A Method for Approximating the Variance of the Sojourn Times in Star-Shaped Queueing Networks

R.D. van der Mei[a,b], A.R. de Wilde[a], and S. Bhulai[a,b]

[a]CWI, Probability and Stochastic Networks
P.O. Box 94079, 1098 SJ Amsterdam, The Netherlands
R.D.van.der.Mei@cwi.nl

[b]VU University, Faculty of Sciences
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
sbhulai@few.vu.nl

Abstract

We study the sojourn times in an open star-shaped queueing network, with a central processor-sharing (PS) node 0 and $M \geq 1$ multi-server First-Come-First Served (FCFS) nodes. Each customer alternatingly visits the central node and one of the other nodes, in the order $0, 1, 0, 2, \ldots, 0, M, 0$, 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: queueing networks, sojourn time, response time, approximation, variability

1 Motivation and Background

The emergence of Web technology has boosted the development of services running in a distributed computing environment, where 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 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 paying EU). For the user-perceived responsiveness of the service, both the mean and variability of the response times are of main importance [12].

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 able to the 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 (see Remark 4.3 for more details). This naturally leads to the problem formulation studied in this paper (see [11]).

Motivated by these observations, 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 Remark 4.1 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 (PF) solution, which immediately leads to a closed-form expression for the $\mathbb{E}[S]$, 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 (higher moments, tail probabilities); see [1] for a survey of the available results on sojourn times in queueing networks. The main result in [1] 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., where 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 have received service. 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. Disney and Koenig [4] give an overview on Bernoulli feedback models. Van den Berg and Boxma [9] consider an M/G/1 system, with either FCFS or PS service, where a customer after receiving service for the i-th 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 delay 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 [6] study queueing systems with delayed feedback, but their focus is merely on queue-length processes, busy periods and several customer flow processes. For star-shaped queueing networks with Markovian routing that possess a PF-solution, Gijsen et al. [10] propose an approximation method for the variance of the sojourn times; in the present paper, where we consider the same model but with deterministic routing, we build upon the insights obtained in [10]. For queueing-network models that do not admit a PF-solution, hardly any exact results on the sojourn-times are known. As an exception, Boxma et al. [2] 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 [9] 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 (discussed above), appears to be very effective (see Section 4 for details), and seems to be applicable for a much broader class of queueing networks than the one analyzed in this paper (see Remark 4.4). 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. This makes the methods developed in this paper valuable from an application point of view (see Remark 4.3 for further comments on the applicability of the model). These observations make the added value of the current paper 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, and subsequently, we develop an approximation for the variance of the sojourn times. In Section 4 the accuracy of the approximations is tested extensively by comparing the performance predictions based on the approximations with simulation results. Finally, in Section 5 we address a number of challenging topics for further research.

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 $B_{\rm PS}$ and with finite first two moments $\beta_{\rm PS}$ and $\beta_{\rm PS}^{(2)}$, respectively. The service time at FCFS-node k is exponentially distributed with mean β_k for $k = 1, \ldots, M$ (see Remark 4.1). The service times at all nodes are mutually independent. The load at the PS-node and at FCFS-node k is given by

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

respectively. Then, as we define $S_{\rm PS}^{(i)}$ as the sojourn time of the i-th 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

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

To ensure stability at each of the nodes, we assume $\rho_{\rm PS} < 1$ 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 $\mathbb{V}{\rm ar}[S]$. The expression for $\mathbb{E}[S]$ follows directly from the fact that the model possesses a product-form solution. However, an exact expression for $\mathbb{V}{\rm ar}[S]$ can not be obtained; therefore, we develop simple approximate expressions for $\mathbb{V}{\rm ar}[S]$.

3.1 Mean Sojourn Times

The network model under consideration possess a product-form (PF) solution. That is, if we define $L_{\rm PS}$ and L_k to be the stationary number of customers at the PS-node and at the k-th FCFS-node, respectively, we have

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

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

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

Then, using Little's Law we get

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

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

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

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

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

and

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

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

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

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

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

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:

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

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:

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

To obtain an approximation for $\mathbb{V}{\rm ar}[S]$ 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: $\mathbb{C}ov(S_i, S_j) = 0$ 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: $\mathbb{C}ov(S_{PS}^{(i)}, S_j) = 0$ 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 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) implement 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.

