Queueing Syst (2009) 61: 305–328
DOI 10.1007/s11134-009-9110-4

Waiting and sojourn times in a multi-server queue with mixed priorities

Sergey Zeltyn · Zohar Feldman · Segev Wasserkrug

Received: 3 February 2008 / Revised: 25 January 2009 / Published online: 5 March 2009

Abstract We consider a multi-server queue with K priority classes. In this system, customers of the P highest priorities ($P < K$) can preempt customers with lower priorities, ejecting them from service and sending them back into the queue. Service times are assumed exponential with the same mean for all classes.

The Laplace–Stieltjes transforms of waiting times are calculated explicitly and the Laplace–Stieltjes transforms of sojourn times are provided in an implicit form via a system of functional equations. In both cases, moments of any order can be easily calculated. Specifically, we provide formulae for the steady state means and the second moments of waiting times for all priority classes. We also study some approximations of sojourn-time distributions via their moments. In a practical part of our paper, we discuss the use of mixed priorities for different types of Service Level Agreements, including an example based on a real scheduling problem of IT support teams.

Keywords Queues · Priority queues · Workforce management · IT support systems · Contact centers

Mathematics Subject Classification (2000) 60K25 · 90B22 · 68M20

1 Introduction

1.1 Motivation

Many service systems differentiate among customers or jobs according to their business value or other factors. As a result, different classes of customers are subject to

S. Zeltyn (✉) · Z. Feldman · S. Wasserkrug
IBM Haifa Research Lab, Haifa University, Mount Carmel, Haifa 31905, Israel
e-mail: sergeyz@il.ibm.com

different Service Level Agreements (SLAs) that guarantee a certain service level for each customer class and often include penalties for their violation.

Customer differentiation gives rise to queues with two main types of priority disciplines: *preemptive* and *non-preemptive*. In the *preemptive* queues, customers with higher priorities can eject lower-priority customers from service. According to the *non-preemptive* discipline, the service that started should be completed without interruption.

Most research on priority queues has been dedicated to either pure non-preemptive priorities (service is never interrupted) or to pure preemptive service disciplines (any customer of higher priority can eject from service any customer of lower priority). However, *systems with mixed priorities* that combine the two disciplines are widespread in various application areas, such as contact centers, health care, and communication networks.

For example, *Information Technology (IT) support systems* have recently enjoyed significant growth both in the volume of operations and in the workforce employed. IT support systems can perform either relatively simple operations (known as first-level support, traditionally provided via call centers) or operations that require more advanced skills and extended service times (second and third-level support). A typical priority discipline in the second and third-level IT support systems is the mixed discipline that is analyzed in our paper. Specifically, this research resulted from a practical scheduling problem for IT support teams in Bangalore, India. A comprehensive description of this problem and practical methods applied in the scheduling solution are provided in Wasserkrug et al. [25]. See Sect. 3.4 for a numerical example based on this project.

In modern *contact centers*, classical telephone-based service is combined with various Internet services, such as chat and mail (Gans et al. [11]). In such centers, we expect non-preemptive priority discipline to be applied between different classes of phone calls and preemptive discipline between high-priority phone calls and non-urgent Internet services.

Additional examples of service systems in which mixed priority service disciplines are relevant include *health care* and *communication networks*. For example, if a health-care system is experiencing heavy load at the stage of initial treatment (for example, in an emergency room or near a battlefield), patients or wounded are divided into classes according to the severity of their problems. In this case, service discipline can be either preemptive or non-preemptive, depending on the classes involved.

From a system design point of view, mixed priorities provide us with flexibility to satisfy multiple SLAs. Specifically, one can choose between pure preemptive, pure non-preemptive, and different mixed service protocols. In addition, mixed priorities are often applied if pure preemptive priorities are preferable according to the SLA, but it is desirable to decrease the number of service interruptions.

1.2 Contribution of the paper

We consider a multi-server system with n servers, K classes of customers, and *mixed priorities*. Class 1 customers have the highest priority and, in general, class i customers have higher priorities than class j customers if $i < j$. Customers of classes

$1, \ldots, P$ are named *preempting* since they can eject lower-priority customers from service. Customers of classes $P+1, \ldots, K$ are named *non-preempting* since they can enter service only if an idle server is available. Assume $K \geq 3$ and $1 \leq P \leq K-2$. The precise description of our service/preemption protocol is provided in Sect. 1.3.

The arrival processes for all classes are independent Poisson processes and the service distribution is exponential with the same mean for all classes. (The same service-time distribution can be a reasonable assumption if all customers are engaged in the same type of service activities and divided into classes according to their business value or urgency of service.)

Our paper provides the following contributions to the theory and applications of priority queues:

- Laplace–Stieltjes Transforms (LSTs) for waiting times of all classes are computed explicitly, giving rise to straightforward calculations of their moments. In particular, we provide a simple formula for mean waiting and sojourn times and derive the second moments of the waiting times.
- In this paper, LSTs for sojourn times are computed implicitly via systems of $(n+1)$ functional equations. The moments of sojourn times are then calculated via systems of $(n+1)$ linear equations. In addition, Sect. 2.3.3 outlines an alternative method of LSTs calculation for sojourn times.
- In the practical part of the paper, we provide some instructive numerical examples and discuss which types of SLAs could give rise to mixed priorities protocols. We study mean waiting times, waiting-time distributions (via LST inversion), and an approximation of the sojourn-time distribution by three moments using the algorithm of Osogami and Harchol-Balter [18].

Remark 1 As we shall see from the detailed description of our service protocol in Sect. 1.3, non-preempting classes $P+1, \ldots, K$ do not affect performance of preempting classes $1, \ldots, P$. Therefore, the performance measures for preempting classes coincide with the ones in the pure preemptive case that have already been analyzed by Tatashev [24] and Segal [21]. Our analysis on preempting classes is mainly based on the above results, providing some extensions, such as explicit expressions for the second moments of the waiting times or the discussion in Sect. 2.3.3. The results for non-preempting classes are new and provide the main theoretical contribution of this paper.

1.3 Service/preemption protocol

In general, we divide customers in the system into three types: *customers in service*, *ejected customers* (who started service, but have been preempted and are currently waiting for service restart), and *waiting customers* (who have not yet started service). To define the service/preemption protocol in a unique way, we must address two key questions: First, which customer is ejected if a preempting customer arrives at the system and all servers are busy? Second, which customer is taken from the queue by a server that finishes service?

Below we describe the service/preemption protocol used in this paper.

Assume that a preempting customer (classes $1, \ldots, P$) arrives at the system and there are customers of lower priorities in service.

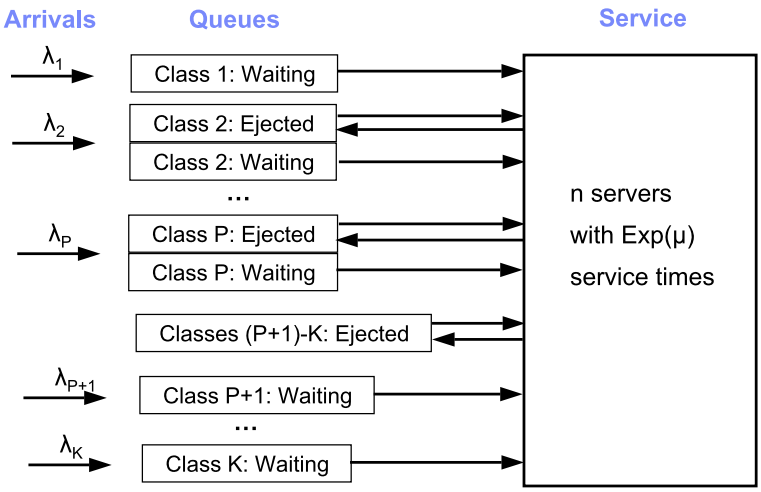

Fig. 1 Schematic representation of our service protocol

- If there are non-preempting customers in service (classes $P+1,\ldots,K$), then the last non-preempting customer who started service is ejected, regardless of priority.
- If there are no non-preempting customers in service, the preempting customer with the lowest priority is ejected from service. If there are several preempting customers of this class, the one that started service last is ejected.

Assume that a server finishes service and there are ejected or waiting customers in the system.

- If there are preempting customers in the queue, the customer with the highest priority is served. If there are several ejected customers of this priority, the customer who was ejected last is taken to service. Otherwise, a waiting customer of the highest priority is served according to FCFS (First Come First Served).
- If there are no preempting customers in the queue, but there are ejected non-preempting customers, then the customer who was ejected last is served, regardless of priority.
- Otherwise, if there are only waiting non-preempting customers in the queue, a waiting customer of the highest priority is served according to FCFS.

Figure 1 depicts the hierarchy of queues in our system.

