Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

ANDREAS BRANDT, MANFRED BRANDT

On a Two-Queue Priority System with Impatience and its Application to a Call Center

Preprint SC 98-21 (August 1998)

On a Two-Queue Priority System with Impatience and its Application to a Call Center[1]

Andreas Brandt

Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin,
Spandauer Str. 1, D-10178 Berlin, Germany
e-mail: brandt@wiwi.hu-berlin.de

Manfred Brandt

Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB),
Takustr. 7, D-14195 Berlin, Germany
e-mail: brandt@zib.de

Abstract

We consider a s-server system with two FCFS queues, where the arrival rates at the queues and the service rate may depend on the number n of customers being in service or in the first queue, but the service rate is assumed to be constant for $n > s$. The customers in the first queue are impatient. If the offered waiting time exceeds a random maximal waiting time I, then the customer leaves the first queue after time I. If I is less than a given deterministic time then he leaves the system else he transits to the end of the second queue. The customers in the first queue have priority. The service of a customer from the second queue will be started if the first queue is empty and more than a given number of servers become idle. For the model being a generalization of the $M(n)/M(n)/s+GI$ system balance conditions for the density of the stationary state process are derived yielding the stability conditions and the probabilities that precisely n customers are in service or in the first queue. For obtaining performance measures for the second queue a system approximation basing on fitting impatience intensities is constructed. The results are applied to the performance analysis of a call center with an integrated voice-mail-server. For an important special case a stochastic decomposition is derived illuminating the connection to the dynamics of the $M(n)/M(n)/s+GI$ system.

Mathematics Subject Classification (MSC 1991): 60K25, 68M20, 60G10.

Keywords: two queues; many-server; server reservation; impatience; occupancy distribution; waiting time distribution; approximate system; $M(n)/M(n)/s+GI$; stochastic decomposition; call center application.

1 Introduction

In this paper we analyze a general two-queue s-server priority system with state dependent arrival and service rates, server reservation and impatient customers in the protected queue. The results are applied to the performance analysis of a call center with an integrated voice-mail-server.

The general model, cf. Figure 1.1., consists of two FCFS queues denoted by Q and Q', respectively, and s servers. The arrival rates λ_n and λ'_n at Q and Q', respectively, are allowed to

[1] This work was supported by a grant from the Siemens AG.

depend on the number n of customers being in service or in Q. The customers in Q are impatient, i.e., each customer arriving at Q has a random maximal waiting time I. If the offered waiting time W^o (i.e., the time which the customer would have to wait for accessing a server if he were sufficiently patient) exceeds I, then the customer leaves Q after time I: If $I < \tau$, where $\tau \in \mathbb{R}_+$ is a given deterministic time (decision parameter), then the customer leaves the system (gets lost), else he transits to the end of Q'. The maximal waiting times are assumed to be i.i.d. with distribution function $C(u) = P(I \leq u)$. As soon as any server is idle, the next customer from Q – provided there is anyone – will be served. The customers in Q' are not impatient. Only if more than a servers, where $a \in \{0, 1, \dots, s-1\}$ is a given parameter, are idle and no customers are waiting in Q then one of the idle servers starts serving a customer from Q' (server reservation for Q being protected). The cumulative rate μ_n of finishing service of a customer is allowed to depend on the number n of customers being in service or in Q, but the rate is assumed to be constant for $n > s$, i.e., the rate of finishing service may only depend on the number of busy servers and additionally whether there are customers waiting for service in Q.

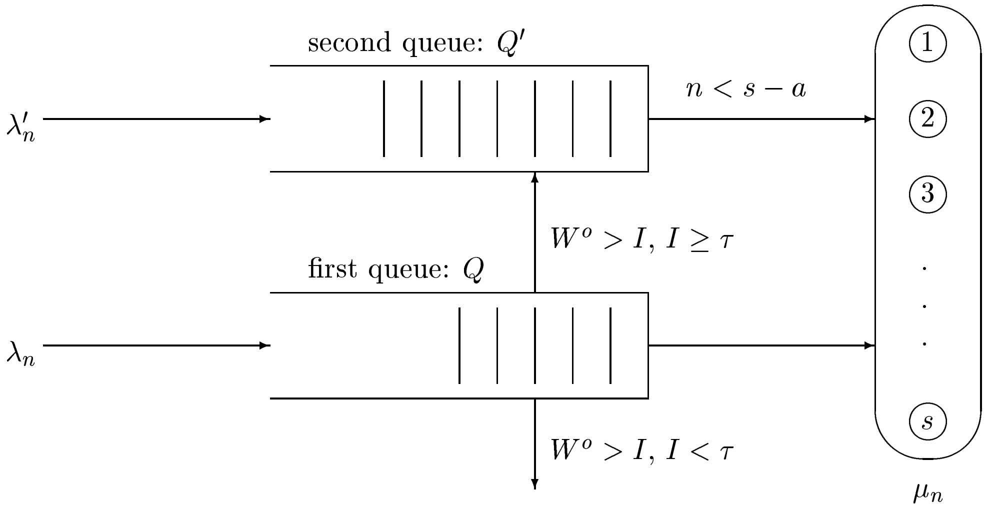

Figure 1.1. Two-queue s-server system with server reservation, impatient customers in the first queue Q, state dependent arrival and service rates, where n denotes the number of customers being in service or in Q.

Note, that for the mathematical analysis – given later in the paper – it is crucial that the arrival and service rates may only depend on the number n of customers being in service or in Q and not on the number of customers in Q'. However, two interesting and very different limiting cases are obtained from this model: In case of $\tau = \infty$, $\lambda'_n \equiv 0$ the model reduces to a $M(n)/M(n)/s+GI$ system (s-server system with impatient customers) which was analyzed by several authors, cf. [BB], [BH], [H], [J1], [J2], [GK], [HS], [W]; in case of $\tau = 0$ the model reduces to a two-queue s-server system with server reservation and transition from the protected to the end of the unprotected queue after waiting time I.

The main results and the organization of the paper are as follows. In Section 2 we derive for the general model, cf. Figure 1.1, a system of balance conditions for the density of the stationary vector process of the number n of customers being in service or in Q, the residual maximal waiting times and original maximal waiting times of customers waiting in Q. The system of balance conditions is just of the same structure as those for the $M(n)/M(n)/s$ queue with impatient customers investigated in [BB]. An application of the results in [BB] and of the conservation principle for stationary point processes yields the stability conditions and an explicit representation for the probabilities that precisely n customers are in service or in Q. However, for the stationary occupancy distribution for Q' no explicit formulae are available. Thus, in Section 3 an approximate system is constructed by replacing the impatience mechanism in Q by waiting place dependent impatience rates. The stability conditions and the probabilities that precisely n customers are in service or in Q are the same as in the exact model. For the factorial moments of the occupancy distribution for Q' in the approximate system a recursive algorithm is developed.

In Section 4 the results of Section 2 and Section 3 are applied to the performance analysis of a call center with an integrated voice-mail-server (VMS). Important performance measures can be computed exactly: stability condition, blocking probability, impatience probability, waiting time distribution in the waiting room etc. The first moment of the occupancy distribution for Q' in the approximate system of the general model, cf. Section 3, yields an approximation for the mean number of calls in the VMS and, in view of Little's formula, also an approximation for the mean waiting time in the VMS. Simulations of the system have shown that this approximation works well. Corresponding numerical results are given.

In the Appendix for a call center consisting of s servers, k waiting places and integrated VMS (special case $\lambda_n = 1\!\!1\{0 \leq n < s+k\}\lambda$ for some positive integer k, $\lambda'_n = 0$, $\mu_n = \min(n,s)\mu$ for $n = 0,1,2,\ldots$ and $I \equiv \tau$ in the model of Figure 1.1) a stochastic decomposition is given in terms of the $M(n)/M(n)/s$ queue with impatient calls. This decomposition result illuminates additionally the connection between the $M(n)/M(n)/s$ queue with impatient calls investigated in [BB] and the stochastics of the system considered in this paper.

2 A system of balance conditions and the stationary occupancy distribution for Q

As above denote by n the number of customers being in service or in Q and by n' the number of customers in Q'. Let in the following $\ell := (n-s)_+$. In view of the system dynamics then there are ℓ customers waiting in Q. We number the waiting customers in Q and Q' according to their positions in Q and Q', respectively. Thus, by the FCFS discipline the first customer in each queue will be potentially the next from this queue for service. Throughout this paper we make the following assumption concerning the arrival and service rates.

Assumption 2.1. *Let λ_n and λ'_n be bounded and either $\lambda_n > 0$ for $n \geq 0$ or there is a positive integer k such that $\lambda_n > 0$ for $0 \leq n < s+k$ and $\lambda_n \equiv 0$ for $n \geq s+k$. Let $\lambda'_n \equiv 0$ for $n < s-a$. Further, let $\mu_0 = 0$, $\mu_n > 0$ for $s-a \leq n \leq s$ and $\mu_n \equiv \mu_* > 0$ for $n > s$.*

Remark 2.2. *The assumption $\lambda'_n \equiv 0$ for $n < s-a$ has technical reasons only. The general case is obtained by redefining the arrival intensities for $n < s-a$: λ_n as the sum of the arrival intensities at Q and Q' as well as $\lambda'_n :\equiv 0$.*

We assume that the system is stable (the stability conditions will be given later) and that $C(u)$ is non-defective and has a continuous density, for the general case see later. Let us introduce the following random variables and probabilities:

$N(t)$ – sum of the number of customers in service and of the number of customers in Q at time t,

$L(t) := (N(t) - s)_+$ – number of customers in Q at time t,

$N'(t)$ – number of customers in Q' at time t,

$(X_1(t), \ldots, X_{L(t)}(t))$ – vector of the residual maximal waiting times of waiting customers in Q ordered according to their positions in Q at time t,

$(I_1(t), \ldots, I_{L(t)}(t))$ – vector of the original maximal waiting times of waiting customers in Q ordered according to their positions in Q at time t,

$p^*(n, n') := P(N(t) = n, N'(t) = n')$

– stationary distribution of the vector $(N(t), N'(t))$,

$$P^*(n, n';\, x_1, \ldots, x_\ell;\, u_1, \ldots, u_\ell) := P(N(t) = n, N'(t) = n';\ X_1(t) \le x_1, \ldots, X_\ell(t) \le x_\ell;\\ I_1(t) \le u_1, \ldots, I_\ell(t) \le u_\ell)$$

– stationary distribution on $(N(t), N'(t)) = (n, n')$,

$p(n) := P(N(t) = n)$ – stationary distribution of $N(t)$,

$$P(n;\, x_1, \ldots, x_\ell;\, u_1, \ldots, u_\ell) := P(N(t) = n;\ X_1(t) \le x_1, \ldots, X_\ell(t) \le x_\ell;\\ I_1(t) \le u_1, \ldots, I_\ell(t) \le u_\ell)$$

– stationary distribution on $N(t) = n$.

Obviously, for fixed $n > s$, $n' \ge 0$ the support of $P^*(n, n'; x_1, \ldots, x_\ell; u_1, \ldots, u_\ell)$ is contained in

$$\Omega_\ell := \{(x_1, \ldots, x_\ell; u_1, \ldots, u_\ell) \in I\!\!R_+^{2\ell} \,:\, u_1 - x_1 \ge \ldots \ge u_\ell - x_\ell \ge 0\}. \tag{2.1}$$

In view of the assumptions on $C(u)$ the densities

$$p^*(n, n'; x_1, \ldots, x_\ell; u_1, \ldots, u_\ell) := \frac{\partial^{2\ell}}{\partial x_1 \cdot \ldots \cdot \partial x_\ell \partial u_1 \cdot \ldots \cdot \partial u_\ell} P^*(n, n'; x_1, \ldots, x_\ell; u_1, \ldots, u_\ell),$$

$$p(n; x_1, \ldots, x_\ell; u_1, \ldots, u_\ell) := \frac{\partial^{2\ell}}{\partial x_1 \cdot \ldots \cdot \partial x_\ell \partial u_1 \cdot \ldots \cdot \partial u_\ell} P(n; x_1, \ldots, x_\ell; u_1, \ldots, u_\ell)$$

are continuous on Ω_ℓ.

