On a General Mixed Priority Queue with Server Discretion

Val Andrei Fajardo,[1*] Steve Drekic[1]

[1]Department of Statistics and Actuarial Science, University of Waterloo,
200 University Avenue West, Waterloo, ON N2L 3G1, Canada
[*]To whom correspondence should be addressed; E-mail: andrei.fajardo@uwaterloo.ca

We consider a single-server queueing system which attends to N priority classes that are classified into two distinct types: (i) urgent: classes which have preemptive resume priority over at least one lower priority class, and (ii) non-urgent: classes which only have non-preemptive priority amongst lower priority classes. While urgent customers have preemptive priority, the ultimate decision on whether to interrupt a current service is based on certain discretionary rules. An accumulating prioritization is also incorporated. The marginal waiting time distributions are obtained and numerical examples comparing the new model to other similar priority queueing systems are provided.

Keywords: Mixed priority queue; accumulating priority; server discretion rules; waiting time distribution; Laplace-Stieltjes transform.

1 Introduction

Service rules which dictate the order of service through the priority (or urgency) of the customers in the system are known as priority disciplines. Systems that employ a priority discipline give preferential treatment to customers of greater urgency in the sense that at a service selection instant, the customer of (or with) the greatest priority is usually selected. To remove the ambiguity in this notion of the "customer with the greatest priority", a mechanism for assigning priorities to the customers is required.

Oftentimes, the customers of a priority queueing system are categorized into a fixed number of distinct priority classes labelled with class indices $1, 2, \ldots, N$. Throughout this paper, we use the symbol $\mathcal{C}_i$ which is to be read as "class-i customer". In general, we say that $\mathcal{C}_i$s are prioritized over $\mathcal{C}_j$s whenever $i < j$. With this setup, one can assign priorities to customers quantitatively by using the so-called *priority functions*, which are generally class-dependent. We denote the priority function for the $\mathcal{C}_k$s by $q_k(t)$, where the argument t represents time.

Much of the existing literature has been focused on the study of priority disciplines for which priority is assigned to each class in a *static* (or fixed) manner. Specifically, under a static priority discipline, the priority functions are of the form

$$q_k(t) = a_k, \quad k = 1, 2, \ldots, N, \tag{1}$$

where the set of constants $\{a_i\}_{i=1}^{N}$ are arranged so that $a_1 > a_2 > \cdots > a_N$. Furthermore, amongst all of the customers belonging to the same class, it is assumed that the oldest such customer is the one with the greatest priority. Hence, within classes, customers are served on a first-come-first-serve (FCFS) basis.

Alternatively, the priority of a customer can be assigned in a non-static or *dynamic* fashion, so that the priority of the customer accumulates (or possibly dissipates) throughout its time in the system. Let ψ_k be the arrival time of a $\mathcal{C}_k$. One such example of a dynamic priority discipline uses priority functions of the form

$$q_k(t) = b_k(t - \psi_k), \quad t \geq \psi_k, \quad k = 1, 2, \ldots, N, \tag{2}$$

where the *priority accumulation rates* $\{b_i\}_{i=1}^N$ are arranged so that $b_1 \geq b_2 \geq \cdots \geq b_N \geq 0$. Kleinrock [17] was the first to consider a priority queue that assigns priority to customers via Eq. (2). His main contribution was a set of recursive equations for the means of the steady-state waiting times for each class. Note that $\{b_i\}_{i=1}^N$ represents a set of parameters for the system, enabling a systems manager to control the mean waiting times of each class by simply fine-tuning these rates. As pointed out by Kleinrock [17], it is precisely this flexibility which makes priority functions like the ones given in Eq. (2) so useful.

Several other researchers have considered various dynamic priority functions, and successfully obtained expressions (or bounds) for the mean waiting times of each class (to name a few, see the papers by Hsu [14], Kanet [16], Netterman and Adiri [18], and Trivedi et al. [22]). However, it is only recently that the paper by Stanford et al. [20] has provided a distributional result for the steady-state waiting times of a dynamic priority queue. In their paper, they considered the same dynamic priority discipline as in Kleinrock [17], which they referred to as the *accumulating priority queue* (APQ) discipline. In order to derive the *Laplace-Stieltjes transform* (LST) of the steady-state class-k waiting time distribution, the authors utilized a new stochastic process which they called the *maximal priority process*. Later in this paper, we too use the maximal priority process to obtain the LSTs of steady-state waiting time distributions for a certain collection of customer classes in our new priority queue.

Another very important distinction of priority queues is based on the decision of whether or not to interrupt the servicing of a customer for another higher priority customer present in the system. In this regard, there are three types of priority queues:

(i) Non-preemptive: service of customers proceeds to completion without any interruptions,

(ii) Preemptive: service of lower priority customers is interrupted for higher priority customers,

(iii) Mixed: subject to some discretionary rules, the service of lower priority customers may or may not be interrupted for higher priority customers.

The literature on all three types of priority queues for which the assignment of priority to customers is static is vast. For a detailed analysis on both static non-preemptive and preemptive priority queues, we refer the reader to the texts by Conway et al. [9], Jaiswal [15], and Takagi [21]. With regards to mixed priority queues, several researchers have previously considered various guidelines and discretion rules to dictate the interruptions of service. A well-known guideline for prescribing interruptions based solely on the class indices is the so-called *preemption distance* (PD) rule. The PD rule allows for preemption only if the difference in the class indices of the two customers under consideration exceeds a specified value. Adiri and Domb [3, 4] and Paterok and Ettl [19] have analyzed static priority queues implementing the PD rule. Mixed priority queues for which the discretion rules are based on the service time of the customer currently in service have also been previously considered. For example, three such discretion rules are:

1. *Proportion-based* (PB) *policy:* Once a certain proportion α, $0 \leq \alpha \leq 1$, of the service time has been successfully rendered, further preemptions are prevented;

2. *Front-end time-based* (FETB) *policy:* Once T time units of service have been successfully rendered, further preemptions are prevented;

3. *Tail-end time-based* (TETB) *policy:* Once the time remaining to successfully complete service is less than τ time units, further preemptions are prevented.

The above threshold-based discretion rules were first studied by Cho and Un [7]. Later, Drekic and Stanford [10] considered a generalized version of these discretion rules by allowing the threshold parameters to be class-dependent. In this paper, we consider a mixed priority model using a further generalization of the above threshold-based discretion rules to dictate the interruptions of service.

Due to their complex nature, the existing literature for dynamic priority queues is predominantly of the non-preemptive type. For example, the priority queues explored by Hsu [14], Kanet [16], Netterman and Adiri [18], and also the model considered by Stanford et al. [20], are all of the non-preemptive type. However, as evidenced in Fajardo and Drekic [13], one can apply similar techniques to those of Stanford et al. [20] to characterize the waiting time distributions for the preemptive variant of the APQ. In regards to other research papers investigating preemptive dynamic priority queues, there are (to our knowledge) only two other papers appearing in the priority queueing literature (i.e., Kleinrock [17] and Trivedi et al. [22]), both of which analyze the preemptive resume case.

In this paper, we consider an $M/G/1$ mixed priority queue where the N distinct priority classes of customers are further classified into two distinct types. Specifically, we refer to those classes which have preemptive resume priority over at least one lower priority class as *urgent* classes, and those which only have non-preemptive priority amongst lower priority classes as *non-urgent*. Also, the assignment of priorities to these two types of classes is different; urgent classes are assigned static priority as in Eq. (1), while non-urgent classes are assigned priority dynamically as in Eq. (2). We provide a detailed description of the model and other preliminaries in the next section.

The resulting priority queueing system is quite general and can be used to model several real world situations. For example, the main motivation of Stanford et al. [20] was to study the effectiveness of triage policies in an emergency room of a hospital. Their model was universally non-preemptive; however, it is quite reasonable to assume that some arriving patients will be more urgent than others and should require a doctor's attention immediately. Our new priority model allows for the consideration of such types of patients with preemptive priority over those which are less urgent. Moreover, in some instances, a doctor may decide to continue the servicing of a lower priority patient even in the midst of an arrival of an urgent-type patient. The new model can also have potential use in computer job scheduling applications, as well as other areas (such as those discussed in Drekic and Stanford [10, 11] and Paterok and Ettl [19]).

The rest of the paper is organized as follows. In Section 2, we introduce the model and the notation used throughout the paper. Section 3 describes the general methodology which is employed for deriving the LSTs of the marginal waiting time distributions. In Section 4, we establish the LSTs for the auxiliary random variables used to obtain the waiting time distributions. Two numerical examples, comparing our new priority system to similar previously-analyzed priority models, are given in Section 5. Lastly, in Section 6, we offer some concluding remarks and directions for future work.

2 Model description and preliminaries

2.1 Setup of the model

We consider a single-server queueing system which attends to N distinct priority classes of customers. The arrival processes for each class of customers form individual and independent Poisson processes, where λ_i denotes the arrival rate for class i, $i = 1, 2, \ldots, N$. We also let $\Lambda_i = \sum_{j=1}^{i} \lambda_j$ for $i = 1, 2, \ldots, N$. The service requirements for each customer are assumed to be class-dependent and independent of the arrival streams. Let $X^{(i)}$ represent the class-i service time random variable whose distribution function (df) and LST are denoted by

$$B^{(i)}(x) = \mathbb{P}(X^{(i)} \leq x) \quad \text{and} \quad \widetilde{B}^{(i)}(s) = \mathbb{E}(e^{-sX^{(i)}}),$$

respectively. In general, unless otherwise specified, we let $Y(x) = 1 - \overline{Y}(x) = \mathbb{P}(Y \leq x)$ and $\widetilde{Y}(s) = \mathbb{E}(e^{-sY})$ represent the df and LST, respectively, of a random variable Y.

We assume that $\mathcal{C}_i$s have priority over $\mathcal{C}_j$s whenever $i < j$. Moreover, the N classes of customers are further classified into two distinct types:

(i) *urgent*: classes which have preemptive resume priority over at least one lower priority class;

(ii) *non-urgent*: classes which only have non-preemptive priority amongst lower priority classes.

In general, we say that there are $0 \leq m \leq N$ urgent classes so that the set $\mathcal{U} \equiv \{i : 1 \leq i \leq m\}$ represents the collection of all urgent classes of customers. Conversely, $\mathcal{N} \equiv \{i : m < i \leq N\}$ denotes the aggregated set of non-urgent classes. For convenience, we refer to urgent and non-urgent customers as class-$\mathcal{U}$ and class-$\mathcal{N}$ customers, to be represented by the symbols $\mathcal{C}_{\mathcal{U}}$ and $\mathcal{C}_{\mathcal{N}}$, respectively.

The assignment of priority to a $\mathcal{C}_{\mathcal{U}}$ differs from that for a $\mathcal{C}_{\mathcal{N}}$. In particular, we use the following class-k priority functions:

- For $k \in \mathcal{U}$:

$$q_k(t) = a_k, \tag{3}$$

where $a_1 > a_2 > \cdots > a_m > 0$.

- For $k \in \mathcal{N}$:

$$q_k(t) = b_k(t - \psi_k), \quad t \geq \psi_k, \tag{4}$$

where $b_{m+1} \geq b_{m+2} \geq \cdots \geq b_N \geq 0$.

It is further assumed that

$$a_m >> b_{m+1}, \tag{5}$$

which guarantees that at no point in time could a $\mathcal{C}_{\mathcal{N}}$ ever have greater priority than a $\mathcal{C}_{\mathcal{U}}$. Moreover, we assume that a $\mathcal{C}_i$ has preemptive resume priority over a $\mathcal{C}_j$ whenever $i < j$ and only if $i \in \mathcal{U}$; otherwise, if $i \in \mathcal{N}$, then the $\mathcal{C}_i$ has only non-preemptive priority over the $\mathcal{C}_j$.

2.2 The service discipline

In this subsection, we describe, in careful detail, the service discipline of the new priority queue. Note that when we speak of a service selection instant, we are referring to an instant in time when a customer departs the system (i.e., after being completely serviced) and the server must subsequently select, from all the remaining customers in the system, the next customer to be serviced. It is important to realize that we do not consider a preemption instant to be a service selection instant.

For priority queueing systems, it is customary to use the following general service guideline:

Priority Service Guideline: At a service selection instant, the customer with the greatest priority enters into service.

We remark that the classical preemptive and classical non-preemptive priority queueing models both employ the Priority Service Guideline. Mixed priority queues, such as the one considered in this paper, also employ the Priority Service Guideline; however, certain policies may further be put into place so as to override the Priority Service Guideline at a specific type of service selection instant. We provide the details to these exceptions later on in this section.

For simplicity, in what follows next, we describe the service discipline from the perspective of a $\mathcal{C}_k$. Note that for each $k \in \{1, 2, \ldots, N\}$, a convenient partition of the remaining $N - 1$ classes can be constructed on the basis of the priority relationship between those classes and class k, namely:

$$
\begin{aligned}
b &\equiv \quad \text{The set of classes which class } k \text{ has priority over,} \\
a_{np} &\equiv \quad \text{The set of classes which have non-preemptive priority over class } k, \\
a_p &\equiv \quad \text{The set of classes which have preemptive priority over class } k, \\
a = a_{np} \cup a_p &\equiv \quad \text{The set of classes which have priority over class } k.
\end{aligned}
$$

To begin, suppose that a $\mathcal{C}_k$ enters into service for the first time. For systems with at least one urgent class (i.e., $m > 0$), a_p must be a non-empty set if $k > 1$, and hence, it is possible for the service of this $\mathcal{C}_k$ to be interrupted by a $\mathcal{C}_{a_p}$. An interruption may take place if there exists a $\mathcal{C}_{a_p}$ with greater priority than the $\mathcal{C}_k$ currently in service. Since $a_p \subset \mathcal{U}$, it follows as a consequence of Eqs. (3) and (5) that any interruption period must commence immediately upon the arrival of the interrupting $\mathcal{C}_{a_p}$ to the system.