Discussion on the service protocol First, in our protocol, customers of the same class are ejected according to the shortest time in service. In many applications, this is reasonable since breaks at early stages of service are preferable to breaks at later stages. These assumptions are also consistent with several papers cited below in Sect. 1.4. (See, for example, [21].) Second, we assume that ejected customers have priority over waiting customers of the same class, since it is preferable to finish jobs that were previously started before taking the new ones to service. Third, our protocol is "fair" in the sense that two ejected customers of the same class restart service

according to the order of their arrival. (A customer who arrives first starts service first among the two customers, is ejected after the second one, and restarts service before the second one.) Finally, we assume that ejected non-preempting customers are treated as a single class, once they start service for the first time. Therefore, the difference between non-preempting classes arises only at the waiting stage. Alternative protocols that distinguish between non-preempting customers at the later stages of service/ejection should be studied too. The choice of a specific protocol can depend on a specific SLA. (See Remark 6 for details.)

1.4 Related work

The research on priority queues was initiated by Cobham [5] more than half a century ago. Classical results on non-preemptive discipline with identically distributed exponential service times for all classes can be found in Davis [6] and Kella and Yechialy [16]. The corresponding results for preemptive priorities are presented in Segal [21], Buzen and Bondi [3], and Tatashev [24].

If the service times have different means and/or are non-exponential, the problem becomes much more complicated. Gail et al. [9, 10] consider both preemptive and non-preemptive disciplines in the case of two classes and exponential service times with different means. Harchol-Balter et al. [13] provide an approximate analysis of a preemptive system with the phase-type service distributions. Finally, Sleptchenko et al. [23] analyze preemptive system with two classes and hyperexponential service times, where some performance measures for hyperexponential subclasses are derived.

Literature on systems with mixed priorities is more scarce than that for pure preemptive or non-preemptive disciplines. Some single-server results are available in Chang [4], Adiri and Domb [2], Simon [22], and Drekic and Stanford [7], but no results for multi-server systems exist, as far as we know.

Finally, distributions of waiting and sojourn times in the priority queues are often derived via their Laplace–Stieltjes transforms. Due to the increase in the computational power of modern computers, inversion of Laplace–Stieltjes transforms is now a much more feasible task than it was in the past. See Abate and Whitt [1] and Appendix A from Jagerman and Melamed [15] that discuss numerical Laplace inversions and several specific algorithms.

2 Theoretical results

2.1 Notation

This section includes a list of basic notation and definitions, used in this paper.

- $\lambda_1, \ldots, \lambda_K$—Poisson arrival rates associated with classes $1, \ldots, K$.
- $\bar{\lambda}_k = \sum_{j=1}^{k} \lambda_j$, $1 \le k \le K$,—aggregated arrival rates associated with classes $1, \ldots, k$.
- $\rho_k = \lambda_k/(\mu n)$, $1 \le k \le K$,—servers' utilization associated with class k.

- $\bar{\rho}_k = \sum_{j=1}^{k} \rho_j$, $1 \le k \le K$,—aggregated utilization associated with class k. We assume that the stability condition $\bar{\rho}_K < 1$ prevails.
- W_k, $1 \le k \le K$—steady-state waiting time of class k. (The waiting time includes the wait before the first service start and also possible wait after ejections from service.)
- V_k, $1 \le k \le K$—steady-state sojourn time of class k. (The sojourn time includes waiting and service times, where the service time can have several phases, due to ejections.)
- $\tilde{V}_k(s)$, $\tilde{W}_k(s)$, $1 \le k \le K$—LST of sojourn and waiting times, respectively. For example, $\tilde{W}_k(s) = \int_0^\infty e^{-st}\, dF_{W_k}(t)$, where $F_{W_k}(\cdot)$ is the corresponding cumulative distribution function.
- M/M/$n(\lambda;\mu)$—the M/M/n queue with arrival rate λ and service rate μ.
- M/M/$n/n(\lambda;\mu)$—the M/M/n/n loss system with arrival rate λ and service rate μ.
- M/M/$n(\lambda_1,\ldots,\lambda_K;\mu)_{np}$—a queueing system with K classes, non-preemptive priorities, service rate μ and vector of Poisson arrival rates $\lambda_1,\ldots,\lambda_K$.
- M/M/$n(\lambda_1,\ldots,\lambda_K;\mu)_{pr}$—a queueing systems with preemptive priorities and parameters that coincide with the ones from the previous definition.
- M/M/$n(\lambda_1,\ldots,\lambda_K;\mu;P)_{mx}$—a queueing systems with mixed priorities and P preempting classes.

2.2 Waiting times

We start with several additional definitions. It is well known [21] that the LST of a busy-period length for M/M/$n(\bar{\lambda}_k,\mu)$ is given by

$$\tilde{B}_k(s) = \frac{s+\bar{\lambda}_k+n\mu-\sqrt{(s+\bar{\lambda}_k+n\mu)^2-4\bar{\lambda}_k n\mu}}{2\bar{\lambda}_k}, \quad 1 \le k \le K. \tag{2.1}$$

Now consider the M/M/$n(\lambda_1,\ldots,\lambda_k,\ldots,\lambda_l;\mu)_{np}$ system with $l \ge k$. The LST of the density of conditional waiting time $(W_k|W_k>0)$ for class k is equal to [16]:

$$\tilde{f}_k^0(s) = \frac{2\cdot(n\mu)\cdot(1-\bar{\rho}_k)}{s+\bar{\lambda}_{k-1}+n\mu-2\bar{\lambda}_k+\sqrt{(s+\bar{\lambda}_{k-1}+n\mu)^2-4\bar{\lambda}_{k-1}n\mu}}, \quad 1 \le k \le K. \tag{2.2}$$

(Let here and in continuation denote $\bar{\lambda}_0 = 0$, $\bar{\rho}_0 = 0$.) Note that (2.2) does not depend on the arrival rates of classes with priorities that are lower than k.

Define the loss probability in M/M/$n/n(\bar{\lambda}_k,\mu)$ by

$$E_k^B = \frac{(\bar{\lambda}_k/\mu)^n/n!}{\sum_{j=0}^{n}(\bar{\lambda}_k/\mu)^j/j!}, \quad 1 \le k \le K, \qquad E_0^B = 1, \tag{2.3}$$

and the delay probability in M/M/n $(\bar{\lambda}_k,\mu)$ by

$$E_k^C = \left[1+\sum_{j=0}^{n-1}\frac{n!(1-\bar{\rho}_k)}{j!(\bar{\lambda}_k/\mu)^{n-j}}\right]^{-1}, \quad 1 \le k \le K.$$

Finally, introduce the following constants (their intuitive meaning is explained in Statements 3 and 4 in Sect. 2.2.1):

$$p_k^0 = 1 - \frac{\bar{\rho}_{k-1}(E_k^B - E_{k-1}^B)}{\rho_k(1 - E_k^B)}, \quad 1 \le k \le P, \tag{2.4}$$

$$p_{np}^0 = 1 - \frac{\bar{\rho}_P(E_K^B - E_P^B)}{(\bar{\rho}_K - \bar{\rho}_P)(1 - E_K^B)}, \tag{2.5}$$

$$p_k = 1 - \bar{\rho}_{k-1}\big(1 - E_{k-1}^B\big), \quad 1 \le k \le P, \tag{2.6}$$

$$p_{np} = 1 - \bar{\rho}_P\big(1 - E_P^B\big). \tag{2.7}$$

Theorem 1 *For the preempting classes* $(1 \le k \le P)$*, the LST of the waiting time is given by*

$$\tilde{W}_k(s) = \big(1 - E_k^C\big) \cdot p_k^0 + \frac{E_k^C p_k \tilde{f}_k^0(s) + (1 - E_k^C) p_k (1 - p_k^0) \tilde{B}_{k-1}(s)}{1 - (1 - p_k)\tilde{B}_{k-1}(s)}. \tag{2.8}$$

For the non-preempting classes $(P + 1 \le k \le K)$*, the LST of the waiting time is given by*

$$\tilde{W}_k(s) = \big(1 - E_K^C\big) \cdot p_{np}^0 + \frac{E_K^C p_{np} \tilde{f}_k^0(s) + (1 - E_K^C) p_{np} (1 - p_{np}^0) \tilde{B}_P(s)}{1 - (1 - p_{np})\tilde{B}_P(s)}. \tag{2.9}$$

Corollary 1 The expected values of the waiting times are equal to

$$\mathrm{E}[W_k] = \frac{1}{n\mu} \cdot \left[\frac{\bar{\rho}_k}{\rho_k} \frac{E_k^C}{1 - \bar{\rho}_k} - \frac{\bar{\rho}_{k-1}}{\rho_k} \frac{E_{k-1}^C}{1 - \bar{\rho}_{k-1}} \right], \quad 1 \le k \le P, \tag{2.10}$$

$$\mathrm{E}[W_k] = \frac{1}{n\mu} \cdot \left[\frac{E_K^C}{(1 - \rho_{k-1})(1 - \bar{\rho}_k)} + \frac{\bar{\rho}_P(E_K^C - E_P^C)}{(1 - \bar{\rho}_P)(\bar{\rho}_K - \bar{\rho}_P)} \right], \quad P + 1 \le k \le K. \tag{2.11}$$

