Math Meth Oper Res (2010) 72:453–476
DOI 10.1007/s00186-010-0328-z

ORIGINAL ARTICLE

On a processor sharing queue that models balking

Qiang Zhen · Johan S. H. van Leeuwaarden · Charles Knessl

Received: 9 April 2009 / Accepted: 24 August 2010 / Published online: 9 September 2010

Abstract We consider the processor sharing $M/M/1$-PS queue which also models balking. A customer that arrives and sees n others in the system "balks" (i.e., decides not to enter) with probability $1-b_n$. If b_n is inversely proportional to $n+1$, we obtain explicit expressions for a tagged customer's sojourn time distribution. We consider both the conditional distribution, conditioned on the number of other customers present when the tagged customer arrives, as well as the unconditional distribution. We then evaluate the results in various asymptotic limits. These include large time (tail behavior) and/or large n, lightly loaded systems where the arrival rate $\lambda \to 0$, and heavily loaded systems where $\lambda \to \infty$. We find that the asymptotic structure for the problem with balking is much different from the standard $M/M/1$-PS queue. We also discuss a perturbation method for deriving the asymptotics, which should apply to more general balking functions.

Keywords Queueing theory · Processor sharing · Balking · Asymptotics

Q. Zhen (✉)
Department of Mathematics and Statistics, University of North Florida, 1 UNF DR, Bldg 14/2731, Jacksonville, FL 32224-7699, USA
e-mail: q.zhen@unf.edu

J. S. H. van Leeuwaarden
Department of Mathematics and Computer Science, Eindhoven University of Technology, Room HG 9.13, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: j.s.h.v.leeuwaarden@tue.nl

C. Knessl
Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 South Morgan (M/C 249), Chicago, IL 60607-7045, USA
e-mail: knessl@uic.edu

1 Introduction

Balking refers to the phenomenon that customers, when forced to wait for service, refuse to join the queue. First introduced by Haight (1957), balking can be specified by a probability distribution associated with the system state. Specifically, a customer that finds n customers in the systems upon arrival, balks with probability $1 - b_n$ and joins the queue with probability b_n. Haight considered several examples of balking functions, including $b_n = 1/(n+1)$, $b_n = 1\{n \leq K\}$ (with $1\{\cdot\}$ the indicator function), and $b_n = e^{-cn}$. The latter case was also studied in detail by Morse (1958). In this paper we shall investigate the effect of balking upon systems with processor sharing (PS).

A model for the round-robin scheduling mechanisms in time-shared computer systems, processor sharing was first introduced by Kleinrock (1964), and refers to the service discipline under which every customer gets a fair share of the server. It is by now well known that PS is intimately related to random order of service (ROS), which refers to the discipline where customers are chosen for service at random. First studied by Vaulot (1946) and Pollaczek (1946), the ROS discipline has a long tradition in queueing theory. Pollaczek obtained the Laplace transform of the distribution of the steady-state waiting time $\mathcal{W}$ in the $M/M/1$-ROS queue, by solving a differential-difference equation. In fact, the latter was almost identical to the differential-difference equation studied by Coffman et al. (1970) for the $M/M/1$-PS queue. Indeed, by comparing these differential-difference equations, it is readily established that (see Cohen 1984)

$$\Pr[\mathcal{V} > t] = C \cdot \Pr[\mathcal{W} > t]$$

with $\mathcal{V}$ the steady-state sojourn time in the $M/M/1$-PS queue and C a constant. A probabilistic argument based on coupling was given in Borst et al. (2003), and the equivalence result was shown to extend to other models as well, including finite capacity queues, repairman problems and networks. We shall show that for the $M/M/1$ queue with balking the equivalence result also holds.

The distribution of $\mathcal{W}$ does not have a simple representation. Pollaczek (1946) was able to invert the Laplace transform and obtained a rather intricate but explicit integral representation for the waiting time distribution. The integral, along with the method of steepest descent, allowed Pollaczek to derive an intriguing asymptotic expression for the tail distribution $\Pr[\mathcal{W} > t]$. This asymptotic expression was rediscovered by Flatto (1997). Morrison (1985) considered the heavy-traffic limit, where the traffic intensity $\rho \to 1$, and derived asymptotic expansions in powers of $1 - \rho$ for the sojourn time distributions. The tail results of Pollaczek and Flatto were related to the heavy-traffic results in Morrison recently in Zhen and Knessl (2007).

In this paper we consider the $M/M/1$-PS (or ROS) queue with balking. The sojourn time distribution (or waiting time distribution in the ROS model) satisfies a differential-difference equation that differs only slightly from the one considered by Pollaczek for the systems without balking. However, the analysis, and also the system behavior, changes drastically. We shall assume that $b_n = 1/(n+1)$, so that the non-balking probability exactly matches the share of a server that the customers get upon arrival.

For this choice of b_n the differential-difference equation allows for an exact and asymptotic analysis. We obtain the following results:

(i) An exact spectral representation for the sojourn time density in terms of generalized Laguerre polynomials.
(ii) An expression for the Laplace-Stieltjes transform of the sojourn time density.
(iii) Asymptotic results for tail probabilities when ρ is fixed; asymptotics for the light-traffic case where $\rho \to 0$; and asymptotics for the heavy-traffic case where $\rho \to \infty$.

Gromoll et al. (2008) have recently investigated a PS queue with customer impatience (or reneging), where a customer may leave without completing service, if that customer's sojourn time would exceed a prescribed limit. In that model a customer who reneges from a PS queue will have already received partial service, which implies loss of work conservation. With balking this is not the case, because a customer simply does not enter the system. Here we make Markovian assumptions and carry out exact and precise asymptotic analyses, while in Gromoll et al. (2008) the authors allow for general service and arrival distributions, but analyze their model more approximately, through fluid limits. For future work, we hope to consider balking and reneging from the viewpoint of the server as two different ways of controlling access to the system.

We shall derive our results by both a perturbation method and by using the exact representations. The former method should also be useful for general non-balking functions b_n, provided that we can write ρb_n in the form $\rho b_n = B(\varepsilon n)$, where ε is a small parameter. Thus ρb_n is a "slowly varying" function of n. For example, this would apply to $b_n = e^{-cn}$ [used by Morse (1958)] if c is small. This limit would also apply to repairman problems (or finite populations queues) where $b_n = M - n$ and M is the customer population. Then $\rho b_n = \rho M(1 - n/M)$ and we would assume that $M \to \infty$ (thus $\varepsilon = M^{-1}$) and $\rho \to 0$, with $\rho M = O(1)$. It is likely that the asymptotic structure of all of these models is quite different, and the perturbation method should clearly show these differences.

1.1 Equivalence relation

We denote the sojourn time of a non-balking customer that arrives to a PS queue with n other customers competing for service by $\mathcal{V}_n$, and the waiting time of a non-balking customer that arrives to a ROS queue with n other customers waiting for service and one additional customer in service by $\mathcal{W}_n$. Then we let b_n and b_n^r be the non-balking probabilities in the PS system and the ROS system, respectively, when there are n customers in the system (including the customer in service).

Proposition 1 *If* $b_0^r = 1$ *and* $b_n^r = b_{n-1}$, $n = 1, 2, \ldots$, *then*

$$\mathcal{V}_n \stackrel{d}{=} \mathcal{W}_n \tag{1.1}$$

and

$$\Pr[\mathcal{V} > t] = C \cdot \Pr[\mathcal{W} > t] \tag{1.2}$$

with

$$C = \frac{1}{\rho} \cdot \frac{1 + \sum_{n=1}^{\infty} \rho^n b_0 \dots b_{n-1}}{1 + \sum_{n=1}^{\infty} \rho^n b_0 \dots b_n}.$$

Proof Borst et al. (2003) made the observation that whenever a service completion occurs in the PS system, each of the customers present is equally likely to be the one that departs due to the memoryless property of the exponential distribution. In that respect, the pool of customers competing for service under PS behaves exactly as the pool of customers waiting for service under ROS. Note that the arrival processes in both systems can be coupled due to the assumption that $b_0^r = 1$ and $b_n^r = b_{n-1}$, $n = 1, 2, \dots$. A similar coupling argument as in Borst et al. (2003) then yields (1.1).