Although it is true that the set of classes in a_p have preemptive priority over class k, the ultimate decision on whether to interrupt the current servicing of the $\mathcal{C}_k$ is made according to the three threshold-based discretion rules: PB, FETB, and TETB. As stated earlier, Drekic and Stanford [10] investigated the class-dependent case by letting α_k, T_k, and τ_k represent the corresponding class-k threshold parameters. We extend this idea one step further by allowing these threshold parameters to also depend on the class of the customer causing the interruption. Thus, we introduce $\alpha_{i,k} \in (0, 1)$, $T_{i,k} \geq 0$, and $\tau_{i,k} \geq 0$ as the corresponding class-k threshold parameters pertaining to a newly-arriving high priority $\mathcal{C}_i$, $i \in a_p$. For any $k > 1$ and $i < j \in a_p$, we further assume that

$$\alpha_{i,k} \geq \alpha_{j,k}, \qquad T_{i,k} \geq T_{j,k}, \qquad \text{and} \qquad \tau_{i,k} \leq \tau_{j,k}. \tag{6}$$

We say that a class-k service becomes *class-i protected* the moment that the service of the $\mathcal{C}_k$ can no longer be preempted by a $\mathcal{C}_i$, $i \in a_p$. Hence, the consequences of Eq. (6) are that a class-k service becomes class-j protected before it becomes class-i protected for $i < j \in a_p$.

Now, if the $\mathcal{C}_k$ is preempted out of service, then we refer to the interval of time starting from the preemption instant up until the moment that the interrupted $\mathcal{C}_k$ finally re-enters into service as an *interruption period.* In this paper, we define an interruption period to consist of the following two components: (i) the time required to completely service the interrupting customer, and (ii) the additional time required to clear the system of all those remaining $\mathcal{C}_{a_p}$s whom, if they had arrived to the system at the time of the preemption, would have also caused an interruption. Hence, at the end of an interruption period, the $\mathcal{C}_k$ re-enters service despite the fact that there may be customers of higher priority in the system (i.e., these are the higher priority customers who either never could, or can no longer cause an interruption to the $\mathcal{C}_k$).

Let $\{\delta_i\}_{i=1}^{\infty}$ represent the sequence of service selection instants. Furthermore, we denote a type-2 service selection instant to refer to a service selection instant which is also the instant in time that an interruption period ends. All other types of service selection instants are referred to as being of type 1. The service discipline for the new priority queue now follows:

- For type-1 service selection instants, the Priority Service Guideline is used to select the next customer for service.

- For type-2 service selection instants, the most recently interrupted customer re-enters into service.

- Preemption instants within the service of a $\mathcal{C}_k$ $(k > 1)$ occur at the arrivals of $\mathcal{C}_{a_p}$s in accordance with the threshold-based discretion rules of PB, FETB, and TETB.

2.3 Service-structure elements and auxiliary random variables

In this subsection, we define several random variables of interest. First of all, we define $W^{(k)}$ as the steady-state class-k waiting time representing the total elapsed time from a $\mathcal{C}_k$'s arrival to the system until the first time it enters service. In addition, the steady-state class-k flow time

$F^{(k)}$ represents the total time the $\mathcal{C}_k$ spends in the system. The main objective of this paper is to derive the LSTs of $W^{(k)}$ and $F^{(k)}$ for each $k = 1, 2, \ldots, N$. To do so, the following random variables, for which we collectively refer to as the service-structure elements, are needed:

Residence period ($R^{(k)}$) $\equiv$ The time elapsed between first entry into service of the $\mathcal{C}_k$ and its departure,

Completion period ($C^{(k)}$) $\equiv$ The total elapsed time between the initial entry of a $\mathcal{C}_k$ into service and the first instant that the server is ready to select the next $\mathcal{C}_k$ for service.

The *utilization factor* associated with our priority queueing model is given by

$$\rho = \sum_{i=1}^{N} \lambda_i \mathbb{E}(X^{(i)}),$$

which we assume satisfies the stability condition $\rho < 1$. In the next section, we derive the LST of $W^{(k)}$, which itself depends on the LSTs of the following two auxiliary random variables:

$\Upsilon_i^{(k)}$ $\equiv$ The interval of time starting with the service of a $\mathcal{C}_i$ ($i \in a$) and ending at the first moment that the server is ready to select the next $\mathcal{C}_k$ for service,

$\Phi_i^{(k)}$ $\equiv$ The interval of time starting with the class-k protected portion of service of a $\mathcal{C}_i$ ($i \in b$) and ending at the first moment that the server is ready to select the next $\mathcal{C}_k$ for service.

Remark 2.1 *For $k \in \mathcal{U}$, the first time that the server is ready to select a $\mathcal{C}_k$ after any one of these time intervals have started represents the first time that the system is clear of all $\mathcal{C}_a$s. However, for the case of $k \in \mathcal{N}$, the first time that the server is ready to select a $\mathcal{C}_k$ represents the first time that the system is clear of all those $\mathcal{C}_a$s which are only of a certain kind (to be introduced in the next section).*

In what follows, we extend the definition of $\Upsilon_i^{(k)}$ to incorporate the case when $i = k$, with the understanding that $\Upsilon_k^{(k)} = C^{(k)}$. To find the LST of the class-k flow time $F^{(k)}$, we use the relation

$$\widetilde{F^{(k)}}(s) = \widetilde{W^{(k)}}(s)\widetilde{R^{(k)}}(s),$$

which readily follows from the independence of $W^{(k)}$ and $R^{(k)}$. The derivations of the LSTs of $C^{(k)}$, $R^{(k)}$, and the two auxiliary random variables are carried out in Section 4.

Lastly, a classical result which we use repeatedly throughout the paper is the well-known functional for the LST of the duration of an $M/G/1$ busy period. In particular, the LST of a busy period in an $M/G/1$ queue with customer arrival rate λ and service time X having df $B(x)$ is given by (e.g., see Conway et al. [9, p. 150]),

$$\widetilde{\Gamma}(s) \equiv \widetilde{\Gamma}(s; \lambda, X) = \widetilde{B}(s + \lambda - \lambda\widetilde{\Gamma}(s)). \tag{7}$$

We also require the *delay* version of this result. Specifically, if the initial service time of the busy period is now X_0 having df $B_0(x)$, then the LST of the duration of this delay busy period is (e.g., see Conway et al. [9, p. 151])

$$\widetilde{\Gamma}_0(s) \equiv \widetilde{\Gamma}_0(s; \lambda, X, X_0) = \widetilde{B_0}(s + \lambda - \lambda\widetilde{\Gamma}(s)). \tag{8}$$

3 Derivation of the waiting time LST

To derive an expression for $\widetilde{W^{(k)}}(s)$, we employ two analytical approaches; one for each of the cases $k \in \mathcal{U}$ and $k \in \mathcal{N}$. The reason for the two separate approaches is the fact that the assignment of priority for a $\mathcal{C}_{\mathcal{U}}$ (which is via Eq. (3)) differs from that for a $\mathcal{C}_{\mathcal{N}}$ (which is via Eq. (4)). For the case $k \in \mathcal{U}$, we apply a similar level-crossing argument to the one used in Paterok and Ettl [19]. As evidenced in their work, the level-crossing method provides a simple approach to obtain the integral equation for the probability density function (pdf) of the steady-state class-k virtual wait. For dynamic priority queues, it is quite difficult to define the class-k virtual wait. However, Stanford et al. [20] developed a general approach to obtain the LSTs of waiting time distributions in a dynamic priority queue which uses the priority functions of Eq. (2). This approach, which takes inspiration from the traditional busy cycle approach used in Conway et al. [9], is what we use to establish $\widetilde{W}^{(k)}(s)$ for $k \in \mathcal{N}$.

3.1 Waiting time LST for $k \in \mathcal{U}$

Let $\{V_k(t), t \geq 0\}$ denote the class-k virtual wait process whose steady-state distribution we characterize as follows:

$$F(x) = \lim_{t\to\infty} \mathbb{P}(V_k(t) \leq x),\ f(x) = \lim_{t\to\infty} \tfrac{\partial}{\partial x}\mathbb{P}(V_k(t) \leq x),\ \text{and } P_0 = \lim_{t\to\infty} \mathbb{P}(V_k(t) = 0),$$

subject to the normalizing condition

$$P_0 + \int_0^\infty f(x)\mathrm{d}x = 1. \tag{9}$$

Note that $\{V_k(t), t \geq 0\}$ is at level 0 only during times that the server is either idle or is attending to a $\mathcal{C}_b$ in its class-k preemptible portion of service. During such times, we say that the system is in a *virtually idle* state. Hence, P_0 represents the long-run fraction of time that the system is virtually idle. Moreover, since the arrivals of the $\mathcal{C}_k$s form a Poisson process, it then follows that

$$\widetilde{W}^{(k)}(s) = \int_{x=0}^\infty e^{-sx}\mathrm{d}F(x) = P_0 + \int_0^\infty e^{-sx} f(x)\mathrm{d}x. \tag{10}$$

To obtain the desired LST, we apply a level-crossing approach to establish an integral equation for $f(x)$. Let $U_t(x)$ and $D_t(x)$ denote the number of up- and down-crossings of level x of the class-k virtual wait process, respectively, during the time interval $(0, t)$. The principle of set balance (e.g., see Brill [6, Section 2.4.6]) states that

$$\lim_{t\to\infty} \frac{\mathbb{E}(D_t(x))}{t} = \lim_{t\to\infty} \frac{\mathbb{E}(U_t(x))}{t}.$$

This fundamental relation between the up- and down-crossing rates of level x is precisely all we need to establish an integral equation for $f(x)$.

To find the up-crossing rate of level x of $\{V_k(t), t \geq 0\}$, we observe that a sample path of $\{V_k(t), t \geq 0\}$ up-jumps in three instances of time: (i) whenever a $\mathcal{C}_k$ arrives to the system, (ii) when a newly-arriving $\mathcal{C}_a$ finds the system in the virtually idle state, and (iii) the moment when a $\mathcal{C}_b$'s service becomes class-k protected. A typical sample path of $\{V_k(t), t \geq 0\}$ is illustrated in Figure 1. It is important to note that depending on the specification of the threshold-based discretion parameters, the service of a $\mathcal{C}_b$ may either be entirely, partially, or not at all class-k protected. In Figure 1, both the first and third waiting $\mathcal{C}_b$s have service times which are entirely class-k protected, whereas the second waiting $\mathcal{C}_b$ has a service time that is only partially class-k protected.

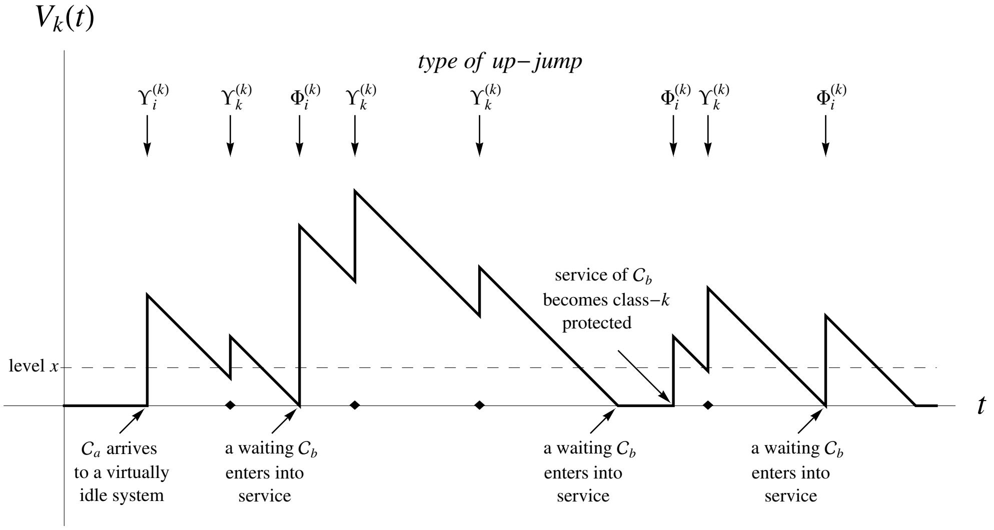

Figure 1: A typical sample path of $\{V_k(t), t \geq 0\}$

Let $\kappa_{k,i}$ denote the probability that the service of a $\mathcal{C}_i$ ($i \in b$) ever becomes class-k protected. Under the PB rule, $\kappa_{k,i} = 1$ as long as $\alpha_{k,i} < 1$ and is zero otherwise. Similarly, under the TETB rule, $\kappa_{k,i} = 1$ if $\tau_{k,i} > 0$ and is zero otherwise. However, for the FETB rule, a class-i service becomes class-k protected only if the service time is greater than $T_{k,i}$, and so $\kappa_{k,i} = 1 - B^{(i)}(T_{k,i})$ under this rule. The next theorem establishes the up- and down-crossing rates of level x.

