End-to-End Response Times in Multi-Provider Service Environments

Arald de Wilde

July 2006

Supervisor CWI:
Prof. dr. Rob van der Mei

Supervisors VU:
Dr. Sandjai Bhulai
Dr. Wojtek Kowalczyk

Preface

This thesis is a result of six months work at the Centrum voor Wiskunde en Informatica (CWI), the national research institute of mathematics and computer science in Amsterdam, the Netherlands. These months reflect the last part of my Master study Business Mathematics and Informatics at the Vrije Universiteit Amsterdam (VUA). Beside a feeling of happiness caused by finshing my study, I cannot deny there is another feeling because I leave people who helped me through this last part of my study. So I want to thank some people.

First of all I want to thank Rob van der Mei who has been my supervisor at CWI. His very active way of explaining difficult stuff, his enthusiasm for queueing networks and his flexible way of coordinating my research has resulted in interesting work every day. Beside Rob, I want to thank the whole crew of PNA 2. I experienced a very nice atmosphere during all the months I have been involved in this group. That is why I want thank my colleagues for being a office mate, for explaining Extend, for teaching me salsa dancing, for lunches, for good talks and for the sometimes remarkable but always nice coffee breaks. Michel, Sindo, Sem, Ton, Regina, Lasse, Pascal, Chrétien, Maaike, Wemke, Matthieu, Urtzi: Thanks!!

Since Sandjai Bhulai and Wojtek Kowalczyk were my supervisors at the VUA these six months, I want to thank them too. Wojtek Kowalczyk for being my second reader and Sandjai Bhulai for his enthusiasm to read this thesis and for his smart suggestions.

Beside these people I want to thank God for helping me through all parts of my life and in this case especially for giving me some insight in queueing theory.

Arald de Wilde

Amsterdam, July 2006

Contents

Chapter 1

Introduction

The dramatic growth of the Internet, the popularity of PCs and the emergence of mobile communications have boosted the development of a wide variety of Web-based services. Typical examples of such services that are offered today are PC banking, on-line gaming applications and airline ticket reservations. And in the near future a variety of new on-line services will be offered including, for example, E-Health applications that enable doctors to access medical data of their patient at any time and any place, and on-line access to Video-on-Demand services that enable the consumer to watch any movie at any time. A typical feature of this type of on-line applications is that a single transaction initiated by the end user induces a sequence of server and database transactions. Therefore, a key factor for the success of this type of services is that the response times observed by the end user are not prohibitively long. This raises the need for the development of quantitative models to analyze and predict the response times experienced by the end-user under anticipated load scenarios.

1.1 Background

Today's service offerings are increasingly based upon combining and integrating information from multiple logically and geographically distributed servers, interconnected by communication networks. Typical examples are location-based information services, where first the geographical location is determined, and subsequently, the information corresponding to this location is retrieved (e.g., restaurants, hotels, weather forecasts). To realize this type of transaction-based services, the service provider (SP) needs to make agreements with other parties involved, including access network operators to provide wireless access, location service provider to give the user's location, and content providers to deliver the requested local information. From the end-user's perspective, the SP is responsible for the billing and the proper functioning of the service. In this environment, the service is offered via a multitude of administrative domains each owned by different parties. This raises the need for SPs to control for guarantee *end-to-end* Quality of Service (QoS) perceived by the paying customer. Such an end-to-end QoS depends on the per-domain QoS. So, since domains are owned by different parties, typically bilateral Service Level Agreements (SLAs) are negotiated. From the perspective of the SP, the key question during such negotiations is: *"What combination of SLAs with other domains leads to desired end-to-end QoS perceived by the end-user?"*. Currently, no satisfactory answers exist for this question. In current practice, the main focus is often on the short time-to-market, i.e., to make the service operational as fast as possible, while performance-related issues are tackled on an ad-hoc basis. However, to avoid customer dissatisfaction, potential performance problems, for example, caused by strong growth of usage, should be anticipated on, in order to timely take appropriate measures (server upgrades, bandwidth upgrades, and modifications to SLAs). This motivates the development of models and techniques to address what-if questions regarding performance under different evolution scenarios, explicitly incorporating the effect of particular parameter choices in the SLAs on the end-to-end

performance in terms of end-user perceived QoS.

1.1.1 Running Example

Throughout the report we consider the Location-based Restaurant Service (LRS) given in [19] as a running example. We use this LRS since it includes the relevant aspects of transaction-based services running in a multi-domain environment. The LRS provides a mobile end user with a list of restaurants in the neighbourhood that meets the user's personal preferences. An LRS service request proceeds along the following steps:

1. The end user uses a mobile device to request suitable restaurants. This typically generates an HTTP request from the mobile device to the application server over the access network.

2. The application server processes the request and sends a location request to the location server. The location server determines the location of the end user and returns the location coordinates to the application server.

3. The application server processes this response and sends a request for restaurants that meet the user's preferences in the neighbourhood of the end user's location to the restaurant server. The restaurant server uses this information to identify a list of suitable restaurants and returns this list to the application server.

4. The application server processes this response, builds an HTML page and sends it to the user as the reply to the HTTP request.

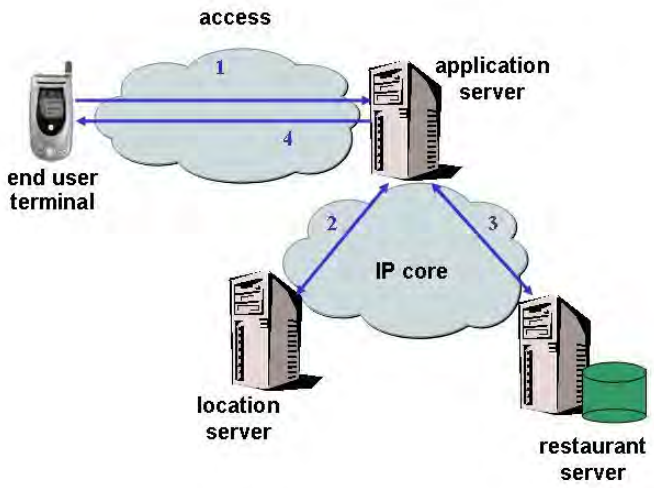

Figure 1.1: Location-based Restaurant Service.

A graphical interpretation of these steps is given in Figure 1.1. An important characteristic of the LRS is that it crosses multiple administrative domains and that multiple parties are involved, each with their own business incentives. The parties involved are the LRS service provider, typically the owner of the application server, the different network providers, the location service provider and the restaurant service provider as presented by Figure 1.2.

It is important to note that in practice each of the stakeholders may be companies with their

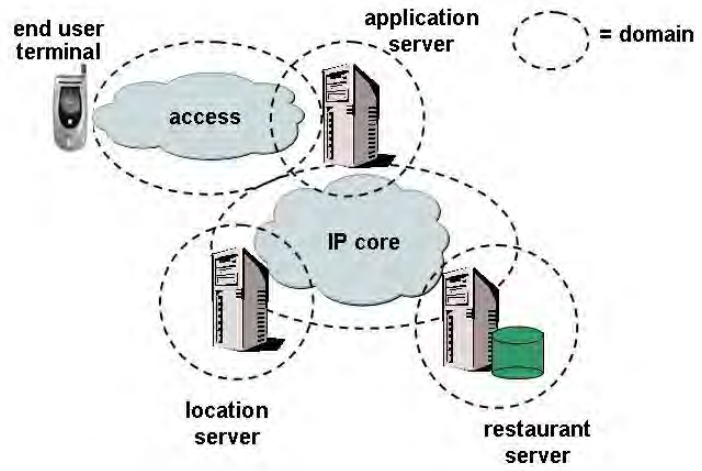

Figure 1.2: Multiple domains involved in the LRS.

own complex and possibly distributed infrastructures. In the running example discussed above, the LRS service provider negotiates SLAs with the other stakeholders without being concerned with how these stakeholders realize the service levels in the negotiated SLAs. Thus, it is the responsibility of the domain owners to realize the service levels agreed upon (e.g., which equipment, over-dimensioning or not, which service contracts, which sub-contractors).
This observation allows a hierarchical modelling framework that can be recursively applied at each abstraction level, consisting of a business party with its direct sub-contractors. Figure 1.3 shows an example tree structure of domain owners (parents in the tree) with SLA relations to their sub-contractors (children in the tree).
If a domain owner is involved in realizing the service, then it can meet the SLAs with its customers - which may be either end users or business customers - by properly operating its own domain in combination with negotiating the right SLAs with its sub-contractors. This raises the need for quantitative models and solution techniques that allow a domain owner to identify the relation between (a) the requirements on the performance of its own domain, (b) the SLAs with its sub-contractors, and (c) the service level offered to its own customers.

We distinguish between two types of SLAs: (a) SLAs with network domains, and (b) SLAs with service domains. Although several parameters can be considered in network SLAs (e.g., availability, network bandwidth, network latency, and packet delay, loss and jitter), in current practice network techniques have developed so well that almost no restrictions are required in the network domain. So we just focus on the performance of server domains. A typical example of such a service domain SLA has the following structure. On the one hand, the client application limits the request rate, and in return the service domain provides a statistical QoS guarantee. The request rate may be limited, for example, by putting a cap on the number of simultaneous TCP connections, or on the average (or maximum) number of requests over a given time interval. In return, the service domain may provide statistical QoS guarantees on ,for example, response times, download times, and availability.

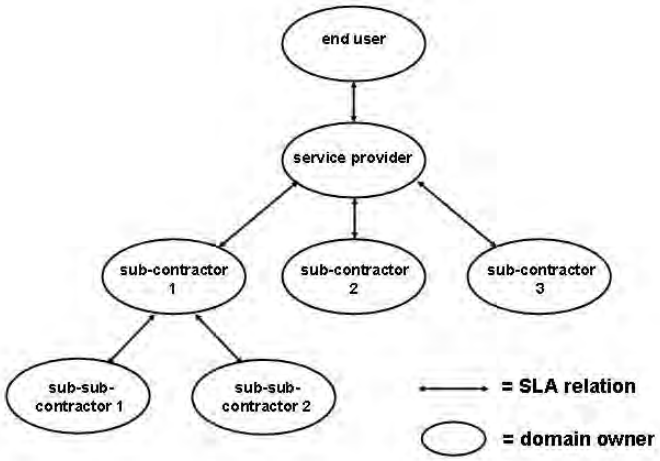

Figure 1.3: Example tree structure of domain owners with SLA relations to their sub-contractors.

1.1.2 Model Formulation

To capture the multi-domain infrastructure in a modelling framework, we map the infrastructure into a queueing network according to [19]. The main entities in such a network are jobs, queueing nodes, service time distributions of jobs per queueing node, and routing schemes. In our approach, a job represents a request and a domain consists of one or more queueing nodes of which the parameters are determined by the SLAs. It is obvious that many specific networks fit this general model. However, to illustrate this general framework, we discuss it for the specific LRS case. It consists of a network of three nodes, representing the application server, the location service and the restaurant service, see Figure 1.4.

Customers represent transactions and arrive at the application server (AS) with arrival rate λ_{AS}. Let us follow the processing steps experienced by a tagged customer T. First T requires service time $B_{AS,1}$ at the AS with mean $\beta_{AS,1}$. Then, T is forwarded to the LS, requiring service time B_{LS} at the LS with mean β_{LS}. Upon departure from the LS node, T returns to the AS, and requires service time $B_{AS,2}$ at the AS with mean $\beta_{AS,2}$. Subsequently, T is routed to the RS, requiring service time B_{RS} at the RS with mean β_{RS}. Next, T returns again to the AS, where it requires service time $B_{AS,3}$ at the AS with mean $\beta_{AS,3}$ before departing from the system. The AS is typically CPU-bound, and is therefore modelled as a Processor Sharing (PS) server; that is, when the server is handling $k > 0$ requests simultaneously, each of these k requests receives a fair share $1/k$ of the total processing capacity. The LS is modelled as a multi-server First Come First Serve (FCFS) node, where the number of servers, c_{LS}, represents the maximum number of simultaneous location lookup requests and the service times represent the response time negotiated in $\mathrm{SLA_{LS}}$. Similarly, the RS is also modelled as a multi-server FCFS node with c_{RS} parallel servers with service times representing the response time negotiated in $\mathrm{SLA_{RS}}$. In this model, the total sojourn time represents the end-to-end response time experienced by an end user of the LRS service, excluding the access network delay.

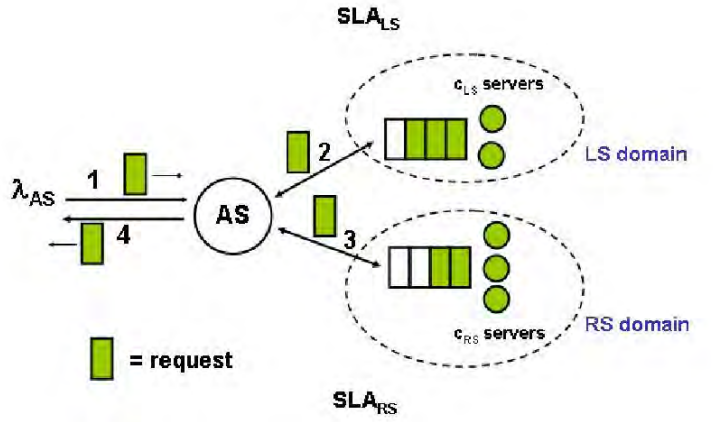

Figure 1.4: End-to-end queueing network model for the service domains.

1.2 Related Work

Much research to sojourn times is already done. Coffman et al. [5] obtain the Laplace-Stieltjes Transform (LST) of the sojourn time distribution for the $M/M/1$-PS node. Morrison derives in [14] an integral representation for this distribution and Ott derives in [15] the sojourn time distribution for the $M/G/1$-PS case where a representation of the probability distribution of the service time is needed. For FCFS nodes the main results which we will use are given by Takács in [17] where he obtains the LST and the first two moments of the total sojourn time in an $M/G/1$-FCFS queue with Bernouilli feedback. The first important results for calculating sojourn times in a queueing network are given by Jackson in [9] where he proposes the so-called product-form networks. To be a product-form network, a network has to fulfill the following conditions: (a) The arrival process is a Poisson process, the arrival rate has to be independent of the number of customers at the nodes in the network. (b) The service times are exponentially distributed, the service rate at a node may only depend on the number of customers in that particular node. (c) The next node visited may depend on the present node, but has to be otherwise independent of the state of the system. Within such Jackson networks, different subclasses are distinguished of which we will use some. An overview of results on sojourn times in queueing networks is given by Boxma and Daduna in [1]. They also give an expression for the LST of the joint probability distribution at the nodes of a customer that traverses a predefined path of nodes in a product-form network. Boxma et al. study in [3] response times in a two-node network with feedback. This work is an extension of the work done in [2]. They propose some solid approximations for the sojourn times with iterative requests. In [3] they show that these approximations perform very well for the first moment of the sojourn times. Gijssen et al. give in [8] exact results for the mean sojourn times and derive approximations for the variances of the sojourn times for a network with an $M/G/1$-PS node and several multi-server FCFS nodes with exponential service times.

Despite their business relevance, surprisingly few papers focus on the QoS via SLA-based solutions. For a recent survey on the state-of-the-art on SLA-based solutions for end-to-end QoS

problems, we refer to the work done by Sorteber and Kure in [16]. Applied to military purposes they distinguish the following two main solutions for SLA-management: an end-to-end solution, where the SLAs are directly negotiated with all parties involved, and a cascaded solution, where the SP only negotiates SLAs with its neighbouring domains. Van der Mei and Meeuwissen make in [19] the definition of SLAs for server domains more explicit by developing performance models to *quantify* the complex relation between SLAs and end-to-end QoS. They have implemented these models in a simulation tool, but it is obvious that queueing methods are preferred.

1.3 Goal and Structure of the Thesis

In view of the problems described above, the main goal of the thesis is to answer the following question:

"What combination of SLAs with other domains leads to desired QoS of the response time?"

In other words: What is the SLA negotiation space? To this end, we analyze the sojourn times in a variety of queueing networks. In cases where it is possible exact analyses are given, in other cases approximations are made.
The remainder of this thesis is organized as follows. In Chapter 2 we consider approximations for variances of sojourn times for a two-node network with as less as possible restrictions. In Chapter 3 we obtain exact expressions for the mean sojourn time, and we derive approximations of the variance of sojourn times in queueing networks with deterministic routing. In Chapter 4 we obtain expressions for mean sojourn times in a network with finite capacity, and with or without a buffer in front of the network. To come back to our running example and to make the investigated models less abstract, in Chapter 5 we apply some mathematical results to the running example given in Section 1.1.2 by calculating some negotiation spaces.

Chapter 2

Two-Node Queueing Networks

In this chapter we analyze the performance of a two-node queueing network with one processor sharing (PS) node and one single-server first-come first-served (FCFS) node. External customers arrive at the PS node according to a Poisson process with rate λ. A departing customer subsequently enters the FCFS node with probability p, and leaves the network with probability $1-p$. Upon departure from the FCFS node, a customer always returns to the PS node. A graphical representation of the network is given in Figure 2.1. All service times at all visits to both

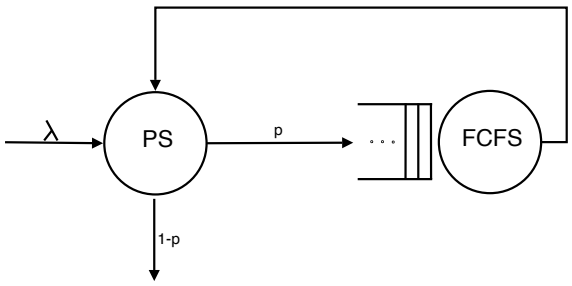

Figure 2.1: The two-node open queueing network with probabilistic routing.

nodes are independent random variables. The service time S_{PS} at the PS node has distribution function $B_{PS}(\cdot)$ and the service time S_F at the FCFS node has distribution function $B_F(\cdot)$ with first moment β_F, second moment $\beta_F^{(2)}$, and third moment $\beta_F^{(3)}$. The PS node has first moment $\mathbb{E}[S_{PS}] = \beta_{PS}$ and squared coefficient of variation c_{PS}^2 which is defined by

$$c_{PS}^2 = \frac{\mathbb{V}\mathrm{ar}[S_{PS}]}{\mathbb{E}[S_{PS}]^2}. \tag{2.1}$$

Since, next to the external arrivals, there are also internal arrivals due to jobs that are fed back with probability p, the mean arrival rate at the PS node is given by

$$\lambda_{PS} := \lambda + p\lambda + p^2\lambda + \cdots = \frac{\lambda}{1-p}.$$

The total load at the PS node is thus given by

$$\rho_{PS} = \frac{\lambda\beta_{PS}}{1-p}.$$

The arrival rate at the FCFS node can be expressed by

$$\lambda_F = \frac{p\lambda}{1-p},$$

so the total load at the FCFS node is given by

$$\rho_F = \frac{p\lambda\beta_F}{1-p}.$$

It is hard to obtain exact results for sojourn times in queueing networks if some form of overtaking occurs. Both the processor sharing and the feedback mechanism induce such overtaking. If the service times at the FCFS node are exponentially distributed, then the joint queue length distribution has a product form, and the mean queue length is easily obtained, yielding the mean sojourn times via Little's formula; but otherwise, even the mean sojourn times are not known. Analytically, the complexity of the problem of obtaining sojourn time results in queues with non-instantaneous feedback was discussed by Foley and Disney in [6].
The network we investigate is related to the work of Boxma et al. in [2, 3]. This work resulted in approximations for the Laplace-Stieltjes transform (LST) of the joint distribution of the total sojourn times. To assess the accuracy of these approximations the authors ran some experiments to compare the mean sojourn times of the simulations with the approximated mean sojourn times. In this chapter we extend these results by deriving the second moment of the total sojourn time.

2.1 Approximation Methods

To express the distribution of the total sojourn time, we want to approximate the Laplace-Stieltjes Transform (LST) of the joint distribution of the total sojourn time S_{PS} in the PS node and S_F in the FCFS node. Denote by $S_{PS}^{(j)}$ the total sojourn time of the first j visits to the PS node and denote by $S_F^{(j)}$ the total sojourn time of the first j visits to the FCFS node with $S_F^{(0)} = 0$. We can write for Re ω_1, Re $\omega_2 \geq 0$:

$$\Psi(\omega_1,\omega_2) = \mathbb{E}[e^{-\omega_1 S_{PS} - \omega_2 S_F}] = \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1 S_{PS}^{(k+1)} - \omega_2 S_F^{(k)}}]. \tag{2.2}$$

Note that taking $\omega_1 = \omega_2$ yields an expression for the total sojourn time S of a customer in the system. We saw already that due to overtaking it is extremely hard to obtain exact solutions, so we present three different approximation methods.

2.1.1 Method I: Independence Assumption (IA)

This approximation method is based on the following assumptions:

Assumption 1 $S_{PS}^{(k+1)}$ and $S_F^{(k)}$ are independent for $k = 0, 1, 2, \ldots$.

