Queueing Syst (2009) 63: 355–381
DOI 10.1007/s11134-009-9142-9

Exact tail asymptotics in a priority queue—characterizations of the preemptive model

Hui Li · Yiqiang Q. Zhao

Received: 1 December 2008 / Revised: 2 September 2009 / Published online: 1 October 2009

Abstract In this paper, we consider the classical preemptive priority queueing system with two classes of independent Poisson customers and a single exponential server serving the two classes of customers at possibly different rates. For this system, we carry out a detailed analysis on exact tail asymptotics for the joint stationary distribution of the queue length of the two classes of customers, for the two marginal distributions and for the distribution of the total number of customers in the system, respectively. A complete characterization of the regions of system parameters for exact tail asymptotics is obtained through analysis of generating functions. This characterization has never before been completed. It is interesting to note that the exact tail asymptotics along the high-priority queue direction is of a new form that does not fall within the three types of exact tail asymptotics characterized by various methods for this type of two-dimensional system reported in the literature. We expect that the method employed in this paper can also be applied to the exact tail asymptotic analysis for the non-preemptive priority queueing model, among other possibilities.

Keywords Exact tail asymptotics · Light tail · Geometric decay · Decay rate · Double QBD process · Preemptive priority queue · Generating functions

Mathematics Subject Classification (2000) 60K25 · 30E15 · 60J10

1 Introduction

Priority queueing systems are important due to their broad range of applications, such as in computer and telecommunications networks. The classical priority queue-

H. Li
Department of Mathematics, Mount Saint Vincent University, Halifax, NS B3M 2J6, Canada

Y.Q. Zhao (✉)
School of Mathematics and Statistics, Carleton University, Ottawa, ON K1S 5B6, Canada
e-mail: zhao@math.carleton.ca

ing system with one server and two classes of customers, preemptive or non-preemptive, has been considered by several researchers. One often referenced article is Miller [37]. In [37], a special structure on the rate matrix, which is a key probabilistic quantity in the analysis using the matrix analytic method, was explored and an efficient computational scheme was suggested based on this structure. Earlier work by Miller and other researchers has been since extended to priority models of more general types, in both continuous and discrete time; for example, Gail, Hantler and Taylor [20, 21], Kao and Narayanan [27], Takine [48], Alfa [2], Isotupa and Stanford [26], Alfa, Liu and He [3], Sleptchenko, Adan and van Houtum [46], Drekic and Woolford [9], and Zhao et al. [53], among many others.

Our purpose in this paper is to characterize tail asymptotics for the classical preemptive priority queueing model with two classes of customers. Specifically, we are interested in the exact tail asymptotics in the joint distribution along the queue length direction of low- and high-priority customers, respectively. We also characterize exact tail asymptotics in the marginal distribution of the queue length of low-priority customers and in the distribution for the total number of customers in the system, respectively. Compared to studies on other aspects of priority queues, there are not many references on tail asymptotic results. Tail asymptotic results for the classical priority model studied here have been reported, but the characterization is not complete. Closely related to this study, a sufficient condition for an exact geometric decay along the low-priority queue direction was reported in Haque [22] and Haque, Liu and Zhao [23] for the same model considered in this paper but with the same service rate for the two servers. In Miyazawa and Zhao [43], a discrete time preemptive priority model in which batch arrivals are allowed for one class of customers was considered, and a region in which the joint distribution has an exact geometric decay along the low-priority queue direction was described. Its special case when batch arrivals are not allowed is equivalent to the model studied in this paper, but again with the same service rate for the two servers. Tail asymptotic results on more general priority queueing models are also available, including Abate and Whitt [1], in which light-tailed, including exact geometric and two types of non-exact-geometric, asymptotics in the waiting time distribution for a low-priority customer were obtained for the continuous time $M/G/1$ priority queueing model. These three types of light-tailed asymptotics are also presented in a discrete time $M/G/1$ model as reported in Maertens, Walraevens and Bruneel [32]. Xue and Alfa [52] obtained exact tail asymptotics in the marginal distribution for the low-priority queue for the $BMAP/PH/1$ priority queue in discrete time by allowing different service rates.

Exact and logarithmic tail asymptotics for a marginal performance metric were also considered for priority queues with a more general type of arrivals, say a Gaussian process, under many-source assumptions. For example, in Delas, Mazumdar and Rosenberg [8], bounds on exact asymptotics were calculated for a two-queue priority model under certain technical assumptions. Bounds on logarithmic asymptotics were obtained in Wischik [50]. More information can be found in van Uitert [49] and Mandjes [35].

A few methods are available for the analysis of exact tail asymptotics, all of which have been proven to have advantages in some aspects of the analysis. The generating function method is a classical method with the advantage of explicitly determining the constant besides the exact decay function. This method is also

valid in principle for the double quasi-birth-and-death (QBD) process, or the random walks in the quarter plane, in which the generating function can be formally expressed based on the theory of Riemann surfaces, algebraic ideas and Riemann–Hilbert boundary value problems. References include Malyshev [33, 34], Flatto and McKean [15], Fayolle and Iasnogorodski [10], Fayolle, King and Mitrani [11], Cohen and Boxma [7], Flatto and Hahn [16], Flatto [14], Fayolle, Iasnogorodski and Malyshev [12], Wright [51], Kurkova and Suhov [29] and Morrison [44]. Theorems based on large deviations can often be presented for more general processes, which is one of the advantages of this approach. The method developed in Borovkov and Mogul'skii [5] focused on tail asymptotics along an arbitrary direction, but not including the two coordinate directions, while the method presented by McDonald [36] and Foley and McDonald [17–19] is applied to the coordinate directions. The matrix-analytic method is an efficient method for exact geometric decay along a coordinate, referred to as the level, direction and also in the marginal distribution of the level for two-dimensional block-structured transition matrices such as the $GI/G/1$ type and its special cases. It also provides conditions on non-exact-geometric tail asymptotics. References on this method include Takahashi, Fujimoto and Makimoto [47], Haque [22], Miyazawa [39], Miyazawa and Zhao [43], Kroese, Scheinhardt and Taylor [28], Haque, Liu and Zhao [23], Motyer and Taylor [45], Li, Miyazawa and Zhao [30], and He, Li and Zhao [24], among others. Recently, Miyazawa [40–42] converted the asymptotic problem in the double QBD process into a non-linear optimization description and confirmed that there are only three types of exact tail asymptotics along a coordinate direction: exact geometric, geometric multiplied by a power function with power $-1/2$ or $-3/2$. Methods based on the analysis of combinatorics, for example, Bousquet-Melou [6], Mishna [38] and Hou and Mansour [25], are also a candidate for exact tail asymptotics in the quarter plane.

Our research extends the previous study to allow asymmetric service rates. The main contributions made in this paper include: (1) an explicit determination of exact light-tailed asymptotics in the joint distribution along the high-priority queue, which is a type of exact tail asymptotic different from the three types reported in the literature for the double QBD process along a coordinate direction, and therefore different from the behaviour in the high-priority marginal distribution; (2) the identification of the regions for the three different types of exact light-tailed asymptotics in the joint distribution along the low-priority queue and in the marginal distribution for the low-priority queue; and (3) a complete characterization of regions for the asymptotic for the total number of customers in the system, in which two additional types, different from the three types mentioned in (2), of tail asymptotics are revealed. To achieve our goal, the generating function approach is employed here since the generating functions involved can be expressed explicitly. It also works well for the reducible case, or for the high-priority queue direction where the irreducible condition required by other methods is not present.

The rest of the paper is organized as follows. Section 2 provides the model description and obtains expressions of the generating functions. The main asymptotic results are stated in Sect. 3. A singularity analysis based on a Tauberian theorem is carried out in Sect. 4. Sections 5, 6 and 7 detail proofs of the main results. Concluding remarks are made in the final section.

2 Model description and generating functions

In this section, we first recall the $M/M/1$ preemptive priority queue with two classes of customers and then obtain expressions for the generating functions of interest in this paper.

In the $M/M/1$ preemptive priority queueing system with two classes of customers, the high- and low-priority classes of customers arrive independently according to two Poisson processes with rates λ_h and λ_ℓ, respectively. High-priority customers have priority in service over low-priority customers. In each of these two classes, customers are served according to the first in, first out (FIFO) queueing discipline. With the preemptive rule, the service of a low-priority customer is interrupted upon the arrival of a high-priority customer that stays at the head of the line waiting for the restart of its service immediately after the last high-priority customer in the system completes its service. Both classes of customers require an exponential amount of service time and are served by a single server with possibly different service rates μ_h and μ_ℓ for high- and low-priority customers, respectively. All service times are independent and also independent of the two arrival processes. If we use the number of high- and low-priority customers, respectively, in the system as two system variables, then we have a continuous time Markov chain $X(t) = \{(X_h(t), X_\ell(t)) : t \geq 0\}$. Throughout the paper, let $\lambda = \lambda_h + \lambda_\ell$ and assume that $\rho = \rho_h + \rho_\ell < 1$ for the stability of the system (Markov chain), where $\rho_h = \frac{\lambda_h}{\mu_h}$ and $\rho_\ell = \frac{\lambda_\ell}{\mu_\ell}$. Under this condition, we denote by $\pi_{i,j}$ the joint stationary probability distribution for the number of high- and low-priority customers, respectively, in the system. The marginal distributions for the high- and low-priority customers are denoted by $\pi_i^{(h)}$ and $\pi_j^{(\ell)}$, respectively. Without loss of generality, throughout the paper we assume that $\lambda_h + \lambda_\ell + \mu_h + \mu_\ell = 1$. We also use the following convention: for two functions $f(n)$ and $g(n)$ of nonnegative integers n, $f(n) \sim g(n)$ means that $\lim_{n\to\infty} f(n)/g(n) = 1$. We also define $\bar{\rho}_h = \frac{\lambda}{\mu_h}$.

The system of balance equations for the preemptive priority queueing system considered in this paper is given by

$$(\lambda_h + \lambda_\ell)\pi_{0,0} = \mu_h \pi_{1,0} + \mu_\ell \pi_{0,1}, \tag{2.1}$$

$$(\lambda_h + \lambda_\ell + \mu_\ell)\pi_{0,j} = \mu_h \pi_{1,j} + \mu_\ell \pi_{0,j+1} + \lambda_\ell \pi_{0,j-1}, \quad j \geq 1, \tag{2.2}$$

$$(\lambda_h + \lambda_\ell + \mu_h)\pi_{i,0} = \mu_h \pi_{i+1,0} + \lambda_h \pi_{i-1,0}, \quad i \geq 1, \tag{2.3}$$

$$(\lambda_h + \lambda_\ell + \mu_h)\pi_{i,j} = \mu_h \pi_{i+1,j} + \lambda_h \pi_{i-1,j} + \lambda_\ell \pi_{i,j-1}, \quad i \geq 1, j \geq 1. \tag{2.4}$$

Consider the discrete time Markov chain obtained through uniformization of the continuous time Markov chain $X(t)$. The above two Markov chains have the same stationary probability distribution $\pi_{i,j}$, and (2.1)–(2.4) are also the stationary equations of the discrete time Markov chain.