Let $\mathcal{N}_p$ and $\mathcal{N}_r$ denote the number of customers at arrival epochs in the PS system and ROS system, respectively. Then, $\Pr[\mathcal{N}_p = n] = \Pr[\mathcal{N}_p = 0]\rho^n b_0 \dots b_{n-1}$ and

$$\Pr[\mathcal{V} > t] = \frac{\sum_{n=0}^{\infty} \rho^n b_0 \dots b_n \Pr[\mathcal{V}_n > t]}{\sum_{n=0}^{\infty} \rho^n b_0 \dots b_n}. \tag{1.3}$$

Similarly, $\Pr[\mathcal{N}_r = n] = \Pr[\mathcal{N}_r = 0]\rho^n b_0^r \dots b_{n-1}^r$ and

$$\Pr[\mathcal{W} > t] = \frac{\sum_{n=0}^{\infty} \rho^{n+1} b_0^r \dots b_{n+1}^r \Pr[\mathcal{W}_n > t]}{\sum_{n=0}^{\infty} \rho^n b_0^r \dots b_n^r}. \tag{1.4}$$

Upon comparing (1.3) and (1.4), and using $b_0^r = 1$, $b_n^r = b_{n-1}$ for $n = 1, 2, \dots$, and $\mathcal{V}_n \stackrel{d}{=} \mathcal{W}_n$, the equivalence relation (1.2) follows. □

The $M/M/1/K$-PS queue can be viewed as a special case of the $M/M/1$-PS queue with balking by choosing $b_n = 1$ if $n \leq K - 1$ and 0 otherwise. In that case the equivalence relation becomes

$$\Pr[\mathcal{V} > t] = \frac{1}{\rho} \cdot \frac{1 - \rho^{K+1}}{1 - \rho^K} \cdot \Pr[\mathcal{W} > t],$$

which was already obtained in Borst et al. (2003). For this $M/M/1/K$-PS queue, Knessl (1993) uses singular perturbation techniques to construct asymptotic approximations to the sojourn time distribution.

We shall consider the $M/M/1$-PS queue with non-balking $b_n = \frac{1}{n+1}$, in which case we have the equilibrium distribution (see Haight 1957)

$$\Pr[\mathcal{N}_p = n] = \frac{e^{-\rho}\rho^n}{n!}, \quad n = 0, 1, \dots \tag{1.5}$$

and

$$\Pr[\mathcal{V} > t] = \frac{e^{\rho}}{e^{\rho} - 1} \cdot \Pr[\mathcal{W} > t]. \tag{1.6}$$

2 Problem statement and summary of results

We consider a processor sharing $M/M/1$ queue, which also models balking. Customers arrive at rate λ and the service rate will be denoted by μ. We can clearly scale time so as to have $\mu = 1$, and then the traffic intensity is $\rho = \lambda/\mu = \lambda$. We recall that $\mathcal{V}_n$ is the sojourn time of a tagged customer that finds n others in the system upon arrival. We then define $\mathbf{V}_n(t) = \Pr[\mathcal{V}_n > t]$. With the PS discipline, each customer receives service at rate $1/n$ when there are n customers in service. When a tagged customer arrives we assume that he/she will enter the system with probability b_n, and "balk" with the remaining probability $1 - b_n$.

The function $\mathbf{V}_n(t)$ satisfies the differential-difference equation

$$\mathbf{V}_n'(t) = \frac{n}{n+1}\,\mathbf{V}_{n-1}(t) - (1 + \rho\, b_n)\;\mathbf{V}_n(t) + \rho\, b_n \mathbf{V}_{n+1}(t) \tag{2.1}$$

with $\mathbf{V}_n(0) = 1$. (There is a slight error in Riordan (1962), where this equation was previously given.) It is reasonable to define $b_0 = 1$ and have b_n a decreasing function of n. Here we assume that

$$b_n = \frac{1}{n+1}, \quad n = 0, 1, 2, 3, \ldots. \tag{2.2}$$

Then we define the sojourn time density $p_n(t)$ by $p_n(t) = -\mathbf{V}_n'(t)$, which satisfies:

$$p_n'(t) = \frac{n}{n+1}\,p_{n-1}(t) - \left(1 + \frac{\rho}{n+1}\right)\,p_n(t) + \frac{\rho}{n+1}\,p_{n+1}(t), \quad t > 0 \tag{2.3}$$

with the initial condition

$$p_n(0) = \frac{1}{n+1}, \quad n \geq 0. \tag{2.4}$$

Clearly (2.2) is a very special case of b_n. We can consider also $b_n = \alpha/(n+1)$, since α can be incorporated into the traffic intensity ρ. However, with even a slight change (such as taking $b_n = 1/(n+\beta)$ with $\beta \neq 1$) it seems that the problem is no longer amenable to exact solution. We shall discuss an asymptotic approach to solving (2.1) (cf. Section 5), which should work also for more general b_n.

We give below various exact and asymptotic expressions for $p_n(t)$.

Theorem 2.1 *The conditional sojourn time density has the following exact expression (spectral representation):*

$$p_n(t) = \sum_{m=1}^{\infty} C_m(\nu_m)\,\phi_m(n, \nu_m)\,e^{\nu_m t} + \sum_{m=1}^{\infty} C_m(\widetilde{\nu}_m)\,\phi_m(n, \widetilde{\nu}_m)\,e^{\widetilde{\nu}_m t}, \tag{2.5}$$

where

$$\nu_m = -1 + \frac{1}{2\,m}\left[-\rho + \sqrt{\rho^2 + 4m\rho}\right], \tag{2.6}$$

$$\widetilde{\nu}_m = -1 + \frac{1}{2\,m}\left[-\rho - \sqrt{\rho^2 + 4m\rho}\right], \tag{2.7}$$

$$C_m(\nu) = \frac{m^{m-1}}{m!}\,\frac{\nu}{\nu - 1}\,e^{-m}, \tag{2.8}$$

and

$$\phi_m(n, \nu) = n!\left(\frac{\nu + 1}{-\rho}\right)^n L_n^{(m-1-n)}\left(\frac{\rho}{(\nu + 1)^2}\right). \tag{2.9}$$

Here $L_n^{(l)}(z)$ is the generalized Laguerre polynomial (*see* Magnus et al. 1966).

If we take the Laplace transform of (2.3) and multiply by $n + 1$, we have

$$\rho\,\widehat{p}_{n+1}(\theta) - [(n + 1)(\theta + 1) + \rho]\,\widehat{p}_n(\theta) + n\,\widehat{p}_{n-1}(\theta) = -1, \tag{2.10}$$

where $\widehat{p}_n(\theta) = \int_0^\infty p_n(t)e^{-\theta t}dt$. Solving the recurrence Eq. (2.10), we obtain another exact expression for $p_n(t)$, in terms of its Laplace transform.

Theorem 2.2 *The Laplace transform of the conditional sojourn time density has the following form:*

$$\widehat{p}_n(\theta) = M\,G_n \sum_{l=0}^{n} \frac{\rho^l}{l!}\,H_l + M\,H_n \sum_{l=n+1}^{\infty} \frac{\rho^l}{l!}\,G_l, \tag{2.11}$$

where

$$M = M(\theta) \equiv \frac{\rho^{r+1}}{(1 + \theta)\,\Gamma(r + 1)}\,e^{r/\theta}, \tag{2.12}$$

$$G_n = G_n(\theta) \equiv \int_0^{\frac{1}{(1+\theta)}} z^n \left(\frac{1}{1 + \theta} - z\right)^r \exp\left(-\frac{\rho\,z}{1 + \theta}\right) dz, \tag{2.13}$$

$$H_n = H_n(\theta) \equiv \int_{\frac{1}{(1+\theta)}}^{\infty} z^n \left(z - \frac{1}{1 + \theta}\right)^r \exp\left(-\frac{\rho\,z}{1 + \theta}\right) dz, \tag{2.14}$$

and

$$r = r(\theta) \equiv \frac{\rho\,\theta}{(1 + \theta)^2}.$$

The first two conditional moments of the sojourn time are

$$\mathcal{M}_n = \int_0^\infty t\, p_n(t)\, dt = \frac{n+\rho}{2} + 1,$$

$$\mathcal{S}_n = \int_0^\infty t^2\, p_n(t)\, dt = \frac{n^2}{3} + \left(2 + \frac{5}{6}\,\rho\right) n + \frac{5}{6}\,\rho^2 + 3\rho + 2.$$

Using (2.11), we obtain the following asymptotic expansions for $p_n(t)$, valid for $\rho > 0$ and n and/or $t \to \infty$. Throughout the paper the notation $f(x) \sim g(x)$ as $x \to x_0$ means that $\lim_{x\to x_0}[f(x)/g(x)] = 1$.

