The final publication is available at Springer via http://dx.doi.org/10.1007/s11009-015-9476-1

Waiting Time Distributions in the Preemptive Accumulating Priority Queue

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

We consider a queueing system in which a single server attends to N priority classes of customers. Upon arrival to the system, a customer begins to accumulate priority linearly at a rate which is distinct to the class to which it belongs. Customers with greater accumulated priority levels are given preferential treatment in the sense that at every service selection instant, the customer with the greatest accumulated priority level is selected next for servicing. Furthermore, the system is preemptive so that the servicing of a customer is interrupted for customers with greater accumulated priority levels. The main objective of the paper is to characterize the waiting time distributions of each class. Numerical examples are also provided which exemplify the true benefit of incorporating an accumulating prioritization structure, namely the ability to control waiting times.

Keywords: Accumulating priority; preemptive priority; dynamic priority queues; maximal priority process; Laplace-Stieltjes transform.

1 Introduction

Within a priority queueing discipline, every customer is assigned a *priority level*, which determines its position inside the queue. Concerning the assignment of priorities to customers, there are generally two kinds of priority disciplines. The first has come to be known as the static (or fixed) priority discipline, wherein a customer's priority level determines the class to which it belongs and is constant with respect to its time spent in the system. These static priority disciplines have been extensively studied in the literature (e.g., see the reference texts by Conway et al. 1967, Jaiswal 1968, and Takagi 1991). The second kind of priority discipline is one in which a customer's priority depends on both its class specification and its time spent in the system. Although many real life systems would permit the usage of a static priority discipline, there are situations where the priority of a customer may change throughout its sojourn in the system (e.g., in a hospital emergency room, the condition of a patient may worsen while waiting to see a physician).

Consider a priority queueing system consisting of N distinct classes of customers, labelled $1, 2, \ldots, N$. Throughout the paper, we use the symbol $\mathcal{C}_i$ which should be read as "class-i customer". In general, a $\mathcal{C}_i$ is prioritized over a $\mathcal{C}_j$ whenever $i < j$. To describe the priority levels of customers, we make use of *priority functions*. Let the priority function for a $\mathcal{C}_k$ at time t be denoted by $q_k(t)$. A priority discipline such that $q_k(t)$ is constant with respect to t for all $k = 1, 2, \ldots, N$ is known as a static priority discipline, satisfying

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

where $\{a_i\}_{i=1}^N$ are real constants such that $a_1 > a_2 > \cdots > a_N$. In addition, within a given class, we assume that customers are served on a first-come-first-serve (FCFS) basis.

Priority disciplines in which $q_k(t)$ is dependent on t have been more or less termed in the literature as dynamic priority disciplines. If τ_k is the arrival time of a $\mathcal{C}_k$, then a dynamic priority

discipline can be characterized (as in Netterman and Adiri, 1979) as having priority functions given by

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

where $\{\phi_i(x)\}_{i=1}^N$ is a sequence of functions satisfying

$$\phi_1(0) \geq \phi_2(0) \geq \cdots \geq \phi_N(0), \tag{3}$$

and

$$\phi_1'(x) \geq \phi_2'(x) \geq \cdots \geq \phi_N'(x) \geq 0 \qquad \text{for all } x > 0. \tag{4}$$

For $i < j$, note that Eq. (3) infers that a $\mathcal{C}_i$ arrives to the system with an initial priority level which is at least as great as the initial priority level of a $\mathcal{C}_j$. Similarly, Eq. (4) implies that a $\mathcal{C}_i$ earns priority at least as fast as a $\mathcal{C}_j$ does. The *general priority service guideline* imposes that at each service completion instant, the customer with the greatest *accumulated* priority level be selected next for service. Hence, for dynamic priority queues employing this guideline, Eq. (3) and Eq. (4) imply that, within a given class, service is administered based on the order of arrival (as in the case of the static priority discipline). To our knowledge, all of the dynamic priority queues which have been previously analyzed in the literature employ the general priority service guideline. We provide a brief history of the literature next.

Jackson (1960, 1961, 1962) was the first to implement a dynamic priority discipline into a discrete-time queueing system. In these articles, he considered priority functions of the form

$$q_k(t) = a_k + (t - \tau_k), \quad t \geq \tau_k, \tag{5}$$

where the initial priority levels were arranged such that $a_1 > a_2 > \cdots > a_N$. He derived bounds for the mean waiting time of a $\mathcal{C}_k$, and in Jackson (1962), he obtained an approximation for the waiting time distribution.

The first to consider a dynamic priority discipline under a continuous-time framework was Kleinrock (1964), who developed a recursion for calculating average waiting times for a system with exponential interarrival and service times using priority functions of the form

$$q_k(t) = b_k \cdot (t - \tau_k), \quad t \geq \tau_k, \tag{6}$$

where the *accumulating priority rates* $\{b_i\}_{i=1}^N$ were arranged so that $b_1 \geq b_2 \cdots \geq b_N \geq 0$. Kleinrock termed this specific dynamic priority discipline as the *delay dependent priority* discipline. Kleinrock and Finkelstein (1967) then extended this work by considering the same $M/M/1$-type priority system with priority functions of the form

$$q_k(t) = b_k \cdot (t - \tau_k)^r, \quad t \geq \tau_k,$$

with $r \geq 0$. A few years later, Holtzman (1971) considered an $M/G/1$-type priority system characterized by Eq. (5) for which he derived both upper and lower bounds for the marginal expected waiting times of each class.

Netterman and Adiri (1979) subsequently analyzed an $M/G/1$-type priority system with a more general priority function in that the only requirement was that $\phi_k(x)$ be concave. In their paper, they obtained an integral recursive function for the expected class-k waiting time. There, the authors pointed out that, in general, the extraction of expected waiting times via their recursive function is quite difficult. Thus, they also obtained upper and lower bounds for the expected waiting times of each class. Others have also found expressions and corresponding bounds of steady-state expected waiting times for more general linearly increasing priority functions (e.g., see Bagchi and Sullivan 1985 and Sharma and Sharma 1994).