The above priority queueing system is a specific case of random walks in the quarter plane studied in Fayolle, Iasnogorodski and Malyshev [12]. For this random

walk, define the following generating functions for the stationary probability vector $\pi_{i,j}$:

$$\varphi_j(x) = \sum_{i=0}^{\infty} \pi_{i,j} x^i, \quad j \geq 0,$$

$$\psi_i(y) = \sum_{j=0}^{\infty} \pi_{i,j} y^j, \quad i \geq 0,$$

$$P(x,y) = \sum_{j=0}^{\infty} \varphi_j(x) y^j = \sum_{i=0}^{\infty} \psi_i(y) x^i.$$

It is clear that $\varphi_0(x) = P(x,0)$ and $\psi_0(y) = P(y,0)$.

Equations (1.3.5) and (1.3.6) of Fayolle, Iasnogorodski and Malyshev [12] provide an expression of the generating function for joint probabilities $\pi_{i,j}$ for $i, j \geq 1$ (excluding $i = 0$ or $j = 0$), in terms of the boundary generating functions. Using a similar argument, we can obtain the following expression for the generating function $P(x,y)$ for the priority model.

Proposition 2.1 *For* $|x| \leq 1$ *and* $|y| \leq 1$,

$$\begin{aligned} &y\big[(\lambda + \mu_h - \lambda_\ell y)x - \mu_h - \lambda_h x^2\big] P(x,y) \\ &\quad = \big[(\mu_h - \mu_\ell)xy - \mu_h y + \mu_\ell x\big]\psi_0(y) + \pi_{0,0}\mu_\ell(y-1)x; \end{aligned} \tag{2.5}$$

$$\begin{aligned} &P(x,1) = \frac{1-\rho_h}{1-\rho_h x}; \qquad P(1,y) = \frac{\mu_\ell}{\lambda_\ell y}\big[\psi_0(y) - (1-\rho)\big]; \\ &P(x,x) = \frac{\mu_\ell(1-\rho)}{\mu_h[1-(\lambda/\mu_h)x]} + \frac{(\mu_h - \mu_\ell)}{\mu_h[1-(\lambda/\mu_h)x]}\psi_0(x); \end{aligned} \tag{2.6}$$

and $\pi_{0,0} = 1 - \rho$.

Proof Equation (2.5) follows from a similar proof to (1.3.5) and (1.3.6) of Fayolle, Iasnogorodski and Malyshev [12]. Elementary manipulations lead to expressions evaluated at various special values. Finally, let $y = \frac{\mu_\ell x}{\mu_h - (\mu_h - \mu_\ell)x}$ in (2.5) such that the coefficient of $\psi_0(y)$ is zero. Then, let $x \to 1$, which implies $y \to 1$, and we obtain the evaluation of $\pi_{0,0}$. □

Remark 2.1 In order to find an explicit expression for the generating function $P(x,y)$, we need to determine $\psi_0(y)$. A general approach is to consider zeros defined by the kernel equation $K(x,y) = 0$, through which we obtain an expression for $\psi_0(y)$. For the priority queue, the kernel $K(x,y) = yK_0(x,y)$ is not irreducible (as a polynomial of x and y), where $K_0(x,y) = (\lambda + \mu_h - \lambda_\ell y)x - \mu_h - \lambda_h x^2$ is called the key kernel. Since the kernel is reducible, (2.5) does not immediately provide a determination of $\varphi_0(x) = P(x,0)$, which will be obtained directly based on balance equations.

In the following, we provide a basic property (Lemma 2.1) of the zeros of the key kernel $K_0(x, y)$, after which we first determine $\varphi_j(x)$ (based on balance equations) and then $\psi_0(y)$.

For each fixed y, consider $K_0(x, y)$ as a polynomial of x, which has two zeros given by

$$x_1(y) = \frac{(\lambda + \mu_h - \lambda_\ell y) - \sqrt{\Delta(y)}}{2\lambda_h}, \qquad x_2(y) = \frac{(\lambda + \mu_h - \lambda_\ell y) + \sqrt{\Delta(y)}}{2\lambda_h}, \tag{2.7}$$

where $\Delta(y) = (\lambda + \mu_h - \lambda_\ell y)^2 - 4\lambda_h\mu_h = \lambda_\ell^2(1 - b_1 y)(1 - b_2 y)/(b_1 b_2)$ with

$$b_1 = \frac{\lambda_\ell}{\lambda_\ell + \mu_h + \lambda_h - 2\sqrt{\lambda_h\mu_h}} = \frac{\lambda_\ell}{\lambda_\ell + (\sqrt{\mu_h} - \sqrt{\lambda_h})^2}, \tag{2.8}$$

$$b_2 = \frac{\lambda_\ell}{\lambda_\ell + \mu_h + \lambda_h + 2\sqrt{\lambda_h\mu_h}} = \frac{\lambda_\ell}{\lambda_\ell + (\sqrt{\mu_h} + \sqrt{\lambda_h})^2}. \tag{2.9}$$

$1/b_1$ and $1/b_2$ are called the branch points of the kernel with $b_1 > b_2$. Therefore, $x_1(y)$ and $x_2(y)$ can also be written as

$$x_1(y) = \frac{(\lambda + \mu_h - \lambda_\ell y) - \lambda_\ell\sqrt{(1 - b_1 y)(1 - b_2 y)/(b_1 b_2)}}{2\lambda_h},$$

$$x_2(y) = \frac{(\lambda + \mu_h - \lambda_\ell y) + \lambda_\ell\sqrt{(1 - b_1 y)(1 - b_2 y)/(b_1 b_2)}}{2\lambda_h}.$$

When $y = 0$, we have

$$x_1 = x_1(0) = \frac{r_0}{\rho_h}, \tag{2.10}$$

$$x_2 = x_2(0) = \frac{1}{r_0}, \tag{2.11}$$

where

$$r_0 = \frac{(\lambda + \mu_h) - \sqrt{(\lambda + \mu_h)^2 - 4\lambda_h\mu_h}}{2\mu_h}. \tag{2.12}$$

Remark 2.2 It can be easily proven that $0 < b_2 < b_1 < 1$.

The following lemma can be easily proved using simple algebra.

Lemma 2.1 *For* $-1 \le y \le 1$, *we have* $0 < x_1(y) < x_2(y)$ *and* $x_1(y) \le 1$. *When* $y = 0$, *we then have* $0 < x_1 < 1 < x_2$.

The generating functions $\varphi_j(x)$ are recursively determined in the following proposition.

Proposition 2.2

$$\varphi_0(x) = \frac{1-\rho}{1-r_0 x} \tag{2.13}$$

and

$$\varphi_j(x) = \frac{a_j}{1-r_0 x} + \frac{\lambda_\ell r_0 x}{\lambda_h(1-r_0 x)} \frac{\varphi_{j-1}(x) - \varphi_{j-1}(x_1)}{x - x_1}, \quad j = 1, 2, \ldots, \tag{2.14}$$

where

$$a_j = \frac{[(\mu_h - \mu_\ell)\pi_{0,j} + \mu_\ell \pi_{0,j+1} + \lambda_\ell \varphi_{j-1}(x_1)]r_0}{\lambda_h}.$$

Proof It follows from (2.3) that

$$\varphi_0(x) = \frac{\mu_h(\pi_{0,0} + \pi_{1,0}x) - (\lambda + \mu_h)\pi_{0,0}x}{\lambda_h x^2 - (\lambda + \mu_h)x + \mu_h} = \frac{\mu_h(\pi_{0,0} + \pi_{1,0}x) - (\lambda + \mu_h)\pi_{0,0}x}{\lambda_h(x - x_1)(x - x_2)}, \tag{2.15}$$

where x_1 and x_2 are given by, respectively, (2.10) and (2.11). Similarly, it follows from (2.4) that

$$\left(\lambda + \mu_h - \frac{\mu_h}{x} - \lambda_h x\right)\varphi_j(x) = (\lambda + \mu_h)\pi_{0,j} - (\mu_h/x)[\pi_{0,j} + \pi_{1,j}x] + \lambda_\ell \varphi_{j-1}(x) - \lambda_\ell \pi_{0,j-1}, \quad j \geq 1.$$

By using (2.2), the above equation can be written as

$$\begin{aligned}&\left(\lambda + \mu_h - \frac{\mu_h}{x} - \lambda_h x\right)\varphi_j(x)\\ &\quad = (\lambda + \mu_\ell + \mu_h - \mu_\ell)\pi_{0,j} - \mu_h \pi_{1,j} - \lambda_\ell \pi_{0,j-1} - \frac{\mu_h}{x}\pi_{0,j} + \lambda_\ell \varphi_{j-1}(x)\\ &\quad = \mu_\ell \pi_{0,j+1} + (\mu_h - \mu_\ell)\pi_{0,j} - \frac{\mu_h}{x}\pi_{0,j} + \lambda_\ell \varphi_{j-1}(x), \quad j \geq 1,\end{aligned}$$

or

$$\varphi_j(x) = \frac{\mu_h \pi_{0,j} + x[\mu_\ell(\pi_{0,j} - \pi_{0,j+1}) - \mu_h \pi_{0,j}] - \lambda_\ell x \varphi_{j-1}(x)}{\lambda_h(x - x_1)(x - x_2)}.$$

For (2.13), notice that $x_1 < 1$ and $\varphi_0(x)$ is analytic inside the unit circle, which implies that $x_1 = r_0/\rho_h$ is also a zero of the numerator of the function on the right-hand side of (2.15). Therefore, (2.15) becomes

$$\varphi_0(x) = \frac{\text{Const}}{1 - r_0 x},$$

where $\text{Const} = 1 - \rho$ is determined in Proposition 2.1.

For (2.14), based on the same argument,

$$\mu_h \pi_{0,j} + x_1\big[\mu_\ell(\pi_{0,j} - \pi_{0,j+1}) - \mu_h \pi_{0,j}\big] - \lambda_\ell x_1 \varphi_{j-1}(x_1) = 0.$$

Therefore,

$$\varphi_j(x) = \frac{\lambda_\ell x\varphi_{j-1}(x) + (\mu_h - \mu_\ell)x\pi_{0,j} + \mu_\ell x\pi_{0,j+1} - \lambda_\ell x_1\varphi_{j-1}(x_1) - (\mu_h - \mu_\ell)x_1\pi_{0,j} - \mu_\ell x_1\pi_{0,j+1}}{\lambda_h x_2(x - x_1)(1 - x/x_2)}$$
$$= \frac{\lambda_\ell x\varphi_{j-1}(x) - \lambda_\ell x\varphi_{j-1}(x_1) + [\lambda_\ell\varphi_{j-1}(x_1) + (\mu_h - \mu_\ell)\pi_{0,j} + \mu_\ell x\pi_{0,j+1}](x - x_1)}{\lambda_h x_2(x - x_1)(1 - r_0 x)},$$

which leads to (2.14). □

In the following, we determine the generating function $\psi_0(y)$.