Theorem 2.3 *For a fixed $\rho > 0$ with $n,\ t \to \infty$, the conditional sojourn time density has the following asymptotic expansions:*

1. *When $n \to \infty$, $n/t > 1$,*

$$p_n(t) = \frac{1}{n} - \frac{\rho}{n(n-t)} + \frac{\rho - 1}{n^2} + O(n^{-3}). \tag{2.15}$$

2. *When $n/t = 1 + \Delta\, t^{-1/2} = 1 + O(t^{-1/2})$,*

$$p_n(t) \sim \frac{1}{2n}\,\mathrm{erfc}\left(-\frac{\Delta}{\sqrt{2}}\right) = \frac{1}{n\sqrt{\pi}} \int_{-\Delta/\sqrt{2}}^\infty e^{-u^2}\, du. \tag{2.16}$$

3. *When $\Lambda_0 < n/t < 1$ with $\Lambda_0 = \left(-\rho + \sqrt{\rho^2 + 4\rho}\right)/2$,*

$$p_n(t) \sim \frac{\Gamma(r_* + 1)\, e^{r_*}}{\sqrt{2\pi}\,(1 - n/t)^{r_*+1}}\, n^{-3/2-\rho\, t/n+\rho\, t^2/n^2} \left(\frac{t}{n}\right)^n e^{n-t}, \tag{2.17}$$

where

$$r_* = r_*\left(\frac{n}{t}\right) \equiv \rho\, \frac{t^2}{n^2}\left(\frac{n}{t} - 1\right).$$

4. *When $n/t = \Lambda_0 + \Lambda/\sqrt{t}$, $\Lambda = O(1)$,*

$$p_n(t) \sim \frac{\sqrt{\rho+4} - \sqrt{\rho}}{4\sqrt{\rho+4}}\, e^{-1}\, \Lambda_0^{-n}\, e^{-t+\Lambda_0 t}\, \mathrm{erfc}\left\{\frac{\Lambda}{\sqrt{2\Lambda_0}}\right\}. \tag{2.18}$$

5. *When $n/t < \Lambda_0$,*

$$p_n(t) \sim \frac{\sqrt{\rho+4} - \sqrt{\rho}}{2\sqrt{\rho+4}}\, e^{-1}\, \Lambda_0^{-n}\, e^{-t+\Lambda_0 t}. \tag{2.19}$$

Expression (2.19) applies also to $t \to \infty$ with $n = O(1)$, and gives the exponential decay rate of the density $p_n(t)$. We note that the right side of (2.19) is precisely the $m = 1$ term in the first sum in (2.5), i.e., $C_1(\nu_1)\,\phi_1(n, \nu_1)\,e^{\nu_1 t}$. If we start with a fixed large n and increase time t from $t = 0$, we traverse cases 1–5 in Theorem 2.3 in the order given. The leading term in (2.15) is $p_n(t) \sim 1/n$ for $t < n$ which corresponds to a uniform distribution. The $O(n^{-2})$ correction term(s) have a singularity as $t \uparrow n$, which indicates that the asymptotics become invalid. We also note that if $t = 0$, (2.15) becomes $p_n(0) = 1/n - 1/n^2 + O(n^{-3})$, which is just the large n expansion of the initial condition $p_n(0) = 1/(n+1)$. As t/n increases through one, there is a transition region (cf. 2.16) and then for $t/n > 1$ (but with $t/n < 1/\Lambda_0$) the density becomes exponentially small, with a rather intricate dependence on the space–time ratio, as given in (2.17). After another transition region where $t/n \approx 1/\Lambda_0$ (cf. 2.18) the density becomes purely exponential in t, which corresponds to the dominant singularity in the Laplace transform $\widehat{p}_n(\theta)$, which occurs at $\theta = \nu_1 < 0$.

We next consider a small traffic intensity, $\rho \to 0^+$. We shall consider the time scales $t = O(\rho^{-1/2})$ and $t = O(1)$, since for $t = O(\rho^{-1})$ the results can be obtained as limiting cases of Theorem 2.3.

Theorem 2.4 *For $\rho \to 0^+$, the conditional sojourn time density has the following asymptotic expansions:*

1. *For $t = \omega/\sqrt{\rho} = O(\rho^{-1/2})$ and $n = O(1)$, we have*

$$p_n(t) \sim e^{-t}\,\rho^{-n/2}\,Q_n(\omega), \tag{2.20}$$

where

$$\begin{aligned} Q_n(\omega) = &\sum_{m=1}^{\infty} (-1)^n\,\frac{n!\,m^{m-n/2-1}}{2\,m!}\,e^{-m}\,L_n^{(m-1-n)}(m)\,e^{\omega/\sqrt{m}} \\ &+ \sum_{m=1}^{\infty} \frac{n!\,m^{m-n/2-1}}{2\,m!}\,e^{-m}\,L_n^{(m-1-n)}(m)\,e^{-\omega/\sqrt{m}}. \end{aligned} \tag{2.21}$$

2. *For $t, n = O(1)$, we have*

$$p_n(t) = p_n^{(0)}(t) + \rho\,p_n^{(1)}(t) + O(\rho^2), \tag{2.22}$$

where

$$p_n^{(0)}(t) = \frac{1}{n+1}\,\frac{1}{2\pi i}\int_{\mathcal{B}} \left[1 - \left(\frac{1}{1+\theta}\right)^{n+1}\right] \frac{e^{\theta t}}{\theta}\,d\theta = \frac{e^{-t}}{n+1}\sum_{l=0}^{n}\frac{t^l}{l!}$$

and

$$p_n^{(1)}(t) = \frac{e^{-t}}{n+1}\left[\frac{t^{n+2}}{(n+2)!}\sum_{l=0}^{n}\frac{1}{l+2} + \sum_{l=1}^{n+1}\frac{t^l}{l!}\left(\frac{1}{n+2} - \frac{1}{n+2-l}\right)\right].$$

Here $\mathcal{B}$ is a vertical contour in the θ-plane with $\Re(\theta) > 0$.

For $n,\ t = O(1)$ the leading term in (2.22) corresponds to the tagged customer entering the system and no further arrivals entering during his/her sojourn time.

The result in (2.20) is obtained by letting $t = \omega/\sqrt{\rho}$ and taking $\rho \to 0$ in the exact expression (2.5). If we let $\omega \to \infty$, the $m = 1$ term in the first summation in (2.21) dominates.

Finally, we consider a large traffic intensity, $\rho \to \infty$. The structure of $p_n(t)$ is different in two cases.

Theorem 2.5 *For $\rho \to \infty$, the conditional sojourn time density has the following asymptotic expansions:*

1. *When $t = T\rho = O(\rho)$ and $n = N\rho = O(\rho)$,*

$$p_n(t) = \rho^{-1}\, P_0(N, T) + \rho^{-2}\, P_1(N, T) + O(\rho^{-3}), \tag{2.23}$$

where

$$P_0(N, T) = \frac{e^{U-T}}{N-U} = \frac{N-U-1}{(N-1)\,(N-U)} \tag{2.24}$$

and $U = U(N, T)$ is defined implicitly by

$$\frac{U}{N-1} = 1 - e^{U-T}. \tag{2.25}$$

If $N = 1$ we obtain the explicit form $P_0(1, T) = e^{-T}$.

2. *When $t = \tau/\rho = O(\rho^{-1})$ and $n = O(1)$,*

$$p_n(t) \sim \int_0^1 (1-\xi)^n\, J_0\left(2\sqrt{\tau}\sqrt{-\xi - \log(1-\xi)}\right) d\xi, \tag{2.26}$$

where $J_0(\cdot)$ is the Bessel function of the first kind.

We shall compute $P_0(N, T)$ and the correction term $P_1(N, T)$ (cf. (5.7)) in Sect. 5 and also give some alternate expressions for $P_0(N, T)$, as infinite series. The expression in (2.23) remains valid for $n = O(1)$ and $t = O(\rho)$, as well as $t = O(1)$ and $n = O(\rho)$. For $N/T \gg 1$ we have $U \sim T$ and then $P_0(N, T) \sim 1/N$ which is consistent with $p_n(0) = 1/(n+1) \sim \rho^{-1}/N$. For $T/N \gg 1$, $U \to -1$ and we obtain $P_0(N, T) \sim e^{-1}\, e^{-T}$, which is consistent with the expansion of $C_1\, \phi_1\, e^{\nu_1 t}$ for $\rho \to \infty$ and $t = O(\rho)$. Note that $\nu_1 \sim -1/\rho$ from (2.6).

