in the network. We give exact expressions for the mean sojourn times. In the absence of exact results for the variance of the

[1]This work was done in the context of the project End-to-end Quality of Service in Next Generation Network (EQUANET), sponsored by SenterNovem.

[2]Corresponding author. E-mail: mei@cwi.nl.

sojourn times, we develop a new method for deriving simple and explicit and approximations for the variance of the sojourn times.

Many distributed applications work as follows. The end user initiates a transaction request that is sent to a front-end server that parses the request and kicks off a server-side script. The script iteratively sends information-retrieval requests to different remote back-end systems and processes these pieces of information upon receipt before sending a response to the end user. The front-end application server processing is usually highly CPU-intensive processing steps, and therefore, is modeled as a Processor Sharing (PS) node; that is, when there are k parallel scripts running on the front-end server, then each script received a fraction $1/k$ of the available processing speed. In contrast, the information systems typically handle the queries in the order of arrival, and are multi-threaded, so that multiple transactions in parallel. Therefore, we model the information system in the back-end First-Come-First Served (FCFS) nodes with multiple servers.
Based on these assumptions, we study the following queueing network model with a single PS node (modeling a front-end application server) and $M \geq 1$ multi-server FCFS nodes (modeling multi-threaded information back-end systems). External customers arrive at the PS node according to a Poisson process. After departing from the PS node a customer proceeds to the k-th FCFS node with probability p_k, and with probability $1-p$, with $p = \sum_{k=1}^{M} p_k$ $(0 < p < 1)$ the customer departs from the system. After each visit to any FCFS node customers are fed back to the PS node. The service time are the PS node are generally distributed, and the service times at the FCFS nodes are exponentially distributed. This model is known to possess a product-form solution for the joint number of customers at the nodes. Hence, the mean sojourn time follows directly from Little's Law. The *variance* of the sojourn times are much more complicated, and no exact expressions can be obtained in the general setting of the model. For this reason in this paper we focus on the development and experimental validation of approximate closed-form expressions for the variance of the sojourn times.

The analysis of the variance of the sojourn time in the present model is complicated due to the fact that *overtaking* may occur, i.e., customers may bypass each other. Overtaking usually destroys any hope for an exact analysis of the higher moments of the sojourn-time distributions (see [2] for a survey of the available results on sojourn times in queueing networks). The main result in [2] is an expression for the Laplace-Stieltjes Transform (LST) of the joint probability distribution of the sojourn times at the nodes of a customer that traverses a predefined path of nodes in a product-form queueing network. Several results are known for single-node queueing systems with instantaneous feedback. For the M/G/1 queue with Bernoulli feedback, Doshi and Kaufmann [6] derive expressions for the LST of the joint distribution of the sojourn times of a customer at its successive passes through the system. Disney and Koenig [5] give an overview on Bernoulli feedback models. Van den Berg and Boxma [1] consider an M/G/1 system, with either FCFS or PS service, where a customer after receiving service for the i-th time is looped back into the system with probability q_i and departs from the system with probability $1-q_i$. For this model, Van den Berg and Boxma [1] analyse the joint distribution of the first i successive sojourn times of a customer (who is fed back at least $i-1$ times), and derive expressions for both the moments of these sojourn times and for the correlations between the successive sojourn times of an arbitrary customer in the system. Fewer results are known for sojourn time distributions for networks with delayed feedback, which occurs in the present model. Foley and Disney [7] study queueing systems with delayed feedback, but their focus is merely on queue length processes, busy period and several customer flow processes.

The results presented in this paper generalize those presented in [10], where we considered a network with a single FCFS node with a single server, and with exponentially distributed service times at both the PS and the FCFS

node, and derived approximate expressions for the variance of the sojourn times. In this context, the contribution of the present paper is two-fold. First, the model considered in this paper is much more generic in several respects (general number of FCFS nodes, multiple servers, and general service-time distributions at the PS node), and hence is much more interesting from an application point of view. Second, the analysis of a queueing network with multiple FCFS nodes introduces several interesting complications due to the impact of cross-correlations in the number of visits to each of the information systems. The queueing-theoretical contribution lies in the fact that this paper provides an effective means to deal with these cross-correlations. These observations make the added value of the current paper compared to [10] evident.

The remainder of this paper is organized as follows. In section 2 the model is described. In section 3 we present exact expressions for the mean sojourn times. In section 4 we develop an approximation for the variance of the sojourn times. In section 5 the accuracy of the approximations is tested by comparing the performance predictions based on the approximations with simulation results. Finally, in section 6 we address a number of challenging topics for further research.

2 Model

Consider an open queueing model with a single customer class, a PS node and $M \geq 1$ multi-server FCFS-nodes with $c_k \geq 1$ servers at FCFS node k $(k = 1, \ldots, M)$. Customers arrive from outside at the PS node according to a Poisson process with rate λ. After service completion at the PS node, the customer proceeds to the k-th FCFS node with probability p_k, and with probability $1-p$, with $p := \sum_{k=1}^{M} p_k$, the customer departs from the system. After receiving service at a FCFS node a customer is always fed back to the PS node. The service time at the PS node is a generally distributed random variable B_{ps} with finite first two moments β_{ps} and $\beta_{ps}^{(2)}$, respectively, and the service times at the FCFS node k are exponentially distributed with mean $\beta_{fcfs,k}$, $k = 1, \ldots, M$. The service times at all nodes are assumed to be mutually independent and independent of the state of the system. For an arbitrary customer denote by N, the random variables indicating the number of *returns* to the PS node, and by N_k the number of visits to k-th FCFS node, before departing from the system. Then clearly N is geometrically distributed with parameter p, i.e., $\text{Prob}\{N = n\} = (1-p)p^n$, for $n = 0, 1, \ldots$. Similarly, it is easily seen that $\text{Prob}\{N_k = i\} = (1-q_k)q_k^i$, for $i = 0, 1, \ldots$, with $q_k := p_k/(1-p+p_k)$. Notice that by definition $N := \sum_{k=1}^{M} N_k$, so that the random variables N and N_k $(k = 1, \ldots, M)$ are not mutually independent. Moreover, note that the total number of visits to the PS node before departing from the system is $N+1$. For notational convenience, define the joint probability distribution of $(N_1, \ldots, N_M)$ as follows: for $n_k = 0, 1, \ldots$, and $k = 1, \ldots, M$,

$$f(n_1, \ldots, n_M) := \text{Prob}\{N_1 = n_1, \ldots, N_M = n_M\}. \tag{1}$$

The load at the PS node and the FCFS nodes is given by

$$\rho_{ps} := \frac{\lambda\ \beta_{ps}}{1-p}, \quad \text{and} \quad \rho_{fcfs,k} := \frac{\lambda\ \beta_{fcfs,k}\ q_k}{c_k(1-q_k)} = \frac{\lambda\ \beta_{fcfs,k}\ p_k}{c_k(1-p)} \quad (k = 1, \ldots, M). \tag{2}$$

To ensure stability of the system it is assumed that $\rho_{ps},\ \rho_{fcfs,k} < 1$ $(k = 1, \ldots, M)$. For $i = 1, 2, \ldots, N+1$, let $S_i^{(ps)}$ denote the sojourn time of the i-th visit to the PS node, and for $j = 1, \ldots, N_k$, denote by $S_j^{(fcfs,k)}$ the duration of the j-th visit to the k-th FCFS node. The total sojourn time is then given by

$$S = \sum_{i=1}^{N+1} S_i^{(ps)} + \sum_{k=1}^{M} \sum_{j=1}^{N_k} S_j^{(fcfs,k)}. \tag{3}$$

3 Mean Sojourn Times

The queueing network model described in Section 2 is a product-form network. Defining L_{ps}, $L_{fcfs,k}$ to be the stationary number of customers at the PS node and at the k-th FCFS node, respectively, we have: For $l \geq 0, l_k \geq 0$ $(k = 1, \ldots, M)$,

$$\text{Prob}\{L_{ps} = l; L_{fcfs,1} = l_1, \ldots, L_{fcfs,M} = l_M\} \;=\; \text{Prob}\{L_{ps} = l\} \prod_{k=1}^{M} \text{Prob}\{L_{fcfs,k} = l_k\} \tag{4}$$

$$= (1 - \rho_{ps})\rho_{ps}^{l} \prod_{k=1}^{M} (1 - \rho_{fcfs,k})\rho_{fcfs,k}^{l_k}. \tag{5}$$

The successive sojourn times of a customer are generally not independent. Nonetheless, the successive sojourn times of a tagged customer at the same node are identically distributed:

Lemma 1
(a) *The successive sojourn times* $S_i^{(ps)}$ $(i = 1, \ldots, N+1)$ *are identically distributed.*
(b) *The successive sojourn times* $S_j^{(fcfs,k)}$ $(j = 1, \ldots, N_k)$, *are identically distributed for each* $k = 1, \ldots, M$.