In case of $0 \le n < s$, $n' \ge 0$ we have the balance equations

$$\begin{aligned}
&(\lambda_n + \lambda'_n + \mu_n) p^*(n, n') \\
&\quad = \mathrm{1\!I}\{n > 0\}\lambda_{n-1} p^*(n-1, n') + \mathrm{1\!I}\{n' > 0\}\lambda'_n p^*(n, n'-1) \\
&\qquad + (\mathrm{1\!I}\{n+1 \ne s-a\} + \mathrm{1\!I}\{n+1 = s-a\}\mathrm{1\!I}\{n' = 0\})\mu_{n+1} p^*(n+1, n') \\
&\qquad + \mathrm{1\!I}\{n = s-a\}\mu_n p^*(n, n'+1)
\end{aligned} \tag{2.2}$$

and in case of $n = s$, $n' \ge 0$ the balance equations

$$(\lambda_s + \lambda'_s + \mu_s) p^*(s, n')$$

$$
\begin{aligned}
&= \lambda_{s-1}p^*(s-1,n') + 1\!\!1\{n'>0\}\lambda'_s p^*(s,n'-1) \\
&\quad + \int_0^\tau p^*(s+1,n';0;u)\mathrm{d}u + 1\!\!1\{n'>0\}\int_\tau^\infty p^*(s+1,n'-1;0;u)\mathrm{d}u \\
&\quad + \mu_* \int_{I\!R_+^2} p^*(s+1,n';x;u)\mathrm{d}x\mathrm{d}u + 1\!\!1\{a=0\}\mu_s p^*(s,n'+1). \qquad (2.3)
\end{aligned}
$$

In case of $n > s$, $n' \geq 0$ and $(x_1,\ldots,x_\ell;u_1,\ldots,u_\ell) \in \Omega_\ell$, in view of (2.1) especially implying $0 \leq x_\ell \leq u_\ell$, we have the balance conditions

$$
\begin{aligned}
&p^*(n,n';x_1,\ldots,x_\ell;u_1,\ldots,u_\ell) \\
&= p^*(n,n';x_1{+}h,\ldots,x_\ell{+}h;u_1,\ldots,u_\ell)(1-h\lambda_n-h\lambda'_n-h\mu_*) \\
&\quad + h1\!\!1\{n'>0\}\lambda'_n p^*(n,n'{-}1;x_1,\ldots,x_\ell;u_1,\ldots,u_\ell) \\
&\quad + h\sum_{i=1}^{\ell+1}\int_0^\tau p^*(n{+}1,n';x_1,\ldots,x_{i-1},0,x_i,\ldots,x_\ell;u_1,\ldots,u_{i-1},u,u_i,\ldots,u_\ell)\mathrm{d}u \\
&\quad + h1\!\!1\{n'>0\}\sum_{i=1}^{\ell+1}\int_\tau^\infty p^*(n{+}1,n'{-}1;x_1,\ldots,x_{i-1},0,x_i,\ldots,x_\ell; \\
&\qquad\qquad\qquad u_1,\ldots,u_{i-1},u,u_i,\ldots,u_\ell)\mathrm{d}u \\
&\quad + h\mu_*\int_{I\!R_+^2} p^*(n{+}1,n';x,x_1,\ldots,x_\ell;u,u_1,\ldots,u_\ell)\mathrm{d}x\mathrm{d}u + o(h), \quad h>0, \quad x_\ell < u_\ell, \qquad (2.4)
\end{aligned}
$$

$$
\begin{aligned}
&p^*(n,n';x_1,\ldots,x_{\ell-1},u_\ell;u_1,\ldots,u_\ell) \\
&= \lambda_{n-1}p^*(n-1,n';x_1,\ldots,x_{\ell-1};u_1,\ldots,u_{\ell-1})c(u_\ell). \qquad (2.5)
\end{aligned}
$$

Unfortunately, there seems to be no explicit solution for (2.2) – (2.5). Hence, we will deal with the marginal system of balance conditions for the customers being in service or in Q.

Summing over $n' \in Z\!\!\!Z_+$ in (2.2) yields that for $0 \leq n < s$

$$
\begin{aligned}
&(\lambda_n + 1\!\!1\{n \neq s-a\}\mu_n + 1\!\!1\{n = s-a\}\mu_n p_0)p(n) \\
&= 1\!\!1\{n>0\}\lambda_{n-1}p(n-1) + 1\!\!1\{n+1 \neq s-a\}\mu_{n+1}p(n+1) \\
&\quad + 1\!\!1\{n+1 = s-a\}\mu_{n+1}p_0 p(n+1), \qquad (2.6)
\end{aligned}
$$

where $p_0 := p^*(s{-}a,0)/p(s{-}a)$ is the conditional probability that Q' is empty conditioned upon $N(t) = s{-}a$. Using the notation

$$
\mu'_n := (1\!\!1\{n = s{-}a\}p_0 + 1\!\!1\{n \neq s{-}a\})\mu_n, \qquad n = 0,1,\ldots \qquad (2.7)
$$

from (2.6) we obtain that for $0 \le n < s$

$$(\lambda_n + \mu'_n)p(n) = 1\!\!1\{n > 0\}\lambda_{n-1}p(n-1) + \mu'_{n+1}p(n+1). \tag{2.8}$$

From (2.3) by summing over $n' \in \mathbb{Z}_+$ and using (2.7) it follows

$$(\lambda_s + \mu'_s)p(s) = \lambda_{s-1}p(s-1) + \int\limits_{\mathbb{R}_+} p(s+1;0;u)\mathrm{d}u + \mu_* \int\limits_{\mathbb{R}^2_+} p(s+1;x;u)\mathrm{d}x\mathrm{d}u. \tag{2.9}$$

Moreover, in case of $n > s$ and $(x_1,\ldots,x_\ell;u_1,\ldots,u_\ell) \in \Omega_\ell$ we have the marginal balance conditions

$$\begin{aligned} & p(n;x_1,\ldots,x_\ell;u_1,\ldots,u_\ell) \\ = \; & p(n;x_1+h,\ldots,x_\ell+h;u_1,\ldots,u_\ell)(1-h\lambda_n-h\mu_*) \\ & + h\sum_{i=1}^{\ell+1} \int\limits_{\mathbb{R}_+} p(n+1;x_1,\ldots,x_{i-1},0,x_i,\ldots,x_\ell;u_1,\ldots,u_{i-1},u,u_i,\ldots,u_\ell)\mathrm{d}u \\ & + h\mu_* \int\limits_{\mathbb{R}^2_+} p(n+1;x,x_1,\ldots,x_\ell;u,u_1,\ldots,u_\ell)\mathrm{d}x\mathrm{d}u + o(h), \quad h>0, \quad x_\ell < u_\ell, \end{aligned} \tag{2.10}$$

$$p(n;x_1,\ldots,x_{\ell-1},u_\ell;u_1,\ldots,u_\ell) = \lambda_{n-1}p(n-1;x_1,\ldots,x_{\ell-1};u_1,\ldots,u_{\ell-1})c(u_\ell). \tag{2.11}$$

The system of equations (2.8) – (2.11) coincides with (2.3), (2.4), (2.7), (2.8) (with different parameters) in [BB], characterizing there the steady state distribution of a s-server queueing system with impatient customers and state dependent arrival and service rates, denoted by $M(n)/M(n)/s+GI$. (The connection to this queueing system is additionally illuminated in the Appendix for an important special case by a stochastic decomposition.)

The assumptions on our system imply that equations (2.3), (2.4), (2.7) and (2.8) in [BB] have exactly one normalized solution. Thus, for $n \le s$ from (2.10) in [BB] we obtain

$$p(n) = g\Big(\prod_{i=0}^{n-1}\lambda_i\Big)\Big(\prod_{i=n+1}^{s}\mu'_i\Big), \qquad n \le s. \tag{2.12}$$

For $n > s$ from (2.17) in [BB] for the density $p(n;x_1,\ldots,x_\ell;u_1,\ldots,u_\ell)$ we find the expression

$$\begin{aligned} & p(n;x_1,\ldots,x_\ell;u_1,\ldots,u_\ell) \\ = \; & 1\!\!1\{(x_1,\ldots,x_\ell;u_1,\ldots,u_\ell)\in\Omega_\ell\}\, g\left(\prod_{i=0}^{n-1}\lambda_i\right)\left(\prod_{i=1}^{\ell}c(u_i)\right)e^{-\mu_*(u_1-x_1)}, \qquad n > s, \end{aligned} \tag{2.13}$$

where $g > 0$ is a normalizing factor. In view of (2.7) from (2.12) it follows

$$p(n) = (1\!\!1\{n < s-a\}p_0 + 1\!\!1\{n \ge s-a\})\, g\left(\prod_{i=0}^{n-1}\lambda_i\right)\Big(\prod_{i=n+1}^{s}\mu_i\Big), \qquad n \le s. \tag{2.14}$$

Analogously to the derivation of equation (3.1) in [BB] from (2.13) in case of $n > s$ for the stationary distribution of $N(t)$ we obtain

$$\begin{aligned} p(n) &= \int\limits_{\mathbb{R}_+^{2\ell}} p(n; x_1, \ldots, x_\ell; u_1, \ldots, u_\ell) \mathrm{d}x_1 \ldots \mathrm{d}x_\ell \mathrm{d}u_1 \ldots \mathrm{d}u_\ell \\ &= g\Big(\prod_{i=0}^{n-1} \lambda_i\Big) \frac{1}{(n-s)!} \int\limits_0^\infty F(\xi)^{n-s} e^{-\xi} \mathrm{d}\xi, \qquad n > s, \end{aligned} \tag{2.15}$$

where

$$F(\xi) := \int\limits_0^{\xi/\mu_*} (1 - C(\eta))\mathrm{d}\eta, \qquad \xi \in \mathbb{R}_+. \tag{2.16}$$

Defining

$$q(n) := \begin{cases} \Big(\prod\limits_{i=0}^{n-1} \lambda_i\Big)\Big(\prod\limits_{i=n+1}^{s} \mu_i\Big), & n = 0, 1, \ldots, s, \\ \Big(\prod\limits_{i=0}^{n-1} \lambda_i\Big) \frac{1}{(n-s)!} \int\limits_0^\infty F(\xi)^{n-s} e^{-\xi} \mathrm{d}\xi, & n = s+1, s+2, \ldots \end{cases} \tag{2.17}$$

we have

$$p(n) = (\mathbb{1}\{n < s-a\}p_0 + \mathbb{1}\{n \geq s-a\})\, g\, q(n), \qquad n = 0, 1, \ldots \tag{2.18}$$

Let α_n be the intensity of customers leaving the system due to impatience and α_n' the intensity of customers transiting from Q into Q' conditioned upon $\ell := n-s > 0$ customers are in Q. In case of $p(n) > 0$ these intensities are given by

$$\begin{aligned} \alpha_n = \frac{1}{p(n)} \sum_{i=1}^{\ell} \int\limits_{\mathbb{R}_+^{2\ell-1}} & \mathbb{1}\{u_i < \tau\} p(n; x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_\ell; u_1, \ldots, u_\ell) \\ & \mathrm{d}x_1 \ldots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \ldots \mathrm{d}x_\ell \mathrm{d}u_1 \ldots \mathrm{d}u_\ell, \end{aligned} \tag{2.19}$$

$$\begin{aligned} \alpha_n' = \frac{1}{p(n)} \sum_{i=1}^{\ell} \int\limits_{\mathbb{R}_+^{2\ell-1}} & \mathbb{1}\{u_i \geq \tau\} p(n; x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_\ell; u_1, \ldots, u_\ell) \\ & \mathrm{d}x_1 \ldots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \ldots \mathrm{d}x_\ell \mathrm{d}u_1 \ldots \mathrm{d}u_\ell. \end{aligned} \tag{2.20}$$

Analogously to the derivation of equation (3.10) in [BB] by taking into account (2.13), (2.17), (2.18) after some algebra for $n > s$ we obtain

$$\alpha_n = \Big(\ell \int\limits_0^\infty F(\xi)^{\ell-1} C(\min(\xi/\mu_*, \tau)-) e^{-\xi} \mathrm{d}\xi\Big) \Big(\int\limits_0^\infty F(\xi)^{\ell} e^{-\xi} \mathrm{d}\xi\Big)^{-1}, \tag{2.21}$$

$$\alpha'_n \;=\; \Big(\ell \int\limits_{\mu_*\tau}^{\infty} F(\xi)^{\ell-1}(C(\xi/\mu_*) - C(\tau-))e^{-\xi}\mathrm{d}\xi\Big)\Big(\int\limits_{0}^{\infty} F(\xi)^{\ell}e^{-\xi}\mathrm{d}\xi\Big)^{-1}, \tag{2.22}$$

where in case of $p(n) = 0$ we define α_n and α'_n by these equations.