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

From (2.2) and Assumption 1 and 2a we obtain for Re ω_1, Re $\omega_2 \geq 0$:

$$\Psi(\omega_1,\omega_2) \approx \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}\mathbb{E}[e^{-\omega_2\sigma_F}]^k. \tag{2.3}$$

In particular, we find from (2.3) the expressions we need to compute the variance of the total sojourn time. The second moment of the sojourn time in the PS node can be obtained by

$$\begin{aligned}
\mathbb{E}[S_{PS}^2] &= \left.\frac{\partial^2\Psi}{\partial\omega_1^2}\right|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\frac{\partial^2}{\partial\omega_1^2}(\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1})\mathbb{E}[e^{-\omega_2\sigma_F}]^k\Big|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k(k+1)\frac{\partial}{\partial\omega_1}(\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^k\mathbb{E}[-\sigma_{PS}e^{-\omega_1\sigma_{PS}}])\mathbb{E}[e^{-\omega_2\sigma_F}]^k\Big|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k(k+1)(k\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k-1}\mathbb{E}[\sigma_{PS}e^{-\omega_1\sigma_{PS}}]^2 \\
&+ \mathbb{E}[e^{-\omega_1\sigma_{PS}}]^k\mathbb{E}[\sigma_{PS}^2e^{-\omega_1\sigma_{PS}}])\mathbb{E}[e^{-\omega_2\sigma_F}]^k\Big|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k(k+1)(k\mathbb{E}[\sigma_{PS}]^2+\mathbb{E}[\sigma_{PS}^2]) \\
&\approx \frac{2p}{(1-p)^2}\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2+\frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
&+ \frac{1-c_{PS}^2}{1-p}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}-\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right).
\end{aligned}$$

We derive the last term using the insensitivity result of sojourn times at the PS node proved by Ott [15]. We also use the approximation for the variance of the sojourn time at the PS node given by Van der Berg and Boxma [18] with c_{PS}^2 as the squared coefficient of variation of one service time. The approach of these authors is based on a linear interpolation between the cases of exponential and deterministic service times. Further we obtain the totals of the sums by

$$\begin{aligned}
\sum_{k=0}^{\infty}(1-p)p^k(k+1)k &= (1-p)p\sum_{k=0}^{\infty}(k+1)kp^{k-1} = (1-p)p\frac{d^2}{dp^2}\sum_{k=0}^{\infty}p^{k+1} \\
&= (1-p)p\frac{d^2}{dp^2}\sum_{k=0}^{\infty}p^k = (1-p)p\frac{d^2}{dp^2}\frac{1}{1-p} = \frac{2p}{(1-p)^2}, \qquad (2.4)
\end{aligned}$$

and, similarly,

$$\sum_{k=0}^{\infty}(1-p)p^k(k+1) = (1-p)\frac{d}{dp}\sum_{k=0}^{\infty}p^k = (1-p)\frac{d}{dp}\frac{1}{1-p} = \frac{1}{1-p}. \qquad (2.5)$$

The second moment of the sojourn time in the FCFS node can be obtained by

$$\begin{aligned}
\mathbb{E}[S_F^2] &= \left.\frac{\partial^2\Psi}{\partial\omega_2^2}\right|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}\frac{\partial^2}{\partial\omega_2^2}(\mathbb{E}[e^{-\omega_2\sigma_F}]^k)\Big|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}k\frac{\partial}{\partial\omega_2}(\mathbb{E}[e^{-\omega_2\sigma_F}]^{k-1}\mathbb{E}[-\sigma_Fe^{-\omega_2\sigma_F}])\Big|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}k\Big((k-1)\mathbb{E}[e^{-\omega_2\sigma_F}]^{k-2}\mathbb{E}[\sigma_Fe^{-\omega_2\sigma_F}]^2 \\
&+ \mathbb{E}[e^{-\omega_2\sigma_F}]^{k-1}\mathbb{E}[\sigma_F^2e^{-\omega_2\sigma_F}]\Big)\Big|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^kk((k-1)\mathbb{E}[\sigma_F]^2+\mathbb{E}[\sigma_F^2]) \\
&= \frac{2p^2}{(1-p)^2}\left(\beta_F+\frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2
\end{aligned}$$

$$+ \quad \frac{p}{1-p}\left[\beta_F^{(2)} + 2\left(\frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2 + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(3)}}{3\beta_F}\right].$$

We derive the last term by using the Pollaczek-Khintchine formula, as for example given in [20], and differentiate this once in $s = 0$ to derive the mean waiting time, and differentiate this twice in $s = 0$ to obtain the second moment of the waiting time. These derivatives can be obtained easily by rewriting the Pollaczek-Khintchine formula in such a way that it contains the LST of the residual service time. Note that the third moment of the service time is needed. We compute the total of the sums by

$$\begin{aligned}\sum_{k=0}^{\infty}(1-p)p^k(k-1)k &= p^2(1-p)\sum_{k=0}^{\infty}p^{k-2}(k-1)k = p^2(1-p)\frac{d^2}{dp^2}\sum_{k=0}^{\infty}p^k \\ &= p^2(1-p)\frac{d^2}{dp^2}\frac{1}{1-p} = \frac{2p^2}{(1-p)^2}, \end{aligned} \tag{2.6}$$

and

$$\begin{aligned}\sum_{k=0}^{\infty}(1-p)p^k k &= (1-p)p\sum_{k=0}^{\infty}p^{k-1}k = p(1-p)\frac{d}{dp}\sum_{k=0}^{\infty}p^k \\ &= p(1-p)\frac{d}{dp}\frac{1}{1-p} = \frac{p}{1-p}. \end{aligned} \tag{2.7}$$

The expectation of S_{PS} times S_F can be obtained by

$$\begin{aligned}\mathbb{E}[S_{PS}\,S_F] &= \left.\frac{\partial^2\Psi}{\partial\omega_1\partial\omega_2}\right|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\frac{\partial}{\partial\omega_1}(\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1})\frac{\partial}{\partial\omega_2}(\mathbb{E}[e^{-\omega_2\sigma_F}]^k)\Big|_{(0,0)} \\ &= \sum_{k=0}^{\infty}(1-p)p^k(k+1)\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^k\mathbb{E}[-\sigma_{PS}e^{-\omega_1\sigma_{PS}}] \\ &\quad\cdot\ k\mathbb{E}[e^{-\omega_2\sigma_F}]^{k-1}\mathbb{E}[-\sigma_F e^{-\omega_2\sigma_F}])\Big|_{(0,0)} \\ &= \sum_{k=0}^{\infty}(1-p)p^k(k+1)k\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F] \\ &= \frac{2p}{(1-p)^2}\frac{\beta_{PS}}{1-\rho_{PS}}\left(\beta_F + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right). \end{aligned}$$

Expressions for $\mathbb{E}[S_{PS}]$ and $\mathbb{E}[S_F]$ are derived in a similar way as the previous results and are given in (2.3) and (2.4) of [3]. Combining these expressions with the previous formulas results in an approximation for the variance of the total sojourn time:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[S_{PS}+S_F] &= \mathbb{E}[(S_{PS}+S_F)^2] - (\mathbb{E}[S_{PS}+S_F])^2 \\ &= \mathbb{E}[S_{PS}^2] + 2\mathbb{E}[S_{PS}\,S_F] + \mathbb{E}[S_F^2] - (\mathbb{E}[S_{PS}] + \mathbb{E}[S_F])^2 \\ &\approx \frac{2p}{(1-p)^2}\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 + \frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\ &\quad + \frac{1-c_{PS}^2}{1-p}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2} - \frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}} - 1 - \rho_{PS})\right) \\ &\quad + 2\left(\frac{2p}{(1-p)^2}\frac{\beta_{PS}}{1-\rho_{PS}}\left(\beta_F + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)\right) \\ &\quad + \frac{2p^2}{(1-p)^2}\left(\beta_F + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2 \end{aligned}$$

$$+ \quad \frac{p}{1-p}\left[\beta_F^{(2)} + 2\left(\frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2 + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(3)}}{3\beta_F}\right]$$

$$- \quad \left(\frac{\beta_{PS}}{(1-p)(1-\rho_{PS})} + \frac{p}{1-p}\left(\beta_F + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)\right)^2.$$

2.1.2 Method II: Short-Circuit Assumption (SC)

This approximation method is based on the following assumptions:

Assumption 1 $S_{PS}^{(k+1)}$ and $S_F^{(k)}$ are independent for $k = 0, 1, 2, \ldots$.

Assumption 2b $S_{PS}^{(k+1)}$ has the same distribution as $\sigma_{PS}^{(k)}$; the total sojourn time after k visits in the PS node short-circuited (i.e., with the FCFS node removed). Similarly, $S_F^{(k)}$ has the same distribution as $\sigma_F^{(k)}$; the total sojourn time after k visits in the FCFS nodes short-circuited (i.e., with the PS node removed).

From (2.2) and Assumptions 1 and 2b we obtain for Re ω_1, Re $\omega_2 \geq 0$:

$$\Psi(\omega_1, \omega_2) \approx \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]. \tag{2.8}$$

Using (2.8) we obtain the expressions we need tot compute the variance of the total sojourn time by

$$\begin{aligned}
\mathbb{E}[S_{PS}^2] &= \left.\frac{\partial^2\Psi}{\partial\omega_1^2}\right|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\left.\frac{\partial^2}{\partial\omega_1^2}(\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}])\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]\right|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k\left.\frac{\partial}{\partial\omega_1}(\mathbb{E}[-\sigma_{PS}^{(k+1)}e^{-\omega_1\sigma_{PS}^{(k+1)}}])\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]\right|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[(\sigma_{PS}^{(k+1)})^2e^{-\omega_1\sigma_{PS}^{(k+1)}}]\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]\Big|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[(\sigma_{PS}^{(k+1)})^2] \\
&\approx \sum_{k=0}^{\infty}\left[(1-p)p^k\frac{c_{PS}^2}{k+1}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{(k+1)\beta_{PS}}{1-\rho_{PS}}\right)^2\right. \\
&+ \left.(1-\frac{c_{PS}^2}{k+1})\left(\frac{2((k+1)\beta_{PS})^2}{(1-\rho_{PS})^2} - \frac{2((k+1)\beta_{PS})^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right)\right] \\
&= \sum_{k=0}^{\infty}(1-p)p^k(k+1)c_{PS}^2\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
&+ \sum_{k=0}^{\infty}(1-p)p^k(k+1)^2\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2} - \frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right) \\
&- \sum_{k=0}^{\infty}(1-p)p^k(k+1)c_{PS}^2\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2} - \frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right) \\
&= \frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2
\end{aligned}$$

$$+ \quad \frac{1+p-(1-p)c_{PS}^2}{(1-p)^2}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}-\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right). \tag{2.9}$$

Here we use the approximation of the second moment of the sojourn time in an $M/G/1$-PS node given by Boxma and Van den Berg in [18] where we see the $k+1$ visits together as a convolution of $k+1$ sojourn times. So the service time is $k+1$ times longer and the squared coefficient of variation must be divided by $k+1$. This can easily be shown by adapting (2.1):

$$\frac{(k+1)\mathbb{V}\mathrm{ar}[S_{PS}]}{((k+1)\mathbb{E}[S_{PS}])^2}=\frac{1}{k+1}c_{PS}^2, \tag{2.10}$$

for $k=0,1,\ldots$. The total sums are calculated using (2.4) and (2.5), which also can be used to calculate the following sum:

$$\begin{aligned}\sum_{k=0}^{\infty}(1-p)p^k(k+1)^2 &= \sum_{k=0}^{\infty}(1-p)p^k(k(k+1)+k+1)\\ &= \sum_{k=0}^{\infty}(1-p)p^k k(k+1)+\sum_{k=0}^{\infty}(1-p)p^k(k+1)\\ &= \frac{2p}{(1-p)^2}+\frac{1}{1-p}=\frac{1+p}{(1-p)^2}.\end{aligned}$$

The second moment of the total sojourn time in the FCFS node can be obtained as follows:

$$\begin{aligned}\mathbb{E}[S_F^2] &= \left.\frac{\partial^2\Psi}{\partial\omega_2^2}\right|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]\left.\frac{\partial^2}{\partial\omega_2^2}(\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}])\right|_{(0,0)}\\ &= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]\left.\frac{\partial}{\partial\omega_2}(\mathbb{E}[-\sigma_F^{(k)}e^{-\omega_2\sigma_F^{(k)}}])\right|_{(0,0)}\\ &= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]\mathbb{E}[(\sigma_F^{(k)})^2e^{-\omega_2\sigma_F^{(k)}}]\Big|_{(0,0)}\\ &= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[(\sigma_F^{(k)})^2]\\ &= p\sum_{k=1}^{\infty}(1-p)p^{k-1}\mathbb{E}[(\sigma_F^{(k)})^2]\\ &= p\frac{(1-p)^2-2(1-p)}{6((1-p)-p\lambda\beta_F)^2((1-p)^2-(1-p)(2+p\lambda\beta_F)+p\lambda\beta_F)}\\ &\cdot \left[2(1-p)(6p\lambda\beta_F^3-6\beta_F^2-6p\lambda\beta_F\beta_F^{(2)}+3\beta_F^{(2)}+p\lambda\beta_F^{(3)})\right.\\ &- \left.(12p\lambda\beta_F^3-12\beta_F^2-6p\lambda\beta_F\beta_F^{(2)}+2(p\lambda)^2\beta_F\beta_F^{(3)}-3(p\lambda)^2(\beta_F^{(2)})^2)\right].\end{aligned}$$

The used expression of the second moment of the sojourn time in an FCFS node with direct feedback is given by Takács in formula (36) of [17]. An expression for $\mathbb{E}[S_{PS}S_F]$ can be derived as follows:

$$\begin{aligned}\mathbb{E}[S_{PS}S_F] &= \left.\frac{\partial^2\Psi}{\partial\omega_1\partial\omega_2}\right|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\frac{\partial}{\partial\omega_1}(\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}])\frac{\partial}{\partial\omega_2}(\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]^k)\Big|_{(0,0)}\\ &= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[-\sigma_{PS}^{(k+1)}e^{-\omega_1\sigma_{PS}^{(k+1)}}]\mathbb{E}[-\sigma_F^{(k)}e^{-\omega_2\sigma_F^{(k)}}]^k\Big|_{(0,0)}\\ &= \sum_{k=0}^{\infty}(1-p)p^k\mathbb{E}[\sigma_{PS}^{(k+1)}]\mathbb{E}[\sigma_F^{(k)}]\end{aligned}$$

$$= \frac{2p}{(1-p)^2}\frac{\beta_{PS}}{1-\rho_{PS}}\left(\frac{\beta_F}{1-\rho_F}+\frac{\lambda}{2}(\beta_F^{(2)}-2\beta_F^2)\frac{p}{1-\rho_F}\right).$$

Expressions for $\mathbb{E}[S_{PS}]$ and $\mathbb{E}[S_F]$ are obtained in a similar way as the previous results and are given in (2.6) and (2.7) of [3]. Combining these expressions with the previous formulas results in an approximation for the variance of the total sojourn time:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[S_{PS}+S_F] &= \mathbb{E}[(S_{PS}+S_F)^2]-(\mathbb{E}[S_{PS}+S_F])^2\\
&= \mathbb{E}[S_{PS}^2]+2\mathbb{E}[S_{PS}S_F]+\mathbb{E}[S_F^2]-(\mathbb{E}[S_{PS}]+\mathbb{E}[S_F])^2\\
&\approx \frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2\\
&+ \frac{1+p-(1-p)c_{PS}^2}{(1-p)^2}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}-\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right)\\
&+ 2\frac{2p}{(1-p)^2}\frac{\beta_{PS}}{1-\rho_{PS}}\left(\frac{\beta_F}{1-\rho_F}+\frac{\lambda}{2}(\beta_F^{(2)}-2\beta_F^2)\frac{p}{1-\rho_F}\right)\\
&+ p\frac{(1-p)^2-2(1-p)}{6((1-p)-p\lambda\beta_F)^2((1-p)^2-(1-p)(2+p\lambda\beta_F)+p\lambda\beta_F)}\\
&\cdot \Big[2(1-p)(6p\lambda\beta_F^3-6\beta_F^2-6p\lambda\beta_F\beta_F^{(2)}+3\beta_F^{(2)}+p\lambda\beta_F^{(3)})\\
&- (12p\lambda\beta_F^3-12\beta_F^2-6p\lambda\beta_F\beta_F^{(2)}+2(p\lambda)^2\beta_F\beta_F^{(3)}-3(p\lambda)^2(\beta_F^{(2)})^2)\Big]\\
&- \left(\frac{\beta_{PS}}{(1-p)(1-\rho_{PS})}+\frac{p}{1-p}\left[\frac{\beta_F}{1-\rho_F}+\frac{\lambda}{2}(\beta_F^{(2)}-2\beta_F^2)\frac{p}{1-\rho_F}\right]\right)^2.
\end{aligned}$$

2.1.3 Method III: Weighted Average Approximation (WA)

A third approximation method is set up as follows: replace the LSTs in the righthand side of (2.2) by weighted sums of LSTs that correspond to the two extremes of short-circuiting (i.e., immediate feedback to the same queue) and independence of successive sojourn times of a customer at the same queue (i.e., feedback after an infinite amount of time):

$$\begin{aligned}
\Psi(\omega_1,\omega_2) &\approx \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]+(1-w)\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}\right)\\
&\cdot \left((1-w)\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]+w\mathbb{E}[e^{-\omega_2\sigma_F}]^k\right), \qquad (2.11)
\end{aligned}$$

with Re ω_1, Re $\omega_2 \geq 0$. In the same way as we did with the two other methods, we compute the second moment of the sojourn time at the PS node by

$$\begin{aligned}
\mathbb{E}[S_{PS}^2] &= \left.\frac{\partial^2\Psi}{\partial\omega_1^2}\right|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\left(w\frac{\partial^2}{\partial\omega_1^2}(\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}])+(1-w)\frac{\partial^2}{\partial\omega_1^2}(\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1})\right)\\
&\cdot \left((1-w)\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]+w\mathbb{E}[e^{-\omega_2\sigma_F}]^k\right)\Big|_{(0,0)}\\
&= \sum_{k=0}^{\infty}(1-p)p^k\left(w\frac{\partial}{\partial\omega_1}(\mathbb{E}[-\sigma_{PS}^{(k+1)}e^{-\omega_1\sigma_{PS}^{(k+1)}}])\right.\\
&+ \left.(1-w)(k+1)\frac{\partial}{\partial\omega_1}(\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^k\mathbb{E}[-\sigma_{PS}e^{-\omega_1\sigma_{PS}}])\right)\\
&\cdot \left((1-w)\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]+w\mathbb{E}[e^{-\omega_2\sigma_F}]^k\right)\Big|_{(0,0)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathbb{E}[(\sigma_{PS}^{(k+1)})^2 e^{-\omega_1\sigma_{PS}^{(k+1)}}]\right)\\
&+ (1-w)(k+1)(k(\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k-1}\mathbb{E}[-\sigma_{PS}e^{-\omega_1\sigma_{PS}}]^2\\
&+ \mathbb{E}[e^{-\omega_1\sigma_{PS}}]^k\mathbb{E}[\sigma_{PS}^2 e^{-\omega_1\sigma_{PS}}])\Big)\left((1-w)\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]+w\mathbb{E}[e^{-\omega_2\sigma_F}]^k\right)\Big|_{(0,0)}\\
&= \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathbb{E}[(\sigma_{PS}^{(k+1)})^2]+(1-w)(k+1)(k\mathbb{E}[\sigma_{PS}]^2+\mathbb{E}[\sigma_{PS}^2])\right)\\
&\approx w\left[\frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2\right.\\
&+ \left.\frac{1+p-(1-p)c_{PS}^2}{(1-p)^2}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}-\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right)\right]\\
&+ (1-w)\left[\frac{2p}{(1-p)^2}\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2+\frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2\right.\\
&+ \left.\frac{1-c_{PS}^2}{1-p}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}-\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right)\right].
\end{aligned}
$$

The second moment of the sojourn time in the FCFS node can be obtained by

$$
\begin{aligned}
\mathbb{E}[S_F^2] &= \frac{\partial^2\Psi}{\partial\omega_2^2}\Big|_{(0,0)} \approx \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]+(1-w)\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}\right)\\
&\cdot \left((1-w)\frac{\partial^2}{\partial\omega_2^2}(\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}])+w\frac{\partial^2}{\partial\omega_2^2}(\mathbb{E}[e^{-\omega_2\sigma_F}]^k)\right)\Big|_{(0,0)}\\
&= \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]+(1-w)\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}\right)\\
&\cdot \left((1-w)\frac{\partial}{\partial\omega_2}(\mathbb{E}[-\sigma_F^{(k)}e^{-\omega_2\sigma_F^{(k)}}])+wk\frac{\partial}{\partial\omega_2}(\mathbb{E}[e^{-\omega_2\sigma_F}]^{k-1}\mathbb{E}[-\sigma_F e^{-\omega_2\sigma_F}])\right)\Big|_{(0,0)}\\
&= \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]+(1-w)\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1}\right)\\
&\cdot \left((1-w)(\mathbb{E}[(\sigma_F^{(k)})^2e^{-\omega_2\sigma_F^{(k)}}])+wk((k-1)\mathbb{E}[e^{-\omega_2\sigma_F}]^{k-2}\mathbb{E}[-\sigma_F e^{-\omega_2\sigma_F}]^2\right.\\
&+ \left.\mathbb{E}[e^{-\omega_2\sigma_F}]^{k-1}\mathbb{E}[(\sigma_F)^2e^{-\omega_2\sigma_F}])\right)\Big|_{(0,0)}\\
&= \sum_{k=0}^{\infty}(1-p)p^k\left((1-w)\mathbb{E}[(\sigma_F^{(k)})^2]+wk((k-1)\mathbb{E}[\sigma_F]^2+\mathbb{E}[\sigma_F^2])\right)\\
&= (1-w)p\frac{(1-p)^2-2(1-p)}{6((1-p)-p\lambda\beta_F)^2((1-p)^2-(1-p)(2+p\lambda\beta_F)+p\lambda\beta_F)}\\
&\cdot \left[2(1-p)(6p\lambda\beta_F^3-6\beta_F^2-6p\lambda\beta_F\beta_F^{(2)}+3\beta_F^{(2)}+p\lambda\beta_F^{(3)})\right.\\
&- \left.(12p\lambda\beta_F^3-12\beta_F^2-6p\lambda\beta_F\beta_F^{(2)}+2(p\lambda)^2\beta_F\beta_F^{(3)}-3(p\lambda)^2(\beta_F^{(2)})^2)\right]\\
&+ w\left(\frac{2p^2}{(1-p)^2}\left(\beta_F+\frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2\right.
\end{aligned}
$$

$$+ \quad \frac{p}{1-p}\left[\beta_F^{(2)} + 2\left(\frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2 + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(3)}}{3\beta_F}\right]\Bigg).$$