Systems where priority levels are decreasing rather than increasing have been studied in the papers by Hsu (1970) and Bagchi (1984). Following along the lines of Kleinrock (1964), these authors considered priority functions as in Eq. (6) with the exception that the rates $\{b_i\}_{i=1}^N$ were arranged such that $0 \geq b_1 \geq b_2 \cdots \geq b_N$ (i.e., the priority level of a $\mathcal{C}_i$ decreases at a slower

rate compared to that of a $\mathcal{C}_j$ whenever $i < j$). They derived recursions for the mean waiting times[1]. Kanet (1982) later considered an $M/G/1$-type priority system for which the classes of customers were divided into two sets: one set of classes whose customers accumulate priority, and the other whose customers' priority levels dissipate throughout time. Specifically, Kanet (1982) considered priority functions as in Eq. (6) with accumulating priority rates

$$b_1 \geq \cdots \geq b_i \geq 0 \geq b_{i+1} \geq \cdots \geq b_N$$

for some $i = 1, 2, \ldots, N$. He obtained a recursion for the steady-state expected waiting times for such a model.

From the mid-1980s to the end of the twentieth century, the literature on dynamic priority queues was nearly non-existent, with the only published work in this area being the paper by Sharma and Sharma (1994). Furthermore, it is clear that the analysis of such priority queues had been essentially focused on deriving expressions or bounds for the steady-state mean waiting times of each class. We believe that the overall complexity of these models is what deterred researchers from determining the distributions of steady-state waiting times.

In a recent paper, almost two decades removed from the last recorded work on the subject, Stanford et al. (2014) revisited the delay dependent priority discipline (i.e., Eq. (6)) and applied it to an $M/G/1$-type priority system. With a newly defined stochastic process, called the *maximal priority process*, the authors shed new light on the speficic structuralization of such a dynamic priority queue. Ultimately, by virtue of the maximal priority process, they derived the Laplace-Stieltjes transform (LST) of the steady-state class-k waiting time distribution. In their paper, they renamed the discipline as the *accumulating priority queue* on the basis that the term "delay dependent" (or "time dependent") had since gained several other meanings in the queueing literature.

[1] Bagchi (1984) points out two errors in Hsu's (1970) derivation of mean waiting times.

Unlike its counterpart (i.e., static priority queues), the existing literature on dynamic priority queueing systems is predominantly non-preemptive in nature. With the exception of Kleinrock (1964) and Kleinrock and Finkelstein (1967), where the authors find expressions for steady-state mean waiting times under the preemptive resume discipline[2], all of the aforementioned works have dealt with non-preemptive systems. It seems that for the preemptive variant, the only other notable publication is that of Trivedi et al. (1984), who considered the resume discipline in Kanet's (1982) mixed model. Once again, the analysis therein focused on finding the steady-state expected waiting times of each class.

In this paper, we adopt the same methodology of Stanford et al. (2014) to obtain the LSTs of the steady-state waiting time distributions associated with the dynamic preemptive priority queueing model defined by the priority functions of Eq. (6). The rest of the paper is organized as follows. In the next section, we present some notation and introduce the fundamental service-structure elements. Section 3 is devoted to the introduction of the maximal priority process for our dynamic preemptive priority queueing model. In Section 4, we first present the notion of a pseudo-interruption period and subsequently derive its LST. Residence periods and gross-service times are studied in Section 5. The marginal waiting time LSTs are established in Section 6. Two numerical examples are provided in Section 7, which exemplify the real benefit of implementing an accumulating prioritization scheme, namely the ability to control waiting times. Finally, in Section 8, we offer some concluding remarks.

2 Model description and preliminaries

A single-server dynamic priority queueing system with N distinct classes is considered. It is assumed that the arrivals of customers for the individual classes form independent Pois-

[2]However, non-preemptive systems were still the main focus of Kleinrock (1964) and Kleinrock and Finkelstein (1967).

son streams at rates $\lambda_1, \lambda_2, \ldots, \lambda_N$. The service times of customers are mutually independent, where the class-k service time is distributed identically to $X^{(k)}$ with distribution function (df) $B^{(k)}(x) = \mathbb{P}(X^{(k)} \leq x)$ and corresponding LST $\widetilde{B}^{(k)}(s) = \int_0^\infty e^{-sx} \mathrm{d}B^{(k)}(x)$. The assignment of priority to customers is done according to the priority functions of Eq. (6) with $b_1 \geq b_2 \geq \cdots \geq b_N \geq 0$. In other words, upon arriving to the system, a customer begins to accumulate priority linearly at a rate that is distinct to the class to which it belongs. It is important to note that customers accumulate priority throughout their entire stay in the system. At a service selection instant (i.e., a departure instant of a customer), the system employs the general priority service guideline.

In addition, our current system is preemptive in nature, meaning that the service of a customer is interrupted for any customer with a greater priority level. Since priority is assigned via Eq. (6), this implies that a preemption does not necessarily occur at the arrival instant of a higher priority customer, but rather at the instant in time that the higher priority customer accumulates a priority level which is equal to that of the customer currently in service. Note that the former situation describes the case of the classical static preemptive priority queue (i.e., interruptions always occur whenever a higher priority customer arrives). It is important to realize that a preemption instant is not considered to be a service selection instant. We review the three traditional preemption disciplines, which specify the nature of the servicing when an interrupted $\mathcal{C}_k$ re-enters service:

(i) Resume: service of the $\mathcal{C}_k$ continues from where it was interrupted.

(ii) Repeat-different: all previous work is lost and a new service time is independently sampled from $B^{(k)}(x)$.

(iii) Repeat-identical: all previous work is lost and service is restarted with the originally sampled service time.