In steady state we remove the conditioning and use (1.5) to get the unconditional sojourn time density for the PS model as

$$p_{PS}(t) = \sum_{n=0}^{\infty} \frac{\rho^n}{n!}\, e^{-\rho}\, p_n(t). \tag{2.27}$$

Then the density $p(t)$ for the ROS model follows from (1.6) as

$$p(t) = (1 - e^{-\rho})\, p_{PS}(t).$$

Note also that the full density, $p_{ROS}(t)$, for the ROS model is $e^{-\rho}\,\delta(t) + p(t)$, since there is a non-zero probability that $\mathcal{W} = 0$. The exact representation for $p_{PS}(t)$ is as follows.

Theorem 2.6 *The unconditional sojourn time density has the exact expression*

$$p_{PS}(t) = \sum_{m=1}^{\infty} C_m(\nu_m)\, \Phi_m(\nu_m)\, e^{\nu_m t} + \sum_{m=1}^{\infty} C_m(\widetilde{\nu}_m)\, \Phi_m(\widetilde{\nu}_m)\, e^{\widetilde{\nu}_m t},$$

where

$$\Phi_m(\nu) = e^{-\rho}\, (-\nu)^{m-1}\, \exp\left(\frac{\rho}{\nu + 1}\right).$$

We also give the asymptotic results for $p_{PS}(t)$ and $p(t)$ for the different scales of ρ and t.

Theorem 2.7 *The unconditional sojourn time density for the PS model and waiting time density for the ROS model have the following asymptotic expansions:*

1. *For ρ fixed with $t \to \infty$, we have*

$$p_{PS}(t) = \frac{p(t)}{1 - e^{-\rho}} \sim \frac{\sqrt{\rho + 4} - \sqrt{\rho}}{2\sqrt{\rho + 4}}\, e^{-1-\rho}\, e^{\rho/\Lambda_0}\, e^{-t}\, e^{\Lambda_0 t}. \tag{2.28}$$

2. *For $\rho \to 0$, there are three scales of t.*
   (a) If $t = \zeta/\rho = O(\rho^{-1})$, we have

$$p_{PS}(t) \sim \rho^{-1}\, p(t) \sim \frac{1}{2}\, e^{-1}\, e^{-\zeta/2}\, e^{-(1-\sqrt{\rho})t}. \tag{2.29}$$

   (b) If $t = \omega/\sqrt{\rho} = O(\rho^{-1/2})$, we have

$$p_{PS}(t) \sim \rho^{-1}\, p(t) \sim e^{-t}\, Q_0(\omega), \tag{2.30}$$

   where

$$Q_0(\omega) = \sum_{m=1}^{\infty} \frac{m^{m-1}}{m!}\, e^{-m}\, \cosh(\omega/\sqrt{m}).$$

(c) If $t = O(1)$, we have

$$p_{PS}(t) = e^{-t}\left[1 + \frac{\rho}{4}(t^2 - 2) + O(\rho^2)\right]. \tag{2.31}$$
$$p(t) = \rho\, e^{-t}\left[1 + \frac{\rho}{4}(t^2 - 4) + O(\rho^2)\right].$$

3. For $\rho \to \infty$ with $t = T\,\rho = O(\rho)$, we have

$$p_{PS}(t) \sim p(t) \sim \frac{1}{\rho}\, e^{-T}. \tag{2.32}$$

For fixed ρ and large t, we removed the condition by using the expansion in the region $t/n > 1/\Lambda_0$ (i.e., (2.19)) in (2.27), thus obtaining (2.28).

For a small traffic intensity ρ, (2.29) on the $t = O(\rho^{-1})$ scale is the limiting case of (2.28), as $\rho \to 0$. For the scale $t = O(\rho^{-1/2})$, we used (2.20) in (2.27). Since ρ is small, the $n = 0$ term dominates, which leads to (2.30). When $t = O(1)$, using (2.22) in (2.27) and the fact that $e^{-\rho} \sim 1 - \rho$ leads to

$$p_{PS}(t) = (1 - \rho)\left[p_0^{(0)}(t) + \rho\, p_1^{(0)}(t) + \rho\, p_0^{(1)}(t) + O(\rho^2)\right],$$

which yields (2.31). We note that if we let $\omega \to 0$ in (2.30), (2.30) reduces to the leading term in (2.31). This indicates that the $t = O(1)$ scale is a special case of the $t = O(\rho^{-1/2})$ scale, for small ρ.

In the case $\rho \to \infty$ with $t = O(\rho)$, we used the leading term in (2.23) in (2.27) and noticed that the infinite sum concentrates near $n = \rho$ (i.e., $N = 1$), which led to (2.32). In fact this result is uniform on both the $t = O(\rho)$ and $t = O(\rho^{-1})$ time scales, for large ρ.

3 Derivations of the exact representations

3.1 Proof of Theorem 2.1

We first derive the spectral representation (2.5) of the conditional sojourn time density. Consider the Eq. (2.3) and assume that $p_n(t)$ has the form $p_n(t) = e^{\nu t}\,\phi(n)$. Then the exponential generating function $G(z)$ of $\phi(n)$,

$$G(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}\,\phi(n), \tag{3.1}$$

satisfies

$$[(\nu + 1)\,z - \rho]\; G'(z) + (\nu + 1 + \rho - z)\, G(z) = 0. \tag{3.2}$$

Here we assumed that $n\,\phi(n-1)$ is finite as $n \to 0$. Solving (3.2), we have

$$G(z) = C \exp\left(\frac{z}{\nu+1}\right)\left(1 - \frac{\nu+1}{\rho}\,z\right)^{-R_0-1}, \quad R_0 = \frac{\rho\,\nu}{(1+\nu)^2}, \tag{3.3}$$

where $C = G(0) = \phi(0)$. Without loss of generality, we let $\phi(0) = C = 1$.

To avoid $\phi(n)$ growing like $n!$ as $n \to \infty$, $G(z)$ must be an entire function of z, so that $-R_0 - 1$ must be a non-negative integer. The eigenvalues ν thus satisfy the quadratic equation $R_0 = -m, m = 1, 2, \ldots$, which leads to the two sets of eigenvalues given by (2.6) and (2.7). We denote by $\phi_m(n, \nu_m)$ and $\phi_m(n, \widetilde{\nu}_m)$ the eigenfunctions corresponding to the eigenvalues ν_m and $\widetilde{\nu}_m$, respectively. Then after some calculation and inversion of (3.1) we find that $\phi_m(n, \nu_m)$ and $\phi_m(n, \widetilde{\nu}_m)$ both satisfy (2.9) with $\nu = \nu_m$ and $\nu = \widetilde{\nu}_m$ respectively.

Thus, we can express the conditional sojourn time density as the spectral representation in (2.5), with only the two coefficient sequences $C_m(\nu_m)$ and $C_m(\widetilde{\nu}_m)$ to be determined.

To determine these coefficients, we can easily establish the following orthogonality relation for the eigenfunctions:

$$\sum_{n=0}^{\infty} \frac{n+1}{n!}\,\rho^n\,\phi_m(n, \nu_m)\,\phi_{m'}(n, \nu_{m'}) = 0, \quad \nu_m \neq \nu_{m'}. \tag{3.4}$$

By the spectral representation (2.5) and the initial condition (2.4), we must have

$$\sum_{m=1}^{\infty} C_m(\nu_m)\,\phi_m(n, \nu_m) + \sum_{m=1}^{\infty} C_m(\widetilde{\nu}_m)\,\phi_m(n, \widetilde{\nu}_m) = \frac{1}{n+1}.$$

Using (3.4), we can easily show that, for any eigenvalue ν_m or $\widetilde{\nu}_m$,

$$C_m(\nu_m) = \frac{\sum_{n=0}^{\infty} \rho^n\,\phi_m(n, \nu_m)/n!}{\sum_{n=0}^{\infty}(n+1)\,\rho^n\,\phi_m^2(n, \nu_m)/n!}. \tag{3.5}$$

Using the generating function (3.1) and (3.3), the numerator in (3.5) is

$$\sum_{n=0}^{\infty} \frac{\rho^n}{n!}\,\phi_m(n, \nu_m) = G_m(\rho) = (-\nu_m)^{m-1}\,\exp\left(\frac{\rho}{\nu_m+1}\right). \tag{3.6}$$

After a lengthy but standard calculation we then obtain

$$\sum_{n=0}^{\infty} \frac{(n+1)}{n!}\,\rho^n\,\phi_m^2(n, \nu_m) = \frac{m!}{m^{m-1}}\,(-\nu_m)^{m-2}\,(1-\nu_m)\exp\left(\frac{\rho}{(\nu_m+1)^2}\right), \tag{3.7}$$

which determines the denominator in (3.5). Using (3.6) and (3.7) in (3.5), we obtain (2.8) with $\nu = \nu_m$. By the same argument, we find that (2.8) is also true for the eigenvalues $\widetilde{\nu}_m$. This completes the (sketched) derivation of Theorem 2.1.

