Stochastic Models, 24:487–501, 2008

ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1080/15326340802232327

Taylor & Francis
Taylor & Francis Group

A METHOD FOR APPROXIMATING THE VARIANCE OF THE SOJOURN TIMES IN STAR-SHAPED QUEUEING NETWORKS

R. D. van der Mei[1,2], A. R. de Wilde[1], and S. Bhulai[1,2]

[1]*Department of Probability and Stochastic Networks, Centre for Mathematics and Computer Science, Amsterdam, The Netherlands*
[2]*Department of Mathematics De Boelelaan, Faculty of Sciences, VU University Amsterdam, Amsterdam, The Netherlands*

□ *We study the sojourn times in an open star-shaped queueing network, with a central processor-sharing (PS) node and multiple multi-server First-Come-First Served (FCFS) nodes. Each customer alternatingly visits the central node and one of the other nodes, before departing from the system. For this model, exact expressions for the mean sojourn time can be easily obtained, but an exact analysis of the variance is not possible. Therefore, we propose a method for deriving simple but accurate approximations for the variance of the sojourn times. Extensive simulations demonstrate that the approximations are extremely accurate for a wide range of parameter values.*

Keywords Approximation; Queueing networks; Response time; Sojourn time; Variability.

Mathematics Subject Classification 90B22; 90B18; 68M10; 60B12; 60J80; 60J85.

1. MOTIVATION AND BACKGROUND

The emergence of Web technology has boosted the development of services running in a distributed computing environment, in which data is collected from diverse and remote information services, and processed before a response is returned to the end user. Examples of such distributed applications are on-line ticketing, electronic banking, on-line shopping, or services built on top of location-based services that allow mobile users to get access to location-dependent information. A typical feature of

Received July 2007; Accepted April 2008
Address correspondence to R. D. van der Mei, CWI, Probability and Stochastic Networks, P.O. Box 94079, Amsterdam, 1098 SJ, The Netherlands; E-mail: r.d.van.der.mei@cwi.nl

such services is that a single transaction initiated by the end user (EU) automatically initiates a fixed predetermined sequence of sub-transactions to be performed by the different information services. 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 EU. For the user-perceived responsiveness of the service, both the mean and the variability of the response times are of main importance (Ref.[14]). A second motivation comes from the growth in large-scale distributed applications that typically cross multiple organizational borders. In such a context, the concept of Service Level Agreements (SLAs) provides an effective means to realize a desired end-to-end level as experienced by EUs. In such a case, it is crucial for the service providers to be able to identify the most cost-effective combination of the SLAs to be negotiated with different third parties while meeting the desired end-to-end performance level to the paying EUs. This naturally leads to the problem formulation studied in this paper[13].

Motivated by these application areas, we model the end-to-end response time as the sojourn time of a customer in an open star-shaped queueing network, where the customers represent EU-initiated transactions. The central node represents an application server responsible for implementing the business logic, which usually consists of highly CPU-intensive processing steps, and is therefore modeled as a processor-sharing (PS) node. The information servers can typically handle a given number of information-retrieval requests in parallel (this number may be negotiated in service-level agreements between the application service provider and the information service provider), while excess requests are temporarily queued up, and are therefore modeled as multi-server FCFS-nodes. Customer arrive according to a Poisson arrival process and visit the nodes in a fixed order $0, 1, 0, 2, \ldots, 0, M, 0$, before departing from the system. The service times at the central node are generally distributed, while the service times at the FCFS-nodes are exponential (see Subsection 4.3 for a discussion about this assumption). For this model we are interested in the mean and the variance of the total sojourn time S of a customer. This model is known to possess a product-form solution, which immediately leads to a closed-form expression for $\mathbb{E}[S]$, by using Little's Law. However, an exact analysis of the *variance* of the sojourn time, $\mathbb{V}\mathrm{ar}[S]$, is not possible; this is caused by the fact that *overtaking* may occur (i.e., customers may bypass each other) at the successive visits to the PS-node.

The phenomenon of overtaking usually destroys any hope for an exact analysis of the sojourn-time distributions (see Ref.[2] for a survey of the available results on sojourn times in queueing networks). The main result in Ref.[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. Another complicating aspect that plays a key role in the model analyzed in the present paper is the phenomenon of feedback, i.e., customers may visit queues multiple times; in fact, in our model the PS-node is visited $M + 1$ times by each customer. Several results are known for single-node queueing systems in which customers may be fed back into the system after having received service. 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[11] consider an M/G/1 system, with either FCFS or PS service, where a customer after receiving service for the ith time is immediately looped back into the system with probability q_i and departs from the system with probability $1 - q_i$. For this model, they analyze the joint probability 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*, i.e., where customers after receiving service at a node are fed back into receiving service at one or more other nodes; note that our model also includes delayed feedback: after receiving service at the PS-node, a customer is fed back into the PS-node after receiving service at one of the FCFS-nodes. Foley and Disney[7] study queueing systems with delayed feedback, but their focus is merely on queue-length processes, busy periods, and several customer flow processes. Van der Mei et al.[12] develop approximations for the variance of the sojourn times in a two-node queueing network with a PS-node and a single-server FCFS-node. Gijsen et al.[8] extend the results in[12] to star-shaped queueing networks with a central PS-node and multiple multi-server FCFS-nodes with exponential service times, and with Markovian routing. In the present paper, where we consider the same model but with deterministic routing, we build upon the insights obtained in Ref.[8], which can be seen as the random counterpart of the present paper. Note that the models in Refs.[8,12] possess a product-form solution, so that an exact expression is known for the expected sojourn times. For queueing-network models that do not admit a product-form solution, hardly any exact results on the sojourn-times are known. As an exception, Boxma et al.[3] develop approximations for the sojourn-time distributions for a two-node model with a PS-node, a single-server FCFS-node, and with Bernoulli feedback.