Proof: We observe that the model under consideration is a multi-class product-form network, where the customer classes are defined as follows. Each customer enters the system (at the PS node) as a class-0 customer, and its class number is incremented from i to $i+1$ any time the customer jumps from one node to the next $(i = 0, 1, \ldots)$. (In this way, for each customer its class indicates the number of node visits since the arrival of the customer in the system.) Then according to the Arrival Theorem for multi-class product-form networks (cf., e.g., Walrand [11] (Theorem 4.4.1)) a jumping customer sees the system in steady state, *regardless* of its class number, which immediately implies the validity of Lemma 1. □

Using Lemma 1, it follows directly from equation (5) and Little's law that

$$E\left[L_{ps}\right] = \frac{\rho_{ps}}{1 - \rho_{ps}}, \tag{6}$$

and

$$E\left[S_i^{(ps)}\right] = \frac{\rho_{ps}}{\frac{\lambda}{(1-p)}(1 - \rho_{ps})} = \frac{\beta_{ps}}{1 - \rho_{ps}}, \quad i = 1, \ldots, N+1. \tag{7}$$

Recall that the total arrival intensity at the PS node equals $\lambda/(1-p)$. Moreover, it is readily verified that for the FCFS nodes we have, for $k = 1, \ldots, M$,

$$E\left[L_{fcfs,k}\right] = \frac{\rho_{fcfs,k}\pi_k}{1 - \rho_{fcfs,k}} + c_k\rho_{fcfs,k} \tag{8}$$

and

$$E\left[S_j^{(fcfs,k)}\right] = \frac{\beta_{fcfs,k}\pi_k}{(1 - \rho_{fcfs,k})c_k} + \beta_{fcfs,k}, \quad j = 1, \ldots, N_k. \tag{9}$$

Here, π_k stands for the probability that an customer arriving at FCFS node k can not be served immediately, and hence has to wait. From standard theory for the $M/M/c$ queue, we have, for $k = 1, \ldots, M$,

$$\pi_k = \frac{c_k^{c_k}\rho_{fcfs,k}^{c_k}}{c_k!}\left[1 + \sum_{n=1}^{c_k - 1}\left(1 - \frac{n}{c_k}\right)\frac{c_k^n\rho_{fcfs,k}^n}{n!}\right]^{-1}. \tag{10}$$

Note that for the special case $c_k = 1$ we have $\pi_k = \rho_{fcfs,k}$. Now, combining (3), (7) and (9) and applying Wald's equation we obtain the following expression for the mean total sojourn time of an arbitrary customer:

$$E[S] = E\left[\sum_{i=1}^{N+1} S_i^{(ps)} + \sum_{k=1}^{M}\sum_{j=1}^{N_k} S_j^{(fcfs,k)}\right] = (E[N]+1)E\left[S_1^{(ps)}\right] + \sum_{k=1}^{M} E[N_k]E\left[S_1^{(fcfs,k)}\right] \tag{11}$$

$$= \frac{1}{(1-p)}\frac{\beta_{ps}}{(1-\rho_{ps})} + \sum_{k=1}^{M}\frac{p_k}{(1-p)}\left[\frac{\beta_{fcfs,k}\pi_k}{(1-\rho_{fcfs,k})c_k} + \beta_{fcfs,k}\right]. \tag{12}$$

4 Variance of the sojourn times: approximations

Analysis of the variance of the total sojourn time, $Var[S]$, is fundamentally more complex than the analysis of the mean. The complexity is caused by the fact that *overtaking* may occur. Overtaking introduces correlation between sojourn times of job visits at the nodes in the queueing network. A formal definition of overtaking is presented by definition 2.2 in [2]. According to this definition the queueing network considered in this paper is not overtake-free. In the absence of exact expressions for the variance of the sojourn times we develop new, closed-form approximations for the variance of the sojourn times. To this end, in 4.2.1 $Var[S]$ is expressed in a convient form. In 4.2.2 we use this expression to derive an approximation for $Var[S]$ for the case of exponential service times at the PS node. Then, in 4.2.3 we extend these expressions for the case of non-exponential service times at the PS node. The accuracy of the results will be extensively studied in Section 5.

4.1 Preliminaries

First, we rewrite the sojourn time variance $Var[S]$ in the following convenient form:

$$Var[S] = Var\left[\sum_{i=1}^{N+1} S_i^{(ps)} + \sum_{k=1}^{M}\sum_{j=1}^{N_k} S_j^{(fcfs,k)}\right] \tag{13}$$

$$= E\left[Var\left[\sum_{i=1}^{N+1} S_i^{(ps)} + \sum_{k=1}^{M}\sum_{j=1}^{N_k} S_j^{(fcfs,k)} | N_1,\ldots,N_M\right]\right]$$

$$+Var\left[E\left[\sum_{i=1}^{N+1} S_i^{(ps)} + \sum_{k=1}^{M}\sum_{j=1}^{N_k} S_j^{(fcfs,k)} | N_1,\ldots,N_M\right]\right] \tag{14}$$

$$= \sum_{n_1=0}^{\infty}\cdots\sum_{n_M=0}^{\infty} Var\left[\sum_{i=1}^{n+1} S_i^{(ps)} + \sum_{k=1}^{M}\sum_{j=1}^{n_k} S_j^{(fcfs,k)}\right] f(n_1,\ldots,n_M)$$

$$+ Var\left[\sum_{i=1}^{N+1} E\left[S_i^{(ps)}\right] + \sum_{k=1}^{M}\sum_{j=1}^{N_k} E\left[S_j^{(fcfs,k)}\right]\right] \tag{15}$$

$$= \sum_{n=0}^{\infty} Var\left[\sum_{i=1}^{n+1} S_i^{(ps)}\right](1-p)p^n + \sum_{k=1}^{M}\sum_{N_k=0}^{\infty} Var\left[\sum_{j=1}^{n_k} S_j^{(fcfs,k)}\right](1-q_k){q_k}^{n_k}$$

$$+2\sum_{k=1}^{M}\sum_{n_1=0}^{\infty}\cdots\sum_{n_M=0}^{\infty} Cov\left[\sum_{i=1}^{n+1} S_i^{(ps)}, \sum_{j=1}^{n_k} S_j^{(fcfs,k)}\right] f(n_1,\ldots,n_M)$$

$$+\sum_{k\neq m}\sum_{n_1=0}^{\infty}\cdots\sum_{n_M=0}^{\infty} Cov\left[\sum_{j=1}^{n_k} S_j^{(fcfs,k)}, \sum_{j=1}^{n_m} S_j^{(fcfs,m)}\right] f(n_1,\ldots,n_M)$$

$$+Var\left[\sum_{i=1}^{N+1} E\left[S_i^{(ps)}\right] + \sum_{k=1}^{M}\sum_{j=1}^{N_k} E\left[S_j^{(fcfs,k)}\right]\right] \tag{16}$$

$$= \sum_{n=0}^{\infty}(n+1)Var\left[S_1^{(ps)}\right](1-p)p^n + \sum_{k=1}^{M}\sum_{n_k=0}^{\infty} n_k Var\left[S_1^{(fcfs,k)}\right](1-q_k){q_k}^{n_k}$$

$$+\sum_{n=0}^{\infty}\sum_{i\neq j} Cov\left[S_i^{(ps)}, S_j^{(ps)}\right](1-p)p^n$$

$$+\sum_{k=1}^{M}\sum_{n_k=0}^{\infty}\sum_{i\neq j} Cov\left[S_i^{(fcfs,k)}, S_j^{(fcfs,k)}\right](1-q_k){q_k}^{n_k}$$

$$+2\sum_{k=1}^{M}\sum_{n_1=0}^{\infty}\cdots\sum_{n_M=0}^{\infty} Cov\left[\sum_{i=1}^{n+1} S_i^{(ps)}, \sum_{j=1}^{n_k} S_j^{(fcfs,k)}\right] f(n_1,\ldots,n_M)$$

$$+\sum_{k\neq m}\sum_{n_1=0}^{\infty}\cdots\sum_{n_M=0}^{\infty} Cov\left[\sum_{j=1}^{n_k} S_j^{(fcfs,k)}, \sum_{j=1}^{n_m} S_j^{(fcfs,m)}\right] f(n_1,\ldots,n_M)$$

$$+Var\left[\sum_{i=1}^{N+1} E\left[S_i^{(ps)}\right] + \sum_{k=1}^{M}\sum_{j=1}^{N_k} E\left[S_j^{(fcfs,k)}\right]\right]. \tag{17}$$