The conservation principle applied to Q' yields that the intensity of all arrivals at Q' (external arrivals and transitions from Q into Q') equals the intensity of customers passing from Q' into service. Since $(1-p_0)p(s-a)$ is the probability that precisely a servers are idle, Q is empty and Q' is not empty we have that $\mu_{s-a}(1-p_0)p(s-a)$ is just the intensity of starting service of a customer from Q'. Hence it holds

$$\sum_{n=s-a}^{\infty} \lambda'_n p(n) + \sum_{n=s+1}^{\infty} \alpha'_n p(n) = \mu_{s-a}(1-p_0)p(s-a). \tag{2.23}$$

From (2.23) and (2.18) for the unknown conditional probability p_0 we obtain the explicit expression

$$p_0 = \frac{\mu_{s-a}q(s-a) - \sum\limits_{n=s-a}^{\infty} \lambda'_n q(n) - \sum\limits_{n=s+1}^{\infty} \alpha'_n q(n)}{\mu_{s-a}q(s-a)}. \tag{2.24}$$

Since the $p(n)$, $n = 0, 1, \ldots$ must sum up to one, in view of (2.18) for the normalizing factor g it follows

$$g = \left(p_0 \sum_{n=0}^{s-a-1} q(n) + \sum_{n=s-a}^{\infty} q(n)\right)^{-1}, \tag{2.25}$$

yielding the stability condition for Q:

$$\sum_{n=s+1}^{\infty} q(n) < \infty. \tag{2.26}$$

Since p_0 must be positive, from (2.24) it follows the stability condition for Q':

$$\sum_{n=s-a}^{\infty} \lambda'_n q(n) + \sum_{n=s+1}^{\infty} \alpha'_n q(n) < \mu_{s-a}q(s-a). \tag{2.27}$$

The case of a general distribution $C(u)$ of the maximal waiting times is obtained by considering $C(u)$ as the limit in distribution of a sequence of non-defective distributions $C_\nu(u)$ with continuous density. In particular the formulae (2.16) – (2.18) and (2.21) – (2.25) remain valid, as well as the considerations concerning the stability conditions. Summarizing the preceding results we obtain the following statement.

Theorem 2.3. *The two-queue s-server system of Figure 1.1 with a general distribution $C(u)$ of the maximal waiting times is stable, i.e., there exists a unique stationary state process of the system, iff* (2.26) *and* (2.27) *are satisfied, where $q(n)$ and α'_n are given by* (2.17) *and* (2.22), *respectively.*

If the stability conditions (2.26) *and* (2.27) *are fulfilled, then the stationary probabilities $p(n)$ that the system is in state $N(t) = n$ are given by* (2.18)*, where $q(n)$, p_0 and g are given by* (2.17)*,* (2.24) *and* (2.25)*, respectively.*

Remark 2.4. *If only Q is stable, i.e., if* (2.26) *is fulfilled but not* (2.27)*, then the formulae* (2.16) – (2.18)*,* (2.21)*,* (2.22) *and* (2.25) *remain valid, where we have to define $p_0 := 0$. We obtain the analysis of a modified model analyzed in [BB], where Q' is replaced by an infinite reservoir of customers.*

Analogously to [BB] formulae can be derived for various performance measures related to Q as impatience probabilities, waiting time distributions etc. However, performance characteristics related to Q' like the mean sojourn time until service in Q' seem to be not available by this method.

The following monotonicity results for the intensities α_n and α'_n, respectively, play a crucial rule in the stochastic interpretation of a system approximation given in the next section.

Lemma 2.5. *The intensities α_n, $n = s+1, s+2, \ldots$ and α'_n, $n = s+1, s+2, \ldots$ increase monotonically with respect to n.*

Proof. 1. From (2.21) for $n = s+1, s+2, \ldots$ we obtain

$$\frac{\alpha_{n+1}}{\alpha_n} = \frac{(\ell+1)\int\limits_0^\infty F(\mu_*\xi)^\ell e^{-\mu_*\xi}\mathrm{d}\xi \int\limits_0^\infty F(\mu_*\xi)^\ell C(\min(\xi,\tau)-)e^{-\mu_*\xi}\mathrm{d}\xi}{\ell\int\limits_0^\infty F(\mu_*\xi)^{\ell+1} e^{-\mu_*\xi}\mathrm{d}\xi \int\limits_0^\infty F(\mu_*\xi)^{\ell-1} C(\min(\xi,\tau)-)e^{-\mu_*\xi}\mathrm{d}\xi}.$$

In view of (2.16) integration by parts yields

$$\frac{\alpha_{n+1}}{\alpha_n} = \frac{\int\limits_0^\infty F(\mu_*\xi)^{\ell-1}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi \int\limits_0^\infty F(\mu_*\xi)^\ell C(\min(\xi,\tau)-)e^{-\mu_*\xi}\mathrm{d}\xi}{\int\limits_0^\infty F(\mu_*\xi)^{\ell}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi \int\limits_0^\infty F(\mu_*\xi)^{\ell-1} C(\min(\xi,\tau)-)e^{-\mu_*\xi}\mathrm{d}\xi}. \tag{2.28}$$

Since α_n is monotonically increasing iff $\alpha_{n+1}/\alpha_n \geq 1$ we conclude that the assertion is equivalent to the non-negativity of the difference Δ of the numerator and denominator of the right-hand side of (2.28). Using Fubini's Theorem we find

$$\begin{aligned}\Delta \;:=\; & \int\limits_0^\infty F(\mu_*\xi)^{\ell-1}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi \int\limits_0^\infty F(\mu_*\xi)^\ell C(\min(\xi,\tau)-)e^{-\mu_*\xi}\mathrm{d}\xi \\ & - \int\limits_0^\infty F(\mu_*\xi)^{\ell}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi \int\limits_0^\infty F(\mu_*\xi)^{\ell-1} C(\min(\xi,\tau)-)e^{-\mu_*\xi}\mathrm{d}\xi\end{aligned}$$

$$= \frac{1}{2}\int_0^\infty\int_0^\infty F(\mu_*\xi)^{\ell-1}F(\mu_*\eta)^{\ell-1}e^{-\mu_*(\xi+\eta)}(F(\mu_*\xi)-F(\mu_*\eta)) \Big(C(\min(\xi,\tau)-)(1-C(\eta))-C(\min(\eta,\tau)-)(1-C(\xi))\Big)\mathrm{d}\xi\mathrm{d}\eta.$$

Since $F(\mu_*\xi)$ and $C(\xi)$ increase monotonically it follows

$$(F(\mu_*\xi)-F(\mu_*\eta))\Big(C(\min(\xi,\tau)-)(1-C(\eta))-C(\min(\eta,\tau)-)(1-C(\xi))\Big)\geq 0, \quad \xi,\eta\in I\!R_+.$$

This and $F(\mu_*\xi)\geq 0$, $\xi\in I\!R_+$ imply that the integrand is non-negative over $I\!R_+^2$. Thus $\Delta\geq 0$.

2. The monotonicity of the sequence α'_n, $n=s+1,s+2,\dots$ will be proved analogously. From (2.22) for $n=s+1,s+2,\dots$ we obtain

$$\frac{\alpha'_{n+1}}{\alpha'_n}=\frac{(\ell+1)\int_0^\infty F(\mu_*\xi)^\ell e^{-\mu_*\xi}\mathrm{d}\xi\int_\tau^\infty F(\mu_*\xi)^\ell(C(\xi)-C(\tau-))e^{-\mu_*\xi}\mathrm{d}\xi}{\ell\int_0^\infty F(\mu_*\xi)^{\ell+1}e^{-\mu_*\xi}\mathrm{d}\xi\int_\tau^\infty F(\mu_*\xi)^{\ell-1}(C(\xi)-C(\tau-))e^{-\mu_*\xi}\mathrm{d}\xi}.$$

Integration by parts yields

$$\frac{\alpha'_{n+1}}{\alpha'_n}=\frac{\int_0^\infty F(\mu_*\xi)^{\ell-1}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi\int_\tau^\infty F(\mu_*\xi)^\ell(C(\xi)-C(\tau-))e^{-\mu_*\xi}\mathrm{d}\xi}{\int_0^\infty F(\mu_*\xi)^{\ell}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi\int_\tau^\infty F(\mu_*\xi)^{\ell-1}(C(\xi)-C(\tau-))e^{-\mu_*\xi}\mathrm{d}\xi}. \tag{2.29}$$

Analogously to the first part the assertion is equivalent to the non-negativity of the difference Δ' of the numerator and denominator of the right-hand side of (2.29). Using Fubini's Theorem we find

$$\begin{aligned}\Delta' &:= \int_0^\infty F(\mu_*\xi)^{\ell-1}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi\int_0^\infty F(\mu_*\xi)^\ell 1\!\mathrm{I}\{\xi\geq\tau\}(C(\xi)-C(\tau-))e^{-\mu_*\xi}\mathrm{d}\xi\\ &\quad -\int_0^\infty F(\mu_*\xi)^{\ell}(1-C(\xi))e^{-\mu_*\xi}\mathrm{d}\xi\int_0^\infty F(\mu_*\xi)^{\ell-1} 1\!\mathrm{I}\{\xi\geq\tau\}(C(\xi)-C(\tau-))e^{-\mu_*\xi}\mathrm{d}\xi\\ &= \frac{1}{2}\int_0^\infty\int_0^\infty F(\mu_*\xi)^{\ell-1}F(\mu_*\eta)^{\ell-1}e^{-\mu_*(\xi+\eta)}(F(\mu_*\xi)-F(\mu_*\eta))\\ &\quad \Big(1\!\mathrm{I}\{\xi\geq\tau\}(C(\xi)-C(\tau-))(1-C(\eta)) - 1\!\mathrm{I}\{\eta\geq\tau\}(C(\eta)-C(\tau-))(1-C(\xi))\Big)\mathrm{d}\xi\mathrm{d}\eta.\end{aligned}$$

Since $F(\mu_*\xi)$ and $C(\xi)$ are non-negative and increase monotonically and since $C(\xi)\leq 1$ it follows that the integrand is non-negative over $I\!R^2$. Hence it holds $\Delta'\geq 0$, finishing the proof.

3 A system approximation by fitting impatience rates

Since for the stationary distribution $P^*(n,n',x_1,\ldots,x_\ell,u_1,\ldots,u_\ell)$ and in particular for the occupancy distribution for Q' no explicit representation seems to be available, approximations are of interest. We construct an approximate system by replacing the impatience mechanism in Q by waiting place dependent impatience rates. Using Lemma 2.5 these rates can be fitted appropriately. For the factorial moments of the occupancy distribution for Q' in the approximate system a recursive algorithm is given.