In the absence of exact results for the variance of the sojourn time on queueing networks in which overtaking may occur, in this paper we propose and validate a new method to develop simple approximations

for the variance of the sojourn time S. The method is based on two simple ideas: (1) merging the $M+1$ different service-time requirements corresponding to different visits of a customer to the PS-node, say $B_1, \ldots, B_{M+1}$, into a single visit with convolved service-time requirement $B_1 + \cdots + B_{M+1}$, and (2) assuming the other cross-correlations between the visit times to be mutually independent. Using these assumptions, we use the results for the sojourn times in single-node systems[11] to obtain a simple approximation for $\mathbb{V}\mathrm{ar}[S]$. To assess the accuracy of the approximations, we have conducted extensive experiments with simulations. The results show that the approximations are extremely accurate for a wide range of parameter settings.

The contribution of this paper is two-fold. First, from a methodological point of view, it is challenging to develop simple but accurate approximate methods to analyze the sojourn-time distributions for queueing networks in which overtaking may occur. The proposed method, which is based on simple ideas, appears to be very effective and seems to be applicable for a much broader class of queueing networks than the one analyzed in this paper. As such, the proposed method can be viewed as an effective way to approximate sojourn times in queueing networks. Second, the star-shaped model discussed in this paper is strongly motivated by the recent emergence of real-time services that operate in large-scale distributed computing environment services, while for these services both the mean and the variance of the response times are crucial for the end-user perception of the quality[14]. This makes the methods developed in this paper valuable from an application point of view.

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, and subsequently, we develop an approximation for the variance of the sojourn times. Finally, in Section 4 the accuracy of the approximations is tested extensively by comparing the performance predictions based on the approximations with simulation results.

2. MODEL

Consider an open queuing network with a single class of customers, a PS-node and $M \geq 1$ FCFS-nodes with c_k servers at node k ($k = 1, \ldots, M$). Customers arrive at the PS-node according to a Poisson process with rate λ; for notational convenience, the PS-node will be referred to as node 0. Each customer visits the nodes in the (fixed) order $0, 1, 0, 2, \ldots, 0, M, 0$, before departing from the system; thus, all FCFS-nodes are visited once, while the PS-node is visited $M+1$ times by each customer. The service time at the PS-node is a generally distributed random variable B_{PS} and with finite first two moments β_{PS} and $\beta_{\mathrm{PS}}^{(2)}$, respectively. The service time at FCFS-node k is exponentially distributed with mean β_k for $k = 1, \ldots, M$. The service times

at all nodes are mutually independent. The load at the PS-node and at FCFS-node k is given by

$$\rho_{\mathrm{PS}} := (M+1)\lambda\beta_{\mathrm{PS}}, \quad \text{and} \quad \rho_k := \frac{\lambda\beta_k}{c_k}, \quad (k = 1, \ldots, M), \tag{1}$$

respectively. Then, as we define $S_{\mathrm{PS}}^{(i)}$ as the sojourn time of the ith visit to the PS-node, for $i = 1, 2, \ldots, M$, and also define S_k as the sojourn time of the only visit to FCFS-node k $(k = 1, 2, \ldots, M)$, the total sojourn time is given by

$$S := \sum_{i=1}^{M+1} S_{\mathrm{PS}}^{(i)} + \sum_{k=1}^{M} S_k. \tag{2}$$

To ensure stability at each of the nodes, we assume $\rho_{\mathrm{PS}} < 1$ and $\rho_k < 1$ $(k = 1, \ldots, M)$.

3. ANALYSIS

In this section we derive expressions for the mean sojourn time $\mathbb{E}[S]$ and the variance $\mathbb{V}\mathrm{ar}[S]$. The expression for $\mathbb{E}[S]$ follows directly from the fact that the model possesses a product-form solution. However, an exact expression for $\mathbb{V}\mathrm{ar}[S]$ cannot be obtained; therefore, we develop simple approximate expressions for $\mathbb{V}\mathrm{ar}[S]$.