Proposition 2.3 *Let*

$$F(y) = \lambda_\ell y^2 - [1 - 2\mu_h + \mu_\ell]y + 2\mu_\ell, \tag{2.16}$$

$T(y) = F(y) - y\sqrt{\Delta(y)}$ *and* $T^*(y) = F(y) + y\sqrt{\Delta(y)}$, *where* $\Delta(y)$ *is defined in* (2.7). *Then, for* $|y| \le 1$, *we have*

$$\psi_0(y) = a\frac{T^*(y)}{1 - \eta_1 y} + b\frac{T^*(y)}{1 - \eta_2 y}, \tag{2.17}$$

where

$$a = \frac{1 - \rho}{2\mu_\ell}\frac{\eta_1}{\eta_1 - \eta_2}, \qquad b = \frac{1 - \rho}{2\mu_\ell}\frac{\eta_2}{\eta_2 - \eta_1}, \tag{2.18}$$

$$\eta_1 = \frac{(1 - 2\mu_h) + \sqrt{(1 - 2\mu_h)^2 + 4(\mu_h - \mu_\ell)\lambda_\ell}}{2\mu_\ell}, \tag{2.19}$$

$$\eta_2 = \frac{(1 - 2\mu_h) - \sqrt{(1 - 2\mu_h)^2 + 4(\mu_h - \mu_\ell)\lambda_\ell}}{2\mu_\ell}. \tag{2.20}$$

When $\mu_\ell = \mu_h$, *we have* $\eta_2 = 0$ *and* $\eta_1 = \rho$. *In this case, the expression for* $\psi_0(y)$ *can be further simplified.*

Proof According to the expression for the generating function $P(x, y)$ given in (2.5), we have

$$P(x, y) = \frac{[(\mu_h - \mu_\ell)xy - \mu_h y + \mu_\ell x]\psi_0(y) + (1 - \rho)\mu_\ell(y - 1)x}{\lambda_h y(x - x_1(y))(x - x_2(y))},$$

where $x_1(y)$ and $x_2(y)$ are the two zeros of the key kernel function given in (2.7). For $-1 \le y \le 1$, $P(x_1(y), y)$ is analytic and nonzero. Therefore, the kernel equal to zero implies that

$$\big[(\mu_h - \mu_\ell)x_1(y)y - \mu_h y + \mu_\ell x_1(y)\big]\psi_0(y) + (1 - \rho)\mu_\ell(y - 1)x_1(y) = 0.$$

The above equation leads to

$$\psi_0(y) = \frac{(1 - y)x_1(y)(1 - \rho)\mu_\ell}{x_1(y)[(\mu_h - \mu_\ell)y + \mu_\ell] - \mu_h y} = \frac{2(1 - y)(1 - \rho)\mu_\ell T^*(y)}{T(y)T^*(y)},$$

by noticing that $x_1(y)x_2(y) = \frac{\mu_h}{\lambda_h}$ and the expression of x_2. For (2.17), we study the kernel equation $T(y)T^*(y) = 0$ for $\psi_0(y)$ to identify the zeros of the kernel, which gives

$$T(y)T^*(y) = 4\mu_\ell^2(1-y)(1-\eta_1 y)(1-\eta_2 y).$$

Substituting the above into $\psi_0(y)$ and then using partial fractions lead to (2.17) for $-1 \le y \le 1$. Finally, notice that as a function of complex numbers, $T^*(y)$ is analytic in $|y| \le 1$ since both branch points $1/b_1 > 1$ and $1/b_2 > 1$. Therefore, (2.17) holds for $|y| \le 1$, which completes the proof of the proposition. □

Remark 2.3 (i) $1/\eta_1$ and $1/\eta_2$ are the two non-unit zeros of the denominator, or the kernel, of the generating function $\psi_0(y)$ for $\pi_{0,n}$ with $\eta_1 > \eta_2$. (ii) Writing the kernel function of $\psi_0(y)$ as a polynomial kernel is more convenient for the analysis of singularities. The three parameter regions, which will be characterized, are described by $F(1/b_1) > 0, = 0$ and < 0, respectively.

Remark 2.4 From the above proof, one can see that $\psi_0(y)$ can be continued as a meromorphic function in the whole area of the complex plane by cutting the line $[1/b_1, \infty)$ if we specify $\sqrt{\Delta(y)}$ to be the branch such that $\sqrt{\Delta(1)} > 0$. Throughout the paper, we will always choose such a branch for a square root function.

As will be seen (from the Key Lemma) the type of exact tail asymptotics relies on the region characterized by the system parameters, or the value of $F(1/b_1)$ at the branch point $1/b_1$. We first simplify the expression of $F(1/b_i)$.

Lemma 2.2

$$F\left(\frac{1}{b_1}\right) = \frac{2}{\lambda_\ell}\left[\left(\lambda+\mu_h-2\sqrt{\lambda_h\mu_h}\right)\left(\mu_h-\mu_\ell-\sqrt{\lambda_h\mu_h}\right)+\lambda_\ell\mu_\ell\right],$$

$$F\left(\frac{1}{b_2}\right) = \frac{2}{\lambda_\ell}\left[\left(\lambda+\mu_h+2\sqrt{\lambda_h\mu_h}\right)\left(\mu_h-\mu_\ell+\sqrt{\lambda_h\mu_h}\right)+\lambda_\ell\mu_\ell\right].$$

$F(1/b_1)$ can be further simplified as

$$F\left(\frac{1}{b_1}\right) = \frac{2\mu_h^2}{\lambda_\ell}\left(1-\sqrt{\rho_h}\right)\left(\rho-\sqrt{\rho_h}\right)$$

if $\mu_\ell = \mu_h$.

Proof The proof follows from the definition of $F(y)$ in (2.16), the expressions of b_1 and b_2 in (2.8) and (2.9), and elementary manipulations. □

Remark 2.5 The above simplification is important since it provides an explicit characterization of the whole stability region according to the type of asymptotics. For convenience, let

$$D = \left(\lambda+\mu_h-2\sqrt{\lambda_h\mu_h}\right)\left(\mu_h-\mu_\ell-\sqrt{\lambda_h\mu_h}\right)+\lambda_\ell\mu_\ell. \tag{2.21}$$

It is clear that $D > 0, = 0$ or < 0 is equivalent to $F(1/b_1) > 0, = 0$ or < 0, respectively.

As a by-product, we have the following equivalent stability condition.

Lemma 2.3 *$\rho < 1$ if and only if $T'(1) < 0$.*

Proof Based on

$$\frac{d}{dy}T(y) = 2\lambda_\ell y - (1 - 2\mu_h + \mu_\ell) - \sqrt{(1 - \mu_\ell - \lambda_\ell y)^2 - 4\lambda_h\mu_h} + \frac{(1 - \mu_\ell - \lambda_\ell y)\lambda_\ell y}{\sqrt{(1 - \mu_\ell - \lambda_\ell y)^2 - 4\lambda_h\mu_h}},$$

we calculate $T'(1) = \lambda_\ell - 2\mu_\ell + \frac{\lambda_\ell(\mu_h+\lambda_h)}{\mu_h-\lambda_h}$, which is negative if and only if $\lambda_\ell\mu_h + 2\lambda_h\mu_\ell < \mu_\ell\mu_h$, or $\rho < 1$. □

3 Main results

In this section, we provide a complete characterization of exact tail asymptotics in the stationary distribution, joint, marginal and for the total number of customers in the system. Proofs of these results will be detailed in later sections.

The following lemma characterizes the tail asymptotics in the two marginal distributions.

Lemma 3.1 *For the classical $M/M/1$ preemptive priority queue with two classes of customers satisfying $\rho < 1$, characterizations of the exact tail asymptotics in the two marginal stationary distributions are given by*:

(1) *For the marginal distribution $\pi_n^{(h)}$ of the high-priority queue, we have $\pi_n^{(h)} \sim (1 - \rho_h)\rho_h^n$. The decay rate in the marginal distribution for the high-priority queue is ρ_h.*
(2) *For the marginal distribution $\pi_n^{(\ell)}$ of the low-priority queue, we have*

$$\pi_n^{(\ell)} = \frac{\mu_\ell}{\lambda_\ell}\pi_{0,n+1},$$

where the characterization of exact tail asymptotics of $\pi_{0,n+1}$ is given by the result for $i = 0$ in Theorem 3.2.

Remark 3.1 The exact tail is well known for the marginal distribution for the queue length of high-priority customers. In fact, the distribution is geometric.

For stating the exact tail asymptotic property along the high-priority queue direction we need to define

$$r_1 = \frac{\lambda_\ell r_0}{\sqrt{(\lambda + \mu_h)^2 - 4\lambda_h\mu_h}}, \tag{3.1}$$

where r_0 is given in (2.12).

Remark 3.2 r_0 and r_1 are the first two main diagonal entries in the rate matrix R studied in Miller [37].

Theorem 3.1 *For the classical $M/M/1$ preemptive priority queue with two classes of customers satisfying $\rho < 1$, the exact tail asymptotics in the joint stationary distribution along the high-priority queue direction is characterized by: for a fixed number $j \geq 0$ of low-priority customers,*

$$\pi_{n,j} \sim (1-\rho)\left(\frac{r_1^j}{j!}\right) n^j r_0^{n-j},$$

where r_0 and r_1 are given, respectively, by (2.12) and (3.1).

Remark 3.3 In [42], Miyazawa provided a description of exact tail asymptotics for the double QBD process under an irreducible condition, which has only three types of exact tail asymptotics. The type presented in the above theorem is not among these three types since the irreducible condition imposed there is not presented here. It would be interesting to investigate how many more types of exact tail asymptotics might exist when the irreducible condition is not satisfied. Intuitively, this result is not surprising since the asymptotic distribution along the low-priority queue direction given the busy period of high-priority customers must be Poisson when the busy period gets large. The extra polynomial prefactor n^j is a property of large deviations along the low-priority queue direction.

Along the low-priority queue direction, the model serves as an example in which all three types of exact tail asymptotics are presented corresponding to the three cases of the dominant singularity in the generating function $\psi_0(y)$: a pole only; a pole and a branch point simultaneously; and a branch point only. To state the result, we partition the whole stability region into three regions according to $D > 0$, $D = 0$ or $D < 0$, respectively, where D is given in (2.21). For the case of $\mu_\ell = \mu_h$, these regions can be simplified as a direct consequence of Lemma 2.2.

Corollary 3.1 *When $\mu_\ell = \mu_h$, the regions $D > 0$, $D = 0$ and $D < 0$ are simplified to $\rho^2 > \rho_h$, $\rho^2 = \rho_h$ and $\rho^2 < \rho_h$, respectively. In this case, $\eta_1 = \rho$.*

For stating the tail asymptotics for $\pi_{i,n}$ along the direction n of the low-priority queue, let

$$u(\eta) = \frac{1-\mu_\ell - (\lambda_\ell/\eta) - \sqrt{[(1-\mu_\ell) - (\lambda_\ell/\eta)]^2 - 4\lambda_h\mu_h}}{2\mu_h}. \tag{3.2}$$

Also, let $C_{\ell,1}$, $C_{\ell,2}$ and $C_{\ell,3}$ be defined by (5.4), which are constants independent of n and i.

