ELSEVIER

Performance Evaluation 49 (2002) 99–110

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Response times in a two-node queueing network with feedback

R.D. van der Mei [a,b,*], B.M.M. Gijsen [a], N. in't Veld [a], J.L. van den Berg [a,c]

[a] *KPN Research, Center of Excellence Quality of Service Control, P.O. Box 421, 2260 AK Leidschendam, The Netherlands*
[b] *Free University, Mathematics and Computer Science, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*
[c] *University of Twente, Mathematical Sciences, P.O. Box 217, 7500 AE Enschede, The Netherlands*

Abstract

The study presented in this paper is motivated by the performance analysis of response times in distributed information systems, where transactions are handled by iterative server and database actions. We model system response times as sojourn times in a two-node open queueing network with a processor sharing (PS) node and a first-come-first-served (FCFS) node. External customers arrive at the PS node according to a Poisson process. After departing from the PS node a customer proceeds to the FCFS node with probability p, and with probability $1-p$ the customer departs from the system. After a visit to the FCFS node, customers are fed back to the PS node. The service requirements at both nodes are exponentially distributed. The model is a Jackson network, admitting a product-from solution for the joint number of customers at the nodes, immediately leading to a closed-form expression for the mean sojourn times in steady-state. The variance of the sojourn times, however, does not admit an exact expression—the complexity is caused by the possibility of overtaking. In this paper we propose a methodology for deriving *simple*, *explicit* and *fast-to-evaluate* approximations for the variance of the sojourn times. Numerical results demonstrate that the approximations are very accurate in most model instances.

Keywords: Queueing networks; Sojourn time; Response time; Feedback; Approximation

1. Motivation and background

This paper is motivated by response time analysis of distributed information systems where transactions are handled by iterative server and database actions. Distributed information systems are commonly implemented in a three-tiered system architecture, with a presentation tier implementing the end-user interface, a business logic tier implementing the service logic and a data tier with information databases and legacy systems. A typical feature of such applications is that a single transaction initiated by the client (at the presentation tier) may initiate a sequence of transactions to be performed on the different system components. Consider for example an on-line service offered by a telephone company that enables the customers to check the status of telephone bills at its home PC with Internet access (see Fig. 1). The client

* Corresponding author. Present address: KPN Research, Department of Network Planning, Room LC 140, P.O. Box 421, 2260 AK Leidschendam, The Netherlands.
E-mail addresses: r.d.vandermei@kpn.com (R.D. van der Mei), b.m.m.gijsen@kpn.com (B.M.M. Gijsen), n.intveld@kpn.com (N. in't Veld), j.l.vandenberg@kpn.com (J.L. van den Berg).

0166-5316/02/$ – see front matter
PII: S0166-5316(02)00121-9

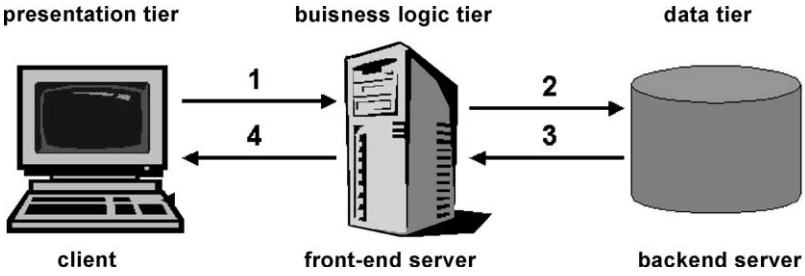

Fig. 1. Three-tiered system architecture for distributed applications.

initiates a transaction request by typing in the proper uniform resource locator (URL), supplemented by a login and password. The request is then sent to a Web server (a front-end server at the business logic tier) that initiates a script (e.g., a common gateway interface (CGI) script, or an active server page (ASP)) that first performs an authentication check at a database, and if successful, sends a query (or a sequence of queries) to the database server (backend server at the data tier) to retrieve the requested information. The Web server then processes the information into a proper format (e.g., hypertext markup language (HTML)) before sending back the response to the client. For the commercial success of distributed information systems the ability to deliver an acceptable quality of service (QoS) level in terms of response times is of key importance.