3.1. Mean Sojourn Times

The network model under consideration is readily send to fall within the class of so-called BMCP-networks[1] and hence, possesses a product-form solution. That is, if we define L_{PS} and L_k to be the steady-state number of customers at the PS-node and at the kth FCFS-node, respectively, then (Ref.[1]):

$$\mathrm{Prob}(L_{\mathrm{PS}} = l; L_1 = l_1; \ldots; L_M = l_M) = \mathrm{Prob}(L_{\mathrm{PS}} = l) \prod_{k=1}^{M} \mathrm{Prob}(L_k = l_k), \tag{3}$$

with $l \geq 0$ and $l_k \geq 0$ for $k = 1, \ldots, M$. For the PS-node we get the equilibrium distribution $\mathrm{Prob}(L_{\mathrm{PS}} = l) = (1 - \rho_{\mathrm{PS}})\rho_{\mathrm{PS}}^l$ $(l = 0, 1, \ldots)$, which implies

$$\mathbb{E}[L_{\mathrm{PS}}] = \frac{\rho_{\mathrm{PS}}}{1 - \rho_{PS}}. \tag{4}$$

Then, using Little's Law we get

$$\mathbb{E}\big[S_{\mathrm{PS}}^{(i)}\big] = \frac{\rho_{\mathrm{PS}}}{(M+1)\lambda(1-\rho_{\mathrm{PS}})} = \frac{\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}}, \tag{5}$$

for $i = 1, \ldots, M+1$. For the FCFS-nodes it is readily verified that the marginal distribution of L_k satisfies the following equations

$$\lambda \mathrm{Prob}(L_k = l_k - 1) = \min(l_k, c_k)\mathrm{Prob}(L_k = l_k)/\beta_k, \tag{6}$$

for $k = 1, \ldots, M$ and $l_k = 0, 1, 2, \ldots$. Iterating gives

$$\mathrm{Prob}(L_k = l_k) = \frac{(c_k\rho_k)^{l_k}}{l_k!}\mathrm{Prob}(L_k = 0), \quad l_k = 0, \ldots, c_k, \tag{7}$$

and

$$\mathrm{Prob}(L_k = c_k + l_k) = \rho_k^{l_k}\frac{(c_k\rho_k)^{c_k}}{c_k!}\mathrm{Prob}(L_k = 0), \quad l_k = 0, 1, 2, \ldots, \tag{8}$$

for $k = 1, \ldots, M$ and where $\mathrm{Prob}(L_k = 0)$ follows from normalization, yielding

$$\mathrm{Prob}(L_k = 0) = \left(\sum_{l=0}^{c_k-1}\frac{(c_k\rho_k)^l}{l!} + \frac{(c_k\rho_k)^{c_k}}{c_k!}\frac{1}{1-\rho_k}\right)^{-1}. \tag{9}$$

From the probabilities in (7) and (8) we can derive the probability π_k that a customer has to wait at FCFS-queue k ($k = 1, \ldots, M$):

$$\begin{aligned}\pi_k &= \sum_{l=0}^{\infty}\mathrm{Prob}(L_k = c_k + l) = \mathrm{Prob}(L_k = c_k)\big[1 + \rho_k + \rho_k^2 + \cdots\big] \\ &= \frac{\mathrm{Prob}(L_k = c_k)}{1-\rho_k} = \frac{(c_k\rho_k)^{c_k}}{c_k!}\left((1-\rho_k)\sum_{l=0}^{c-1}\frac{(c_k\rho_k)^l}{l!} + \frac{(c_k\rho_k)^{c_k}}{c_k!}\right)^{-1}. \end{aligned} \tag{10}$$

Then it is readily seen that the mean sojourn time at FCFS-node k is given by

$$\mathbb{E}[S_k] = \frac{\beta_k}{(1-\rho_k)c_k}\pi_k + \beta_k, \tag{11}$$

for $k = 1, \ldots, M$. Combining (2), (5), and (11) we obtain the following expression for the mean total sojourn time of an arbitrary customer:

$$\mathbb{E}[S] = (M+1)\mathbb{E}\left[S_{\mathrm{PS}}^{(1)}\right] + \sum_{k=1}^{M} \mathbb{E}[S_k] \frac{(M+1)\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}} + \sum_{k=1}^{M} \left(\frac{\beta_k}{(1-\rho_k)c_k} \pi_k + \beta_k \right), \tag{12}$$

where π_k is given in (10).

3.2. Variance of the Sojourn Times

In this section an approximation for the variance of the sojourn times in a queueing network with deterministic routing will be derived. To start, we write the variance of the sojourn times in the following general form:

$$\mathbb{V}\mathrm{ar}[S] = \mathbb{V}\mathrm{ar}\left[\sum_{i=1}^{M+1} S_{\mathrm{PS}}^{(i)} + \sum_{k=1}^{M} S_k \right]. \tag{13}$$

To obtain an approximation for $\mathbb{V}\mathrm{ar}[S]$ we make the following simplifying assumptions:

Approximation Assumption 1 (AA1). The total sojourn time of a customer at the PS-node equals the sojourn time of an $M/G/1$-PS model with arrival rate λ and where the service-times are distributed as the convolution $B_1 + \cdots + B_{M+1}$.