We next define the class-k *waiting time*, $W^{(k)}$, as the total elapsed time from a $\mathcal{C}_k$'s arrival to the first time this customer goes into service. We also define the class-k *flow time*, $F^{(k)}$, as the total time spent in the system for a $\mathcal{C}_k$. The main objective of this paper is to establish the LST corresponding to the steady-state distribution of $W^{(k)}$, which we denote as $\widetilde{W}^{(k)}(s)$, for the three preemption disciplines above. We are also concerned with identifying the distributions of other key random variables, which we refer to as the service-structure elements. In fact, the LSTs of these random variables are required in order to obtain $\widetilde{W}^{(k)}(s)$. We define these service-structure elements with respect to a $\mathcal{C}_k$ as follows:

Residence period $R^{(k)} \equiv$	The time elapsed between first entry to service of a $\mathcal{C}_k$ and its departure.
Gross service time $G^{(k)} \equiv$	The total amount of time that the server spends servicing a $\mathcal{C}_k$ before its departure from the system.
Interruption period $A^{(k)} \equiv$	The time between a preemption instant and the instant in which the interrupted $\mathcal{C}_k$ next returns to service.

With these definitions in place, the stability condition of our priority queueing model is known to be

$$\overline{U} = \sum_{i=1}^{N} \rho_i = \sum_{i=1}^{N} \lambda_i \mathbb{E}(G^{(i)}) < 1, \tag{7}$$

where $\overline{U}$ is known as the *utilization factor*. The stability condition given by Eq. (7) is assumed throughout the paper. We also remark that some important relationships do exist amongst the service-structure elements. For example, we note that $R^{(k)}$ is comprised of $G^{(k)}$ and possibly several interruption periods $A^{(k)}$. As in the classical static preemptive priority queue, these interruption periods are independent and identically distributed (iid) regardless of the specific preemption discipline in place. Furthermore, due to independence, the LST of $F^{(k)}$ can be expressed as

$$\widetilde{F}^{(k)}(s) = \widetilde{W}^{(k)}(s)\widetilde{R}^{(k)}(s). \tag{8}$$

Fig. 1 illustrates the relationships between the service-structure elements.

We end this section with two more items pertaining to our adopted notation and one final remark on the name we give to our priority queueing model. First of all, unless otherwise specified, we denote the LST of a given random variable Y by $\widetilde{Y}(s) = \mathbb{E}(e^{-sY})$. Secondly, we point out that only those customers who belong to class i for any $i \in \{1, 2, \ldots, k-1\}$ can cause a preemption to a $\mathcal{C}_k$. Thus, for convenience, we adopt the convention of Conway et al. (1967) by referring to the aggregation of classes $\{1, 2, \ldots, k-1\}$ as class a, whose aggregated arrival rate we denote by $\Lambda_{k-1} = \sum_{i=1}^{k-1} \lambda_i$.

Remark 2.1 *Our current model represents the preemptive version of the model considered by Stanford et al. (2014). As mentioned in the introduction, Stanford et al. (2014) coined their model as the accumulating priority queue. In this paper, we refer to their model as the* non-preemptive accumulating priority queue *(NPAPQ). Similarly, we refer to our model as the* pre-emptive accumulating priority queue *(PAPQ).*

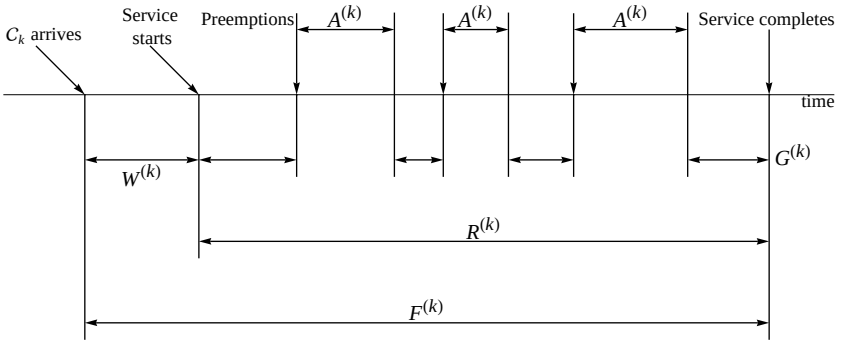


Figure 1: Depiction of the service-structure elements for a preemptive priority queue

3 The maximal priority process

In this section, we define an upper bound $M_k(t)$ for the accumulated priority level of any $\mathcal{C}_k$ potentially present in the system at time $t > 0$. We say potentially present since for $b_k >$

0, this upper bound has the virtue of always being positive during busy periods, even in the absence of $\mathcal{C}_k$s. The collection of these upper bounds (i.e., one for each class, so N in total) is what Stanford et al. (2014) referred to as the *maximal priority process*, which in general, is an N-dimensional stochastic process. Later in this section, we show that these upper bounds form the least upper bounds to the accumulated priority levels of customers when given only (certain) partial information to the system. Nevertheless, the real importance of this process is that it provides a useful structuralization for both the busy periods and the customers serviced within them. In terms of the PAPQ, the maximal priority process allows us to analyze the service-structure elements described in the previous section, and ultimately provides a means of obtaining the LST of the class-k waiting time distribution.

As the PAPQ allows for the preemption of customers, the maximal priority process defined here is slightly different than the one given by Stanford et al. (2014) for the NPAPQ. We define $Q_i(t)$ to be the priority level of the oldest $\mathcal{C}_i$ at time t. Note that our definition of $Q_i(t)$ is such that $Q_i(t) < 0$ means that there are no $\mathcal{C}_i$s present in the system at time t, and that the next $\mathcal{C}_i$ arrives to the system at time $t + Q_i(t)/b_i$. Moreover, let $\chi(t)$ and $Q_\vee(t)$ indicate the class and priority level, respectively, of the customer in service at time t. Clearly, for any t during a busy period, we have that

$$\chi(t) = \underset{1\le i\le N}{\arg\max}\{Q_i(t)\} \quad \text{and} \quad Q_\vee(t) = \max_{1\le i\le N}\{Q_i(t)\}.$$

For any t during an idle period, we further define $\chi(t) = Q_\vee(t) = 0$. Our definition of the maximal priority process for the PAPQ now follows.