The second moments of the waiting times are calculated via

$$\mathrm{E}[W_k]^2 = \frac{2}{(n\mu)^2} \cdot \left[\frac{E_k^C (1 - \bar{\rho}_{k-1}\bar{\rho}_k)}{(1 - \bar{\rho}_k)^2 (1 - \bar{\rho}_{k-1})^3} + \frac{E_k^C \bar{\rho}_{k-1}(1 - E_{k-1}^C)}{(1 - \bar{\rho}_k)(1 - \bar{\rho}_{k-1})^3} + \frac{(1 - p_k^0 + (p_k^0 - p_k)E_k^C)(1 - \bar{\rho}_{k-1} + p_k\bar{\rho}_{k-1})}{p_k^2 (1 - \bar{\rho}_{k-1})^3} \right], \quad 1 \le k \le P, \tag{2.12}$$

$$\mathrm{E}[W_k]^2 = \frac{2}{(n\mu)^2} \cdot \left[\frac{E_K^C (1 - \bar{\rho}_{k-1}\bar{\rho}_k)}{(1 - \bar{\rho}_k)^2 (1 - \bar{\rho}_{k-1})^3} + \frac{E_K^C \bar{\rho}_P (1 - E_P^C)}{(1 - \bar{\rho}_P)^2 (1 - \bar{\rho}_k)(1 - \bar{\rho}_{k-1})} + \frac{(1 - p_{np}^0 + (p_{np}^0 - p_{np})E_K^C)(1 - \bar{\rho}_P + p_{np}\bar{\rho}_P)}{p_{np}^2 (1 - \bar{\rho}_P)^3} \right], \quad P + 1 \le k \le K. \tag{2.13}$$

Remark 2 Higher moments of waiting times can be derived via differentiation of (2.8) and (2.9).

Remark 3 Formulae (2.8) and (2.10) are well known, see Tatashev [24] and Buzen and Bondi [3], respectively.

2.2.1 *Proof of Theorem 1*

Remark 4 Due to work-conservation and the assumption of the same service rate for all classes, the steady-state distribution of overall number-in-system in $\mathrm{M/M}/n(\lambda_1,\ldots,\lambda_K;\mu;P)_{mx}$ coincides with the steady-state distribution in $\mathrm{M/M}/n(\bar{\lambda}_K,\mu)$. Moreover, if we consider a preempting class $k \le P$, the distribution of number-in-system for classes $1,\ldots,k$ coincides with $\mathrm{M/M}/n(\bar{\lambda}_k,\mu)$.

Denote by I_k the steady-state number of service interruptions for class k, $1 \le k \le K$. (Of course, $I_1 \equiv 0$.) The waiting time of class k can then be represented via the following sum:

$$W_k = W_k(0) + W_k(1) + \cdots + W_k(I_k), \tag{2.14}$$

where $W_k(0)$ is the waiting time before the first service start and $W_k(i)$, $1 \le i \le I_k$, is the waiting time between the ith interruption and the next service start.

Below we show that the following four statements prevail.

Statement 1 *For the preempting classes $2 \le k \le P$, distribution of $W_k(i)$, $i \ge 1$, coincides with the distribution of busy-period length in $\mathrm{M/M}/n(\bar{\lambda}_{k-1},\mu)$ with LST given by $\tilde{B}_{k-1}(s)$ in (2.1). For the non-preempting classes $P+1 \le k \le K$, it coincides with the distribution of busy-period length in $\mathrm{M/M}/n(\bar{\lambda}_P,\mu)$ with LST given by $\tilde{B}_P(s)$.*

Statement 2 *For the preempting classes $1 \le k \le P$, distribution of $W_k(0)$ coincides with the distribution of the kth class waiting time in $\mathrm{M/M}/n(\lambda_1,\ldots,\lambda_k;\mu)_{np}$. A corresponding random variable is zero with probability $1-E_k^C$ and has LST $\tilde{f}_k^0(s)$, provided in formula (2.2), otherwise. For the non-preempting classes $P+1 \le k \le K$, distribution of $W_k(0)$ coincides with the distribution of the kth class waiting time in $\mathrm{M/M}/n(\lambda_1,\ldots,\lambda_k,\ldots,\lambda_K;\mu)_{np}$. (A corresponding random variable is zero with probability $1-E_K^C$ and has LST $\tilde{f}_k^0(s)$, otherwise.)*

Statement 3 *The probability to finish service without interruption given the service has been started* after positive wait *is given by p_k from formula (2.6) for the preemptive classes and by p_{np} from (2.7) for the non-preempting classes. Formally,*

$$\mathrm{P}\{I_k = 0 \mid W_k(0) > 0\} = p_k \stackrel{\Delta}{=} 1 - \bar{\rho}_{k-1}\bigl(1 - E_{k-1}^B\bigr), \quad 2 \le k \le P, \tag{2.15}$$

$$\mathrm{P}\{I_k = 0 \mid W_k(0) > 0\} = p_{np} \stackrel{\Delta}{=} 1 - \bar{\rho}_P\bigl(1 - E_P^B\bigr), \quad P+1 \le k \le K. \tag{2.16}$$

Note that due to the memoryless properties of our system, the left part of (2.15) and (2.16) is also equal to $\mathrm{P}\{I_k = i+1 \mid I_k \ge i\}$, for $i \ge 0$.

Statement 4 *The steady-state probability to finish service without interruption given the service has been started* immediately upon arrival *is given by* p_k^0 *from formula* (2.4) *for the preempting classes, and by* p_{np}^0 *from* (2.5) *for the non-preempting classes. Formally,*

$$\mathrm{P}\{I_k = 0 \mid W_k(0) = 0\} = p_k^0, \quad 2 \le k \le P, \tag{2.17}$$

$$\mathrm{P}\{I_k = 0 \mid W_k(0) = 0\} = p_{np}^0, \quad P+1 \le k \le K. \tag{2.18}$$

We shall prove Statements 1–4 for the non-preempting classes. The proofs for preempting classes are similar.

Proof of Statement 1 Assume that a non-preempting customer is interrupted. According to the protocol in Sect. 1.3, this customer is then placed at the head of the queue of non-preemptive customers. At the moment of ejection, there are no preempting customers in the queue. (Otherwise, the customer in consideration would be ejected earlier.) Therefore, the customer restarts service once the number of service terminations after his/her ejection exceeds the number of new arrivals of preempting customers.

Formally, if we define by A_i^P interarrival times of preempting customers that are $\exp(\bar{\lambda}_P)$ distributed, and by S_i times between service terminations ($\exp(n\mu)$ distributed), then the waiting time till service restart is equal to $S_1 + \cdots + S_U$, where

$$U = \min\{i \ge 1 : S_1 + \cdots + S_i < A_1 + \cdots A_i^P\}.$$

To complete the proof, observe that $S_1 + \cdots + S_U$ is distributed exactly as the busy-period length in M/M/$n(\bar{\lambda}_P, \mu)$. □

Proof of Statement 2 Assume that a customer of priority k, $P+1 \le k \le K$, encounters wait (all servers are busy) upon arrival. Denote by Q_k the number of customers who are waiting in the queue upon arrival of our customer and who will enter service before him.

Using an argument that is similar to the one in the proof of Statement 1, it is easy to show that a busy-period of M/M/$n(\bar{\lambda}_{k-1}, \mu)$ should elapse before our customer moves one step forward in the queue. The waiting time distribution till first service start (conditional on positive wait) is then given by

$$W_k(0) \stackrel{d}{=} \sum_{i=1}^{Q_k+1} B_i, \tag{2.19}$$

where B_i are iid random variables with LST $\tilde{B}_{k-1}(s)$ given by (2.1).

Customers that constitute Q_k can be divided into two types:

Type 1: Customers of classes $1, \ldots, k$ that arrived before our customer and were waiting in the queue upon his/her arrival;

Type 2: Customers of classes $k+1, \ldots, K$ that were ejected and were waiting for service restart at the time of arrival of our customer.