Remark 3.4 The tail asymptotic for $\pi_{0,n}$ is obtained based on asymptotic analysis of the generating function $\psi_0(y)$, and for $i > 0$, the tail asymptotic of $\pi_{i,n}$ is derived by the mathematical induction based on balance equations and the tail asymptotic result for $\pi_{0,n}$, which leads to a natural definition of $u(\eta)$.

Theorem 3.2 *For the classical $M/M/1$ preemptive priority queue with two classes of customers satisfying $\rho < 1$, characterizations of the exact tail asymptotics in the joint stationary distribution along the low-priority queue direction are given below for a fixed number $i \geq 0$ of high-priority customers*:

(1) (*Exact geometric decay*) *In the region defined by $D > 0$,*

$$\pi_{i,n} \sim C_{\ell,1}\big[u(\eta_1)\big]^i \eta_1^n,$$

where η_1 and $u(\eta_1)$ are given, respectively, by (2.19) *and* (3.2).

(2) (*Geometric decay with prefactor $n^{-1/2}$*) *In the region defined by $D = 0$,*

$$\pi_{i,n} \sim C_{\ell,2}\big(\sqrt{\rho_h}\big)^i n^{-1/2} b_1^n,$$

where $b_1 = \eta_1$ is given by (2.19) *or* (2.8).

(3) (*Geometric decay with prefactor $n^{-3/2}$*) *In the region defined by $D < 0$,*

$$\pi_{i,n} \sim C_{\ell,3}\big(1 + i\,\tilde{B}\big)\big(\sqrt{\rho_h}\big)^i n^{-3/2} b_1^n,$$

where b_1 is given by (2.8) *and*

$$\tilde{B} = \frac{\mu_\ell - \mu_h - \mu_\ell b_1 + \sqrt{\lambda_h \mu_h}}{\sqrt{\lambda_h \mu_h}}.$$

Remark 3.5 The three types of tail asymptotics correspond to positive recurrent, null recurrent, and transient properties, respectively, in the probabilistic method and the matrix analytic method.

Let $\pi_n^{(T)} = \sum_{n=i+j} \pi_{i,j}$ be the probability distribution of the total number of customers in the system. To characterize the exact tail asymptotics for $\pi_n^{(T)}$, we further partition each of the three regions according to whether or not $\bar{\rho}_h \geq 1$. This makes sense since intuitively given that the number of high-priority customers is greater than zero, then $\bar{\rho}_h$ is simply the traffic load to the total of customers in the system, which should play an important role for the tail asymptotic. Let $C_{t,1a}$, $C_{t,1b}$, $C_{t,1c}$, $C_{t,2a}$, $C_{t,2b}$, $C_{t,2c}$, $C_{t,3a}$ and $C_{t,3b}$ be constants given by (6.1), (6.2), (6.3), (6.4), (6.2), (6.5), (6.6) and (6.2), respectively, which are independent of n.

Theorem 3.3 *For the classical $M/M/1$ preemptive priority queue with two classes of customers satisfying $\rho < 1$, if $\mu_h = \mu_\ell$, then*

$$\pi_n^{(T)} = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots.$$

If $\mu_h \neq \mu_\ell$, then the exact tail asymptotic in the stationary distribution $\pi_n^{(T)}$ of the total number of customers in the system is characterized below.

(1) *In the region defined by* $D > 0$, *we have* $b_1 < \eta_1$. *Three cases exist*:
   (a) *If* (i) $\bar{\rho}_h \geq 1$; *or* (ii) $\bar{\rho}_h < 1$ *and* $\bar{\rho}_h < \eta_1$, *then*

$$\pi_n^{(T)} \sim C_{t,1a}\eta_1^n,$$

   where η_1 *is given by* (2.19).
   (b) *If* $\bar{\rho}_h < 1$ *and* $\bar{\rho}_h > \eta_1$, *then*

$$\pi_n^{(T)} \sim C_{t,1b}(\bar{\rho}_h)^n.$$

   (c) *If* $\bar{\rho}_h < 1$ *and* $\bar{\rho}_h = \eta_1$, *then*

$$\pi_n^{(T)} \sim C_{t,1c}n\eta_1^n,$$

   where η_1 *is given by* (2.19).
(2) *In the region defined by* $D = 0$, *we have* $b_1 = \eta_1$. *Three cases exist*:
   (a) *If* $\bar{\rho}_h \geq 1$, *then*

$$\pi_n^{(T)} \sim C_{t,2a}n^{-1/2}b_1^n,$$

   where b_1 *is given by* (2.8).
   (b) *If* $\bar{\rho}_h < 1$ *and* $\bar{\rho}_h \neq \sqrt{\rho_h}$, *then*

$$\pi_n^{(T)} \sim C_{t,2b}(\bar{\rho}_h)^n.$$

   (c) *If* $\bar{\rho}_h < 1$ *and* $\bar{\rho}_h = \sqrt{\rho_h}$, *then* $\bar{\rho}_h = b_1 = \eta_1$ *and*

$$\pi_n^{(T)} \sim C_{t,2c}\,n^{1/2}b_1^n,$$

   where b_1 *is given by* (2.8).
(3) *In the region defined by* $D < 0$, *two cases exist*:
   (a) *If* $\bar{\rho}_h \geq 1$, *then*

$$\pi_n^{(T)} \sim C_{t,3a}n^{-3/2}b_1^n,$$

   where b_1 *is given by* (2.8).
   (b) *If* $\bar{\rho}_h < 1$, *then*

$$\pi_n^{(T)} \sim C_{t,3b}(\bar{\rho}_h)^n.$$

Remark 3.6 When $\mu_h = \mu_\ell$, the result is obvious since the dynamic is the same as that for the $M/M/1$ queue with the traffic intensity ρ. When $\mu_h \neq \mu_\ell$, the decay is determined by the dominant singularity of $P(x, x)$, which is either the pole $1/\eta_1$, or the pole $1/\bar{\rho}_h$ or the branch point $1/b_1$. The prefactor is simply a constant (which leads to the type of exact geometric decay); $n^{-1/2}$; $n^{-3/2}$; n; or $n^{1/2}$ when the dominant singularity is a simple pole (either $1/\eta_1$ or $1/\bar{\rho}_h$); both a simple pole and a branch point ($\eta_1 = b_1 \neq \bar{\rho}_h$ while the case $\bar{\rho}_h = b_1 \neq \eta_1$ is not possible); a branch point only ($1/b_1$); a double pole ($\eta_1 = \bar{\rho}_h$); or both a double pole and a branch point ($\eta_1 = \bar{\rho}_h = b_1$), respectively.

Remark 3.7 It should be noted that two additional types of exact tail asymptotics are presented here for the total number of customers in the system, although it is believed that in general, the three types (with a constant, $n^{-1/2}$ and $n^{-3/2}$ prefactors) of exact tail asymptotics are the only asymptotics along any directions including the main diagonal direction. This is not a contradiction since the asymptotic for the total number of customers in the system is not the same as that along the diagonal direction.

4 Analysis of singularities and asymptotic expansions

The analysis of exact tail asymptotics along the low-priority queue direction and for the total number of customers in the system in the stationary distribution $\pi_{i,j}$ of the priority queueing model relies on the analysis of the singularities of the generating function $\psi_0(y)$, which is the focus of this section. Asymptotics of the coefficients of $\psi_0(y)$ are obtained by using a Tauberian-type theorem, such as Theorem 4 in Bender [4] or Corollary 2 in Flajolet and Odlyzko [13], which is restated in the following lemma for convenience. The key idea used in the following analysis is the same as that used in Lieshout and Mandjes [31].

For a function $f(y)$ that is analytic at $y = 0$, we denote the coefficient of y^k in the Taylor expression of $f(y)$ by $C_k[f(y)]$. A function is denoted by $o(k(n))$ if $\lim_{n\to\infty} o(k(n))/k(n) = 0$.

Lemma 4.1 (Flajolet and Odlyzko) *Assume that $f(z)$ is analytic in $\Delta(\phi, \varepsilon) = \{z : |z| \le 1 + \varepsilon, |\mathrm{Arg}\,(z-1)| \ge \phi, \varepsilon > 0, 0 < \phi < \pi/2\}$ except at $z = 1$ and*

$$f(z) \sim K(1-z)^s \quad \text{as } z \to 1 \text{ in } \Delta(\phi, \varepsilon).$$

Then as $n \to \infty$: (i) *If $s \notin \{0, 1, 2, \ldots\}$,*

$$f_n \sim \frac{K}{\Gamma(-s)} n^{-s-1}.$$

(ii) *If s is a nonnegative integer, then*

$$f_n = o\big(n^{-s-1}\big).$$

Remark 4.1 $\Delta(\phi, \varepsilon)$ is a domain of the form of an indented disk.

The key goal is to locate the dominant singularity, which determines the decay, and to characterize the nature of the dominant singularity, which determines the prefactor and the singularity coefficient. Changes in values of the system parameters make a significant impact on the above properties. Therefore, the following lemmas play a key role in the asymptotic analysis.

Lemma 4.2 *For the non-unit zeros $1/\eta_1$ and $1/\eta_2$ of the denominator of $\psi_0(y)$, where η_1 and η_2 are given in (2.19) and (2.20) respectively, we have* (1) *both η_1 and η_2 are real*; (2) $\eta_1 > 0$; (3) $\eta_1 > \eta_2$; *and* (4) $\eta_2 < 0$, $\eta_2 = 0$, *or* $\eta_2 > 0$ *if and only if* $\mu_\ell < \mu_h$, $\mu_\ell = \mu_h$, *or* $\mu_\ell > \mu_h$, *respectively.*

Proof It is clear from the expressions of η_1 and η_2 and $(1-2\mu_h)^2+4(\mu_h-\mu_\ell)\lambda_\ell = [\lambda-(\mu_\ell-\mu_h)]^2+4(\mu_\ell-\mu_h)\lambda_h$. □

Remark 4.2 $1/\eta_1$ and $1/\eta_2$ are zeros of $T(y)$ and/or $T^*(y)$, and $T(y)$ and $T^*(y)$ have no other non-unit zeros. We should note that $|\eta_2|>\eta_1$ is possible when $\eta_2<0$. In this case, we prove that η_2 is removable whenever it is a candidate for the dominant singularity. Therefore, η_2 does not contribute to the decay in any asymptotic expression.

The following properties on the function $T(y)$ follow from calculations and using standard methods in calculus.

Lemma 4.3 (1) *$T'(y)$ is an increasing function over the interval* $[0, 1/b_1)$; (2) *$T(y)$ is a decreasing function over the interval* $(-\infty, 0]$ *and does not have any zero in* $(-\infty, 0]$; *and* (3) *in the interval* $[0, 1/b_1]$, $y=1$ *is a zero of* $T(y)$, *and there is at most another zero that lies in the subinterval* $(1, 1/b_1]$.