Theorem 3.1 *The up- and down-crossing rates of level x are given by*

$$\lim_{t\to\infty}\frac{\mathbb{E}(U_t(x))}{t} = P_0\sum_{i=1}^{k}\lambda_i\overline{\Upsilon}_i^{(k)}(x) + \sum_{i=k+1}^{N}\kappa_{k,i}\lambda_i\overline{\Phi}_i^{(k)}(x) + \lambda_k\int_{y=0}^{x}\overline{\Upsilon}_k^{(k)}(x-y)f(y)\,\mathrm{d}y, \quad x>0 \tag{11}$$

and

$$\lim_{t\to\infty}\frac{\mathbb{E}(D_t(x))}{t} = f(x), \quad x>0. \tag{12}$$

Proof. We present intuitive explanations for each term of Eq. (11). For $i \in a$ or $i = k$, the rate of up-jumps caused by a $\mathcal{C}_i$ arriving to a virtually idle system is simply $\lambda_i P_0$. Furthermore, only the proportion $\overline{\Upsilon}_i^{(k)}(x)$ of these up-jumps lead to an up-crossing of level x. The rate at which a $\mathcal{C}_i$ ($i \in b$) arrives to the system that eventually induces a delay to the $\mathcal{C}_k$s is $\lambda_i\kappa_{k,i}$. Such arrivals eventually result in up-jumps of $\{V_k(t), t \geq 0\}$ which cross level x with probability $\overline{\Phi}_i^{(k)}(x)$. Finally, the long-run probability of an up-jump occurring from level y is $f(y)dy$, and the probability that an up-crossing of level x occurs from level y is $\overline{\Upsilon}_k^{(k)}(x-y)$. The justification of Eq. (12) is similar to that for the down-crossing rate of the virtual wait process in an $M/G/1$ queue (e.g., see Brill [6, Theorem 3.3 and Corollary 3.2]). □

From the principle of set balance, we equate Eqs. (11) and (12) to yield an integral equation for $f(x)$, namely,

$$f(x) = P_0\sum_{i=1}^{k}\lambda_i\overline{\Upsilon}_i^{(k)}(x) + \sum_{i=k+1}^{N}\lambda_i\kappa_{k,i}\overline{\Phi}_i^{(k)}(x) + \lambda_k\int_{y=0}^{x}\overline{\Upsilon}_k^{(k)}(x-y)f(y)\,\mathrm{d}y, \quad x>0. \tag{13}$$

By multiplying Eq. (13) by e^{-sx} and integrating x over $(0,\infty)$, we obtain

$$\int_{x=0}^{\infty}e^{-sx}f(x)\,\mathrm{d}x = \frac{P_0\big(\sum_{i=1}^{k}\lambda_i(1-\widetilde{\Upsilon}_i^{(k)}(s))\big) + \sum_{i=k+1}^{N}\lambda_i\kappa_{k,i}(1-\widetilde{\Phi}_i^{(k)}(s))}{s-\lambda_k+\lambda_k\widetilde{C}^{(k)}(s)}.$$

It follows from Eq. (10) that for $k \in \mathcal{U}$,

$$\widetilde{W}^{(k)}(s) = \frac{P_0\big(s+\sum_{i=1}^{k-1}\lambda_i(1-\widetilde{\Upsilon}_i^{(k)}(s))\big) + \sum_{i=k+1}^{N}\lambda_i\kappa_{k,i}(1-\widetilde{\Phi}_i^{(k)}(s))}{s-\lambda_k+\lambda_k\widetilde{C}^{(k)}(s)}. \tag{14}$$

Alternatively, by defining $W_{BP}^{(k)}$ as the waiting time of a $\mathcal{C}_k$ who arrives to the system during a busy period and incurs a positive wait time, we have that

$$\widetilde{W}^{(k)}(s) = P_0 + (1 - P_0)\widetilde{W}_{BP}^{(k)}(s).$$

From Eq. (10), it must be that $\widetilde{W}_{BP}^{(k)}(s) = \int_{x=0}^{\infty} e^{-sx} f(x)\mathrm{d}x/(1 - P_0)$. Moreover, an expression for $\mathbb{E}(W^{(k)})$ can be obtained by multiplying Eq. (13) by x and integrating x over $(0, \infty)$, leading to

$$\mathbb{E}(W^{(k)}) = \frac{P_0 \sum_{i=1}^{k-1} \lambda_i \mathbb{E}\big((\Upsilon_i^{(k)})^2\big) + \lambda_k \mathbb{E}\big((C^{(k)})^2\big) + \sum_{i=k+1}^{N} \lambda_i \kappa_{k,i} \mathbb{E}\big((\Phi_i^{(k)})^2\big)}{2\big(1 - \lambda_k \mathbb{E}(C^{(k)})\big)}. \tag{15}$$

We next proceed to establish a formula for P_0. Observe that

$$\int_0^{\infty} f(x)\,\mathrm{d}x = \frac{P_0 \sum_{i=1}^{k} \lambda_i \mathbb{E}(\Upsilon_i^{(k)}) + \sum_{i=k+1}^{N} \lambda_i \kappa_{k,i} \mathbb{E}(\Phi_i^{(k)})}{1 - \lambda_k \mathbb{E}(C^{(k)})}.$$

It readily follows, using the normalizing condition Eq. (9), that

$$P_0 = \frac{1 - \lambda_k \mathbb{E}(C^{(k)}) - \sum_{i=k+1}^{N} \lambda_i \kappa_{k,i} \mathbb{E}(\Phi_i^{(k)})}{1 + \sum_{i=1}^{k-1} \lambda_i \mathbb{E}(\Upsilon_i^{(k)})}. \tag{16}$$

We end the current subsection with a remark on the level-crossing approach used here and the one employed by Paterok and Ettl [19].

Remark 3.2 *The level-crossing analysis of $\{V_k(t), t \geq 0\}$ carried out by Paterok and Ettl [19] differs slightly from the one we use here. While their approach compares the expected number of up- and down-crossings of level x of $\{V_k(t), t \geq 0\}$ within a single regeneration cycle, our level-crossing analysis compares the long-run up- and down-crossing rates of level x. The latter level-crossing approach was first introduced by Brill [5], whereas the former approach was independently developed by Cohen [8].*

3.2 Waiting time LST for $k \in \mathcal{N}$

By definition, a $\mathcal{C}_{\mathcal{N}}$ can never preempt another customer out of service. Therefore, any $\mathcal{C}_{\mathcal{N}}$ who arrives to the system during a busy period must necessarily wait a positive amount of time before entering into service. Moreover, only those $\mathcal{C}_{\mathcal{N}}$s who arrive to the system during idle periods enter into service immediately upon arrival, without experiencing any wait. From these observations, an expression for the class-k waiting time LST is given by

$$\widetilde{W}^{(k)}(s) = (1-\rho) + \rho \widetilde{W}^{(k)}_{BP}(s), \quad k \in \mathcal{N}. \tag{17}$$

Let $P^{(k)}_{BP}$ be the accumulated priority (immediately prior to entering into service for the first time) of a $\mathcal{C}_k$ arriving to the system during a busy period. Since priority is assigned to a $\mathcal{C}_k$ via Eq. (4), the following simple relation holds:

$$P^{(k)}_{BP} = b_k \times W^{(k)}_{BP},$$

from which it follows that

$$\widetilde{W}^{(k)}_{BP}(s) = \widetilde{P}^{(k)}_{BP}(s/b_k). \tag{18}$$

Hence, to obtain the waiting time LST, we seek to derive $\widetilde{P}^{(k)}_{BP}(s)$. To do so, we make use of the so-called *maximal priority process*, which was first introduced by Stanford et al. [20, Section 3]. This stochastic process provides a useful structuralization of the general busy period and the customers serviced within it. We devote the next few subsections to its definition and some of its useful properties and results.

3.2.1 The maximal priority process

Upon arrival to the system, a $\mathcal{C}_k$ ($k \in \mathcal{N}$) begins to accumulate priority linearly at rate b_k. In this subsection, we define a specific upper bound for the accumulated priority of any $\mathcal{C}_k$ potentially present in the system at any time $t > 0$. We say "potentially present" since for

$b_k > 0$, this upper bound has the property of being positive during every busy period, even if none of the customers present in the system belong to class k. The collection of these upper bounds (i.e., $N - m$ in total, one for each $k \in \mathcal{N}$) is what Stanford et al. [20] referred to as the maximal priority process.

Stanford et al. [20] defined the maximal priority process in terms of the service commencement times and departure instants of the system. Since the current priority model allows for a $\mathcal{C}_{\mathcal{N}}$ to be preempted out of service, we require a slightly more general definition of the maximal priority process. Our definition of the maximal priority process follows below.

Definition 3.1 *The maximal priority process is an $(N - m)$-dimensional stochastic process $\mathcal{M}(t) = \{(M_{m+1}(t), M_{m+2}(t), \ldots, M_N(t)), t \geq 0\}$, satisfying the following conditions:*

1. *The sample path of $M_k(t)$ for each $k \in \mathcal{N}$ is continuous with respect to t, except possibly when t corresponds to a service selection instant.*

2. *$\mathcal{M}(t) = (0, 0, \ldots, 0)$ for all t corresponding to idle periods.*

3. *For all t during the service of any customer,*

$$\frac{\mathrm{d}M_k(t)}{\mathrm{d}t} = b_k, \quad k \in \mathcal{N}.$$

4. *At the sequence of service selection instants $\{\delta_i\}_{i=1}^{\infty}$,*

$$M_k(\delta_i^+) = \left\{ \begin{array}{ll} \min\{M_k(\delta_i^-), q_\vee(\delta_i^+)\} & \textit{if } \delta_i \textit{ is of type 1} \\ M_k(\delta_i^-) & \textit{if } \delta_i \textit{ is of type 2} \end{array} \right., \tag{19}$$

where $q_\vee(t)$ represents the greatest (accumulated) priority amongst all the customers present at time t, which is zero during idle periods. In Eq. (19), *note that*

$$M_k(\delta_i^-) = \lim_{\epsilon \to 0} M_k(\delta_i - \epsilon), \quad M_k(\delta_i^+) = \lim_{\epsilon \to 0} M_k(\delta_i + \epsilon), \quad \textit{and} \quad q_\vee(\delta_i^+) = \lim_{\epsilon \to 0} q_\vee(\delta_i + \epsilon).$$

In what follows, we (artificially) set $b_{N+1} = 0$ (which correspondingly implies that $M_{N+1}(t) = 0$ for all $t > 0$). Definition 3.1 simply implies that during busy periods, $M_k(t)$ increases linearly at rate b_k and down-jumps at some of the service selection instants. Figure 2 illustrates a typical sample path of the maximal priority process for a 5-class mixed priority queue with $m = 2$. In Figure 2, the actual accumulated priorities of the customers present in the system are given by the thin lines.

Suppose that δ represents a type-1 service selection instant for which at least one component of $\mathcal{M}(t)$ down-jumps (or, equivalently, δ represents an instant for which a down-jump in the first component $M_{m+1}(t)$ occurs). It then follows (from the Priority Service Guideline) that if there are any customers present at time δ, the $\mathcal{C}_{\mathcal{N}}$ with the greatest accumulated priority enters into service. Thus, the following two statements about the system at time δ must necessarily be true: (i) the system is clear of all $\mathcal{C}_{\mathcal{U}}$s, and (ii) the system is clear of all previously-interrupted customers.

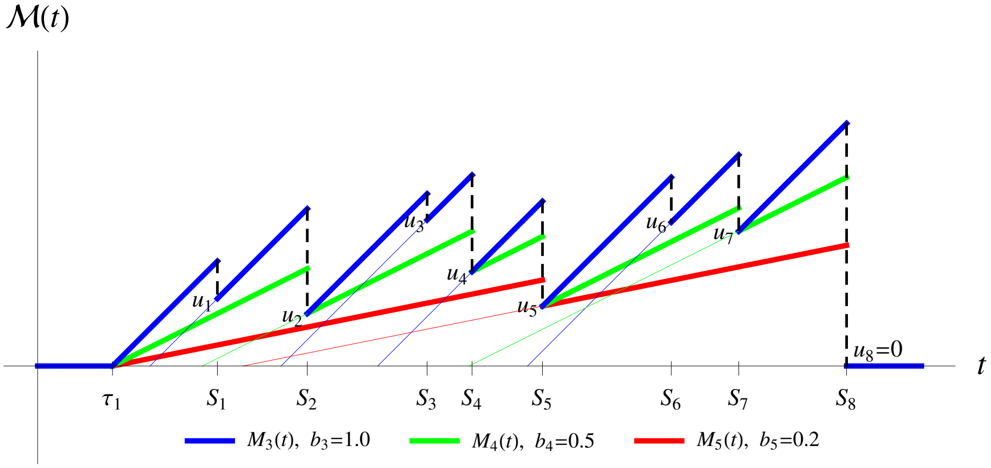

Figure 2: A typical sample path of $\{\mathcal{M}(t), t \geq 0\}$ for a 5-class mixed priority queue with $m = 2$ (i.e., $\mathcal{N} = \{3, 4, 5\}$)

Let S_i denote the i-th instant in time such that $M_{m+1}(t)$ down-jumps. In other words, S_i represents the i-th type-1 service selection instant satisfying the same two requirements as

δ above, namely (i) and (ii). We refer to S_i as the i-th service selection instant for a $\mathcal{C}_{\mathcal{N}}$. Furthermore, let $\mathbf{S} = \{S_i\}_{i=1}^{\infty}$ be the sequence of service selection instants for the $\mathcal{C}_{\mathcal{N}}$s. It is important to note that S_i represents the service commencement of a $\mathcal{C}_{\mathcal{N}}$ only if there are still customers who remain in the system at S_i. Otherwise, S_i represents the end of a busy period, which is signalled by a down-jump of $M_{m+1}(t)$ to level 0 (e.g., see S_8 in Figure 2).