Note that in the M/M/$n(\lambda_1,\ldots,\lambda_K;\mu)_{np}$ system, the queue consists of Type 1 customers only and the waiting time can be also represented via (2.19). In our M/M/$n(\lambda_1,\ldots,\lambda_K;\mu;P)_{mx}$ queue, every Type 2 customer was ejected by a preempting customer with the same exponential service distribution. Therefore, distribution of Q_k is the same for the two systems and Statement 2 prevails. □

Proof of Statement 3 Let q_i, $0 \le i \le n-1$, denote the probability that a non-preempting customer *does not* finish service without interruption given i customers in the system upon arrival. Conditioning on the next event (our customer finished service, some other customer finished service, arrival of a preemptive customer), we get the following system of equation for q_i:

$$\begin{cases} q_0 = \frac{\bar\lambda_P}{\bar\lambda_P+\mu} q_1, \\ q_1 = \frac{\bar\lambda_P}{\bar\lambda_P+2\mu} q_2 + \frac{\mu}{\bar\lambda_P+2\mu} q_0, \\ \cdots \\ q_{k-1} = \frac{\bar\lambda_P}{\bar\lambda_P+k\mu} q_k + \frac{(k-1)\mu}{\bar\lambda_P+k\mu} q_{k-2}, \\ \cdots \\ q_{n-1} = \frac{\bar\lambda_P}{\bar\lambda_P+n\mu} + \frac{(n-1)\mu}{\bar\lambda_P+n\mu} q_{n-2}. \end{cases} \tag{2.20}$$

Now define by $E_P^B(k)$, $0 \le k \le n$, the loss probability in M/M/$k/k(\bar\lambda_P,\mu)$. (Let $E_P^B(0)=0$ and note that $E_P^B(n) \stackrel{\Delta}{=} E_P^B$ by (2.3).) The solution of (2.20) is then given by

$$q_k = \frac{E_P^B}{E_P^B(k)}, \quad 0 \le k \le n-1. \tag{2.21}$$

To prove (2.21), define $\bar R_P = \bar\lambda_P/\mu$ and use the following relations between the M/M/n/n probabilities that can be easily derived from (2.3):

$$\frac{E_P^B(k)}{E_P^B(k-1)} = \frac{\bar R_P(1-E_P^B(k))}{k}, \quad k \ge 1, \tag{2.22}$$

$$\frac{E_P^B(k)}{E_P^B(k-2)} = \frac{\bar R_P \cdot (\bar R_P - (\bar R_P + k)E_P^B(k))}{(k-1)k}, \quad k \ge 2. \tag{2.23}$$

Now we can substitute (2.21) into (2.20) and check that the equations in (2.20) prevail. For example, the first equation is equivalent to $q_0 = \frac{\bar R_P}{\bar R_P+1}$ or $E_P^B(1) = \frac{\bar R_P}{1+\bar R_P}$, which immediately follows from the M/M/n/n loss probability definition.

Then equations in (2.20) are equivalent to

$$q_{k-1} = \frac{\bar R_P}{\bar R_P + k} q_k + \frac{k-1}{\bar R_P + k} q_{k-2}, \quad 1 \le k \le n. \tag{2.24}$$

(Define $q_n \stackrel{\Delta}{=} 1$.) Substituting (2.21) into (2.24), we get

$$\frac{E_P^B(k)}{E_P^B(k-1)} = \frac{\bar{R}_P}{\bar{R}_P + k} + \frac{k-1}{\bar{R}_P + k} \cdot \frac{E_P^B(k)}{E_P^B(k-2)}$$

that can be verified via (2.22) and (2.23).

Finally, a non-preempting customer who starts service after queueing has the same probability of finishing service without interruption, as the customer who observed $n-1$ customers in the system upon arrival. This probability is equal to

$$1 - q_{n-1} = 1 - \bar{\rho}_P\left(1 - E_P^B\right).$$

The last equality follows from (2.21), (2.22) and $\bar{\rho}_P = \bar{R}_P/n$. □

Proof of Statement 4 Assume that a non-preempting customer started service immediately. The probability that this customer will be interrupted at least once is then given by:

$$q_{np}^0 = \sum_{i=0}^{n-1} \frac{E_P^B(n)}{E_P^B(i)} \cdot \frac{\bar{R}_K^i/i!}{\sum_{j=0}^{n-1} \bar{R}_K^j/j!} \tag{2.25}$$

where the first term in the product is taken from (2.21) and the second term is the probability to encounter i customers in service upon arrival given the service started immediately. Then, using definition (2.3),

$$q_{np}^0 = \frac{\bar{R}_P^n/n!}{(\sum_{j=0}^{n} \bar{R}_P^j/j!)(\sum_{j=0}^{n-1} \bar{R}_K^j/j!)} \cdot \sum_{i=0}^{n-1}\left(\frac{\bar{R}_K}{\bar{R}_P}\right)^i \cdot \sum_{j=0}^{i} \frac{\bar{R}_P^j}{j!}.$$

Interchanging sums, we get

$$\begin{aligned} q_{np}^0 &= \frac{\bar{R}_P^n/n!}{(\sum_{j=0}^{n} \bar{R}_P^j/j!)(\sum_{j=0}^{n-1} \bar{R}_K^j/j!)} \cdot \sum_{j=0}^{n-1} \frac{\bar{R}_P^j}{j!} \cdot \sum_{i=j}^{n-1}\left(\frac{\bar{R}_K}{\bar{R}_P}\right)^i \\ &= \frac{\bar{R}_P^n/n!}{(\sum_{j=0}^{n} \bar{R}_P^j/j!)(\sum_{j=0}^{n-1} \bar{R}_K^j/j!)} \cdot \frac{\bar{R}_P}{\bar{R}_K - \bar{R}_P} \cdot \sum_{j=0}^{n-1} \frac{\bar{R}_P^j}{j!} \cdot \left(\left(\frac{\bar{R}_K}{\bar{R}_P}\right)^n - \left(\frac{\bar{R}_K}{\bar{R}_P}\right)^j\right) \\ &= \frac{\bar{\lambda}_P}{(\bar{\lambda}_K - \bar{\lambda}_P) \cdot (1 - E_K^B)} \cdot \frac{\bar{R}_P^n/n!}{(\sum_{j=0}^{n} \frac{\bar{R}_P^j}{j!})(\sum_{j=0}^{n} \frac{\bar{R}_K^j}{j!})} \cdot \left(\left(\frac{\bar{R}_K}{\bar{R}_P}\right)^n - \left(\frac{\bar{R}_K}{\bar{R}_P}\right)^j\right), \end{aligned} \tag{2.26}$$

where the last equation follows from definition (2.3). We can then calculate $(E_K^B - E_P^B)$ using (2.3) and check that it is equal to the product of the last two terms of (2.26). The last observation shows that

$$q_{np}^0 = \frac{\bar{\lambda}_P(E_K^B - E_P^B)}{(\bar{\lambda}_K - \bar{\lambda}_P)(1 - E_K^B)}$$

and completes the proof of Statement 4. □

Define by $Geom(p)$, $p > 0$, the geometric distribution with parameter p, supported on the set $\{0, 1, 2, \ldots\}$. Combining Statements 3 and 4 with independence between probabilities to finish service after the first service start and after interruptions $1, \ldots, I_k$ we get that:

- Distribution of $\{I_k \mid W_k(0) > 0\}$ is $Geom(p_k)$ for $2 \le k \le P$, and $Geom(p_{np})$ for $P + 1 \le k \le K$.
- Distribution of $\{I_k \mid W_k(0) = 0\}$, $2 \le k \le P$, is zero with probability p_k^0 and $1 + Geom(p_k)$ with probability $1 - p_k^0$. For $P + 1 \le k \le K$, distribution of $\{I_k \mid W_k(0) = 0\}$ is zero with probability p_{np}^0 and $1 + Geom(p_{np})$ with probability $1 - p_{np}^0$.

Now, using Statements 1 and 2, we observe that the following random variable will have the same distribution as the waiting time of a non-preempting class k:

$$W_k = J_K^C \cdot \left(W_k^{q0} + \sum_{i=1}^{G_{np}} B_P^i \right) + \left(1 - J_K^C\right) \cdot \left(1 - J_{np}^0\right) \sum_{i=1}^{G_{np}+1} B_P^i \tag{2.27}$$

where J_K^C and J_{np}^0 are Bernoulli random variables with parameters E_K^C and p_{np}^0, respectively, $G_{np} \sim Geom(p_{np})$, W_k^{q0} has LST $\tilde{f}_k^0(s)$, B_P^i, $i \ge 1$ has LST $\tilde{B}_P(s)$, and all random variables above are independent.

It is easy to show that if $\{Y_i\}_{i=1}^{\infty}$ are iid random variables with LST $\tilde{f}_Y(s)$, N has $Geom(P)$ distribution and $X = \sum_{i=1}^{N} Y_i$, then LST of X is given by:

$$\tilde{f}_X(s) = \frac{p}{1 - (1 - p)\tilde{f}_Y(s)}. \tag{2.28}$$

Now (2.9) follows from representation (2.27), (2.28) and well-known LST properties.

Proof of Corollary 1 Formula (2.11) can be derived via the straightforward differentiation of (2.9) and a well-known relation between the delay probability in M/M/n and the loss probability in M/M/n/n:

$$E_k^B = \frac{E_k^C(1 - \bar{\rho}_k)}{1 - \bar{\rho}_k E_k^C}, \quad 1 \le k \le K,$$

that enables us to present expressions in (2.4)–(2.7) via M/M/n delay probabilities.

Formulae (2.12) and (2.13) are derived via computation of the second derivatives of (2.8) and (2.9), respectively, in the origin. (We omit details of these straightforward but tedious calculations; (2.12) and (2.13) were verified via symbolic differentiation software.) □

2.3 Analysis of sojourn times

Distributions of sojourn times cannot be derived automatically from distributions of waiting times. The reason is that service and waiting times are dependent: longer ser-

vice times would imply more service interruptions and longer waiting times. Therefore, we need to use a different technique to derive LST and moments of sojourn times.