An expression for $\mathbb{E}[S_{PS}S_F]$ can be derived as follows:

$$\begin{aligned}
\mathbb{E}[S_{PS}S_F] &= \left.\frac{\partial^2\Psi}{\partial\omega_1\partial\omega_2}\right|_{(0,0)} \\
&\approx \sum_{k=0}^{\infty}(1-p)p^k\left(w\frac{\partial}{\partial\omega_1}(\mathbb{E}[e^{-\omega_1\sigma_{PS}^{(k+1)}}]) + (1-w)\frac{\partial}{\partial\omega_1}(\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^{k+1})\right) \\
&\cdot \quad \left((1-w)\frac{\partial}{\partial\omega_2}(\mathbb{E}[e^{-\omega_2\sigma_F^{(k)}}]) + w\frac{\partial}{\partial\omega_2}(\mathbb{E}[e^{-\omega_2\sigma_F}]^k)\right)\Bigg|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k\Big(w\mathbb{E}[-\sigma_{PS}^{(k+1)}e^{-\omega_1\sigma_{PS}^{(k+1)}}] \\
&+ \quad (1-w)(k+1)\mathbb{E}[e^{-\omega_1\sigma_{PS}}]^k\mathbb{E}[-\sigma_{PS}e^{-\omega_1\sigma_{PS}}]\Big) \\
&\cdot \quad \left((1-w)\mathbb{E}[-\sigma_F^{(k)}e^{-\omega_2\sigma_F^{(k)}}] + wk\mathbb{E}[e^{-\omega_2\sigma_F}]^{k-1}\mathbb{E}[-\sigma_F e^{-\omega_2\sigma_F}]\right)\Bigg|_{(0,0)} \\
&= \sum_{k=0}^{\infty}(1-p)p^k\left(w\mathbb{E}[\sigma_{PS}^{(k+1)}] + (1-w)(k+1)\mathbb{E}[\sigma_{PS}]\right)\left((1-w)\mathbb{E}[\sigma_F^{(k)}] + wk\mathbb{E}[\sigma_F]\right) \\
&= \sum_{k=0}^{\infty}(1-p)p^k\Big(w(1-w)\mathbb{E}[\sigma_{PS}^{(k+1)}]\mathbb{E}[\sigma_F^{(k)}] + w^2\mathbb{E}[\sigma_{PS}^{(k+1)}]k\mathbb{E}[\sigma_F] \\
&+ \quad (1-w)^2(k+1)\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F^{(k)}] + (1-w)w(k+1)k\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F]\Big) \\
&= \sum_{k=0}^{\infty}(1-p)p^k\Big(w(1-w)(k+1)\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F^{(k)}] + w^2(k+1)k\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F] \\
&+ \quad (1-w)^2(k+1)\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F^{(k)}] + (1-w)w(k+1)k\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F]\Big) \\
&= \sum_{k=0}^{\infty}(1-p)p^k\Big((1-w)(k+1)\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F^{(k)}] + w(k+1)k\mathbb{E}[\sigma_{PS}]\mathbb{E}[\sigma_F]\Big) \\
&= (1-w)\frac{2p}{(1-p)^2}\frac{\beta_{PS}}{1-\rho_{PS}}\left(\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}\left(\beta_F^{(2)} - 2\beta_F^2\right)\frac{p}{1-\rho_F}\right) \\
&+ \quad w\frac{2p}{(1-p)^2}\frac{\beta_{PS}}{1-\rho_{PS}}\left(\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}\left(\beta_F^{(2)} - 2\beta_F^2\right)\frac{p}{(1-p)(1-\rho_F)}\right) \\
&= \frac{2p}{(1-p)^2}\frac{\beta_{PS}}{1-\rho_{PS}}\left(\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}\left(\beta_F^{(2)} - 2\beta_F^2\right)\frac{p}{1-\rho_F}\left(1-w+\frac{w}{1-p}\right)\right).
\end{aligned}$$

An expression for $\mathbb{E}[S_{PS}]$ and $\mathbb{E}[S_F]$ is obtained in a similar way as previous results and is given in (2.9) and (2.10) of [3]. Combining these expressions results in an approximation for the variance of the total sojourn time:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[S_{PS}+S_F] &= \mathbb{E}[(S_{PS}+S_F)^2] - (\mathbb{E}[S_{PS}+S_F])^2 \\
&= \mathbb{E}[S_{PS}^2] + 2\mathbb{E}[S_{PS}S_F] + \mathbb{E}[S_F^2] - (\mathbb{E}[S_{PS}] + \mathbb{E}[S_F])^2 \\
&\approx w\left[\frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2\right. \\
&+ \quad \left.\frac{1+p-(1-p)c_{PS}^2}{(1-p)^2}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2} - \frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}} - 1 - \rho_{PS})\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ (1-w)\Bigg[\frac{2p}{(1-p)^2}\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 + \frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
&+ \frac{1-c_{PS}^2}{1-p}\left(\frac{2\beta^2}{(1-\rho_{PS})^2} - \frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right)\Bigg] \\
&+ 2\frac{2p}{(1-p)^2}\left(\frac{\beta_{PS}}{1-\rho_{PS}}\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}\left(\beta_F^{(2)}-2\beta_F^2\right)\frac{p}{1-\rho_F}\left(1-w+\frac{w}{1-p}\right)\right) \\
&+ (1-w)p\frac{(1-p)^2-2(1-p)}{6((1-p)-p\lambda\beta_F)^2((1-p)^2-(1-p)(2+p\lambda\beta_F)+p\lambda\beta_F)} \\
&\cdot \Big[2(1-p)(6p\lambda\beta_F^3 - 6\beta_F^2 - 6p\lambda\beta_F\beta_F^{(2)} + 3\beta_F^{(2)} + p\lambda\beta_F^{(3)}) \\
&- (12p\lambda\beta_F^3 - 12\beta_F^2 - 6p\lambda\beta_F\beta_F^{(2)} + 2(p\lambda)^2\beta_F\beta_F^{(3)} - 3(p\lambda)^2(\beta_F^{(2)})^2)\Big] \\
&+ w\Bigg(\frac{2p^2}{(1-p)^2}\left(\beta_F + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2 \\
&+ \frac{p}{1-p}\left[\beta_F^{(2)} + 2\left(\frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(2)}}{2\beta_F}\right)^2 + \frac{\rho_F}{1-\rho_F}\frac{\beta_F^{(3)}}{3\beta_F}\right]\Bigg) \\
&- \Bigg(\frac{\beta_{PS}}{(1-p)(1-\rho_{PS})} \\
&+ \frac{p}{1-p}\left[\frac{\beta_F}{1-\rho_F} + \frac{\lambda}{2}(\beta_F^{(2)}-2\beta_F^2)\frac{p}{1-\rho_F}\left(1-w+\frac{w}{1-p}\right)\right]\Bigg)^2.
\end{aligned}
$$

2.2 Numerical Results

To assess the accuracy of the previously derived approximations, we have performed numerical experiments, comparing the approximations with simulations. We have checked the accuracy of the approximations for many parameter combinations, by varying the arrival rate λ, the loads ρ_{PS} and ρ_F, the variability in the service-time distributions c_{PS}^2 and c_F^2 and the feedback probability p. Denoting the point estimations of the variance based on simulations by $\mathbb{V}\mathrm{ar}_s[S]$, and denoting the calculated values of the variance by $\mathbb{V}\mathrm{ar}_a[S]$, the relative error of the variance is defined as

$$\Delta\% = \frac{\mathbb{V}\mathrm{ar}_a[S] - \mathbb{V}\mathrm{ar}_s[S]}{\mathbb{V}\mathrm{ar}_s[S]} \cdot 100\%. \tag{2.12}$$

Each configuration is ran ten times where the runs are taken long enough to ensure that all the confidence intervals were at most 15% of the point estimator value. In Table 2.1 we give numerical results for different cases with exponentially distributed service times at both nodes.
From Table 2.1 it becomes clear that if both nodes are highly asymmetrically loaded, WA performs very well for exponentially distributed service times. However, when the loads are nearly equal, the performance of the approximations is less accurate. In Table 2.2 we present numerical results for different cases with uniformly distributed service times at both nodes over the interval $[0;\ 2\beta_{PS}]$ for the PS node and $[0;\ 2\beta_F]$ for the FCFS node. When the service times are uniformly distributed, Table 2.2 shows that even for equally loaded nodes WA performs very well. This indicates that for c^2s below one WA performs better than for c^2s above one. We will investigate this further using Gamma distributed service times.
The results in Table 2.3 show that when unequal loads are used, WA performs very well for several c^2s. But when the nodes are symmetrically loaded, the results are even worse than the results given in Table 2.1, in which we use exponentially distributed service times.
In Table 2.4 we investigate te influence of several squared coefficients of variation. This table

p	λ	ρ_{PS}	ρ_F	c^2_{PS}	c^2_F	$\mathbb{V}$ar Sim	IA	$\Delta\%$	SC	$\Delta\%$	WA	$\Delta\%$
0.2	0.72	0.9	0.1	1	1	415.38	364.87	-12.16	418.74	0.81	415.60	0.05
0.5	0.45	0.9	0.1	1	1	1066.37	737.33	-30.86	1082.07	1.47	1073.76	0.69
0.8	0.18	0.9	0.1	1	1	6724.48	3380.48	-49.73	6827.83	1.54	6775.40	0.76
0.2	0.08	0.1	0.9	1	1	129966.23	114220.84	-12.11	136555.45	5.07	136500.39	5.03
0.5	0.05	0.1	0.9	1	1	158946.82	98005.20	-38.34	153548.40	-3.40	153207.30	-3.61
0.8	0.04	0.1	0.9	1	1	161109.20	77195.38	-52.09	161571.23	0.29	160745.20	-0.23
0.2	0.64	0.8	0.8	1	1	576.70	510.42	-11.49	581.98	0.92	562.41	-2.48
0.5	0.40	0.8	0.8	1	1	870.61	666.67	-23.43	885.52	1.71	788.10	-9.48
0.8	0.16	0.8	0.8	1	1	4481.16	2979.17	-33.52	4575.27	2.10	3784.33	-15.55

Table 2.1: Sojourn time variances of a network with exponentially distributed service times, compared with three approximations.

p	λ	ρ_{PS}	ρ_F	c^2_{PS}	c^2_F	$\mathbb{V}$ar Sim	IA	$\Delta\%$	SC	$\Delta\%$	WA	$\Delta\%$
0.2	0.72	0.9	0.1	0.33	0.33	265.93	216.79	-18.48	270.69	1.79	267.62	0.64
0.5	0.45	0.9	0.1	0.33	0.33	807.64	500.35	-38.05	845.26	4.66	837.03	3.64
0.8	0.18	0.9	0.1	0.33	0.33	5804.96	2786.82	-51.99	6235.84	7.42	6183.34	6.52
0.2	0.08	0.1	0.9	0.33	0.33	75006.54	52658.84	-29.79	74960.68	-0.06	74887.98	-0.16
0.5	0.05	0.1	0.9	0.33	0.33	103771.71	46031.49	-55.64	107954.83	4.03	107505.39	3.60
0.8	0.04	0.1	0.9	0.33	0.33	146747.27	37247.44	-74.62	141428.54	-3.62	140347.16	-4.36
0.2	0.64	0.8	0.8	0.33	0.33	335.08	272.93	-18.55	351.05	4.77	328.29	-2.02
0.5	0.40	0.8	0.8	0.33	0.33	570.83	408.02	-28.52	676.50	18.51	558.11	-2.23
0.8	0.16	0.8	0.8	0.33	0.33	3195.89	2058.39	-35.59	4156.17	30.05	3133.58	-1.95

Table 2.2: Sojourn time variances of a network with uniformly distributed service times, compared with three approximations.

shows it is hard to say that for larger squared coefficients of variation the error is also larger. Then, the question arises why the results of Table 2.2 are a little more precise. We did some other experiments with uniformly distributed service times which also show that the results are a little more exact than the same experiments with Gamma distributed service times and a squared variance of variation of 1/3. A reason for this is a different skewness of the distributions whereby the third moments differ. That can explain the better results of Table 2.2.
Comparing the tables shows that even with advanced approximation methods it is not easy to obtain precise approximations for a network with two generally distributed nodes. However, in Tables 2.1–2.3 approximation methods SC and WA perform very well for low p-values, and performs quite well for high p-values. Since from an application point-of-view high p-values are less relevant, these results can be useful in some cases. Together with the results in Table 2.4 we can conclude that the approximation methods SC and WA at least can be used to obtain an indication for variances of the sojourn times.

p	λ	ρ_{PS}	ρ_F	c^2_{PS}	c^2_F	$\mathbb{V}$ar Sim	IA	$\Delta\%$	SC	$\Delta\%$	WA	$\Delta\%$
0.2	0.72	0.9	0.1	1.67	1.67	583.39	513.69	-11.95	567.53	-2.72	564.33	-3.27
0.5	0.45	0.9	0.1	1.67	1.67	1223.67	975.50	-20.28	1320.07	7.88	1311.67	7.19
0.8	0.18	0.9	0.1	1.00	4.56	7154.76	3392.08	-52.59	6830.41	-4.53	6778.41	-5.26
0.2	0.08	0.1	0.9	1.00	4.56	764243.17	833931.86	9.12	772466.96	1.08	772533.82	1.08
0.5	0.05	0.1	0.9	4.56	4.56	505953.86	699811.37	38.32	533439.90	5.43	534051.49	5.55
0.8	0.04	0.1	0.9	1.67	4.56	301897.28	532870.33	76.51	292326.06	-3.17	294513.48	-2.45
0.2	0.64	0.8	0.8	1.67	1.67	853.39	812.53	-4.79	863.88	1.23	848.44	-0.58
0.5	0.40	0.8	0.8	1.67	1.67	1163.54	977.95	-15.95	1117.26	-3.98	1046.49	-10.06
0.8	0.16	0.8	0.8	1.67	4.56	10294.56	9816.29	-4.65	6483.60	-37.02	7593.79	-26.23

Table 2.3: Sojourn time variances of a network with Gamma distributed service times, compared with three approximations.

p	λ	ρ_{PS}	ρ_F	c^2_{PS}	c^2_F	$\mathbb{V}$ar Sim	IA	$\Delta\%$	SC	$\Delta\%$	WA	$\Delta\%$
0.50	0.50	0.80	0.80	0.10	0.10	285.99	197.67	-30.88	390.21	36.44	305.33	6.76
0.50	0.50	0.80	0.80	0.30	0.30	348.77	235.45	-32.49	427.19	22.48	343.80	-1.43
0.50	0.50	0.80	0.80	0.50	0.50	410.59	275.69	-32.86	465.45	13.36	383.91	-6.50
0.50	0.50	0.80	0.80	1.00	1.00	566.13	387.04	-31.63	566.74	0.11	491.18	-13.24
0.50	0.50	0.80	0.80	1.50	1.50	713.37	513.75	-27.98	676.05	-5.23	608.11	-14.75
0.50	0.50	0.80	0.80	2.00	2.00	844.06	655.82	-22.30	793.39	-6.00	734.40	-12.99
0.50	0.50	0.80	0.80	4.00	4.00	1525.55	1377.69	-9.69	1343.05	-11.96	1329.05	-12.88
0.50	0.50	0.80	0.80	8.00	8.00	3224.03	3558.72	10.38	2827.76	-12.29	2927.11	-9.21
0.50	0.50	0.80	0.80	16.00	16.00	8186.87	10869.90	32.77	7338.76	-10.36	7698.12	-5.97

Table 2.4: Sojourn time variances of a network with equally loaded nodes with Gamma distributed service times where c^2 increases, compared with three approximations.

Chapter 3

Queueing Networks with Deterministic Routing

Although there are many applications where a model with one PS node and one FCFS node with Bernoulli feedback, as considered in previous chapter, is very useful, it is the smallest network that is possible. So in this chapter we investigate more extended networks. Van der Mei et al. give in [8] exact expressions for the mean sojourn times and approximations for the variances in a queueing network with one PS node and several FCFS nodes with Markovian routing. We will adapt some of these results to a network with deterministic routing. Note that the running example given in Section 1.1.2 is one of such networks. A graphical representation of a general network with deterministic routing is given in Figure 3.1.

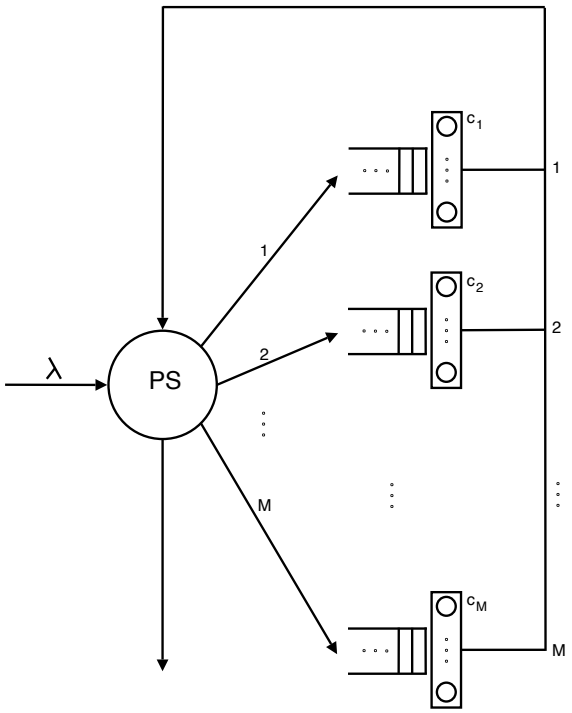

Figure 3.1: Network with deterministic routing which consists of one PS node and several multi-server FCFS nodes.

3.1 Sojourn Times with an $M/G/1$ PS Node and $M/M/c$ FCFS Nodes

We assume that customers arrive at the network according to a Poisson process with rate λ. The service time S_{PS} at the PS node is represented as a generally distributed random variable with distribution $B_{PS}(\cdot)$ and with first moment β_{PS}. The load at the PS node is given by

$$\rho_{PS} = (M+1)\lambda\beta_{PS},$$

where $M \in \mathbb{N}$ represents the number of FCFS nodes in the network. For an FCFS node k with exponentially distributed service time distribution $B_k(\cdot)$ and β_k as first moment of the service time, the load is given by

$$\rho_k = \frac{\lambda\beta_k}{c_k}, \tag{3.1}$$

with $c_k \in \mathbb{N}$ the number of servers at the FCFS node for $k = 1, \ldots, M$. Then, as we define $S_{PS}^{(i)}$ as the sojourn time of the i-th visit to the PS node, for $i = 1, \ldots, M+1$, and also define S_k as the sojourn time of the only visit to the FCFS node k, for $k = 1, \ldots, M$, the total sojourn time is given by

$$S = \sum_{i=1}^{M+1} S_{PS}^{(i)} + \sum_{k=1}^{M} S_k. \tag{3.2}$$

3.1.1 Expectation of the Sojourn Times

Since the network with deterministic routing fulfils each of the conditions of a product-form network, we may use the properties of such a network. So, by defining L_{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

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

with $l \geq 0$ and $l_k \geq 0$ for $k = 1, \ldots, M$. For the PS node we get the equilibrium distribution $\mathbb{P}(L_{PS} = l) = (1 - \rho_{PS})\rho_{PS}^l$, whereby it follows that

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

Then, using Little's law we get

$$\mathbb{E}[S_{PS}^{(i)}] = \frac{\rho_{PS}}{(M+1)\lambda(1-\rho_{PS})} = \frac{\beta_{PS}}{1-\rho_{PS}}, \tag{3.3}$$

for $i = 1, \ldots, M+1$. For the FCFS nodes we derive $\mathbb{P}(L_k = l_k)$ from the equilibrium probabilities

$$\lambda\mathbb{P}(L_k = l_k - 1) = \min(l_k, c_k)\mathbb{P}(L_k = l_k)/\beta_k,$$

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

$$\mathbb{P}(L_k = l_k) = \frac{(c_k\rho_k)^{l_k}}{l_k!}\mathbb{P}(L_k = 0), \qquad l_k = 0, \ldots, c_k, \tag{3.4}$$

and

$$\mathbb{P}(L_k = c_k + l_k) = \rho_k^{l_k}\frac{(c_k\rho_k)^{c_k}}{c_k!}\mathbb{P}(L_k = 0), \qquad l_k = 0, 1, 2, \ldots, \tag{3.5}$$

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

$$\mathbb{P}(L_k = 0) = \left(\sum_{l=0}^{c_k - 1} \frac{(c_k \rho_k)^l}{l!} + \frac{(c_k \rho_k)^{c_k}}{c_k!} \frac{1}{1 - \rho_k} \right)^{-1}.$$

From the probabilities in (3.4) and (3.5) we can derive the probability π_k that a customer has to wait at FCFS queue k:

$$\begin{aligned} \pi_k &= \mathbb{P}(L_k = c_k) + \mathbb{P}(L_k = c_k + 1) + \mathbb{P}(L_k = c_k + 2) + \ldots \\ &= \mathbb{P}(L_k = c_k)[1 + \rho_k + \rho_k^2 + \ldots] = \frac{\mathbb{P}(L_k = c_k)}{1 - \rho_k} \\ &= \frac{(c_k \rho_k)^{c_k}}{c_k!} \left((1 - \rho_k) \sum_{l=0}^{c-1} \frac{(c_k \rho_k)^l}{l!} + \frac{(c_k \rho_k)^{c_k}}{c_k!} \right)^{-1}. \end{aligned} \tag{3.6}$$

If a customer enters an FCFS $M/M/c_k$ queue with mean service time β_k, then the conditional number of customers in the system, given that the customer has to wait, equals the number of customers in an FCFS $M/M/1$ queue with service time β_k/c_k. Thereby we can derive the following expression of the waiting time for an FCFS $M/M/c_k$ queue:

$$\mathbb{E}[S_k] = \frac{\beta_k}{(1 - \rho_k)c_k} \pi_k + \beta_k, \tag{3.7}$$

for $k = 1, \ldots, M$. Combining (3.2), (3.3) and (3.7) we obtain the following expression for the mean total sojourn time of an arbitrary customer:

$$\begin{aligned} \mathbb{E}[S] &= \mathbb{E}\left[\sum_{i=1}^{M+1} S_{PS}^{(i)} + \sum_{k=1}^{M} S_k \right] \\ &= (M+1)\mathbb{E}[S_{PS}^{(1)}] + \sum_{k=1}^{M} \mathbb{E}[S_k] \\ &= \frac{(M+1)\beta_{PS}}{1 - \rho_{PS}} + \sum_{k=1}^{M} \left(\frac{\beta_k}{(1 - \rho_k)c_k} \pi_k + \beta_k \right), \end{aligned} \tag{3.8}$$

where π_k is given in (3.6).