3.2 Proof of Theorem 2.2

We use a discrete Green's function to derive (2.11). Consider the recurrence Eq. (2.10). The discrete Green's function $\mathcal{G}(\theta; n, l)$ satisfies

$$\begin{aligned} &\rho\,\mathcal{G}(\theta; n+1, l) - [(n+1)(1+\theta) + \rho]\,\mathcal{G}(\theta; n, l) \\ &\quad + n\,\mathcal{G}(\theta; n-1, l) = -\delta(n, l), \quad (n, l \geq 0) \end{aligned} \tag{3.8}$$

where $\delta(n, l) = 1\{n = l\}$ is the Kronecker delta. To construct the Green's function we need two linearly independent solutions to

$$\rho\,\mathcal{G}_H(\theta; n+1, l) - [(n+1)(1+\theta) + \rho]\,\mathcal{G}_H(\theta; n, l) + n\,\mathcal{G}_H(\theta; n-1, l) = 0, \tag{3.9}$$

which is the homogeneous version of (3.8).

We seek solutions of (3.9) $\mathcal{G}_H(\theta; n) = G_n$ in the form of contour integrals $G_n = \int_{\mathcal{D}} z^n g(z) dz$, where the function $g(z) = g(z; \theta)$ and the path $\mathcal{D}$ of integration in the complex z-plane are to be determined. Using the above form in (3.9) we obtain an ODE for $g(z)$ whose solution is

$$g(z) = \left(z - \frac{1}{1+\theta}\right)^r \exp\left(-\frac{\rho\, z}{1+\theta}\right), \quad \text{where } r = \frac{\rho\,\theta}{(1+\theta)^2}.$$

If the path of integration $\mathcal{D}$ is chosen as the segment $[0,\, 1/(1+\theta)]$ of the real axis, then we obtain G_n as in (2.13). We note that G_n decays as $n \to \infty$, and by scaling $z = (1 - y/n)/(1+\theta)$ and using the Laplace method, we find that G_n is asymptotically given by

$$G_n \sim \frac{\Gamma(r+1)}{n^{r+1}\,(1+\theta)^{n+r+1}}\, e^{-r/\theta}, \quad n \to \infty. \tag{3.10}$$

However, G_n becomes infinite as $n \to -1$, and $n\,G_{n-1}$ goes to a nonzero limit as $n \to 0$. Thus G_n is not an acceptable solution to (3.9) at $n = 0$.

To construct a second solution to (3.9), we consider another path of the integration, the real interval $[1/(1+\theta),\ \infty)$. Thus, we have another solution of (3.9), H_n, which is given by (2.14). H_n is finite as $n \to -1$, but grows as $n \to \infty$. From (2.14) we find that H_n grows roughly like $n!$ for n large; more precisely

$$H_n \sim n!\, n^r \left(\frac{1+\theta}{\rho}\right)^{n+r+1}, \quad n \to \infty. \tag{3.11}$$

Thus, the discrete Green's function can be represented by

$$\mathcal{G}(\theta; n, l) = \begin{cases} H_l\, G_n\, \mathcal{G}_0 & \text{if } n \geq l \\ G_l\, H_n\, \mathcal{G}_0 & \text{if } 0 \leq n < l, \end{cases} \tag{3.12}$$

which has acceptable behavior both at $n = 0$ and as $n \to \infty$. Here $\mathcal{G}_0$ depends only upon θ and l.

To determine $\mathcal{G}_0$, we let $n = l$ in (3.9) and use the fact that both G_l and H_l satisfy (3.9) with $n = l$. Then we can infer a simple difference equation for the discrete Wronskian $G_l\, H_{l+1} - G_{l+1}\, H_l$, whose solution we write as

$$G_l\, H_{l+1} - G_{l+1}\, H_l = \frac{l!}{\rho^{\,l}\, \mathcal{G}_1}, \tag{3.13}$$

where $\mathcal{G}_1 = \mathcal{G}_1(\theta)$ depends upon θ only. Then using (3.12) in (3.9) with $n = l$ shows that $\mathcal{G}_0$ and $\mathcal{G}_1$ are related by $\mathcal{G}_0 = \rho^{l-1}\, \mathcal{G}_1 / l!$.

Letting $l \to \infty$ in (3.13) and using the asymptotic results in (3.10) and (3.11), we determine $\mathcal{G}_1$ and then obtain $\mathcal{G}_0$ as

$$\mathcal{G}_0 = \frac{\rho^{r+l+1}}{l!\, \Gamma(r+1)\, (1+\theta)}\, e^{r/\theta}.$$

Then, we multiply (3.8) by the solution $\widehat{p}_l(\theta)$ to (2.10) and sum over all $l \geq 0$. After some manipulation this yields

$$\widehat{p}_n(\theta) = \sum_{l=0}^{\infty} \mathcal{G}(\theta; n, l),$$

which is equivalent to (2.11). Taking the inverse Laplace transform gives the conditional sojourn time density $p_n(t)$ as the contour integral

$$p_n(t) = \frac{1}{2\pi i} \int_{\mathcal{B}} \widehat{p}_n(\theta) e^{\theta t} d\theta, \tag{3.14}$$

where $\mathcal{B}$ is a vertical contour in the complex θ-plane, with $\Re(\theta) \geq 0$. The form in (3.14) is more useful than the spectral representation for obtaining asymptotic results in various limits, such as n, t simultaneously large.

The first two moments $\mathcal{M}_n$ and $\mathcal{S}_n$ can be computed by expanding $\widehat{p}_n(\theta)$ about $\theta = 0$, or by using (2.3) to derive simple difference equations for these moments.

4 Asymptotic results for fixed ρ and $\rho \to 0$

4.1 Proof of Theorem 2.3

We first assume that the traffic intensity ρ is fixed. We sketch the main points in deriving Theorem 2.3. We first consider $n, t \to \infty$ with $n > t$ and use the result in (2.11). To obtain a two term approximation, we need the correction terms in the approximations in (3.10) and (3.11), which are given by

$$G_n = \frac{e^{-r/\theta}}{n^{r+1}\,(1+\theta)^{n+r+1}} \left[\Gamma(r+1) + \frac{1}{n}\left(\frac{r}{\theta}\,\Gamma(r+2) - \frac{1}{2}\,\Gamma(r+3)\right) + O(n^{-2})\right] \tag{4.1}$$

and

$$H_n = n!\, n^r \left(\frac{1+\theta}{\rho}\right)^{n+r+1} \left[1 - \frac{r^2}{n\,\theta} + O(n^{-2})\right]. \tag{4.2}$$

From (4.1) and (4.2), we note that the first term in (2.11) dominates the second, and thus the Laplace transform is asymptotically given by

$$\widehat{p}_n(\theta) \sim M\, G_n \sum_{l=0}^{n} \frac{\rho^l}{l!}\, H_l \tag{4.3}$$

$$\sim I_1 + I_2, \quad n \to \infty, \tag{4.4}$$

where

$$I_1 = \left[\frac{1}{1+\theta} + \frac{(1+\theta)\,A}{\rho\,\theta\,n}\right] \int_0^1 (1+\theta)^{-n\,y} (1-y)^r\, dy,$$

$$I_2 = -\frac{\rho}{(1+\theta)^3\, n} \int_0^1 (1+\theta)^{-n\,y} (1-y)^{r-1}\, dy,$$

and

$$A = A(\theta) \equiv \frac{r\,(r+1)}{2\theta}\,[2r - \theta\,(r+2)].$$

Here we used the Euler–Maclaurin summation formula to approximate the sums in (2.11) by integrals. By scaling $\theta = s/n = O(1/n)$ and noting that

$$\frac{(1+\theta)\,A}{\rho\,\theta} = \rho - 1 + \frac{s}{2n}(2\rho^2 - 9\rho + 2) + O(n^{-2})$$

and $r \sim \rho\, s/n$, I_1 becomes, for $n \to \infty$ with s fixed,

$$I_1 \sim \left(1 + \frac{\rho - 1}{n}\right) \int_0^1 e^{-s\,y}\, dy = \left(1 + \frac{\rho - 1}{n}\right) \frac{1 - e^{-s}}{s}.$$