Consider the two-queue s-server system of Figure 1.1, but with the following modification of the impatience mechanism in Q: Let the waiting places for Q be numbered by $i = 1,2,\ldots$ A customer waiting on place i in Q is impatient: He leaves Q with rate $\beta_i+\beta'_i$, where he leaves the system with probability $\beta_i/(\beta_i+\beta'_i)$ and transits to the end of Q' with probability $\beta'_i/(\beta_i+\beta'_i)$, cf. Figure 3.1. The customers behind him move up in Q according to the FCFS discipline. Conditioned upon $n-s > 0$ customers are in Q, the cumulative rates $\tilde{\alpha}_n$ of leaving the system due to impatience and $\tilde{\alpha}'_n$ of transiting from Q into Q' are given by

$$\tilde{\alpha}_n = \sum_{i=1}^{n-s} \beta_i, \qquad \tilde{\alpha}'_n = \sum_{i=1}^{n-s} \beta'_i, \qquad n = s+1, s+2, \ldots \tag{3.1}$$

From Lemma 2.5 we know that the corresponding intensities α_n and α'_n in the original model, cf. (2.21), (2.22), increase monotonically with respect to n. Thus the fitting

$$\tilde{\alpha}_n := \alpha_n, \qquad \tilde{\alpha}'_n := \alpha'_n, \qquad n = s+1, s+2, \ldots \tag{3.2}$$

provides uniquely determined waiting place dependent impatience rates β_i and β'_i, $i = 1,2,\ldots$

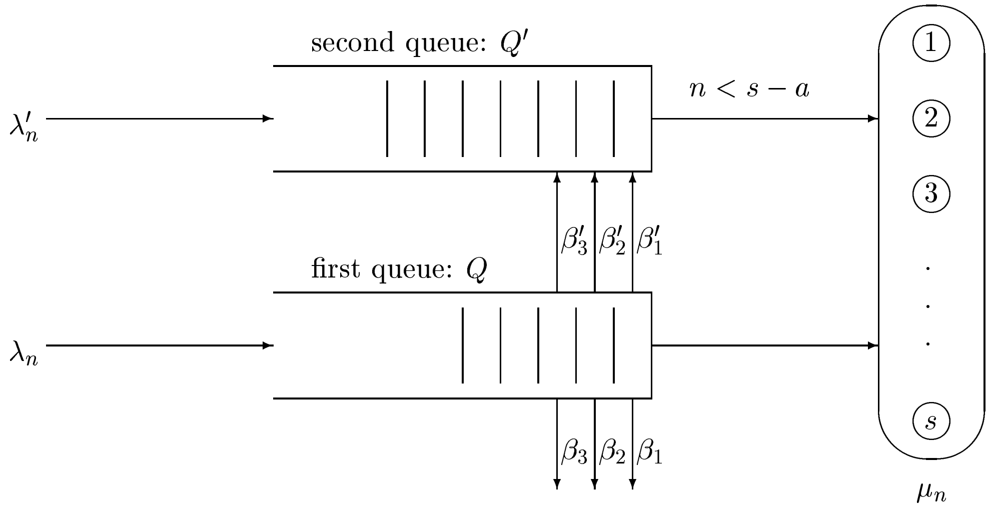

Figure 3.1. Approximate system: The two-queue s-server system with waiting place dependent impatience rates β_i and β'_i.

Denote by $\tilde{p}(n,n')$ the stationary probability that precisely n customers are in service or in Q

and n' customers are waiting in Q'. Then the balance equations read

$$\begin{aligned}
&(\lambda_n + \lambda'_n + \alpha_n + \alpha'_n + \mu_n)\tilde{p}(n,n') \\
&\quad = \mathbb{1}\{n > 0\}\lambda_{n-1}\tilde{p}(n-1,n') + \mathbb{1}\{n' > 0\}\lambda'_n\tilde{p}(n,n'-1) \\
&\qquad + (\mathbb{1}\{n+1 \neq s-a\} + \mathbb{1}\{n+1 = s-a\}\mathbb{1}\{n' = 0\})\mu_{n+1}\tilde{p}(n+1,n') \\
&\qquad + \mathbb{1}\{n = s-a\}\mu_n\tilde{p}(n,n'+1) \\
&\qquad + \alpha_{n+1}\tilde{p}(n+1,n') + \mathbb{1}\{n' > 0\}\alpha'_{n+1}\tilde{p}(n+1,n'-1), \qquad (n,n') \in \mathbb{Z}_+^2, \qquad (3.3)
\end{aligned}$$

where $\alpha_n := \alpha'_n := 0$ for $n = 0, 1, \ldots, s$. The normalizing condition reads

$$\sum_{(n,n') \in \mathbb{Z}_+^2} \tilde{p}(n,n') = 1. \qquad (3.4)$$

The two-dimensional system of equations (3.3), (3.4) can be solved numerically, in principle. By means of the $\tilde{p}(n,n')$ approximations for relevant performance measures can be computed. But instead of dealing with $\tilde{p}(n,n')$ being of some numerical complexity we will rather deal with the factorial moments of the number of customers in Q'. Define by

$$f(z,n) := \sum_{n'=0}^{\infty} \tilde{p}(n,n')z^{n'}, \quad n = 0, 1, \ldots \qquad (3.5)$$

the partial generating functions of the stationary distribution $\tilde{p}(n,n')$. For fixed $z \in \mathbb{C}$ with $|z| \leq 1$ by multiplying (3.3) by $z^{n'}$ and summing over $\{n, n+1, \ldots\} \times \{0, 1, \ldots\}$ we obtain

$$\begin{aligned}
&(\alpha_n + \alpha'_n + \mu_n)f(z,n) - \mathbb{1}\{n > 0\}\lambda_{n-1}f(z,n-1) \\
&\qquad - \mathbb{1}\{n = s-a\}\mu_{s-a}(f(z,s-a) - f(0,s-a)) \\
&\quad = \mathbb{1}\{n \leq s-a\}\mu_{s-a}\frac{1-z}{z}(f(z,s-a) - f(0,s-a)) \\
&\qquad - (1-z)\sum_{i=n}^{\infty}\lambda'_i f(z,i) - (1-z)\sum_{i=n+1}^{\infty}\alpha'_i f(z,i), \quad n = 0, 1, \ldots \qquad (3.6)
\end{aligned}$$

For $n = 0$ equation (3.6) simplifies to

$$\mu_{s-a}(f(z,s-a) - f(0,s-a)) = z\sum_{i=s-a}^{\infty}\lambda'_i f(z,i) + z\sum_{i=s+1}^{\infty}\alpha'_i f(z,i). \qquad (3.7)$$

For $z = 1$ from (3.6), (3.7) it follows

$$\begin{aligned}
&(\alpha_n + \alpha'_n + \mu_n)f(1,n) - \mathbb{1}\{n > 0\}\lambda_{n-1}f(1,n-1) \\
&\quad = \mathbb{1}\{n = s-a\}\mu_{s-a}(f(1,s-a) - f(0,s-a)), \quad n = 0, 1, \ldots, \qquad (3.8)
\end{aligned}$$

$$\mu_{s-a}(f(1,s-a) - f(0,s-a)) = \sum_{i=s-a}^{\infty} \lambda_i' f(1,i) + \sum_{i=s+1}^{\infty} \alpha_i' f(1,i). \tag{3.9}$$

Using the conservation principle one can show that the balance equations (3.8) and (3.9) also hold in the exact model, cf. Figure 1.1, where one has to replace the probabilities $f(1,n)$ that precisely n customers are in service or in Q by $p(n)$ and $f(0,s-a)$ $(= \tilde{p}(s-a,0))$ by $p^*(s-a,0)$, cf. (2.23). These facts and $p^*(s-a,0) = p_0 p(s-a)$ yield

$$f(1,n) = p(n), \quad n = 0,1,\ldots, \tag{3.10}$$

$$f(0,s-a) = p_0 p(s-a). \tag{3.11}$$

From (3.10) and (3.11) we see that the fitting of the impatience rates, cf. (3.2), implies a fitting of the probabilities that precisely n customers are in service or in Q, meeting the aim of a system approximation. Moreover, for the approximate system again we have the stability conditions (2.26) and (2.27).

Since $\tilde{p}(n,n') = 0$ for $0 \le n < s-a$ and $n' > 0$ from (3.5), (3.10) we conclude

$$f(z,n) \equiv p(n), \quad n = 0,1,\ldots,s-a-1. \tag{3.12}$$

Denoting by $f_j(n)$ the j-th derivative of $f(z,n)$ at $z = 1$, the j-th factorial moment f_j of the number of customers in Q' is given by

$$f_j = \sum_{n=s-a}^{\infty} f_j(n), \quad j = 1,2,\ldots \tag{3.13}$$

By taking the j-fold derivate of (3.6) at $z = 1$ for $j = 1,2,\ldots$ and $n = s-a+1,\ldots$ we obtain

$$f_j(n) = \frac{\lambda_{n-1} f_j(n-1) + \sum_{i=n}^{\infty} \lambda_i' j f_{j-1}(i) + \sum_{i=n+1}^{\infty} \alpha_i' j f_{j-1}(i)}{\alpha_n + \alpha_n' + \mu_n}\,. \tag{3.14}$$

For $j = 1,2,\ldots$ let us define

$$g_j(s-a) := 0, \tag{3.15}$$

$$g_j(n) := \frac{\lambda_{n-1} g_j(n-1) + \sum_{i=n}^{\infty} \lambda_i' j f_{j-1}(i) + \sum_{i=n+1}^{\infty} \alpha_i' j f_{j-1}(i)}{\alpha_n + \alpha_n' + \mu_n}, \quad n = s-a+1,\ldots \tag{3.16}$$

Then from (3.14) and (3.16) for $j = 1,2,\ldots$ and $n = s-a+1,\ldots$ it follows

$$f_j(n) - g_j(n) = \frac{\lambda_{n-1}}{\alpha_n + \alpha_n' + \mu_n} \left(f_j(n-1) - g_j(n-1)\right). \tag{3.17}$$

A look at (3.8), (3.10) and (3.17) shows that for $n = s{-}a, \dots$ the quantities $f_j(n){-}g_j(n)$ and $p(n)$ coincide up to a factor being independent on n. Thus, in view of $g_j(s{-}a) = 0$, for $n = s{-}a, \dots$ we find

$$f_j(n) = g_j(n) + \frac{p(n)}{p(s{-}a)} f_j(s{-}a). \tag{3.18}$$

For $j = 1, 2, \dots$ the j-fold differentiation of (3.7) at $z = 1$ yields

$$\mu_{s-a} f_j(s{-}a) = \sum_{i=s-a}^{\infty} \lambda'_i (j f_{j-1}(i) + f_j(i)) + \sum_{i=s+1}^{\infty} \alpha'_i (j f_{j-1}(i) + f_j(i)). \tag{3.19}$$

From (3.19), (3.18) in view of (2.23) it follows

$$\mu_{s-a} p_0 f_j(s{-}a) = \sum_{i=s-a}^{\infty} \lambda'_i (j f_{j-1}(i) + g_j(i)) + \sum_{i=s+1}^{\infty} \alpha'_i (j f_{j-1}(i) + g_j(i)). \tag{3.20}$$

Summarizing, for the factorial moments of the occupancy distribution for Q' from (3.10), (3.15), (3.16), (3.20), (3.18) and (3.13) we obtain the following recursion.