3.1.2 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:

$$\mathbb{V}\mathrm{ar}[S] = \mathbb{V}\mathrm{ar}\left[\sum_{i=1}^{M+1} S_{PS}^{(i)} + \sum_{k=1}^{M} S_k \right]. \tag{3.9}$$

Then we need the following assumptions to limit the complexity of the problem:

Assumption 1: The total sojourn time of a customer in the PS node can be approximated by the sum of $M + 1$ independent identically distributed sojourn times.

Assumption 2: The total arrival process at the PS node is a Poisson process with rate $(M+1)\lambda$.

Assumption 3: The sojourn times S_i and S_j are uncorrelated, with $i \neq j$. So, $\text{Cov}[S_i, S_j] \approx 0$. The sojourn times $S_{PS}^{(i)}$ and S_j are uncorrelated for all i, j. So, $\text{Cov}[S_{PS}^{(i)}, S_j] \approx 0$.

To approximate the variance of the sojourn times in the PS node we use the expression of Van den Berg and Boxma in [18] for the second moment of the sojourn time of an $M/G/1$-PS node. Similar as in (2.9), we adapt the expression of Van den Berg and Boxma by considering the $M+1$ visits together resulting in a convolution of $M+1$ sojourn times. Thus, also the squared coefficient of variation must be divided by $M+1$, which can be proved using (2.10) and substituting $k = M$. Using this for the total arrival time at the PS node in our model, we derive for the second moment of the sojourn times

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

The variance of the sojourn time with general service times at the PS node can now be obtained by

$$\mathbb{V}\text{ar}\left[\sum_{i=1}^{M+1} S_{PS}^{(i)}\right] = \mathbb{E}[S_{PS}^2] - ((M+1)\mathbb{E}[S_{PS}^{(1)}])^2. \tag{3.10}$$

Using assumption 2, c_{PS}^2 can be derived from real data or depends on which distribution is used. By following assumption 3, the variance of the total sojourn time in an FCFS node with exponentially distributed service times can be expressed as follows:

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\sum_{k=1}^{M} S_k\right] &= \sum_{k=1}^{M}\mathbb{V}\text{ar}[S_k] + 2\sum_{i=1}^{M}\sum_{j=i+1}^{M}\mathbb{C}\text{ov}[S_i, S_j] \\
&\approx \sum_{k=1}^{M}\left(\mathbb{E}[W_k^2] - (\mathbb{E}[W_k])^2 + \beta_k^2\right). \\
&= \sum_{k=1}^{M}\left(\pi_k\frac{2\beta_k^2}{c_k^2(1-\rho_k)^2} - \pi_k^2\frac{\beta_k^2}{c_k^2(1-\rho_k)^2} + \beta_k^2\right) \\
&= \sum_{k=1}^{M}\left(\frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2} + \beta_k^2\right),
\end{aligned} \tag{3.11}$$

where W_k represents the waiting time at queue $k = 1, \ldots, M$. Using the assumptions and by substituting (3.11) and (3.10) in (3.9), an explicit expression is derived for the variance of the total sojourn time in a queueing network with general service times at the PS node:

$$\begin{aligned}
\mathbb{V}\text{ar}[S] &\approx (M+1)c_{PS}^2\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
&+ ((M+1)^2-(M+1)c_{PS}^2)\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right) \\
&- \left(\frac{(M+1)\beta_{PS}}{1-\rho_{PS}}\right)^2 + \sum_{k=1}^{M}\left(\beta_k^2 + \frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2}\right).
\end{aligned} \tag{3.12}$$

Numerical Results

In this section the accuracy of (3.12) is assessed. We have performed numerous numerical experiments and checked the accuracy of approximations 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 are run ten times where the lengths of the runs are taken long enough to ensure reliable results. The relative error is obtained by using (2.12). In all tables the variance obtained by (3.12) is compared with simulation results. To illustrate the rate of accuracy, we compare the simulation results also with a simple, straightforward approximation, which completely ignores dependencies between sojourn times in the PS node. This simple approximation is given by

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathrm{simple}}[S] &:= (M+1)\left[c_{PS}^2\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2\right.\\ &+ (1-c_{PS}^2)\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right)\\ &- \left.\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2\right]+\sum_{k=1}^{M}\left(\beta_k^2+\frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2}\right). \qquad (3.13)\end{aligned}$$

Single-Server Case: In Table 3.1 the results of the variance are given for a queueing network with deterministic routing, exponential service times at the PS node and exponential service times at the two identical single-server FCFS nodes.

β_{PS}	β_F	$\mathbb{V}\mathrm{ar}_s[S]$	$\mathbb{V}\mathrm{ar}_a[S]$	$\Delta\mathbb{V}\mathrm{ar}\%$	$\mathbb{V}\mathrm{ar}_a[S]$ (simple)	$\Delta\mathbb{V}\mathrm{ar}\%$ (simple)
0.1	0.2	0.24	0.24	-1.11	0.17	-30.26
0.1	0.5	2.15	2.11	-1.74	2.04	-4.97
0.1	0.8	32.82	32.11	-2.17	32.04	-2.38
0.2	0.2	2.22	2.28	2.65	0.35	-84.04
0.2	0.5	4.21	4.15	-1.27	2.23	-47.00
0.2	0.8	34.26	34.15	-0.33	32.23	-5.94
0.3	0.2	119.51	117.85	-1.39	0.87	-99.27
0.3	0.5	122.82	119.72	-2.52	2.75	-97.76
0.3	0.8	145.34	149.72	3.01	32.75	-77.47

Table 3.1: Sojourn time variances of a queueing network with exponential service times and two identical single-server FCFS nodes.

We see that our approximation for the variance performs very well. As expected, the simple approximation as given in (3.13) consistently and strongly underestimates the variance of the total sojourn time; it appears to be an inaccurate lower bound. To investigate the impact of asymmetry of the loads on the accuracy of the approximations, we have also considered a variety of parameter combinations with unequal loads of the two FCFS node. Table 3.2 presents the results for this case where we have the same network as in Table 3.1, except instead of using two identically single server nodes, the loads of the two FCFS nodes are taken asymmetrically.
We see that the results in Table 3.2 demonstrate that for asymmetric cases the relative error is still very low. Again the estimate is sometimes higher and sometimes lower than the simulated value. It does not seem to make any difference whether the loads of the nodes are very different or close to each other. Asymmetric loads do not cause the approximation to perform significantly worse. This could be expected since the approximation contains no covariance terms. Consequently the formula given in (3.12) can be applied to asymmetric loads as well.

To validate our approximation for a network with more than two FCFS nodes, we also include the results for a case with five FCFS nodes in Table 3.3. For the PS node as for the FCFS nodes we

β_{PS}	β_F		$\mathbb{V}\text{ar}_s[S]$	$\mathbb{V}\text{ar}_a[S]$	$\Delta\mathbb{V}\text{ar}\%$	$\mathbb{V}\text{ar}_a[S]$ (simple)	$\Delta\mathbb{V}\text{ar}\%$ (simple)
0.1	0.2	0.8	15.87	16.17	1.93	16.10	1.49
0.1	0.3	0.7	5.54	5.74	3.56	5.67	2.31
0.1	0.4	0.6	2.87	2.8	-2.29	2.74	-4.71
0.2	0.2	0.8	18.24	18.21	-0.12	16.29	-10.66
0.2	0.3	0.7	7.85	7.78	-0.83	5.86	-25.34
0.2	0.4	0.6	4.97	4.85	-2.42	2.92	-41.14
0.3	0.2	0.8	142.08	133.78	-5.84	16.81	-88.17
0.3	0.3	0.7	129.34	123.35	-4.63	6.37	-95.07
0.3	0.4	0.6	122.14	120.41	-1.42	3.44	-97.18

Table 3.2: Sojourn time variances of a queueing network with exponential service times and two asymmetrically loaded single-server FCFS nodes.

use exponential service times with several service rates. As we can see, even for very high loads the approximation is still accurate. And as expected, the simple approximations again underestimate the variance of the total sojourn time in the case with five FCFS nodes. Since the more advanced approximations behave very well in systems with five FCFS nodes, we expect that the approximation will also behave well in systems with an arbitrary number of FCFS nodes, because a larger number of FCFS nodes will reduce the cross-correlations.

β_{PS}	β_F					$\mathbb{V}\text{ar}_s[S]$	$\mathbb{V}\text{ar}_a[S]$	$\Delta\mathbb{V}\text{ar}\%$	$\mathbb{V}\text{ar}_a[S]$ (simple)	$\Delta\mathbb{V}\text{ar}\%$ (simple)
0.05	0.2	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 3.3: Sojourn time variances of a queueing network with exponential service times and five single-server FCFS nodes.

Multi-Server Case: Now that we have seen that our approximation given in (3.12) is very accurate for all different configurations in networks with $M/M/1$ FCFS nodes, we evaluate the accuracy of our approximation for multi-server FCFS nodes. We first consider the model with three symmetric $M/M/c$ queues. As before, the PS node and FCFS nodes have exponentially distributed service times. As shown in Table 3.4, also for symmetric multi-server queues our approximation performs well. To extend these results we consider in Table 3.5 the case of asymmetric multi-server nodes. As expected the relative error of this case is still very good. So for the case of exponential service times at the PS node and at the FCFS nodes we can say that even in the worst case our approximation as given in (3.12) performs very well and can be used for every scenario in a queueing network with deterministic routing to obtain the variance of the sojourn time.
Now that we are convinced that our approximation performs well for exponential service times at the PS node, the question arises how this formula performs for other service time distributions. Table 3.6 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_{PS}^2 of the service time distribution at the PS node is varied as 0,

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

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	0.1	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 3.5: Sojourn time expectations and variances with three asymmetric multi-server FCFS nodes.

4, and 16. These last service times are deterministic for the case $c_{PS}^2 = 0$ and Gamma distributed 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 $\beta_{PS} = \Gamma/\lambda^* = 0.3$.
We see in Table 3.6 that our approximation also holds for general service times at the PS node. To extend these results Table 3.7 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 $c_2 = 3$. And if we look at the results, also for these cases the approximation given in (3.12) is still accurate. So we can conclude that our approximation covers a wide range of different configurations of the PS node and FCFS nodes and therefore is a reliable formula to obtain the variance in a network with deterministic routing with just two restrictions: a Poisson distributed arrival rate and exponentially distributed FCFS service times.

3.2 Sojourn Times with an $M/G/1$ PS Node and $M/G/\infty$ FCFS Nodes

Although it is very hard to obtain an expression for the expectation and variance of the sojourn time in a network which consists of multi-server FCFS nodes with general service times, using the assumption that there are infinitely servers available, an exact result of the mean sojourn time and a good approximation of the variance of the sojourn times can be determined. The mean and variance of the sojourn time of an FCFS node in any $M/G/\infty$ node equals the mean and variance of the service time. So, if we denote the first moment with β_k and second moment with $\beta_k^{(2)}$ of

β_{PS}	c_{PS}^2	β_F		$\mathbb{V}\mathrm{ar}_s[S]$	$\mathbb{V}\mathrm{ar}_a[S]$	$\Delta\mathbb{V}\mathrm{ar}\%$	$\mathbb{V}\mathrm{ar}_a[S]$ (simple)	$\Delta\mathbb{V}\mathrm{ar}\%$ (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 3.6: Sojourn time variances for a queuing network with general service times at the PS node and with two asymmetrically loaded $M/M/1$ nodes.

β_{PS}	c_{PS}^2	β_F		$\mathbb{V}\mathrm{ar}_s[S]$	$\mathbb{V}\mathrm{ar}_a[S]$	$\Delta\mathbb{V}\mathrm{ar}\%$	$\mathbb{V}\mathrm{ar}_a[S]$ (simple)	$\Delta\mathbb{V}\mathrm{ar}\%$ (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 3.7: Sojourn time variances for a queueing network with general service times at the PS node and with two asymmetrical multi-server FCFS nodes.

the generally distributed random service time distribution $B_k(\cdot)$ of the k-th FCFS node, the mean and variance of the sojourn times can be obtained for $k = 1, \ldots, M$. Note that ρ_k has no meaning anymore since $c_k \to \infty$. Further, we still use the three assumptions as given before.

3.2.1 Expectation of the Sojourn Times

Since a network with deterministic routing, general service times and infinite servers at the FCFS nodes fulfills the conditions of a product-form network we can write the closed-form expression for the expectation of the sojourn time as follows:

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{M+1} S_{PS}^{(i)} + \sum_{k=1}^{M} S_k\right] = \frac{(M+1)\beta_{PS}}{1-\rho_{PS}} + \sum_{k=1}^{M} \beta_k. \tag{3.14}$$

3.2.2 Variance of the Sojourn Times

If we use the approximation in (3.10) for the sojourn time in the PS node with general distributed service times, the following approximation can be used to determine the variance of the sojourn time in a deterministic routing network with general service times at the PS node and FCFS nodes

with infinitely many servers:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[S] &\approx (M+1)c_{PS}^2\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\ &+ ((M+1)^2-(M+1)c_{PS}^2)\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right) \\ &- \left(\frac{(M+1)\beta_{PS}}{1-\rho_{PS}}\right)^2+\sum_{k=1}^{M}\left(\beta_k^{(2)}-\beta_k^2\right). \end{aligned} \tag{3.15}$$

Numerical Results

In Table 3.8 the accuracy of (3.15) is assessed. It presents the results of the variance for a queueing network with deterministic routing, general service times at the PS node and general service times at the five infinite-server FCFS nodes. The service time of the PS node is Gamma distributed with parameters shape $\Gamma = 1/16$ and scale $\lambda^* = 5/4$ for $\beta_{PS} = 0.05$ and with parameters $\Gamma = 1/16$ and $\lambda^* = 12/5$ for $\beta_{PS} = 0.15$. The service times of the k-th FCFS node are uniformly distributed over the interval $[0;\ 2\beta_k]$. From now on we do not compare the results with a simple approximation anymore, since it is obvious that even with a lowly loaded PS node, leaving out covariances cannot result in accurate approximations. Table 3.8 makes clear that the approximation of variances of the sojourn times given by (3.15) performs very well.

β_{PS}	β_F					$\mathbb{V}\mathrm{ar}_s[S]$	$\mathbb{V}\mathrm{ar}_a[S]$	$\Delta\mathbb{V}\mathrm{ar}\%$
0.05	0.2	0.2	0.2	0.2	0.2	0.65	0.66	1.46
0.05	0.5	0.5	0.5	0.5	0.5	1.00	1.01	1.25
0.05	0.8	0.8	0.8	0.8	0.8	1.65	1.66	0.95
0.05	0.1	0.3	0.5	0.7	0.9	1.13	1.14	1.21
0.05	0.7	0.5	0.9	0.3	0.1	1.13	1.14	1.53
0.15	0.2	0.2	0.2	0.2	0.2	483.25	453.17	-6.22
0.15	0.5	0.5	0.5	0.5	0.5	439.28	453.52	3.24
0.15	0.8	0.8	0.8	0.8	0.8	443.70	454.17	2.36
0.15	0.1	0.3	0.5	0.7	0.9	443.68	453.66	2.25
0.15	0.7	0.5	0.9	0.3	0.1	444.46	453.66	2.07

Table 3.8: Sojourn time variances of a queueing network with Gamma distributed service times at the PS node and five uniformly distributed $M/G/\infty$ nodes.

3.2.3 Variance of the Sojourn Times with $c_k < \infty$

Since $c_k = \infty$ with $k = 1, \ldots, M$ is not a realistic assumption in many cases, the question arises for which c_k the approximation given in (3.15) is still valid. One answer to this question can be given by determining the smallest c_k for which the probability that all servers are busy approaches zero, or in other words, determining the smallest c_k for which π_k remains very small, with π_k given in (3.6). In Figure 3.2 the decline of π_k is given for several β_Fs with $\lambda = 1$. We see that just a few extra servers are needed to let π_k go from 1 to 0. The question that has to be answered now is for which π_k the approximation of the variance of the sojourn times are still accurate. We found this answer with trial and error and can conclude that for $\pi_k \leq 0.05$ the approximation given in (3.15) performs very well. This value of π_k seems extreme low, but we can see in Figure 3.2, with $\pi_k = 0.05$, still a plausible number of servers are given.
Configurations of the FCFS nodes are given in Table 3.9 where we use $M/G/c_k$ FCFS servers in the network with the highest c_k for $\pi_k \leq 0.05$. As in Table 3.8, the service time of the PS node is Gamma distributed with shape $\Gamma = 1/16$ and scale $\lambda^* = 5/4$ for $\beta_{PS} = 0.05$. Since we want to study the number of servers at the FCFS nodes, we keep the service time of the PS node constant at $\beta_{PS} = 0.05$. And the service times of the k-th FCFS node are uniformly distributed over the

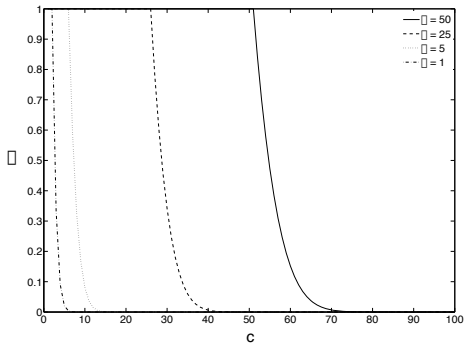

Figure 3.2: The decline of π_k for several β_ks with $\lambda = 1$ shows that just a few extra servers are needed to let π_k go from 1 to 0.

interval $[0;\ 2\beta_k]$. These are very interesting results since the number of servers is not extremely

β_{PS}	β_F					c					$\mathbb{V}\mathrm{ar}_s[S]$	$\mathbb{V}\mathrm{ar}_a[S]$	$\Delta\mathbb{V}\mathrm{ar}\%$
0.05	2	2	2	2	2	6	6	6	6	6	7.27	7.26	-0.10
0.05	5	5	5	5	5	10	10	10	10	10	42.49	42.26	-0.54
0.05	10	10	10	10	10	17	17	17	17	17	168.17	167.26	-0.54
0.05	2	5	10	2	5	6	10	17	6	10	53.47	53.26	-0.40
0.05	10	2	5	2	10	17	6	10	6	17	78.31	78.26	-0.06

Table 3.9: Sojourn time variances of a queueing network with Gamma distributed service times at the PS node and five uniformly distributed $M/G/c$ nodes.

high compared to the service rate. Now the question arises what happens if the number of servers are taken more smaller. We found that it is hard to maintain accurate results for variances of the sojourn time if c_k is reduced, so we conclude that for c_k for which $\pi_k = 0.05$ the approximation given in (3.15) is very useful. This usefulness can be illustrated by the fact that the restriction $\pi_k \leq 0.05$ will become more and more a weak restriction when the number of servers increases. If we keep β_k constant for $k = 1, \ldots, M$, and we vary the number of servers, we can calculate the load ρ_k as given in (3.1) for a calculated λ whereby $\pi_k = 0.05$ using the formula given in (3.6). Table 3.10 shows that for an example for several values of c_k and with constant $\beta_k = 1$, λ and ρ_k increase when c_k increases.

c_k	λ	ρ_k
1	0.05	0.05
5	1.90	0.38
10	5.29	0.53
15	9.04	0.60
20	13.00	0.65
30	21.25	0.71
40	29.77	0.74
50	38.47	0.77
100	83.37	0.83

Table 3.10: When c_k increases, the load needed to keep π_k at 0.05 increases too when β_k is a constant.