In this paper we focus on the processing of transactions at the front-end server and the backend servers. The front-end server that implements the business logic typically executes a script that involves a sequence of database accesses in addition to highly CPU-intensive processing steps. Therefore, the front-end server is typically assumed to be CPU-bound, i.e., its processing capacity is limited by the CPU speed. Alternatively, the backend servers typically handle the queries in the order in which they arrive. We analyse response times of a transaction requests by modelling them as sojourn times of a customer in a queueing network. More specifically, we consider a delayed feedback queueing network with a processor sharing (PS) node and a first-come-first-served (FCFS) node. External customers arrive at the PS node according to a Poisson process. After departing from the PS node a customer proceeds to the FCFS node with probability p, and with probability $1 - p$ the customer departs from the system. After each visit to the FCFS node customers are fed back to the PS node. The mean sojourn time follows directly from the product-form solution for the joint number of customers at both nodes, combined with Little's law. In this paper we focus on the *variance* of the total sojourn time.

The analysis of the variance of the sojourn time in the present model is complicated due to the fact that *overtaking* may occur. Overtaking usually destroys any hope for an exact analysis of the higher moments of the sojourn-time distributions. We refer to Boxma and Daduna [2] for an excellent 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 successive sojourn times 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 [5] derive expressions for the LST of the joint distribution of the sojourn times of a customer at its successive passes through the system. We refer to Disney and Koenig [4] for 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 kth time

is looped back into the system with probability p_k and departs from the system with probability $1 - p_k$. For this model, van den Berg and Boxma [1] analyse the joint distribution of the first k successive sojourn times of a customer (who is fed back at least $k - 1$ times). In particular, they derive expressions for the moments of these sojourn times, and for the correlations between the successive sojourn times of an arbitrary customer in the system. Less results are known for sojourn-time distributions for networks with *delayed* feedback (which occurs in the present model). In [6], Foley and Disney studied queueing systems with delayed feedback, but their focus is merely on queue length processes, busy period and several customer flow processes.

In the absence of exact solutions, we propose a new methodology for developing simple, explicit and fast-to-evaluate approximations. We explore the specific structure of the network by combining known results for (instantaneous) feedback queues and product-form queueing networks. We emphasise that the methodology for developing approximations is applicable in a much more general context than the specific model considered here. Numerical results demonstrate that the approximations are highly accurate in most model instances.

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

2. Model

We consider an open queueing model with a single customer class and two nodes: a PS node and a FCFS node (see Fig. 2). External customers arrive at the PS node according to a Poisson process with rate λ. After service completion at the PS node, the customer proceeds to the FCFS node with probability p, and with probability $1 - p$ the customer departs from the system. Customers leaving the FCFS node are always fed back to the PS node. The service times at each node are exponentially distributed with mean β_{ps} and β_{fcfs}, respectively. The successive service times at both nodes are assumed to be mutually independent and independent of the state of the system. The load at the PS node and the FCFS node is given by

$$\rho_{\text{ps}} := \frac{\lambda \beta_{\text{ps}}}{1 - p} \quad \text{and} \quad \rho_{\text{fcfs}} := \frac{\lambda \beta_{\text{fcfs}} p}{1 - p}, \tag{1}$$

respectively. For an arbitrary customer denoted by N the random variable indicating the number of visits to the FCFS node before departing from the system. Then it is readily seen that N is geometrically

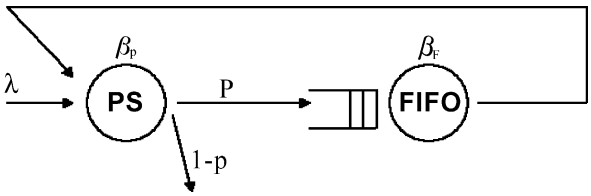

Fig. 2. Illustration of the two-node queueing network model.

distributed with parameter p, i.e., $\text{Prob}\{N = n\} = (1-p)p^n$, for $n = 0, 1, \ldots$. For $i = 1, 2, \ldots, N+1$, let $S_i^{(\text{ps})}$ denote the sojourn time of the ith visit to the PS node, and for $j = 1, \ldots, N$, denote by and $S_j^{(\text{fcfs})}$ the duration of the jth visit to the FCFS node. The total sojourn time is then given by

$$S = \sum_{i=1}^{N+1} S_i^{(\text{ps})} + \sum_{j=1}^{N} S_j^{(\text{fcfs})}. \tag{2}$$

To ensure stability of the system it is assumed that $\rho_{\text{ps}}, \rho_{\text{fcfs}} < 1$.