Algorithm 3.1. *Let*

$$f_0(n) := p(n), \qquad n = s{-}a, s{-}a{+}1, \dots \tag{3.21}$$

Then for $j = 1, 2, \dots$ the factorial moments f_j of the occupancy distribution for Q' in the two-queue s-server system of Figure 3.1 are given by the recursion

$$g_j(s{-}a) := 0, \tag{3.22}$$

$$g_j(n) := \frac{1}{\alpha_n + \alpha'_n + \mu_n} \Big(\lambda_{n-1} g_j(n{-}1) + \sum_{i=n}^{\infty} \lambda'_i j f_{j-1}(i) + \sum_{i=n+1}^{\infty} \alpha'_i j f_{j-1}(i) \Big),$$
$$n = s{-}a{+}1, s{-}a{+}2, \dots, \tag{3.23}$$

$$f_j(s{-}a) := \frac{1}{\mu_{s-a} p_0} \Big(\sum_{i=s-a}^{\infty} \lambda'_i (j f_{j-1}(i) + g_j(i)) + \sum_{i=s+1}^{\infty} \alpha'_i (j f_{j-1}(i) + g_j(i)) \Big), \tag{3.24}$$

$$f_j(n) := g_j(n) + \frac{p(n)}{p(s{-}a)} f_j(s{-}a), \quad n = s{-}a{+}1, s{-}a{+}2, \dots, \tag{3.25}$$

$$f_j = \sum_{n=s-a}^{\infty} f_j(n). \tag{3.26}$$

4 Application: Performance analysis of an inbound call center with an integrated voice-mail-server

Call centers are installed for several different services and businesses, e.g. for catalog orders on a service 800 base, hotline calls related to specific products, travel agencies, telebanking and many more. The automatic call distribution (ACD) software distributes – in accordance to flexible rules – the calls arriving at a call center to agents who will provide a service or product, cf. [P], [ST], [DPW], [G], [HHP]. Customers may abandon due to impatience if they have to wait too long for service. By integrating a voice-mail-server this effect can be smoothed: after some waiting time the customer will be informed or will get the offer that he will be recalled later by the system when enough idle agents are available.

In this section we apply the results of the last two sections to the performance analysis of a call center with an integrated voice-mail-server. The model is as follows: Consider a call center consisting of a group of s agents, k waiting places (i.e., $s+k$ lines) and an integrated voice-mail-server (VMS) of infinite capacity, cf. Figure 4.1.

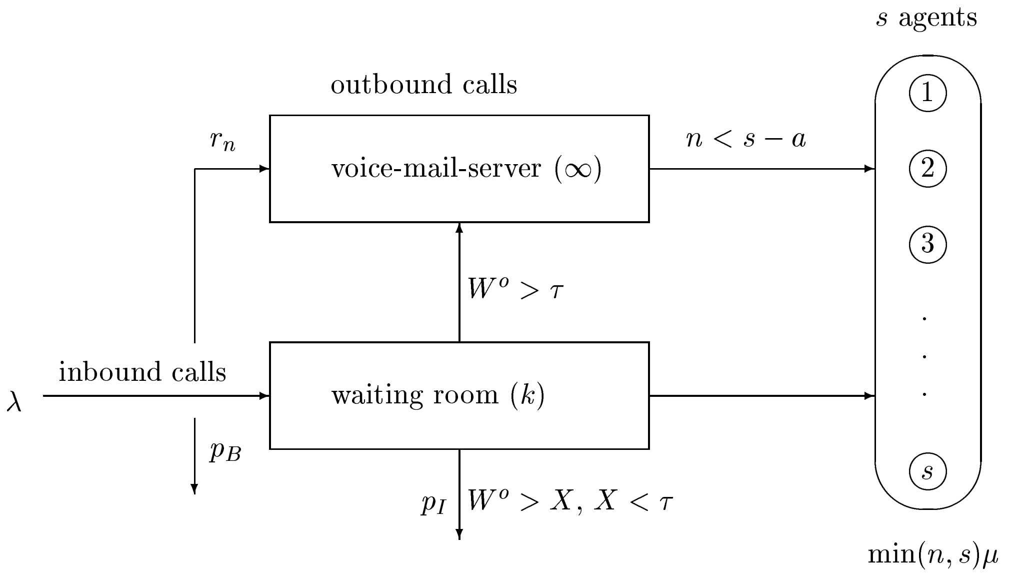

Figure 4.1. Call center with impatient inbound calls and overflow into a voice-mail-server (VMS) occurring directly at the arrival (state dependent on the number n of calls being in service or in the waiting room) or after a deterministic maximal waiting time τ.

Let n denotes the number of calls being in service or in the waiting room. At the system there arrive inbound calls from outside according to a Poisson process with intensity λ. If at the arrival of a call we have $n < s$, then there is an idle agent, and the service begins immediately. If at the arrival of a call we have $s \leq n < s+k$, then all agents are busy, and with probability $r_n \in [0,1)$ the arriving call goes immediately into the VMS, otherwise it begins to wait for service in the waiting room. If at the arrival of a call we have $n = s+k$, then there is no free line, and the call gets lost (blocking). The calls waiting in the waiting room are served in a FCFS manner. But they abandon after a random time X or will be transferred into the voice-mail-server after a constant time τ, the technical maximal waiting time, according to the following mechanism. Each call arriving at the system has a random individual maximal waiting time X. If $X < \tau$ and the waiting time exceeds X, then the call gets lost due to individual impatience. If $X \geq \tau$ and the waiting time exceeds τ, then the call is transferred into the VMS. The calls waiting in the VMS are also served in a FCFS manner, but they are not impatient. If more than a agents, where $a \in \{0, 1, \ldots, s-1\}$ is the outbound parameter, are idle and no calls are waiting in the waiting room then one of the idle agents serves a call from the VMS provided there is anyone. This implies that the service of a call from the VMS will be started at those moments when the service of any call is just finishing, a+1 agents become idle, no calls are waiting in the waiting room and there are calls waiting in the VMS. The service times of all calls are assumed to be i.i.d. and exponentially distributed with parameter μ. Also, the individual maximal waiting times are assumed to be i.i.d. with distribution function $P(X \leq x)$. Further, the arrival stream, the service times and the individual maximal waiting times are assumed to be mutually independent.

Remark 4.1. *The possibility of an immediate transition into the VMS at a call arrival with state dependent probability r_n corresponds to the situation that the customer has information about the queue length and hence decides whether to wait or to be recalled later. The technical maximal waiting time τ means that after time τ the call will be cut short by the system and the customer will be recalled later. By an appropriate change of X one can also model the more realistic situation that the customer will be recalled with a given probability later (modelling the decision of the customer and/or the ability of the system).*

Remark 4.2. *In view of the FCFS queuing discipline for the waiting room and since the technical maximal waiting times are deterministic, the call on the first waiting place is the next potential call for service and for being transferred into the VMS.*

Defining

$$\lambda_n := 1\!\!1\{n < s\}\lambda + 1\!\!1\{s \leq n < s+k\}\lambda(1-r_n), \quad n \geq 0, \tag{4.1}$$

$$\lambda'_n := 1\!\!1\{s \leq n < s+k\}\lambda r_n, \quad n \geq 0, \tag{4.2}$$

$$\mu_n := \min(n, s)\mu, \quad n \geq 0, \qquad \mu_* := s\mu, \tag{4.3}$$

$$I \ := \min(X, \tau) \tag{4.4}$$

the call center model proves to be a special case of the general model of Section 1, cf. Figure 1.1.

4.1 Performance measures for the call center

For the model considered several performance measures are of interest, e.g. the blocking probability, the probability of leaving the system due to impatience, waiting times in the waiting room and in the VMS etc.

From (2.17), (2.18), (4.1) and (4.3) for the stationary probabilities $p(n)$ that precisely n calls are in service or in the waiting room we obtain

$$p(n) = (1\!\!1\{0 \leq n < s-a\}p_0 + 1\!\!1\{s-a \leq n \leq s+k\})\, g\, q(n), \quad n \geq 0, \tag{4.5}$$

where

$$q(n)=\begin{cases} s!\mu^s\frac{(\lambda/\mu)^n}{n!}, & 0\le n\le s,\\ \lambda^n\Big(\prod\limits_{i=s}^{n-1}(1-r_i)\Big)\frac{1}{(n-s)!}\int\limits_0^\infty F(\xi)^{n-s}e^{-\xi}\mathrm{d}\xi, & s<n\le s+k,\\ 0 & \text{elsewhere,}\end{cases} \tag{4.6}$$

$F(\xi)$ is given by (2.16) and where in view of (4.4) we have

$$C(u)=1\!\!\mathrm{I}\{u<\tau\}P(X\le u)+1\!\!\mathrm{I}\{u\ge\tau\},\quad u\ge 0. \tag{4.7}$$

From (2.21), (2.22), (4.6), (2.16) and (4.7) for $s<n\le s+k$ we find the intensities α_n of calls leaving the system due to impatience and α_n' of calls transiting from the waiting room into the VMS conditioned upon $n-s$ calls are waiting in the waiting room

$$\alpha_n = \frac{\lambda^n}{q(n)}\Big(\prod_{i=s}^{n-1}(1-r_i)\Big)\frac{1}{(n-s-1)!}\int\limits_0^\infty F(\xi)^{n-s-1}C(\min(\xi/\mu_*,\tau)-)e^{-\xi}\mathrm{d}\xi, \tag{4.8}$$

$$\alpha_n' = \frac{\lambda^n}{q(n)}\Big(\prod_{i=s}^{n-1}(1-r_i)\Big)\frac{e^{-\mu_*\tau}}{(n-s-1)!}F(\mu_*\tau)^{n-s-1}(1-C(\tau-)). \tag{4.9}$$

For the probability $p_0:=p^*(s-a,0)/p(s-a)$ that the VMS is empty on the condition that precisely a agents are idle and no calls are waiting in the waiting room from (2.24), (4.2), (4.3) and (4.6) it follows

$$p_0=\frac{(s-a)\mu q(s-a)-\lambda\sum\limits_{n=s}^{s+k-1}r_nq(n)-\sum\limits_{n=s+1}^{s+k}\alpha_n'q(n)}{(s-a)\mu q(s-a)}. \tag{4.10}$$

From (2.25), (4.6) for the normalizing factor we obtain

$$g=\Big(p_0\sum_{n=0}^{s-a-1}q(n)+\sum_{n=s-a}^{s+k}q(n)\Big)^{-1}. \tag{4.11}$$

For the call center the stability condition (2.26) for Q is always fulfilled, in view of (4.6). Thus, according to (2.27), (4.2), (4.6) and (4.3) the stability condition reads

$$\lambda\sum_{n=s}^{s+k-1}r_nq(n)+\sum_{n=s+1}^{s+k}\alpha_n'q(n)<(s-a)\mu q(s-a). \tag{4.12}$$

Remark 4.3. *If* (4.12) *is fulfilled then* (4.12) *may be multiplied by g, i.e., $q(n)$ may be replaced by $p(n)$ in* (4.12). *The resulting inequality has the following interpretation: In the steady state the intensity of all calls entering and hence leaving the VMS is smaller than the intensity of all calls being served from state $s-a$ on.*

Remark 4.4. *The model considered in [BB], Section 4, may be obtained from our model as the limiting case where we have equality in* (4.12)*, cf. Remark 2.4. However, a complete mathematical analysis for our model as given for the model in [BB] seems to be not available, cf. the comment after formula* (2.5)*.*

In the following we assume that (4.12) is fulfilled, i.e., that the call center of Figure 4.1 is stable. Now, formulae for several performance measures of interest will be derived. (Since the system dynamics of the call center considered here are different to the dynamics of the model in [BB] not all notation used here coincides with that in [BB].)

The following *call intensities* related to the system are of special interest:

Λ_A – intensity of all calls accepted by the system,
Λ_I – intensity of all calls accepted by the system and which will get lost due to individual impatience later,
Λ_W – intensity of all calls accepted by the system and which have to wait for service in the waiting room,
Λ_V – intensity of all calls accepted by the system and which go immediately at their arrival or will be transferred later into the VMS.