3.3 Comparison with Markovian Routing

To assess the differences between deterministic routing and Markovian routing, we compare the variances of two different sets of runs where we once use deterministic routing and once use Markovian routing. The network consists of one PS node with Gamma distributed service times which has $c_{PS}^2 = 4$, one $M/M/2$ FCFS node and one $M/M/3$ FCFS node. All nodes have the same load. We compare the approximation given in (3.12) for a network with deterministic routing, with a variant of the formula given in (29) of Van der Mei et al. in [8] for a network with Markovian routing. By substituting the expression for the variance of the sojourn time at the PS node already used in (2.9), we propose the following improvement of the approximation given by Van der Mei et al. in [8] for the variance of the total sojourn time in a network with Markovian routing:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_M[S] \;\approx\;& \frac{c_{PS}^2}{1-p}\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
&+ \frac{1+p-(1-p)c_{PS}^2}{(1-p)^2}\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}-\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right) \\
&- \left(\frac{1}{1-p}\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
&+ \sum_{k=1}^{M}\frac{q_k}{1-q_k}\left(\beta_k^2+\frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2}\right) \\
&+ \sum_{k=1}^{M}\frac{2q_k^2\pi_k\beta_k^2\left((1-\rho_k)^2c_k^2+\pi_k(2-\pi_k)\right)}{(1-q_k)^2(1-q_k\rho_k+q_k)(1-\rho_k)^2c_k^2} \\
&+ \sum_{k=1}^{M}\frac{q_k}{(1-q_k)^2}\left(\frac{\beta_{PS}}{1-\rho_{PS}}+\beta_k+\frac{\pi_k\beta_k}{c_k(1-\rho_k)}\right)^2 \\
&+ \sum_{k\neq m}\frac{p_kp_m}{(1-p)^2}\left(\frac{\beta_{PS}}{1-\rho_{PS}}+\beta_k+\frac{\pi_k\beta_k}{c_k(1-\rho_k)}\right) \\
&\cdot \left(\frac{\beta_{PS}}{1-\rho_{PS}}+\beta_m+\frac{\pi_m\beta_m}{c_m(1-\rho_m)}\right), \qquad (3.16)
\end{aligned}$$

where $p=\sum_{k=1}^{M}p_k$ and $q_k = p_k/(1-p+p_k)$ for $k=1,\ldots,M$, and $\rho_{PS}=\lambda\beta_{PS}(1-p)^{-1}$. This approximation differs from the expression given in [8] since there is no explicit covariance term for the sojourn times in the PS node. If we define $N=0,1,2,\ldots$ as the number of returns to the PS node then, because N is geometrically distributed, we know that the expectation of total visits to the PS node equals

$$\mathbb{E}[N]+1=\frac{p}{1-p}+1=\frac{1}{1-p}.$$

To get the same mean number of customers at each node as with deterministic routing, we have to set the expectation of the number of visits to the PS node at $M+1$, so $p=M/(M+1)$. With $p_k = 1/(M+1)$ we can obtain $q_k = 1/2$. Since the expectation of the number of visits to the PS node equals $(1-p)^{-1}=M+1$, we approximate the sojourn times in the PS node by using a convolution of $M+1$ independent identically distributed service times. This makes clear that we only review the first three terms of (3.16) which represent the variances of the PS node according to (3.12). The other variance and covariance terms still remain as derived in [8].

In Table 3.11 the simulation results of the variance of the sojourn times are given by $\mathbb{V}\mathrm{ar}_{Ds}[S]$ and $\mathbb{V}\mathrm{ar}_{Ms}[S]$ for the network with deterministic routing and the network with Markovian routing, respectively. The approximated values are given by $\mathbb{V}\mathrm{ar}_{Da}[S]$ and $\mathbb{V}\mathrm{ar}_{Ma}[S]$ for the network with deterministic routing and the network with Markovian routing, respectively. It becomes clear that both approximations perform very well, even for high loads. To show the differences between the

β_{PS}	β_F		$\mathbb{V}\mathrm{ar}_{Ds}[S]$	$\mathbb{V}\mathrm{ar}_{Da}[S]$	$\Delta\mathbb{V}\mathrm{ar}_D\%$	$\mathbb{V}\mathrm{ar}_{Ms}[S]$	$\mathbb{V}\mathrm{ar}_{Ma}[S]$	$\Delta\mathbb{V}\mathrm{ar}_M\%$
0.033	0.2	0.3	0.15	0.15	-0.15	0.65	0.65	-0.41
0.067	0.4	0.6	0.63	0.63	0.04	2.78	2.78	-0.09
0.100	0.6	0.9	1.57	1.56	-0.34	6.92	6.98	0.85
0.133	0.8	1.2	3.22	3.22	-0.07	14.47	14.63	1.11
0.167	1.0	1.5	6.38	6.31	-1.02	28.43	28.85	1.50
0.200	1.2	1.8	12.86	12.75	-0.79	57.36	57.70	0.59
0.233	1.4	2.1	29.16	28.84	-1.10	126.34	126.05	-0.23
0.267	1.6	2.4	86.06	82.86	-3.72	349.65	341.05	-2.46
0.300	1.8	2.7	433.76	429.52	-0.98	1695.63	1631.09	-3.81

Table 3.11: Sojourn time variances of two queueing network with deterministic routing and Markovian routing, respectively. Both networks have symmetrically loaded nodes where the PS node has Gamma distributed service times and the two multi-server FCFS nodes have exponentially distributed service times.

variances of both networks, Figure 3.3 presents the same results in a graphical way.

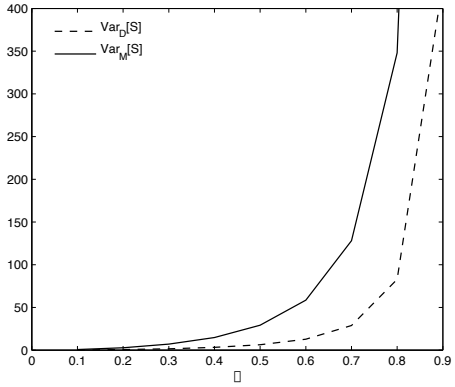

Figure 3.3: Variances of sojourn times of a queueing network with deterministic routing and a queueing network with Markovian routing.

Figure 3.3 also shows that the variance in a network with Markovian routing grows faster than the variance in a network with deterministic routing. In general, we can prove that the variance in a network with Markovian routing is larger than when the customers are deterministically routed for $\mathbb{E}[N] = M$. With the approximation of the variance in a queueing network with deterministic routing given in (3.12) and the expressions for p, p_k and q_k as given before we get

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_M[S] \;\approx\;& (M+1)c_{PS}^2\left(1+\frac{2+\rho_{PS}}{2-\rho_{PS}}\right)\left(\frac{\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
+\;& ((2M+1)(M+1)-(M+1)c_{PS}^2)\left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2}-\frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})}(e^{\rho_{PS}}-1-\rho_{PS})\right) \\
-\;& \left(\frac{(M+1)\beta_{PS}}{1-\rho_{PS}}\right)^2 \\
+\;& \sum_{k=1}^{M}\left(\beta_k^2+\frac{\pi_k(2-\pi_k)\beta_k^2}{c_k^2(1-\rho_k)^2}\right) \\
+\;& \sum_{k=1}^{M}\frac{2\pi_k\beta_k^2\left((1-\rho_k)^2c_k^2+\pi_k(2-\pi_k)\right)}{(3/2-1/2\rho_k)(1-\rho_k)^2c_k^2}
\end{aligned}$$

$$
\begin{aligned}
&+ \quad \sum_{k=1}^{M} \frac{1}{2} \left(\frac{\beta_{PS}}{1-\rho_{PS}} + \beta_k + \frac{\pi_k \beta_k}{c_k(1-\rho_k)} \right)^2 \\
&+ \quad \sum_{k \neq m} \left(\frac{\beta_{PS}}{1-\rho_{PS}} + \beta_k + \frac{\pi_k \beta_k}{c_k(1-\rho_k)} \right) \\
&\cdot \quad \left(\frac{\beta_{PS}}{1-\rho_{PS}} + \beta_m + \frac{\pi_m \beta_m}{c_m(1-\rho_m)} \right) \\
&= \quad \mathbb{V}\mathrm{ar}_D[S] + (M^2+M) \left(\frac{2\beta_{PS}^2}{(1-\rho_{PS})^2} - \frac{2\beta_{PS}^2}{\rho_{PS}^2(1-\rho_{PS})} (e^{\rho_{PS}} - 1 - \rho_{PS}) \right) \\
&+ \quad \sum_{k=1}^{M} \frac{2\pi_k \beta_k^2 \left((1-\rho_k)^2 c_k^2 + \pi_k(2-\pi_k) \right)}{(3/2 - 1/2\rho_k)(1-\rho_k)^2 c_k^2} \\
&+ \quad \sum_{k=1}^{M} \frac{1}{2} \left(\frac{\beta_{PS}}{1-\rho_{PS}} + \beta_k + \frac{\pi_k \beta_k}{c_k(1-\rho_k)} \right)^2 \\
&+ \quad \sum_{k \neq m} \left(\frac{\beta_{PS}}{1-\rho_{PS}} + \beta_k + \frac{\pi_k \beta_k}{c_k(1-\rho_k)} \right) \\
&\cdot \quad \left(\frac{\beta_{PS}}{1-\rho_{PS}} + \beta_m + \frac{\pi_m \beta_m}{c_m(1-\rho_m)} \right). \qquad (3.17)
\end{aligned}
$$

Since the term of the second moment of an PS node with deterministic service times is larger than zero and since covariance terms in the last part of (3.17) are larger than zero, we can say that for a network with $\mathbb{E}[N] = M$, the approximated variance is larger when Markovian routing is applied than when deterministic routing is used.

Chapter 4

Queueing Network with Admission Control

It is not hard to imagine that when many customers arrive in a short time, it will have high impact on the total service time in a network. To guarantee a certain quality of service, admission control can be used to make sure that the system keeps functioning in a correct way. Admission control is the simple practice of discriminating which traffic is admitted into a network in the first place. Admission control can be applied to computer networks, call center networks and many other networks where some restrictions are defined to guarantee a quality of service. A special case is web admission control. In [7], Gijsen et al. give a solid control scheme to apply admission control at web-based services. A more mathematically founded control scheme for predictable service response times for web access is given by Chen et al. in [4].
In this chapter we derive exact expressions for some performance measures in a queueing network with exponentially distributed FCFS nodes and generally distributed PS nodes where a maximum number of customers is allowed. When a customer arrives and the network is full, this customer is blocked and disappears. Beside these exact expressions we also derive some approximations for performance measures when an arriving customer is kept in a buffer for the case the network is full. An example of such a network with buffer is given in Figure 4.1. We will use such a network to obtain some simulation results.

4.1 Network Without Buffer

Consider an arbitrary open queueing network which can hold at most K customers. The network may consist of PS nodes with generally distributed service times. Further, the network may consist of multi-server FCFS nodes with exponentially distributed service times. In total the network consists of M nodes with service times B_i which has first moment β_i for $i = 1, \ldots, M$. A routing matrix can be constructed for which we define for $i, j = 1, \ldots, M$ $p_{i,j}$ as the probability that a customer moves from node i to node j, with $p_{0,i}$ the probability that an external arrival enters node i, and with $p_{i,M+1}$ the probability that a customers leaves the network from node i. Customers arrive according a Poisson process with rate λ. The load of node i is given by

$$\rho_i = \frac{\lambda \Lambda_i \beta_i}{c_i},$$

with Λ_i the mean number of visits to node i of an arbitrary customer which is given by the unique solution of

$$\Lambda_i = p_{0,i} + \sum_{k=1}^{M} p_{k,i} \Lambda_k,$$

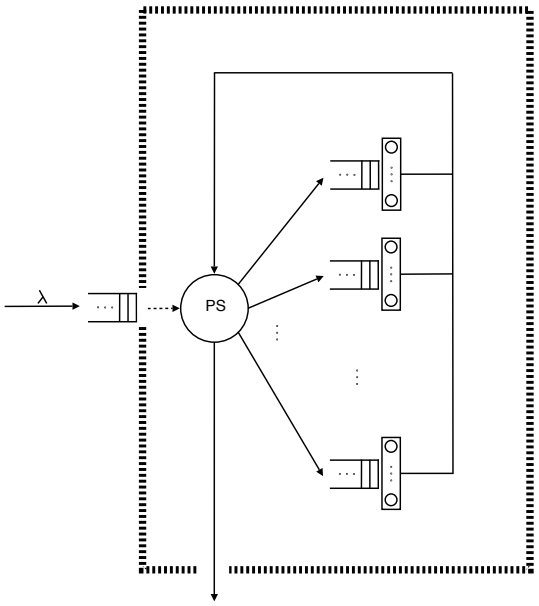

Figure 4.1: A queueing network with admission control in which customers can wait in a finite buffer if the network is full.

for $\Lambda_i \geq 0$. Due to insensitivity results for the first moment of the sojourn time at an PS node as proved by Ott in [15], we consider an PS node as a node with one server. Some extra proofs for the PS node are given in Appendix A. Note that in that case the load at an PS node i is given by $\rho_i = \lambda \Lambda_i \beta_i$. Then, as we define $S_i^{(j)}$ as the sojourn time of the j-th visit to the i-th node, the total sojourn time is given by

$$S = \sum_{i=1}^{M} \sum_{j=1}^{N} S_i^{(j)}.$$

To obtain the probability distribution of the number of customers in the system we first define the function g over the number of customers n_k at node $k = 1, \ldots, M$:

$$g(n_k) = \begin{cases} \frac{c_k^{n_k} \rho_k^{n_k}}{n_k!}, & n_k = 0, \ldots, c_k \\ \frac{\rho_k^{n_k} c_k^{c_k}}{c_k!}, & n_k = c_k, c_k + 1, \ldots . \end{cases} \tag{4.1}$$

4.1.1 Steady-state Probability Distribution

Consider an open network where at maximum K customers are allowed. Jackson proved in [10] that the equilibrium state joint probability function is given by

$$p(n_1, \ldots, n_M) = \frac{\prod_{j=1}^{M} g(n_j)}{\sum_{i=0}^{K} \sum_{n_1 + \cdots + n_M = i} \prod_{j=1}^{M} g(n_j)} = \frac{\prod_{j \in \Omega} g(n_j)}{\sum_{i=0}^{K} \sum_{\#_i} \prod_{j \in \Omega} g(n_j)}, \tag{4.2}$$

where we define $\#_k$ as the sum being taken over each combination where $n_1 + n_2 + \cdots + n_M = k$ for $k \in \mathbb{N}$ and where Ω represents the set of the M nodes: $\Omega = \{1, \ldots, M\}$. The probability that l customers are in the network is thus given by

$$\mathbb{P}(L = l) = \frac{\sum_{\#_l} \prod_{j \in \Omega} g(n_j)}{\sum_{i=0}^{K} \sum_{\#_i} \prod_{j \in \Omega} g(n_j)}.$$

For the $M/G/1$-PS node, and thus also for an $M/M/1$ queueing node, we prove this result in an alternative way in Appendix A.1.

4.1.2 Blocking Probability

By PASTA, the probability that when an arriving customer is blocked because there are already K customers in the network, equals the probability that there are K customers in the network. So, the blocking probability can be derived as follows:

$$\begin{aligned}\mathbb{P}(\text{customer is blocked}) &= \mathbb{P}(L=K)\\ &= \frac{\sum_{\#_K}\prod_{j\in\Omega} g(n_j)}{\sum_{i=0}^{K}\sum_{\#_i}\prod_{j\in\Omega} g(n_j)}.\end{aligned} \tag{4.3}$$

4.1.3 Mean Number of Customers in the Network

Now that we have derived the steady-state distribution we can obtain the mean number of customers in the network by

$$\begin{aligned}\mathbb{E}[L] &= \sum_{l=0}^{K} l\mathbb{P}(L=l)\\ &= \sum_{l=0}^{K} l\frac{\sum_{\#_l}\prod_{j\in\Omega} g(n_j)}{\sum_{i=0}^{K}\sum_{\#_i}\prod_{j\in\Omega} g(n_j)}\\ &= \sum_{l=0}^{K} \frac{\sum_{\#_l}(n_0+\cdots+n_M)\prod_{j\in\Omega} g(n_j)}{\sum_{i=0}^{K}\sum_{\#_i}\prod_{j\in\Omega} g(n_j)}\\ &= \sum_{l=0}^{K} \frac{\sum_{\#_l} n_0\prod_{j\in\Omega} g(n_j)}{\sum_{i=0}^{K}\sum_{\#_i}\prod_{j\in\Omega} g(n_j)}\\ &+ \quad \cdots \quad + \sum_{l=0}^{K} \frac{\sum_{\#_l} n_M\prod_{j\in\Omega} g(n_j)}{\sum_{i=0}^{K}\sum_{\#_i}\prod_{j\in\Omega} g(n_j)}\\ &= \mathbb{E}[L_0]+\cdots+\mathbb{E}[L_M],\end{aligned} \tag{4.4}$$

where $\mathbb{E}[L_i]$, for $i=1\ldots,M$, represents the mean number of customers at node i.

4.1.4 Expectation of the Sojourn Times

As a variant of the arrival theorem given by Lavenberg and Reiser in [13] and conform the PASTA property, we can say that when a customer leaves a station in the network, the joint distribution of the numbers of customers in all stations at this jump epoch equals the steady-state distribution of customers in the same network but with one customer less. So, with exponentially distributed service times we define the probability that an arriving customer has to wait at station $k=1,\ldots,M$ by

$$\pi_k = \frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k\geq c_k}\prod_{j\in\Omega} g(n_j)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega} g(n_j)},$$

and we define the number of customers already in the queue when an arbitrary customer arrives by

$$\mathbb{E}[L_k^{q*}] = \frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k\geq c_k}(n_k-c_k)\prod_{j\in\Omega} g(n_j)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega} g(n_j)}.$$

Since a customer always asks for service time β_k, and a customer by average π_k of the time finds all servers busy and thus has to wait β_k/c_k till a server is ready, and has to wait till all the customers already in the queue are served, we obtain the mean sojourn time at the k-th station by

$$\mathbb{E}[S_k] = \beta_k + \pi_k \frac{\beta_k}{c_k} + \mathbb{E}[L_k^{q*}]\frac{\beta_k}{c_k}. \tag{4.5}$$

With (4.5) we can write for the total mean sojourn time at a given station:

$$\begin{aligned}
\Lambda_k \mathbb{E}[S_k] &= \Lambda_k\beta_k + \Lambda_k\pi_k\frac{\beta_k}{c_k} + \Lambda_k\mathbb{E}[L_k^{q*}]\frac{\beta_k}{c_k} \\
&= \Lambda_k\beta_k(1-\pi_k) + \Lambda_k\beta_k\pi_k + \Lambda_k\mathbb{E}[L_k^{q*}]\frac{\beta_k}{c_k} + \Lambda_k\pi_k\frac{\beta_k}{c_k} \\
&= \Lambda_k\beta_k(1-\pi_k) + \Lambda_k(\mathbb{E}[L_k^{q*}] + \pi_k c_k)\frac{\beta_k}{c_k} + \Lambda_k\pi_k\frac{\beta_k}{c_k} \\
&= \Lambda_k\beta_k\frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k<c_k}\prod_{j\in\Omega}g(n_j)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&+ \frac{\Lambda_k\beta_k}{c_k}\frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k\geq c_k} n_k\prod_{j\in\Omega}g(n_j)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&+ \frac{\Lambda_k\beta_k}{c_k}\frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k\geq c_k}\prod_{j\in\Omega}g(n_j)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&= \Lambda_k\beta_k\frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k<c_k}\prod_{j\in\Omega}g(n_j)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&+ \frac{\Lambda_k\beta_k}{c_k}\frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k\geq c_k}(n_k+1)\prod_{j\in\Omega}g(n_j)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&= \frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k<c_k}\frac{1}{\lambda}(n_k+1)g(n_k)\frac{c_k\rho_k}{n_k+1}\prod_{i\in\Omega\setminus k}g(n_i)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&+ \frac{\sum_{l=0}^{K-1}\sum_{\#_l:n_k\geq c_k}\frac{1}{\lambda}(n_k+1)g(n_k)\rho_k\prod_{i\in\Omega\setminus k}g(n_i)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&= \frac{\sum_{l=0}^{K-1}\sum_{\#_l}\frac{1}{\lambda}(n_k+1)g(n_k+1)\prod_{i\in\Omega\setminus k}g(n_i)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)}.
\end{aligned}$$

In the last two steps we use the definition of g given in (4.1) for $n_k < c_k$ and $n_k \geq c_k$. Using this result we get an explicit expression for the total mean sojourn time of a customer in the network by

$$\begin{aligned}
\mathbb{E}[S] &= \Lambda_0\mathbb{E}[S_0] + \cdots + \Lambda_M\mathbb{E}[S_M] \\
&= \frac{\sum_{l=0}^{K-1}\sum_{\#_l}\frac{1}{\lambda}(n_0+1)g(n_0+1)\prod_{i\in\Omega\setminus 0}g(n_i)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&+ \cdots + \frac{\sum_{l=0}^{K-1}\sum_{\#_l}\frac{1}{\lambda}(n_M+1)g(n_M+1)\prod_{i\in\Omega\setminus M}g(n_i)}{\sum_{l=0}^{K-1}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)} \\
&= \frac{\sum_{l=1}^{K}\frac{1}{\lambda}\sum_{\#_l}(n_0+\ldots+n_M)\prod_{j\in\Omega}g(n_j)}{\sum_{i=0}^{K-1}\sum_{\#_i}\prod_{j\in\Omega}g(n_j)} \\
&= \frac{\sum_{l=1}^{K}\frac{l}{\lambda}\sum_{\#_l}\prod_{j\in\Omega}g(n_j)}{\sum_{i=0}^{K-1}\sum_{\#_i}\prod_{j\in\Omega}g(n_j)}.
\end{aligned} \tag{4.6}$$