Equation (13) follows from definition (3), and (14) follows directly from the classical $Var[U] = E[Var[U|V]] + Var[E[U|V]]$ by taking $U := \sum_{i=1}^{N+1} S_i^{(ps)} + \sum_{k=1}^{M}\sum_{j=1}^{N_k} S_j^{(fcfs,k)}$ and $V := \{N_1 = n_1, \ldots, N_M = n_M\}$. Equation (15) is then obtained by conditioning with respect to the event V. Subsequently, (16) follows from Lemma 1 and classical rules for the variance of random variables, and (17) is obtained from Lemma 1.

The quantities $E\left[S_i^{(ps)}\right]$ and $E\left[S_j^{(fcfs,k)}\right]$ in (17), which are independent of i and j, respectively, are given by equations (7) and (9). Since we are considering a product form network, the distribution function of the number of customers in each of the queues is known. Hence, the sojourn time distribution for each of the FCFS nodes is also known. Specifically, sojourn times at FCFS queue k behaves as an $M/M/c_k$-FCFS queue. Thus, for the sojourn time variance at the FCFS nodes we have (cf. [4]):

$$Var\left[S_1^{(fcfs,k)}\right] = \beta_{fcfs,k}^2 + \frac{\pi_k(2-\pi_k)\beta_{fcfs,k}^2}{c_k^2(1-\rho_{fcfs,k})^2}. \tag{18}$$

Further, given Lemma 1 and the fact that the number of visits to the PS node always equals the number of visits to the FCFS nodes plus one (see also section 2) we have after some standard variance calculus:

$$Var\left[\sum_{i=1}^{N+1} E\left[S_i^{(ps)}\right] + \sum_{k=1}^{M}\sum_{j=1}^{N_k} E\left[S_j^{(fcfs,k)}\right]\right] = Var\left[\sum_{k=1}^{M} N_k\left(E\left[S_1^{(ps)}\right] + E\left[S_1^{(fcfs,k)}\right]\right)\right]$$

$$= \sum_{k=1}^{M} Var\left[N_k\right]\left(E\left[S_1^{(ps)}\right] + E\left[S_1^{(fcfs,k)}\right]\right)^2$$

$$+ \sum_{k\neq m} Cov\left[N_k, N_m\right]\left(E\left[S_1^{(ps)}\right] + E\left[S_1^{(fcfs,k)}\right]\right)\left(E\left[S_1^{(ps)}\right] + E\left[S_1^{(fcfs,m)}\right]\right). \tag{19}$$

Since N_k is geometrically distributed with parameter q_k, we know that $Var[N_k] = \frac{q_k}{(1-q_k)^2}$. Further, we can express the sum of covariances between N_k and N_m in terms of $Var[N]$ and $\sum_{k=1}^{M} Var[N_k]$: $Var[N] = \sum_{k=1}^{M} Var[N_k] + \sum_{k \neq m} Cov[N_k, N_m]$. Then, with some calculus we find that $\sum_{k \neq m} Cov[N_k, N_m] = \sum_{k \neq m} \frac{p_k p_m}{(1-p)^2}$. Substituting these expressions in equation (19), we obtain:

$$Var\left[\sum_{i=1}^{N+1} E\left[S_i^{(ps)}\right] + \sum_{k=1}^{M}\sum_{j=1}^{N_k} E\left[S_j^{(fcfs,k)}\right]\right] = \sum_{k=1}^{M} \frac{q_k}{(1-q_k)^2}\left(E\left[S_1^{(ps)}\right] + E\left[S_1^{(fcfs,k)}\right]\right)^2 + \sum_{k \neq m} \frac{p_k p_m}{(1-p)^2}\left(E\left[S_1^{(ps)}\right] + E\left[S_1^{(fcfs,k)}\right]\right)\left(E\left[S_1^{(ps)}\right] + E\left[S_1^{(fcfs,m)}\right]\right). \quad (20)$$

Hence, it remains to develop approximations for $Var\left[S_1^{(ps)}\right]$, $Cov\left[S_i^{(ps)}, S_j^{(ps)}\right]$, $Cov\left[S_i^{(fcfs,k)}, S_j^{(fcfs,k)}\right]$, for any $i \neq j$, $k = 1, \ldots, M$ and $Cov\left[S_i^{(fcfs,k)}, S_j^{(fcfs,m)}\right]$, $Cov\left[S_i^{(ps)}, S_j^{(fcfs,k)}\right]$, for any $i, j = 1, 2, \ldots$, $k \neq m$.

4.2 The case of exponential service times at the PS node

To start, we use the following approximation assumption.

Approximation Assumption 1 (AA1)
The total arrival process at PS node is a Poisson process with rate $\lambda/(1-p)$.

In general it is known that Approximation Assumption 1 is not true for non-acyclic queueing networks, not even under the assumption that the service times are exponentially distributed. The violation of the Poisson assumption is caused by the feedback loop, implying dependent interarrival times at the nodes. Based on Approximation Assumption 1, we obtain the following approximate expression for the variance of the sojourn times at the PS node (cf. [9]):

$$Var\left[S_1^{(ps)}\right] \approx \frac{2+\rho_{ps}}{2-\rho_{ps}}\left(\frac{\beta_{ps}}{1-\rho_{ps}}\right)^2. \quad (21)$$

Van den Berg and Boxma [1] derive exact expressions for the covariance of the successive sojourn times for single-server FCFS and for PS queues with *direct feedback*, where customers upon receiving service are immediately fed back into the system (with some probability). We emphasize that the model discussed in section 2 implements a *delayed feedback* mechanism: upon departing from the PS node, a customer is first processed by a FCFS-node (if not leaving the system immediately) before returning to the PS node. Similarly, after leaving any FCFS node, a customer is first processed at least once by the PS node before returning to the FCFS node.

Approximation Assumption 2 (AA2)
(a) *The covariance of the successive sojourn times of a customer at the PS node in the network with delayed feedback may be approximated by those in a single* $M/M/1$ *PS node with direct feedback.*
(b) *The covariance of the successive sojourn times of a customer FCFS node* k *in the network with delayed feedback may be approximated by those in a single* $M/M/c_k$ *FCFS node with direct feedback* $(k = 1, \ldots, M)$.