These intensities are given by

$$\Lambda_A = \lambda(1-p(s+k)), \qquad \Lambda_I = \sum_{n=s+1}^{s+k} \alpha_n p(n), \tag{4.13}$$

$$\Lambda_W = \lambda \sum_{n=s}^{s+k-1} (1-r_n)p(n), \qquad \Lambda_V = \lambda \sum_{n=s}^{s+k-1} r_n p(n) + \sum_{n=s+1}^{s+k} \alpha'_n p(n). \tag{4.14}$$

Corresponding *probabilities* are:

p_B – blocking probability that a call arriving from outside is not accepted,
p_I – impatience probability that a call accepted by the system will get lost due to individual impatience later,
p_W – probability that a call accepted by the system has to wait for service in the waiting room,
p_V – probability that a call accepted by the system goes immediately at its arrival or will be transferred later into the VMS.

Obviously, we obtain

$$p_B = p(s+k), \qquad p_I = \Lambda_I/\Lambda_A, \qquad p_W = \Lambda_W/\Lambda_A, \qquad p_V = \Lambda_V/\Lambda_A. \tag{4.15}$$

By the intensity conversation principle the intensity of all calls accepted by the system and which will not get lost due to individual impatience equals the intensity of all served calls. In view of (4.3) this yields the following alternative expression for the impatience probability p_I:

$$1 - p_I = \frac{1}{\Lambda_A} \sum_{n=0}^{s+k} \min(n,s)\mu p(n). \tag{4.16}$$

The *distribution of the waiting time in the waiting room* of a typical call accepted by the system on the condition that it has to wait in the waiting room is given by Theorem 3.2 in [BB], formula (3.11), where Λ has to be replaced by Λ_W since Λ_W is the intensity of calls accepted by the system which have to wait in the waiting room. According to (2.16), (4.1) therefore we find

$$1 - W_W(x) = \frac{\lambda p(s)}{\Lambda_W}(1 - C(x)) \sum_{j=0}^{k-1} \Big(\prod_{i=s}^{s+j} (1 - r_i) \Big) \frac{1}{j!} \int_{\mu_* x}^{\infty} (\lambda F(\xi))^j e^{-\xi} \mathrm{d}\xi, \quad x \in I\!R_+. \tag{4.17}$$

By Little's formula for the mean waiting time $E\,W_W$ in the waiting room of accepted calls on the condition that they have to wait in the waiting room we obtain

$$E\,W_W = \frac{1}{\Lambda_W} \sum_{n=s+1}^{s+k} (n-s)p(n). \tag{4.18}$$

In the special case of (4.1) – (4.4), the first moment f_1 of the occupancy distribution for Q' in the model of Figure 3.1 yields an approximation for the mean number of calls in the VMS. Therefore, by Little's formula

$$E\,\tilde{W}_V := f_1/\Lambda_V \tag{4.19}$$

provides an *approximation for the mean waiting time of the calls waiting in the VMS.* In the special case of (4.1) – (4.4) for the first moment f_1 of the number of customers in Q' in the model of Figure 3.1, the Algorithm 3.1 simplifies as follows:

$$g_1(s-a) := 0, \tag{4.20}$$

$$g_1(n) := \frac{p(n)}{\lambda_{n-1} p(n-1)} \Big(\lambda_{n-1} g_1(n-1) + \lambda \sum_{i=\max(n,s)}^{s+k-1} r_i p(i) + \sum_{i=\max(n,s)+1}^{s+k} \alpha_i' p(i) \Big),$$

$$n = s-a+1, s-a+2, \ldots, s+k, \tag{4.21}$$

$$f_1 = \sum_{i=s-a+1}^{s+k} g_1(i) + \frac{\sum_{i=s-a}^{s+k} p(i)}{(s-a)\mu p_0 p(s-a)} \Big(\lambda \sum_{i=s}^{s+k-1} r_i \big(p(i) + g_1(i)\big) + \sum_{i=s+1}^{s+k} \alpha_i' \big(p(i) + g_1(i)\big) \Big). \tag{4.22}$$

4.2 Numerical examples in case of exponential individual maximal waiting times

Consider the case of exponential individual maximal waiting times X with parameter α, i.e.

$$C(u) = 1 - 1\!\!\mathrm{I}\{u < \tau\} e^{-\alpha u}, \qquad u \geq 0. \tag{4.23}$$

Exploiting the special structure of (4.23), from the explicit formulae of Section 4.1 numerically stable algorithms have been derived for the performance measures p_B, p_I, p_W, p_V, EW_W and for the approximate mean waiting time $E\tilde{W}_V$ in the VMS. These algorithms and a simulation of the call center have been implemented. In Table 4.2, numerical and simulation results are given for a call center with the parameters $\lambda/\mu = 100$, $1/\mu = 300$, $1/\alpha = 180$, $\tau = 20$ and $r_n = 1 - 0,98^{n-s+1}$ for $s \leq n < s+k$. The parameters s, k, a vary accordingly.

s	k	a	p_B	p_I	p_W	p_V	EW_W	$E\tilde{W}_V$	EW_V^*	EW_V^{**}
100	4	2	0.04549	0.00951	0.25928	0.01782	6.61	93.47	98.83	96.30
100	4	4	0.04448	0.00929	0.25325	0.01740	6.61	175.14	180.69	179.40
100	4	6	0.04369	0.00912	0.24857	0.01708	6.61	299.12	296.97	305.14
100	8	2	0.01071	0.02236	0.45991	0.06035	8.75	404.96	429.95	419.52
100	8	4	0.00992	0.02069	0.42557	0.05584	8.75	793.21	828.90	815.16
100	8	6	0.00936	0.01952	0.40132	0.05266	8.75	1763.62	1974.91	1827.20
100	12	2	0.00044	0.02579	0.51414	0.07472	9.03	668.73	651.46	669.72
100	12	4	0.00040	0.02345	0.46743	0.06793	9.03	1438.47	1552.97	1447.97
100	12	6	0.00038	0.02185	0.43544	0.06328	9.03	4847.53	3473.95	4969.33
105	4	2	0.02620	0.00562	0.16308	0.01027	6.20	73.84	74.26	75.48
105	4	4	0.02558	0.00548	0.15913	0.01002	6.20	129.45	137.43	131.90
105	4	6	0.02512	0.00538	0.15618	0.00983	6.20	203.72	212.06	207.82
105	8	2	0.00543	0.01230	0.27096	0.03142	8.17	203.73	204.36	210.29
105	8	4	0.00505	0.01143	0.25182	0.02920	8.17	318.35	313.49	325.57
105	8	6	0.00479	0.01083	0.23867	0.02768	8.17	476.76	477.43	485.53
105	12	2	0.00021	0.01384	0.29595	0.03783	8.42	265.76	246.30	270.69
105	12	4	0.00020	0.01268	0.27114	0.03466	8.42	404.28	378.10	409.73
105	12	6	0.00018	0.01190	0.25450	0.03253	8.42	599.39	593.44	610.47
110	4	2	0.01318	0.00292	0.09015	0.00526	5.83	60.72	62.08	61.55
110	4	4	0.01285	0.00285	0.08791	0.00512	5.83	102.14	100.88	103.35
110	4	6	0.01262	0.00280	0.08630	0.00503	5.83	153.58	160.93	154.82
110	8	2	0.00240	0.00591	0.13976	0.01440	7.62	134.65	135.54	138.52
110	8	4	0.00224	0.00553	0.13063	0.01346	7.62	197.45	194.02	201.76
110	8	6	0.00214	0.00527	0.12456	0.01284	7.62	273.63	279.20	279.35
110	12	2	0.00009	0.00651	0.14951	0.01688	7.84	163.21	158.93	165.90
110	12	4	0.00008	0.00602	0.13820	0.01560	7.84	232.04	233.18	236.92
110	12	6	0.00008	0.00570	0.13080	0.01477	7.84	315.57	317.05	319.81

Table 4.2. Blocking probability p_B, impatience probability p_I, probabilities p_W of waiting in the waiting room and p_V of immediate or later transition into the VMS, the mean waiting time EW_W in the waiting room as well as the approximate mean waiting time $E\tilde{W}_V$ and the simulated mean waiting times EW_V^* and EW_V^{**} in the VMS for the inbound call center with integrated VMS for the case of s agents, k waiting places and outbound parameter a. EW_V^* and EW_V^{**} were obtained by simulating the system with 10^6 and 10^8 evaluated arrivals, respectively, starting from the empty system after 10^4 and 10^6 non-evaluated arrivals, respectively.

In the examples considered the call center works in the domain of critical loading. Comparing the mean waiting times $E\,W_V^*$ and $E\,W_V^{**}$ in the VMS obtained by simulating 10^6 and 10^8 arrivals at the system, respectively, we see that the waiting times W_V in the VMS of accepted calls on the condition that they will wait for service in the VMS vary very strongly. Not before simulating a relatively large number of arrivals at the system (e.g. 10^8 as in Table 4.2) the mean waiting times in the VMS obtained from simulation seem to be stable for the parameters chosen in Table 4.2. This causes from the fact that in critically loaded systems a strong fluctuation (large variance) has to be expected. Moreover, the sequence of waiting times is strongly dependent. Thus, long simulation runs are necessary for obtaining stable statistics. However, the values of $E\,W_V^{**}$ demonstrate that in case of exponential individual maximal waiting times the approximation $E\,\tilde{W}_V$ given by (4.19) works well. The relative error of the approximate mean waiting time $E\,\tilde{W}_V$ compared to the simulated mean waiting time $E\,W_V^{**}$ is less than 4% within Table 4.2. Also for many other examples – not reported here – we found that the approximation works well and hence can be used successfully. On the other hand, in the examples considered the mean waiting time $E\,W_V$ in the VMS does not give much information about quantiles of the distribution of the waiting times W_V in the VMS due to the high variability of the waiting times in the VMS. For the parameters of Table 4.2 it holds $E\,\tilde{W}_V < E\,W_V^{**}$. It seems that in case of exponential individual maximal waiting times the approximation $E\,\tilde{W}_V$ is always less than the exact value $E\,W_V$.

Numerical results also show the impact of the operational strategy (outbound parameter a) on the system performance. If a is chosen not too small, then among the parameters considered only the mean waiting time $E\,W_V$ in the VMS depends very strongly on a. This observation bases on the fact that in the special case of (4.1) – (4.4) only Q' in Figure 1.1 modelling the VMS may be unstable. Moreover, the numerical results illustrate the tradeoff between the blocking probability p_B and the impatience probability p_I if k varies.

The implemented algorithms can be used for studying the effects arising by integrating a VMS into a call center and also as basis for optimizing call centers with integrated VMS.

A Appendix. The special case $r_n \equiv 0$, $X \equiv \infty$ of the call center: A stochastic decomposition by a $M(n)/M(n)/s+GI$ system

Consider the call center of Section 4 in the special case of $r_n \equiv 0$ and $X \equiv \infty$, cf. Figure A.1, i.e., the model of Section 1 in case of $\lambda_n = 1\!\!1\{0 \le n < s+k\}\lambda$ for some positive integer k, $\lambda'_n = 0$, $\mu_n = \min(n, s)\mu$ for $n = 0, 1, 2, \ldots$ and $I \equiv \tau$. In this case the dynamics simplify as follows: An arriving inbound call from outside is accepted if $n < s+k$ calls are in service or in the waiting room, where in case of $s \le n < s+k$ it begins to wait for service in the waiting room. If the k waiting places are occupied then it gets lost (blocking). If the offered waiting time W^o of a call waiting in the waiting room exceeds the deterministic technical maximal waiting time τ, then the call will be transferred into the voice-mail-server (VMS) after waiting time τ. The calls in the VMS are served accordingly to the mechanism described in Section 4.