4.2 Network With Buffer

Consider the same open network as considered in the previous section where a maximum of K customers are allowed. The only difference is that if a customer is blocked, he waits in a buffer with space for N customers. An exact expression for the mean sojourn time is hard to obtain since the second moment of the state-dependent sojourn time in the network is needed. Therefore, we approximate the probability distribution by obtaining the equilibrium equations as if the probabilities are memoryless like a queue with exponentially distributed service times. So, the equilibrium equations for the network can by approximation be written as:

$$\mathbb{P}(L=l) \approx \begin{cases} \mathbb{P}(L=0)\sum_{\#_l}\prod_{i\in\Omega}\rho_i^{n_i}, & l=0,\ldots,K, \\ \\ \Phi^{l-K}\,\mathbb{P}(L=K), & l=K,\ldots,N, \end{cases} \tag{4.7}$$

with

$$\Phi = \lambda\frac{\mathbb{E}_{\rm c}[S]}{K}, \tag{4.8}$$

where $\mathbb{E}_c[S]$ represents the sojourn time of one cycle, in other words: the time till a customer returns to the node where he is started, in a closed network which depends on K. To obtain this expression, we calculate the expectation of the sojourn time at a given station in a closed network with constant K customers with the same argumentation as done in (4.5): using the steady state distribution of a closed network as proved by Jackson in [10] given by

$$p_{\rm c}(n_0,\ldots,n_M) = \frac{\prod_{i\in\Omega}g(n_i)}{\sum_{\#_K}\prod_{i\in\Omega}g(n_i)},$$

for $n_0,\ldots,n_M \in \mathbb{N}$. We can derive the probability that an arriving customer has to wait at station $k=1,\ldots,M$ by

$$\phi_k = \frac{\sum_{\#_{K-1;n_k\geq c_k}}\prod_{i\in\Omega}g(n_i)}{\sum_{\#_{K-1}}\prod_{i\in\Omega}g(n_i)}.$$

The number of customers already waiting for service if a customer arrives is given by

$$\mathbb{E}_{\rm c}[L_k^{q*}] = \frac{\sum_{\#_{K-1;n_k\geq c_k}}(n_k-c_k)\prod_{i\in\Omega}g(n_i)}{\sum_{\#_{K-1}}\prod_{i\in\Omega}g(n_i)},$$

for $k=1,\ldots,M$. Then the mean sojourn time at station k can be obtained by

$$\mathbb{E}_{\rm c}[S_k] = \beta_k + \phi_k\frac{\beta_k}{c_k} + \mathbb{E}_{\rm c}[L_k^{q*}]\frac{\beta_k}{c_k}, \tag{4.9}$$

for $k=1,\ldots,M$. With (4.9) we can write for the total sojourn time at a given station:

$$\begin{aligned}
\Lambda_k\mathbb{E}_{\rm c}[S_k] &= \Lambda_k\beta_k + \Lambda_k\phi_k\frac{\beta_k}{c_k} + \Lambda_k\mathbb{E}_{\rm c}[L_k^{q*}]\frac{\beta_k}{c_k} \\
&= \Lambda_k\beta_k(1-\phi_k) + \Lambda_k(\mathbb{E}_{\rm c}[L_k^{q*}] + \phi_k c_k)\frac{\beta_k}{c_k} + \Lambda_k\phi_k\frac{\beta_k}{c_k} \\
&= \Lambda_k\beta_k\frac{\sum_{\#_{K-1:n_k<c_k}}\prod_{j\in\Omega}g(n_j)}{\sum_{\#_{K-1}}\prod_{j\in\Omega}g(n_j)} \\
&+ \frac{\Lambda_k\beta_k}{c_k}\frac{\sum_{\#_{K-1:n_k\geq c_k}}n_k\prod_{j\in\Omega}g(n_j)}{\sum_{\#_{K-1}}\prod_{j\in\Omega}g(n_j)} \\
&+ \frac{\Lambda_k\beta_k}{c_k}\frac{\sum_{\#_{K-1:n_k\geq c_k}}\prod_{j\in\Omega}g(n_j)}{\sum_{\#_{K-1}}\prod_{j\in\Omega}g(n_j)}
\end{aligned}$$

$$
\begin{aligned}
&= \Lambda_k \beta_k \frac{\sum_{\#_{K-1}: n_k < c_k} \prod_{j\in\Omega} g(n_j)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)} \\
&+ \frac{\Lambda_k \beta_k}{c_k} \frac{\sum_{\#_{K-1}: n_k \geq c_k} (n_k+1) \prod_{j\in\Omega} g(n_j)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)} \\
&= \frac{\sum_{\#_{K-1}: n_k < c_k} \frac{1}{\lambda}(n_k+1) g(n_k) \frac{c_k \rho_k}{n_k+1} \prod_{i\in\Omega\setminus k} g(n_i)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)} \\
&+ \frac{\sum_{\#_{K-1}: n_k \geq c_k} \frac{1}{\lambda}(n_k+1) g(n_k) \rho_k \prod_{i\in\Omega\setminus k} g(n_i)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)} \\
&= \frac{\sum_{\#_{K-1}} \frac{1}{\lambda}(n_k+1) g(n_k+1) \prod_{i\in\Omega\setminus k} g(n_i)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)},
\end{aligned}
$$

for $k = 1, \ldots, M$. Using this result we get an explicit expression for the total sojourn time of a customer in one cycle of a closed network with $K = 1, 2, \ldots$ customers by

$$
\begin{aligned}
\mathbb{E}_{\mathrm{c}}[S] &= \Lambda_0 \mathbb{E}_{\mathrm{c}}[S_0] + \cdots + \Lambda_M \mathbb{E}_{\mathrm{c}}[S_M] \\
&= \frac{\sum_{\#_{K-1}} \frac{1}{\lambda}(n_0+1) g(n_0+1) \prod_{i\in\Omega\setminus k} g(n_i)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)} \\
&+ \cdots + \frac{\sum_{\#_{K-1}} \frac{1}{\lambda}(n_M+1) g(n_M+1) \prod_{i\in\Omega\setminus k} g(n_i)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)} \\
&= \frac{\frac{1}{\lambda} \sum_{\#_K} (n_0 + \ldots + n_M) \prod_{j\in\Omega} g(n_j)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)} \\
&= \frac{K}{\lambda} \frac{\sum_{\#_K} \prod_{j\in\Omega} g(n_j)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)}.
\end{aligned} \tag{4.10}
$$

Substituting (4.10) into (4.8) gives

$$
\Phi = \frac{\sum_{\#_K} \prod_{j\in\Omega} g(n_j)}{\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)}. \tag{4.11}
$$

This expression makes clear that for $\Phi < 1$ the system is stable and $\lim_{l\to\infty} \mathbb{P}(L = l) = 0$ for $N \to \infty$. In other words the buffer will not always be full when the network is in steady state. On the other hand, if $\Phi \geq 1$, then the probability that l customers are in the system will increase. Using previous results we rewrite (4.7) as

$$
\mathbb{P}(L = l) \approx \begin{cases} \mathbb{P}(L = 0) \sum_{\#_l} \prod_{j\in\Omega} g(n_j), & l = 0, \ldots, K, \\ \mathbb{P}(L = 0) \frac{\left(\sum_{\#_K} \prod_{j\in\Omega} g(n_j)\right)^{l-K+1}}{\left(\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)\right)^{l-K}}, & l = K, \ldots, N. \end{cases} \tag{4.12}
$$

As a result we get the probability that no customers are in the system, or in other words the normalization factor, by

$$
\begin{aligned}
\mathbb{P}(L = 0) &\approx \left[\sum_{l=0}^{K} \sum_{\#_l} \prod_{j\in\Omega} g(n_j) + \sum_{j=K+1}^{K+N} \frac{\left(\sum_{\#_K} \prod_{j\in\Omega} g(n_j)\right)^{j-K+1}}{\left(\sum_{\#_{K-1}} \prod_{j\in\Omega} g(n_j)\right)^{j-K}} \right]^{-1} \\
&= \left[\sum_{l=0}^{K} \sum_{\#_l} \prod_{j\in\Omega} g(n_j) + \sum_{j=1}^{N} \Phi^j \sum_{\#_K} \prod_{j\in\Omega} g(n_j) \right]^{-1}.
\end{aligned} \tag{4.13}
$$

For the $M/G/1$-PS node, and thus also for an $M/M/1$ queueing node, we prove this result in an alternative way in Appendix A.2.

4.2.1 Blocking Probability

The probability that an arriving customer is blocked because there are already K customers in the network and N customers in the buffer, equals the probability that there are $K+N$ customers in the entire system. So, by PASTA, the blocking probability $\mathbb{P}$(customer is blocked) can be approximated as follows:

$$\begin{aligned}\mathbb{P}(\text{customer is blocked}) &= \mathbb{P}(L=K+N)\\ &\approx \frac{\left(\sum_{\#_K}\prod_{j\in\Omega} g(n_j)\right)^{N+1}}{\left(\sum_{\#_{K-1}}\prod_{j\in\Omega} g(n_j)\right)^{N}}\mathbb{P}(L=0)\\ &= \Phi^N\mathbb{P}(L=0)\sum_{\#_K}\prod_{j\in\Omega} g(n_j), \qquad (4.14)\end{aligned}$$

with Φ given in (4.11).

Numerical Results

To check the accuracy of (4.14), several experiments are done with a queueing network that consists of one PS node and two multi-server FCFS nodes. So for low as well high values of K and for low as well high values of ρ, the probability is derived by the formula given in (4.14) and is obtained by several simulations. The two FCFS nodes have 2 and 3 servers, respectively. The number of runs is taken large enough to ensure reliable results.
Table 4.1 shows that the computations correspond very well with the simulations. We simulated the same situation again, but with five back-end $M/M/c$ nodes. The five FCFS nodes have 2, 3, 1, 4 and 2 servers, respectively. We see in Table 4.2 that for low loads as well for high loads our proposed expressions seem to be accurate approximations.

4.2.2 Mean Number of Customers

Since we have derived the steady-state distribution we can obtain the mean number of customers in the entire system by

$$\mathbb{E}[L] = \sum_{l=0}^{K+N} l\,\mathbb{P}(L=l), \qquad (4.15)$$

with $\mathbb{P}(L=l)$ given in (4.12). If we exclude customers in the buffer, we get the mean number of customers which are just in the network by

$$\mathbb{E}[L_{\text{netw}}] = \sum_{l=0}^{K+N} \min\{l,K\}\mathbb{P}(L=l). \qquad (4.16)$$

The mean number of customers that are in the buffer can be computed by

$$\mathbb{E}[L_{\text{buf}}] = \sum_{l=0}^{N} l\,\mathbb{P}(L=K+l). \qquad (4.17)$$

Extending (4.16) we can write for the mean number of customers at node $k=1,2,\ldots$:

$$\mathbb{E}[L_k] = \left(\sum_{l=0}^{K}\sum_{\#l} n_k\prod_{j\in\Omega} g(n_j) + \sum_{j=1}^{N}\Phi^j\sum_{\#_K} n_k\prod_{j\in\Omega} g(n_j)\right)\mathbb{P}(L=0).$$

β_{PS}	β_F		K	N	$\mathbb{P}_s(L = K + N)$	$\mathbb{P}_a(L = K + N)$
1.0	2.0	1.0	1	1	0.88	0.88
1.0	2.0	1.0	1	5	0.88	0.88
1.0	2.0	1.0	1	20	0.88	0.88
1.0	2.0	1.0	2	1	0.79	0.79
1.0	2.0	1.0	2	5	0.79	0.79
1.0	2.0	1.0	2	20	0.79	0.79
1.0	2.0	1.0	5	1	0.70	0.70
1.0	2.0	1.0	5	5	0.69	0.69
1.0	2.0	1.0	5	20	0.69	0.69
1.0	2.0	1.0	10	1	0.67	0.67
1.0	2.0	1.0	10	5	0.67	0.67
1.0	2.0	1.0	10	20	0.67	0.67
1.0	2.0	1.0	15	1	0.67	0.67
1.0	2.0	1.0	15	5	0.67	0.67
1.0	2.0	1.0	15	20	0.67	0.67
0.2	0.6	0.4	1	1	0.55	0.52
0.2	0.6	0.4	1	5	0.47	0.46
0.2	0.6	0.4	1	20	0.46	0.46
0.2	0.6	0.4	2	1	0.26	0.24
0.2	0.6	0.4	2	5	0.10	0.08
0.2	0.6	0.4	2	20	0.02	0.01
0.2	0.6	0.4	5	1	0.02	0.02
0.2	0.6	0.4	5	5	0.00	0.00
0.2	0.6	0.4	5	20	0.00	0.00
0.2	0.6	0.4	10	1	0.00	0.00
0.2	0.6	0.4	10	5	0.00	0.00
0.2	0.6	0.4	10	20	0.00	0.00
0.2	0.6	0.4	15	1	0.00	0.00
0.2	0.6	0.4	15	5	0.00	0.00
0.2	0.6	0.4	15	20	0.00	0.00

Table 4.1: Probability that an arriving customer will be blocked by a queueing network with space for K customers, with a buffer for N customers and with two back-end nodes.

Numerical Results

To validate the approximation for the mean number of customers as given in (4.15), experiments are done in the same way as the previous simulations with a queueing network which consist of one PS node and two multi-server FCFS nodes which have 2 and 3 servers, respectively. The number of runs is taken large enough to ensure reliable results and the relative error is computed by

$$\Delta\mathbb{E}\% = \frac{\mathbb{E}_a[S] - \mathbb{E}_s[S]}{\mathbb{E}_s[S]} \cdot 100\%, \tag{4.18}$$

with $\mathbb{E}_a[S]$ the approximated value and $\mathbb{E}_s[S]$ the simulated value of the expectation.
Table 4.3 shows very promising results. To be sure these results can be generalized we simulated the same situation again, but with five back-end $M/M/c$ nodes. The five FCFS nodes have each 2, 3, 1, 4 and 2 servers, respectively. The result we see in Table 4.4 is that for low loads as well for high loads our proposed expression seems to be an accurate approximation for the mean number of customers in a system with a buffer.

4.2.3 Expectation of the Sojourn Time

With the approximations for the expectation of the number of customers in a network with capacity for K customers with a buffer for maximum N customers, we get the expectation of the sojourn time in the entire system by the sum of the mean time spend in the buffer plus the mean time spend in the network. By looking at the system as if it has an exponentially distributed total

β_{PS}	β_F					K	N	$\mathbb{P}_s(L=K+N)$	$\mathbb{P}_a(L=K+N)$
1.0	2.0	1.0	2.0	1.0	2.0	1	1	0.85	0.85
1.0	2.0	1.0	2.0	1.0	2.0	1	5	0.85	0.85
1.0	2.0	1.0	2.0	1.0	2.0	1	20	0.85	0.85
1.0	2.0	1.0	2.0	1.0	2.0	2	1	0.71	0.71
1.0	2.0	1.0	2.0	1.0	2.0	2	5	0.70	0.70
1.0	2.0	1.0	2.0	1.0	2.0	2	20	0.70	0.70
1.0	2.0	1.0	2.0	1.0	2.0	5	1	0.42	0.41
1.0	2.0	1.0	2.0	1.0	2.0	5	5	0.37	0.37
1.0	2.0	1.0	2.0	1.0	2.0	5	20	0.37	0.36
1.0	2.0	1.0	2.0	1.0	2.0	10	1	0.22	0.22
1.0	2.0	1.0	2.0	1.0	2.0	10	5	0.18	0.18
1.0	2.0	1.0	2.0	1.0	2.0	10	20	0.15	0.15
1.0	2.0	1.0	2.0	1.0	2.0	15	1	0.15	0.14
1.0	2.0	1.0	2.0	1.0	2.0	15	5	0.12	0.11
1.0	2.0	1.0	2.0	1.0	2.0	15	20	0.09	0.10
0.2	0.6	0.4	0.2	0.1	0.4	1	1	0.60	0.57
0.2	0.6	0.4	0.2	0.1	0.4	1	5	0.55	0.55
0.2	0.6	0.4	0.2	0.1	0.4	1	20	0.55	0.54
0.2	0.6	0.4	0.2	0.1	0.4	2	1	0.35	0.33
0.2	0.6	0.4	0.2	0.1	0.4	2	5	0.22	0.20
0.2	0.6	0.4	0.2	0.1	0.4	2	20	0.17	0.17
0.2	0.6	0.4	0.2	0.1	0.4	5	1	0.07	0.07
0.2	0.6	0.4	0.2	0.1	0.4	5	5	0.01	0.01
0.2	0.6	0.4	0.2	0.1	0.4	5	20	0.00	0.00
0.2	0.6	0.4	0.2	0.1	0.4	10	1	0.01	0.01
0.2	0.6	0.4	0.2	0.1	0.4	10	5	0.00	0.00
0.2	0.6	0.4	0.2	0.1	0.4	10	20	0.00	0.00
0.2	0.6	0.4	0.2	0.1	0.4	15	1	0.00	0.00
0.2	0.6	0.4	0.2	0.1	0.4	15	5	0.00	0.00
0.2	0.6	0.4	0.2	0.1	0.4	15	20	0.00	0.00

Table 4.2: Probability that an arriving customer will be blocked by a queueing network with space for K customers, with a buffer for N customers and with five back-end nodes.

service time and with usage of (4.17), we approximate the sojourn time in the buffer by

$$\begin{aligned}
\mathbb{E}[S_{\text{buf}}] &\approx \mathbb{E}[L_{\text{buf}}]\frac{\mathbb{E}_{\text{c}}[S](K)}{K} \\
&\approx \frac{\sum_{\#_K}\prod_{j\in\Omega} g(n_j)}{\sum_{\#_{K-1}}\prod_{j\in\Omega} g(n_j)}\sum_{l=0}^{N} l\frac{\left(\sum_{\#_K}\prod_{j\in\Omega} g(n_j)\right)^{l+1}}{\left(\sum_{\#_{K-1}}\prod_{j\in\Omega} g(n_j)\right)^{l}}\mathbb{P}(L=0) \\
&= \sum_{l=0}^{N} l\Phi^{l+1}\sum_{\#_K}\prod_{j\in\Omega} g(n_j)\mathbb{P}(L=0). \qquad (4.19)
\end{aligned}$$

According to the method used in Section 4.1.4 we approximate the probability that a customer has to wait at node $k=1,\ldots,M$ by

$$\pi_k \approx \left(\sum_{l=0}^{K}\sum_{\#l;n_k\geq c_k}\prod_{j\in\Omega} g(n_j) + \sum_{j=1}^{N-1}\Phi^j\sum_{\#_K;n_k\geq c_k}\prod_{j\in\Omega} g(n_j)\right)\mathbb{P}(L^*=0),$$

with the normalization constant for a network without the arriving customer given by

$$\mathbb{P}(L^*=0) = \left[\sum_{l=0}^{K}\sum_{\#_l}\prod_{j\in\Omega} g(n_j) + \sum_{j=1}^{N-1}\Phi^j\sum_{\#_K}\prod_{j\in\Omega} g(n_j)\right]^{-1}.$$

β_{PS}	β_F		K	N	$\mathbb{E}_s[L]$	$\mathbb{E}_a[L]$	$\Delta\mathbb{E}[L]\%$
1.0	2.0	1.0	1	1	1.87	1.88	-0.47
1.0	2.0	1.0	1	5	5.87	5.88	-0.22
1.0	2.0	1.0	1	20	20.87	20.87	-0.01
1.0	2.0	1.0	2	1	2.75	2.77	-0.81
1.0	2.0	1.0	2	5	6.74	6.77	-0.48
1.0	2.0	1.0	2	20	21.74	21.76	-0.10
1.0	2.0	1.0	5	1	5.57	5.60	-0.53
1.0	2.0	1.0	5	5	9.56	9.60	-0.44
1.0	2.0	1.0	5	20	24.56	24.59	-0.12
1.0	2.0	1.0	10	1	10.51	10.53	-0.18
1.0	2.0	1.0	10	5	14.51	14.54	-0.21
1.0	2.0	1.0	10	20	29.51	29.52	-0.04
1.0	2.0	1.0	15	1	15.50	15.51	-0.04
1.0	2.0	1.0	15	5	19.50	19.51	-0.04
1.0	2.0	1.0	15	20	34.50	34.49	0.03
0.2	0.6	0.4	1	1	1.39	1.40	-1.18
0.2	0.6	0.4	1	5	4.92	5.11	-3.76
0.2	0.6	0.4	1	20	19.82	20.08	-1.27
0.2	0.6	0.4	2	1	1.64	1.65	-0.62
0.2	0.6	0.4	2	5	3.39	3.39	0.20
0.2	0.6	0.4	2	20	8.35	7.59	9.88
0.2	0.6	0.4	5	1	1.98	1.99	-0.38
0.2	0.6	0.4	5	5	2.08	2.08	0.20
0.2	0.6	0.4	5	20	2.09	2.08	0.45
0.2	0.6	0.4	10	1	2.06	2.06	-0.12
0.2	0.6	0.4	10	5	2.06	2.05	0.21
0.2	0.6	0.4	10	20	2.06	2.06	-0.26
0.2	0.6	0.4	15	1	2.06	2.06	0.06
0.2	0.6	0.4	15	5	2.06	2.06	0.13
0.2	0.6	0.4	15	20	2.06	2.06	0.04

Table 4.3: Expectation of the number of customers in a queueing network with space for K customers, with a buffer for N customers and with two back-end nodes.