Thus, taking the inverse Laplace transforms $\mathcal{L}^{-1}$ of I_1 and I_2 yields

$$\mathcal{L}^{-1}(I_1) \sim \frac{1}{n} + \frac{\rho - 1}{n^2} \qquad (n > t) \tag{4.5}$$

and

$$\begin{aligned} \mathcal{L}^{-1}(I_2) &\sim -\frac{\rho}{n^2} \frac{1}{2\pi i} \int_{\mathcal{B}} e^{s\,t/n} \left[\int_0^1 \frac{e^{-s\,y}}{(1-y)^{1-\rho s/n}}\, dy \right] ds \\ &\sim -\frac{\rho}{n^2} \int_0^1 \frac{\delta(t/n - y)}{1 - y}\, dy \\ &= -\frac{\rho}{n\,(n-t)} \qquad (n > t). \end{aligned} \tag{4.6}$$

Here we also used the fact that

$$\begin{aligned} \mathcal{L}^{-1}\left[\frac{1 - e^{-s}}{s}\right]\left(\frac{t}{n}\right) &= \frac{1}{n}\left[1\left\{\frac{t}{n} > 0\right\} - 1\left\{\frac{t}{n} > 1\right\}\right] \\ &= \frac{1}{n}\, 1\{0 < t < n\}, \end{aligned}$$

where $1\{\cdot\}$ is an indicator function. Then inverting (4.4) with (4.5) and (4.6) leads to (2.15).

This analysis suggests that $p_n(t)$ is approximately zero in the range $t/n > 1$. We shall show that in this sector the density is exponentially small. Before doing this, we first investigate the transition region, where $t \approx n$.

Thus, we consider $n, t \to \infty$ with $n/t = 1 + \Delta\, t^{-1/2} = 1 + O(t^{-1/2})$. We can still use (4.3) but now scale $l = y\sqrt{n} = O(\sqrt{n})$, and approximate the sum by

$$\widehat{p}_n(\theta) \sim \frac{1}{(1+\theta)^{n+1}\,(\sqrt{n})^{r+1}} \int_0^\infty (1+\theta)^{\sqrt{n}\,y}\, y^r\, dy.$$

Scaling $\theta = \varpi/\sqrt{n} = O(1/\sqrt{n})$, and noting that $(1+\theta)^{-n-1} \sim e^{\varpi^2/2-\sqrt{n}\,\varpi}$ and $(1+\theta)^{\sqrt{n}\,y} \sim e^{\varpi\, y}$, the inverse Laplace transform leads to

$$p_n(t) \sim \frac{1}{n} \int_0^\infty \frac{1}{2\pi i} \int_{\mathcal{B}} e^{\varpi^2/2}\, e^{(-\sqrt{n}+y+t/\sqrt{n})\,\varpi}\, d\varpi\, dy. \tag{4.7}$$

Note that in this range of (n, t),

$$\frac{t}{\sqrt{n}} - \sqrt{n} = \left(1 - \frac{n}{t}\right) \frac{t}{\sqrt{n}} = -\Delta \frac{\sqrt{t}}{\sqrt{n}} \sim -\Delta = O(1).$$

Then by using the identity

$$\frac{1}{2\pi i} \int_{\mathcal{B}} e^{C_0 \varpi^2 + C_1 \varpi}\, d\varpi = \frac{1}{2\sqrt{\pi\, C_0}} \exp\left(-\frac{C_1^2}{4C_0}\right)$$

and noting that $\Delta = (n-t)/\sqrt{t}$, we explicitly evaluate the integral over ϖ in (4.7) and obtain (2.16).

Now we consider $n,\ t \to \infty$ with $t > n$. We rewrite (2.11) as

$$\widehat{p}_n(\theta) = M\, G_n \sum_{l=0}^{\infty} \frac{\rho^l}{l!}\, H_l \ + \ M \sum_{l=n+1}^{\infty} \frac{\rho^l}{l!}\, (H_n\, G_l - H_l\, G_n). \tag{4.8}$$

The first sum can be calculated exactly by using (2.14), which yields

$$\sum_{l=0}^{\infty} \frac{\rho^l}{l!}\, H_l = e^r\, \Gamma(r+1) \left(-\frac{1+\theta}{\rho\, \theta}\right)^{r+1}. \tag{4.9}$$

The result in (4.9) holds for $\theta < 0$ and $\theta > \theta_p = -1 + [-\rho + \sqrt{\rho^2 + 4\rho}]/2$, since $\Gamma(r(\theta)+1)$ has a simple pole at $\theta = \theta_p$. We note that $\theta_p = \nu_1$, which is the first eigenvalue in (2.6). The second sum in (4.8) is negligible in view of (3.10), (3.11) and (4.9), and the fact that $\theta < 0$. Using (2.12), (3.10) and (4.9) in the first sum of (4.8), then taking the inverse Laplace transform, we have

$$p_n(t) \sim \frac{1}{2\pi i} \int_{\mathcal{B}} h(\theta)\, e^{t\, f(\theta)}\, d\theta, \tag{4.10}$$

where $f(\theta) = \theta - \log(1+\theta)\, n/t$ and

$$h(\theta) = \frac{\Gamma(r+1)\, e^r}{(1+\theta)\, (-\theta)^{r+1}\, n^{r+1}}.$$

For $t \to \infty$ and n/t fixed we evaluate (4.10) by the saddle point method. There is a saddle point at $\theta = \theta_s \equiv n/t - 1 < 0$, which satisfies $f'(\theta) = 0$. Hence, using the saddle point method gives

$$p_n(t) \sim \frac{h(\theta_s)}{\sqrt{2\pi t f''(\theta_s)}}\, e^{t\, f(\theta_s)}, \quad n, t \to \infty$$

and this leads to (2.17), where $r_* = r(\theta_s)$. This analysis indicates that (2.17) only holds for $n,\ t \to \infty$ with $n/t < 1$ and $\theta_s > \theta_p$, so that $n/t = 1 + \theta_s > 1 + \theta_p = \Lambda_0$.

There is a transition region where $n/t = \Lambda_0 + \Lambda/\sqrt{t}$ with $\Lambda = O(1)$. We still use (4.10) and note that $h(\theta)$ has a simple pole at $\theta = \theta_p$, and the saddle point θ_s of $f(\theta)$ is now close to θ_p. We expand the integrand in (4.10) about $\theta = \theta_p$ using

$$h(\theta) \sim \frac{\sqrt{\rho+4} - \sqrt{\rho}}{2\sqrt{\rho+4}}\, e^{-1}\, \frac{1}{\theta - \theta_p}$$

and $f(\theta) = f(\theta_p) + f'(\theta_p)(\theta - \theta_p) + \frac{1}{2} f''(\theta_p)(\theta - \theta_p)^2 + \cdots$. Using $1 + \theta_p = \Lambda_0$ and $n/t \sim \Lambda_0$, and scaling $\theta - \theta_p = S/\sqrt{t}$, (4.10) asymptotically becomes

$$p_n(t) \sim \Lambda_0^{-n}\, e^{-t + \Lambda_0 t}\, \frac{1}{2\pi i} \int_{\mathcal{B}} \frac{1}{S}\, \exp\left[-\frac{\Lambda}{\Lambda_0}\, S + \frac{1}{2\Lambda_0}\, S^2 \right] dS,$$

where $\Re(S) > 0$ on the contour $\mathcal{B}$. Then we use the identity

$$\frac{1}{2\pi i} \int_{\mathcal{B}} \frac{1}{S}\, e^{-A\, S + \alpha\, S^2/2} dS = \frac{1}{\sqrt{2\pi}} \int_{A/\sqrt{\alpha}}^{\infty} e^{-u^2/2} du,$$

with $A = \Lambda/\Lambda_0$ and $\alpha = \Lambda_0^{-1}$, to eventually obtain (2.18).

Finally, for the scale $n,\ t \to \infty$ with $n/t < \Lambda_0$, the pole at $\theta = \theta_p$ dominates the asymptotic behavior of $p_n(t)$, and (2.19) is obtained by evaluating the residue at the dominant pole in (4.10). This concludes the derivation of Theorem 2.3.