Now, based on Approximation Assumption 2 we approximate the covariances between the successive sojourn times at the same node (i.e., $Cov\left[S_i^{(ps)}, S_j^{(ps)}\right]$ and $Cov\left[S_i^{(fcfs,k)}, S_j^{(fcfs,k)}\right]$, for $i \neq j$ and any k). It is easily verified that by using the exact results for single-server FCFS systems with *direct* feedback, derived from equations (9.13) and (3.17)

in [1], and conditioning on the event that a customer arriving at FCFS node k has to wait, we obtain the following approximations: for $1 \leq i < n, 1 \leq j \leq n-i$ and $1 \leq k \leq M$,

$$Cov\left[S_i^{(fcfs,k)}, S_{i+j}^{(fcfs,k)}\right] \approx \pi_k(\rho_{fcfs,k}(1-q_k)+q_k)^{j-1}\pi_k\left(\beta_{fcfs,k}^2 + \pi_k(2-\pi_k)\frac{\beta_k^2}{c_k^2(1-\rho_{fcfs,k})^2}\right), \tag{22}$$

and similarly, for $1 \leq i < n+1, 1 \leq j \leq n+1-i$,

$$Cov\left[S_i^{(ps)}, S_{i+j}^{(ps)}\right] \approx \frac{\rho_{ps}\beta_{ps}^2}{(1-\rho_{ps})^2(2-\rho_{ps}-p+\rho_{ps}p)^{j+1}}. \tag{23}$$

Approximation Assumption 3 (AA3)
The sojourn times $S_i^{(ps)}$ and $S_j^{(fcfs,k)}$ are uncorrelated; for $i = 1, \ldots, N+1$, $j = 1, \ldots, N$ and $k = 1, \ldots, M$:

$$Cov\left[S_i^{(ps)}, S_j^{(fcfs,k)}\right] \approx 0. \tag{24}$$

In general, Approximation Assumption 3 is known to be not true. However, the product-form solution for the present model, see (5), implies that the *number of customers* at both nodes *are* independent in equilibrium. Also, the sojourn time at the FCFS queues is closely related to the number of customers at that node: if a customer finds $n_{fcfs,k}$ customers at the k-th FCFS node upon arrival, then the sojourn time simply consists of $n_{fcfs,k}+1$ independent successive exponential phases each with rate $1/\beta_{fcfs,k}$, which results in an Erlang distribution with shape parameter $n_{fcfs,k}+1$ and rate parameter $1/\beta_{fcfs,k}$. For the PS node, the correlation between the sojourn times and number of customers present upon arrival is less clear, and intuitively seems to be weaker than for FCFS nodes. These observations suggest that the cross-correlation terms are rather small. In our previous work for a queueing network with only one FCFS node we performed a variety of simulation experiments to validate this conjecture, and we found that the cross-correlation coefficients (between PS and FCFS nodes) were about a factor two smaller than the correlation coefficient for successive sojourn times at the PS node. Also we found that the correlation coefficient for sojourn times at the FCFS node were about three times larger than the PS node correlation coefficient. These results confirm the conjecture that the cross-correlation terms for the sojourn times of visits to *different* nodes are *indeed negligible* compared to the correlation terms of successive visits to the *same* node. For queueing networks with multiple FCFS nodes the impact of ignoring cross-correlations on the approximation accuracy is even less, as the numerical results in section 5 will demonstrate.

Finally, substituting the exact formula for the variance of the sojourn time in the FCFS nodes and approximations (20)-(24) in the expression for $Var[S]$ in (17) we obtain the following approximation for the variance of the sojourn time for the case of exponential service times at the PS node:

$$\begin{aligned} Var_{exp}[S] \;\approx\; & \frac{1}{1-p}\,\frac{2+\rho_{ps}}{2-\rho_{ps}}\left(\frac{\beta_{ps}}{1-\rho_{ps}}\right)^2 \\ & + \frac{2p\rho_{ps}\beta_{ps}^2}{(2-\rho_{ps}-p+p\rho_{ps})(1-p)^2(2-\rho_{ps})(1-\rho_{ps})^2} \\ & + \sum_{k=1}^{M}\frac{q_k}{1-q_k}\left(\beta_{fcfs,k}^2 + \frac{\pi_k(2-\pi_k)\beta_{fcfs,k}^2}{c_k^2(1-\rho_{fcfs,k})^2}\right) \\ & + \sum_{k=1}^{M}\frac{2{q_k}^2\pi_k\beta_{fcfs,k}^2\left((1-\rho_{fcfs,k})^2c_k^2+\pi_k(2-\pi_k)\right)}{(1-q_k)^2(1-q_k\rho_{fcfs,k}+q_k)(1-\rho_{fcfs,k})^2c_k^2} \end{aligned} \tag{25}$$

$$+ \quad \sum_{k=1}^{M} \frac{q_k}{(1-q_k)^2} \left(\frac{\beta_{ps}}{1-\rho_{ps}} + \beta_{fcfs,k} + \frac{\pi_k \beta_{fcfs,k}}{c_k(1-\rho_{fcfs,k})} \right)^2$$

$$+ \quad \sum_{k \neq m} \frac{p_k p_m}{(1-p)^2} \left(\frac{\beta_{ps}}{1-\rho_{ps}} + \beta_{fcfs,k} + \frac{\pi_k \beta_{fcfs,k}}{c_k(1-\rho_{fcfs,k})} \right) \left(\frac{\beta_{ps}}{1-\rho_{ps}} + \beta_{fcfs,m} + \frac{\pi_m \beta_{fcfs,m}}{c_m(1-\rho_{fcfs,m})} \right).$$

In the next subsection we will extend the approximations to the case of general service times at the PS node.

4.3 The case of general service times at the PS node

For the case of general service times at the PS node we adopt the assumptions AA1, AA2 and AA3. Based on AA1, we can approximate $Var\left[S_1^{(ps)}\right]$ by the variance of the sojourn time in an $M/G/1$-PS system with arrival rate $\lambda/(1-p)$ and service-time distribution B_{ps}. Van den Berg and Boxma [1] propose the following simple approximation for the second moment of the sojourn time $S_{M/G/1}$ in an $M/G/1$-PS with mean service time $\beta_{M/G/1}$, load ρ and squared coefficient of variation $c^2_{M/G/1}$, which is a linear interpolation between the cases of exponential and deterministic service times, respectively:

$$E_{app}\left[S_{M/G/1}\right]^2 = c^2_{M/G/1}\left(1+\frac{2+\rho}{2-\rho}\right)\frac{\beta^2_{M/G/1}}{(1-\rho)^2} + (1-c^2_{M/G/1})\left(\frac{2\beta^2_{M/G/1}}{(1-\rho)^2} - \frac{2\beta^2_{M/G/1}}{\rho^2(1-\rho)}(e^{\rho}-1-\rho)\right). \tag{26}$$

Using this expression, we approximate the second moment of $S_1^{(ps)}$ for an arbitrary visit of a customer to the PS by

$$E\left[S_1^{(ps)}\right]^2 \approx c^2_{ps}\left(1+\frac{2+\rho_{ps}}{2-\rho_{ps}}\right)\left(\frac{\beta_{ps}}{1-\rho_{ps}}\right) + (1-c^2_{ps})\left(\frac{2\beta^2_{ps}}{(1-\rho_{ps})^2} - \frac{2\beta^2_{ps}}{\rho^2_{ps}(1-\rho_{ps})}(e^{\rho_{ps}}-1-\rho_{ps})\right), \tag{27}$$

and hence,

$$Var\left[S_1^{(ps)}\right] \approx E\left[S_1^{(ps)}\right]^2 - \left(\frac{\beta_{ps}}{1-\rho_{ps}}\right)^2, \tag{28}$$

where c^2_{ps} is the squared coefficient of variation of the service times at the PS node.

In general, the variance of the sojourn times in an $M/G/1$-PS system depends on the third moment of the service-time distribution at that node, whereas the simple approximation in (26) only depends on the first two moments of the service-time distribution at the PS node. For sake of simplicity, we adopt (26) in our approximations. We refer to [1] for a discussion on the refined approximations that do take into account third moments of the service times.
In the adoptation of AA2(a) we assume that the covariance of the successive sojourn times at the PS node with delayed feedback is still approximated by a single $M/M/1$-PS node, rather than a $M/G/1$-PS node. The reason for this is that although an exact analysis of the covariances of the sucessive sojourn times for the $M/G/1$-PS feedback system is possible, the expressions are not explicit (see [1] for details) and require the solution of a non-linear set of equations. Since our goal is to develop closed-form approximations for the variance of the sojourn times we adopt approximation (23).