Remark 4.3 If $\eta_2<0$, it follows from the above lemma that $T^*(1/\eta_2)=0$ and $T(1/\eta_2)\neq 0$.

The following lemma provides information about the dominant singularity in each case.

Lemma 4.4 (Key Lemma) *For the property of* $1/\eta_1$, *there are three cases*:

(1) *In the region defined by* $D>0$, $1<1/\eta_1<1/b_1$ *and* $1/\eta_1$ *is a zero of* $T(y)$ (*but not* $T^*(y)$).
(2) *In the region defined by* $D=0$, $1<1/\eta_1=1/b_1$ *and* $1/\eta_1$ *is a zero of both* $T(y)$ *and* $T^*(y)$.
(3) *In the region defined by* $D<0$, $1<1/\eta_1<1/b_1$ *and* $1/\eta_1$ *is a zero of* $T^*(y)$ (*but not* $T(y)$).

Proof If $D>0$ or equivalently $F(\frac{1}{b_1})>0$, from $T(1)=0$ and $T'(1)<0$, we can easily show that $T(y)$ has a zero in $(1, 1/b_1)$, which should be either $1/\eta_1$ or $1/\eta_2$. Suppose $T(1/\eta_2)=0$, then $\eta_2>b_1>0$ and $T^*(1/\eta_2)\neq 0$. Since $T(1)=T(1/\eta_2)=0$ and $T'(y)$ is an increasing function, we have $T'(1/\eta_2)>0$. From

$$\frac{d}{dy}T(y)T^*(y)=4\mu_l^2\big[-(1-\eta_1 y)(1-\eta_2 y)-\eta_2(1-y)(1-\eta_1 y)-\eta_1(1-y)(1-\eta_2 y)\big],$$

we obtain $T^*(1/\eta_2)T'(1/\eta_2)=4\mu_l^2[-\eta_2(1-1/\eta_2)(1-\eta_1/\eta_2)]<0$, hence $T^*(1/\eta_2)<0$. On the other hand, $T^*(1/b_1)=F(1/b_1)>0$, which implies $T^*(y)$ has another zero other than $1/\eta_1$. The contradiction shows $T(1/\eta_1)=0$.

If $F(\frac{1}{b_1})=0$, then both $T(\frac{1}{b_1})=0$ and $T^*(\frac{1}{b_1})=0$ since $y\sqrt{\Delta(y)}=0$. In this case, either $b_1=\eta_1$ or $b_1=\eta_2$. If $b_1=\eta_2$, then $T^*(1/\eta_1)=0$ and $T(1/\eta_1)\neq 0$. Sim-

ilarly, from $\frac{d}{dy}[T(y)T^*(y)]$ we obtain $T(1/\eta_1)T^{*\prime}(1/\eta_1) = 4\mu_l^2[-\eta_1(1-1/\eta_1)\times(1-\eta_2/\eta_1)] > 0$. Since $T^*(1/\eta_1) = 0$, $T(1/\eta_1) < 0$, which yields $T^{*\prime}(1/\eta_1) < 0$. On the other hand, it can be shown easily that $\lim_{y\to 1/b_1^-} T^{*\prime}(y) = -\infty$, which implies that there must exist $1/\eta_1 < \zeta < 1/b_1$, such that $T^*(\zeta) > 0$ since $T^*(1/b_1) = 0$. Hence $T^*(y)$ has another zero other than $1/\eta_1$. The contradiction shows $b_1 = \eta_1$.

If $F(\frac{1}{b_1}) < 0$, since $T^*(0) > 0$ and $T^*(1/b_1) < 0$, $T^*(y)$ has a zero in $(0, 1/b_1)$, which is either $\frac{1}{\eta_1}$ or $\frac{1}{\eta_2}$. We show that it cannot be $\frac{1}{\eta_2}$. Otherwise, by Lemma 4.2 we would have $\frac{1}{\eta_1} < \frac{1}{\eta_2} < \frac{1}{b_1}$. It then follows from Remark 4.2 that either $T(1/\eta_1) = 0$ or $T^*(1/\eta_1) = 0$. If $T(1/\eta_1) = 0$, then $T'(1/\eta_1) \neq 0$ which would yield another zero of $T(y)$ in interval $(1, 1/b_1)$ since $T(1) = 0$, $T'(1) < 0$ and $T(1/b_1) < 0$. This contradicts (3) of Lemma 4.3. Hence $T^*(1/\eta_1) = 0$ must hold. So, in either case, $T^*(1/\eta_1) = 0$ and $1 < \frac{1}{\eta_1} < \frac{1}{b_1}$. □

Corollary 4.1 *$\eta_2 \neq b_1$, and either $|\eta_2| < b_1$ or $T^*(1/\eta_2) = 0$.*

Proof From the Key Lemma, we see that when $F(1/b_1) \geq 0$, we have $\eta_2 \neq b_1$. In the case of $F(1/b_1) < 0$, we also have $\eta_2 \neq b_1$, which follows from $T(1/b_1) = T^*(1/b_1) = F(1/b_1) \neq 0$ and Remark 4.2. Now, assume $|\eta_2| \geq b_1$. By the same argument used to show $T^*(1/\eta_1) = 0$ in the proof of (3) of the Key Lemma, we obtain $T^*(1/\eta_2) = 0$ when $\eta_2 > b_1$. On the other hand, according to Remark 4.3, we also have $T^*(1/\eta_2) = 0$ when $\eta_2 < 0$. This completes the proof of the corollary. □

Remark 4.4 The three cases of $F(y)$ in the proposition correspond to the three types of exact tail asymptotics as shown later: (i) exact geometric (the simple pole $1/\eta_1$ is the dominant singularity of $\psi_0(y)$); (ii) geometric with a prefactor $n^{-1/2}$ (the simple pole $1/\eta_1$ and the branch point $1/b_1$ coincide as the dominant singularity); and (iii) geometric with a prefactor $n^{-3/2}$ (the branch point $1/b_1$ is the dominant singularity).

5 Exact tail asymptotics for the low-priority queue

For the low-priority queue, we consider two cases: (1) exact tail asymptotics in the joint distribution along the low-priority queue direction, which is stated in Theorem 3.2; and (2) exact tail asymptotics in the marginal distribution of the low-priority queue, which is stated in Lemma 3.1(2) and proved to have the same form as that in (1).

It follows from the Key Lemma that the types of exact tail asymptotics rely on the region characterized by $D > 0$, $D = 0$ or $D < 0$.

We first prove a proposition.

Proposition 5.1 *Let $\eta \neq 0$ and $\eta \neq b_1$, and assume $T^*(1/\eta) = 0$ when $\eta > b_1$ or $\eta < 0$, then*

$$C_n\left[\frac{T^*(y)}{1-\eta y}\right] \sim \sigma(\eta) n^{-3/2} b_1^n, \tag{5.1}$$

where

$$\sigma(\eta) = \frac{\lambda_\ell \sqrt{1 - \frac{b_2}{b_1}}}{2\sqrt{b_1 b_2}\sqrt{\pi}(\eta - b_1)}. \tag{5.2}$$

Proof Let

$$\Delta(\phi, \varepsilon) = \big\{ b_1 y : |b_1 y| \le 1 + \varepsilon, \big|\mathrm{Arg}\,(b_1 y - 1)\big| \ge \phi, \varepsilon > 0, 0 < \phi < \pi/2 \big\}. \tag{5.3}$$

First assume $\eta > b_1$ or $\eta < 0$. It follows from direct algebraic manipulations and $T^*(1/\eta) = 0$ that

$$C_n\left[\frac{T^*(y)}{1-\eta y}\right] = C_n\left[\frac{F(y) - F(1/\eta)}{1-\eta y}\right] - \frac{1}{\eta} C_n\big[\sqrt{H(y)}\big] + \frac{1}{\eta} C_n\left[\frac{S(y)}{\sqrt{H(y)} + \sqrt{H(1/\eta)}}\right],$$

where

$$H(y) = \frac{\lambda_\ell^2 (1 - b_2 y)(1 - b_1 y)}{b_1 b_2}, \qquad S(y) = \frac{H(y) - H(1/\eta)}{1 - \eta y}.$$

Clearly, the first term is a polynomial of degree 1, hence $C_n[\frac{F(y)-F(1/\eta)}{1-\eta y}] = 0$ for $n > 1$. For the second term, it is easy to see that $\sqrt{H(y)}$ is analytic in $\Delta(\phi, \varepsilon)$ except at $b_1 y = 1$ and $\sqrt{H(y)} \sim \lambda_\ell \sqrt{\frac{1-b_2/b_1}{b_1 b_2}} \sqrt{1 - b_1 y}$ as $b_1 y \to 1$ in $\Delta(\phi, \varepsilon)$. By applying the Tauberian theorem (Lemma 4.1) to the second term, we obtain

$$-\frac{1}{\eta} C_n\big[\sqrt{H(y)}\big] \sim \frac{\lambda_\ell}{2\eta\sqrt{\pi}} \sqrt{\frac{1 - b_2/b_1}{b_1 b_2}}\, n^{-3/2} b_1^n.$$

For the third term, since

$$\frac{S(y)\sqrt{H(y)}}{\sqrt{H(y)} + \sqrt{H(1/\eta)}} + \frac{S(y)\sqrt{H(1/\eta)}}{\sqrt{H(y)} + \sqrt{H(1/\eta)}} = S(y)$$

and $S(y)$ is a polynomial of degree 1, we have for $n \ge 2$,

$$C_n\left[\frac{S(y)}{\sqrt{H(y)} + \sqrt{H(1/\eta)}}\right] = -\frac{1}{\sqrt{H(1/\eta)}} C_n\left[\frac{S(y)\sqrt{H(y)}}{\sqrt{H(y)} + \sqrt{H(1/\eta)}}\right].$$

It is easy to see that $\frac{S(y)\sqrt{H(y)}}{\sqrt{H(y)}+\sqrt{H(1/\eta)}}$ is analytic in $\Delta(\phi, \varepsilon)$ except at $b_1 y = 1$ and $\frac{S(y)\sqrt{H(y)}}{\sqrt{H(y)}+\sqrt{H(1/\eta)}} \sim \frac{S(1/b_1)\lambda_\ell \sqrt{\frac{1-b_2/b_1}{b_1 b_2}} \sqrt{1-b_1 y}}{\sqrt{H(1/\eta)}}$ as $b_1 y \to 1$ in $\Delta(\phi, \varepsilon)$. By the Tauberian theorem (Lemma 4.1), we obtain

$$C_n\left[\frac{S(y)}{\sqrt{H(y)} + \sqrt{H(1/\eta)}}\right] \sim -\frac{\lambda_\ell \sqrt{\frac{1-b_2/b_1}{b_1 b_2}}}{2\sqrt{\pi}(1 - \eta/b_1)} n^{-3/2} b_1^n.$$

Combining the three asymptotics gives (5.1).