Definition 3.1 *The maximal priority process is a N-dimensional stochastic process $\mathcal{M}(t) = \{(M_1(t), M_2(t), \ldots, M_N(t)), t \ge 0\}$, satisfying the following conditions:*

1. *The sample path of $M_k(t)$ for each $k = 1, 2, \ldots, N$ 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 class of customer,*

$$\frac{\mathrm{d}M_k(t)}{\mathrm{d}t} = \min\{b_k, b_{\chi(t)}\}.$$

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

$$M_k(\delta_i^+) = \min\{M_k(\delta_i^-), Q_\vee(\delta_i^+)\},$$

where $M_k(t^-) = \lim_{\epsilon\to 0} M_k(t-\epsilon)$*,* $M_k(t^+) = \lim_{\epsilon\to 0} M_k(t+\epsilon)$*, and* $Q_\vee(t^+) = \lim_{\epsilon\to 0} Q_\vee(t+\epsilon)$.

In what follows, let us also (artificially) define $b_{N+1} = 0$ and $M_{N+1}(t) = 0$ for all $t > 0$. Definition 3.1 simply states that during busy periods $M_k(t)$ increases linearly at the rate corresponding to the smallest of b_k and $b_{\chi(t)}$, and down-jumps at some of the service selection instants (i.e., customer departure instants). Fig. 2 illustrates a typical sample path of $\mathcal{M}(t)$ for a 3-class PAPQ, where the bold thick lines represent the components of $\mathcal{M}(t)$ and the thin lines represent the actual priority levels of the customers. Furthermore, the intersects between the thin lines and the t-axis represent the times customers enter the queue with priority level zero.

We next make the following observations about $\mathcal{M}(t)$:

(i) Observe that $M_1(t) = Q_\vee(t)$ for all $t > 0$, and, just as $Q_\vee(t)$ does, $M_1(t)$ down-jumps at every service selection instant.

(ii) Once a $\mathcal{C}_k$ commences service, its priority level is represented by $M_k(t)$ up until its departure from the system.

(iii) The periods between successive down-jumps of $M_N(t)$ partition the general busy period.

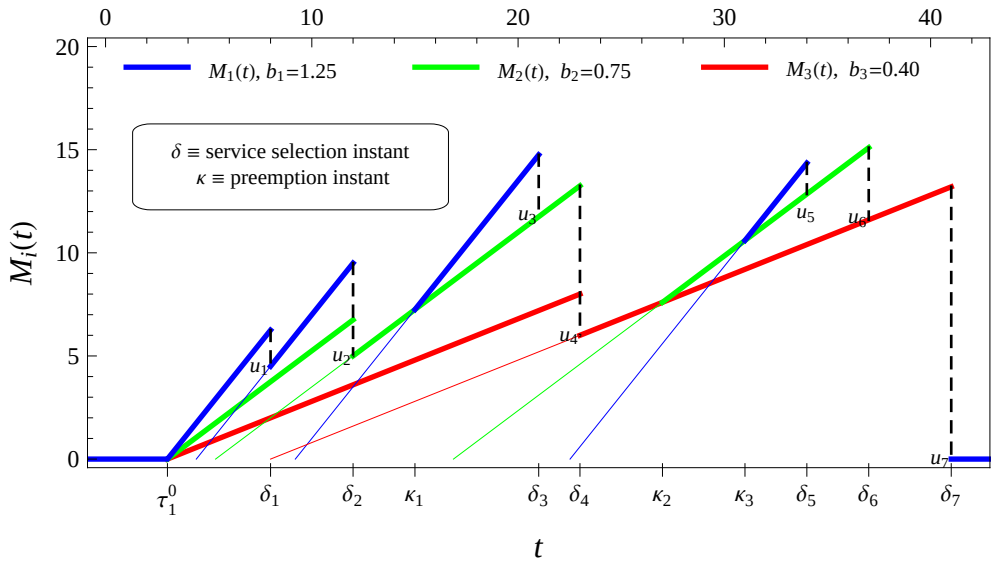


Figure 2: $\mathcal{M}(t)$ in a typical busy period of the PAPQ for $N = 3$

Observation (i) explains why $M_1(t)$ yields a least upper bound for class-1 priority levels at time t. In other words, all class-1 priority levels must be less than the priority level of the customer currently in service; a situation where a $\mathcal{C}_1$'s priority level is greater than $Q_\vee(t)$ for some time t is impossible as it would imply the occurence of a prior violation of the service discipline (i.e., either through a preemption that should have occurred before time t or an incorrect customer selection at a previous service selection instant). We proceed next to describe the type of least upper bounds that the other components provide for their respective classes' priority levels. First of all, we stress that one is able to (progressively) draw $\mathcal{M}(t)$ given only the following pieces of information:

(a) the sequence of busy period commencement times $\{\tau_i^0\}_{i=1}^\infty$;

(b) the sequence of service selection instants $\{\delta_i\}_{i=1}^\infty$, and for each of these, the priority level of the incoming service $u_i = Q_\vee(\delta_i^+)$;

(c) the sequence of preemption instants $\{\kappa_i\}_{i=1}^\infty$; and

(d) the class of the customer entering (or re-entering) service (i.e., $\chi(\tau_i^0)$, $\chi(\delta_i)$, and $\chi(\kappa_i)$ for all $i = 1, 2, \ldots$).

In particular, $\mathcal{M}(t)$ represents the collection of least upper bounds to the accumulated priorities of each class given only the partial information (a)–(d). Of course, to draw these sample paths, one must also keep in mind the fundamental characteristics of the system, namely: customers accumulate priority according to Eq. (6), customers arrive with an initial priority level of zero, and preemptions occur whenever a higher priority customer's priority level matches that of the customer currently in service. Note that the resulting $M_k(t)$ provides the least upper bound of class-k accumulated priority levels which would not lead to a violation of the service discipline similar to that described for $M_1(t)$ above. For example, one is able to reproduce the sample path in Fig. 2 given only the information found in Table 1. Finally, we emphasize that $M_k(t)$ generally does not represent the priority level of the oldest $\mathcal{C}_k$ at time t, $Q_k(t)$ – it only does so for t corresponding to a class-k residence period.