Finally, these observations lead to the following expression for $Var[S]$ for the case of general service times at the PS node:

$$Var_{gen}[S] \quad \approx \quad \frac{1}{1-p}\left\{ c^2_{ps}\left(1+\frac{2+\rho_{ps}}{2-\rho_{ps}}\right)\frac{\beta_{ps}}{(1-\rho_{ps})^2} + (1-c^2_{ps})\left(\frac{2\beta^2_{ps}}{(1-\rho_{ps})^2} - \frac{2\beta^2_{ps}}{\rho^2_{ps}(1-\rho_{ps})}(e^{\rho_{ps}}-1-\rho_{ps})\right)\right\}$$

$$
\begin{aligned}
&- \quad \frac{1}{1-p}\left(\frac{\beta_{ps}}{1-\rho_{ps}}\right)^2 \\
&+ \quad \frac{2p\rho_{ps}\beta_{ps}^2}{(2-\rho_{ps}-p+p\rho_{ps})(1-p)^2(2-\rho_{ps})(1-\rho_{ps})^2} \\
&+ \quad \sum_{k=1}^{M}\frac{q_k}{1-q_k}\left(\beta_{fcfs,k}^2+\frac{\pi_k(2-\pi_k)\beta_{fcfs,k}^2}{c_k^2(1-\rho_{fcfs,k})^2}\right) \\
&+ \quad \sum_{k=1}^{M}\frac{2{q_k}^2\pi_k\beta_{fcfs,k}^2\left((1-\rho_{fcfs,k})^2c_k^2+\pi_k(2-\pi_k)\right)}{(1-q_k)^2(1-q_k\rho_{fcfs,k}+q_k)(1-\rho_{fcfs,k})^2c_k^2} \\
&+ \quad \sum_{k=1}^{M}\frac{q_k}{(1-q_k)^2}\left(\frac{\beta_{ps}}{1-\rho_{ps}}+\beta_{fcfs,k}+\frac{\pi_k\beta_{fcfs,k}}{c_k(1-\rho_{fcfs,k})}\right)^2 \\
&+ \quad \sum_{k\neq m}\frac{p_kp_m}{(1-p)^2}\left(\frac{\beta_{ps}}{1-\rho_{ps}}+\beta_{fcfs,k}+\frac{\pi_k\beta_{fcfs,k}}{c_k(1-\rho_{fcfs,k})}\right)\left(\frac{\beta_{ps}}{1-\rho_{ps}}+\beta_{fcfs,m}+\frac{\pi_m\beta_{fcfs,m}}{c_m(1-\rho_{fcfs,m})}\right).
\end{aligned}
\tag{29}
$$

In the next section we assess the accuracy of the approximations.

5 Numerical Results

To validate the accuracy of the approximations for the variance of the sojourn times proposed in Section 4, we have performed extensive numerical experiments, comparing the approximations with simulations. To this end, we have checked the accuracy of approximations for many parameter combinations, by varying the arrival rate, the service-times distributions, the asymmetry in the loads of the nodes, the numbers of servers at the FCFS nodes, and the values of the routing probabilities p_k. From the simulations, we have calculated the point estimates for the variance of the sojourn times, and 95% confidence intervals (C.I.'s). We calculated the confidence intervals for the sojourn time variance using the Jackknife method (see [8]). For each parameter case we ran 10 simulation runs. The runs lengths were taken long enough to ensure that all the confidence intervals were at most 15% of the point estimator value. In the tables below we present results for a subset of the parameter cases that we validated. Denoting the point estimations based on simulations by "simulation", and the approximated values by "approx", the relative error of the approximations is defined as

$$
\Delta\% \quad = \frac{approx - simulation}{simulation} * 100\%. \tag{30}
$$

The results of the validation experiments will be discussed below. In section 5.1 we give the results for the case of exponential service times at the PS node. In section 5.2 we discuss the results for non-exponential service times at the PS node.

5.1 Exponential service times at PS node

To assess the accuracy of the approximations developed in Section 4, one might question whether including covariance terms in the approximation (i.e., the last summation in (25)), which make the approximation slightly more complex, indeed lead to a higher level of accuracy. To illustrate the 'added value' of including covariance terms in the approximation we also compare it to a simple, straightforward approximation, which completely *ignores* dependencies between successive sojourn times of a tagged customer in the PS or FCFS nodes. In particular, the simple approximation is the same as approximation (25) without the covariance terms, resulting in the expression:

$$Var_{simple}[S] \approx \frac{1}{1-p}\,\frac{2+\rho_{ps}}{2-\rho_{ps}}\left(\frac{\beta_{ps}}{1-\rho_{ps}}\right)^2 + \sum_{k=1}^{M}\frac{q_k}{1-q_k}\left(\beta_{fcfs,k}^2 + \frac{\pi_k(2-\pi_k)\beta_{fcfs,k}^2}{c_k^2(1-\rho_{fcfs,k})^2}\right)$$
$$+ \sum_{k=1}^{M}\frac{q_k}{(1-q_k)^2}\left(\frac{\beta_{ps}}{1-\rho_{ps}} + \beta_{fcfs,k} + \frac{\pi_k\beta_{fcfs,k}}{c_k(1-\rho_{fcfs,k})}\right)^2.$$

Throughout we will denote the results of this simple approximation by "simple".

5.1.1 Single server at FCFS nodes

Let us first consider the accuracy of the approximations for models with single-server FCFS nodes, i.e., $c_1 = \cdots = c_M = 1$. To start, consider the model with $M = 2$ identical FCFS nodes (i.e. $\beta_{fcfs,1} = \beta_{fcfs,2} =: \beta_{fcfs}$), $\lambda = 1$ and where the routing probabilities to the FCFS nodes are $p_1 = p_2 = \frac{p}{2}$. Table 1 shows the simulated and approximated value of the $Var[S]$ for various combinations of β_{ps} and β_{fcfs}. Note that in these symmetric models the load values of the FCFS nodes are the same, i.e. $\rho_{fcfs,1} = \rho_{fcfs,2} =: \rho_{fcfs}$. The results in Table 1 show that the approximations in

p	β_{ps}	β_{fcfs}	ρ_{ps}	ρ_{fcfs}	simulation	95% C.I.	approx	Δ%	simple	Δ%
0.2	0.4	1.6	0.5	0.2	4.86	(4.84, 4.89)	4.92	1.2	4.54	-6.7
0.2	0.4	4	0.5	0.5	43.50	(42.89, 44.10)	43.52	0.1	39.11	-10
0.2	0.4	6.4	0.5	0.8	642.5	(631.7, 653.4)	643.7	0.2	559.9	-13
0.5	0.1	0.4	0.2	0.2	1.11	(1.10, 1.12)	1.11	0.2	0.87	-21
0.5	0.25	1	0.5	0.5	19.19	(18.94, 19.44)	19.31	0.7	14.21	-26
0.5	0.4	0.4	0.8	0.2	40.45	(38.91, 42.00)	41.15	1.7	28.29	-30
0.5	0.1	1.6	0.2	0.8	247.3	(234.3, 260.2)	244.1	-1.3	163.1	-34
0.5	0.4	1.6	0.8	0.8	338.3	(326.8, 349.9)	340.4	0.6	232.7	-31
0.8	0.1	0.1	0.5	0.2	3.01	(2.99, 3.03)	3.03	0.5	1.66	-45
0.8	0.1	0.25	0.5	0.5	13.1	(13.1, 13.2)	13.12	-0.1	7.21	-45
0.8	0.1	0.4	0.5	0.8	159.0	(153.3, 164.7)	158.8	-0.1	74.41	-53

Table 1: Sojourn time variance with two identical FCFS nodes: approximations versus simulations.

(25) are extremely accurate. The relative error of the approximation does not exceed 2%. As expected the "simple" approximation consistently and strongly underestimates the variance of the total sojourn time; it appears to be an inaccurate lower bound. The relative error of the "simple" approximation becomes higher for higher p when load is fixed. When p is increased, the expected number of times a job will be fed back grows. When p is increased, the correlation between the successive sojourn times of a job tends to increase. Since this rough approximation omits the covariance of successive sojourn times of a job, the accuracy of the "simple" approximation degrades as p increases. Also, the relative error of the "simple" approximation becomes higher for higher load when p is fixed. When the load increases, the covariance of successive sojourn times of a job grows faster than the variance. As a result, the covariance part in the total variance will increase more than proportionally in comparison with the variance for increasing load. Because the covariance is not taken into account in the "simple" approximation, the relative error grows. These observations also hold for the other cases. Therefore, we conclude that the "simple" approximation is too inaccurate

and the additional complexity of the approximation presented in section 3 (due to inclusion of the covariance terms) is justified, for increasing the approximation accuracy.
To summarize, Table 1 shows that the approximation works very well in these symmetric cases. Also the errors are positive for some cases and negative for other cases, and all lie in the 95% confidence interval.