3. Analysis

The queueing network model described in Section 2 is a Jackson network, and as such has the following product-form solution (with L_{ps} and L_{fcfs} denoting the number of customers at the PS node and at the FCFS node, respectively):

$$\text{Prob}\{L_{\text{ps}} = l; L_{\text{fcfs}} = l'\} = \text{Prob}\{L_{\text{ps}} = l\}\,\text{Prob}\{L_{\text{fcfs}} = l'\} \tag{3}$$

$$\text{Prob}\{L_{\text{ps}} = l; L_{\text{fcfs}} = l'\} = (1-\rho_{\text{ps}})\rho_{\text{ps}}^{l}(1-\rho_{\text{fcfs}})\rho_{\text{fcfs}}^{l'}, \quad l, l' \geq 0. \tag{4}$$

In general, the arrivals at both nodes are not Poisson, and 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* $S_i^{(\text{ps})}$ $(i = 1, \ldots, N+1)$ *are identically distributed.*
(b) *The successive sojourn times* $S_j^{(\text{fcfs})}$ $(j = 1, \ldots, N)$ *are identically distributed.*

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 [10, Theorem 4.4.1]) sees the system in steady-state, *regardless* of its class number, which immediately implies the validity of Lemma 1. □

3.1. Mean sojourn times: exact expressions

Using the above lemma, it follows directly from Eq. (4) and Little's law that

$$E[L_{\text{ps}}] = \frac{\rho_{\text{ps}}}{1-\rho_{\text{ps}}}, \tag{5}$$

$$E[S_i^{(\text{ps})}] = \frac{\rho_{\text{ps}}}{(\lambda/(1-p))(1-\rho_{\text{ps}})} = \frac{\beta_{\text{ps}}}{1-\rho_{\text{ps}}}, \quad i = 1, \ldots, N+1. \tag{6}$$

Recall that the total arrival intensity at the PS node equals $\lambda/(1-p)$. Similarly, for the FCFS node we have

$$E[L_{\text{fcfs}}] = \frac{\rho_{\text{fcfs}}}{1-\rho_{\text{fcfs}}} \tag{7}$$

and

$$E[S_j^{(\text{fcfs})}] = \frac{\rho_{\text{fcfs}}}{(p\lambda/(1-p))(1-\rho_{\text{fcfs}})} = \frac{\beta_{\text{fcfs}}}{1-\rho_{\text{fcfs}}}, \quad j = 1, \ldots, N. \tag{8}$$

Combining (6) and (8) and applying Wald's equation we obtain the following expression for the mean total sojourn time of an arbitrary customer:

$$E[S] = E\left[\sum_{i=1}^{N+1} S_i^{(\text{ps})} + \sum_{i=1}^{N} S_i^{(\text{fcfs})}\right] = (E[N]+1)E[S_1^{(\text{ps})}] + E[N]E[S_1^{(\text{fcfs})}] \tag{9}$$

$$= \frac{1}{(1-p)}\frac{\beta_{\text{ps}}}{(1-\rho_{\text{ps}})} + \frac{p}{(1-p)}\frac{\beta_{\text{fcfs}}}{(1-\rho_{\text{fcfs}})}. \tag{10}$$

3.2. Variance of the sojourn times: approximations

Analysis of the variance of the total sojourn time is fundamentally more complex. The complexity is caused by the fact that *overtaking* may occur. In the absence of exact expressions for the variance of the sojourn times we develop new, simple and fast approximations for the variance of the sojourn times. The accuracy of the results will be demonstrated in the next section.

To start we rewrite the sojourn time variance $\text{Var}[S]$ in the following convenient form:

$$\text{Var}[S] = \text{Var}\left[\sum_{i=1}^{N+1} S_i^{(\text{ps})} + \sum_{j=1}^{N} S_j^{(\text{fcfs})}\right] \tag{11}$$

$$= E\left[\text{Var}\left[\sum_{i=1}^{N+1} S_i^{(\text{ps})} + \sum_{j=1}^{N} S_j^{(\text{fcfs})} | N\right]\right] + \text{Var}\left[E\left[\sum_{i=1}^{N+1} S_i^{(\text{ps})} + \sum_{j=1}^{N} S_j^{(\text{fcfs})} | N\right]\right] \tag{12}$$

$$= \sum_{n=0}^{\infty} \text{Var}\left[\sum_{i=1}^{n+1} S_i^{(\text{ps})} + \sum_{j=1}^{n} S_j^{(\text{fcfs})}\right](1-p)p^n + \text{Var}\left[\sum_{i=1}^{N+1} E[S_i^{(\text{ps})}] + \sum_{j=1}^{N} E[S_j^{(\text{fcfs})}]\right] \tag{13}$$