Approximation Assumption 2 (AA2). The sojourn times at the FCFS-nodes are uncorrelated: $\mathbb{C}\mathrm{ov}(S_i, S_j) = 0$ for all $i, j = 1, \ldots, M$ with $i \neq j$.

Approximation Assumption 3 (AA3). The sojourn times at the FCFS-nodes and the sojourn times at the PS-node are uncorrelated: $\mathbb{C}\mathrm{ov}(S_{\mathrm{PS}}^{(i)}, S_j) = 0$ for $i = 1, \ldots, M+1$, $j = 1, \ldots, M$.

The motivation behind AA1 is as follows: consider a single-node PS-model with Poisson arrivals in which each customer is visiting the PS-node $M+1$ times in a row, with generally distributed service-time requirements $B_1, \ldots, B_{M+1}$, respectively. Then, when a tagged customer T is fed back into the node *immediately* after having received service B_i $(i = 1, \ldots, M)$, the total amount of service time required by T is $B_1 + \cdots + B_{M+1}$. Hence, the evolution of this system is stochastically identical to the classical $M/G/1$-PS system (without feedback) where the service-time requirement is the convolution $B_1 + \cdots + B_{M+1}$. In this context, note that the network model under consideration (see Section 2) implements delayed feedback to the PS-node, in the sense that customers are fed back

into the PS-node after some amount of delay D_i, being the sojourn time at FCFS-node i ($i = 1, \ldots, M$). In other words, from the perspective of the PS-node AA1 approximates a system with delayed feedback to the PS-node by a system with immediate feedback to the PS-node. Assumptions AA2 and AA3 are motivated by the fact that the system possesses a product-form solution, which implies that the steady-state number of customers at the PS and FCFS-nodes are mutually independent. Although this does not imply that the sojourn times of a given customer at the FCFS-nodes are independent (!), one may suspect that the dependence is rather weak.

To approximate the variance of the sojourn times in the PS-node, we use the approximate expression in Ref.[11] for the second moment of the sojourn time $S_{M/G/1}$ of an $M/G/1$-PS system, with load ρ and where the service time has mean $\beta_{M/G/1}$ and squared coefficient of variation $c^2_{M/G/1}$. For $\rho < 1$,

$$\mathbb{E}\left[S^2_{M/G/1}\right] \approx c^2_{M/G/1}\left(1 + \frac{2+\rho}{2-\rho}\right)\frac{\beta^2_{M/G/1}}{(1-\rho)^2} + (1 - c^2_{M/G/1})\left(\frac{2\beta^2_{M/G/1}}{(1-\rho)^2} - \frac{2\beta^2_{M/G/1}}{\rho^2(1-\rho)}(e^\rho - 1 - \rho)\right). \quad (14)$$

Note that the approximation is based on a linear inter- or extrapolation of the known exact results for the cases of deterministic and exponential service times, where $c^2_{M/G/1}$ is the interpolating factor. Using this approximation for the model under consideration (Assumption AA1), the total service-time distribution at the PS-node is the $(M+1)$-fold convolution of the service-time distribution at the PS-node, and hence has mean $(M+1)\beta_{\mathrm{PS}}$ and squared coefficient of variation $c^2_{\mathrm{PS}}/(M+1)$. Hence, the second moment of the total sojourn time at the PS-node, $S_{\mathrm{PS}} := \sum_{i=1}^{M+1} S^{(i)}_{\mathrm{PS}}$, is approximated by the following expression:

$$\begin{aligned}
&\mathbb{E}[S^2_{\mathrm{PS}}] \\
&\quad\approx \frac{c^2_{\mathrm{PS}}}{M+1}\left(1 + \frac{2+\rho_{\mathrm{PS}}}{2-\rho_{\mathrm{PS}}}\right)\left(\frac{(M+1)\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}}\right)^2 \\
&\qquad + \left(1 - \frac{c^2_{\mathrm{PS}}}{M+1}\right)\left(\frac{2((M+1)\beta_{\mathrm{PS}})^2}{(1-\rho_{\mathrm{PS}})^2} - \frac{2((M+1)\beta_{\mathrm{PS}})^2}{\rho^2_{\mathrm{PS}}(1-\rho_{\mathrm{PS}})}(e^{\rho_{\mathrm{PS}}} - 1 - \rho_{\mathrm{PS}})\right) \\
&\quad= (M+1)c^2_{\mathrm{PS}}\left(1 + \frac{2+\rho_{\mathrm{PS}}}{2-\rho_{\mathrm{PS}}}\right)\left(\frac{\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}}\right)^2 + \left((M+1)^2 - (M+1)c^2_{\mathrm{PS}}\right) \\
&\qquad \times \left(\frac{2\beta^2_{\mathrm{PS}}}{(1-\rho_{\mathrm{PS}})^2} - \frac{2\beta^2_{\mathrm{PS}}}{\rho^2_{\mathrm{PS}}(1-\rho_{\mathrm{PS}})}(e^{\rho_{\mathrm{PS}}} - 1 - \rho_{\mathrm{PS}})\right), \quad (15)
\end{aligned}$$