Table 1: Partial information (a)–(d) required to recreate $\mathcal{M}(t)$ of Fig. 2

	τ_0^1	δ_1	δ_2	κ_1	δ_3	δ_4	κ_2	κ_3	δ_5	δ_6	δ_7
t	3	8	12	15	21	23	27	31	34	37	41
$Q_\vee(t)$	0	4.5	5	–	11.75	6	–	–	12.85	11.6	0
$\chi(t)$	1	1	2	1	2	3	2	1	2	3	0

3.1 Structuralization of the general busy period and its customers

Following the convention of Stanford et al. (2014), we make the following definitions. First of all, we say that a waiting $\mathcal{C}_j$ (for $j \leq k$) is at *level-k accreditation* at time t if its priority level lies within the interval $[M_{k+1}(t), M_k(t))$. Since priority is earned linearly throughout time, it must be that the graph representing the priority level of customers at level-k accreditation at time t must have intersected $M_{k+1}(\cdot)$ at instants in time occurring before t. We refer to these

instants in time as *level-k accreditation instants*. Lastly, suppose at service selection instant δ that a $\mathcal{C}_j$ (for $j \leq k$) enters into service for the first time. Then, $Q_\vee(\delta^+)$ (i.e., the priority level of this $\mathcal{C}_j$ immediately prior to entering service for the first time) must lie within one of the following intervals:

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

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

$$Q_\vee(\delta^+) \in [M_{m+1}(\delta^-), M_m(\delta^-)) \quad \text{for} \quad m = j, j+1, \ldots, N.$$

In this paper, we use the symbol $\mathcal{C}^{(acc:m)}$ to denote a customer who is served at level-m accreditation. Note that a $\mathcal{C}^{(acc:m)}$ must belong to class i for some $i \in \{1, 2, \ldots, m\}$, and that when necessary, we use the symbol $\mathcal{C}_i^{(acc:m)}$ to refer to a $\mathcal{C}_i$ who is served at level-m accreditation. For example, the service selection instants δ_1, δ_2, and δ_4 of Fig. 2 represent the service commencements of a $\mathcal{C}^{(acc:1)}$, a $\mathcal{C}^{(acc:2)}$, and a $\mathcal{C}^{(acc:3)}$, respectively. The following result is crucial to our analysis of the PAPQ.

Lemma 3.1 *Suppose that at service selection instant δ, a $\mathcal{C}^{(acc:m)}$ enters into service with priority level $Q_\vee(\delta^+)$. Then, the magnitude of the down-jump of $M_m(t)$ occurring at time δ has an exponential distribution with rate $\sum_{i=1}^m \lambda_i/b_i$.*

Proof. From Definition 3.1, $M_m(t)$ will down-jump at δ to the level corresponding to the greatest priority level. In particular, the magnitude of the down-jump is given by

$$\min_{1 \leq i \leq m} \{M_m(\delta^-) - Q_i(\delta^-)\}.$$

The result follows since $M_m(\delta^-) - Q_i(\delta^-)$ has an exponential distribution with rate λ_i/b_i for all $i = 1, 2, \ldots, m$, which is independent of $M_m(\delta^-) - Q_j(\delta^-)$ for $j \neq i$. $\square$

Remark 3.2 *Since a $\mathcal{C}^{(acc:m)}$ can only belong to one class in the set $\{1, 2, \ldots, m\}$, this implies that one $\mathcal{C}^{(acc:m)}$ may accumulate priority linearly at a rate which is different to another $\mathcal{C}^{(acc:m)}$ (i.e., if they belong to different classes). However, the result in Lemma 3.1 holds true regardless of the specific class to which the $\mathcal{C}^{(acc:m)}$ belongs.*

The previous definition and Lemma 1 pertain to a $\mathcal{C}_j$ who is selected for service at a departure instant of another customer. However, it is also possible for a $\mathcal{C}_j$ to enter into service by preempting a $\mathcal{C}_i$ (for $i > j$) out of service. Specifically, suppose that a $\mathcal{C}_j$ enters into service at time κ, corresponding to a preemption instant of a $\mathcal{C}_{k+1}$. Then, from Definition 3.1, we have that the priority level of the interrupting $\mathcal{C}_j$ upon entry into service is such that

$$Q_{\vee}(\alpha^{+}) = M_{k+1}(\kappa) = M_k(\kappa) = \cdots = M_j(\kappa) = \cdots = M_1(\kappa).$$

We call such a $\mathcal{C}_j$ who preempts a $\mathcal{C}_\ell$ (for $\ell > j$) out of service as a *class-ℓ interrupting customer*, denoted by $\mathcal{C}^{(int:\ell)}$. Therefore, a $\mathcal{C}_j$ who arrives during a busy period must either be a $\mathcal{C}^{(acc:\ell)}$ for some $\ell \geq j$ or a $\mathcal{C}^{(int:\ell)}$ for some $\ell > j$. The next result specifies the rate at which a preemption occurs.

Lemma 3.3 *The rate of preemption for the servicing of a $\mathcal{C}_k$ is $\Lambda_{k-1}^{(k)} = \sum_{i=1}^{k-1} \lambda_i^{(k)}$, where $\lambda_i^{(k)} = \lambda_i(1 - b_k/b_i)$.*

Proof. Suppose that at time t, a $\mathcal{C}_k$ enters into service with a priority level of $u \geq 0$. Hence, there can be no $\mathcal{C}_i$ (for $i \in a$, where a is the aggregation of classes $\{1, 2, \ldots, k-1\}$, as defined earlier) with a priority level equal to u at time t. Next, define T_i to be the time, starting from t, until the first $\mathcal{C}_i$ accumulates a priority level of u. It follows from the memoryless property that T_i has an exponential distribution with rate λ_i. Furthermore, let Y_i represent the time, starting from t, until the priority level of the $\mathcal{C}_i$ first matches that of the $\mathcal{C}_k$ in service. It is then quite straightforward to show that $Y_i = T_i(1 - b_k/b_i)^{-1}$. □

In addition to providing the above classifications of customers, the maximal priority process also produces special subperiods of the overall busy period, which we refer to as *level-k accreditation intervals*. In general, a level-k accreditation interval starts in one of three ways:

(i) at the moment when a $\mathcal{C}_k$ or a $\mathcal{C}_a$ arrives to an empty system, thereby initiating a busy period;

(ii) when a $\mathcal{C}_k^{(acc:\ell)}$ or a $\mathcal{C}_a^{(acc:\ell)}$ for $\ell > k$ enters into service for the first time; or

(iii) at the moment when a $\mathcal{C}_k^{(int:\ell)}$ or a $\mathcal{C}_a^{(int:\ell)}$ preempts a $\mathcal{C}_\ell$ (for $\ell > k$) out of service.