Remark 3.1 (Capturing correlations between the sojourn times at the PS-node) *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., $S_{PS}^{(i)}$, $i = 1, \ldots, M$) 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; note that this is the case since we do not take heavy-tailed service-time distributions into account, but instead consider service-time distributions with finite second moments. 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., $S_{PS}^{(i)}$, $i = 1, \ldots, M$) 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 (to be discussed in Section 4 below) that the inclusion of these correlations via AA1 is highly effective.*

To approximate the variance of the sojourn times in the PS-node, we use the approximate expression in [9] 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 $c_{M/G/1}^2$. For $\rho < 1$,

$$\begin{aligned}\mathbb{E}[S_{M/G/1}^2] \approx\ & c_{M/G/1}^2\left(1 + \frac{2+\rho}{2-\rho}\right)\frac{\beta_{M/G/1}^2}{(1-\rho)^2} \\ & + (1 - c_{M/G/1}^2)\left(\frac{2\beta_{M/G/1}^2}{(1-\rho)^2} - \frac{2\beta_{M/G/1}^2}{\rho^2(1-\rho)}(e^\rho - 1 - \rho)\right).\end{aligned} \tag{14}$$

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 $c_{M/G/1}^2$ 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 $(M+1)\beta_{\rm PS}$ and squared coefficient of variation $c_{\rm PS}^2/(M+1)$. Hence, the second moment of the total sojourn time at the PS-node, $S_{\rm PS} := \sum_{i=1}^{M+1} S_{\rm PS}^{(i)}$, is approximated by the following expression:

$$\begin{aligned}\mathbb{E}[S_{\rm PS}^2] &\approx \frac{c_{\rm PS}^2}{M+1}\left(1 + \frac{2+\rho_{\rm PS}}{2-\rho_{\rm PS}}\right)\left(\frac{(M+1)\beta_{\rm PS}}{1-\rho_{\rm PS}}\right)^2 \\ &\quad + \left(1 - \frac{c_{\rm PS}^2}{M+1}\right)\left(\frac{2((M+1)\beta_{\rm PS})^2}{(1-\rho_{\rm PS})^2} - \frac{2((M+1)\beta_{\rm PS})^2}{\rho_{\rm PS}^2(1-\rho_{\rm PS})}(e^{\rho_{\rm PS}} - 1 - \rho_{\rm PS})\right) \\ &= (M+1)c_{\rm PS}^2\left(1 + \frac{2+\rho_{\rm PS}}{2-\rho_{\rm PS}}\right)\left(\frac{\beta_{\rm PS}}{1-\rho_{\rm PS}}\right)^2 \\ &\quad + \left((M+1)^2 - (M+1)c_{\rm PS}^2\right)\left(\frac{2\beta_{\rm PS}^2}{(1-\rho_{\rm PS})^2} - \frac{2\beta_{\rm PS}^2}{\rho_{\rm PS}^2(1-\rho_{\rm PS})}(e^{\rho_{\rm PS}} - 1 - \rho_{\rm PS})\right),\end{aligned} \tag{15}$$

where $c^2_{\rm PS}$ 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

$$\mathbb{V}\mathrm{ar}\Big[\sum_{i=1}^{M+1} S^{(i)}_{\rm PS}\Big] = \mathbb{E}[S^2_{\rm PS}] - \big((M+1)\mathbb{E}[S^{(1)}_{\rm PS}]\big)^2. \tag{16}$$

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

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

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 $\mathbb{V}\mathrm{ar}[S]$:

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

In the next section the accuracy of this approximation is evaluated.

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 to ensure that the accuracy of the simulations themselves are within 1% is the real value; for compactness of the presentation, the confidence intervals are not shown here. 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 Remark 3.1 that in general the successive visits to the PS-node *are* dependent):

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

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 backend 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 $\beta_1, \ldots, \beta_5$. The seventh column shows the "exact" values of $\mathbb{V}\mathrm{ar}[S]$ obtained via simulations (denote $\mathbb{V}\mathrm{ar}_s[S]$), the eighth column shows $\mathbb{V}\mathrm{ar}_a[S]$ according to (18), and the ninth column shows the corresponding relative error defined in (19). The tenth and eleventh column show the results for the "simple" approximation defined in (20), and the corresponding relative error.