Let

$$\pi_k^j = \frac{(\bar{\lambda}_k/\mu)^j}{j!} \cdot \left[\sum_{i=0}^{n-1} \frac{(\bar{\lambda}_k/\mu)^i}{i!} + \frac{(\bar{\lambda}_k/\mu)^n}{n!(1-\bar{\rho}_k)} \right]^{-1}, \quad 1 \le k \le P,\ 0 \le j \le n,$$

denote the steady-state probabilities of number-in-system in M/M/n ($\bar{\lambda}_k$, μ). In addition, denote by $\tilde{V}_k(s, j)$, $0 \le j \le n$, the LST of the sojourn time distribution for preempting class k, $1 \le k \le P$, given j customers from classes $1-k$ in the system upon arrival. Let $\tilde{V}_a(s, j)$, $0 \le j \le n-1$, denote the LST of the sojourn time distribution for non-preempting class k, $P+1 \le k \le K$, given j customers from *all* classes $1-K$ in the system upon arrival. (This distribution does not depend on a customer's class, see Sect. 2.3.1.) Finally, $\tilde{V}_a(s, n)$ is the LST of the sojourn time distribution for non-preempting class $P+1$, given all servers busy and no queue in the system upon arrival.

Theorem 2 *The LST of the steady-state sojourn time distribution is given by*:

$$\tilde{V}_k(s) = \sum_{j=0}^{n-1} \pi_k^j \tilde{V}_k(s, j) + \frac{\pi_k^n \tilde{V}_k(s, n)}{1 - \bar{\rho}_k \tilde{B}_{k-1}(s)}, \quad 1 \le k \le P, \tag{2.29}$$

$$\tilde{V}_k(s) = \sum_{j=0}^{n-1} \pi_K^j \tilde{V}_a(s, j) + E_K^C \cdot \tilde{f}_k^0(s) \cdot \tilde{V}_a(s, n-1), \quad P+1 \le k \le K, \tag{2.30}$$

where functions $\tilde{B}_{k-1}(s)$ and $\tilde{f}_k^0(s)$ were defined in formulae (2.1) and (2.2), respectively, and Laplace–Stieltjes transforms $\tilde{V}_k(s, j)$ in (2.29) can be calculated via the system of $n+1$ equations, provided by (2.31)–(2.32). Specifically, for $0 \le j \le n-1$

$$-j\mu \tilde{V}_k(s, j-1) + \left[s + \bar{\lambda}_{k-1} + (j+1)\mu \right] \tilde{V}_k(s, j) - \bar{\lambda}_{k-1} \tilde{V}_k(s, j+1) = \mu, \tag{2.31}$$

$$\tilde{V}_k(s, n) = \tilde{B}_{k-1}(s) \tilde{V}_k(s, n-1). \tag{2.32}$$

(*Assume* $\tilde{V}_k(s, -1) = 0$.)

Finally, the Laplace–Stieltjes transforms $\tilde{V}_a(s, j)$ in (2.30) are computed by the system of equations (2.33)–(2.34). For $0 \le j \le n-1$

$$-j\mu \tilde{V}_a(s, j-1) + \left[s + \bar{\lambda}_P + (j+1)\mu \right] \tilde{V}_a(s, j) - \bar{\lambda}_P \tilde{V}_a(s, j+1) = \mu, \tag{2.33}$$

and

$$\tilde{V}_a(s, n) = \tilde{B}_P(s) \tilde{V}_a(s, n-1). \tag{2.34}$$

(*Assume* $\tilde{V}_a(s, -1) = 0$.)

Remark 5 Note that formula (2.29), (2.31) and (2.32) are known from Segal [21].

2.3.1 *Proof of Theorem 2*

The proof of Theorem 2 is based on the method of Segal [21]. Since the proof for preempting classes is almost identical to [21] (only formulae modification for $\mu \neq 1$ is needed), we shall consider the non-preempting classes $P+1 \leq k \leq K$. First, we shall prove formulae (2.33) and (2.34) for LSTs of conditional distributions, and then proceed to (2.30).

Proof of (2.33)–(2.34) Assume that a non-preempting customer from class k, $P+1 \leq k \leq K$, arrived, observed $0 \leq j < n$ customers in the system and started service immediately. Define by $f_a(t, j)$, $0 \leq j < n$, density of sojourn time for this customer. This density does not depend on class k since, according to our service protocol, the class of a non-preempting customer is not taken into account after the first service start. Assume that a non-preempting customer from class $P+1$ arrived, observed n customers in the system and was placed at the head of the queue. Define by $f_a(t, n)$ the corresponding density of sojourn time.

Conditioning on the next event (service of our customer, service of other customer, and arrival of preemptive customer) we get:

$$f_a(t, j) = \mu e^{-\mu t} e^{-(j\mu+\bar{\lambda}_P)t} + j \int_0^t \mu e^{-\mu x} e^{(-j\mu+\bar{\lambda}_P)x} f_a(t-x, j-1)\, dx + \int_0^t \bar{\lambda}_P e^{-\bar{\lambda}_P x} e^{-(j+1)\mu x} f_a(t-x, j+1)\, dx, \quad 0 \leq j \leq n-1. \tag{2.35}$$

(Note that the last term of (2.35) for $j = n-1$ corresponds to ejection of our customer. In this case, according to the service protocol in Sect. 1.3, the customer is placed at the head of the queue of ejected customers and his/her residual sojourn time distribution is identical to the sojourn time distribution of class $P+1$ customer that observed n customers in the system upon arrival.)

Applying Laplace–Stieltjes transform to (2.35), we derive (2.33) via straightforward manipulations.

Equation (2.34) prevails, because if a customer from class $P+1$ is at the first place in the queue, it takes him/her the busy-period length of M/M/$n(\bar{\lambda}_P, \mu)$ to start service. (See the proof of Theorem 1 and the definition of the service protocol in Sect. 1.3.) □

Proof of (2.30) As mentioned previously in Sect. 2.2.1, the number of jobs in our system is identical to the number-in-system in M/M/$n(\bar{\lambda}_K, \mu)$. According to the PASTA principle [26], an arriving customer encounters $0, 1, \ldots, n-1$ customers with probabilities $\pi_K^0, \pi_K^1, \ldots, \pi_K^{n-1}$. The conditional density of the sojourn time is equal to $f_a(t, j)$, $0 \leq j \leq n-1$ in this case, implying the first term on the right side of (2.30). A non-preempting customer must wait with probability E_K^C and, in this case, the LST of his waiting time till the first service start is given by $\tilde{f}_k^0(s)$ defined in (2.2). (See Statement 2 from the proof of Theorem 1.) The sojourn time after the first service start is distributed according to $f_a(t, n-1)$, implying the second term on the right side of (2.30). □

2.3.2 *Moments of sojourn times*

Theorem 2 can be used to derive the moments of sojourn times. Differentiating Laplace–Stieltjes transforms $\tilde{V}_k(s)$ at the origin, we get systems of linear equations for these moments that are presented below.

In our calculations, we need the moments of the busy-period length of M/M/n ($\bar{\lambda}_k$,μ). Denote the lth moment, $l \geq 1$, by B_k^l. Their values can be calculated via the differentiation of (2.1) at the origin. For example, the first three moments are given by:

$$B_k^1 = \frac{1}{n\mu(1-\bar{\rho}_k)}, \qquad B_k^2 = \frac{2}{(n\mu)^2(1-\bar{\rho}_k)^3}, \qquad B_k^3 = \frac{6(1+\bar{\rho}_k)}{(n\mu)^3(1-\bar{\rho}_k)^5}.$$

Preempting classes Let $M_k^l(j)$, $1 \leq k \leq P$, $0 \leq j \leq n$, $l \geq 1$, denote the lth moment of sojourn time given j customers of classes $1,\ldots,k$ upon arrival. Differentiating (2.31) and (2.32), we derive the system of $(n+1)$ linear equations (2.36)–(2.37) for these moments:

$$-j\mu M_k^l(j-1) + \bigl[\bar{\lambda}_{k-1} + (j+1)\mu\bigr] M_k^l(j) - \bar{\lambda}_{k-1} M_k^l(j+1) = l M_k^{l-1}(j), \quad (2.36)$$

where $0 \leq j \leq n-1$ and by convention $M_k^l(-1) = 0$, and

$$M_k^l(n) = M_k^l(n-1) + \sum_{i=0}^{l-1} \binom{l}{i} M_k^i(n-1) B_{k-1}^{l-i}, \quad (2.37)$$

where $M_k^0(j) = 1$. Now the unconditional formulae for moments of sojourn times can be obtained via the differentiation of (2.29).

First and second moments for preempting classes

$$E[V_k] = \sum_{j=0}^{n-1} \pi_k^j M_k^1(j) + \frac{\pi_k^n M_k^1(n)}{1-\bar{\rho}_k} + \frac{\pi_k^n \bar{\rho}_k B_{k-1}^1}{(1-\bar{\rho}_k)^2},$$

$$E\bigl[V_k^2\bigr] = \sum_{j=0}^{n-1} \pi_k^j M_k^2(j) + \frac{\pi_k^n M_k^2(n)}{1-\bar{\rho}_k} + \frac{\pi_k^n \bar{\rho}_k (B_{k-1}^2 + 2B_{k-1}^1 M_k^1(n))}{(1-\bar{\rho}_k)^2} + \frac{2\pi_k^n (\bar{\rho}_k B_{k-1}^1)^2}{(1-\bar{\rho}_k)^3}.$$

(Formula (2.10) provides a simpler alternative for the computation of means.)