Regardless of how it starts, a level-k accreditation interval always ends once the system becomes clear of the initial customer and all $\mathcal{C}^{(acc:i)}$s for $i = 1, 2, \ldots, k$ (i.e., all customers who have become level k or more accredited). Let u_0 denote the priority level of the initial customer of a level-k accreditation interval. Then, u_0 is strictly positive for level-k accreditation intervals starting according to (ii) and (iii), and $u_0 = 0$ otherwise. We note that the distribution of the length of an accreditation interval depends only on the class to which the intial customer belongs and not on the specific value of u_0 (see Stanford et al., 2014, Lemma 4.3). A recursive scheme for the LST corresponding to the distribution of the duration of a level-k accreditation interval is provided in the next section, but before that, we end this section with one final important result.

It follows from Definition 3.1 that a level-k accreditation interval has the virtue that throughout the entire interval, $M_{k+1}(t)$ and $M_k(t)$ increase with rates b_{k+1} and b_k, respectively. Moreover, a level-k accreditation interval is partitioned by subperiods which are defined by the successive down-jumps of $M_k(t)$. Except for the final one, these down-jumps correspond to the service selection instants of a $\mathcal{C}^{(acc:k)}$; the final down-jump represents either the end of a busy period, the commencement of service of a $\mathcal{C}^{(acc:\ell)}$, or the re-entry into service of an interrupted $\mathcal{C}_\ell$ for some $\ell > k$. For a level-k accreditation interval with an initial priority level of u_0, we say

that a $\mathcal{C}_i$ for $i \leq k$ *arrives-to-the-interval* if its priority level becomes equal to u_0 before the end of the interval. Fig. 3 illustrates a level-k accreditation interval with four class-i arrivals-to-the-interval.

Lemma 3.4 *The steady-state proportion of $\mathcal{C}_i$s (for $i \leq k$) that arrive-to-the-interval and are served at level-k accreditation is* $(b_k - b_{k+1})/b_i$.

Proof. Consider a level-k accreditation interval with an initial priority level of u_0. Suppose that the accreditation interval has an overall duration of T and that it has n subperiods defined by the successive down-jumps of $M_k(t)$. Let $\{T_j, j = 1, 2, \ldots, n\}$ denote the duration of these subperiods (e.g., see Fig. 3). Now, observe first that the proportion of T for which a $\mathcal{C}_i$ arrives-to-the-interval and fails to become level-k accredited is given by b_{k+1}/b_i. For example, the fourth $\mathcal{C}_i$ to arrive-to-the-interval in Fig. 3 arrives within this proportion, and thus is not serviced in this interval. Secondly, we observe that there are disjoint time periods of length $T_j(1 - b_k/b_i)$ for each $j = 1, 2, \ldots, n$, such that a $\mathcal{C}_i$ arrival-to-the-interval during any one of these time periods would lead to a level-$(k-1)$ accreditation for the arriving customer. As a result, the proportion of T for which a $\mathcal{C}_i$ arrives-to-the-interval and fails to become level-$(k-1)$ accredited is given by b_k/b_i. Therefore, the proportion of T for which a $\mathcal{C}_i$ arrives-to-the-interval and fails to become level-$(k-1)$ accredited but yet succeeds in becoming level-k accredited is $(b_k - b_{k+1})/b_i$. Note that a $\mathcal{C}_i$ such as the one previously described is precisely one that is serviced at level-k accreditation (e.g., see the second $\mathcal{C}_i$ who arrives-to-the-interval in Fig. 3). The result follows because the above proportions and the fact that the class-i arrivals-to-the-interval form a Poisson process with rate λ_i hold true for every level-k accreditation interval. □

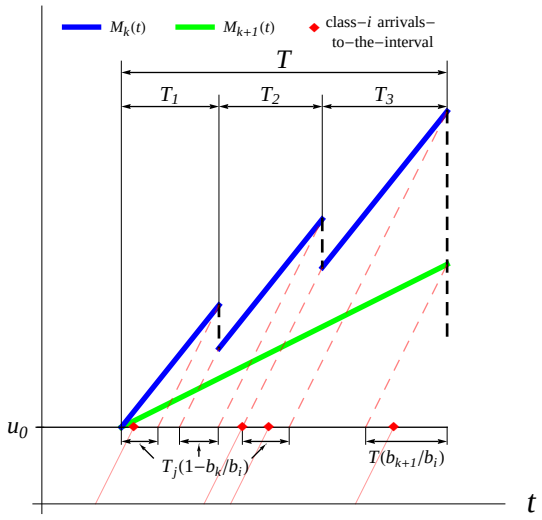


Figure 3: Supplemental illustration of a level-k accreditation interval for the proof of Lemma 3. Note that T_2 is initiated by a $\mathcal{C}^{(acc:k)}$ not belonging to class i.

4 Interruption periods and pseudo-interruption periods

We begin with the class-$(k+1)$ interruption period $A^{(k+1)}$. It is clear that only a $\mathcal{C}_a^{(int:k+1)}$ or a $\mathcal{C}_k^{(int:k+1)}$ can initiate a class-$(k+1)$ interruption period, and further that such a period ends as soon as the system is clear of all higher priority customers whose priority level exceeds that of the interrupted $\mathcal{C}_{k+1}$. From the previous section, such customers are referred to as $\mathcal{C}^{(acc:i)}$s for some $i \leq k$. Furthermore, from the previous section, we acknowledge that $A^{(k+1)}$ is merely a level-k accreditation interval of type (iii).