The main reason for defining $\mathbf{S}$, however, is stated in the next observation. The maximal priority process defined for the non-urgent classes in our new priority queue behaves identically to the maximal priority process for the non-preemptive priority queue considered by Stanford et al. [20]. In other words, we can similarly analyze the waiting times for a $\mathcal{C}_{\mathcal{N}}$ of the new priority queue as we would for a customer in the non-preemptive dynamic priority queue considered by Stanford et al. [20]. In this equivalent non-preemptive priority queue, $\mathbf{S}$ would play the role of the sequence of departure instants of the customers, while $C^{(k)}$, $\Upsilon_i^{(k)}$, and $\Phi_i^{(k)}$ would serve as the effective service times.

Essential to our analysis are four important properties of the maximal priority process, which we describe below. We remark that these properties were first derived by Stanford et al. [20]. We do not provide the proofs of these properties but instead direct interested readers to Stanford et al. [20, Theorems 3.1 and 7.2] for their proofs. The four properties are as follows:

(P.1) The accumulated priorities of the $\mathcal{C}_{\mathcal{N}}$s still present in the queue at time t are distributed as independent Poisson processes, each with rate λ_i/b_i on the intervals $[0, M_i(t))$ for $i \in \mathcal{N}$.

(P.2) The accumulated priorities of the $\mathcal{C}_{\mathcal{N}}$s still present in the queue at time t are distributed as independent Poisson processes, each with piecewise constant rate zero on the interval $[M_{m+1}, \infty)$ and rate $\sum_{j=m+1}^{k} \lambda_j/b_j$ on the interval $[M_{k+1}(t), M_k(t))$ for $k \in \mathcal{N}$.

(P.3) A waiting $\mathcal{C}_{\mathcal{N}}$ whose priority, at time t, lies in the interval $[M_{k+1}(t), M_k(t))$ belongs to class i with probability $(\lambda_i/b_i)/(\sum_{j=m+1}^{k} \lambda_j/b_j)$, independently of the class of all other customers present in the queue.

(P.4) The statements (P.1)–(P.3) above also hold at any random time δ that is a stopping time for the raw filtration of $\mathcal{M}(t)$.

We end this introduction to the maximal priority process by giving an interpretation of the type of upper bounds that $\mathcal{M}(t)$ provides. First of all, for each $k \in \mathcal{N}$, $M_k(t)$ is the least upper bound of class-k accumulated priorities which would not result in a violation of the service discipline. Secondly, one can think of $\mathcal{M}(t)$ as the collection of these least upper bounds for accumulated priorities that one would sketch when given only the following three pieces of information:

(a) the sequence of busy period commencement times $\{\tau_i\}_{i=1}^{\infty}$,

(b) the sequence $\mathbf{S}$ of service selection instants for the $\mathcal{C}_{\mathcal{N}}$s, and

(c) for each $i = 1, 2, \ldots$, the value $u_i = q_{\vee}(S_i^+)$ corresponding to the greatest accumulated priority at each service selection instant S_i.

To sketch $\mathcal{M}(t)$, one must also bear in mind some of the fundamental characteristics of the priority system, namely that $\mathcal{C}_k$s accumulate priority via Eq. (4), $\mathcal{C}_{\mathcal{N}}$s arrive to the system with zero initial priority, and $\mathcal{C}_{\mathcal{N}}$s cannot preempt service. For example, one can reproduce the sample path of $\mathcal{M}(t)$ in Figure 2 given only τ_1 and the pairs (S_i, u_i) for $i = 1, 2, \ldots, 8$.

3.2.2 Classification of the $\mathcal{C}_{\mathcal{N}}s$

Following the convention of Stanford et al. [20], we introduce some fundamental terminology pertaining to the $\mathcal{C}_{\mathcal{N}}$s arriving during busy periods. First of all, we say that a waiting $\mathcal{C}_j$ (for $j = m+1, m+2, , \ldots, k$) is at *level-$k$ accreditation* at time t if its accumulated priority lies in the interval $[M_{k+1}(t), M_k(t))$. Since priority is earned linearly throughout time, it is clear that the accumulated priority of customers at level-k accreditation must have intersected $M_{k+1}(\cdot)$ at time epochs which we refer to as *level-k accreditation instants*. Similarly, we say that a $\mathcal{C}_j$ becomes *level-k accredited* once its accumulated priority moves into the interval $[M_{k+1}(\cdot), M_k(\cdot))$ (i.e., at the corresponding accreditation instant).

Since $M_j(t)$ represents an upper bound for class-j accumulated priorities, it is obvious that the greatest accreditation a waiting $\mathcal{C}_j$ may attain is level-j accreditation. In other words, if we suppose that δ represents the first time that a $\mathcal{C}_j$ is admitted into service, then $q_\vee(\delta^+)$ (i.e., the priority of this customer upon entering into service) may only lie in one of the following intervals (assuming $j \leq k$):

$$[0, M_N(\delta^-)), [M_N(\delta^-), M_{N-1}(\delta^-)), \ldots, [M_{k+1}(\delta^-), M_k(\delta^-)), \ldots, [M_{j+1}(\delta^-), M_j(\delta^-)).$$

In addition, we say that this $\mathcal{C}_j$ is *served at level-k accreditation* if

$$q_\vee(\delta^+) \in [M_{k+1}(\delta^-), M_k(\delta^-)).$$

We use the symbol $\mathcal{C}_j^{(acc:k)}$ to refer to a $\mathcal{C}_j$ who is served at level-k accreditation. Whenever the knowledge of the specific class of customer is not required, we omit the subscript j and simply use $\mathcal{C}^{(acc:k)}$. An important result pertaining to the proportion of $\mathcal{C}_k$s who arrive during busy periods and are $\mathcal{C}^{(acc:k)}$s is provided in the next lemma.

Lemma 3.3 *The steady-state probability that a $\mathcal{C}_k$ who arrives during a busy period and is serviced at level-k accreditation (i.e., is also a $\mathcal{C}^{(acc:k)}$) is given by $1 - b_{k+1}/b_k$ for any $k \in \mathcal{N}$.*

Proof. Within every busy period, there are intervals of time during which if a $\mathcal{C}_k$ arrives within them, then it eventually would be serviced at level-k accreditation. It is not difficult to see that for every busy period, the ratio of the sum of the lengths of these intervals over the length of the busy period is always $1 - b_{k+1}/b_k$. The result then follows from the fact that $\mathcal{C}_k$s arrive to the system according to a Poisson process. □

A related notion to the general busy period has to do with level-specific accreditation intervals. A *level-k accreditation interval* is a period of time that either starts at the beginning of a busy period, or when a $\mathcal{C}^{(acc:\ell)}$ for $\ell > k$ enters into service for the first time. Regardless of how it starts, a level-k accreditation interval always ends once the system becomes clear of both the initial customer and all $\mathcal{C}^{(acc:i)}$s for $i = m+1, m+2, \ldots, k$ (i.e., all customers that have become at least level-k accredited).

Note that if δ represents the service selection instant for a $\mathcal{C}^{(acc:\ell)}$ ($\ell > k$), then this implies that $M_{k+1}(t)$ must have down-jumped at time δ (i.e., $q_{\vee}(\delta^{+}) < M_{k+1}(\delta^{-})$). In addition, if there are still customers present at the end of the ensuing level-k accreditation interval, then clearly, at this same instant, another $\mathcal{C}^{(acc:\ell)}$ for $\ell > k$ will commence service. Therefore, we observe that during busy periods, the commencement/termination instants of level-k accreditation intervals coincide with the service selection instants **S** for which $M_{k+1}(t)$ down-jumps. In other words, during busy periods, the level-k accreditation intervals are the time periods between successive down-jumps of $M_{k+1}(t)$. It is also obvious that a termination instant of a level-k accreditation interval which clears the system of all customers does not also represent a commencement instant of the next level-k accreditation interval, but rather signals the end of the busy period. Figure 3 illustrates the general structure of a level-4 accreditation interval for a 6-class mixed priority queue with $m = 2$.

Within a level-k accreditation interval, we note further that $M_k(t)$ down-jumps at instants corresponding to the service selection instants of all the $\mathcal{C}^{(acc:k)}$s. However, a down-jump of $M_k(t)$ also marks the commencement/termination of a level-$(k-1)$ accreditation interval. Therefore, a level-k accreditation interval is partitioned by a sequence of level-$(k-1)$ accreditation intervals. This suggests that it may be possible to view a level-k accreditation interval as a delay busy period of $\mathcal{C}^{(acc:k)}$s, whose effective service times are level-$(k-1)$ accreditation intervals. We show that this is precisely the case in Section 4.

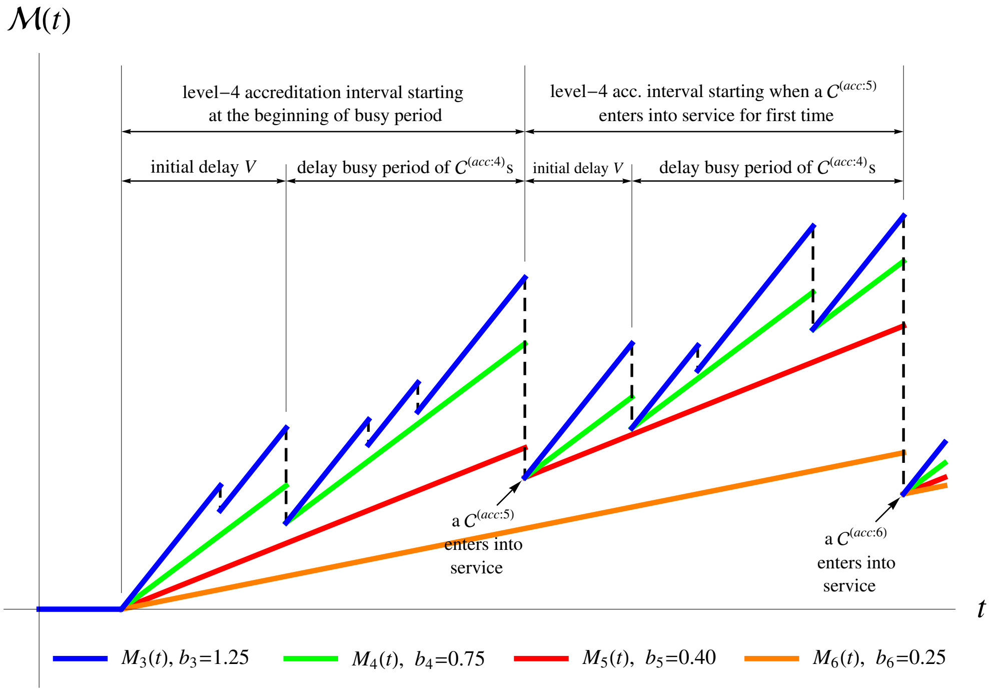

Figure 3: Level-4 accreditation intervals in a 6-class mixed priority queue with $m = 2$ (i.e., $\mathcal{N} = \{3, 4, 5, 6\}$)

We next proceed to establish the relation between level-k accreditation intervals and the previously-introduced auxiliary variables (including the completion periods). First of all, observe that of the service selection instants $\mathbf{S}$, only those resulting in a down-jump of $M_{k+1}(t)$ represent the possible selection instants for a $\mathcal{C}_{k+1}$. As a result, the end of a level-k accreditation interval also represents the instant in time that the server is ready to select a $\mathcal{C}_{k+1}$ for service. Hence, the distribution of the level-k accreditation interval depends on the class of the initial customer and is given by the corresponding auxiliary random variable. Table 1 summarizes the distributions of the types of level-k accreditation intervals, including the distribution of the initiating level-$(k-1)$ accreditation interval, which we denote by V and refer to as the initial delay of the interval.

Table 1: Distributions of the level-k accreditation intervals

Initial customer of level-k accreditation interval	Initial Delay V	Entire Interval
$\mathcal{C}_i$ for $i = 1, 2, \ldots, k$	$\Upsilon_i^{(k)}$	$\Upsilon_i^{(k+1)}$
$\mathcal{C}_{k+1}$	$\Phi_{k+1}^{(k)}$	$C^{(k+1)}$
$\mathcal{C}_i$ for $i = k+2, k+3, \ldots, N$	$\Phi_i^{(k)}$	$\Phi_i^{(k+1)}$

Finally, we end this subsection with the most vital distributional result for our overall expression of $\widetilde{W}^{(k)}(s)$. First of all, we define u_{int} to be the initial priority level of the level-k accreditation interval. Clearly, $u_{int} = 0$ if the level-k accreditation interval starts at the beginning of a busy period, and $u_{int} > 0$ if the initial customer is a $\mathcal{C}^{(acc:\ell)}$ for $\ell > k$. It is obvious that any customer who is serviced during a level-k accreditation interval must have had an accumulated priority that was greater than u_{int} immediately prior to entering service. Furthermore, the accumulated priority of a $\mathcal{C}^{(acc:k)}$ may be decomposed into two independent components – namely, the initiating priority level u_{int} and the additional priority accumulated during the accreditation interval after having accumulated priority u_{int}. Figure 4 illustrates such a decomposition of the accumulated priority for a $\mathcal{C}^{(acc:4)}$ in a 5-class mixed priority queue with

$m = 2$.