Non-preempting classes The lth moments of sojourn times for non-preempting classes $P+1 \leq k \leq K$ can be computed via the differentiation of (2.30):

$$\mathrm{E}\bigl[V_k^l\bigr] = \sum_{j=0}^{n-1} \pi_K^j M_a^l(j) + E_K^C \sum_{i=0}^{l} \binom{l}{i} M_{np}^i M_a^{l-i}(n-1), \quad (2.38)$$

where $M_a^l(j)$, $0 \le j \le n-1$, $l \ge 1$, are sojourn time moments of non-preempting classes, conditioned on j customers in the system upon arrival. Conditional moments $M_a^l(j)$ are computed via (2.39)–(2.40):

$$-j\mu M_a^l(j-1) + \big[\bar{\lambda}_P + (j+1)\mu\big] M_a^l(j) - \bar{\lambda}_P M_a^l(j+1) = l M_a^l(j),$$
$$0 \le j \le n-1, \tag{2.39}$$

$$M_a^l(n) = M_a^l(n-1) + \sum_{i=0}^{l-1} \binom{l}{i} M_a^l(n-1) B_P^{l-i}. \tag{2.40}$$

Finally,

$$M_{np}^i = \mathrm{E}\big[(W_k)^i \mid W_k > 0\big],$$

are conditional moments of the waiting time in a pure non-preemptive system that can be derived via the differentiation of (2.2). Specifically, the first three moments are given by:

$$M_{np}^1 = \frac{1}{n\mu(1-\bar{\rho}_{k-1})(1-\bar{\rho}_k)},$$
$$M_{np}^2 = \frac{2(1-\bar{\rho}_{k-1}\bar{\rho}_k)}{(n\mu)^2(1-\bar{\rho}_k)^2(1-\bar{\rho}_{k-1})^3},$$
$$M_{np}^3 = \frac{6}{(n\mu)^3} \cdot \frac{(1+\bar{\rho}_{k-1})(1+\bar{\rho}_{k-1}\bar{\rho}_k^2) - 4\bar{\rho}_{k-1}\bar{\rho}_k}{(1-\bar{\rho}_k)^3(1-\bar{\rho}_{k-1})^5}.$$

Special cases. First and second moments for non-preempting classes

$$\mathrm{E}[V_k] = \sum_{j=0}^{n-1} \pi_K^j M_a^1(j) + E_K^C \cdot \left[\frac{1}{n\mu(1-\bar{\rho}_{k-1})(1-\bar{\rho}_k)} + M_a^1(n-1)\right],$$

$$\mathrm{E}\big[V_k^2\big] = \sum_{j=0}^{n-1} \pi_K^j M_a^2(j) + E_K^C \cdot \left[\frac{2(1-\bar{\rho}_{k-1}\bar{\rho}_k)}{(n\mu)^2(1-\bar{\rho}_k)^2(1-\bar{\rho}_{k-1})^3} + \frac{2M_a^1(n-1)}{n\mu(1-\bar{\rho}_{k-1})(1-\bar{\rho}_k)} + M_a^2(n-1)\right].$$

(Formula (2.11) above provides a simpler alternative for means.)

2.3.3 *Laplace–Stieltjes transforms of sojourn-time distribution: outline of alternative calculation*

In Sect. 2.3 LSTs of sojourn times were derived implicitly. (However, explicit expressions could be developed via inversion of coefficient matrices from (2.31)–(2.34).) An approach that is similar to the one used in Tatashev [24] enables an alternative derivation of explicit LSTs. Below we outline this derivation and finish with a brief

discussion. We consider non-preempting priorities; the calculations for preempting priorities are similar.

If I_{np} denotes the number of service interruptions for a non-preempting customer, then the service time has $I_{np}+1$ phases. Introduce the following four distributions and the corresponding LSTs:

- LST $\tilde{g}^{l}_{np}(s)$ corresponds to the last phase of service time, given this phase started after positive wait.
- LST $\tilde{g}^{l0}_{np}(s)$ corresponds to the last phase of service time, given this phase started immediately upon arrival.
- LST $\tilde{g}^{e}_{np}(s)$ corresponds to a phase of service time that was terminated by ejection, given this phase started after positive wait.
- LST $\tilde{g}^{e0}_{np}(s)$ corresponds to a phase of service time that was terminated by ejection, given this phase started immediately upon arrival.

Then the following random variable will have the same distribution as the sojourn time of a non-preempting customer from class k, $P+1 \le k \le K$:

$$V_k = J^C_K \cdot \left(V^{q0}_k + \sum_{i=1}^{G_{np}} B^i_P + S^l_{np} + \sum_{i=1}^{G_{np}} S^{e,i}_{np} \right) + \left(1 - J^C_K\right) \cdot J^0_{np} \cdot S^{l0}_{np}$$
$$+ \left(1 - J^C_K\right) \cdot \left(1 - J^0_{np}\right) \left(\sum_{i=1}^{G_{np}+1} B^i_P + S^l_{np} + \sum_{i=1}^{G_{np}} S^{e,i}_{np} + S^{e0}_{np} \right). \tag{2.41}$$

Here S^l_{np} has LST $\tilde{g}^l_{np}(s)$, S^{l0}_{np} has LST $\tilde{g}^{l0}_{np}(s)$, S^{e0}_{np} has LST $\tilde{g}^{e0}_{np}(s)$, and random variables $S^{e,i}_{np}$, $i \ge 1$ have LST $\tilde{g}^e_{np}(s)$. Definitions of other random variables in (2.41) are explained in the comments to representation (2.27). All random variables in consideration are independent.

Then from LST properties, we get the following Laplace–Stieltjes transform of sojourn time for non-preempting class k:

$$\tilde{V}_k(s) = \left(1 - E^C_K\right) p^0_{np} \tilde{g}^{e0}_{np}(s)$$
$$+ \frac{E^C_K p_{np} \tilde{f}^0_k(s) \tilde{g}^l_{np}(s) + (1 - E^C_K) p_{np} (1 - p^0_{np}) \tilde{B}_P(s) \tilde{g}^{e0}_{np}(s) \tilde{g}^l_{np}(s)}{1 - (1 - p_{np}) \tilde{B}_P(s) \tilde{g}^l_{np}(s)}.$$

We now describe how to compute four LSTs of service-time phases.

We start with a general remark. Let F denote a general phase-time distribution with generator matrix R and vector of starting probabilities $\bar{q}$. It is known [17] that LST of F is given by $\bar{q}'[sI - R]^{-1}\bar{r}$, where $\bar{r} = -R\bar{1}$ is the vector of transition rates to the absorbing state. Let Y denote exponential random variable with rate μ, which is independent of F. We are interested in the LSTs of two conditional distributions, $\{Y \mid Y \le F\}$ and $\{Y \mid Y > F\}$. Straightforward calculations provide us with:

$$\tilde{f}_{\{Y|Y\le F\}}(s) = \frac{\mu}{\mu + s} \cdot \frac{1 - \bar{q}'[(s+\mu)I - R]^{-1}\bar{r}}{1 - \bar{q}'[\mu I - R]^{-1}\bar{r}}, \tag{2.42}$$

$$\tilde{f}_{\{Y|Y>F\}}(s) = \frac{\mu}{\mu+s} \cdot \frac{\bar{q}'[(s+\mu)I - R]^{-1}\bar{r}}{\bar{q}'[\mu I - R]^{-1}\bar{r}}. \tag{2.43}$$

Define F_1 to be the length of the *idle period* of the M/M/$n(\bar{\lambda}_P, \mu)$ queue (an idle period is an interval between two busy-periods). Let F_2 denote the time till the start of the next busy-period of M/M/$n(\bar{\lambda}_P, \mu)$, if the starting number-in-system probabilities are $q_{2i} = (\bar{\lambda}_K^i / i!)/(\sum_{j=0}^{n-1} \bar{\lambda}_K^j / j!)$, $0 \le i \le n-1$. Note that F_1 and F_2 are phase-type distributions with the same generator matrix and two different starting distributions: $(0, \ldots, 0, 1)'$ for F_1 and $\bar{q}_2 = (q_{20}, q_{21}, \ldots, q_{2,n-1})$ for F_2. Observe that the service of a non-preempting customer is interrupted whenever there are n preempting customers in the system, which corresponds to the start of the M/M/$n(\bar{\lambda}_P, \mu)$ busy-period. Hence,

$$\tilde{g}_{np}^{l}(s) = \tilde{f}_{\{Y|Y\le F_1\}}(s), \qquad \tilde{g}_{np}^{e}(s) = \tilde{f}_{\{Y|Y>F_1\}}(s),$$
$$\tilde{g}_{np}^{l0}(s) = \tilde{f}_{\{Y|Y\le F_2\}}(s), \qquad \tilde{g}_{np}^{e0}(s) = \tilde{f}_{\{Y|Y>F_2\}}(s)$$

and can be calculated via (2.42) and (2.43).