To establish a recursive scheme for $\widetilde{A}^{(k+1)}(s)$, recall that a level-k accreditation interval is partitioned by subperiods which are defined by the successive down-jumps of $M_k(t)$. It turns out that these time periods are either themselves level-$(k-1)$ accreditation intervals or class-k residence periods. For example, if the initial customer is a $\mathcal{C}_k$ (which from Lemma 3.3 occurs with probability $\lambda_k^{(k+1)}/\Lambda_k^{(k+1)}$), then the initial subperiod is merely a class-k residence period $R^{(k)}$. On the other hand, if the initial customer is a $\mathcal{C}_a$ (which from Lemma 3.3 occurs with probability $\Lambda_{k-1}^{(k+1)}/\Lambda_k^{(k+1)}$), then the initial subperiod is indeed a level-$(k-1)$ accreditation

interval of type (iii). This level-$(k-1)$ accreditation interval has all of the same characteristics as a class-k interruption period $A^{(k)}$ (i.e., it is initiated by a $\mathcal{C}_a$ and terminates once the system is clear of all $\mathcal{C}^{(acc:i)}$s for $i < k$), with the exception that a $\mathcal{C}_k$ has not actually been preempted (i.e., in this case, a $\mathcal{C}_{k+1}$ is being preempted). As a result, we define our first kind of *pseudo-interruption period*:

$$A^{(m)}_{p_{k+1}} \text{ (for } m \leq k+1) \equiv \begin{array}{l}\text{A class-}m\text{ pseudo-interruption period initiating} \\ \text{with the preemption of a class-}(k+1)\text{ customer.}\end{array}$$

We stress that $A^{(m)}_{p_{k+1}}$ is a level-$(m-1)$ accreditation interval of type (iii). Thus, if the initial customer is a $\mathcal{C}_a$, then the initial subperiod is $A^{(k)}_{p_{k+1}}$.

For the subsequent subperiods of $A^{(k+1)}$, we understand from the previous section that they can only be initiated by either a $\mathcal{C}_a^{(acc:k)}$ or a $\mathcal{C}_k^{(acc:k)}$. Similar to the initial subperiod, if a $\mathcal{C}_k^{(acc:k)}$ enters into service (which from Lemma 3.1 occurs with probability $(\lambda_k/b_k)/\sum_{i=1}^{k}\lambda_i/b_i$), then the ensuing subperiod is $R^{(k)}$. On the contrary, if the initial customer is a $\mathcal{C}_a^{(acc:k)}$, then the subperiod is a level-$(k-1)$ accreditation interval. Again, it turns out that this level-$(k-1)$ accreditation interval bears all the same characteristics as $A^{(k)}$ with the exception that no customer is actually being preempted. This leads us to our second kind of pseudo-interruption period:

$$A^{(m)}_{np} \text{ (for } m = 1, 2, \ldots, N) \equiv \begin{array}{l}\text{A class-}m\text{ pseudo-interruption period not initiating} \\ \text{at a preemption instant, but instead at the} \\ \text{commencement of service of a } \mathcal{C}_i^{(acc:\ell)} \text{ for } i < m \\ \text{and any } \ell \geq m.\end{array}$$

We stress that $A^{(m)}_{np}$ is a level-$(m-1)$ accreditation interval of type (ii). Thus, if a $\mathcal{C}_a^{(acc:k)}$ enters into service, then a subperiod $A^{(k)}_{np}$ ensues.

Our previous observations suggest that $A^{(k+1)}$ may be viewed as a delay busy period which services two kinds of customers (i.e., $\mathcal{C}_k^{(acc:k)}$s and $\mathcal{C}_a^{(acc:k)}$s), whose respective initial delay and service time LSTs are given by

$$\widetilde{V}^{(k)}_{p_{k+1}}(s) = \sum_{i=1}^{k-1} \frac{\lambda_i^{(k+1)}}{\Lambda_k^{(k+1)}} \widetilde{A}^{(k)}_{p_{k+1}}(s) + \frac{\lambda_k^{(k+1)}}{\Lambda_k^{(k+1)}} \widetilde{R}^{(k)}(s), \tag{9}$$

and

$$\Phi_k(s) = \frac{\sum_{i=1}^{k-1} \lambda_i/b_i}{\sum_{i=1}^{k} \lambda_i/b_i} \widetilde{A}_{np}^{(k)}(s) + \frac{\lambda_k/b_k}{\sum_{i=1}^{k} \lambda_i/b_i} \widetilde{R}^{(k)}(s). \tag{10}$$

In order to show this, we make an important connection between $(M_k(t), M_{k+1}(t))$ during level-k accreditation intervals and the maximal priority process of the $M/G/1$ queue with accumulating priority and blocking introduced by Fajardo and Drekic (2015). This model represents a FCFS $M/G/1$ queue, whose customers, upon arrival to the system, accumulate priority linearly at rate $\xi_1 > 0$. The blocking of customers occurs near the end of a busy period of the queue. In particular, at the beginning of each busy period, an *accreditation threshold* increases linearly at rate ξ_2, where $\xi_1 > \xi_2 \geq 0$, so that only those customers whose priority levels surpass this accreditation threshold are serviced; customers who fail to surpass this threshold depart the system without ever being serviced. The maximal priority process for this model is a two-dimensional stochastic process $(M(t), \Theta(t))$, where $M(t)$ provides the least upper bound of accumulated priorities similar to $\mathcal{M}(t)$ defined in Definition 3.1 and $\Theta(t)$ gives the value of the accreditation threshold at time t. Two important observations follow.