where c_{PS}^2 stands for the squared coefficient of variation of the (individual) service-time distribution at the PS-node. The variance of the sojourn time with general service times at the PS-node can now be obtained by

$$\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^{M+1} S_{\mathrm{PS}}^{(i)}\right] = \mathbb{E}\left[S_{\mathrm{PS}}^2\right] - \left((M+1)\mathbb{E}\left[S_{\mathrm{PS}}^{(1)}\right]\right)^2. \tag{16}$$

Using AA2 and AA3, the variance of the total sojourn time at the FCFS-nodes can be approximated as follows:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\sum_{k=1}^{M} S_k\right] &= \sum_{k=1}^{M} \mathbb{V}\mathrm{ar}[S_k] + 2\sum_{i=1}^{M}\sum_{j=i+1}^{M} \mathbb{C}\mathrm{ov}[S_i, S_j] \\
&\approx \sum_{k=1}^{M} \left(\mathbb{E}[W_k^2] - (\mathbb{E}[W_k])^2 + \beta_k^2\right) \\
&= \sum_{k=1}^{M} \left(\pi_k \frac{2\beta_k^2}{c_k^2(1-\rho_k)^2} - \pi_k^2 \frac{\beta_k^2}{c_k^2(1-\rho_k)^2} + \beta_k^2\right) \\
&= \sum_{k=1}^{M} \left(\frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2} + \beta_k^2\right),
\end{aligned} \tag{17}$$

where W_k represents the waiting time at queue $k = 1, \ldots, M$, and can be derived using (7)–(10). Using the assumption AA3 and substituting (17) and (16) in (13) leads to the following approximate expression for $\mathbb{V}\mathrm{ar}[S]$:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\mathrm{app}}[S] := {} & (M+1)c_{\mathrm{PS}}^2\left(1 + \frac{2+\rho_{\mathrm{PS}}}{2-\rho_{\mathrm{PS}}}\right)\left(\frac{\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}}\right)^2 \\
& + \left((M+1)^2 - (M+1)c_{\mathrm{PS}}^2\right) \\
& \times \left(\frac{2\beta_{\mathrm{PS}}^2}{(1-\rho_{\mathrm{PS}})^2} - \frac{2\beta_{\mathrm{PS}}^2}{\rho_{\mathrm{PS}}^2(1-\rho_{\mathrm{PS}})}(e^{\rho_{\mathrm{PS}}} - 1 - \rho_{\mathrm{PS}})\right) \\
& - \left(\frac{(M+1)\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}}\right)^2 + \sum_{k=1}^{M}\left(\beta_k^2 + \frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2}\right).
\end{aligned} \tag{18}$$

In the context of assumption AA1, it is interesting to realize that the sojourn times corresponding to the successive visits to the PS-node (i.e., $S_{\mathrm{PS}}^{(i)}$, $i = 1, \ldots, M$) may be highly dependent, even though the corresponding service-time requirements B_i $(i = 1, \ldots, M)$ are mutually independent. To give an intuitive explanation for this, note that if the sojourn time of a customer during a visit to the PS-node is very long, then most likely there will have been many other customers at the PS-node

at the same time. Hence, if the customer is fed back into the PS-node, possibly after some amount of delay, then the number of customers present at that moment is likely to be still relatively high, which implies that the sojourn time of the next visit will also be relatively long. These arguments intuitively explain the fact that the correlations between the successive visits to the PS-node (i.e., $S_{\mathrm{PS}}^{(i)}$, $i = 1, \ldots, M$) may be significant. We emphasize that by taking assumption AA1, we implicitly take the cross-correlations between the successive visits to the PS-node into account; in fact, the numerical results discussed below show that the inclusion of these correlations via AA1 is highly effective.

The question arises how the accuracy of the approximations degrades if the correlations between the successive visits to the PS-node are neglected, so that the successive arrivals to the PS-node are assumed to be independent (with rate $(M+1)\lambda$). In this context, the numerical results in Table 3 in Ref.[12] show that if the arrivals at the PS are highly correlated and are extremely bursty, then an approximation based on independent arrivals may become inaccurate. On the other hand, the results in Tables 1 and 2 in Ref.[12] show that if the successive visits to the central node are sufficiently separated in time, such an approximation is still very accurate.