In the following we will give a stochastic decomposition yielding an alternative approach for determining the distribution of the number of calls being in service or in the waiting room. The decomposition bases on a particular $M(n)/M(n)/s+GI$ system, analyzed in [BB], and on a birth and death process. The idea is to consider the LCFS discipline instead of the FCFS discipline for the VMS. This modification does not influence the occupancy distribution and other characteristics being of interest, but it offers a stochastic analysis. It seems that such kinds of arguments have been used for the first time in [GK] for the analysis of the busy period distribution in the $M/GI/1/\infty$ system. Similar and more general stochastic decomposition techniques were developed since that time for various queueing system, e.g. for the $M/GI/1/$-processor-sharing system in [KY], [Y], for polling systems in [KLS]. Further sources are [B], [FC], [M], [BFL].

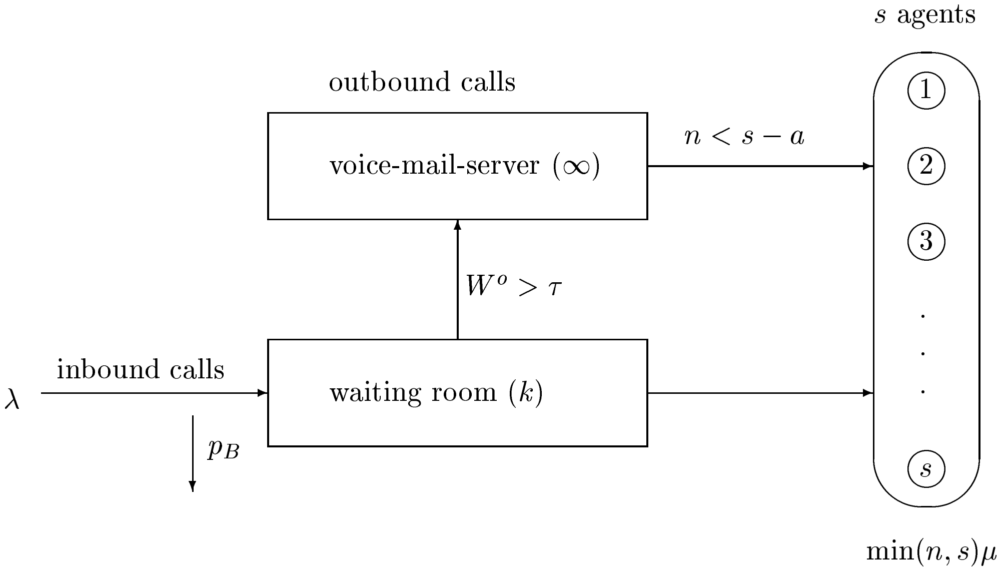

Figure A.1. Call center with calls being transferred into the voice-mail-server after a deterministic maximal waiting time τ.

A.1 A stochastic decomposition

Consider now the VMS with the LCFS queueing discipline, i.e., the last arrived call is the next for service. By the independence and distributional assumptions the process $(N(t), N'(t))$, $t \in I\!R$ of the vector of the number of calls being in service or in the waiting room and of the number of calls in the VMS has the same distribution as in case of the FCFS discipline for the VMS. The LCFS discipline results into a stochastic cycle-decomposition as follows. Assuming that the system is in steady state – the corresponding stability condition will be given later – then the system becomes empty infinitely often almost surely. Let $T^{(i)}_{s-a-1,s-a}$, $i \in Z\!\!\!Z$ be the time instants where $N(t)$ jumps from $s-a-1$ to $s-a$. Since at these time instants the VMS and the waiting room are empty, we conclude by the exponentiality of the service times that the $T^{(i)}_{s-a-1,s-a}$, $i \in Z\!\!\!Z$ form a renewal process and hence divide the time axis into cycles. The distribution of the duration Z of a typical cycle is given by

$$P(Z \le t) = P\Big(\inf\{t : t > 0,\, N(t) = N(t-)+1 = s-a\} \,\Big|\, N(0) = N(0-)+1 = s-a\Big).$$

Since $N(t)$ jumps downwards and upwards only by one, i.e., $N(t) - N(t-) \in \{-1, 0, 1\}$, $t \in I\!R$, it follows that after a jump from $s-a-1$ to $s-a$ there is a jump from $s-a$ to $s-a-1$ before the next jump from $s-a-1$ to $s-a$ may occur. This implies that every cycle consists of two intervals of duration D and G, respectively, cf. Figure A.2,

$$Z = D + G. \tag{A.1}$$

Here D is the duration from the beginning of a cycle until the first entrance of $N(t)$ into $s{-}a{-}1$, i.e., of a jump from $s-a$ to $s-a-1$, and G the subsequent duration until $N(t)$ jumps into $s-a$, i.e., until the end of the cycle where a jump from $s-a-1$ to $s-a$ occurs. Because of the exponential service times, the time instants where $N(t)$ jumps from $s{-}a{-}1$ to $s{-}a$ or from $s{-}a$ to $s-a-1$ form an alternating renewal process with alternating phases D and G, cf. e.g. [A], p. 130, or [T]. In the following we assume that a typical cycle starts at $t = 0$, i.e., we consider the Palm-distribution of $N(t)$ with respect to the embedded point process $T^{(i)}_{s-a-1,s-a}$, $i \in \mathbb{Z}$, cf. e.g. [FKAS], Section 1.5. Let $(N^0(t),\ t \in \mathbb{R})$ be a corresponding version of this process, cf. Figure A.2.

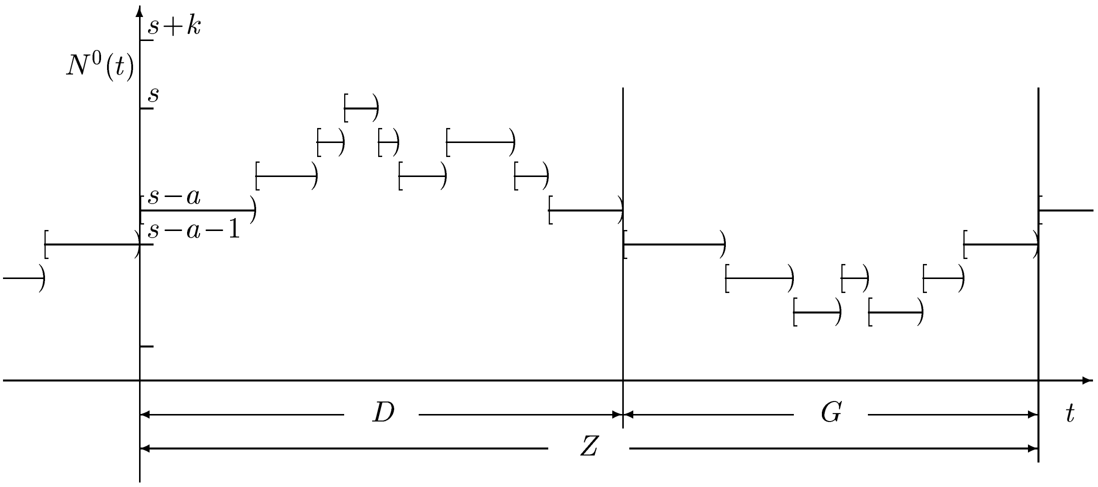

Figure A.2. Typical cycle of length Z, splitting in two phases of length D and G, respectively.

Now let C be the first time instant after $t = 0$ where $a{+}1$ agents become idle and no calls are in the waiting room, i.e. the first time instant where a call from the VMS could go into service. During the interval $(0, C]$ there transit M (≥ 0) calls into the VMS (from the waiting room due to exceeding the maximal waiting time τ). Denote these calls by $Call_1, \ldots, Call_M$. If $M = 0$ then $C = D$, and after this time it follows the second part of the cycle of duration G, after which a new cycle starts. If $M > 0$ then $C < D$, and $Call_M$ (being the last call arrived at the VMS) goes at the time instant C into service. In this case $N^0(C) = s-a$, and there is no jump of $N^0(t)$ at the time instant C. Consider now the case $M > 0$ in detail. Denote by D_M the time from the time instant C until $a{+}1$ agents become idle and no calls are in the waiting room anew and only the calls $Call_1, \ldots, Call_{M-1}$ are present in the VMS. Then in view of the LCFS discipline D_M is of the same probabilistic structure as D, i.e., $D =^{\mathcal{D}} D_{M-1}$, where $=^{\mathcal{D}}$ denotes equality in distribution. (During $(C, C{+}D_M]$ all calls have been served which arrived during that time at the VMS). At the time instant $C{+}D_M$ as next the call $Call_{M-1}$ goes in service. The time D_{M-1} until $a{+}1$ agents become idle and no calls are in the waiting room anew and only the calls $Call_1, \ldots, Call_{M-2}$ are present in the VMS has – by the same arguments as above – the same distribution as D, i.e., $D =^{\mathcal{D}} D_{M-1}$. Continuing this scheme, each of the calls $Call_{M-2}, \ldots, Call_1$ initiates with the beginning of its service a corresponding cycle of duration

$D_{M-2}, \ldots, D_1$, having the same distribution as D. Thus it follows

$$D =^{\mathcal{D}} C + \sum_{i=1}^{M} D_i. \tag{A.2}$$

In view of the system dynamics, the independence and exponentiality assumptions, the D_i can be chosen from a sequence of i.i.d. random variables $D_1, D_2, \ldots$ with $D_i =^{\mathcal{D}} D$, and M is independent on the sequence $(D_i)_{i=1}^{\infty}$. Applying Wald's identity from (A.2) we obtain

$$E\,D = E\,C + E\,M \cdot E\,D. \tag{A.3}$$

The process $N(t)$, $t \in I\!\!R$ is a regenerative process concerning the epochs $T^{(i)}_{s-a-1,s-a}$, $i \in Z\!\!\!Z$. Thus the stationary distribution $p(n) = P(N(t) = n)$, $n = 0, \ldots, s{+}k$ of the number of calls being in service or in the waiting room is given by the fraction of the sojourn time in the state n during a cycle, cf. [A], p. 126,

$$p(n) = \frac{E\,T_n^Z}{E\,Z} = \frac{E\,T_n^Z}{E\,D + E\,G}, \qquad n = 0, \ldots, s{+}k, \tag{A.4}$$

where T_n^Z is the sojourn time of $N^0(t)$ in state n during the cycle $(0, Z]$.

A.2 Stability condition and stationary distribution $p(n)$

In order to compute the stationary distribution $p(n)$ via the right-hand side of (A.4) a computation of $E\,T_n^Z$, $E\,C$, $E\,M$ and $E\,G$ is necessary, in view of (A.3). As mentioned in the introduction, these quantities can be computed using results for the $M(n)/M(n)/s{+}GI$ system given in [BB] and for birth and death processes.

Determination of $E\,G$.
The random time G corresponds to the first passage time into the state $s{-}a$ of a birth and death process $\overline{N}(t)$ with initial condition $\overline{N}(0) = s{-}a{-}1$ and birth and death rates $\lambda_n := \lambda$, $\mu_n := n\mu$, $n = 0, \ldots, s-a$, respectively, i.e., $G = \inf\{t : t > 0, \overline{N}(t) = s-a\}$, $\overline{N}(0) = s-a-1$. From the theory of birth and death processes we find

$$E\,G = \frac{(s{-}a{-}1)!}{\lambda(\lambda/\mu)^{s-a-1}} \sum_{j=0}^{s-a-1} \frac{(\lambda/\mu)^j}{j!}. \tag{A.5}$$

Determination of $E\,C$.
Let us consider the call center without VMS, i.e., accepted calls for which the offered waiting time W^o exceeds τ will get lost due to impatience later, cf. Figure A.3. This system is a special $M(n)/M(n)/s{+}GI$ system with parameters

$$\lambda_n := 1\!\!\mathrm{I}\{0 \le n < s{+}k\}\lambda, \quad \mu_n := \min(n, s)\mu, \quad n = 0, 1, \ldots, \qquad I \equiv \tau$$

and is analyzed in [BB].