β_{PS}	$beta_1$	β_2	β_3	β_4	β_5	$\mathbb{V}\mathrm{ar}_s[S]$	$\mathbb{V}\mathrm{ar}_a[S]$	$\Delta\%$	$\mathbb{V}\mathrm{ar}_{simple}[S]$	$\Delta\%$
0.05	0.2	0.2	0.2	0.2	0.2	0.41	0.39	−5.65	0.60	44.87
0.05	0.5	0.5	0.5	0.5	0.5	5.16	5.08	−1.70	5.28	2.32
0.05	0.8	0.8	0.8	0.8	0.8	82.82	80.08	−3.31	80.28	−3.06
0.05	0.1	0.3	0.5	0.7	0.9	87.37	87.72	0.39	87.92	0.63
0.05	0.9	0.7	0.5	0.3	0.1	89.57	87.72	−2.07	87.92	−1.83
0.05	0.5	0.1	0.7	0.3	0.9	92.83	87.72	−5.51	87.92	−5.28
0.05	0.8	0.8	0.8	0.8	0.8	79.61	80.08	0.59	80.28	0.85
0.15	0.1	0.3	0.5	0.7	0.9	171.74	181.40	5.63	87.92	−48.80
0.15	0.9	0.7	0.5	0.3	0.1	186.44	181.40	−2.70	87.92	−52.84
0.15	0.5	0.1	0.7	0.3	0.9	183.25	181.40	−1.01	87.92	−52.02
0.15	0.1	0.3	0.5	0.7	0.9	173.06	181.40	4.82	87.92	−49.20

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

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 backend nodes

To evaluate the accuracy of our approximation for multi-server FCFS-nodes, we first consider the model with three symmetric multi-server queues with $c_1 = c_2 = c_3 = c$. As before, $\lambda = 1$, and the PS-node and FCFS-nodes have exponentially distributed service times.

β_{PS}	β_F	c	$\mathbb{V}\text{ar}_s[S]$	$\mathbb{V}\text{ar}_a[S]$	$\Delta\mathbb{V}\text{ar}\%$	$\mathbb{V}\text{ar}_a[S]$ (simple)	$\Delta\mathbb{V}\text{ar}\%$ (simple)
0.05	0.4	2	0.53	0.53	−0.05	0.52	−2.73
0.10	1.0	2	5.03	4.94	−1.86	4.72	−6.15
0.20	1.6	2	67.82	69.82	2.95	51.98	−23.35
0.05	1.6	2	53.91	51.70	−4.11	51.69	−4.13
0.20	0.2	2	17.64	18.26	3.51	0.43	−97.58
0.05	0.8	4	1.94	1.95	0.61	1.94	−0.12
0.10	2.0	4	13.22	13.22	0.04	13.01	−1.59
0.20	3.6	4	284.82	289.07	1.49	271.24	−4.77
0.05	7.2	8	373.58	376.90	0.89	376.88	0.88
0.10	1.6	8	7.85	7.95	1.24	7.73	−1.51
0.20	4.0	8	62.89	66.49	5.72	48.65	−22.64

Table 2: Sojourn time variances with three symmetric multi-server FCFS-nodes.

As shown in Table 2, also for symmetric multi-server queues our approximation performs well. To extend these results we consider in Table 3 the case of multi-server nodes with unequal numbers of servers. And as expected the relative error of this case is still very low, with errors typically less than 4%.

β_{PS}	β_F			(c_1, c_2, c_3)			$\mathbb{V}\text{ar}_s[S]$	$\mathbb{V}\text{ar}_a[S]$	$\Delta\%$	$\mathbb{V}\text{ar}_a[S]$ (simple)	$\Delta\%$ (simple)
0.05	0.1	0.4	0.9	1	2	3	1.04	1.04	0.25	1.03	−1.12
0.10	0.4	1.2	3.2	1	2	3	24.21	23.49	−2.98	23.28	−3.87
0.20	0.6	0.8	0.6	1	2	3	20.71	21.57	4.18	3.74	−81.95
0.05	3.2	1.8	0.8	4	3	2	28.68	29.03	1.23	29.02	1.18
0.10	1.6	2.5	0.2	2	5	1	23.62	24.05	1.82	23.84	0.91
0.20	2.7	3.2	3.5	3	4	5	148.02	142.94	−3.43	125.11	−15.48

Table 3: Sojourn time expectations and variances with three asymmetric multi-server FCFS-nodes.