To investigate the impact of asymmetry in the number of visits per FCFS node on the accuracy of the approximations, we have also considered a variety of parameter combinations with unequal visits per node. In this second case the loads of both FCFS nodes are still equal, but the probabilities of a visit to each FCFS nodes are not equal (and thus the service times are unequal too). This represents, for example, a database system that authenticates a request (visit to FCFS node 2) and then retrieves information from a database (visit to FCFS node 1) once or several times. We have chosen the relative distribution of visits to each of the FCFS nodes as: $p_1 = \frac{3}{4}p, p_2 = \frac{1}{4}p$. Table 2 presents the results for case 2. In the third case we consider an even more asymmetric network scenario, where the

p	β_{ps}	$\beta_{fcfs,1}$	$\beta_{fcfs,2}$	ρ_{ps}	ρ_{fcfs}	simulation	95% C.I.	approx	Δ%	Δ_{simple}%
0.2	0.4	1.05	3.15	0.5	0.20	5.46	(5.40, 5.51)	5.46	0.0	-8.2
0.2	0.4	2.7	8.1	0.5	0.51	56.85	(55.08, 58.62)	56.81	-0.1	-14.1
0.2	0.4	4.3	12.9	0.5	0.81	899.53	(860.92, 938.14)	881.07	-2.1	-12.2
0.5	0.1	0.3	0.9	0.2	0.23	1.66	(1.65, 1.67)	1.66	0.0	-23.3
0.5	0.25	0.6	1.8	0.5	0.45	15.87	(15.71, 16.02)	15.72	-0.9	-30.4
0.5	0.4	0.3	0.9	0.8	0.23	42.72	(41.34, 44.09)	42.31	-1.0	-29.1
0.5	0.1	1.05	3.15	0.2	0.79	249.29	(238.68, 259.91)	245.78	-1.4	-27.8
0.5	0.4	1.05	3.15	0.8	0.79	337.35	(320.88, 353.82)	337.66	0.1	-18.8
0.8	0.1	0.07	0.21	0.5	0.21	3.16	(3.12, 3.19)	3.19	1.2	-42.2
0.8	0.1	0.17	0.51	0.5	0.51	14.61	(14.49, 14.74)	14.7	0.6	-48.7
0.8	0.1	0.27	0.81	0.5	0.81	186.98	(181.18, 192.79)	192.49	2.9	41.0

Table 2: Sojourn time variance with two identically loaded FCFS nodes, but with different service times: approximations versus simulations.

number of visits to the FCFS nodes *and* the loads of the FCFS nodes are taken asymmetric. The results for this case are presented in Table 3. The relative distribution of visits to each of the FCFS nodes remains: $p_1 = \frac{3}{4}p, p_2 = \frac{1}{4}p$. The results in Tables 2 and 3 demonstrate that for asymmetric cases the relative error is still very low, smaller than 3% and all approximation results are within the confidence intervals. Again, the estimation is sometimes higher and sometimes lower than the centre of the confidence interval. It does not seem to make any difference whether the loads of the nodes are very different, e.g. 0.20 – 0.80, or close to each other. Observing that the relative errors are very low, we impute the difference in sign to the randomness of the simulation. Asymmetric loads do not cause the approximation to perform significantly worse. This could be expected, as the approximation contains separate covariance terms for the PS-node and the FCFS-nodes. Consequently the formulas can adapt to asymmetric loads. Again, the accuracy is at least an order of magnitude better than in the case of the "simple" approximation.

To validate the approximation for a network with more than two FCFS nodes, we also include the results for a case with five FCFS nodes in Table 4. For notational convenience, define $\underline{\beta}_{fcfs} := (\beta_{fcfs,1}, \ldots, \beta_{fcfs,N})$, and $\underline{\rho}_{fcfs} := (\rho_{fcfs,1}, \ldots, \rho_{fcfs,N}.)$. As expected, the "simple" approximation still underestimates the variance of the total

p	β_{ps}	$\beta_{fcfs,1}$	$\beta_{fcfs,2}$	ρ_{ps}	$\rho_{fcfs,1}$	$\rho_{fcfs,2}$	simulation	approx	Δ%	Δ_{simple}%
0.2	0.4	2.65	8	0.5	0.5	0.5	53.48	53.83	0.6	-7.5
0.2	0.4	1.1	12.8	0.5	0.21	0.8	579.53	566.34	-2.3	-7.2
0.2	0.4	4.3	3.2	0.5	0.81	0.2	250.43	243.74	-2.7	-14.6
0.5	0.25	0.67	2	0.5	0.5	0.5	21.88	22.14	1.2	-22.3
0.5	0.25	0.27	3.2	0.5	0.2	0.8	186.15	181.01	-2.8	-18.9
0.5	0.25	1.07	0.8	0.5	0.8	0.2	97.99	97.81	-0.2	-28.8
0.8	0.04	0.17	0.51	0.2	0.51	0.51	10.00	10.03	0.3	-42.7
0.8	0.04	0.07	0.8	0.2	0.21	0.8	74.27	75.74	2.0	-33.9
0.8	0.04	0.27	0.2	0.2	0.81	0.2	59.3	60.73	2.4	-45.6

Table 3: Sojourn time variance with two asymmetrically loaded FCFS nodes: approximations versus simulations.

p_1	p_2	p_3	p_4	p_5	ρ_{ps}	$\underline{\rho}_{fcfs}$					sim	approx	Δ%	Δ_{simple}%
0.1	0.1	0.1	0.1	0.1	0.4	0.2	0.2	0.2	0.2	0.2	7.09	7.11	0.4	-30.8
0.1	0.1	0.1	0.1	0.1	0.4	0.8	0.8	0.8	0.8	0.8	1344.2	1352.8	0.6	-33.3
0.1	0.1	0.1	0.1	0.1	0.4	0.8	0.64	0.48	0.32	0.16	297.8	297.3	-0.2	-23.0
0.1	0.1	0.1	0.1	0.1	0.4	0.9	0.2	0.2	0.2	0.2	1081.5	1070.9	-1.0	-16.6
0.4	0.1	0.1	0.1	0.1	0.5	0.8	0.2	0.2	0.2	0.2	74.6	77.4	3.7	-47.3
0.4	0.1	0.1	0.1	0.1	0.5	0.8	0.8	0.8	0.8	0.8	823.5	837.7	1.7	-55.4
0.3	0.2	0.15	0.1	0.05	0.5	0.8	0.53	0.4	0.27	0.13	98.3	100.8	2.5	-51.6
0.3	0.2	0.15	0.1	0.05	0.5	0.6	0.6	0.6	0.6	0.6	124.6	126.1	1.2	-53.6

Table 4: Sojourn time variance with five FCFS nodes: approximations versus simulations.

sojourn time in this case with five FCFS nodes. As can be seen the more advanced approximation behaves very well in systems with five FCFS nodes. The relative error is not larger than 4%. We expect that the approximation will also behave well in systems with another number of FCFS nodes, because a larger number of FCFS nodes will reduce the cross-correlations and the correlations between subsequent visits to each FCFS node. This conjecture is supported by our efforts to find a worst-case scenario for the approximation. In fact, we did not find any parameter scenario where the approximation was less accurate than the worst-case scenario that we found in our previous work, with a feedback network with only one FCFS node. The worst-case scenario presented in our previous work was a pathological scenario in which the arrival processes at the PS node and FCFS node are highly non-Poisson by taking the external arrival rate λ close to 0 and p close to 1. We demonstrated that the approximation tends to become less accurate when rate λ very low and p very high, but in several cases still acceptable. However, the cases for which the approximation becomes poor are quite pathological and less relevant from a practical point of view.