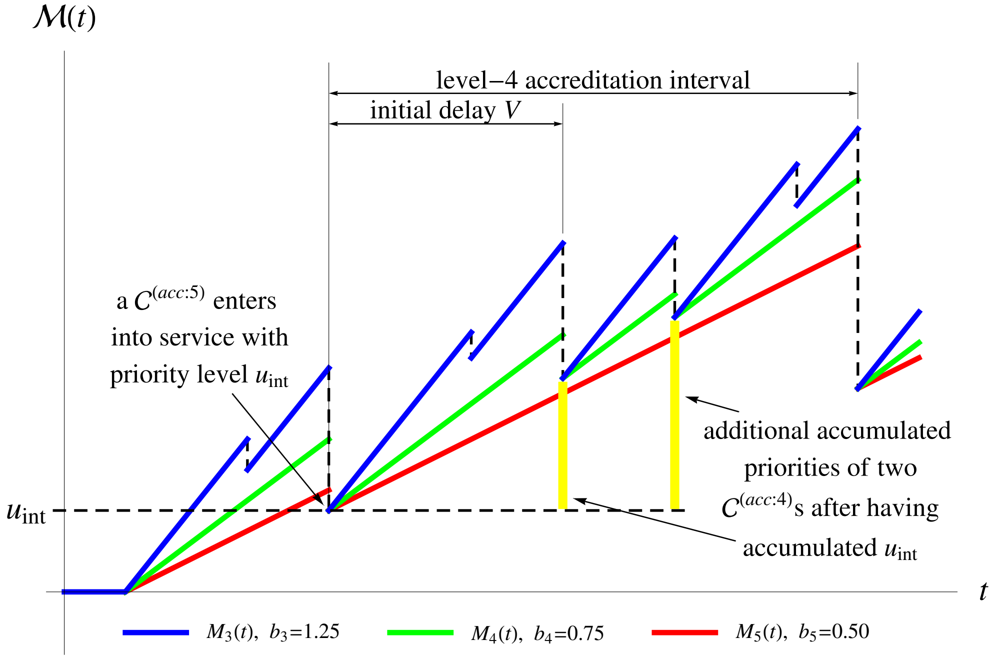

Figure 4: Decomposition of the accumulated priority for a $\mathcal{C}^{(acc:4)}$ in a 5-class mixed priority queue with $m = 2$ (i.e., $\mathcal{N} = \{3, 4, 5\}$)

Let $\mathcal{P}^{(acc:k)}$ denote the random variable representing the additional priority that a $\mathcal{C}^{(acc:k)}$ accumulates in a level-k accreditation interval after having accumulated the initial priority level. The LST of $\mathcal{P}^{(acc:k)}$, associated with an initial delay V, is given by

$$\widetilde{\mathcal{P}}^{(acc:k)}(s) \equiv \widetilde{\mathcal{P}}^{(acc:k)}(s; V) = \frac{\left(1 - \gamma_k^{(k+1)}\mu_{k,1}\right)\left(\widetilde{\mathcal{A}}(b_{k+1}s) - \widetilde{V}(b_k s)\right)}{\mathbb{E}(V)\left(1 - \frac{b_{k+1}}{b_k}\right)\left(b_k s - \gamma_k\left(1 - \widetilde{\beta}^{(k)}(b_k s)\right)\right)}, \tag{20}$$

where $\widetilde{\mathcal{A}}(s) = \widetilde{\Gamma}_0(s; \gamma_k^{(k+1)}, \beta^{(k)}, V)$ from Eq. (8),

$$\beta^{(k)}(x) = \sum_{i=m+1}^{k} \frac{\lambda_i(b_k/b_i)}{\gamma_k}\Upsilon_i^{(k)}(x), \quad \gamma_k = \sum_{i=m+1}^{k} \lambda_i(b_k/b_i), \quad \gamma_k^{(k+1)} = \gamma_k(1 - b_{k+1}/b_k),$$

and $\mu_{k,i}$ represents the i-th moment of the random variable (to be denoted by $\beta^{(k)}$) whose df is $\beta^{(k)}(x)$. Eq. (20) was first presented and proven by Stanford et al. [20]. Fajardo and Drekic [12] later provided an alternate derivation of the result using level-crossing methodology for a related $M/G/1$ queue working under a particular blocking policy known as the q-policy. To explain Eq. (20), we observe that from properties (P.2) and (P.4), it must be that the down-jumps of $M_k(t)$ during the level-k accreditation interval are exponentially distributed with parameter $\sum_{j=m+1}^{k} \lambda_j/b_j$. This ultimately leads to the key observation that the distribution of $\mathcal{P}^{(acc:k)}/b_k$ (i.e., the additional wait after having accumulated priority level u_{int}) is equivalent to that of the wait experienced by customers arriving during delay busy periods in an $M/G/1$ queue under the q-policy with the following parameters (e.g., see Fajardo and Drekic [12, Section 5]):

$$\left.\begin{array}{rl} \text{(i) arrival rate:} & \gamma_k, \\ \text{(ii) service time df:} & \beta^{(k)}(x), \\ \text{(iii) initial delay df:} & V(x), \\ \text{(iv) blocking proportion:} & q = b_{k+1}/b_k. \end{array}\right\} \tag{21}$$

The first moment of $\mathcal{P}^{(acc:k)}$ works out to be

$$\mathbb{E}(\mathcal{P}^{(acc:k)}) = b_k \left(\frac{\mathbb{E}(V^2)}{2\mathbb{E}(V)} \cdot \left[1 + \frac{b_{k+1}/b_k}{1 - \gamma_k^{(k+1)} \mu_{k,1}} \right] \right. \\ \left. + \frac{\gamma_k \mu_{k,2}}{2(1 - \gamma_k \mu_{k,1})} \cdot \left[1 - \left(\frac{b_{k+1}/b_k}{1 - \gamma_k^{(k+1)} \mu_{k,1}} \right)^2 \right] \right). \tag{22}$$

We close this subsection with the following remark.

Remark 3.4 *Note the fact that a $\mathcal{C}^{(acc:k)}$ must belong to one of the classes in $\{m+1, m+2, \ldots, k\}$. This of course implies that one $\mathcal{C}^{(acc:k)}$ may accumulate priority linearly at a different rate from another $\mathcal{C}^{(acc:k)}$ (i.e., if they each belong to two different classes). Nonetheless, the distribution of $\mathcal{P}^{(acc:k)}$ remains the same regardless of the specific class to which the $\mathcal{C}^{(acc:k)}$ belongs.*

3.2.3 A recursion for the waiting time LST

Let $P_{acc}^{(k)}$ be the accumulated priority of a $\mathcal{C}_k^{(acc:k)}$. Similarly, we define $P_{unacc}^{(k)}$ as the accumulated priority of a $\mathcal{C}_k^{(acc:\ell)}$ for some $\ell > k$. For convenience, let $\mathcal{C}_k^{(acc:>k)}$ denote a $\mathcal{C}_k^{(acc:\ell)}$ for some $\ell > k$. It therefore follows from Lemma 3.3 that

$$\widetilde{P}_{BP}^{(k)}(s) = \frac{b_k - b_{k+1}}{b_k}\widetilde{P}_{acc}^{(k)}(s) + \frac{b_{k+1}}{b_k}\widetilde{P}_{unacc}^{(k)}(s). \tag{23}$$

To obtain a recursion for (23), it follows from Remark 3.4 that $\mathcal{C}_k^{(acc:>k)}$s have an accumulated priority that is identically distributed to that of a $\mathcal{C}_{k+1}$ who arrives during a busy period, so that $\widetilde{P}_{unacc}^{(k)}(s) = \widetilde{P}_{BP}^{(k+1)}(s)$. This result is an intuitive one as both types of customers have the property that their accumulated priorities are always bounded by $M_{k+1}(t)$. We may now rewrite Eq. (23) as

$$\widetilde{P}_{BP}^{(k)}(s) = \frac{b_k - b_{k+1}}{b_k}\widetilde{P}_{acc}^{(k)}(s) + \frac{b_{k+1}}{b_k}\widetilde{P}_{BP}^{(k+1)}(s), \tag{24}$$

thereby achieving a recursive relation.

To obtain $\widetilde{P}_{acc}^{(k)}(s)$, we must consider whether the level-k accreditation interval in which the $\mathcal{C}_k^{(acc:k)}$ is serviced starts at the beginning of a busy period or at the service commencement of a $\mathcal{C}^{(acc:\ell)}$ for some $\ell > k$. We define $P_{acc,0}^{(k)}$ to be the accumulated priority of a $\mathcal{C}_k^{(acc:k)}$ serviced within a level-k accreditation interval that starts at the beginning of the busy period. We obtain the LST of $P_{acc,0}^{(k)}$ using the relation

$$\widetilde{P}_{acc,0}^{(k)}(s) = \widetilde{\mathcal{P}}^{(acc:k)}(s; V_0^{(k)}), \tag{25}$$

where $V_0^{(k)}$ is the random variable whose distribution is defined via its LST

$$\widetilde{V}_0^{(k)}(s) = \sum_{i=1}^{k} \frac{\lambda_i}{\Lambda_N}\widetilde{\Upsilon}_i^{(k)}(s) + \sum_{i=k+1}^{N} \frac{\lambda_i}{\Lambda_N}\widetilde{\Phi}_i^{(k)}(s).$$

To understand Eq. (25), note that the initial priority level of a level-k accreditation interval is zero. Therefore, the accumulated priority of a $\mathcal{C}_k^{(acc:k)}$ serviced within these kinds of level-k accreditation intervals is simply equal to the priority accumulated during the interval. Furthermore, the initial delay V_0 is a level-$(k-1)$ accreditation interval which can be initiated by any customer arriving to an empty system.

Similarly, let $P^{(k)}_{acc,1}$ represent the accumulated priority of a $\mathcal{C}^{(acc:k)}_k$ serviced within a level-k accreditation interval initiated by a $\mathcal{C}^{(acc:\ell)}$ for some $\ell > k$. An expression for the LST of $P^{(k)}_{acc,1}$ is given by

$$\widetilde{P}^{(k)}_{acc,1}(s) = \frac{\sum_{j=m+1}^{k} \pi^{(k)}_j \widetilde{P}^{(k+1)}_{BP}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s;\Upsilon^{(k)}_j) + \sum_{j=k+1}^{N} \pi^{(k)}_j \widetilde{P}^{(j)}_{BP}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s;\Phi^{(k)}_j)}{\sum_{j=m+1}^{N} \pi^{(k)}_j}, \tag{26}$$

where $\pi^{(k)}_j$ is the long-run fraction of time that the system processes a level-k accreditation interval initiated by a $\mathcal{C}_j$ ($j \in \mathcal{N}$) arriving to the system during a busy period. To understand Eq. (26), recall that the priority of a $\mathcal{C}^{(acc:k)}_k$ serviced within a level-k accreditation interval starting at the service commencement of a $\mathcal{C}^{(acc:\ell)}$ for some $\ell > k$ can be decomposed into two independent components: u_{int}, the accumulated priority of the initiating $\mathcal{C}^{(acc:\ell)}$, and $\mathcal{P}^{(acc:k)}$, the additional priority accumulated after having accumulated the initial priority level u_{int}. Hence, the accumulated priority of such a $\mathcal{C}^{(acc:k)}_k$ has LST which takes on the general form

$$\widetilde{P}^{(k)}_{acc,1}(s;V) = \widetilde{u}_{int}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s;V),$$

where V is the initial delay of the level-k accreditation interval.

The distributions of both u_{int} and V depend solely on the class of the initial customer. In particular, if the initial customer is of class j for $m < j \leq k$, then $\widetilde{u}_{int}(s) = \widetilde{P}^{(k+1)}_{BP}(s)$ and $\widetilde{V}(s) = \widetilde{\Upsilon}^{(k)}_j(s)$. Otherwise, for $j > k$, $\widetilde{u}_{int}(s) = \widetilde{P}^{(j)}_{BP}(s)$ and $\widetilde{V}(s) = \widetilde{\Phi}^{(k)}_j(s)$. If we define $\pi^{(k)}_0$ as the long-run fraction of time that the system spends processing a level-k accreditation interval initiated by a customer who arrived to an empty queue, then it must be that

$$\widetilde{P}^{(k)}_{acc}(s) = \frac{1}{\rho}\big(\pi^{(k)}_0 \widetilde{P}^{(k)}_{acc,0}(s) + (\rho - \pi^{(k)}_0)\widetilde{P}^{(k)}_{acc,1}(s)\big). \tag{27}$$

Eqs. (24)–(27) together provide a recursive method to obtain $\widetilde{P}^{(k)}_{BP}(s)$.