The question arises how this formula performs for non-exponential service-time distributions at the PS-node. Table 4 shows the results of the simulated and approximated values of $\mathbb{V}\text{ar}[S]$ for

a variety of parameters, where the service times of the FCFS-node are exponentially distributed and the squared coefficient of variation $c_{\rm PS}^2$ 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 $c_{\rm PS}^2 = 0$ and gamma-distributed for the cases $c_{\rm PS}^2 = 4$ and $c_{\rm PS}^2 = 16$. To obtain $c_{\rm PS}^2 = 4$, we use a gamma distribution with shape parameter $\gamma = 1/4$ and scale parameter $\lambda = 5/2$ for $\beta_{\rm PS} = \gamma/\lambda = 0.1$, and we use a gamma-distribution with $\gamma = 1/4$ and $\lambda = 5/6$ for $\beta_{\rm PS} = \gamma/\lambda = 0.3$. To obtain $c_{\rm PS}^2 = 16$, we use a gamma distribution with $\gamma = 1/16$ and $\lambda = 5/8$ for $\beta_{\rm PS} = \gamma/\lambda = 0.1$, and we use a gamma distribution with $\gamma = 1/16$ and $\lambda = 5/24$ for $\beta_{\rm PS} = \gamma/\lambda = 0.3$.

$\beta_{\rm PS}$	$c_{\rm PS}^2$	β_F		$\mathbb{V}{\rm ar}_s[S]$	$\mathbb{V}{\rm ar}_a[S]$	$\Delta\%$	$\mathbb{V}{\rm ar}_a[S]$ (simple)	$\Delta\%$ (simple)
0.1	0	0.1	0.9	82.15	81.05	−1.33	81.01	−1.38
0.3	0	0.8	0.5	85.89	86.81	1.06	17.12	−80.06
0.1	0	0.5	0.3	1.25	1.22	−1.76	1.19	−4.87
0.3	0	0.9	0.1	147.49	150.82	2.25	81.14	−44.99
0.1	4	0.1	0.9	80.83	81.33	0.62	81.17	0.42
0.3	4	0.8	0.5	274.00	278.46	1.63	19.61	−92.84
0.1	4	0.5	0.3	1.54	1.50	−2.68	1.34	−13.15
0.3	4	0.9	0.1	331.30	342.47	3.37	83.62	−74.76
0.1	16	0.1	0.9	81.55	82.16	0.75	81.63	0.10
0.3	16	0.8	0.5	831.15	853.41	2.68	27.07	−96.74
0.1	16	0.5	0.3	2.37	2.33	−1.86	1.80	−24.19
0.3	16	0.9	0.1	871.49	917.42	5.27	91.09	−89.55

Table 4: Sojourn time variances for a queuing network with general service times at the PS-node and two single-server FCFS nodes.

We observe in Table 4 that our approximation also keeps standing for general service times at the PS-node. To extend these results Table 5 presents the results of almost exactly the same network, with as exception that both FCFS-nodes become $M/M/c$ nodes. The first FCFS-node has two servers and the second FCFS-node has three servers, so $c_1 = 2$ and $c_2 = 3$. The results show that for these cases the approximation given in (18) is still accurate.

To summarize, the numerical results presented in Tables **??**–5 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 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 Remark 3.1), and (2) the idea of neglecting all other cross-correlations between

$\beta_{\rm PS}$	$c^2_{\rm PS}$	β_F		$\mathbb{V}{\rm ar}_s[S]$	$\mathbb{V}{\rm ar}_a[S]$	$\Delta\mathbb{V}{\rm ar}\%$	$\mathbb{V}{\rm ar}_a[S]$ (simple)	$\Delta\mathbb{V}{\rm ar}\%$ (simple)
0.1	0	0.2	2.7	80.82	85.66	5.99	85.62	5.94
0.3	0	1.6	1.5	88.82	89.70	0.99	20.02	−77.46
0.1	0	1.0	0.9	2.43	2.43	−0.02	2.39	−1.61
0.3	0	1.8	0.3	149.31	152.38	2.05	82.69	−44.62
0.1	4	0.2	2.7	88.50	85.94	−2.90	85.78	−3.08
0.3	4	1.6	1.5	272.00	281.35	3.44	22.50	−91.73
0.1	4	1.0	0.9	2.71	2.71	−0.23	2.55	−6.18
0.3	4	1.8	0.3	330.14	344.03	4.21	85.18	−74.20
0.1	16	0.2	2.7	89.60	86.77	−3.16	86.24	−3.76
0.3	16	1.6	1.5	820.26	856.30	4.39	29.97	−96.35
0.1	16	1.0	0.9	3.57	3.54	−0.79	3.01	−15.66
0.3	16	1.8	0.3	920.45	918.98	−0.16	92.64	−89.93