5.1.2 Multiple servers at FCFS nodes

To check the accuracy of the approximations for models with multiple servers at the FCFS nodes, we first consider the following symmetric model with multiple servers at the FCFS nodes: $\lambda = 1$, $M = 3$, $c_1 = c_2 = c_3 =: c$, $p_1 = p_2 = p_3 = 0.3$, $\beta_{fcfs,1} = \beta_{fcfs,2} = \beta_{fcfs,3} =: \beta_{fcfs}$. Note that $\rho_{fcfs,1} = \rho_{fcfs,2} = \rho_{fcfs,3} =: \rho_{fcfs}$. Table 5 shows the results for a variety of combinations of β_{PS}, β_{fcfs}, c, ρ_{PS} and ρ_{fcfs}. The parameters have been varied in such a way that the approximations are tested for a broad range of load combinations of the PS node and the FCFS nodes. The results in Table 5 show that our approximation also works very well for models with multi-server

β_{PS}	β_{fcfs}	c	ρ_{PS}	ρ_{fcfs}	sim	approx	Δ%	simple	Δ_{simple}%
0.02	0.33	2	0.20	0.50	22.90	23.67	3.35	9.50	-58.4
0.02	0.53	2	0.20	0.50	271.43	269.03	-0.88	93.56	-65.5
0.05	0.13	2	0.50	0.20	5.88	6.08	3.53	2.39	-59.3
0.05	0.53	2	0.50	0.80	306.54	306.97	0.14	107.43	-64.9
0.08	0.13	2	0.80	0.20	47.85	48.90	2.20	14.3	-70.0
0.08	0.33	2	0.80	0.50	87.11	90.55	3.95	30.96	-64.5
0.01	0.33	4	0.10	0.25	11.85	11.90	0.46	5.33	-55.0
0.01	1.20	4	0.10	0.90	1680.76	1637.24	-2.59	550.25	-67.3
0.03	0.13	4	0.25	0.90	2.67	2.70	1.18	1.17	-56.0
0.08	1.20	4	0.75	0.90	1829.87	1842.08	0.67	629.58	-65.6
0.09	0.13	4	0.90	0.75	230.90	235.73	2.09	59.96	-74.0
0.09	1.00	4	0.90	0.75	592.27	718.54	3.79	246.18	-64.4
0.02	1.67	10	0.20	0.50	286.38	290.27	1.36	129.57	-54.8
0.05	0.67	10	0.50	0.20	56.97	57.64	1.17	25.33	-55.5
0.08	0.67	10	0.80	0.20	123.59	128.96	4.34	48.69	-60.6
0.08	1.67	10	0.80	0.50	430.20	440.47	2.39	184.36	-57.2

Table 5: Sojourn time variance with three symmetric multi-server FCFS nodes: approximations versus simulations.

FCFS nodes, and show that the simple approximation is strongly outperformed, reducing the error by two orders of magnitude.

Next, we consider the accuracy of the approximations for asymmetric models with multi-server FCFS nodes. The results are shown in Tables 6 and 7. Table 6 shows the results for the model with $\lambda = 1$ and $c_1 = c_2 = c_3 = c$. Table 7 shows the results for a variety of parameter settings in which the numbers of servers are also asymmetric.

p_1	p_2	p_3	β_{PS}	$\underline{\beta}_{fcfs}$			c	ρ_{PS}	$\underline{\rho}_{fcfs}$			sim	approx	$\Delta\%$	simple	$\Delta\%$
0.1	0.2	0.3	0.08	0.80	0.40	0.27	1	0.20	0.20	0.20	0.20	1.97	1.99	0.87	1.38	-29.8
0.1	0.2	0.3	0.20	4.00	2.00	1.33	2	0.50	0.50	0.50	0.50	51.39	52.73	2.61	35.62	-30.7
0.1	0.2	0.4	0.24	7.20	3.60	1.80	3	0.80	0.80	0.80	0.80	675.26	685.68	1.54	390.88	-42.1
0.1	0.2	0.4	0.06	2.40	1.20	0.60	4	0.20	0.20	0.20	0.20	13.85	13.85	0.01	9.37	-32.3
0.1	0.3	0.4	0.10	5.00	1.67	1.25	5	0.50	0.50	0.50	0.50	122.15	126.17	3.29	71.35	-41.9
0.1	0.3	0.4	0.16	9.60	3.20	2.40	6	0.80	0.80	0.80	0.80	949.43	987.05	3.96	494.13	-48.0
0.1	0.3	0.5	0.02	1.40	0.47	0.28	7	0.20	0.20	0.20	0.20	25.66	25.84	0.69	12.77	-50.2
0.2	0.3	0.5	0.05	2.00	1.33	1.00	8	0.50	0.50	0.50	0.50	208.88	213.75	2.33	95.02	-54.5

Table 6: Sojourn time variance with three symmetric multi-server FCFS nodes: approximations versus simulations.

p_1	p_2	p_3	$\underline{\beta}_{fcfs}$			c_1	c_2	c_3	ρ_{PS}	$\underline{\rho}_{fcfs}$			sim	approx	$\Delta\%$
0.1	0.2	0.3	3.00	2.00	1.00	1	2	3	0.50	0.75	0.50	0.25	135.49	131.78	-2.8
0.1	0.2	0.4	2.25	2.70	1.69	1	3	5	0.75	0.75	0.60	0.45	211.61	216.96	-2.5
0.1	0.3	0.4	1.80	0.40	0.15	1	8	9	0.75	0.90	0.08	0.03	638.54	642.47	0.6
0.1	0.3	0.4	1.80	3.20	1.35	1	8	9	0.75	0.90	0.60	0.30	904.46	909.08	0.5
0.3	0.3	0.3	0.40	0.67	0.53	6	4	3	0.80	0.20	0.50	0.80	194.92	199.34	2.3

Table 7: Sojourn time variance with three asymmetric multi-server FCFS nodes: approximations versus simulations.

Tables 6 and 7 demonstrate that the approximations work very well for multi-server nodes at the FCFS nodes.

5.2 General service times at PS node

In this subsection we assess the accuracy of the approximations for non-exponential service-time distributions at the PS node. To this end, we first consider a model with $M = 2$ FCFS nodes, with $c_1 = c_2 = 1$ and routing probabilities $p_1 = p/4$ and $p_2 = 3p/4$. Table 8 shows the results the simulated and approximated values of $Var[S]$ for a variety of parameter, where the squared coefficient of variation of the service-time distribution at the PS node (i.e., c_{PS}^2) is varied as 0, 1, 4 and 16. The service times are deterministic for the case $c_{PS}^2 = 0$ and exponential for $c_{PS}^2 = 1$. For the other cases the service times at the PS node are $H_2(p_{ps}, \beta_{ps,1}, \beta_{ps,2})$-distributed; here the notation $H_2(p_{ps}, \beta_{ps,1}, \beta_{ps,2})$ means that samples from the hyper-exponential distribution are drawn from an exponential distribution with parameter $\beta_{ps,1}$ with probability p_{ps} and with probability $1 - p_{ps}$ the sample is drawn from an exponential distribution with parameter $\beta_{ps,2}$. The parameters $p_{ps}, \beta_{ps,1}$ and $\beta_{ps,2}$ are chosen such that the squared coefficient of variation of the service times equals approximately $c_{PS}^2 = 4$ and 16. The precise parameter values of the H_2 distribution are as follows. The parameters for the cases with $c_{PS}^2 = 4$ are $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.75, 0.1, 1.3)$, $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.75, 0.06, 0.82)$, $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.75, 0.01, 0.13)$, for the cases $p = 0.2$, $p = 0.5$, respectively. These H_2 parameters result in an actual squared coeficient of variation of $c_{PS}^2 = 4.38$, $c_{PS}^2 = 4.47$ and $c_{PS}^2 = 4.38$, respectively. Similarly, for cases

denoted by $c^2_{PS} = 16$ the parameters are $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.95, 0.15, 5.15)$, $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.95, 0.09, 3.29)$, $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.95, 0.015, 0.515)$, for the cases $p = 0.2$, $p = 0.5$ and $p = 0.8$, respectively. Hence, the precise H_2 parameters result in an actual squared coeficient of variation of $c^2_{PS} = 15.84$, $c^2_{PS} = 16.56$, respectively $c^2_{PS} = 15.84$.

p	β_{ps}	c^2_{PS}	$\beta_{fcfs,1}$	$\beta_{fcfs,2}$	simulation	approx	$\Delta\%$
0.2	0.40	0	2.65	8.00	52.60	52.82	0.4
0.2	0.40	1	2.65	8.00	53.76	53.83	0.1
0.2	0.40	4	2.65	8.00	56.98	57.23	0.4
0.2	0.40	16	2.65	8.00	68.18	68.81	0.9
0.5	0.25	0	0.67	2.00	21.43	21.50	0.3
0.5	0.25	1	0.67	2.00	22.08	22.14	0.2
0.5	0.25	4	0.67	2.00	24.45	24.32	-0.5
0.5	0.25	16	0.67	2.00	31.88	31.95	0.2
0.8	0.25	0	0.17	0.51	9.91	10.02	1.1
0.8	0.25	1	0.17	0.51	9.93	10.03	0.8
0.8	0.25	4	0.17	0.51	9.95	10.08	1.1
0.8	0.25	16	0.17	0.51	10.12	10.23	1.1

Table 8: Sojourn time variance for general service times at the PS node: approximations versus simulations.