We end this section with the derivation of the steady-state probabilities $\pi^{(k)}_j$ for $j \in \{0, m+1, m+2, \ldots, N\}$. First of all, it is clear that any $\mathcal{C}_j$ ($j > k$) arriving during a busy period will eventually initiate a level-k accreditation interval with an initial delay of $\Phi_j^{(k)}$. Hence, we have

$$\pi_j^{(k)} = \rho \frac{\lambda_j \mathbb{E}(\Phi_j^{(k)})}{1-\gamma_k^{(k+1)}\mu_{k,1}}, \qquad j > k. \tag{28}$$

Next, for a $\mathcal{C}_j$ ($m < j \leq k$) to initiate a level-k accreditation interval, this customer must be served at level-ℓ accreditation for some $\ell > k$. The probability of such a $\mathcal{C}_j$ arriving to the system is $\rho(b_{k+1}/b_j)$. Furthermore, since the initial delay of the resulting level-k accreditation interval is $\Upsilon_j^{(k)}$, we have that

$$\pi_j^{(k)} = \rho \frac{\lambda_j (b_{k+1}/b_j) \mathbb{E}(\Upsilon_j^{(k)})}{1-\gamma_k^{(k+1)}\mu_{k,1}}, \qquad m < j \leq k. \tag{29}$$

Finally, a $\mathcal{C}_j$ arriving to an empty system initiates a level-k accreditation interval whose initial delay is either $\Upsilon_j^{(k)}$ if $j \leq k$ or $\Phi_j^{(k)}$ if $j > k$. Thus,

$$\pi_0^{(k)} = \frac{1-\rho}{1-\gamma_k^{(k+1)}\mu_{k,1}} \left[\sum_{j=1}^{k} \lambda_j \mathbb{E}(\Upsilon_j^{(k)}) + \sum_{j=k+1}^{N} \lambda_j \mathbb{E}(\Phi_j^{(k)}) \right]. \tag{30}$$

Since level-k accreditation intervals partition the general busy period, it is clear that $\pi_0^{(k)} + \sum_{j=m+1}^{N} \pi_j^{(k)} = \rho$.

4 Characterization of the service-structure elements and auxiliary random variables

In this section, we derive expressions for the LSTs of class-k completion periods, residence periods, and the auxiliary random variables introduced earlier in the paper. Since the preemptive resume service discipline is a work-conserving one, it is straightforward to show that the LSTs of the class-k ($k \in \mathcal{U}$) auxiliary random variables are given by

$$\widetilde{\Upsilon}_i^{(k)}(s) = \widetilde{B}^{(i)}\big(s + \Lambda_{k-1}(1 - \widetilde{\Upsilon}_{1:k-1}^{(k)}(s))\big), \qquad i \in a \tag{31}$$

and

$$\widetilde{\Phi}_i^{(k)}(s) = \widetilde{Z}_k^{(i)}\big(s + \Lambda_{k-1}(1 - \widetilde{\Upsilon}_{1:k-1}^{(k)}(s))\big), \qquad i \in b, \tag{32}$$

where $\widetilde{\Upsilon}_{1:k-1}^{(k)} = \widetilde{\Gamma}\big(s;\Lambda_{k-1},\sum_{i=1}^{k-1}(\lambda_i/\Lambda_{k-1})X^{(i)}\big)$ from Eq. (7) is the busy period LST of $\mathcal{C}_a$s and $Z_k^{(i)}$ represents the class-k protected portion of a class-i service. Table 2 reports the various forms of $Z_k^{(i)}$ and $\widetilde{Z}_k^{(i)}(s)$ under each of the three threshold-based discretion rules. Moreover, the class-k completion period LST is simply given by

$$\widetilde{C}^{(k)}(s) = \widetilde{\Upsilon}_k^{(k)}(s) = \widetilde{B}^{(k)}\big(s+\Lambda_{k-1}(1-\widetilde{\Upsilon}_{1:k-1}^{(k)}(s))\big). \tag{33}$$

Table 2: Various forms of $Z_k^{(i)}$ and its corresponding LST

Threshold Rule	$Z_k^{(i)}$	$\widetilde{Z}_k^{(i)}$
PB	$(1-\alpha_{k,i})X^{(i)}$	$\widetilde{B}^{(i)}\big((1-\alpha_{k,i})s\big)$
FETB	$(X^{(i)}-T_{k,i})\vert(X^{(i)}>T_{k,i})$	$\big(\int_{x=T_{k,i}}^{\infty} e^{-s(x-T_{k,i})}\mathrm{d}B^{(i)}(x)\big)/\big(1-B^{(i)}(T_{k,i})\big)$
TETB	$\min\{X^{(i)},\tau_{k,i}\}$	$e^{-s\tau_{k,i}}(1-B^{(i)}(\tau_{k,i}))+\int_{x=0}^{\tau_{k,i}} e^{-sx}\mathrm{d}B^{(i)}(x)$

For the case $k\in\mathcal{N}$, both $\widetilde{\Upsilon}_i^{(k)}(s)$ and $\widetilde{\Phi}_i^{(k)}(s)$ are obtained recursively. Specifically, we have the following recursive schemes for each $k\geq m+1$:

$$\widetilde{\Upsilon}_i^{(k+1)}(s) = \widetilde{\Upsilon}_i^{(k)}\big(s+\gamma_k^{(k+1)}(1-\widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s))\big), \qquad i\leq k \tag{34}$$

and

$$\widetilde{\Phi}_i^{(k+1)}(s) = \widetilde{\Phi}_i^{(k)}\big(s+\gamma_k^{(k+1)}(1-\widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s))\big), \qquad i>k+1, \tag{35}$$

where $\widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s) = \widetilde{\Gamma}(s;\gamma_k^{(k+1)},\beta^{(k)})$ from Eq. (7). Furthermore, the class-$(k+1)$ completion period LST is given by

$$\widetilde{C}^{(k+1)}(s) = \widetilde{\Upsilon}_{k+1}^{(k+1)}(s) = \widetilde{\Phi}_{k+1}^{(k)}\big(s+\gamma_k^{(k+1)}(1-\widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s))\big). \tag{36}$$

The respective starting points for the recursive expressions given in Eqs. (34)–(36) are $\widetilde{\Upsilon}_i^{(m+1)}(s)$ for all $i\leq m+1$, $\widetilde{\Phi}_i^{(m+1)}(s)$ for all $i>m+2$, and $\widetilde{\Phi}_{m+2}^{(m+1)}(s)$. Since $\mathcal{U}$ also represents the set of classes which have priority over class $m+1$, it turns out that the formulas of $\widetilde{\Upsilon}_i^{(k)}(s)$, $\widetilde{\Phi}_i^{(k)}(s)$, and $\widetilde{C}^{(k)}(s)$ given by Eqs. (31)–(33) also hold true when $k=m+1$. Note that in using Eq.

(32) with $k = m+1$, it is necessary to define the threshold parameters $\alpha_{m+1,i} = 0$, $T_{m+1,i} = 0$, and $\tau_{m+1,i} = \infty$ for all $i > m+1$.

The above formulas illustrate the fact that a level-k accreditation interval is merely a delay busy period of $\mathcal{C}^{(acc:k)}$s whose service times are level-$(k-1)$ accreditation intervals, corresponding to $\Upsilon_i^{(k)}$ for $i = m+1, m+2, \ldots, k$. This result follows from the observation that during a level-k accreditation interval, the k-th and $(k+1)$-th components of the maximal priority process $(M_{k+1}(t), M_k(t))$ behave like the maximal priority process of that for an $M/G/1$ queue under the q-policy whose parameters are given by Eq. (21).

To obtain $\widetilde{R}^{(k)}(s)$, we require the joint transform of the preemptible and non-preemptible periods of a class-k service time. In particular, similar to the analysis conducted by Drekic and Stanford [10], we segment the class-k service time $X^{(k)}$ into its preemptible portion $X_p^{(k)}$ and its non-preemptible (or protected) portion $X_{p_0}^{(k)}$. For our new model, however, we must further partition the preemptible portion $X_p^{(k)}$ as follows:

$$X_p^{(k)} = X_{p_{k-1}}^{(k)} + X_{p_{k-2}}^{(k)} + \cdots + X_{p_1}^{(k)},$$

where $X_{p_i}^{(k)}$, $i \in a$, represents the portion of the class-k service time which is preemptible only by a $\mathcal{C}_j$ with $j \in \{1, 2, \ldots, i\}$. It is important to note that $X_{p_i}^{(k)} = 0$ for $i \in a_{np}$. Furthermore, for the purpose of formulating a single expression for $\widetilde{R}^{(k)}(s)$ that holds true for both $k \in \mathcal{U}$ and $k \in \mathcal{N}$, we define $\alpha_{i,k} = 0$, $T_{i,k} = 0$, and $\tau_{i,k} = \infty$ if $i = k$ or if $i < k$ and $i \in \mathcal{N}$.

If we let $\mathbf{s} = [s_1, s_2, \ldots, s_{k-1}, s_0]$ be a $k-$dimensional row vector, then the joint transform of all the portions of $X^{(k)}$ is given by

$$\Theta^{(k)}(\mathbf{s}) = \mathbb{E}(e^{-s_1 X_{p_1}^{(k)} - s_2 X_{p_2}^{(k)} - \cdots - s_{k-1} X_{p_{k-1}}^{(k)} - s_0 X_{p_0}^{(k)}}).$$

We remark that the above transform depends on the specific threshold-based discretion rule in effect for the $\mathcal{C}_k$s. Hence, we have three expressions for $\Theta^{(k)}(\mathbf{s})$, each of which is readily obtained by conditioning on $X^{(k)} = x$ and subsequently characterizing $X_{p_i}^{(k)}$ via the corresponding threshold parameters $\alpha_{i,k}$, $T_{i,k}$, and $\tau_{i,k}$ for each $i \in a$. The expressions for $\Theta^{(k)}(\mathbf{s})$ are as follows:

$$\begin{aligned}\text{(PB)} \quad \Theta^{(k)}(\mathbf{s}) &= \int_{x=0}^{\infty} e^{-(\sum_{i=1}^{k-1} s_i(\alpha_{i,k}-\alpha_{i+1,k})+s_0(1-\alpha_{1,k}))x} \mathrm{d}B^{(k)}(x) \\ &= \widetilde{B}^{(k)}\big(\textstyle\sum_{i=1}^{k-1} s_i(\alpha_{i,k}-\alpha_{i+1,k})+s_0(1-\alpha_{1,k})\big), \end{aligned} \tag{37}$$

$$\begin{aligned}\text{(FETB)} \quad \Theta^{(k)}(\mathbf{s}) = \sum_{i=1}^{k-1} e^{-\sum_{j=i+1}^{k-1}(s_j-s_{j-1})T_{j,k}} \int_{x=T_{i+1,k}}^{T_{i,k}} e^{-s_i x} \mathrm{d}B^{(k)}(x) \\ + e^{-(\sum_{j=2}^{k-1}(s_j-s_{j-1})T_{j,k}+(s_1-s_0)T_{1,k})} \int_{x=T_{1,k}}^{\infty} e^{-s_0 x} \mathrm{d}B^{(k)}(x), \end{aligned} \tag{38}$$

and

$$\begin{aligned}\text{(TETB)} \quad \Theta^{(k)}(\mathbf{s}) = \sum_{i=1}^{k-1} e^{-(\sum_{j=2}^{i}(s_{j-1}-s_j)\tau_{j,k}+(s_0-s_1)\tau_{1,k})} \int_{x=\tau_{i,k}}^{\tau_{i+1,k}} e^{-s_i x} \mathrm{d}B^{(k)}(x) \\ + \int_{x=0}^{\tau_{1,k}} e^{-s_0 x} \mathrm{d}B^{(k)}(x). \end{aligned} \tag{39}$$

During a class-k residence period, only those $\mathcal{C}_a$s participating in the interruption periods extend the overall residence period. Therefore, we obtain

$$\widetilde{R}^{(k)}(s) = \Theta^{(k)}\big(\textstyle\sum_{i=1}^{k-1} \mathbf{1}_i(s+\Lambda_i(1-\widetilde{A}_{p_i}^{(k)}(s))) + s\mathbf{1}_k\big), \tag{40}$$

where $\mathbf{1}_i$ denotes a k-dimensional row vector whose i-th entry is one and all other entries are zero, and $A_{p_i}^{(k)}$ represents an interruption period occurring within the $X_{p_i}^{(k)}$ portion of the class-k service time (i.e., an interruption period in which only $\mathcal{C}_j$s for $j \leq i$ can participate). From Eq. (7), we ultimately have

$$\widetilde{A}_{p_i}^{(k)}(s) = \widetilde{\Gamma}\big(s; \Lambda_i, \textstyle\sum_{j=1}^{i}(\lambda_j/\Lambda_i)X^{(j)}\big). \tag{41}$$

5 Numerical examples

We now present two numerical examples which illustrate the potential use of our mixed priority queueing model. Our first example takes inspiration from the example found in Stanford et al. [20]. The Canadian Triage and Acuity Scale (CTAS) provides five priority classifications for the triage assessment of patients arriving to a hospital emergency room. Furthermore, each class is given a "time to assessment" standard and an accompanying compliance target, which specifies the desired proportion of that class's patients to meet the standard. Table 3 reports these time to assessment standards along with their compliance targets, as indicated in Stanford et al. [20, p. 299].