Table 5: Sojourn time variances for a queueing network with general service times at the PS-node and with two asymmetrically multi-server FCFS-nodes.

pairs of visits to the nodes.

We conclude this section with a number of remarks.

Remark 4.1 (Non-exponential service times at the FCFS nodes)
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 (PF) 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 PF 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 [2] for pioneering contributions in that direction.*

In this context, it should also be noted that, in practice, for many applications the information services are expected to handle huge amounts of transactions per time unit, and can process many transactions simultaneously, which means that the numbers of servers at the FCFS-nodes c_k ($k = 1, \ldots, M$) are often very large. For multi-server nodes it is well-known that the probability that an incoming customer finds at least one of the servers available is high (based on the principle of economies of scale), and hence, that multi-server FCFS-nodes can be well-approximated by infinite-

server nodes, in which case a PF solution does exist for generally distributed service times; in fact, even insensitivity holds with respect to the service-time distributions. Note also that by assuming $c_k = \infty$ $(k = 1, \ldots, M)$*, the right-hand side of equation (17), and hence of (18), is simplified to* $\sum_{k=1}^{M} \left(\beta_k^{(2)} - \beta_k^2\right)$*, which for the special case of exponential service times can be further simplified to* $\sum_{k=1}^{M} \beta_k^2$*, where* $\beta_k^{(2)}$ *stands for the second moment of the service-time distribution at FCFS-node* k $(k = 1, \ldots, M)$*. Hence, from an application point of view, the assumption of exponential service times at the FCFS-nodes may be negligible, especially if the information servers implement a high level of concurrent processing.*

To quantify how many servers are needed for the assumption of $c_k = \infty$ *to make sense, we have found by trial and error that approximating a multi-server FCFS-node by an infinite-server node leads to rather accurate approximations if the probability that a customers has to wait (i.e.,* $\pi_k, k = 1, \ldots, M$*) does not exceed 0.05. For example, we found by using the classical results for the* $M/M/c_k$*-queue that when the mean service times at FCFS-node* k *is* $\beta_k = 1$*, the maximum load-per-server* ρ_k *varies as follows as a function of the number of servers: if* $c_k = 1$ *then* $\rho_k^{(\max)} = 0.05$*, if* $c_k^{(\max)} = 5$ *then* $\rho_k^{(\max)} = 0.38$*, if* $c_k = 10$ *then* $\rho_k^{(\max)} = 0.53$*, if* $c_k = 50$ *then* $\rho_k^{(\max)} = 0.77$ *and if* $c_k = 100$ *then* $\rho_{(\max)} = 0.83$*. Thus, for example, if an FCFS-node can handle 50 customers in parallel, the assumption that* $c_k = \infty$ *is fairly realistic when the load-per-server does exceed 77%, which is a very high threshold value that is met in most practical situations. These observations indicate that in many cases the assumption of exponential service times does pose a main restriction on the applicability of the results.*

Remark 4.2 (Discussion of the sources of inaccuracy)
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 [9], 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 [8] 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 [9] on immediate feedback may become inaccurate. Finally, assumptions AA2 and AA3 may be 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.