4.2 Proof of Theorem 2.4

For the time scale $t = O(\rho^{-1/2})$, from the spectral representation we note that as $\rho \to 0$ the eigenvalues are

$$\nu_m = -1 + \frac{\sqrt{\rho}}{\sqrt{m}} + O(\rho), \quad \widetilde{\nu}_m = -1 - \frac{\sqrt{\rho}}{\sqrt{m}} + O(\rho).$$

Then the eigenfunctions are asymptotically given by

$$\phi_m(n, \nu_m) \sim (-1)^n\, n!\, m^{-n/2}\, \rho^{-n/2}\, L_n^{(m-1-n)}(m)$$

and $\phi_m(n, \widetilde{\nu}_m) \sim (-1)^n \phi_m(n, \nu_m)$. Thus, all the eigenvalues contribute to $p_n(t)$ on the scale $t = O(\rho^{-1/2})$ and $n = O(1)$, and we obtain (2.20).

Now we consider the scale $n,\ t = O(1)$ with $\rho \to 0$ and use the result in (2.11). Since $r = O(\rho)$, G_n in (2.13) becomes

$$G_n \sim \int_0^{\frac{1}{1+\theta}} z^n \, dz = \frac{1}{(n+1)\,(1+\theta)^{n+1}}, \quad \rho \to 0. \tag{4.11}$$

Using (4.11), (3.11) with $r \to 0$ and $M \sim \rho/(1+\theta)$ in (2.11), we find that the first sum dominates the second and we obtain

$$\widehat{p}_n(\theta) \sim M\, G_n \sum_{l=0}^{n} \frac{\rho^l}{l!}\, H_l \sim \frac{1}{n+1}\,\frac{1}{\theta}\left[1 - \frac{1}{(1+\theta)^{n+1}}\right], \quad \rho \to 0.$$

Then we invert the Laplace transform over time, and obtain explicitly the leading term $p_n^{(0)}(t)$ in (2.22). The derivation of the correction term in (2.22) is similar and we omit it.

5 Asymptotic results for $\rho \to \infty$

We shall use a singular perturbation approach to derive the asymptotic approximations for large traffic intensities, $\rho \to \infty$. This method should be useful for analyzing models with more general balking probabilities. We have verified that the same results can be obtained by asymptotically expanding the exact representations in Theorems 2.1 and 2.2. In addition to being applicable to problems that cannot be explicitly solved, the perturbation method leads to a quicker derivation of the asymptotics.

We first consider the scale $t = T\rho = O(\rho)$ and $n = N\rho = O(\rho)$, and expand $p_n(t)$ in powers of ρ^{-1}, as in (2.23). Using (2.23) in the recurrence equation (2.3), the leading term $P_0(N, T)$ satisfies

$$\frac{\partial P_0}{\partial T} + \frac{N-1}{N}\frac{\partial P_0}{\partial N} = -\frac{1}{N}\, P_0 \tag{5.1}$$

with the initial condition $P_0(N, T) = 1/N$. We solve this first order PDE by the method of characteristics. The family of characteristics is given by

$$T = N + \log|1 - N| + \text{constant},$$

where the constant indexes the family. The characteristic $T = N + \log(1 - N)$ goes through the origin $(N, T) = (0, 0)$, along the parabola $T = -N^2/2$. The general solution to (5.1) is

$$P_0(N, T) = \frac{1}{N-1}\mathcal{F}\left((N-1)\, e^{N-T}\right). \tag{5.2}$$

Using the initial condition in (5.2), we determine the function $\mathcal{F}(\cdot)$ from

$$\mathcal{F}\left((N-1)\,e^{N}\right) = \frac{N-1}{N}.$$

If we denote by $N_* = N_*(N,T)$ the solution to

$$(N_*-1)\,e^{N_*} = (N-1)\,e^{N-T}, \tag{5.3}$$

P_0 in (5.2) becomes

$$P_0(N,T) = \frac{1}{N-1}\,\frac{N_*-1}{N_*}. \tag{5.4}$$

Setting $N_* = N-U$ in (5.3) leads to (2.25). Thus, (5.4) can be rewritten as (2.24).

Alternately, we can rewrite (2.24) more explicitly, in terms of an infinite series. From (2.25), we let $U = N-1+U_0$, where $U_0 = U_0(N,T) = (1-N)\,e^{U-T}$. Then U_0 can be expressed in terms of the Lambert W-function (see Corless et al. 1996), which satisfies

$$e^{-U_0}\,U_0 = (1-N)\,e^{N-T-1} \equiv z.$$

We use the series expansion of the Lambert W-function to obtain U_0 as

$$U_0 = \sum_{m=1}^{\infty} \frac{(-m)^{m-1}}{m!}\,z^m,$$

where the series converges for $|z| < e^{-1}$. Thus, U has the following series expansion

$$U(N,T) = N-1+\sum_{m=1}^{\infty} \frac{m^{m-1}}{m!}\,(1-N)^m\,e^{m(N-T-1)}, \tag{5.5}$$

which converges for $|1-N|\,e^{N-T} < 1$. The series is always convergent for $N \le 1$, but diverges for $N > 1$, if $T < N + \log(N-1)$. For example, if $T = 0$ the series converges only for $N < N_c \doteq 1.2784$, where $(N_c-1)\,e^{N_c} = 1$. Using (5.5) in (2.24), we have an alternate series expression for $P_0(N,T)$:

$$P_0(N,T) = \frac{\sum_{m=1}^{\infty} m^{m-1}\,(1-N)^{m-1}\,e^{m(N-T-1)}/m!}{1-\sum_{m=1}^{\infty} m^{m-1}\,(1-N)^{m}\,e^{m(N-T-1)}/m!}. \tag{5.6}$$

By further expanding (2.3) on the $n = O(\rho)$, $t = O(\rho)$ scale we find that the correction term $P_1(N,T)$ satisfies the following PDE

$$\frac{\partial P_1}{\partial T} = \frac{1-N}{N}\,\frac{\partial P_1}{\partial N} - \frac{1}{N}\,P_1 + \frac{N+1}{2N}\,\frac{\partial^2 P_0}{\partial N^2} + \frac{N-1}{N^2}\,\frac{\partial P_0}{\partial N} + \frac{1}{N^2}\,P_0,$$

with the initial condition $P_1(N,0) = -1/N^2$. This follows from expanding $p_n(0) = 1/(n+1) = 1/(N\rho+1)$ in powers of ρ^{-1}.

After a lengthy but routine calculation we find that

$$P_1(N,T) = -\frac{(\xi-1)(2\xi-3)}{2\xi^5} + \frac{3(\xi-1)(2\xi^2-2\xi-3)}{2\xi^5(\xi+\eta-1)} - \frac{(\xi-1)^2(2\xi^3+2\xi^2-5\xi-15)}{2\xi^5(\xi+\eta-1)^2} - \frac{(\xi-1)^3(2\xi^2+4\xi+3)}{2\xi^5(\xi+\eta-1)^3} - \frac{2(\xi-1)(2\xi-3)}{\xi^5(\xi+\eta-1)} \log\left|\frac{\xi+\eta-1}{\xi-1}\right|, \tag{5.7}$$

where $\xi = N - U,\ \eta = U$.

Now we consider some special cases. If $N = 1$, then $U \to 0$ by (2.25). Thus, $\xi \to 1$, $\eta \to 0$ and

$$\frac{\xi-1}{\xi+\eta-1} \to e^{-T}.$$

Then (5.7) reduces to the explicit result

$$P_1(1,T) = \left(2T - \frac{9}{2}\right) e^{-T} + 8\,e^{-2T} - \frac{9}{2}\,e^{-3T}.$$

From (5.6) it follows that $P_0(1,T) = e^{-T}$.

If $N = 0$ and $T \to \infty$, then $U \to -1$, $\eta \to -1$, and $\xi \sim 1 - e^{-1}\,e^{-T}$. Hence, from (5.7), $P_0(0,T) \sim e^{-1}\,e^{-T}$ and $P_1(0,T) \sim (2T-3)\,e^{-1}\,e^{-T}$. From (2.5) we obtain for $n = 0$, $t = T\rho \to \infty$

$$C_1(\nu_1)\,\phi_1(0,\nu_1)\,e^{\nu_1 t} \sim \rho^{-1}\,e^{-1}\,e^{-T} + \rho^{-2}\,(2T-3)\,e^{-1}\,e^{-T}.$$

This agrees with (5.7) and shows that for $N = 0$ and $T \gg 1$ ($t \gg \rho$) only the first eigenvalue ν_1 contributes to the expansion of $p_n(t)$.