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

$$+ \sum_{n=0}^{\infty} 2\,\text{Cov}\left[\sum_{i=1}^{n+1} S_i^{(\text{ps})}, \sum_{j=1}^{n} S_j^{(\text{fcfs})}\right](1-p)p^n + \text{Var}[N](E[S_1^{(\text{ps})}] + E[S_1^{(\text{fcfs})}])^2 \tag{14}$$

$$= \sum_{n=0}^{\infty}(n+1)\mathrm{Var}[S_1^{(\mathrm{ps})}](1-p)p^n + \sum_{n=0}^{\infty} n\,\mathrm{Var}[S_1^{(\mathrm{fcfs})}](1-p)p^n$$
$$+ \sum_{n=0}^{\infty}\sum_{i\neq k}\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_k^{(\mathrm{ps})}](1-p)p^n + \sum_{n=0}^{\infty}\sum_{j\neq l}\mathrm{Cov}[S_j^{(\mathrm{fcfs})}, S_l^{(\mathrm{fcfs})}](1-p)p^n$$
$$+ \sum_{n=0}^{\infty}\sum_{i=1}^{n+1}\sum_{j=1}^{n} 2\,\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_j^{(\mathrm{fcfs})}](1-p)p^n + \mathrm{Var}[N](E[S_1^{(\mathrm{ps})}] + E[S_1^{(\mathrm{fcfs})}])^2 \tag{15}$$

$$= \frac{1}{1-p}\mathrm{Var}[S_1^{(\mathrm{ps})}] + \frac{p}{1-p}\mathrm{Var}[S_1^{(\mathrm{fcfs})}] + \sum_{n=0}^{\infty}\sum_{i\neq k}\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_k^{(\mathrm{ps})}](1-p)p^n$$
$$+ \sum_{n=0}^{\infty}\sum_{j\neq l}\mathrm{Cov}[S_j^{(\mathrm{fcfs})}, S_l^{(\mathrm{fcfs})}](1-p)p^n + \sum_{n=0}^{\infty}\sum_{i=1}^{n+1}\sum_{j=1}^{n} 2\,\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_j^{(\mathrm{fcfs})}](1-p)p^n$$
$$+ \frac{p}{(1-p)^2}(E(S_1^{(\mathrm{ps})}) + E[S_1^{(\mathrm{fcfs})}])^2. \tag{16}$$

Eq. (11) follows from Eq. (2), and (12) follows directly from the classical $\mathrm{Var}[U] = E[\mathrm{Var}[U|V]] + \mathrm{Var}[E[U|V]]$ by taking $U := \sum_{i=1}^{N+1} S_i^{(\mathrm{ps})} + \sum_{j=1}^{N} S_j^{(\mathrm{fcfs})}$ and $V := N$. Eq. (13) is then obtained by conditioning with respect to the event $\{N = n\}$. Subsequently, (14) follows from Lemma 1 and classical rules for the variance of random variables, and (15) is obtained from Lemma 1. Finally, Eq. (16) can be obtained via standard calculus.

$E[S_1^{(\mathrm{ps})}]$ and $E[S_1^{(\mathrm{fcfs})}]$ in (16) are given by Eqs. (6) and (8). From Lemma 1 and the sojourn time variance of an M/M/1-FCFS (cf. [3]), we have $\mathrm{Var}[S_1^{(\mathrm{fcfs})}] = (\beta_{\mathrm{fcfs}}/1-\rho_{\mathrm{fcfs}})^2$. Hence, it remains to develop approximations for $\mathrm{Var}[S_1^{(\mathrm{ps})}]$, $\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_j^{(\mathrm{ps})}]$, $\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_j^{(\mathrm{fcfs})}]$ and $\mathrm{Cov}[S_i^{(\mathrm{fcfs})}, S_j^{(\mathrm{fcfs})}]$, for any $i \not\equiv j$.