4. VALIDATION

To assess the accuracy of the approximation of $\mathbb{V}\mathrm{ar}[S]$ in (18), we have performed extensive simulation experiments, comparing the approximations to simulation results for many parameter combinations, by varying the arrival rate, the service time distributions, the asymmetry in the loads of the nodes, and the numbers of servers at the FCFS-nodes. All simulations were run 15 times where the lengths of the runs are taken long enough to ensure that the confidence intervals are sufficiently narrow. Denoting the point estimations based on simulations by "simulation" and the approximated values by "approximation", the relative error of the approximations is defined as

$$\Delta\% := \frac{\text{approximation} - \text{simulation}}{\text{simulation}} \times 100\%. \tag{19}$$

To illustrate the accuracy of the results, we also compare the approximation defined in (18) with the following simple approximation, which would directly follow from the results in (14) by simply assuming that the arrival processes at all nodes are Poisson (which is clearly not the case in the model under consideration) and the duration of *all* visits

to any node are mutually independent (recall from Section 3.2 that in general the successive visits to the PS-node *are* dependent):

$$\mathbb{V}\mathrm{ar}_{\mathrm{simple}}[S] := (M+1)\Bigg[c_{\mathrm{PS}}^2\left(1+\frac{2+\rho_{\mathrm{PS}}}{2-\rho_{\mathrm{PS}}}\right)\left(\frac{\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}}\right)^2+\left(1-c_{\mathrm{PS}}^2\right) \times\left(\frac{2\beta_{\mathrm{PS}}^2}{(1-\rho_{\mathrm{PS}})^2}-\frac{2\beta_{\mathrm{PS}}^2}{\rho_{\mathrm{PS}}^2(1-\rho_{\mathrm{PS}})}(e^{\rho_{\mathrm{PS}}}-1-\rho_{\mathrm{PS}})\right) -\left(\frac{\beta_{\mathrm{PS}}}{1-\rho_{\mathrm{PS}}}\right)^2\Bigg]+\sum_{k=1}^{M}\left(\beta_k^2+\frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2}\right). \quad (20)$$

Here, the first term between brackets is the approximated variance of the total sojourn time of the $(M+1)$ visits to the PS-node, and follows directly from (14) and (16), while the second term is the approximated variance of the sojourn times at the FCFS-nodes, which follows directly from the approximation in (17). The results of our validation experiments are outlined below.

4.1. Single Servers at the FCFS-Nodes

Consider a model with $\lambda = 1$ and with $M = 5$ asymmetric FCFS nodes with means $\beta_1, \ldots, \beta_5$, and where the service times are exponentially distributed. Table 1 shows the results for a variety of values of β_{PS} and $\beta_1, \ldots, \beta_5$. Column 7 shows the "exact" values of $\mathbb{V}\mathrm{ar}[S]$ obtained via simulations (denoted $\mathbb{V}\mathrm{ar}_{\mathrm{sim}}[S]$), and column 8 shows half the width of the corresponding 95% confidence interval (indicated as C.I.), i.e., $\mathbb{V}\mathrm{ar}_{\mathrm{sim}}[S] \pm$ C.I. represents the 95% confidence interval. Column 9 shows $\mathbb{V}\mathrm{ar}_{\mathrm{app}}[S]$ according to (18), and column 10 shows the corresponding relative error defined in (19). Columns 11 and 12 show the results for the "simple" approximation defined in (20), and the corresponding relative error.

The results presented in Table 1 show that the results are very accurate, even for high loads, with errors typically less than 5%.

4.2. Multiple Servers at the FCFS-Nodes

To evaluate the accuracy of our approximation for asymmetric multi-server FCFS-nodes, consider the model with two multi-server queues with $c_1 = 2$ and $c_2 = 3$, and with $\lambda = 1$. Table 2 shows the results of the simulated and approximated values of $\mathbb{V}\mathrm{ar}[S]$ for a variety of parameters, where the service times of the FCFS-nodes are exponentially distributed and the squared coefficient of variation c_{PS}^2 of the service time distribution at the PS-node is varied as 0, 4, and 16. These last service times are deterministic for the case $c_{\mathrm{PS}}^2 = 0$ and gamma-distributed

TABLE 1 Sojourn time variances of a queueing network with five single-server FCFS-nodes with unequal service-time distributions