Next, we consider short time scales, with $n = O(1)$ and $t = \tau/\rho = O(\rho^{-1})$. Expanding the conditional sojourn time density $p_n(t)$ in the form $p_n(t) = \mathcal{Q}_n(\tau) + O(\rho^{-1})$ and using equation (2.3), we have

$$\mathcal{Q}_n'(\tau) = \frac{1}{n+1}\left[\mathcal{Q}_{n+1}(\tau) - \mathcal{Q}_n(\tau)\right],$$

with the initial condition $\mathcal{Q}_n(0) = 1/(n+1)$. Taking the Laplace transform over the time variable τ with $\widehat{\mathcal{Q}}_n(s) = \int_0^\infty \mathcal{Q}_n(\tau)\,e^{-\tau s}\,d\tau$, we obtain the following difference equation for $\widehat{\mathcal{Q}}_n(s)$:

$$\widehat{\mathcal{Q}}_{n+1}(s) - [(n+1)\,s + 1]\ \widehat{\mathcal{Q}}_n(s) = -1. \tag{5.8}$$

Solving (5.8) and inverting the Laplace transform, we have

$$\mathcal{Q}_n(\tau) = \frac{1}{2\pi i} \int_{\mathcal{B}} \frac{e^{\tau s}}{s} \sum_{j=0}^{\infty} s^{-j} \frac{\Gamma(n+1+1/s)}{\Gamma(n+j+2+1/s)}\, ds. \tag{5.9}$$

Using the identity

$$\int_0^1 t^{x-1} (1-t)^{y-1}\, dt = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)}, \quad x,\ y > 0,$$

we can rewrite (5.9) as

$$\mathcal{Q}_n(\tau) = \int_0^1 (1-z)^n \frac{1}{2\pi i} \int_{\mathcal{B}} e^{\tau s} \frac{e^{z/s}\,(1-z)^{1/s}}{s}\, ds\, dz.$$

Then by the inverse Laplace transform

$$\mathcal{L}^{-1}\left(\frac{1}{s} e^{W/s}\right) = J_0(2\sqrt{\tau\,|W|}),$$

where $W = z + \log(1-z) < 0$ and $J_0(\cdot)$ is the Bessel function, we obtain (2.26). By expanding $\mathcal{Q}_n(\tau)$ for $n \to \infty$ and $\tau \to \infty$, with $\tau = O(n^2)$ we obtain

$$\mathcal{Q}_n(\tau) \sim \int_0^{\infty} e^{-nz} J_0\left(z\sqrt{2\tau}\right) dz = \frac{1}{\sqrt{n^2+2\tau}} = \frac{1}{\rho\sqrt{N^2+2T}}. \tag{5.10}$$

Then we can easily show that $\rho^{-1} P_0(N,T)$, when expanded for $(N,T) \to (0,0)$, gives the same result as in (5.10), which verifies the matching between the long time (T-scale) and short time (τ-scale) results.

We note that by expanding (2.5)–(2.9) for $n = N\rho \to \infty$ and $t = T\rho \to \infty$ we find that $e^{\nu_m t} \sim e^{-mT}$ while $e^{\widetilde{\nu}_m t}$ is exponentially small as $\rho \to \infty$. Then from (2.5), on the (N,T) scale, we obtain

$$p_n(t) \sim \rho^{-1} \sum_{m=1}^{\infty} e^{m(N-1)} (1-N)^{m-1} \frac{m^m}{m!} e^{-mT}.$$

We can show that this series is equivalent to $P_0(N,T)$ in (5.6).

6 Discussion

To summarize, we have obtained both exact and asymptotic results for the $M/M/1$-PS model with non-balking probability $b_n = 1/(n+1)$. We compare our results to the standard model, where $b_n = 1$. First, the spectral representation of $p_n(t)$ for the two models is very different as the standard model has a purely continuous spectrum (see also Guillemin and Boyer 2001) while the balking model has a purely discrete one.

We recently studied (see Zhen and Knessl 2010) $p_n(t)$ for the standard model asymptotically, and found that if $\rho = \lambda/\mu < 1$ and $n,\ t \to \infty$ the asymptotic expansion is different according as $n/t > 1-\rho$, $n/t \approx 1-\rho$, $0 < n/t < 1-\rho$, $n = O(t^{2/3})$, and $n = O(1)$. The scale $n = O(t^{2/3})$ is important in obtaining the tail of the unconditional density, which for the standard PS model has the form (see Pollaczek 1946 and Cohen 1984)

$$p_{PS}(t) \sim \alpha_2\, t^{-5/6}\, e^{-\alpha_0 t}\, e^{-\alpha_1 t^{1/3}},$$

where $\alpha_0 = (1-\sqrt{\rho})^2$ and α_1 and α_2 are constants. In contrast, for the model with balking Theorem 2.3 shows that the structure of $p_n(t)$ is different in three main sectors of the (n,t) plane ($n/t > 1$, $\Lambda_0 < n/t < 1$ and $0 < n/t < \Lambda_0$), with two transition regions connecting them. For $t \to \infty$ with $0 \le n/t < \Lambda_0$ the asymptotics of $p_n(t)$ are governed by the eigenvalue with the largest real part and we obtain the purely exponential behavior in (2.19), which leads to (2.28) for the unconditional density $p_{PS}(t)$. Thus for the model with balking the scale $n = O(t^{2/3})$ is absent.

If $\rho > 1$ the standard PS model has an algebraic tail, with $p_{PS}(t) \sim \alpha_3\, t^{-\rho/(\rho-1)}$, so that the mean sojourn time is finite for $\rho < 2$, the second moment is finite for $\rho < 3/2$, etc. Then the approximation

$$p_n(t) \sim \frac{1}{n}\left[1 + (\rho-1)\,\frac{t}{n}\right]^{-\frac{\rho}{\rho-1}}$$

applies for n and/or $t \to \infty$. This situation is similar to the model with balking in the limit $\rho \to \infty$. Here the tail will be purely exponential, but for n and/or $t \to \infty$ we have the approximation in (2.23), which is quite unlike the three sectors in Theorem 2.3.

Acknowledgments Knessl was partly supported by NSF grant DMS 05-03745 and NSA grant H 98230-08-1-0102. Van Leeuwaarden was supported by a VENI grant from The Netherlands Organization for Scientific Research (NWO).

References

Borst SC, Boxma OJ, Morrison JA, Núñez-Queija R (2003) The equivalence between processor sharing and service in random order. Oper Res Lett 31:254–262

Coffman EG Jr, Muntz RR, Trotter H (1970) Waiting time distributions for processor-sharing systems. J ACM 17:123–130
Cohen JW (1984) On processor sharing and random service (Letter to the editor). J Appl Prob 21:937–937
Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE (1996) On the Lambert *W* function. Adv Comput Math 5:329–359
Flatto L (1997) The waiting time distribution for the random order service M/M/1 queue. Ann Appl Prob 7:382–409
Gromoll HC, Robert P, Zwart B (2008) Fluid limits for processor-sharing queues with impatience. Math Oper Res 33:375–402
Guillemin F, Boyer J (2001) Analysis of the M/M/1 queue with processor sharing via spectral theory. Queueing Syst 39:377–397
Haight FA (1957) Queueing with balking. Biometrika 44:360–369
Kleinrock L (1964) Analysis of a time-shared processor. Nav Res Logist Q 11:59–73
Knessl C (1993) On the sojourn time distribution in a finite capacity processor shared queue. J Assoc Comput Mach 40:1238–1301
Magnus W, Oberhettinger F, Soni RP (1966) Formulas and Theorems for the Special Functions of Mathematical Physics. Springer, New York
Morrison JA (1985) Response-time distribution for a processor-sharing system. SIAM J Appl Math 45:152–167
Morse PM (1958) Queues, Inventories and Maintenance. Wiley, New York
Pollaczek F (1946) La loi d'attente des appels téléphoniques. C R Acad Sci Paris 222:353–355
Riordan J (1962) Stochastic Service Systems. Wiley, New York
Vaulot E (1946) Délais d'attente des appels téléphoniques traités au hasard. C R Acad Sci Paris 222:268–269
Zhen Q, Knessl C (2010) On sojourn times in the M/M/1-PS model, conditioned on the number of other users. Appl Math Res Express AMRX 2009:142–167
Zhen Q, Knessl C (2007) Asymptotic expansions for the conditional sojourn time distribution in the M/M/1-PS queue. Queueing Syst 57:157–168