We define the number of customers which are already in the queue when an arbitrary customer arrives by

$$\mathbb{E}[L_k^{q*}] \approx \left(\sum_{l=0}^{K} \sum_{\#l;n_k \geq c_k} (n_k - c_k) \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N-1} \Phi^j \sum_{\#_K;n_k \geq c_k} (n_k - c_k) \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0),$$

for $k = 1, \ldots, M$. By assuming Poisson arrivals at each node we can approximate, according to (4.5), the mean sojourn time at the k-th station by

$$\mathbb{E}[S_k^\dagger] \approx \beta_k + \pi_k \frac{\beta_k}{c_k} + \mathbb{E}[L_k^{q*}] \frac{\beta_k}{c_k}, \tag{4.20}$$

for $k = 1, \ldots, M$. Note that this is just a naive approximation since (4.20) uses PASTA which is not correct. But we can use this expression to get another expression which will help us to get a nice approximation. With (4.20) we can write for the total sojourn time at a given station:

$$\begin{aligned}
\Lambda_k \mathbb{E}[S_k^\dagger] &= \Lambda_k \beta_k + \Lambda_k \pi_k \frac{\beta_k}{c_k} + \Lambda_k \mathbb{E}[L_k^{q*}] \frac{\beta_k}{c_k} \\
&= \Lambda_k \beta_k (1 - \pi_k) + \Lambda_k (\mathbb{E}[L_k^{q*}] + \pi_k c_k) \frac{\beta_k}{c_k} + \Lambda_k \pi_k \frac{\beta_k}{c_k} \\
&\approx \Lambda_k \beta_k \left(\sum_{l=0}^{K} \sum_{\#l;n_k < c_k} \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N-1} \Phi^j \sum_{\#_K;n_k < c_k} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0)
\end{aligned}$$

β_{PS}	β_F						K	N	$\mathbb{E}_s[L]$	$\mathbb{E}_a[L]$	$\Delta\mathbb{E}[L]\%$
1.0	2.0	1.0	2.0	1.0	2.0		1	1	1.83	1.85	-0.77
1.0	2.0	1.0	2.0	1.0	2.0		1	5	5.82	5.84	-0.39
1.0	2.0	1.0	2.0	1.0	2.0		1	20	20.82	20.83	-0.06
1.0	2.0	1.0	2.0	1.0	2.0		2	1	2.63	2.66	-1.04
1.0	2.0	1.0	2.0	1.0	2.0		2	5	6.57	6.64	-1.01
1.0	2.0	1.0	2.0	1.0	2.0		2	20	21.57	21.63	-0.26
1.0	2.0	1.0	2.0	1.0	2.0		5	1	4.89	4.92	-0.70
1.0	2.0	1.0	2.0	1.0	2.0		5	5	8.44	8.66	-2.61
1.0	2.0	1.0	2.0	1.0	2.0		5	20	23.26	23.58	-1.36
1.0	2.0	1.0	2.0	1.0	2.0		10	1	8.47	8.48	-0.21
1.0	2.0	1.0	2.0	1.0	2.0		10	5	11.49	11.66	-1.50
1.0	2.0	1.0	2.0	1.0	2.0		10	20	24.80	25.50	-2.74
1.0	2.0	1.0	2.0	1.0	2.0		15	1	11.93	11.90	0.25
1.0	2.0	1.0	2.0	1.0	2.0		15	5	14.77	14.82	-0.35
1.0	2.0	1.0	2.0	1.0	2.0		15	20	26.77	27.52	-2.70
0.2	0.6	0.4	0.2	0.1	0.4		1	1	1.48	1.51	-1.84
0.2	0.6	0.4	0.2	0.1	0.4		1	5	5.19	5.41	-4.04
0.2	0.6	0.4	0.2	0.1	0.4		1	20	20.17	20.40	-1.13
0.2	0.6	0.4	0.2	0.1	0.4		2	1	1.89	1.92	-1.81
0.2	0.6	0.4	0.2	0.1	0.4		2	5	4.56	4.81	-5.20
0.2	0.6	0.4	0.2	0.1	0.4		2	20	17.42	18.72	-6.97
0.2	0.6	0.4	0.2	0.1	0.4		5	1	2.74	2.74	-0.23
0.2	0.6	0.4	0.2	0.1	0.4		5	5	3.31	3.28	1.06
0.2	0.6	0.4	0.2	0.1	0.4		5	20	3.61	3.41	6.09
0.2	0.6	0.4	0.2	0.1	0.4		10	1	3.21	3.22	-0.24
0.2	0.6	0.4	0.2	0.1	0.4		10	5	3.28	3.28	0.02
0.2	0.6	0.4	0.2	0.1	0.4		10	20	3.29	3.30	-0.24
0.2	0.6	0.4	0.2	0.1	0.4		15	1	3.28	3.29	-0.21
0.2	0.6	0.4	0.2	0.1	0.4		15	5	3.29	3.30	-0.24
0.2	0.6	0.4	0.2	0.1	0.4		15	20	3.29	3.30	-0.22

Table 4.4: Expectation of the number of customers in a queueing network with space for K customers, with a buffer for N customers and with five back-end nodes.

$$
\begin{aligned}
&+ \quad \frac{\Lambda_k \beta_k}{c_k} \left(\sum_{l=0}^{K} \sum_{\#l;n_k \geq c_k} n_k \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N-1} \Phi^j \sum_{\#_K;n_k \geq c_k} n_k \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) \\
&+ \quad \frac{\Lambda_k \beta_k}{c_k} \left(\sum_{l=0}^{K} \sum_{\#l;n_k \geq c_k} \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N-1} \Phi^j \sum_{\#_K;n_k \geq c_k} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) \\
&= \quad \Lambda_k \beta_k \left(\sum_{l=0}^{K} \sum_{\#l;n_k < c_k} \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N-1} \Phi^j \sum_{\#_K;n_k < c_k} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) \\
&+ \quad \frac{\Lambda_k \beta_k}{c_k} \left(\sum_{l=0}^{K} \sum_{\#_l;n_k \geq c_k} (n_k + 1) \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N-1} \Phi^j \sum_{\#_K;n_k \geq c_k} (n_k + 1) \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) \\
&= \quad \left(\sum_{l=0}^{K} \sum_{\#_l;n_k < c_k} \frac{1}{\lambda} (n_k + 1) g(n_k) \frac{c_k \rho_k}{n_k + 1} \prod_{i \in \Omega \setminus k} g(n_i) \right) \mathbb{P}(L^* = 0) \\
&+ \quad \left(\sum_{j=1}^{N-1} \Phi^j \sum_{\#_K;n_k < c_k} \frac{1}{\lambda} (n_k + 1) g(n_k) \frac{c_k \rho_k}{n_k + 1} \prod_{i \in \Omega \setminus k} \right) \mathbb{P}(L^* = 0)
\end{aligned}
$$

$$
\begin{aligned}
&+ \left(\sum_{l=0}^{K} \sum_{\#_l; n_k \geq c_k} \frac{1}{\lambda}(n_k+1) g(n_k) \rho_k \prod_{i \in \Omega \setminus k} g(n_i) \right) \mathbb{P}(L^* = 0) \\
&+ \left(\sum_{j=1}^{N-1} \Phi^j \sum_{\#_K; n_k \geq c_k} \frac{1}{\lambda}(n_k+1) g(n_k) \rho_k \prod_{i \in \Omega \setminus k} g(n_i) \right) \mathbb{P}(L^* = 0) \\
&= \left(\sum_{l=0}^{K} \sum_{\#_l} \frac{1}{\lambda}(n_k+1) g(n_k+1) \prod_{i \in \Omega \setminus k} g(n_i) \right) \mathbb{P}(L^* = 0) \\
&+ \left(\sum_{j=1}^{N-1} \Phi^j \sum_{\#_K} \frac{1}{\lambda}(n_k+1) g(n_k+1) \prod_{i \in \Omega \setminus k} g(n_i) \right) \mathbb{P}(L^* = 0),
\end{aligned}
$$

for $k = 1, \ldots, M$. Using this result we get an explicit expression for the total sojourn time of a customer in the network by

$$
\begin{aligned}
\mathbb{E}[S_{\text{netw}}] &\approx \sum_{k \in \Omega} \Lambda_k \mathbb{E}[S_k^\dagger] \\
&= \left(\sum_{l=1}^{K} \sum_{\#_l} \frac{1}{\lambda}(n_0 + \ldots + n_M) \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) && (4.21) \\
&+ \left(\Phi \sum_{\#_K} \frac{K}{\lambda} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) && (4.22) \\
&+ \left(\sum_{j=1}^{N-1} \Phi^j \sum_{\#_K} \frac{K}{\lambda} \Phi \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) && (4.23) \\
&= \left(\sum_{l=1}^{K} \sum_{\#_l} \frac{l}{\lambda} \prod_{j \in \Omega} g(n_j) + \Phi \sum_{\#_K} \frac{K}{\lambda} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) \\
&+ \left(\sum_{j=2}^{N} \Phi^j \sum_{\#_K} \frac{K}{\lambda} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0) \\
&= \left(\sum_{l=1}^{K} \sum_{\#_l} \frac{l}{\lambda} \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N} \Phi^j \sum_{\#_K} \frac{K}{\lambda} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0). && (4.24)
\end{aligned}
$$

The result of (4.24) is obtained by the fact that although it is mathematically correct in (4.21) to take the sum over $\#_{K+1}$ instead of $\#_K$, the fact that no more than K customers are allowed in the network implies that the $(K+1)$-th customer has to wait in the buffer as expressed by (4.22). The same holds for (4.23). In this expression it is also mathematically correct to take the sum over $\#_{K+1}$ instead of $\#_K$, but since this is practically impossible, the $(K+1)$-th customer has to wait. For more clarity, compare these expressions with the probability function given by (4.12). For completeness, we can derive approximations for the mean sojourn times at a given node by

$$
\mathbb{E}[S_k] \approx \left(\sum_{l=1}^{K} \sum_{\#_l} \frac{n_k}{\lambda} \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N} \Phi^j \sum_{\#_K} \frac{n_k}{\lambda} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0).
$$

Combining (4.19) and (4.24) gives an approximation for the total sojourn time:

$$
\mathbb{E}[S] \approx \sum_{l=0}^{N} l \Phi^{l+1} \sum_{\#_K} \prod_{j \in \Omega} g(n_j) \mathbb{P}(L = 0)
$$

$$+ \quad \left(\sum_{l=1}^{K} \sum_{\#_l} \frac{l}{\lambda} \prod_{j \in \Omega} g(n_j) + \sum_{j=1}^{N} \Phi^j \sum_{\#_K} \frac{K}{\lambda} \prod_{j \in \Omega} g(n_j) \right) \mathbb{P}(L^* = 0). \tag{4.25}$$

Numerical Results

To validate the expression for the total sojourn time as given in (4.25), experiments are again done in the same way as previous simulations with a queueing network that consist of one PS node and two multi-server FCFS nodes. The two FCFS nodes still have 2 and 3 servers, respectively. Table 4.5 shows that the approximation for the mean sojourn time in the system performs very well for different configurations of the system.

β_{PS}	β_F		K	N	$\mathbb{E}_s[S]$	$\mathbb{E}_a[S]$	$\Delta\mathbb{E}[S]\%$
1.0	2.0	1.0	1	1	16.01	16.05	-0.25
1.0	2.0	1.0	1	5	49.87	49.78	0.18
1.0	2.0	1.0	1	20	177.37	175.73	0.93
1.0	2.0	1.0	2	1	13.34	12.91	3.28
1.0	2.0	1.0	2	5	32.19	32.05	0.43
1.0	2.0	1.0	2	20	103.88	103.40	0.47
1.0	2.0	1.0	5	1	18.27	18.19	0.49
1.0	2.0	1.0	5	5	31.11	30.59	1.70
1.0	2.0	1.0	5	20	79.93	79.23	0.89
1.0	2.0	1.0	10	1	31.85	32.10	-0.77
1.0	2.0	1.0	10	5	43.92	43.25	1.54
1.0	2.0	1.0	10	20	89.33	89.06	0.31
1.0	2.0	1.0	15	1	46.56	46.81	-0.53
1.0	2.0	1.0	15	5	58.57	57.92	1.12
1.0	2.0	1.0	15	20	103.62	102.62	0.97
0.2	0.6	0.4	1	1	2.86	2.96	-3.28
0.2	0.6	0.4	1	5	9.12	9.47	-3.66
0.2	0.6	0.4	1	20	36.67	37.23	-1.51
0.2	0.6	0.4	2	1	2.10	2.18	-3.57
0.2	0.6	0.4	2	5	3.47	3.44	0.82
0.2	0.6	0.4	2	20	7.93	7.20	10.10
0.2	0.6	0.4	5	1	2.01	2.02	-0.63
0.2	0.6	0.4	5	5	2.05	2.06	-0.32
0.2	0.6	0.4	5	20	2.05	2.06	-0.09
0.2	0.6	0.4	10	1	2.06	2.06	-0.03
0.2	0.6	0.4	10	5	2.06	2.06	0.11
0.2	0.6	0.4	10	20	2.06	2.07	-0.40
0.2	0.6	0.4	15	1	2.06	2.07	-0.35
0.2	0.6	0.4	15	5	2.06	2.06	0.13
0.2	0.6	0.4	15	20	2.06	2.06	-0.01

Table 4.5: Expectation of the sojourn times in a queueing network with space for K customers, with a buffer for N customers and with two back-end nodes.

To check if it is allowed to generalize these results, we simulated the same situation again but with five back-end $M/M/c$ nodes. The five FCFS nodes have each 2, 3, 1, 4 and 2 servers, respectively. We see in Table 4.6 that for low loads as well for high loads our proposed expressions seem to be accurate approximations for the mean sojourn time in the system.

Since it seems that the given approximations perform very well, we did some extra experiments to search for a worst case scenario. We found that for each configuration the approximations perform very well except for one situation: we found in a network with one PS node and two single-server FCFS nodes for N large, for $K = 1$ and for low loads, that the approximation performs worse with a maximum error up to 30%. But since this is not a realistic situation, we may conclude that the proposed approximations can be used as a forecasting tool which can be used in SLAs.

β_{PS}	β_F					K	N	$\mathbb{E}_s[S]$	$\mathbb{E}_a[S]$	$\Delta\mathbb{E}[S]\%$
1.0	2.0	1.0	2.0	1.0	2.0	1	1	12.22	12.21	0.05
1.0	2.0	1.0	2.0	1.0	2.0	1	5	38.42	38.49	-0.17
1.0	2.0	1.0	2.0	1.0	2.0	1	20	137.42	137.18	0.17
1.0	2.0	1.0	2.0	1.0	2.0	2	1	9.02	8.79	2.67
1.0	2.0	1.0	2.0	1.0	2.0	2	5	21.87	21.73	0.64
1.0	2.0	1.0	2.0	1.0	2.0	2	20	71.79	71.53	0.36
1.0	2.0	1.0	2.0	1.0	2.0	5	1	8.35	8.38	-0.38
1.0	2.0	1.0	2.0	1.0	2.0	5	5	13.39	13.35	0.28
1.0	2.0	1.0	2.0	1.0	2.0	5	20	36.64	36.72	-0.21
1.0	2.0	1.0	2.0	1.0	2.0	10	1	10.85	10.96	-1.04
1.0	2.0	1.0	2.0	1.0	2.0	10	5	13.95	14.01	-0.45
1.0	2.0	1.0	2.0	1.0	2.0	10	20	29.31	29.89	-1.95
1.0	2.0	1.0	2.0	1.0	2.0	15	1	13.95	13.99	-0.27
1.0	2.0	1.0	2.0	1.0	2.0	15	5	16.74	16.66	0.48
1.0	2.0	1.0	2.0	1.0	2.0	15	20	29.49	30.14	-2.15
0.2	0.6	0.4	0.2	0.1	0.4	1	1	3.52	3.54	-0.42
0.2	0.6	0.4	0.2	0.1	0.4	1	5	11.44	11.88	-3.74
0.2	0.6	0.4	0.2	0.1	0.4	1	20	44.37	44.82	-1.01
0.2	0.6	0.4	0.2	0.1	0.4	2	1	2.79	2.84	-1.48
0.2	0.6	0.4	0.2	0.1	0.4	2	5	5.64	5.76	-2.15
0.2	0.6	0.4	0.2	0.1	0.4	2	20	20.95	22.34	-6.24
0.2	0.6	0.4	0.2	0.1	0.4	5	1	2.92	2.94	-0.92
0.2	0.6	0.4	0.2	0.1	0.4	5	5	3.23	3.22	0.28
0.2	0.6	0.4	0.2	0.1	0.4	5	20	3.42	3.29	3.84
0.2	0.6	0.4	0.2	0.1	0.4	10	1	3.22	3.22	0.01
0.2	0.6	0.4	0.2	0.1	0.4	10	5	3.27	3.28	-0.28
0.2	0.6	0.4	0.2	0.1	0.4	10	20	3.28	3.29	-0.47
0.2	0.6	0.4	0.2	0.1	0.4	15	1	3.28	3.28	0.09
0.2	0.6	0.4	0.2	0.1	0.4	15	5	3.29	3.28	0.12
0.2	0.6	0.4	0.2	0.1	0.4	15	20	3.29	3.29	-0.17

Table 4.6: Expectation of the sojourn times in a queueing network with space for K customers, with a buffer for N customers and with five back-end nodes.

4.3 Buffer Size Comparisons

By usage of the promising results for admission control networks with and without a buffer given in previous two sections, we can make some comparisons. To show what the influence is of the buffer size, the left graph of Figure 4.2 gives some iso-loss curves for several blocking probabilities p. We use the same network as used in Table 4.3. So the model consists of one $M/M/1$-PS node, an $M/M/2$ FCFS node and an $M/M/3$ FCFS node with deterministic routing. Further, the following static parameters are used: arrival rate $\lambda = 1$, service time at the PS node $\beta_1 = 0.3$, service time at the $M/M/2$ node $\beta_2 = 1.6$ and service time at the $M/M/3$ node $\beta_3 = 0.9$. For a certain blocking probability the values of K and N are calculated. As an example, it becomes clear that to obtain a blocking probality of 0.5 just three customers are allowed in the network when the buffer size is less than two and that just two customers are allowed when the buffer size is larger than two. Consequently, to guarantee that the blocking probability is less than 0.5, K must be higher than three when N is less than two, and K must be higher than two when N is higher than two. Since differences in buffer size and network space must influence total sojourn times, the right graph of Figure 4.2 gives the delay curves belonging to the values of the iso-loss curves. For given N we obtain the value of K in the left graph where in the right graph for this K and this N the sojourn time is given by the curve with the same p-value as in the left graph. So, for example, the right graph of Figure 4.2 shows that for a blocking probability of 0.5 the sojourn time increases linearly when the buffer size increases.
The graphs of Figure 4.2 show the importance of determining right buffer sizes and blocking probabilities. When the blocking probability is taken too high, the number of allowed customers

in the network will be too small so waiting occurs mainly in the buffer. When the blocking probability is taken too small, the buffer size of the network hardly influences the total sojourn time, so the number of allowed customers in the network will be too high with as result that waiting occurs mainly in the network.

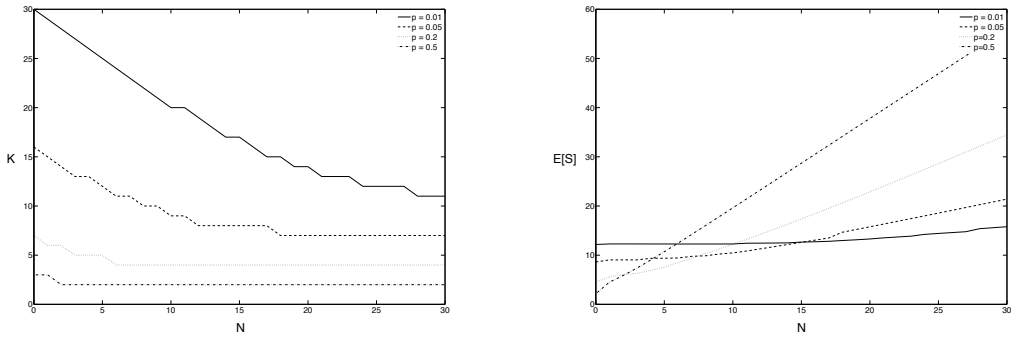

Figure 4.2: Iso-loss curves for several blocking probabilities (left) with associated delay curves (right).

Chapter 5

Application to Service Level Agreements

In this chapter we give an example how to compute some performance measures which can be used in SLAs. We apply some results obtained at networks with deterministic routing in Chapter 3, but of course results from other chapters can be used as well.