β_{PS}	β_1	β_2	β_3	β_4	β_5	$\mathbb{V}\text{ar}_{sim}[S]$	C.I.	$\mathbb{V}\text{ar}_{app}[S]$	$\Delta_{app}\%$	$\mathbb{V}\text{ar}_{simple}[S]$	$\Delta_{simple}\%$
0.05	0.2	0.2	0.2	0.2	0.2	0.41	0.00	0.39	−5.65	0.60	44.87
0.05	0.5	0.5	0.5	0.5	0.5	5.16	0.04	5.08	−1.70	5.28	2.32
0.05	0.8	0.8	0.8	0.8	0.8	82.82	4.19	80.08	−3.31	80.28	−3.06
0.05	0.1	0.3	0.5	0.7	0.9	87.37	6.76	87.72	0.39	87.92	0.63
0.05	0.9	0.7	0.5	0.3	0.1	89.57	11.15	87.72	−2.07	87.92	−1.83
0.05	0.5	0.1	0.7	0.3	0.9	92.83	8.12	87.72	−5.51	87.92	−5.28
0.05	0.8	0.8	0.8	0.8	0.8	79.61	5.07	80.08	0.59	80.28	0.85
0.15	0.1	0.3	0.5	0.7	0.9	171.74	9.86	181.40	5.63	87.92	−48.80
0.15	0.9	0.7	0.5	0.3	0.1	186.44	18.97	181.40	−2.70	87.92	−52.84
0.15	0.5	0.1	0.7	0.3	0.9	183.25	19.51	181.40	−1.01	87.92	−52.02
0.15	0.1	0.3	0.5	0.7	0.9	173.06	15.21	181.40	4.82	87.92	−49.20

for the cases $c_{PS}^2 = 4$ and $c_{PS}^2 = 16$. To obtain $c_{PS}^2 = 4$, we use a gamma distribution with shape parameter $\gamma = 1/4$ and scale parameter $\lambda = 5/2$ for $\beta_{PS} = \gamma/\lambda = 0.1$, and we use a gamma-distribution with $\gamma = 1/4$ and $\lambda = 5/6$ for $\beta_{PS} = \gamma/\lambda = 0.3$. To obtain $c_{PS}^2 = 16$, we use a gamma distribution with $\gamma = 1/16$ and $\lambda = 5/8$ for $\beta_{PS} = \gamma/\lambda = 0.1$, and we use a gamma distribution with $\gamma = 1/16$ and $\lambda = 5/24$ for $\beta_{PS} = \gamma/\lambda = 0.3$. The results in Table 2 show that for these cases the approximation given in (18) is still very accurate.

To summarize, the numerical results presented in Tables 1–2 show that the approximations defined in (18) are highly accurate for a wide range of parameter combinations. Most interestingly, the approximation by far outperforms the accuracy of the simple approximation; this underlines the relevance of assumption AA1, which leads to a dramatic decrease

TABLE 2 Sojourn time variances for a queueing network with general service times at the PS-node and with two asymmetric multi-server FCFS-nodes, with $c_1 = 2$ and $c_2 = 3$

β_{PS}	c_{PS}^2	β_1	β_2	$\mathbb{V}\text{ar}_{sim}[S]$	C.I.	$\mathbb{V}\text{ar}_{app}[S]$	$\Delta_{app}\mathbb{V}\text{ar}\%$	$\mathbb{V}\text{ar}_{simple}[S]$	$\Delta_{simple}\%$
0.1	0	0.2	2.7	80.82	3.39	85.66	5.99	85.62	5.94
0.3	0	1.6	1.5	88.82	4.05	89.70	0.99	20.02	−77.46
0.1	0	1.0	0.9	2.43	0.01	2.43	−0.02	2.39	−1.61
0.3	0	1.8	0.3	149.31	5.39	152.38	2.05	82.69	−44.62
0.1	4	0.2	2.7	88.50	3.24	85.94	−2.90	85.78	−3.08
0.3	4	1.6	1.5	272.00	12.89	281.35	3.44	22.50	−91.73
0.1	4	1.0	0.9	2.71	0.01	2.71	−0.23	2.55	−6.18
0.3	4	1.8	0.3	330.14	14.09	344.03	4.21	85.18	−74.20
0.1	16	0.2	2.7	89.60	3.76	86.77	−3.16	86.24	−3.76
0.3	16	1.6	1.5	820.26	53.90	856.30	4.39	29.97	−96.35
0.1	16	1.0	0.9	3.57	0.01	3.54	−0.79	3.01	−15.66
0.3	16	1.8	0.3	920.45	63.10	918.98	−0.16	92.64	−89.93

in the accuracy of the approximation. Apparently, the "silver bullet" is based on (1) the simple idea of merging the visits to the PS-node with requirements $B_1, \ldots, B_{M+1}$ into a single visit of duration $B_1 + \cdots + B_{M+1}$, thereby accurately capturing the correlations between the successive sojourn times of a customer to the PS-node (see Section 4.3), and (2) the idea of neglecting all other cross-correlations between pairs of visits to the nodes.

4.3. Discussion and Remarks

The assumption that the service-time distributions at the FCFS-nodes are exponentially distributed was mainly made for technical reasons: under this assumption the model under consideration is known to have a product-form solution, which immediately gives an exact expression for the mean sojourn times. Whether or not this assumption affects the accuracy of the approximations depends on the specific application scenario. In some cases this assumption is acceptable, especially in application areas where a rough indication of the *mean* service time is the best one can get in the first place, which is common practice. In other cases more detailed information about the service-time distribution may be available, and the exponentiality assumption may be found to be far from realistic. As a consequence, in those cases, an extension of the results to general service-time distributions is needed, in which case no product-form solution exists. In those cases, even expressions for the expected sojourn times cannot be obtained, which opens up a challenging area for further research on approximate methods; see for example, Ref.[3] for pioneering contributions in that direction.