Next, assume $0 < \eta < b_1$. In this case, we have

$$\frac{1}{n^{-3/2}b_1^n}C_n\left[\frac{T^*(y)}{1-\eta y}\right] = \frac{1}{n^{-3/2}b_1^n}C_n\left[\frac{F(y)}{1-\eta y}\right] + \frac{1}{n^{-3/2}b_1^n}C_n\left[\frac{y\sqrt{H(y)}}{1-\eta y}\right]$$
$$\to -\frac{(1/b_1)\lambda_\ell\sqrt{\frac{1-b_2/b_1}{b_1b_2}}}{2\sqrt{\pi}(1-\eta/b_1)} = \sigma(\eta),$$

since $C_n[\frac{F(y)}{1-\eta y}] \sim o(n^{-3/2}b_1^n)$ by a direct power series expansion at zero and

$$C_n\left[\frac{y\sqrt{H(y)}}{1-\eta y}\right] \sim -\frac{(1/b_1)\lambda_\ell\sqrt{\frac{1-b_2/b_1}{b_1b_2}}}{2\sqrt{\pi}(1-\eta/b_1)}n^{-3/2}b_1^n$$

by using Lemma 4.1 with the facts that $0 < \eta < b_1$, $\frac{y\sqrt{H(y)}}{1-\eta y}$ is analytic in the domain $\Delta(\phi,\varepsilon)$ except at $b_1 y = 1$ and $\frac{y\sqrt{H(y)}}{1-\eta y} \sim \frac{(1/b_1)\lambda_\ell\sqrt{\frac{1-b_2/b_1}{b_1b_2}}\sqrt{1-b_1 y}}{1-\eta/b_1}$. This completes the proof of the proposition. □

Theorem 3.2 consists of two parts: $i = 0$ and $i > 0$. We first characterize the exact tail asymptotic for $\pi_{0,n}$ (Part I or $i = 0$), based on which and the induction we further obtain the tail asymptotic for $\pi_{i,n}$ for an arbitrary $i > 0$ (Part II). Define

$$C_{\ell,1} = 2aF\left(\frac{1}{\eta_1}\right), \qquad C_{\ell,2} = a\frac{\lambda_\ell\sqrt{1-b_2/b_1}}{\sqrt{\pi}b_1\sqrt{b_1b_2}}, \qquad C_{\ell,3} = a\sigma(\eta_1) + b\sigma(\eta_2), \tag{5.4}$$

where a and b, η_1, η_2, b_1, b_2, $F(\cdot)$ and $\sigma(\cdot)$ are given in (2.18), (2.19), (2.20), (2.8), (2.9), (2.16), and (5.2), respectively.

Proof of Part I of Theorem 3.2 This is the case of $i = 0$.

1. In this case, $b_1 < \eta_1$, $T(\frac{1}{\eta_1}) = 0$ and $T^*(\frac{1}{\eta_2}) = 0$ according to the Key Lemma. Clearly, $\frac{T^*(y)}{1-\eta_1 y}$ is analytic in $\Delta(\phi,\varepsilon) = \{\eta_1 y : |\eta_1 y| \le 1+\varepsilon, |\mathrm{Arg}(\eta_1 y - 1)| \ge \phi, \varepsilon > 0, 0 < \phi < \pi/2\}$ except at $\eta_1 y = 1$. By Lemma 4.1,

$$aC_n\left[\frac{T^*(y)}{1-\eta_1 y}\right] \sim C_{\ell,1}\eta_1^n.$$

On the other hand, according to Corollary 4.1 and Proposition 5.1, we have

$$C_n\left[\frac{T^*(y)}{1-\eta_2 y}\right] \sim \sigma(\eta_2)n^{-3/2}b_1^n.$$

Therefore,

$$\pi_{0,n} \sim C_{\ell,1}\eta_1^n.$$

2. In this case, $b_1 = \eta_1$ and $T(\frac{1}{\eta_1}) = 0$ according to the Key Lemma. Clearly,

$$C_n\left[\frac{T^*(y)}{1-\eta_1 y}\right] = C_n\left[\frac{F(y)-F(1/b_1)}{1-b_1 y}\right] + C_n\left[\frac{\lambda_\ell y\sqrt{(1-b_2 y)(1-b_1 y)}}{\sqrt{b_1 b_2}(1-b_1 y)}\right],$$

where $C_n[\frac{F(y)-F(1/b_1)}{1-b_1 y}] = 0$ for $n > 1$ since $\frac{F(y)-F(1/b_1)}{1-b_1 y}$ is a polynomial of degree 1. For the second term, since $\frac{\lambda_\ell y\sqrt{(1-b_2 y)(1-b_1 y)}}{\sqrt{b_1 b_2}(1-b_1 y)}$ is analytic in $\Delta(\phi,\varepsilon)$ given in (5.3) except at $b_1 y = 1$ and $\frac{\lambda_\ell y\sqrt{(1-b_2 y)(1-b_1 y)}}{\sqrt{b_1 b_2}(1-b_1 y)} \sim \frac{(\lambda_\ell/b_1)\sqrt{1-b_2/b_1}}{\sqrt{b_1 b_2}\sqrt{1-b_1 y}}$ as $b_1 y \to 1$ in $\Delta(\phi,\varepsilon)$, we have (by Lemma 4.1) $C_n[\frac{\lambda_\ell y\sqrt{(1-b_2 y)(1-b_1 y)}}{\sqrt{b_1 b_2}(1-b_1 y)}] \sim \frac{\lambda_\ell\sqrt{1-b_2/b_1}}{\sqrt{\pi}b_1\sqrt{b_1 b_2}} n^{-1/2} b_1^n$, which gives that

$$C_n\left[\frac{T^*(y)}{1-\eta_1 y}\right] \sim \frac{\lambda_\ell\sqrt{1-b_2/b_1}}{\sqrt{\pi}b_1\sqrt{b_1 b_2}} n^{-1/2} b_1^n.$$

On the other hand, it follows from Corollary 4.1 and Proposition 5.1 that

$$C_n\left[\frac{T^*(y)}{1-\eta_2 y}\right] \sim \sigma(\eta_2) n^{-3/2} b_1^n.$$

Combining the above two asymptotics gives

$$\pi_{0,n} \sim C_{\ell,2} n^{-1/2} b_1^n.$$

3. In this case, $0 < b_1 < \eta_1 < 1$, $T^*(\frac{1}{\eta_1}) = 0$, and either $T^*(\frac{1}{\eta_2}) = 0$ or $0 < \eta_2 < b_1$ according to the Key Lemma, Corollary 4.1 and Remark 4.3. The tail asymptotic of $\pi_{0,n}$ follows from Proposition 5.1. □

Remark 5.1 $C_{\ell,3} > 0$ as we expect.

Remark 5.2 As we have seen (Corollary 3.1), when $\mu_\ell = \mu_h$, the three regions: $D > 0$, $D = 0$ and $D < 0$ are characterized by $\rho^2 > \rho_h$, $\rho^2 = \rho_h$ and $\rho^2 < \rho_h$, respectively. In this case, $\eta_1 = \rho$ and $b = 0$.

We now characterize the tail asymptotics for $\pi_{i,n}$ for $i > 0$, Part II of Theorem 3.2, in which the positivity of the coefficients is guaranteed by the following lemma.

Let $u(\eta)$ be given by (3.2), which is the smaller root of the equation: $\mu_h[t(\eta)]^2 - [1-\mu_\ell-(\lambda_\ell/\eta)]t(\eta) + \lambda_h = 0$.

Lemma 5.1 *Let*

$$w(\eta) = \frac{1-\mu_h-\mu_\ell\eta-\lambda_\ell/\eta}{\mu_h}.$$

If $T(\frac{1}{\eta_1}) = 0$, *then* $w(\eta_1) = u(\eta_1) > 0$; *and if* $T(\frac{1}{\eta_1}) \neq 0$, *we have* $w(b_1) > 0$.

Proof If $T(\frac{1}{\eta_1}) = 0$, we have $F(\frac{1}{\eta_1}) = (1/\eta_1)\sqrt{\Delta(1/\eta_1)}$, or

$$\eta_1 F\left(\frac{1}{\eta_1}\right) = \sqrt{(1-\mu_\ell-\lambda_\ell/\eta_1)^2 - 4\lambda_h\mu_h}\,,$$

from which

$$\begin{aligned} u(\eta_1) &= \frac{1-\mu_\ell-\lambda_\ell/\eta_1 - \sqrt{(1-\mu_\ell-\lambda_\ell/\eta_1)^2-4\lambda_h\mu_h}}{2\mu_h} \\ &= \frac{1-\mu_\ell-\lambda_\ell/\eta_1-\eta_1 F(\frac{1}{\eta_1})}{2\mu_h} = w(\eta_1). \end{aligned}$$

If $T(\frac{1}{\eta_1}) \neq 0$, we have $F(\frac{1}{b_1}) = \frac{(1-\mu_\ell)}{b_1} - \frac{\lambda_\ell}{b_1^2} - 2(\frac{(1-\mu_h)}{b_1} - \mu_\ell - \frac{\lambda_\ell}{b_1^2}) = \frac{b_1(1-\mu_\ell)-\lambda_\ell}{b_1^2} - \frac{2\mu_h w(b_1)}{b_1} < 0$, from which $\frac{2\mu_h w(b_1)}{b_1} = \frac{b_1(1-\mu_\ell)-\lambda_\ell}{b_1^2} - F(\frac{1}{b_1}) > 0$, since $1-\mu_\ell - \lambda_\ell/b_1 = 2\sqrt{\lambda_h\mu_h} > 0$. □

We also have the following simplifications.

Lemma 5.2 *$u(b_1) = \sqrt{\rho_h}$ and*

$$u(\rho) = \begin{cases} \frac{\rho_h}{\rho}, & \text{if } \rho^2 \geq \rho_h, \\ \rho, & \text{if } \rho^2 < \rho_h. \end{cases}$$

Proof Notice that $1/b_1$ is a branch point, or $\sqrt{[(1-\mu_\ell)-(\lambda_\ell/b_1)]^2 - 4\lambda_h\mu_h} = 0$, we have

$$u(b_1) = \frac{1-\mu_\ell-(\lambda_\ell/b_1)}{2\mu_h} = \frac{2\sqrt{\lambda_h\mu_h}}{2\mu_h} = \sqrt{\rho_h}.$$

The second property follows from

$$u(\rho) = \frac{\rho^2+\rho_h - \sqrt{[\rho^2-\rho_h]^2}}{2\rho}.$$ □

The proof of Part II of Theorem 3.2 is given below.

Proof of Part II of Theorem 3.2 This is the case of $i \geq 1$. In this case, the theorem can be proved by using the result for the case of $i = 0$ and the mathematical induction on i. Since the proofs to cases (1) and (2) are similar, we only provide details for case (1).