Consider the running example as given in Subsection 1.1.2. Suppose the Location-based Restaurant Service (LRS) has a Service Level Agreement (SLA) with its mobile end-users that the mean waiting time till they get a list of restaurants is less than 5 seconds with a maximum standard deviation of 2 seconds. Suppose the customers arrive according to a Poisson process at the application server (AS) with $\lambda_{AS} = 1$, and that the service times $B_{AS,1}$ and $B_{AS,2}$ of AS are exponentially distributed with mean service time $\beta_{AS,1} = \beta_{AS,2} = 0.1$ seconds. With these variables we are able to answer the key question given in Chapter 1: *"What combination of SLAs with other domains leads to desired QoS of the response time?"* Or more specifically: What is the SLA negotiation space of the LRS?

Mean Response Times: Since the total response time can be split up in the sojourn times of AS, the sojourn time of the Location Server (LS) and the sojourn time of the Restaurant Server (RS), we compute, using the first part of equation (3.8), that the total mean sojourn time in the AS equals 0.43 seconds. Since the mean response time must be less than 5 seconds, by average a tagged customer T can spend with a maximum of 4.57 seconds in the other servers. We know

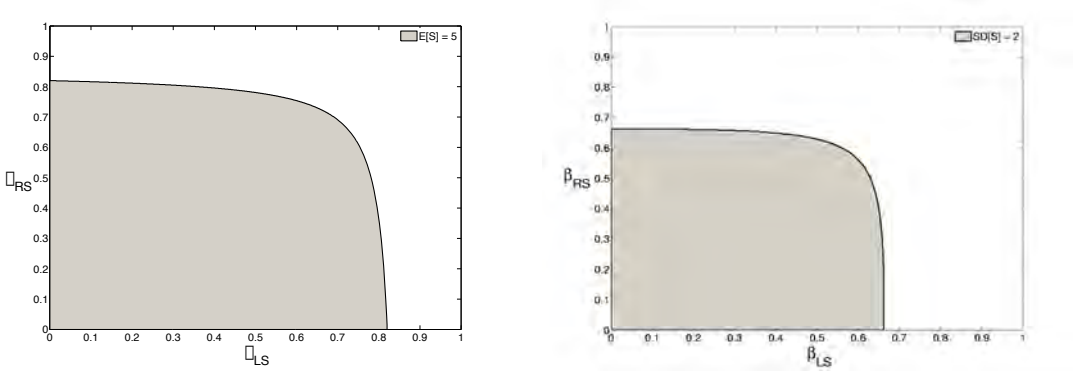

Figure 5.1: The SLA negotiation space with the range of all combinations of β_{LS} and β_{RS} where $\mathbb{E}[S] < 5$ (left) and the range of all combinations of β_{LS} and β_{RS} for which $\mathrm{SD}[S] < 2$ (right).

that the LS and the RS handle just one request a time. If there are more requests, T has to wait and will be served according to a First Come First Served (FCFS) discipline. With the second part of (3.8) we can compute all combinations of the mean service time β_{LS} of the LS and the mean service time β_{RS} of the RS, where the total sojourn time of T in both databases equals or is smaller than 4.57 seconds. With naive intuition we would expect the range of these service times is bounded by a linear curve since the higher the service time in one server, the lower it must be in the other server. But in the left graph of Figure 5.1 is to see that this range is non-linear. This will be explained the last part of this chapter. Using this negotation space and depending on the costs of a certain service time of a database, the LRS can minimize the total costs by making an SLA with LS using the optimal β_{LS} and making an SLA with RS using the optimal β_{RS}.

Standard Deviation of Response Times: By applying the same procedure as done in previous paragraph, we can determine the negotiation space of the LRS with the restriction that the standard deviation of the total sojourn time is smaller than or equals 2 seconds. By substituting the given parameters in (3.10), we obtain the total variance of the sojourn time in AS. Using this result in (3.9), we can compute all combinations of β_{LS} and β_{RS} where the standard deviation is smaller than or equals 2 seconds. The second graph of Figure 5.1 gives the range of these service times. If we need the two restrictions that the expectation of the sojourn time of T is at maximum 5 seconds and the sojourn times have a maximum standard deviation of 2 second, we keep just the smallest range of the combinations of β_{LS} and β_{RS}. It is obvious this range is given in the right graph of Figure 5.1.

In Figure 5.2 examples are given for some upper bounds of $\mathbb{E}[S]$ in the left graph and for some upper bounds of $\mathbb{V}\mathrm{ar}[S]$ in the right graph, where $(M+1)\beta_{AS} = 0.1$, $c_{LS} = 4$ and $c_{RS} = 2$ with $M = 2$. When the AS has the restriction that it must keep the expectation and the standard deviation below a certain level, the intersection of both value sets must be taken. In Figure 5.2, for example, an intersection of the values has to be taken if the expectation of the sojourn time may not exceed 2 seconds and the standard deviation may not exceed 1 second. Figure 5.2 shows also that the lower $\mathbb{E}[S]$, the "more linear" the border of the negotiation space. This can be explained by the fact that for low $\mathbb{E}[S]$ the service times must be low with as result that sojourn times almost equal service times, so almost no queueing exists. The case that for an high upper bound of the sojourn time the negotiation space has almost the form of a square can be explained by the fact that the longer customers have to wait in one queue of a network with deterministic routing, the shorter they have to wait in the other queue as long as the service time of the second queue is less than the first queue.

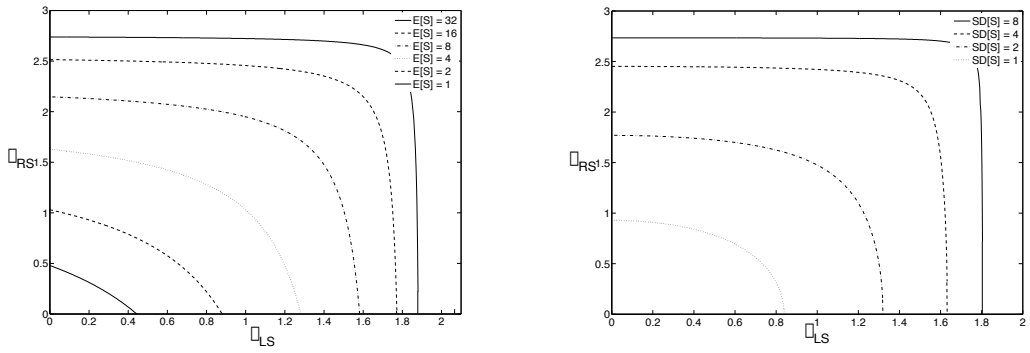

Figure 5.2: SLA negotiation spaces in terms of upper bounds for combinations of β_{LS} and β_{RS} for several values of $\mathbb{E}[S]$ (left) and upper bounds for combinations of β_{LS} and β_{RS} for several values of $\mathrm{SD}[S]$ (right).

Chapter 6

Summary and Further Research

6.1 Summary

In this thesis we have considered the key question *"What combination of Service Level Agreements (SLAs) with other domains leads to desired QoS of the response time?"* To this end, we analyzed the response time performance of several queueing networks.

In Chapter 2 we derived approximations for the second moment of the sojourn time in a network with one generally distributed PS node and one generally distributed FCFS node with Bernouilli feedback. Therefore we used approximations for the Laplace-Stieltjes Transforms (LSTs) of the total sojourn time. This probability distribution is approximated in three ways; by seeing the nodes as independent nodes, by seeing the nodes as short-circuited nodes and by an approximation which gives a weight to the previous two approximation methods. Although it is possible to obtain explicit approximations for the second moment of the sojourn time, simulation results show that they are more accurate for some configurations than for others. But from an application point-of-view it is the question if that matters.

For a network with deterministic routing which consists of one $M/G/1$-PS node and several $M/M/c$ FCFS back-end nodes, we derived in Chapter 3 exact expressions for the mean sojourn time and obtained accurate approximations for the variance of the sojourn times. Beside these results we obtained an improvement of an expression for the variance of the sojourn times in a network with Markovian routing as given in [8].

In Chapter 4 we investigated a network with a given routing matrix that may consist of several $M/G/1$-PS nodes and $M/M/c$ FCFS nodes. When K customers are in the network, new arriving customers will be blocked and disappear. For this network we derive exact expressions for the blocking probability, mean number of customers in the network and mean sojourn time in the network. As a variant of this network we investigated the availability of a buffer where customers can wait if they are blocked by the system. For this network accurate approximations are obtained for several performance measures.

In Chapter 5 we implemented some obtained results to answer the key question as given before. This resulted in several SLA negotiation spaces.

6.2 Further Research

There are many ways to extend the research described in this thesis. First, it seems to be possible to get accurate approximations for the mean sojourn time in a network which consists of multiple generally distributed back-end nodes using the LSTs of Chapter 2. Second, from an application

point-of-view it is interesting to investigate not only networks with Markovian or deterministic routing, but also networks with fork-join or fork-or types of construction. And networks with load balancing are also very challenging to investigate. Third, extending the results of networks with admission control by deriving approximations for networks with generally distributed FCFS nodes is also a way to develop founded results. Fourth, investigating the possibility to apply Buzen's convolution algorithm for computing efficiently the normalization constants of a network with finite capacity as explained in [11] and [12], is also very interesting.

Appendix A

Admission Control Results

In this appendix we prove some major results of Section 4.1 in an alternative way for networks that consist of $M/G/1$-PS nodes. For these computations we need some results of closed networks where the number of customers in the network is constant at K. Jackson proved in [10] that the steady-state distribution of the customers in the network is given by

$$p(n_0, n_1 \ldots, n_M) = \frac{\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_K}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}, \tag{A.1}$$

where we use the same notation as used in Chapter 4. From (A.1) we can derive the mean number of customers at each node by

$$\mathbb{E}_c[L_k](K) = \frac{\sum_{\#_K} n_k \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_K}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}, \tag{A.2}$$

for $k = 1,\ldots,M$. According to PASTA, by using the fact that the sojourn time at a node of an arriving customer equals the sum of service times of all customers already at the node plus the service time of the new customer self, we can derive from (A.2) the mean sojourn times of a customer at each PS node by

$$\begin{aligned}\mathbb{E}_c[S_k](K) &= \beta_k(\mathbb{E}_c[L_k](K-1)+1)\\ &= \beta_k\left(1+\frac{\sum_{\#_{K-1}} n_k \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}\right).\end{aligned} \tag{A.3}$$

With usage of (A.3) we can write for $\Lambda_k\mathbb{E}_c[S_k](K)$:

$$\begin{aligned}\Lambda_k\mathbb{E}_c[S_k](K) &= \Lambda_k\beta_k\left(1+\frac{\sum_{\#_{K-1}} n_k \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}\right)\\ &= \frac{\Lambda_k\beta_k\sum_{\#_{K-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}} + \frac{\Lambda_k\beta_k\sum_{\#_{K-1}} n_k\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}\\ &= \frac{\sum_{\#_{K-1}}(n_k+1)(\Lambda_k\beta_k)^{n_k+1}\prod_{i\in\Omega\setminus k}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}.\end{aligned} \tag{A.4}$$

With this result we can derive the mean total cycle time by

$$\begin{aligned}\mathbb{E}_c[S](K) &= \sum_{k\in\Omega}\Lambda_K\mathbb{E}_c[S_k](K)\\ &= \frac{\sum_{\#_{K-1}}(n_0+1)(\Lambda_0\beta_0)^{n_0+1}\prod_{i\in\Omega\setminus 0}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}\end{aligned}$$

$$
\begin{aligned}
&+ \quad \cdots + \frac{\sum_{\#_{K-1}} (n_M+1)(\Lambda_M\beta_M)^{n_M+1} \prod_{i\in\Omega\setminus M}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}} \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}} \\
&= \frac{\sum_{\#_K} (n_1+\cdots+n_M) \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}} \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}} \\
&= K\frac{\sum_{\#_K} \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{K-1}} \prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}. \qquad \text{(A.5)}
\end{aligned}
$$

A.1 Network without Buffer

With the previous results we can prove the probability distribution for a network in which at maximum K customers are allowed and which consists of several $M/G/1$-PS nodes. Consider an open network in which at maximum K customers are allowed and all remaining customers are blocked. Generalizing the insensitivity results given by Ott in [15] for the PS node to the network we consider, it is allowed to reason in the following way: if the network is in a state with l customers, the network will come in state $l+1$ with mean arrival rate λ and the network will come in state $l-1$ with mean service rate $l/\mathbb{E}_c[S](l)$ where $\mathbb{E}_c[S](l)$ represents the mean sojourn time of one cycle in a closed network with l customers. So we propose the following equilibrium equations:

$$
\begin{aligned}
&\lambda\mathbb{P}(L=l-1) = \frac{l}{\mathbb{E}_c[S](l)}\mathbb{P}(L=l), \qquad l=0,\ldots,K \\
&\sum_{l=0}^{K}\mathbb{P}(L=l) = 1.
\end{aligned}
$$

Iteration gives

$$
\begin{aligned}
\mathbb{P}(L=l) &= \frac{\lambda^l}{\prod_{i=1}^{l}\frac{i}{\mathbb{E}_c[S](i)}}\mathbb{P}(L=0) \\
&= \frac{\lambda^l}{l!}\prod_{i=1}^{l}\mathbb{E}_c[S](i)\mathbb{P}(L=0) \\
&= \frac{\lambda^l}{l!}\prod_{i=1}^{l} i\frac{\sum_{\#_i}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\sum_{\#_{i-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}\mathbb{P}(L=0) \\
&= \lambda^l\frac{\prod_{i=1}^{l}\sum_{\#_i}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\prod_{i=1}^{l}\sum_{\#_{i-1}}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}\mathbb{P}(L=0) \\
&= \lambda^l\frac{\left(\sum_{\#_l}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}\right)\prod_{i=1}^{l-1}\sum_{\#_i}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}{\prod_{i=0}^{l-1}\sum_{\#_i}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}}\mathbb{P}(L=0) \\
&= \lambda^l\sum_{\#_l}\prod_{i\in\Omega}(\Lambda_i\beta_i)^{n_i}\mathbb{P}(L=0) \\
&= \sum_{\#_l}\prod_{i\in\Omega}\rho_i^{n_i}\mathbb{P}(L=0),
\end{aligned}
$$

for $l=0,\ldots,K$. This in combination with the second equilibrium equation gives the steady-state distribution of the number of customers in an open network with finite capacity:

$$
\mathbb{P}(L=l) = \frac{\sum_{\#_l}\prod_{i\in\Omega}\rho_i^{n_i}}{\sum_{i=0}^{K}\sum_{\#_i}\prod_{i\in\Omega}\rho_i^{n_i}}, \qquad \text{(A.6)}
$$

for $l = 0, \ldots, K$. It is easy to see that the joint probability function of the number of customers in each queue is given by

$$p(n_0, n_1, \ldots, n_M) = \frac{\prod_{i\in\Omega} \rho_i^{n_i}}{\sum_{i=0}^{K} \sum_{\#_i} \prod_{i\in\Omega} \rho_i^{n_i}},$$

which equals expression (4.2). So we have proved the expression given by Jackson in [9] in an alternative way using other results of that paper.

A.2 Network with Buffer

Using the results of the previous section, we can approximate the probability distribution of a network in which at maximum K customers are allowed, but in which remaining customers can wait in a buffer of size N. We derive this approximation by obtaining the equilibrium equations as if the probabilities are memoryless like a queue with exponentially distributed service times. So, the equilibrium equations for the network can by approximation be written as:

$$\lambda \mathbb{P}(L = l-1) = \frac{\min\{l, K\}}{\mathbb{E}_c[S](\min\{l, K\})} \mathbb{P}(L = l), \qquad l = 1, \ldots, K+N,$$
$$\sum_{l=0}^{K} \mathbb{P}(L = l) = 1.$$

Iteration gives

$$\begin{aligned} \mathbb{P}(L = l) &= \frac{\lambda^l}{\prod_{i=1}^{l} \frac{i}{\mathbb{E}_c[S](i)}} \mathbb{P}(L = 0) \\ &= \mathbb{P}(L = 0) \sum_{\#_l} \prod_{i\in\Omega} \rho_i^{n_i}, \end{aligned} \tag{A.7}$$

for $l = 0, \ldots, K$, and

$$\mathbb{P}(L = K + l) = \left(\frac{\lambda}{\frac{K}{\mathbb{E}_c[S](K)}} \right)^l \mathbb{P}(L = K) \tag{A.8}$$

$$\begin{aligned} &= \left(\frac{\lambda \sum_{\#_K} \prod_{i\in\Omega} (\Lambda_i \beta_i)^{n_i}}{\sum_{\#_{K-1}} \prod_{i\in\Omega} (\Lambda_i \beta_i)^{n_i}} \right)^l \sum_{\#_K} \prod_{i\in\Omega} \rho_i^{n_i} \mathbb{P}(L = 0) \\ &= \frac{\lambda^{l(1-K)} \left(\sum_{\#_K} \prod_{i\in\Omega} \rho_i^{n_i} \right)^l}{\left(\sum_{\#_{K-1}} \prod_{i\in\Omega} (\Lambda_i \beta_i)^{n_i} \right)^l} \sum_{\#_K} \prod_{i\in\Omega} \rho_i^{n_i} \mathbb{P}(L = 0) \\ &= \frac{\left(\sum_{\#_K} \prod_{i\in\Omega} \rho_i^{n_i} \right)^{l+1}}{\left(\lambda^{K-1} \sum_{\#_{K-1}} \prod_{i\in\Omega} (\Lambda_i \beta_i)^{n_i} \right)^l} \mathbb{P}(L = 0) \\ &= \frac{\left(\sum_{\#_K} \prod_{i\in\Omega} \rho_i^{n_i} \right)^{l+1}}{\left(\sum_{\#_{K-1}} \prod_{i\in\Omega} \rho_i^{n_i} \right)^l} \mathbb{P}(L = 0), \end{aligned} \tag{A.9}$$

for $l = 0, \ldots, N$. If we define

$$\Phi = \frac{\sum_{\#_K} \prod_{i\in\Omega} \rho_i^{n_i}}{\sum_{\#_{K-1}} \prod_{i\in\Omega} \rho_i^{n_i}},$$

then by combining (A.7) and (A.9) we get the probability that no customers are in the system, or in other words the normalization factor:

$$\begin{aligned}\mathbb{P}(L=0) &= \left[\sum_{l=0}^{K}\sum_{\#_l}\prod_{i\in\Omega}\rho_i^{n_i} + \sum_{j=K+1}^{K+N}\frac{\left(\sum_{\#_K}\prod_{i\in\Omega}\rho_i^{n_i}\right)^{j-K+1}}{\left(\sum_{\#_{K-1}}\prod_{i\in\Omega}\rho_i^{n_i}\right)^{j-K}}\right]^{-1} \\ &= \left[\sum_{l=0}^{K}\sum_{\#_l}\prod_{i\in\Omega}\rho_i^{n_i} + \sum_{j=1}^{N}\Phi^j\sum_{\#_K}\prod_{i\in\Omega}\rho_i^{n_i}\right]^{-1}.\end{aligned}$$

These results can also explicitly be found in Section 4.2 for $M/M/c$ queues, but previous results prove that they are also valid for $M/G/1$-PS nodes which shows that it is allowed to generalize the results of Section 4.2 to $M/G/1$-PS nodes.

Bibliography

[1] O.J. Boxma and H. Daduna. Sojourn times in queueing networks. *Stochastic Analysis of Computer and Communication Systems*, pages 401–450, 1990.

[2] O.J. Boxma, B.M.M. Gijsen, R.D. van der Mei, and J.A.C. Resing. Sojourn time approximations in two-node queueing networks. In *Proceedings 2nd international working conference on Performance Modelling and Evaluation of Heterogeneous Networks, HETNETs*, 2004.

[3] 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 of the 19th International Teletraffic Congress - ITC 19*, pages 1121–1133. Beijing, 2005.

[4] X. Chen, P. Mohapatra, and H. Chen. An admission control scheme for predictable server response time for web accesses. In *World Wide Web*, pages 545–554, 2001.

[5] E.G. Coffman, R.R. Muntz, and H. Trotter. Waiting time distributions for processor-sharing systems. *Journal of the ACM*, 17(1):123–130, 1970.

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

[7] B.M.M. Gijsen, P.J. Meulenhoff, M.A. Blom, R.D. van der Mei, and B.D. van der Waaij. Web admission control: improving performance of web-based services. In *Proceedings Computer Measurements Group international conference, CMG*, 2004.

[8] B.M.M. Gijsen, R.D. van der Mei, P. Engelberts, J.L. van den Berg, and K.M.C. van Wingerden. Response times in queueing networks with feedback. *Performance Evaluation*, 63:743–758, 2006.

[9] J.R. Jackson. Networks of waiting lines. *Operations Research*, 5:518–521, 1957.

[10] J.R. Jackson. Jobshop-Like queueing systems. *Management Science*, 10(1):131–142, 1963.

[11] L. Kleinrock. *Queuing Systems Volume II: Computer Applications*. John Wiley & Sons, New York, 1976.

[12] S.S. Lavenberg, editor. *The Computer Performance Modeling Handbook*. Academic Press, San Diego, CA, 1983.

[13] S.S. Lavenberg and M. Reiser. Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *Journal Applied Probability*, 17(4):1048–1061, 1980.

[14] J.A. Morrison. Response-time distribution for a processor-sharing system. *SIAM Journal on Applied Mathematics*, 45(1):152–167, 1985.

[15] T.J. Ott. The sojourn time distribution in the M/G/1 queue with processor sharing. *Journal of Applied Probability*, 21:360–378, 1984.

[16] I. Sorteberg and O. Kure. The use of service level agreements in tactical military coalition force networks. *IEEE Communication Magazine*, pages 107– 114, 2005.

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

[18] 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(4):365–402, 1991.

[19] 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, 2006. *To appear*.

[20] R.W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, New York, 1989.