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

For non-acyclic queueing networks, Approximation Assumption 1 is known to be not true in general, 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. [7]):

$$\mathrm{Var}[S_1^{(\mathrm{ps})}] \approx \frac{2+\rho_{\mathrm{ps}}}{2-\rho_{\mathrm{ps}}}\left(\frac{\beta_{\mathrm{ps}}}{1-\rho_{\mathrm{ps}}}\right)^2. \tag{17}$$

van den Berg and Boxma [1] derived exact expressions for the covariance between the successive sojourn times for single-server FCFS and PS queues with *direct feedback*, where customers upon receiving service are immediately fed back into the system (with some probability). We emphasise that the model discussed in Section 2 implements a *delayed feedback* mechanism: upon departing from the FCFS node, a customer

is first processed by the PS-node before returning to the FCFS node. Similarly, after leaving the PS node, a customer is first processed at the FCFS node before returning to the PS node.

Approximation Assumption 2.

(a) The covariance between the successive sojourn times of a customer at the PS node in the network with delayed feedback is equal to those in a single M/M/1 PS node with direct feedback.
(b) The covariance between the successive sojourn times of a customer at the FCFS node in the network with delayed feedback is equal to those in a single M/M/1 FCFS node with direct feedback.

Now, based on Approximation Assumption 2 we approximate the covariances between the successive sojourn times at the same node (i.e., $\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_j^{(\mathrm{ps})}]$ and $\mathrm{Cov}[S_i^{(\mathrm{fcfs})}, S_j^{(\mathrm{fcfs})}]$ $(i \neq j)$) by the exact results for systems with *direct* feedback, derived from Eq. (9.13), respectively (3.17) in [1]. This leads to the following approximations: for $1 \leq i < n, 1 \leq k \leq n-i$,

$$\mathrm{Cov}[S_i^{(\mathrm{fcfs})}, S_{i+k}^{(\mathrm{fcfs})}] \approx \frac{\rho_{\mathrm{fcfs}}(\rho_{\mathrm{fcfs}}(1-p)+p)^{k-1}\beta_{\mathrm{fcfs}}^2}{(1-\rho_{\mathrm{fcfs}})^2} \tag{18}$$

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

$$\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_{i+k}^{(\mathrm{ps})}] \approx \frac{\rho_{\mathrm{ps}}\beta_{\mathrm{ps}}^2}{(1-\rho_{\mathrm{ps}})^2(2-\rho_{\mathrm{ps}}-p+\rho_{\mathrm{ps}}p)^{k+1}}. \tag{19}$$

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

$$\mathrm{Cov}[S_i^{(\mathrm{ps})}, S_j^{(\mathrm{fcfs})}] \approx 0. \tag{20}$$

In general, Approximation Assumption 3 is known to be not true. However, the product-form solution for the present model, see (4), implies that the *number of customers* at both nodes *are* independent. Also, the sojourn time at the FCFS queue is closely related to the number of customers at that node: if a customer finds n_{fcfs} customers at the FCFS node upon arrival, then the sojourn time simply consists of $n_{\mathrm{fcfs}}+1$ independent successive exponential phases each with rate $\beta_{\mathrm{fcfs}}^{-1}$, which is an Erlang distribution with shape parameter $n_{\mathrm{fcfs}}+1$ and rate parameter $\beta_{\mathrm{fcfs}}^{-1}$. 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. To support this conjecture, we have performed a variety of simulation experiments, calculating the correlation coefficients between the successive sojourn times. For several cases (with varying loads and feedback probabilities) we found that the cross-correlation coefficient were about a factor 2, 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.

Finally, substituting the exact formula for the variance of the sojourn time in the FCFS node and approximations (17)–(20) in the expression for Var[S] in (16) we obtain the following approximation for the variance of the sojourn time:

$$\begin{aligned}\mathrm{Var}[S] \approx{}& \frac{1}{1-p}\frac{2+\rho_{\mathrm{ps}}}{2-\rho_{\mathrm{ps}}}\left(\frac{\beta_{\mathrm{ps}}}{1-\rho_{\mathrm{ps}}}\right)^2 + \frac{p}{1-p}\left(\frac{\beta_{\mathrm{fcfs}}}{1-\rho_{\mathrm{fcfs}}}\right)^2 \\ &+ \frac{2p\rho_{\mathrm{ps}}\beta_{\mathrm{ps}}^2}{(2-\rho_{\mathrm{ps}}-p+p\rho_{\mathrm{ps}})(1-p)^2(2-\rho_{\mathrm{ps}})(1-\rho_{\mathrm{ps}})^2} \\ &+ \frac{2p^2\rho_{\mathrm{fcfs}}\beta_{\mathrm{fcfs}}^2}{(1-p)^2(1-p\rho_{\mathrm{fcfs}}+p)(1-\rho_{\mathrm{fcfs}})^2} + \frac{p}{(1-p)^2}\left(\frac{\beta_{\mathrm{ps}}}{1-\rho_{\mathrm{ps}}}+\frac{\beta_{\mathrm{fcfs}}}{1-\rho_{\mathrm{fcfs}}}\right)^2. \end{aligned} \tag{21}$$