Remark 4.3 (Application of the model in multi-provider service environments)
The model analyzed in this paper is also motivated from large-scale applications that run in multi-provider service environments, an area that will experience dramatic growth over the next few years. The emergence of Web Services and Service-Oriented Architectures (SOAs) allows services to be built on top of existing services (e.g., location-aware services that are built on top of location services for mobile networks), each of which may be owned by different third parties. In this context, to deliver good end-to-end quality to its customers, an application service provider (ASP) typically negotiates Service Level Agreements (SLAs) with these third parties. Such SLAs typically include statistical response-time guarantees in return to a maximum number of parallel information-retrieval requests that can be handled in parallel; request in excess of this maximum may be temporarily buffered. In this multi-provider situation, for the ASP it is essential to know not only what are the end-to-end response times observed by its paying customers under anticipated load scenarios, but also what combinations of SLAs *with third parties are needed to meet requirements on the end-to-end mean and the variability of the user-perceived response times. This set of possible combinations of SLAs that lead to a desired end-to-end performance level is referred to as the* SLA negotiation space *[11]. To identify the SLA negotiation space, the relation is needed between combinations of SLAs on the one hand and the end-to-end response times on the other hand. To this end, the model discussed in this paper can be applied by incorporating the SLA parameters into the model (e.g., by relating the maximum number of parallel connections to the number of servers at an FCFS-node, and by relating the statistical response-time guarantee to the service times to the corresponding FCFS node). The reader is referred to [11] for a detailed discussion on this. In this context, the approximations proposed in this paper are of great practical interest.*

Remark 4.4 (Extension to a broader class of models)
The method to derive the approximation of the variance of the sojourn times for the current model, with a central PS node and multiple FCFS nodes, can be easily extended to a broader class of star-shaped models. The assumptions AA1–AA3 allow one to replace some of the FCFS nodes by PS nodes (with general service-time distributions). One can even replace the FCFS nodes by any service discipline that falls into the class of BCMP networks (e.g., the infinite server node and the generalized PS [3]) such that the product form is preserved. However, one may need to develop

approximations for the second moment of the sojourn times of such nodes similar to (14) for the case of processor sharing.

5 Topics for Further Research

The method presented in this paper shows a remarkable accuracy, and the basic ideas behind method proposed in this paper seem to be applicable to a much broader class of queueing networks than the one studied in this paper. Therefore, it is interesting to investigate to what extent the approximation technique can be generalized to different variants of the model. First, both from an application and queueing theoretical point of view it would be interesting to extend the results to obtain approximate expressions for tail probabilities and higher moments of the total sojourn times. In principle, such an approximation can be based on assumptions AA1–AA3. In this context, note that the marginal sojourn-time distributions for the FCFS nodes can be obtained from (7)–(10). However, one is required to develop new approximations for the complete distribution of the total sojourn time in the PS node, which in turn requires the development of approximate expressions for the sojourn time in an $M/G/1$ PS model, extending (14).

Second, the star shape of the model which was mainly motivated from applications in which a central application server implements the business logic retrieving and parsing information from different information services may be generalized in different ways, including for example a multi-layered tree of services and multiple central nodes. This could lead to multiple visits to the same queue, which may be handled by using results of feedback queues and perhaps extensions of [5] and [7] for the single-server queue, where the main challenge is to extend these results to multi-server queues.

Third, the assumption that the service times at the FCFS-nodes are exponentially distributed may be relaxed, which is particularly interesting from a queueing-theoretical perspective (see also Remark 4.1). Recall that in that case the model no longer possesses a PF solution, so that even exact for the mean sojourn times are generally unknown.

Finally, 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.

References

[1] O.J. Boxma and H. Daduna. Sojourn times in queueing networks. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, pages 401–450. North Holland,

1990.

[2] O.J. Boxma, R.D. van der Mei, J.A.C. Resing, and K.M.C. van Wingerden. Sojourn-time approximations in a two-node queueing network. In *Proceedings 19th International Teletraffic Congress, ITC-19 (Beijing)*, August 2005.

[3] J.W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12:245–284, 1979.

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

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

[6] R.D. Foley and R.L. Disney. Queues with delayed feedback. *Advances in Applied Probability*, 15:162–182, 1983.

[7] L. Takács. A single-server queue with feedback. *The Bell System Technical Journal*, 42:505–519, 1963.

[8] J.L. van den Berg. *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.

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

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

[11] R.D. van der Mei and H.B. Meeuwissen. Modelling end-to-end quality-of-service for transaction-based services in a multidomain environments. In *Proceedings IEEE International Conference on Web Services, ICWS (Chicago)*, pages 453–460, September 2006.

[12] W. Vogels. Learning from the Amazon technology platform. *ACM Queue*, 4, 2006.