Discussion Comparing the methods of this section with Theorem 2 and Sect. 2.3.2, we conclude that Theorem 2 provides more convenient tools for moments calculation. However, LST calculation in this section enables additional insights into sojourn time structure and could possibly be generalized for other service systems.

3 Numerical examples

3.1 Mean waiting times

Mean waiting and mean sojourn times are wide-spread performance measures in many real-world service systems. We compare mean waiting times for systems with pure non-preemptive, pure preemptive, and mixed priorities, respectively, via the following instructive example.

Example 1 $K = 5$, $P = 2$, $n = 2$, $\mu = 1$, $\lambda_3 = 1.2$ and $\lambda_k = 0.1$, $k = 1, 2, 4, 5$ (see Table 1).

Example 1 illustrates a situation in which mixed priorities are preferable to the alternatives from an SLA point of view. Assume that the SLA for the mean waiting

Table 1 Example 1. Mean waiting times

Priorities	1	2	3	4	5
Non-Preemptive	0.37	0.42	1.32	4.74	7.11
Preemptive	0.00	0.02	1.12	5.83	9.16
Mixed	0.00	0.02	1.37	4.80	7.17

Table 2 Example 1. Service-level agreement for waiting times

Class 1	90% of jobs wait less than 30 minutes
Class 2	90% of jobs wait less than 1 hour
Class 3	80% of jobs wait less than 5 hours
Class 4	80% of jobs wait less than 8 hours
Class 5	80% of jobs wait less than 12 hours

times is given by the vector $(0.2, 0.4, 2, 5, 8)$. (The mean waiting time does not exceed 0.2 for class 1, 0.4 for class 2, etc.) The non-preemptive discipline then provides insufficient SL for classes 1 and 2 and the preemptive discipline implies unsatisfactory performance for classes 4 and 5. In contrast, the mixed discipline enables satisfaction of the SLA for all classes. We also assume that the arrival rate of "middle priority" class 3 is significantly larger than for the other classes. (According to our practical experience, this is a typical situation in many applications.) In such cases, a moderate service-level decrease for a "middle" class enables significant improvement for the other classes.

3.2 Distribution of waiting times

As mentioned in Sect. 1.1, the tail probability $\mathrm{P}\{W_k > T\}$ (or $\mathrm{P}\{V_k > T\}$) is probably the most popular performance measure in service applications. In this section, we examine the behavior of the waiting time tail-probabilities for Example 1 from Sect. 3.1.

Assume that the SLA is formulated via Table 2 (let our time-units be hours).

Figure 2 compares between survival functions (tail probabilities) for the three priority disciplines, considered in Sect. 3.1. (The arrows show SLAs for the corresponding class.) The algorithm for the distribution calculation via the LST inversion has been designed using the approach in Jagerman and Melamed [15]. As in Sect. 3.1, we observe that only the mixed priorities discipline satisfies the SLA in Table 2.

3.3 Moments-based approximations of sojourn time distribution

In general, SLAs with waiting times are prevalent in systems that involve direct interaction with customers (for example, call centers), while SLAs with sojourn times are typical in systems without such interaction. As exhibited in Sect. 2.3, it is more difficult to compute the LSTs of sojourn times than the LSTs of waiting times. However, it is relatively easy to compute sojourn-time moments. In Fig. 3, we approximate the sojourn-time survival functions for classes 1 and 4 from Example 1, via the first three moments using the algorithm presented in Osogami and Harchol-Balter [18, 19]. This algorithm fits a general distribution to an Erlang–Coxian distribution by three moments. (We checked this algorithm for the sojourn times of the pure non-preemptive priorities and it implies a reasonable-to-excellent fit for various examples.)

If performance conditions for classes 1 and 4 are given in Table 3, then only the mixed priorities enable SLA satisfaction (see Fig. 3).

Remark 6 As mentioned in Sect. 1.1, the mixed priority discipline can be also relevant if pure preemptive priorities are an appropriate option from an SLA point of

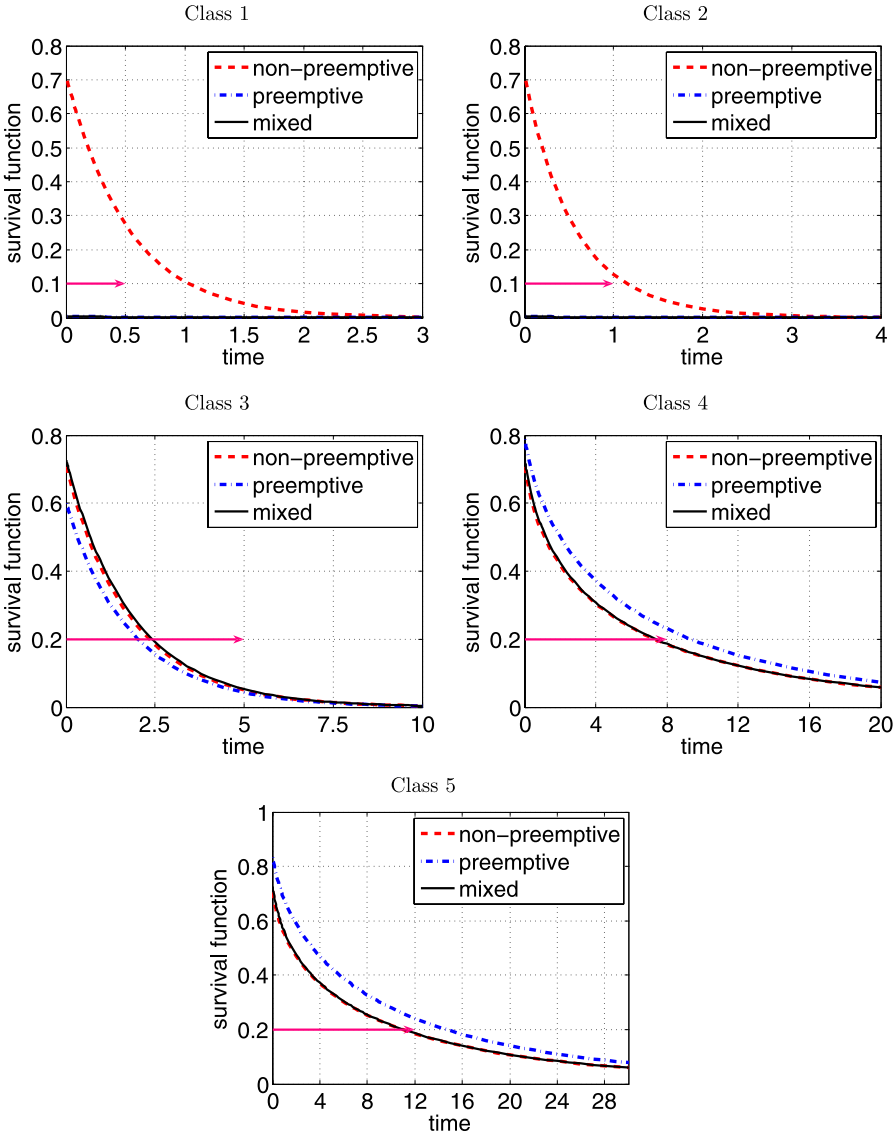

Fig. 2 Example 1. Waiting time distributions

Table 3 Example 1. Service-level agreement for sojourn times

Class 1	90% of jobs are served within 2.5 hours
Class 4	80% of jobs are served within 9 hours

view, but one would still like to decrease the number of service interruptions. Examples from Sects. 3.2 and 3.3 justify this approach. Assume, for example, that the first two classes demand a very high service level and that SLAs for classes 3–5 are

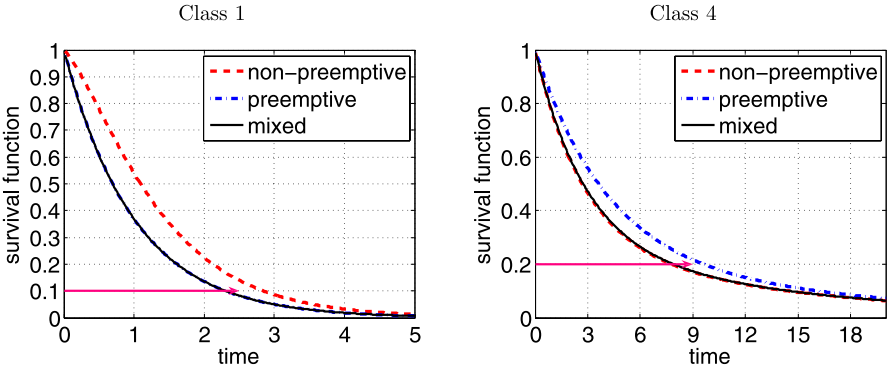

Fig. 3 Example 1. Sojourn time distributions

satisfied for all priority disciplines in consideration. The priority structure defined in Example 1 is then a good choice. In this case, one could also consider service protocols that differentiate between the non-preempting classes after the first service start, providing performance that is closer to the pure preemptive case than our protocol.