Remark 3.1. We re-emphasise that the model and analysis presented in this paper can be extended to a much wider range of models, without complicating the essence of the analysis. In particular, the general method of developing approximations for queueing networks with feedback is to ignore cross-variances and to calculate the variance of total sojourn times from variances of individual nodes and the covariances between sojourn times of successive visits at each node. This method is expected to achieve accurate results also for other queueing model where customers frequently visit a 'central node' (such as the PS node in the model presented in this paper), as is the case in many server and database system applications. Currently, we are exploring the possibilities for extending the application of this method.

4. Numerical results

To assess the accuracy of the approximations for the variance of the sojourn times proposed in Section 3, we have performed numerous 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 mean service times at both nodes (β_{fcfs}, β_{ps}), and the value of the feedback probability p. From the simulations, we have calculated the point estimates for the variance of the sojourn times, and 95% confidence intervals (CIs). Denoting the point estimations based on simulations by "simulation", and the approximated values by "approx.", the relative error of the approximations is defined as

$$\Delta\% = \frac{\text{approx.} - \text{simulation}}{\text{simulation}} \times 100\%. \tag{22}$$

One might question whether including covariance terms in the approximation only makes the approximation more complex, without gaining a higher level of accuracy. In order 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 (21) without the covariance terms. In Table 1 we will denote the results of this simple approximation by "simple". Table 1 shows the results for the case $\beta_{\mathrm{fcfs}} = 1$ and $\rho_{\mathrm{fcfs}} = \rho_{\mathrm{ps}}$, for a variety of combinations of ρ_{ps}, p, β_{ps} and λ (implicitly). The results presented in Table 1 show that the approximations are highly accurate for all parameter combinations considered, with a worst-case error of only 3%. Further, the results show that

Table 1
Variance of the sojourn times: approximations versus simulations

ρ_{ps}	p	β_{ps}	Simulation	CI	Approx.	Δ%	Simple	Δ%
0.2	0.2	0.2	1.2	(1.21, 1.23)	1.2	0.2	1.2	−2.8
0.2	0.5	0.5	10.2	(10.1, 10.3)	10.1	−0.7	9.6	−6.3
0.2	0.8	0.8	123.3	(122.0, 124.7)	123.5	0.2	113.6	−7.9
0.4	0.2	0.2	2.3	(2.24, 2.27)	2.3	1.5	2.2	−4.5
0.4	0.5	0.5	19.7	(19.6, 19.8)	19.6	−0.3	17.4	−11.7
0.4	0.8	0.8	244.5	(241.2, 247.7)	244.3	−0.1	204.4	−16.4
0.6	0.2	0.2	5.3	(5.25, 5.35)	5.4	2.7	5.0	−6.4
0.6	0.5	0.5	48.3	(47.7, 48.9)	48.7	0.7	40.2	−16.8
0.6	0.8	0.8	628.4	(619.7, 637.0)	621.5	−1.1	467.1	−25.7
0.8	0.2	0.2	24.8	(22.9, 23.6)	23.2	−0.3	20.4	−12.2
0.8	0.5	0.5	215.9	(212.3, 219.5)	218.2	1.1	166.7	−22.8
0.8	0.8	0.8	2956.7	(2904.5, 3009.0)	2868.7	−3.0	1906.7	35.5

the approximation is much more accurate than the simple, straightforward approximation, which shows errors of up to 35%.