Finally, Table 9 shows the results for the same model as in Table 8, but with multiple servers: $c_1 = 2$ and $c_2 = 3$. For the results shown in Table 9 the H_2-distribution for $c^2_{PS} = 4$ and $p = 0.2$ the parameters were $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.6, 0.01, 1.6)$, which results in an actual squared coeficient of variation of $c^2_{PS} = 3.91$. For cases denoted by $c^2_{PS} = 4$ and $p = 0.8$ the parameters were $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.6, 0.001, 0.4)$, which results in an actual squared coeficient of variation of $c^2_{PS} = 3.96$. For cases denoted by $c^2_{PS} = 16$ and $p = 0.2$ the parameters were $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.9, 0.05, 6)$, which results in an actual squared coeficient of variation of $c^2_{PS} = 16.32$. Finally, for cases denoted by $c^2_{PS} = 16$ and $p = 0.8$ the parameters were $(p_{ps}, \beta_{ps,1}, \beta_{ps,2}) = (0.9, 0.01, 1.51)$, which results in an actual squared coeficient of variation of $c^2_{PS} = 16.82$.

The results presented in Table 8 and 9 demonstrate that our approximation (29) is also highly accurate for non-exponential service-time distributions at the PS node, with errors of at most a few percent.

Remark 1:
The numerical results presented in Tables 1 to 9 show that the approximation for the variance of the sojourn times in (29) are highly accurate for a remarkably broad range of parameter combinations. Apparently, our closed-form approximation covers the main factors that determine the variance of the total sojourn times of customers in the system.

Remark 2:
Despite the remarkable accuracy of the approximation in (29), almost by definition there are parameter combinations for which the accuracy of the approximation degrades. For the approximation in (29) there are several sources of inaccuracy, which open possibilities for further reducing the inaccuracy of the approximations, at the expense of the

p	β_{ps}	c^2_{PS}	$\beta_{fcfs,1}$	$\beta_{fcfs,2}$	simulation	approx	$\Delta\%$
0.2	0.64	0	5.34	24.00	159.3	161.3	1.2
0.2	0.64	1	5.34	24.00	179.9	181.8	1.1
0.2	0.64	4	5.34	24.00	254.2	253.0	-0.5
0.2	0.64	16	5.34	24.00	532.9	527.5	-1.0
0.2	0.64	0	2.13	38.40	837.4	830.4	-0.8
0.2	0.64	1	2.13	38.40	863.9	850.8	-1.5
0.2	0.64	4	2.13	38.40	923.5	922.92	-0.1
0.2	0.64	16	2.13	38.40	1205.6	1197.3	-0.7
0.2	0.64	0	8.53	9.60	314.3	313.9	-0.7
0.2	0.64	1	8.53	9.60	336.6	334.4	-1.5
0.2	0.64	4	8.53	9.60	399.5	406.3	1.7
0.2	0.64	16	8.53	9.60	697.3	680.8	-2.4
0.8	0.16	0	0.33	1.50	71.3	73.7	3.4
0.8	0.16	1	0.33	1.50	76.7	78.8	2.8
0.8	0.16	4	0.33	1.50	96.3	96.7	0.4
0.8	0.16	16	0.33	1.50	162.4	159.8	-1.6
0.8	0.16	0	0.13	2.40	169.6	171.9	1.3
0.8	0.16	1	0.13	2.40	173.1	177.0	2.3
0.8	0.16	4	0.13	2.40	190.9	195.1	2.2
0.8	0.16	16	0.13	2.40	260.1	257.9	-0.8
0.8	0.16	0	0.53	0.60	124.9	127.5	2.0
0.8	0.16	1	0.53	0.60	130.9	132.6	1.3
0.8	0.16	4	0.53	0.60	149.8	150.7	0.6
0.8	0.16	16	0.53	0.60	218.6	213.6	-2.3

Table 9: Sojourn time variance for general service times at the PS node: approximations versus simulations.

simplicity of the approximation. The first source of inaccuracy stems from the Poisson assumption in AA1, which is generally not the case in networks with feedback loops. Similarly, approximating the results for delayed feedback by known results for non-delayed feedback (AA2), and neglecting the covariance of the successive sojourn times at the different nodes (AA3) are not generally true and hence additional sources of inaccuracy. In the context of the approximations for non-exponential service times at the PS node there are additional sources of inaccuracy. First, the covariance terms in equation (23) are only valid for the case of exponential service times at the PS node (considered in isolation) [1], but not for general non-exponential service times. Second, in general the approximations for $Var[S]$ also depend on the third moments of the service-time distribution at the PS node, whereas approximation (29) only depends on the first two moments of the service-time distributions at the PS node. We refer to [1] for refinements on the approximation for $Var[S]$ in an isolated PS node.

6 Topics for Further Research

The results presented above lead to a number of topics for further research. First, in this paper customers traverse routes through the queueing network according to a Bernoulli feedback scheme, where customers after departing from the PS either leave the system or jump to FCFS node k with probability p_k. An extension of this model, which is very interesting from an application point of view, is to assume deterministic routing, where customers visit the queues in a fixed predetermined order. In this context, notice that product-form solutions, and hence closed-form expressions for the mean sojourn times $E[S]$, also exist for deterministic routing schemes, and moreover, that the covariance results from [1] are applicable to non-Markovian routing schemes. Extension of the results towards deterministic routing schemes is a challenging topic for further research. Second, another model extension that is very interesting from an application point of view is the inclusion of multiple customer types that may each be gouverned by different routing schemes and / or service times. In this context, notice that product-form solutions and hence exact results for $E[S]$, still exist for multiple customer classes under several additional assumptions. Third, in many applications the maximum number of requests that a server will handle simultaneously is limited to some fixed maximum in order to protect the server-side system from getting overloaded. This type of limitations may be included in the model by a token-based mechanism, where customers may need to wait to get access to a token before entering the system. Extension of the model and the results to include the impact of limitations in the number of customers in the system is an interesting topic for further research. Fourth, it is assumed here that the service times at the FCFS nodes are exponentially distributed, whereas in practice the processing times may be non-exponential. Notice that the case of non-exponential service times at the FCFS nodes is fundamentally more complex, and does not admit a product-form solution, so that exact expressions for $E[S]$ can not even be obtained. Extension of the results to incorporate non-exponential service-time distributions for FCFS nodes is an open research topic. Finally, the methodology developed in this paper is new and the results are remarkably accurate. Therefore, it is a challenging topic for further research to investigate to what extent the methodolgy can be applied in a more general context, e.g. application to non-product form queueing networks.

References

[1] J.L. van den Berg and O.J. Boxma (1991). The M/G/1 queue with processor sharing and its relation to a feedback queue. *Queueing Systems* **9**, 365-402.

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

[3] E.G. Coffman, R.R. Muntz and H. Trotter (1970). Waiting time distributions for Processor Sharing Systems. *Journal of the Association for Computing Machinery* **17**, 123-130.

[4] J.W. Cohen (1969). The single server queue. *North-Holland Publishing Company*, Amsterdam.

[5] R.L. Disney and D. Koenig (1985). Queueing networks: a survey of their random processes. *SIAM Rev.* **27**, 335-403.

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

[7] R.D. Foley and R.L. Disney (1983). Queues with delayed feedback. *Advances in Applied Probability* **15**, 162-182.

[8] R.G. Miller (1974). The jackknife - a review. *Biometrika* **61**, 1-15.

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

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

[11] J. Walrand (1988). An introduction to queueing networks. *Prentice-Hall International Editions*, New Jersey.

[12] R.W. Wolff (1989). Stochastic Modeling and the Theory of Queues. *Prenctice Hall International Editions*, New Jersey.