Despite the remarkable accuracy of the approximation in (18), almost by definition there are parameter combinations for which the accuracy of the approximation degrades. For the approximation in (18) there are several sources of inaccuracy, which open possibilities for further reducing the inaccuracy of the approximations, at the expense of the simplicity of the approximation. The first source of inaccuracy stems from the approximation in (14), which is based on an approximation of the second moment of the sojourn times in an $M/G/1$ PS-node with general service-time distributions, proposed in Ref.[11], which generally depend on the complete distribution of the service times. The approximation in (14) is a simple linear interpolation between exact results known for the cases of deterministic and exponential service-time distributions. We suspect that the accuracy of (14) will degrade if the service-time distribution at the PS-node is very erratic (i.e., c_{PS}^2 large). In those cases, one may use the refined approximation in Ref.[10] that takes into account third moments of the service-time distributions. The second source of inaccuracy stems from the assumption that the service-time requirements $B_1, \ldots, B_M$ at the PS-node

are merged into a single visit to the PS-node with service-time requirement $B_1 + \cdots + B_{M+1}$; in this way, the variance of the sojourn time in the model with delayed feedback (discussed in Section 3) is approximated by the sojourn time in a model with immediate feedback (14). This source of inaccuracy may manifest itself strongly when the loads of the FCFS-nodes are extremely high, so that the sojourn times at these nodes is very long, which implies that the correlation between successive visits of the same customer to the PS-node will be weak, in which case the approximation based on the results in Ref.[11] on immediate feedback may become inaccurate. Finally, assumptions AA2 and AA3 may potentially become inaccurate whenever the correlation between successive visits to the FCFS-nodes and the cross-correlations between visits to the PS-node and the FCFS-nodes are no longer negligible. This type of correlations may only occur when both the PS-node and one or more FCFS-nodes have the same loads and are very heavily loaded. But even in those cases we found that it is hard to find model instances for which the approximations are inaccurate.

REFERENCES

1. Baskett, F.; Chandy, K.M.; Muntz, R.R.; Palacios, G. Open, closed, and mixed networks of queues with different classes of customers. Journal of the ACM **1985**, *22*, 248–260.
2. Boxma, O.J.; Daduna, H. Sojourn times in queueing networks. In *Stochastic Analysis of Computer and Communication Systems*; Takagi, H., Ed.; North Holland, 1990; 401–450.
3. Boxma, O.J.; van der Mei, R.D.; Resing, J.A.C.; van Wingerden, K.M.C. Sojourn-time approximations in a two-node queueing network. In *Proceedings 19th International Teletraffic Congress, ITC-19 (Beijing)*, August 2005.
4. Cohen, J.W. The multiple phase service network with generalized processor sharing. Acta Informatica **1979**, *12*, 245–284.
5. Disney, R.L.; Koenig, D. Queueing networks: a survey of their random processes. SIAM Rev. **1985**, *27*, 335–403.
6. Doshi, B.T.; Kaufman, J.S. Sojourn time in an M/G/1 queue with Bernoulli feedback. In *Queueing Theory and its Applications – Liber Amicorum for J.W. Cohen*; Boxma, O.J., Syski, R., Eds.; North-Holland: Amsterdam, 1988; 207–233.
7. Foley, R.D.; Disney, R.L. Queues with delayed feedback. Advances in Applied Probability **1983**, *15*, 162–182.
8. Gijsen, B.M.M.; van der Mei, R.D.; Engelberts, P.; van den Berg, J.L.; van Wingerden K.M.C. Sojourn-time approximations in queueing networks with feedback. Performance Evaluation **2006**, *63*, 743–758.
9. Takács, L. A single-server queue with feedback. The Bell System Technical Journal **1963**, *42*, 505–519.
10. van den Berg, J.L. Sojourn times in Feedback and Processor Sharing Systems. PhD thesis, University of Utrecht, The Netherlands, 1990. A copy of this thesis is available from the author upon request.
11. van den Berg, J.L.; Boxma, O.J. The M/G/1 queue with processor sharing and its relation to a feedback queue. Queueing Systems **1991**, *9*, 365–402.

12. van der Mei, R.D.; Gijsen, B.M.M.; in 't Veld, N.; van den Berg, J.L. Response times in a two-node queueing network with feedback. Performance Evaluation **2002**, *49*, 99–110.
13. van der Mei, R.D.; Meeuwissen, H.B. Modelling end-to-end quality-of-service for transaction-based services in a multi-domain environments. In *Proceedings IEEE International Conference on Web Services, ICWS (Chicago)* pp. 453–460, September 2006.
14. Vogels, W. Learning from the Amazon technology platform. ACM Queue **2006**, *4*, 14–22.