3.4 Practical example: scheduling of the IT support team

In Sect. 1.1, we mentioned a scheduling project for IT support teams. Below is a numerical example based on real data from this application. (Specifically, arrival rates, service rate, and priority structure are derived from this project. However, the problem contained additional features not covered in this example and that required the use of simulation in the scheduling algorithm. See Wasserkrug et al. [25] for details.)

An IT support system with five service classes (known as "problem tickets") is considered. The two classes with the highest priorities have much stricter SLAs than the three lower-priority classes.

The same service-time distribution for all classes of customers has been a working assumption in the field. Our statistical analysis for one of the classes has shown that the mean service time can be approximated by 100 minutes. Although the service time is not exponential, its coefficient of variation (standard deviation divided by mean) is smaller than 1 and according to well-known rule of thumb based on the Khintchine–Pollaczek formula and Allen–Cunneen approximation (see Hall [12]), we assume that the exponential service model will provide upper bounds for mean waiting and sojourn times.

Assume that the Poisson arrival rate depends on the day of the week and is constant over four-hour intervals. Arrival rate significantly differs among classes, class 3 constitutes more than 60% of overall arrivals. The arrival rate and, consequently, the scale of the problem is relatively small. However, scheduling methods for many IT support teams must be provided.

Assume that the SLA is formulated in Table 4.

Table 4 Practical example. Service-level agreement for waiting times

Class 1	90% of jobs wait less than 30 minutes
Class 2	90% of jobs wait less than 1 hour
Class 3	80% of jobs wait less than 5 hours
Class 4	80% of jobs wait less than 7 hours
Class 5	80% of jobs wait less than 9 hours

We want to provide a schedule with the minimal number of working hours per week that enables SLA satisfaction on a weekly level. As in the previous examples, we compare three priority disciplines: pure non-preemptive, pure preemptive, and mixed priorities with two preempting classes ($P = 2$). The last policy is the one applied in the field. The weekly tail probability of wait for each class and policy is calculated as the weighted average of the steady-state probabilities per interval.

We do not calculate the exact optimal staffing level, applying instead the following heuristics. Start with the minimal staffing that guarantees the stability of the system at each four-hour interval. We then consider a class that does not satisfy the corresponding SLA, add a server to the interval where the addition of a server implies the largest service-level improvement, and recalculate the tail probabilities for all classes. We stop when the SLA is satisfied for all classes.

Applying this method, we get 312 working hours per week for the mixed discipline, 328 working hours for the pure preemptive discipline, and 456 working hours for the pure non-preemptive discipline. Hence, the mixed discipline is significantly more efficient than the non-preemptive discipline (32% improvement). It is also more efficient than the preemptive discipline (5% improvement), while resulting in less service interruptions. We get similar results if SLA on mean waiting times is considered.

4 Possible future research

Other protocols with mixed priorities It would be interesting to explore alternatives to our service protocol, formulated in Sect. 1.3. For example, one could consider several versions of service protocols that differentiate between non-preempting customers after their first start and not only during initial waiting in queue, as we do.

Different service-time means and non-exponential service times If one needs to generalize the assumption of exponential service times with the same service mean for all classes, more complicated numerical methods and approximations are needed. For example, van der Heijden et al. [14] and Jagerman and Melamed [15] develop approximations for exponential service time with different means and non-exponential service times, respectively.

Incorporating abandonment In many systems (e.g., call centers or communication networks) customers/jobs can abandon if they stay in a queue too long. Rosenshmidt [20] provides exact and asymptotic analysis of such systems with pure non-preemptive or preemptive priorities. The generalization of her results for mixed priorities could be important for these applications.

Time-varying arrival rate Such queues are prevalent in practice and their analysis poses a challenge. A common approach is to approximate the time-varying arrival-rate by a piecewise-constant function, and then apply steady-state results during periods when the arrival rate is assumed constant, as we have done in Sect. 3.4. However, if the arrival rate is fast-varying with respect to the durations of services, this approach can be flawed. Feldman et al. [8] present a promising time-varying methodology for single-class queues.

Asymptotic analysis In service systems with a large number of servers (some call centers have thousands of agents working simultaneously), approximate methods for performance estimation and staffing can be appropriate. For example, Rosenshmidt [20] explores priority systems with abandonment in the so-called QED operational regime. A specific important research direction would be to check asymptotic equivalence of systems with mixed priorities, under certain conditions, to pure preemptive or non-preemptive ones.

Acknowledgements The authors are deeply grateful to Professor Avishai Mandelbaum from the Technion – Israel Institute of Technology for valuable comments on several paper drafts. Publicly available Matlab code of Takayuki Osogami [19] has been used in Sect. 3.3.

References

1. Abate, J., Whitt, W.: A unified framework for numerically inverting Laplace transforms. INFORMS J. Comput. **18**(4), 408–421 (2006)
2. Adiri, I., Domb, I.: A single-server queueing system working under mixed priority disciplines. Oper. Res. **30**, 97–115 (1982)
3. Buzen, J., Bondi, A.: The response times of priority classes under preemptive resume in M/M/*m* queues. Oper. Res. **31**, 456–465 (1983)
4. Chang, W.: Queuing with nonpreemptive and preemptive-resume priorities. Oper. Res. **13**, 1020–1022 (1965)
5. Cobham, A.: Priority assignment in waiting line problems. Oper. Res. **2**, 70–76 (1954)
6. Davis, R.: Waiting-time distribution of a multi-server, priority queueing system. Oper. Res. **14**, 133–136 (1966)
7. Drekic, S., Stanford, D.A.: Threshold-based interventions to optimize performance in preemptive priority queues. Queueing Syst. **35**, 289–315 (2000)
8. Feldman, Z., Mandelbaum, A., Massey, W., Whitt, W.: Staffing of time-varying queues to achieve time-stable performance. Manag. Sci. **54**(2), 324–338 (2008)
9. Gail, H., Hantler, S., Taylor, B.: Analysis of a non-preemptive priority multiserver queue. Adv. Appl. Probab. **20**, 852–879 (1988)
10. Gail, H., Hantler, S., Taylor, B.: On a preemptive Markovian queue with multiple servers and two priority classes. Math. Oper. Res. **17**, 365–391 (1992)
11. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: a tutorial and literature review. Invited review paper. Manuf. Serv. Oper. Manag. **5**(2), 79–141 (2003)
12. Hall, R.W.: Queueing Methods: For Services and Manufacturing. Prentice Hall, New York (1997)
13. Harchol-Balter, M., Osogami, T., Scheller-Wolf, A., Wierman, A.: Multi-server queueing systems with multiple priority classes. Queueing Syst. Theory Appl. **51**, 331–360 (2005)
14. van der Heijden, M., van Harten, A., Sleptchenko, A.: Approximations for Markovian multi-class queues with preemptive priorities. Oper. Res. Lett. **32**(3), 273–282 (2004)
15. Jagerman, D.L., Melamed, B.: Models and approximations for call center design. Methodol. Comput. Appl. Probab. **5**, 159–181 (2003)
16. Kella, O., Yechiali, U.: Waiting times in the non-preemptive priority M/M/c queue. Stoch. Models **1**, 257–262 (1985)

17. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models. The Johns Hopkins University Press, Baltimore and London (1981)
18. Osogami, T., Harchol-Balter, M.: Closed form solutions for mapping general distributions to quasi-minimal PH distributions. Perform. Eval. **63**, 524–552 (2006)
19. Osogami, T.: Online code repository. Available at http://www.cs.cmu.edu/~osogami/code/index.html
20. Rosenshmidt, L.: On priority queues with impatient customers: stationary and time-varying analysis. M.Sc. thesis, Technion. Available at http://iew3.technion.ac.il/serveng/References/references.html (2007)
21. Segal, M.: A multiserver system with preemptive priorities. Oper. Res. **18**, 316–323 (1970)
22. Simon, B.: Priority queues with feedback. J. ACM **31**(1), 134–149 (1984)
23. Sleptchenko, A., van Harten, A., van der Heijden, M.: An exact solution for the state probabilities of the multi-class, multi-server queue with preemptive priorities. Queueing Syst. Theory Appl. **50**(1), 81–107 (2005)
24. Tatashev, A.G.: Calculation of the distribution of the waiting time in a multiple-channel queueing system with fixed priorities. Eng. Cybern. **6**, 59–62 (1984)
25. Wasserkrug, S., Taub, S., Zeltyn, S., Gilat, D., Lipets, V., Feldman, Z., Mandelbaum, A.: Creating shift schedules for IT delivery center workers. Int. J. Serv. Oper. Inform. **3**, 242–257 (2008)
26. Wolff, R.W.: Poisson arrivals see time averages. Oper. Res. **30**, 223–231 (1982)