1. For $i = 1$, the balance equation is

$$\mu_h \frac{\pi_{1,n}}{\eta_1^n} = (\lambda_h+\lambda_\ell+\mu_\ell)\frac{\pi_{0,n}}{\eta_1^n} - \mu_\ell\eta_1\frac{\pi_{0,n+1}}{\eta_1^{n+1}} - \frac{\lambda_\ell}{\eta_1}\frac{\pi_{0,n-1}}{\eta_1^{n-1}}. \tag{5.5}$$

It follows from the result for the case of $i = 0$ in Theorem 3.2 and $w(\eta_1) = u(\eta_1)$ (Lemma 5.1) that

$$\begin{aligned}\pi_{1,n} &\sim \frac{1}{\mu_h} 2aF\left(\frac{1}{\eta_1}\right)\left(\lambda_h + \lambda_\ell + \mu_\ell - \mu_\ell\eta_1 - \frac{\lambda_\ell}{\eta_1}\right)\eta_1^n \\ &= 2aF\left(\frac{1}{\eta_1}\right)w(\eta_1)\eta_1^n = C_{\ell,1}u(\eta_1)\eta_1^n,\end{aligned}$$

which is the conclusion of the theorem.

Assume when $i \le k$,

$$\pi_{i,n} \sim C_{\ell,1}\big[u(\eta_1)\big]^i \eta_1^n.$$

We prove that the conclusion is also true for $i = k+1$. Based on the balance equation

$$\mu_h\pi_{k+1,n} = (\lambda_h + \lambda_\ell + \mu_h)\pi_{k,n} - \lambda_h\pi_{k-1,n} - \lambda_\ell\pi_{k,n-1}, \quad n \ge 1, i \ge 1, \tag{5.6}$$

and the inductive assumption, we have

$$\begin{aligned}\mu_h\frac{\pi_{k+1,n}}{\eta_1^n} &= (\lambda_h + \lambda_\ell + \mu_h)\frac{\pi_{k,n}}{\eta_1^n} - \lambda_h\frac{\pi_{k-1,n}}{\eta_1^n} - \frac{\lambda_\ell}{\eta_1}\frac{\pi_{k,n-1}}{\eta_1^{n-1}} \\ &\sim C_{\ell,1}\big[u(\eta_1)\big]^{k-1}\left\{\left(1 - \mu_\ell - \frac{\lambda_\ell}{\eta_1}\right)u(\eta_1) - \lambda_h\right\} \\ &= C_{\ell,1}\big[u(\eta_1)\big]^{k-1}\mu_h\big[u(\eta_1)\big]^2 = \mu_h C_{\ell,1}\big[u(\eta_1)\big]^{k+1},\end{aligned}$$

since $w(\eta_1) = u(\eta_1)$ and $\mu_h[u(\eta_1)]^2 - [1 - \mu_\ell - (\lambda_\ell/\eta_1)]u(\eta_1) + \lambda_h = 0$.

2. The proof is similar to that for 1.

3. In this case, applying case (3) of the result for $i = 0$ in Theorem 3.2 to (5.5) gives

$$\mu_h\pi_{1,n} \sim C_{\ell,3}\left[1 - \mu_h - \mu_\ell b_1 - \frac{\lambda_\ell}{b_1}\right]n^{-3/2}b_1^n,$$

or

$$\pi_{1,n} \sim C_{\ell,3}c_1 n^{-3/2}b_1^n,$$

where $c_1 = w(b_1)$. Assume for $i \le k$, we have

$$\pi_{i,n} \sim C_{\ell,3}c_i\, n^{-3/2}b_1^n.$$

Then, clearly, it follows from (5.6) and the inductive assumption that $\frac{\pi_{k+1,n}}{n^{-3/2}b_1^n} \to C_{\ell,3}c_{k+1}$, where $c_{i+1} = 2\sqrt{\rho_h}c_i - \rho_h c_{i-1}$, $i = 1, 2, \ldots, k$, with $c_0 = 1$ and $c_1 = w(b_1)$, since $2\sqrt{\rho_h} = \frac{1-\mu_\ell-\frac{\lambda_\ell}{b_1}}{\mu_h}$. Notice that since $\{c_i\}$ for $i \ge 2$ satisfies the homogenous second order recursive relation and $(2\sqrt{\rho_h})^2 - 4\rho_h = 0$ with the boundary conditions $c_0 = 1$ and $c_1 = w(b_1)$, c_i has the solution of the following form:

$c_i = (\tilde{A} + i\tilde{B})(\sqrt{\rho_h})^i$ with $\tilde{A} = 1$ and $(\tilde{A} + \tilde{B})\sqrt{\rho_h} = w(b_1)$, or

$$\tilde{B} = \frac{w(b_1)}{\sqrt{\rho_h}} - 1 = \frac{\frac{1-\mu_h-\mu_\ell b_1-\lambda_\ell/b_1}{\mu_h}}{\sqrt{\rho_h}} - 1$$
$$= \frac{1-\mu_h-\mu_\ell b_1-\lambda-\mu_h+2\sqrt{\lambda_h\mu_h}}{\mu_h\sqrt{\rho_h}} - 1 > 0,$$

since $(\lambda + \mu_h - 2\sqrt{\lambda_h\mu_h})(\mu_h - \mu_\ell - \sqrt{\lambda_h\mu_h}) + \lambda_\ell\mu_\ell < 0$ implies $\mu_\ell - \mu_h + \sqrt{\lambda_h\mu_h} > \frac{\lambda_\ell\mu_\ell}{\lambda+\mu_h-2\sqrt{\lambda_h\mu_h}} = \mu_\ell b_1$. □

The exact tail asymptotic in the marginal distribution of the low-priority queue is characterized by Lemma 3.1(2), which is proved below:

Proof of Lemma 3.1(2) It follows from

$$P(1, y) = \frac{\mu_\ell}{\lambda_\ell y}\big[\psi_0(y) - (1-\rho)\big].$$

□

6 Exact tail asymptotics for the total number of customers in the system

In this section, we characterize the tail asymptotics for the distribution of the total number of customers in the system. We first prove the following lemma.

Lemma 6.1 *$\overline{\rho}_h > b_1$ if and only if $\overline{\rho}_h \neq \sqrt{\rho_h}$.*

Proof This is clear from $\frac{\lambda}{\mu_h} > b_1$ if and only if $(\frac{\lambda}{\mu_h} - \sqrt{\rho_h})^2 > 0$. □

Remark 6.1 The three regions: $D > 0$, $D = 0$ and $D < 0$ are further partitioned into subregions according to whether or not $\overline{\rho}_h < 1$.

The tail asymptotics of the distribution $\pi_{i,j}$ along the direction of the total number of customers in the system are characterized in Theorem 3.3. Let

$$C_{t,1a} = \frac{(\mu_h - \mu_\ell)\eta_1}{\mu_h|\eta_1 - \overline{\rho}_h|}C_{\ell,1}, \tag{6.1}$$

$$C_{t,1b} = C_{t,2b} = C_{t,3b} = \frac{\mu_\ell(1-\rho)}{\mu_h} + \frac{(\mu_h - \mu_\ell)}{\mu_h}\psi_0\left(\frac{1}{\overline{\rho}_h}\right), \tag{6.2}$$

$$C_{t,1c} = \frac{(\mu_h - \mu_\ell)}{\mu_h}C_{\ell,1}, \tag{6.3}$$

$$C_{t,2a} = \frac{(\mu_\ell - \mu_h)}{\mu_h}\frac{b_1}{\overline{\rho}_h - b_1}C_{\ell,2}, \tag{6.4}$$

$$C_{t,2c} = \frac{2(\mu_h - \mu_\ell)}{\mu_h} C_{\ell,2}, \tag{6.5}$$

$$C_{t,3a} = \frac{(\mu_\ell - \mu_h)}{\mu_h} \frac{b_1}{\overline{\rho}_h - b_1} C_{\ell,3}. \tag{6.6}$$

Proof of Theorem 3.3 According to (2.6) of Proposition 2.1,

$$P(x,x) = \frac{\mu_\ell(1-\rho)}{\mu_h[1-(\lambda/\mu_h)x]} + \frac{(\mu_h - \mu_\ell)}{\mu_h[1-(\lambda/\mu_h)x]} \psi_0(x).$$

The rest of the proof follows from Lemma 6.1, Theorem 3.2 for the case of $i = 0$, Lemma 4.1, and detailed calculations. □

7 Exact tail asymptotics for the high-priority queue

In this section, we establish the results in Lemma 3.1(1) and Theorem 3.1. The former is well known. Therefore, we only need to characterize the exact tail asymptotics in the joint distribution along the high-priority queue direction.

To characterize the tail asymptotics in $\pi_{i,j}$ for a fixed $j \geq 0$, we analyze the generating function $\varphi_j(x)$ given in (2.13) and (2.14). First, we give a relationship between r_0 and r_1, which are defined in (2.12) and (3.1), respectively.

Lemma 7.1

$$\frac{\lambda_\ell}{\lambda_h} \frac{r_0}{1 - x_1 r_0} = \frac{r_0}{r_1},$$

where x_1 is defined in (2.10).

Proof It follows from the definition of r_0, r_1 and x_1, and elementary manipulations. □

The proof of Theorem 3.1 is based on the following lemma and Lemma 4.1.

Lemma 7.2 *For $j \geq 0$,*

$$\varphi_j(x) \sim (1-\rho)\left(\frac{r_1}{r_0}\right)^j \frac{1}{(1-r_0 x)^{j+1}}, \quad \text{as } r_0 x \to 1. \tag{7.1}$$

Proof We use the induction to prove the lemma. Since $\varphi_0(x) = \frac{1-\rho}{1-r_0 x}$, (7.1) is true for $j = 0$. Assume that (7.1) is true for $j = k$, we then show it is also true for $j = k+1$. From (2.14),

$$\varphi_{k+1}(x) = \frac{a_{k+1}}{1 - r_0 x} + \frac{\lambda_\ell r_0 x}{\lambda_h (1 - r_0 x)} \frac{\varphi_k(x) - \varphi_k(x_1)}{x - x_1},$$

where a_{k+1} is a constant depending on k. Since $0 < x_1 < 1 < 1/r_0$, we have, according to the inductive assumption and Lemma 7.1, that

$$\lim_{xr_0\to 1}\frac{\varphi_{k+1}(x)}{(1-r_0x)^{-(k+2)}}$$
$$=\lim_{xr_0\to 1}\left[a_{k+1}(1-r_0x)^{k+1}+\frac{\lambda_\ell r_0x}{\lambda_h}\frac{(1-r_0x)^{k+1}\varphi_k(x)-(1-r_0x)^{k+1}\varphi_k(x_1)}{x-x_1}\right]$$
$$=\frac{\lambda_\ell}{\lambda_h}\frac{(1-\rho)(\frac{r_1}{r_0})^k}{1/r_0-x_1}=(1-\rho)\left(\frac{r_1}{r_0}\right)^{k+1},$$

which is equivalent to (7.1) for $j=k+1$. This completes the proof of the lemma. □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1 We use induction to prove the theorem.