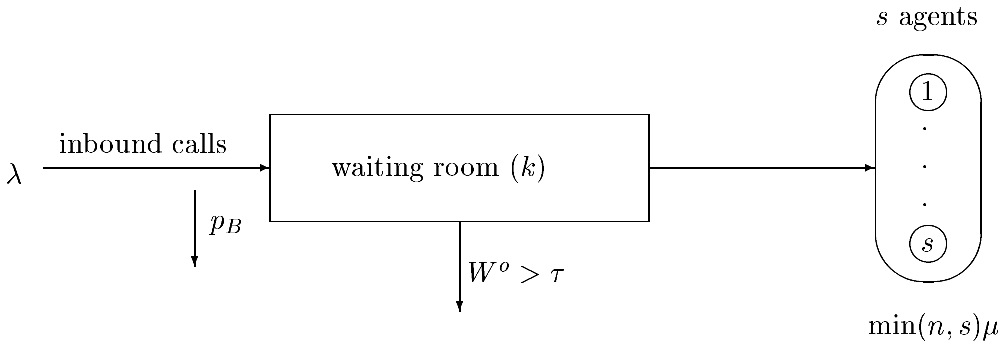

Figure A.3. The call center without VMS.

Let the corresponding quantities in this system be equipped with a tilde, e.g. $\tilde{N}(t)$, $\tilde{C}$, $\tilde{D}$, $\tilde{G}$, $\tilde{Z}$, $\tilde{T}_n^Z$ and $\tilde{p}(n)$. Since there is no VMS, it holds $\tilde{D} = \tilde{C}$. Furthermore, since in both models (Figure A.2 and Figure A.3) the dynamics of G and $\tilde{G}$ and of C and $\tilde{C}$ are the same we obtain

$$\tilde{G} = G, \qquad \tilde{C} = C. \tag{A.6}$$

Analogously to (A.4) it holds

$$\tilde{p}(n) = \frac{E\,\tilde{T}_n^Z}{E\,\tilde{C} + E\,\tilde{G}}, \qquad n = 0, \ldots, s{+}k, \tag{A.7}$$

where the $\tilde{p}(n)$ are explicitly given in [BB]. Since $\tilde{G}$ is just the sojourn time of $\tilde{N}(t)$ in the set $\{0, \ldots, s{-}a{-}1\}$ during the cycle from (A.7) we find

$$\frac{E\,\tilde{G}}{E\,\tilde{C} + E\,\tilde{G}} = \sum_{n=0}^{s-a-1} \tilde{p}(n),$$

and in view of (A.6) it follows

$$E\,C = \Big(1 - \sum_{n=0}^{s-a-1} \tilde{p}(n)\Big)\Big(\sum_{n=0}^{s-a-1} \tilde{p}(n)\Big)^{-1} E\,G. \tag{A.8}$$

Using formula (3.5) in [BB] from (A.5) for $E\,G$ we obtain the alternative expression

$$E\,G = \frac{1}{(s{-}a)\mu\tilde{p}(s{-}a)} \sum_{n=0}^{s-a-1} \tilde{p}(n). \tag{A.9}$$

Combining now (A.8), (A.9) for $E\,C$ we find the representation

$$E\,C = \frac{1}{(s{-}a)\mu\tilde{p}(s{-}a)} \Big(1 - \sum_{n=0}^{s-a-1} \tilde{p}(n)\Big). \tag{A.10}$$

Determination of EM.
For the model of Figure A.3 according to formula (3.8) in [BB] the intensity $\tilde{\lambda}^{(I)}$ of all accepted calls which will get lost due to impatience later is given by

$$\tilde{\lambda}^{(I)} = \lambda(1-\tilde{p}(s+k)) - \sum_{n=1}^{s+k} \min(n,s)\mu\tilde{p}(n). \tag{A.11}$$

In view of the dynamics, for the number $\tilde{M}$ of calls lost due to impatience during $(0,\tilde{C}]$ we have the identity $\tilde{M} = M$. Since during $(\tilde{C},\tilde{C}+\tilde{G}]$ no calls get lost due to impatience, taking into account (A.6), it follows

$$\tilde{\lambda}^{(I)} = \frac{E\tilde{M}}{E\tilde{C}+E\tilde{G}} = \frac{EM}{EC+EG}.$$

Combining this with (A.9), (A.10) we obtain

$$EM = \frac{\tilde{\lambda}^{(I)}}{(s-a)\mu\tilde{p}(s-a)}. \tag{A.12}$$

In view of (A.3), (A.10), (A.12) and (A.9) the quantities ED and EG are determined. For finding an explicit representation for the $p(n)$ from (A.4) the quantities ET_n^Z have to be determined yet.

Determination of ET_n^Z.
In case of $n \in \{0,1,\ldots,s-a-1\}$ from (A.7), (A.6) and using the fact that $\tilde{T}_n^Z = T_n^Z$ it follows

$$ET_n^Z = (EC+EG)\,\tilde{p}(n), \qquad n = 0,\ldots,s-a-1. \tag{A.13}$$

In case of $n \in \{s-a,\ldots,s+k\}$ from the stochastic decomposition, cf. (A.2) and (A.6), we obtain

$$ET_n^Z = E\tilde{T}_n^Z + EM \cdot ET_n^Z.$$

This and (A.7), (A.6) yield

$$ET_n^Z = \frac{E\tilde{T}_n^Z}{1-EM} = \frac{EC+EG}{1-EM}\,\tilde{p}(n), \qquad n = s-a,\ldots,s+k. \tag{A.14}$$

Thus the ET_n^Z, $n = 0,1,\ldots,s+k$ can be computed.

From (A.4), (A.3), (A.13) and (A.14) we obtain

$$p(n) = \frac{EC+EG}{EC+(1-EM)EG}\Big(1 - \mathbb{I}\{n < s-a\}EM\Big)\,\tilde{p}(n), \qquad n = 0,1,\ldots,s+k. \tag{A.15}$$

Finally, using (A.9), (A.10) and (A.12) from (A.15) it follows that in the model of Figure A.1 the stationary distribution of the number of calls being in service or in the waiting room is given by

$$p(n) = \frac{(s-a)\mu\tilde{p}(s-a) - 1\!\!\mathrm{I}\{n < s-a\}\tilde{\lambda}^{(I)}}{(s-a)\mu\tilde{p}(s-a) - \tilde{\lambda}^{(I)} \sum\limits_{i=0}^{s-a-1} \tilde{p}(i)}\, \tilde{p}(n), \qquad n = 0, 1, \ldots, s+k. \tag{A.16}$$

In the special case of $r_n \equiv 0$ and $X \equiv \infty$, (A.16) is equivalent to (4.5), (4.6), where the probability p_0 that the VMS is empty on the condition that precisely a agents are idle and no calls are waiting in the waiting room is given by

$$p_0 = 1 - EM = \frac{(s-a)\mu\tilde{p}(s-a) - \tilde{\lambda}^{(I)}}{(s-a)\mu\tilde{p}(s-a)}. \tag{A.17}$$

In view of the preceding considerations the stability condition reads $EM < 1$, in view of (A.12) being equivalent to

$$\tilde{\lambda}^{(I)} < (s-a)\mu\tilde{p}(s-a). \tag{A.18}$$

In the special case considered here the condition (A.18) is equivalent to (4.12), cf. Remark 4.3.

References

[A] Asmussen, S., Applied probability and queues. J. Wiley & Sons, Chichester 1987.

[BH] Baccelli, F., Hebuterne, G., On queues with impatient customers. Performance '81. F.J. Kylstra (editor). North-Holland Publishing Company, (1981) 159–179.

[B] Boxma, O.J., Workloads and waiting times in single-server systems with multiple customer classes. Queueing Systems 5, No. 1–3 (1989) 185–214.

[BB] Brandt, A., Brandt, M., On the $M(n)/M(n)/s$ queue with impatient calls. Submitted for publication.

[BBSW] Brandt, A., Brandt, M., Spahl, G., Weber, D., Modelling and optimization of call distribution systems. Proc. 15th Int. Teletraffic Cong. (ITC 15), Washington, DC, USA (1997) 133–144.

[BFL] Brandt, A., Franken, P., Lisek, B., Stationary stochastic models. Akademie Verlag, Berlin; Wiley, Chichester 1990.

[C] Cravis, H., Traffic engineering with an ACD. TE&M, July (1990) 56–59.

[DPW] Dumas, G., Perkins, M., White, C., Improving efficiency of PBX-based call centers: Combining inbound and outbound agents with automatic call sharing. Proc. 15th Int. Switching Symposium Berlin, (1995) 346–350.

[FC] Fuhrmann, S., Cooper, R.B., Stochastic decomposition in $M/G/1$ queue with generalized vacations. Operations Research 33 (1985) 1117–1129.

[FKAS] Franken, P., König, D., Arndt, U., Schmidt, V., Queues and point processes. Akademie Verlag, Berlin 1981; Wiley, Chichester 1982.

[G] Gable, R.A., Inbound call centers: Design, implementation and management. Artech House, Boston, London 1993.

[GK] Gnedenko, B.W., Kowalenko, I.N., Einführung in die Bedienungstheorie (first edition in Russian, Nauka, 1966). Akademie Verlag, Berlin 1971, 1974.

[HHP] Harvey, D.E., Hogan, S.M., Payseur, J.Y., Call center solutions. AT&T Technical Journal, Sept./Oct. (1991) 36–44.

[H] Haugen, F.R.B., Queueing systems with several input streams and time out. Telektronikk No. 2 (1978) 100–106.

[HS] Haugen, R.B., Skogan, E., Queueing systems with stochastic time out. IEEE Trans. Commun. vol. COM-28 (1980) 1984–1989.

[J1] Jurkevič, O.M., On the investigation of many-server queueing systems with bounded waiting time (in Russian). Izv. Akad. Nauk SSSR Techničeskaja kibernetika No. 5 (1970) 50–58.

[J2] Jurkevič, O.M., On many-server systems with stochastic bounds for the waiting time (in Russian). Izv. Akad. Nauk SSSR Techničeskaja kibernetika No. 4 (1971) 39–46.

[KLS] Konheim, A.G., Levy, H., Srinivasan, M.M., Descendant set: An efficient approach for the analysis of polling systems. IEEE Trans. Comm. 42 (1994) 1245–1253.

[KY] Kitaev, M.Y., Yashkov, S.F., Distribution of the conditional sojourn time of requests in shared-processor systems. Izv. Akad. Nauk SSSR Techničeskaja kibernetika No. 4 (1978) 211–215.

[M] Mei, R.D., Polling systems in heavy traffic: Higher moments in the delay. Proc. 15th Int. Teletraffic Cong. (ITC 15), Washington, DC, USA (1997) 275–296.

[P] Perry, M., Performance modelling of automatic call distributors. Ph. D. Thesis, North Carolina State University, 1991.

[PN] Perry, M., Nilsson, A., Performance modelling of automatic call distributors: Assignable grade of service staffing. Proc. 14th Int. Switching Symposium Yokohama (1992) 294–298.

[ST] So, K.C., Tang, C., Operational strategies for managing congestion in service systems. Working paper, Graduate School of Management University of California, Irvine, California, 1993.

[T] Tijms, H.C., Stochastic models. J. Wiley & Sons, Chichester 1994.

[Y] Yashkov, S.F., A derivation of response time distribution for a $M/G/1$ processor sharing queue. Problems Contr. Info. Theory 12 (1983) 133–148.

[W] Wallstrøm, B., A queueing system with time-outs and random departure. Proc. ITC 8 Melbourne (1976), paper 231.