Table 3: CTAS key performance indicators

Category	Class	Time to Assessment	Compliance Target (%)
1	Resuscitation	Immediate	98
2	Emergent	15 minutes	95
3	Urgent	30 minutes	90
4	Less Urgent	60 minutes	85
5	Not Urgent	120 minutes	80

As an attempt to meet these standards, we model an emergency room whose 5 classes of patients are defined by the CTAS by invoking a mixed priority queueing scheme with $m = 3$ (i.e., $\mathcal{U} = \{1, 2, 3\}$ and $\mathcal{N} = \{4, 5\}$). The service times corresponding to each patient class are assumed to be exponentially distributed with mean times of 30 minutes for class 1, 20 minutes for classes 2 and 3, and 10 minutes for classes 4 and 5. We assume further that the server (or doctor) implements a PB rule to govern how preemptions to patients take place. For the Resuscitation class, we assume that $\alpha_{1,i} = 1$ for $i = 2, 3, 4, 5$ (i.e., $\mathcal{C}_1$s always preempt lower priority customers). We consider several different values for the other threshold parameters such as $\alpha_{2,i}$ for $i = 3, 4, 5$ and $\alpha_{3,i}$ for $i = 4, 5$. The remaining parameters of the system correspond to the accumulating priority rates of the $\mathcal{C}_\mathcal{N}$s for which we assume $b_4 = 1$ and $0 \leq b_5 \leq 1$.

For each $k = 1, 2, \ldots, 5$, we are interested in calculating $P(W^{(k)} \leq t_k)$, where t_k denotes the class-k time to assessment standard given in Table 3. To do this, we numerically invert $\widetilde{W^{(k)}}(s)$ by employing the EULER and POST-WIDDER algorithms of Abate and Whitt [2] with their suggested parameter settings (and found that the two methods produced equivalent results). We remark that in conducting the numerical inversions, there were several instances for which implicit functionals of LSTs (resembling those of an $M/G/1$ busy period) had to be evaluated at complex arguments. This was performed following the iterative procedure outlined in Abate and Whitt [1]. In addition to reporting the desired probabilities, we provide the mean class-k waiting times and flow times for $k = 1, 2, \ldots, 5$. The results under three separate settings are tabulated to 4 decimal places of accuracy in Table 4. Note also that the reported values are given in scaled multiples of 10 minutes.

In their example, Stanford et al. [20] analyzed a 2-class APQ, modelling only CTAS classes 4 and 5. In our treatment, we utilized the same arrival rates and service rates for the two lowest priority classes as in their example. Moreover, they determined that without the presence of the three highest priority classes, the CTAS 4 and 5 compliance targets were both met as long as the accumulating priority rate of the lowest class did not exceed 0.5. As evidenced by the results in Table 4, this is not the case for our 5-class priority model. In fact, of the three settings considered, only in Setting 3, where the arrival rates of the 3 highest priority classes are the smallest, were all the CTAS compliance targets satisfied. It is also interesting to observe the changes in the mean flow times under the various settings.

In our second example, we consider the 9-class mixed priority queue studied by Paterok and Ettl [19, pp. 1157–1159]. The arrival rates and service time distributions, including the *priority group* of each class, are given in Table 5. Priority groups are used to specify the type of priority that the higher priority customers have over lower priority ones. In particular, a $\mathcal{C}_i$ has preemptive priority over a $\mathcal{C}_j$ $(i < j)$ if they belong to different priority groups; otherwise, the $\mathcal{C}_i$ has only non-preemptive priority over the $\mathcal{C}_j$. It is straightforward to obtain these specific priority relations using our mixed priority model. For example, if we define $\alpha_{(r,s)}$, $T_{(r,s)}$, and $\tau_{(r,s)}$ for all $1 \leq r < s \leq 3$ as the threshold-based discretion parameters between priority groups

Table 4: Performance measures in Example 1 under various settings

Setting 1 ($\rho = 0.863$)

$\alpha_{2,3} = 0.9$, $\alpha_{2,4} = 1$, $\alpha_{2,5} = 1$, $\alpha_{3,4} = 0.5$, $\alpha_{3,5} = 0.75$, and $b_5 = 0.10$

Class k	λ_k	$P(W^{(k)} \leq t_k)$	$\mathbb{E}(W^{(k)})$	$\mathbb{E}(F^{(k)})$
1	0.001	0.9970	0.0090	3.0090
2	0.01	0.9885	0.0511	2.0571
3	0.02	0.9815	0.2775	2.3204
4	0.4	0.8873	2.7217	3.7671
5	0.4	0.6590	11.7522	12.8085

Setting 2 ($\rho = 0.833$)

$\alpha_{2,3} = 0.75$, $\alpha_{2,4} = 0.9$, $\alpha_{2,5} = 1$, $\alpha_{3,4} = 0.25$, $\alpha_{3,5} = 0.5$, and $b_5 = 0.30$

Class k	λ_k	$P(W^{(k)} \leq t_k)$	$\mathbb{E}(W^{(k)})$	$\mathbb{E}(F^{(k)})$
1	0.001	0.9970	0.0090	3.0090
2	0.005	0.9931	0.0361	2.0421
3	0.01	0.9832	0.4128	2.4341
4	0.4	0.8308	3.1880	4.2054
5	0.4	0.7781	7.5744	8.5980

Setting 3 ($\rho = 0.815$)

$\alpha_{2,3} = 0.5$, $\alpha_{2,4} = 0.75$, $\alpha_{2,5} = 1$, $\alpha_{3,4} = 0.25$, $\alpha_{3,5} = 0.5$, and $b_5 = 0.275$

Class k	λ_k	$P(W^{(k)} \leq t_k)$	$\mathbb{E}(W^{(k)})$	$\mathbb{E}(F^{(k)})$
1	0.001	0.9970	0.0090	3.0090
2	0.001	0.9958	0.0433	2.0494
3	0.005	0.9891	0.3652	2.3733
4	0.4	0.8795	2.6638	3.6709
5	0.4	0.8175	6.4787	7.4888

(e.g., $\tau_{i,j} = \tau_{(r,s)}$ whenever a $\mathcal{C}_i$ belongs to priority group r and a $\mathcal{C}_j$ belongs to priority group s), then the desired priority relations are achieved by considering a 9-class mixed priority model with $m = 6$ and the following threshold parameters: $\alpha_{(r,s)} = 1$, $T_{(r,s)} = \infty$, and $\tau_{(r,s)} = 0$ for all $r < s$. We note that in their analysis, Paterok and Ettl [19] used a 15-class priority queue for which the arrival rates of six of the classes were set equal to zero in order to obtain the desired priority relations.

Table 5: Parameters of the Paterok and Ettl [19] example

Class k	Priority Group	λ_k	$\mathbb{E}(X^{(k)})$	Service Time Distribution
1	1	0.062	0.5	Exponential
2	1	0.040	1.0	Erlang-2
3	2	0.020	4.0	Erlang-2
4	2	0.010	3.0	Erlang-3
5	2	0.030	5.0	Exponential
6	2	0.020	4.0	Erlang-2
7	3	0.003	3.0	Exponential
8	3	0.005	6.0	Erlang-3
9	3	0.010	5.0	Erlang-2

We define the weighted average flow time as $\overline{F} = \sum_{i=1}^{9}(\lambda_i/\Lambda_9)\mathbb{E}(F^{(i)})$, and similarly let $\overline{F}_i$ represent the weighted average flow time of classes belonging to priority group i, $i = 1, 2, 3$. In our numerical study, we report the expected flow times of each class, as well as the weighted average flow times under various settings for each of the threshold-based discretion rules. The results for the original Paterok and Ettl [19] setting (referred to as the resume-IPF case, where IPF denotes "interrupted processing first") are tabulated to 3 decimal places of accuracy in Table 6. The results for the PB, FETB, and TETB rules are provided in Tables 7, 8, and 9, respectively.

For the $\mathcal{C}_\mathcal{N}$s, we implement accumulating priority rates of the form $b_7 = 1$, $b_8 = e^{-x}$, and $b_9 = e^{-2x}$ for some $x \geq 0$. We note that as $x \to \infty$, the resulting accumulating prioritization becomes equivalent to that of the static non-preemptive priority service discipline. Conversely, with $x = 0$, the $\mathcal{C}_\mathcal{N}$s are serviced according to their order of arrival (i.e., regardless of the specific class to which they belong). As a consequence of having $x = 0$, the mean waiting times for each class belonging to the lowest priority group would all be identical – a potentially desirable setting. In Tables 6–9, we compute mean flow times for each of the non-urgent classes using $x = 0.1, 1, 10$. We emphasize that by fine-tuning the parameter x, a systems manager is able to achieve a desired balance between the two extremes of FCFS and static non-preemptive priority between the $\mathcal{C}_{\mathcal{N}}$s. We also note that the mean flow times of the $\mathcal{C}_{\mathcal{U}}$s are unaffected by the choice of x.

It is evident from the results in Tables 7–9 that the new priority model is quite flexible. In testing several different parameter values for each of the threshold-based discretion rules, we are, in some instances, able to achieve a lower overall weighted average flow time $\overline{F}$. Furthermore, if instead a systems manager is more concerned with reducing the average flow time of the lowest priority group $\overline{F}_3$, and is less concerned with minimizing $\overline{F}$, then it is clear that our priority model can achieve this objective while still maintaining reasonable weighted average flow times for both $\overline{F}_1$ and $\overline{F}_2$.

6 Concluding remarks

In this paper, we introduced a new general mixed priority queueing model for which we obtained the LST of the steady-state distribution of the class-k waiting time. This model is quite flexible in supplying a systems manager the ability to control both the waiting time distributions and the flow time distributions of each class. This control is administered through the fine-tuning of the threshold-based discretion parameters and the accumulating priority rates $\{b_i\}_{i=m+1}^{N}$.

Furthermore, under various parameter settings, our mixed priority queueing model includes a number of previously-analyzed priority queueing models as special cases. For example, by setting $m = 0$, our priority model exactly becomes the one considered by Stanford et al. [20].

Table 6: Mean flow times in Example 2 under the original Paterok and Ettl [19] setting

		Paterok and Ettl (resume-IPF case)	
Class k		$\mathbb{E}(F^{(k)})$	
1		0.547	
2		1.051	
3		5.999	
4		5.150	
5		7.820	
6		7.695	
	$x = 10$	$x = 1$	$x = 0.1$
7	9.982	10.154	10.649
8	15.422	15.591	15.819
9	14.562	14.429	14.203
$\overline{F}$	4.443	4.443	4.445
$\overline{F}_3$	14.037	14.039	14.060
	$\overline{F}_1 = 0.744$	$\overline{F}_2 = 7.000$	

Table 7: Mean flow times in Example 2 under PB rule

			PB rule			
		$\alpha_{(1,2)} = \alpha_{(2,3)} = 0.70$, $\alpha_{(1,3)} = 0.85$			$\alpha_{(1,2)} = \alpha_{(2,3)} = 0.50$, $\alpha_{(1,3)} = 0.75$	
Class k		$\mathbb{E}(F^{(k)})$			$\mathbb{E}(F^{(k)})$	
1		0.675			0.901	
2		1.188			1.432	
3		5.945			5.952	
4		5.124			5.156	
5		7.760			7.781	
6		7.680			7.754	
	$x = 10$	$x = 1$	$x = 0.1$	$x = 10$	$x = 1$	$x = 0.1$
7	9.388	9.560	10.055	8.992	9.165	9.659
8	14.235	14.404	14.632	13.443	13.612	13.841
9	13.572	13.440	13.214	12.913	12.780	12.554
$\overline{F}$	4.405	4.405	4.407	4.478	4.478	4.480
$\overline{F}_3$	13.059	13.061	13.081	12.407	12.409	12.429
	$\overline{F}_1 = 0.876$	$\overline{F}_2 = 6.957$		$\overline{F}_1 = 1.109$	$\overline{F}_2 = 6.989$	

Table 8: Mean flow times in Example 2 under FETB rule

	FETB rule					
	$T_{(1,2)} = T_{(2,3)} = 5, T_{(1,3)} = 10$			$T_{(1,2)} = T_{(2,3)} = 2, T_{(1,3)} = 4$		
Class k	$\mathbb{E}(F^{(k)})$			$\mathbb{E}(F^{(k)})$		
1	0.922			1.435		
2	1.455			2.006		
3	6.037			6.072		
4	5.244			5.331		
5	7.815			7.911		
6	7.828			8.010		
	$x = 10$	$x = 1$	$x = 0.1$	$x = 10$	$x = 1$	$x = 0.1$
7	9.622	9.794	10.289	8.964	9.136	9.631
8	14.261	14.430	14.658	12.721	12.890	13.119
9	13.700	13.567	13.341	12.468	12.336	12.110
$\overline{F}$	4.584	4.584	4.586	4.783	4.783	4.785
$\overline{F}_3$	13.176	13.178	13.198	11.955	11.957	11.977
	$\overline{F}_1 = 1.131 \quad \overline{F}_2 = 7.053$			$\overline{F}_1 = 1.659 \quad \overline{F}_2 = 7.153$		

Table 9: Mean flow times in Example 2 under TETB rule

	TETB rule					
	$\tau_{(1,2)} = \tau_{(2,3)} = 1.0, \tau_{(1,3)} = 0.50$			$\tau_{(1,2)} = \tau_{(2,3)} = 2.0, \tau_{(1,3)} = 0.15$		
Class k	$\mathbb{E}(F^{(k)})$			$\mathbb{E}(F^{(k)})$		
1	0.587			0.682		
2	1.094			1.196		
3	5.936			5.902		
4	5.088			5.059		
5	7.766			7.751		
6	7.643			7.638		
	$x = 10$	$x = 1$	$x = 0.1$	$x = 10$	$x = 1$	$x = 0.1$
7	9.418	9.590	10.085	9.063	9.236	9.731
8	14.765	14.934	15.162	14.197	14.366	14.594
9	13.916	13.784	13.557	13.398	13.265	13.039
$\overline{F}$	4.384	4.384	4.386	4.381	4.381	4.383
$\overline{F}_3$	13.402	13.404	13.424	12.897	12.899	12.920
	$\overline{F}_1 = 0.786 \quad \overline{F}_2 = 6.943$			$\overline{F}_1 = 0.884 \quad \overline{F}_2 = 6.924$		

By setting $m = N$ and assigning threshold parameters to be $\alpha_{i,k} = \alpha_k$, $T_{i,k} = T_k$, and $\tau_{i,k} = \tau_k$, our priority model is equivalent to the one considered by Drekic and Stanford [10]. Moreover, by setting $m = N$ and using threshold parameters of the form

$$\alpha_{i,k} = \begin{cases} 1 & \text{if } k - i \geq d \\ 0 & \text{otherwise} \end{cases}, \quad T_{i,k} = \begin{cases} \infty & \text{if } k - i \geq d \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad \tau_{i,k} = \begin{cases} 0 & \text{if } k - i \geq d \\ \infty & \text{otherwise} \end{cases},$$

our priority model is equivalent to the one using the PD rule (resume-IPF case) as analyzed by Paterok and Ettl [19], where d is the so-called preemption distance parameter. Finally, it is also evident that the classical non-preemptive and preemptive priority queues, as well as the $\sum_{i=1}^{N} M_i/G_i/1$ FCFS queue, are all special cases of our general model.