For the results presented in Table 1 it is assumed that the load on both nodes are equal. To investigate the impact of asymmetry in the load per node on the accuracy of the approximations, we have also considered a variety of parameter combinations with unequal load per node, with the assumption that $\beta_{fcfs} = 1$. The results are shown in Table 2. The results in Table 2 demonstrate that for these scenarios the relative error is still very low, smaller than 3% and almost all approximation results are within the CIs. Again, the estimation is sometimes higher and sometimes lower than the centre of the CI. It does not seem to make

Table 2
Variance of the sojourn times: approximations versus simulations

ρ_{ps}	ρ_{fcfs}	p	β_{ps}	Simulation	CI	Approx.	Δ%
0.20	0.80	0.2	0.05	16.6	(16.4, 16.9)	16.7	0.2
0.20	0.80	0.5	0.13	111.9	(110.2, 113.5)	114.6	2.5
0.20	0.80	0.8	0.20	1204.2	(1139.4, 1269.0)	1203.6	−0.1
0.80	0.20	0.2	0.80	61.5	(60.5, 62.6)	61.5	−0.1
0.80	0.20	0.5	2.00	979.0	(944.2, 1013.7)	964.2	−2.0
0.80	0.20	0.8	3.20	15603.3	(14495.9, 16710.7)	15514.4	−0.6
0.65	0.85	0.2	0.15	32.0	(31.1, 32.9)	32.0	−0.1
0.65	0.85	0.5	0.38	246.7	(237.3, 256.1)	241.8	−2.0
0.65	0.85	0.8	0.61	2701.3	(2552.7, 2849.9)	2758.2	2.1
0.85	0.65	0.2	0.26	20.0	(19.8, 20.2)	20.0	−0.3
0.85	0.65	0.5	0.65	269.5	(262.1, 276.8)	267.8	−0.6
0.85	0.65	0.8	1.05	4179.5	(3963.3, 4395.7)	4097.4	−2.0
0.20	0.35	0.2	0.11	1.6	(1.58, 1.60)	1.6	0.5
0.20	0.35	0.5	0.29	11.1	(11.0, 11.2)	11.2	0.7
0.20	0.35	0.8	0.46	117.5	(115.6, 119.4)	119.2	1.4
0.35	0.20	0.2	0.35	1.97	(1.96, 1.98)	2.0	−0.0
0.35	0.20	0.5	0.88	21.9	(21.7, 22.1)	21.8	−0.2
0.35	0.20	0.8	1.40	316.2	(311.3, 321.1)	311.9	−1.4

Table 3
Variance of the sojourn times: approximations versus simulations

λ	ρ_{ps}	p	β_{ps}	Simulation	Approx.	Δ%	Simple	Δ%
0.01	0.4	0.95	2	6552.3	6605.2	1	4555	−44
0.01	0.7	0.95	3.5	119491.3	112253.8	−6	57376	−108
0.01	0.95	0.95	4.75	11795228	10126905	−14	3936619	−200

any difference whether the loads of the two nodes are very different, e.g., 0.20–0.80, or close to each other, e.g., 0.65–0.80 and 0.20–0.35. Considering 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 worse. This could be expected, as the approximation contains separate covariance terms for the PS and the FCFS nodes. Consequently, the formulas can adapt to asymmetric loads.

Each approximation, almost by definition, may become inaccurate at some regions of the parameter space. To identify those regions, notice that the formula for the variance of the sojourn times at the PS are based on the assumption that the arrival process is Poisson. However, in general, the arrival process at the PS node is non-Poisson, since the two-node network considered here is not acyclic. Hence, one may expect the approximation to perform worse when the arrival process at the PS node are highly non-Poisson. To this end, we construct a pathological scenario in which the arrival processes at the two nodes are highly non-Poisson by taking the external arrival rate λ close to 0 and p close to 1. This results bursts of arrivals at the nodes. Further, we chose β_{ps} to be negligibly small in order to emphasise the fact that the inaccuracy holds in particular for the PS node. Table 3 shows the results for various parameter combinations (the CIs are omitted here). The results in Table 3 indeed demonstrate that the approximation tends to become less accurate when rate λ becomes very low and p becomes very high, but in several cases still acceptable. When comparing the results to those for the simple approximation introduced earlier in this section, one can see that the errors for approximation (21) are significantly smaller. We emphasise that the scenarios presented in Table 3 are quite pathological, from a practical point of view.