By the induction and (2.14), we can easily obtain that $\varphi_j(x)$, $j\geq 0$, is analytic in the region $\Delta(\phi,\varepsilon)=\{x:|r_0x|\leq 1+\varepsilon,\ |\mathrm{Arg}\,(r_0x-1)|\geq\phi,\ \varepsilon>0,\ 0<\phi<\pi/2\}$ except at $r_0x=1$. By Lemmas 4.1 and 7.2, we have

$$\frac{C_n[\varphi_j(x)]}{r_0^n}\sim(1-\rho)\left(\frac{r_1}{r_0}\right)^j\frac{n^{(j+1)-1}}{\Gamma(j+1)}=(1-\rho)\left(\frac{r_1}{r_0}\right)^j\frac{n^j}{j!},\quad j\geq 0,$$

that is

$$\pi_{n,j}\sim(1-\rho)\left(\frac{r_1^j}{j!}\right)n^jr_0^{n-j},\quad j\geq 0,$$

which completes the proof. □

8 Concluding remarks

In this paper, for the classical preemptive priority queueing system with two types of customers we provided a complete characterization of exact tail asymptotics for the following cases.

(1) Along the high-priority queue direction, which is a type of exact tail asymptotic different from the three types reported in the literature under the condition of irreducibility. This example provides motivation to further investigate the behaviour of the tail along a coordinate direction for the double QBD process if the irreducible condition is not satisfied.
(2) Along the low-priority queue direction. Our study is an extension and the completion of the previous research, for both symmetric and asymmetric service times. This case serves as an example that reveals all three types of exact tail asymptotics when the irreducible condition is satisfied. The singularity analysis enabled us to better understand how these three types arise, which provides us with a possible new angle from which to have a new or better understanding of concepts, such as α-positivity, recurrence and transience, or large deviation paths, or geometric properties of the impact of the boundary behaviour on the asymptotics,

in other methods. It would be interesting to further detail those possible relationships. Also, this example provides many insightful properties, which could be independent of this specific model and lead us to a characterization of exact tail asymptotics along a coordinate direction for a general double QBD process. Details of such a characterization are not available in the literature.

(3) For the total number of customers in the systems, we found that a total of five types of exact tail asymptotics exist, which is not surprising since this is not the same as the tail asymptotic problem along the main diagonal direction in the double QBD process, which can only have three types of tail asymptotics in general.

(4) In the marginal distribution for the queue of low-priority customers. The result in this case is consistent with the directional asymptotics for the low-priority customer queue.

References

1. Abate, J., Whitt, W.: Asymptotics for $M/G/1$ low-priority waiting-time tail probabilities. Queueing Syst. **25**, 173–233 (1997)
2. Alfa, A.S.: Matrix-geometric solution of discrete time $MAP/PH/1$ priority queue. Nav. Res. Logist. **45**, 23–50 (1998)
3. Alfa, A.S., Liu, B., He, Q.-M.: Discrete-time analysis of $MAP/PH/1$ multiclass general preemptive priority queue. Nav. Res. Logist. **50**, 662–682 (2003)
4. Bender, E.: Asymptotic methods in enumeration. SIAM Rev. **16**, 485–513 (1974)
5. Borovkov, A.A., Mogul'skii, A.A.: Large deviations for Markov chains in the positive quadrant. Russ. Math. Surv. **56**, 803–916 (2001)
6. Bousquet-Mélou, M.: Walks in the quarter plane: Kreweras' algebraic model. Ann. Appl. Probab. **15**, 1451–1491 (2005)
7. Cohen, J.W., Boxma, O.J.: Boundary Value Problems in Queueing System Analysis. North-Holland, Amsterdam (1983)
8. Delas, S., Mazumdar, R.R., Rosenberg, C.P.: Tail asymptotics for HOL priority queues handling a large number of independent stationary sources. Queueing Syst. **40**, 183–204 (2002)
9. Drekic, S., Woolford, D.G.: A preemptive priority queue with balking. Eur. J. Oper. Res. **164**, 387–401 (2005)
10. Fayolle, G., Iasnogorodski, R.: Two coupled processors: the reduction to a Riemann–Hilbert problem. Z. Wahrscheinlichkeitstheor. Verw. Geb. **47**, 325–351 (1979)
11. Fayolle, G., King, P.J.B., Mitrani, I.: The solution of certain two-dimensional Markov models. Adv. Appl. Probab. **14**, 295–308 (1982)
12. Fayolle, G., Iasnogorodski, R., Malyshev, V.: Random Walks in the Quarter-Plane. Springer, New York (1991)
13. Flajolet, P., Odlyzko, A.: Singularity analysis of generating functions. SIAM J. Discrete Math. **3**, 216–240 (1990)
14. Flatto, L.: Two parallel queues created by arrivals with two demands II. SIAM J. Appl. Math. **45**, 861–878 (1985)
15. Flatto, L., McKean, H.P.: Two queues in parallel. Commun. Pure Appl. Math. **30**, 255–263 (1977)
16. Flatto, L., Hahn, S.: Two parallel queues created by arrivals with two demands I. SIAM J. Appl. Math. **44**, 1041–1053 (1984)
17. Foley, R.D., McDonald, D.R.: Join the shortest queue: stability and exact asymptotics. Ann. Appl. Probab. **11**, 569–607 (2001)
18. Foley, R.D., McDonald, R.D.: Large deviations of a modified Jackson network: stability and rough asymptotics. Ann. Appl. Probab. **15**, 519–541 (2005)
19. Foley, R.D., McDonald, R.D.: Bridges and networks: exact asymptotics. Ann. Appl. Probab. **15**, 542–586 (2005)
20. Gail, H.R., Hantler, S.L., Taylor, B.A.: Analysis of a non-preemptive priority multiserver queue. Adv. Appl. Probab. **20**, 852–879 (1988)

21. Gail, H.R., Hantler, S.L., Taylor, B.A.: On preemptive Markovian queue with multiple servers and two priority classes. Math. Oper. Res. **17**, 365–391 (1992)
22. Haque, L.: Tail behaviour for stationary distributions for two-dimensional stochastic models. Ph.D. Thesis, Carleton University, Ottawa, ON, Canada (2003)
23. Haque, L., Liu, L., Zhao, Y.Q.: Sufficient conditions for a geometric tail in a QBD process with countably many levels and phases. Stoch. Models **21**(1), 77–99 (2005)
24. He, Q., Li, H., Zhao, Y.Q.: Light-tailed behaviour in QBD process with countably many phases. Stoch. Models **25**, 50–75 (2009)
25. Hou, Q.-H., Mansour, T.: Kernel method and linear recurrence system. J. Comput. Appl. Math. **216**, 227–242 (2008)
26. Isotupa, K.P.S., Stanford, D.A.: An infinite-phase quasi-birth-and-death model for the non-preemptive priority $M/PH/1$ queue. Stoch. Models **18**, 378–410 (2002)
27. Kao, E.P.C., Narayanan, K.S.: Computing steady-state probabilities of a non-preemptive priority multiserver queue. ORSA J. Comput. **2**, 211–218 (1990)
28. Kroese, D.P., Scheinhardt, W.R.W., Taylor, P.G.: Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. Ann. Appl. Probab. **14**(4), 2057–2089 (2004)
29. Kurkova, I.A., Suhov, Y.M.: Malyshev's theory and JS-queues. Asymptotics of stationary probabilities. Ann. Appl. Probab. **13**, 1313–1354 (2003)
30. Li, L., Miyazawa, M., Zhao, Y.: Geometric decay in a QBD process with countable background states with applications to a join-the-shortest-queue model. Stoch. Models **23**, 413–438 (2007)
31. Lieshout, P., Mandjes, M.: Asymptotic analysis of Lévy-driven tandem queues. Queueing Syst. **60**, 203–226 (2008)
32. Maertens, T., Walraevens, J., Bruneel, H.: Priority queueing systems: from probability generating functions to tail probabilities. Queueing Syst. **55**, 27–39 (2007)
33. Malyshev, V.A.: An analytical method in the theory of two-dimensional positive random walks. Sib. Math. J. **13**, 1314–1329 (1972)
34. Malyshev, V.A.: Asymptotic behaviour of stationary probabilities for two-dimensional positive random walks. Sib. Math. J. **14**, 156–169 (1973)
35. Mandjes, M.: Large Deviations for Gaussian Queues: Modelling Communication Networks. Wiley, New York (2007)
36. McDonald, D.R.: Asymptotics of first passage times for random walk in an orthant. Ann. Appl. Probab. **9**, 110–145 (1999)
37. Miller, D.R.: Computation of steady-state probabilites for $M/M/1$ priority queues. Oper. Res. **29**(5), 945–958 (1981)
38. Mishna, M.: Classifying lattice walks restricted to the quarter plane. J. Comb. Theory Ser. A **116**, 460–477 (2009)
39. Miyazawa, M.: The Markov renewal approach to $M/G/1$ type queues with countably many background states. Queueing Syst. **46**, 177–196 (2004)
40. Miyazawa, M.: Doubly QBD process and a solution to the tail decay rate problem. In: Proceedings of the Second Asia-Pacific Symposium on Queueing Theory and Network Applications, Kobe, Japan (2007)
41. Miyazawa, M.: Two sided DQBD process and solutions to the tail decay rate problem and their applications to the generalized join shortest queue. In: Yue, W., Takahashi, Y., Takagi, H. (eds.) Advances in Queueing Theory and Network Applications, pp. 3–33. Springer, New York (2009)
42. Miyazawa, M.: Tail decay rates in double QBD processes and related reflected random walks. Math. Oper. Res. **34**(3), 547–575 (2009)
43. Miyazawa, M., Zhao, Y.Q.: The stationary tail asymptotics in the $GI/G/1$ type queue with countably many background states. Adv. Appl. Probab. **36**(4), 1231–1251 (2004)
44. Morrison, J.A.: Processor sharing for two queues with vastly different rates. Queueing Syst. **57**, 19–28 (2007)
45. Motyer, A.J., Taylor, P.G.: Decay rates for quasi-birth-and-death process with countably many phases and tri-diagonal block generators. Adv. Appl. Probab. **38**, 522–544 (2006)
46. Sleptchenko, A., Adan, I.J.B.F., van Houtum, G.J.: Joint queue length distribution of multi-class, single-server queues with preemptive priorities. Preprint (2004)
47. Takahashi, Y., Fujimoto, K., Makimoto, N.: Geometric decay of the steady-state probabilities in a quasi-birth-and-death process with a countable number of phases. Stoch. Models **17**(1), 1–24 (2001)
48. Takine, T.: A nonpreemptive priority $MAP/G/1$ queue with two classes of customers. J. Oper. Res. Soc. Jpn. **39**, 266–290 (1996)

49. van Uitert, M.J.G.: Generalized processor sharing queues. Ph.D. Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands (2003)
50. Wischik, D.: Sample path large deviations for queues with many inputs. Ann. Appl. Probab. **11**, 379–404 (2001)
51. Wright, P.: Two parallel processors with coupled inputs. Adv. Appl. Probab. **24**, 986–1007 (1992)
52. Xue, J., Alfa, A.S.: Tail probability of low-priority queue length in a discrete-time priority $BMAP/PH/1$ queue. Stoch. Models **21**, 799–820 (2005)
53. Zhao, J.-A., Li, B., Cao, X.-R., Ahmad, I.: A matrix-analytic solution for the $DBMAP/PH/1$ priority queue. Queueing Syst. **53**, 127–145 (2006)