In terms of future work, a possible extension to this model involves the case where the urgent class of customers also accumulates priority via Eq. (2). Furthermore, a variation of our mixed priority queueing model which employs a preemptive repeat (identical or different) service discipline may also be considered. Due to the non-work-conserving nature of the repeat service rule, however, such a model would likely necessitate more involved recursive schemes to obtain the class-k waiting time LST.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix. Moments of the service-structure elements and auxiliary random variables

The first two moments of the auxiliary random variables introduced in Section 2.3 can be obtained in a straightforward fashion by either differentiating their corresponding LSTs, or by applying the well-known formulas for the first two moments of an $M/G/1$ delay busy period (e.g., see Conway et al. [9, p. 151]) with the appropriate parameters. Letting $\overline{U}_k = \sum_{i=1}^{k} \lambda_i \mathbb{E}(X^{(i)})$, we obtain for $k = 1, 2, \ldots, m+1$:

$$\begin{aligned}
\mathbb{E}(\Upsilon_i^{(k)}) &= \frac{\mathbb{E}(X^{(i)})}{1-\overline{U}_{k-1}}, \qquad i \le k, \\
\mathbb{E}\big((\Upsilon_i^{(k)})^2\big) &= \frac{\sum_{j=1}^{k-1} \lambda_j \mathbb{E}\big((X^{(j)})^2\big)}{(1-\overline{U}_{k-1})^3}\mathbb{E}(X^{(i)}) + \frac{\mathbb{E}\big((X^{(i)})^2\big)}{(1-\overline{U}_{k-1})^2}, \qquad i \le k, \\
\mathbb{E}(\Phi_i^{(k)}) &= \frac{\mathbb{E}(Z_k^{(i)})}{1-\overline{U}_{k-1}}, \qquad i > k, \\
\mathbb{E}\big((\Phi_i^{(k)})^2\big) &= \frac{\sum_{j=1}^{k-1} \lambda_j \mathbb{E}\big((X^{(j)})^2\big)}{(1-\overline{U}_{k-1})^3}\mathbb{E}(Z_k^{(i)}) + \frac{\mathbb{E}\big((Z_k^{(i)})^2\big)}{(1-\overline{U}_{k-1})^2}, \qquad i > k.
\end{aligned}$$

For the case $k > m+1$, the first two moments are computed recursively. In particular, we have for $k = m+1, m+2, \ldots, N$:

$$\begin{aligned}
\mathbb{E}(\Upsilon_i^{(k+1)}) &= \frac{\mathbb{E}(\Upsilon_i^{(k)})}{1-\gamma_k^{(k+1)}\mu_{k,1}}, \qquad i \le k, \\
\mathbb{E}\big((\Upsilon_i^{(k+1)})^2\big) &= \frac{\gamma_k^{(k+1)}\mu_{k,2}}{(1-\gamma_k^{(k+1)}\mu_{k,1})^3}\mathbb{E}(\Upsilon_i^{(k)}) + \frac{\mathbb{E}\big((\Upsilon_i^{(k)})^2\big)}{(1-\gamma_k^{(k+1)}\mu_{k,1})^2}, \qquad i \le k, \\
\mathbb{E}(\Phi_i^{(k+1)}) &= \frac{\mathbb{E}(\Phi_i^{(k)})}{1-\gamma_k^{(k+1)}\mu_{k,1}}, \qquad i > k+1, \\
\mathbb{E}\big((\Phi_i^{(k+1)})^2\big) &= \frac{\gamma_k^{(k+1)}\mu_{k,2}}{(1-\gamma_k^{(k+1)}\mu_{k,1})^3}\mathbb{E}(\Phi_i^{(k)}) + \frac{\mathbb{E}\big((\Phi_i^{(k)})^2\big)}{(1-\gamma_k^{(k+1)}\mu_{k,1})^2}, \qquad i > k+1, \\
\mathbb{E}(\Upsilon_{k+1}^{(k+1)}) &= \frac{\mathbb{E}(\Phi_{k+1}^{(k)})}{1-\gamma_k^{(k+1)}\mu_{k,1}}, \\
\mathbb{E}\big((\Upsilon_{k+1}^{(k+1)})^2\big) &= \frac{\gamma_k^{(k+1)}\mu_{k,2}}{(1-\gamma_k^{(k+1)}\mu_{k,1})^3}\mathbb{E}(\Phi_{k+1}^{(k)}) + \frac{\mathbb{E}\big((\Phi_{k+1}^{(k)})^2\big)}{(1-\gamma_k^{(k+1)}\mu_{k,1})^2}.
\end{aligned}$$

Similarly, the following expression for the first moment of $A_{p_i}^{(k)}$ is obtained:

$$\mathbb{E}(A_{p_i}^{(k)}) = \frac{\overline{U}_i}{\Lambda_i(1-\overline{U}_i)}, \qquad i < k.$$

For $k = 1, 2, \ldots, N$, expressions for the first two moments of $Z_k^{(i)}$ and the mean of $R^{(k)}$ under each threshold-based discretion rule are as follows:

PB rule

$$\begin{aligned}
\mathbb{E}(Z_k^{(i)}) &= (1-\alpha_{k,i})\mathbb{E}(X^{(i)}), \qquad i>k,\\
\mathbb{E}\big((Z_k^{(i)})^2\big) &= (1-\alpha_{k,i})^2\mathbb{E}\big((X^{(i)})^2\big), \qquad i>k,\\
\mathbb{E}(R^{(k)}) &= \mathbb{E}(X^{(k)})\Big[\sum_{i=1}^{k-1}\big(1+\Lambda_i\mathbb{E}(A_{p_i}^{(k)})\big)\cdot(\alpha_{i,k}-\alpha_{i+1,k})+(1-\alpha_{1,k})\Big].
\end{aligned}$$

FETB rule

$$\begin{aligned}
\mathbb{E}(Z_k^{(i)}) &= \bigg(\int_{x=T_{k,i}}^{\infty}(x-T_{k,i})\,\mathrm{d}B^{(i)}(x)\bigg)/(1-B^{(i)}(T_{k,i})), \qquad i>k,\\
\mathbb{E}\big((Z_k^{(i)})^2\big) &= \bigg(\int_{x=T_{k,i}}^{\infty}(x-T_{k,i})^2\,\mathrm{d}B^{(i)}(x)\bigg)/(1-B^{(i)}(T_{k,i})), \qquad i>k,\\
\mathbb{E}(R^{(k)}) &= \mathbb{E}(X^{(k)})+\sum_{i=1}^{k-1}\Bigg[\big(B^{(k)}(T_{i,k})-B^{(k)}(T_{i+1,k})\big)\sum_{j=i+1}^{k-1}\Lambda_j\mathbb{E}(A_{p_j}^{(k)})\cdot(T_{j,k}-T_{j+1,k})\\
&\qquad +\Lambda_i\mathbb{E}(A_{p_i}^{(k)})\bigg((T_{i,k}-T_{i+1,k})\cdot(1-B^{(k)}(T_{1,k}))+\int_{x=T_{i+1,k}}^{T_{i,k}}(x-T_{i+1,k})\,\mathrm{d}B^{(k)}(x)\bigg)\Bigg].
\end{aligned}$$

TETB rule

$$\begin{aligned}
\mathbb{E}(Z_k^{(i)}) &= \int_{x=0}^{\tau_{k,i}}x\,\mathrm{d}B^{(i)}(x)+\tau_{k,i}(1-B^{(i)}(\tau_{k,i})), \qquad i>k,\\
\mathbb{E}\big((Z_k^{(i)})^2\big) &= \int_{x=0}^{\tau_{k,i}}x^2\,\mathrm{d}B^{(i)}(x)+\tau_{k,i}^2(1-B^{(i)}(\tau_{k,i})), \qquad i>k,\\
\mathbb{E}(R^{(k)}) &= \mathbb{E}(X^{(k)})+\sum_{i=1}^{k-1}\Bigg[\Lambda_i\mathbb{E}(A_{p_i}^{(k)})\int_{x=\tau_{i,k}}^{\tau_{i+1,k}}(x-\tau_{i,k})\,\mathrm{d}B^{(k)}(x)\\
&\qquad +\big(B^{(k)}(\tau_{i+1,k})-B^{(k)}(\tau_{i,k})\big)\sum_{j=1}^{i-1}\Lambda_j\mathbb{E}(A_{p_j}^{(k)})(\tau_{j+1,k}-\tau_{j,k})\Bigg].
\end{aligned}$$

References

[1] ABATE, J., AND WHITT, W. Solving probability transform functional equations for numerical inversion. *Operations Research Letters 12* (1992), 275–281.

[2] Abate, J., and Whitt, W. Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing 7* (1995), 36–43.

[3] Adiri, I., and Domb, I. A single server queueing system working under mixed priority disciplines. *Operations Research 30* (1982), 97–115.

[4] Adiri, I., and Domb, I. Mixing of non-preemptive and preemptive repeat priority disciplines. *European Journal of Operational Research 18* (1984), 86–97.

[5] Brill, P. *System-point theory in exponential queues*. PhD thesis, University of Toronto, Toronto, Canada, 1975.

[6] Brill, P. H. *Level Crossing Methods in Stochastic Models*. Springer, New York, 2008.

[7] Cho, Y. Z., and Un, C. Analysis of the *M/G/1* queue under a combined preemptive/nonpreemptive priority discipline. *IEEE Transactions on Communications 41* (1993), 132–141.

[8] Cohen, J. On up-and downcrossings. *Journal of Applied Probability* (1977), 405–410.

[9] Conway, R. W., Maxwell, W. L., and Miller, L. W. *Theory of Scheduling*. Addison-Wesley, Reading, 1967.

[10] Drekic, S., and Stanford, D. A. Threshold-based interventions to optimize performance in preemptive priority queues. *Queueing Systems 35* (2000), 289–315.

[11] Drekic, S., and Stanford, D. A. Reducing delay in preemptive repeat priority queues. *Operations Research 49* (2001), 145–156.

[12] Fajardo, V. A., and Drekic, S. Controlling the workload of *M/G/1* queues via the q-policy. *European Journal of Operational Research 243* (2015), 607–617.

[13] FAJARDO, V. A., AND DREKIC, S. Waiting time distributions in the preemptive accumulating priority queue. *Methodology and Computing in Applied Probability* (in press).

[14] HSU, J. A continuation of delay-dependent queue disciplines. *Operations Research 18* (1970), 733–738.

[15] JAISWAL, N. K. *Priority Queues*. Academic Press, New York, 1968.

[16] KANET, J. A mixed delay dependent queue discipline. *Operations Research 30* (1982), 93–96.

[17] KLEINROCK, L. A delay dependent queue discipline. *Naval Research Logistics Quarterly 11* (1964), 329–341.

[18] NETTERMAN, A., AND ADIRI, I. A dynamic priority queue with general concave priority functions. *Operations Research 27* (1979), 1088–1100.

[19] PATEROK, M., AND ETTL, M. Sojourn time and waiting time distributions for *M/GI/1* queues with preemption-distance priorities. *Operations Research 42* (1994), 1146–1161.

[20] STANFORD, D. A., TAYLOR, P., AND ZIEDINS, I. Waiting time distributions in the accumulating priority queue. *Queueing Systems 77* (2014), 297–330.

[21] TAKAGI, H. *Queueing Analysis, Volume 1, Vacation and Priority Systems, Part 1*. North Holland, Amsterdam, 1991.

[22] TRIVEDI, S. K., JAIN, M., AND SHARMA, G. C. A delay dependent queue with preemption. *Indian Journal of Pure and Applied Mathematics 15* (1984), 1296–1301.