5. Topics for further research

Finally, we address a number of topics for further research. First, in the present paper it is assumed that there is a single FCFS node, representing an information database. In practice, however, information is to be retrieved for different information systems, which could be modelled by multiple FCFS nodes. Similarly, we have assumed the presence of a single PS node, representing a CPU-bound server executing heavy scripting processing. In practice, multiple servers may be implemented in the system, which could be modelled by taking into account multiple PS nodes, or a generalised PS (GPS) node. Extension of the results presented here is an interesting and promising topic for further research. Second, in this paper customers traverse routes through the queueing network according to a Bernoulli feedback scheme. However, the covariance results from [1] are more widely applicable. For example, the approximation developed in this paper can be extended to models where customer follow fixed routes. Another potential model extension is to include multiple customer types that may each be governed by different feedback schemes. 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 needed to enter 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. Further, it is assumed here that the service times are exponentially distributed, whereas in practice the processing times may be far from exponential, and may even be heavy tailed. Extension of the results to incorporate non-exponential service-time distributions is an open research topic (in particular, for the FCFS node). Finally, the methodology developed in this paper is new and the results are very accurate. Therefore, it is a challenging topic for further research to investigate to what extent the methodology can be applied in a more general context, e.g., application to non-product form queueing networks.

References

[1] J.L. van den Berg, O.J. Boxma, The M/G/1 queue with processor sharing and its relation to a feedback queue, Queueing Syst. 9 (1991) 365–402.
[2] O.J. Boxma, H. Daduna, Sojourn times in queueing networks, in: H. Takagi (Ed.), Stochastic Analysis of Computer and Communication Systems, North-Holland, Amsterdam, 1990, pp. 401–450.
[3] J.W. Cohen, The Single Server Queue, North-Holland, Amsterdam, 1969.
[4] R.L. Disney, D. Koenig, Queueing networks: a survey of their random processes, SIAM Rev. 27 (1985) 335–403.
[5] B.T. Doshi, 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, Amsterdam, 1988, pp. 207–233.
[6] R.D. Foley, R.L. Disney, Queues with delayed feedback, Adv. Appl. Prob. 15 (1983) 162–182.
[7] T.J. Ott, The sojourn time distribution in the M/G/1 queue with processor sharing, J. Appl. Prob. 21 (1984) 360–378.
[8] J. Walrand, An Introduction to Queueing Networks, Prentice-Hall, Englewood Cliffs, NJ, 1988.

R.D. van der Mei (1966) received his M.Sc. degrees in Mathematics and in Econometrics both from the Free University of Amsterdam, and his Ph.D. from Tilburg University, The Netherlands. Subsequently, he became a postdoctoral fellow at Rutgers University, New Brunswick, USA, and Columbia University in New York City. In 1996, Dr. van der Mei joined AT&T Labs as a Senior Technical Staff Member, Department of Network Planning and Performance Analysis. In 1999 he joined KPN Research, where he is currently a Senior Researcher at the Quality of Service Excellence Center. In part-time, he is also affiliated with the Free University, Amsterdam, Department of Mathematics and Computer Science. His research interests include performance Modeling of computer-communication networks, queueing theory, TCP performance Modeling, Web server performance, and more recently, he has focused on performance modeling of distributed system architectures.

B.M.M. Gijsen (1970) received his M.Sc. degree in Computer Science from the Technical University Eindhoven, The Netherlands, and his master of technological design degree in Applied Mathematics from the University of Twente. In 1996 he joined KPN Research, Leidschendam, The Netherlands, where he is currently working at the Quality of Service Center of Excellence. His research interests include performance Modeling and evaluation of fixed and mobile packet switched communication networks. Recently, his focus extended to performance modelling and evaluation of distributed information systems at the borders of communication networks.

N. in't Veld (1977) received her M.Sc. degree in Econometrics and Operations Research from the Free University of Amsterdam in 2002. Currently, she is a simulation consultant for Incontrol Enterprise Dynamics in Maarssen, The Netherlands. Her research interests include performance analysis of communication networks and simulations.

J.L. van den Berg (1961) received the M.Sc. and Ph.D. degree in Mathematics from the University of Utrecht, The Netherlands, in 1986 and 1990, respectively. From 1986, he worked at the Centre for Mathematics and Computer Science (CWI), Amsterdam. In 1990, he joined KPN Research, Leidschendam, The Netherlands, as a member of technical staff in the Department Network Planning. Since 1997, he is a Senior Research Member and is currently leader of the QoS group within KPN Research. He is particularly working on performance modelling, evaluation and dimensioning of fixed and mobile communication networks. In these fields, he has co-operated within many international projects, e.g. in the European RACE, ACTS and COST research programmes. Since April 2001, he is (part time) Associate Professor within the Stochastic Operations Research group of the faculty of Mathematical Sciences, University of Twente.