Important Observation 1 A level-k accreditation interval is partitioned by subperiods defined by the successive down-jumps of $M_k(t)$. The down-jumps of $M_k(t)$ during a level-k accreditation interval are exponentially distributed with rate $\sum_{i=1}^{k} \lambda_i/b_i$. The time from the start of the interval to the first time that $M_k(t)$ down-jumps, which we denote by V, depends on the initial customer of interval. Furthermore, the distribution of V may differ from that of the times between one down-jump of $M_k(t)$ to the next, which always has LST $\Phi_k(s)$. Lastly, if δ represents the end of a subperiod, then δ also represents the end of the level-k accreditation interval if

$$\min_{1 \leq i \leq k} \{M_k(\delta^-) - Q_i(\delta^-)\} > M_k(\delta^-) - M_{k+1}(\delta^-).$$

Important Observation 2 It follows from Important Observation 1 that the evolution of $(M_k(t), M_{k+1}(t))$ throughout a level-k accreditation interval is equivalent to that of the maximal priority process $(M(t), \Theta(t))$ during busy periods of the FCFS $M/G/1$ queue with accumulating priority and blocking having the following characteristics:

(i) initial delay LST of $\widetilde{V}(s)$;

(ii) service time LST of $\Phi_k(s)$;

(iii) arrival rate of $\gamma_k = \sum_{i=1}^{k} \lambda_i (b_k/b_i)$;

(iv) accumulating priority rate of $\xi_1 = b_k$; and

(v) accreditation threshold rate of $\xi_2 = b_{k+1}$.

We exploit the connection outlined in Important Observation 2 to obtain two fundamental results: the distribution of the duration of a level-k accreditation interval and the distribution of the accumulated priority earned by a $\mathcal{C}^{(acc:k)}$ during a level-k accreditation interval. In particular, it follows from Important Observation 2 that the distribution of the duration of a level-k accreditation interval has corresponding LST (see Fajardo and Drekic, 2015, Theorem 3.1)

$$\widetilde{\mathcal{A}}_k(s) \equiv \widetilde{\mathcal{A}}_k(s; V) = \widetilde{V}\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big), \tag{11}$$

where

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

and $\eta_k(s)$ satisfies

$$\eta_k(s) = \Phi_k\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big). \tag{12}$$

Our previous arguments show that for this specific level-k accreditation interval, the distribution of V has LST $\widetilde{V}_{p_{k+1}}^{(k)}(s)$ as given by Eq. (9). Moreover, from Eq. (11), we observe

that

$$\widetilde{A}^{(k+1)}(s) = \widetilde{A}^{(k+1)}_{p_{k+1}}(s) = \widetilde{\mathcal{A}}_k(s; V^{(k)}_{p_{k+1}}). \tag{13}$$

Eq. (13) also leads to the following recursive scheme which starts with $\widetilde{A}^{(1)}_{p_{k+1}}(s) = 1$ and holds for all $m = 1, 2, \ldots, k$:

$$\widetilde{A}^{(m+1)}_{p_{k+1}}(s) = \frac{\Lambda^{(k+1)}_{m-1}}{\Lambda^{(k+1)}_{m}} \widetilde{A}^{(m)}_{p_{k+1}}\big(s + \gamma^{(m+1)}_m(1 - \eta_m(s))\big) + \frac{\lambda^{(k+1)}_m}{\Lambda^{(k+1)}_{m}} \widetilde{R}^{(m)}\big(s + \gamma^{(m+1)}_m(1 - \eta_m(s))\big). \tag{14}$$

The above recursion requires that both $R^{(k)}$ and $A^{(m)}_{np}$ for $m = 1, 2, \ldots, k$ be priorly established. The former is the subject of the next section. Consider $A^{(k+1)}_{np}$, which represents a level-k accreditation interval which begins with the service of a $\mathcal{C}^{(acc:\ell)}_k$ or $\mathcal{C}^{(acc:\ell)}_a$ for some $\ell > k$. It follows from Lemma 3.1 that the initial subperiod is $A^{(k)}_{np}$ with probability $(\sum_{i=1}^{k-1} \lambda_i/b_i)/(\sum_{i=1}^{k} \lambda_i/b_i)$, and is $R^{(k)}$ with probability $(\lambda_k/b_k)/(\sum_{i=1}^{k} \lambda_i/b_i)$. In other words, for the level-k accreditation interval $A^{(k+1)}_{np}$, V has LST $\widetilde{V}^{(k)}_{np}(s) = \Phi_k(s)$. Therefore, we have that

$$\widetilde{A}^{(k+1)}_{np}(s) = \widetilde{\mathcal{A}}_k(s; V^{(k)}_{np}) = \eta_k(s), \tag{15}$$

which again yields a recursive scheme starting with $\widetilde{A}^{(1)}_{np}(s) = 1$.

The recursive schemes of Eqs. (13) and (15) establish the LSTs of level-k accreditation intervals of types (iii) and (ii), respectively. Hence, all that remains is to establish a recursion for a level-k accreditation interval of type (i). This leads us to our final pseudo-interruption period:

$A^{(m)}_{p_0}$ (for $m = 1, 2, \ldots, N$) $\equiv$ A class-m pseudo-interruption period not initiating at a preemption instant, but instead at the arrival of a $\mathcal{C}_i$ for $i < m$ to an empty system.

We consider $A^{(k+1)}_{p_0}$ and remark that the initial subperiod is either $R^{(k)}$ with probability λ_k/Λ_k or $A^{(k)}_{p_0}$ with probability Λ_{k-1}/Λ_k. Hence, for this level-k accreditation interval, the initial

subperiod V has LST

$$\widetilde{V}_{p_0}^{(k)}(s) = \frac{\lambda_k}{\Lambda_k}\widetilde{R}^{(k)}(s) + \frac{\Lambda_{k-1}}{\Lambda_k}\widetilde{A}_{p_0}^{(k)}(s).$$

Thus,

$$\widetilde{A}_{p_0}^{(k+1)}(s) = \widetilde{\mathcal{A}}_k(s; V_{p_0}^{(k)}), \tag{16}$$

and starting with $\widetilde{A}_{p_0}^{(1)}(s) = 1$, a recursive representation for $\widetilde{A}_{p_0}^{(k+1)}(s)$ is given by

$$\widetilde{A}_{p_0}^{(k+1)}(s) = \frac{\Lambda_{k-1}}{\Lambda_k}\widetilde{A}_{p_0}^{(k)}\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big) + \frac{\lambda_k}{\Lambda_k}\widetilde{R}^{(k)}\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big). \tag{17}$$

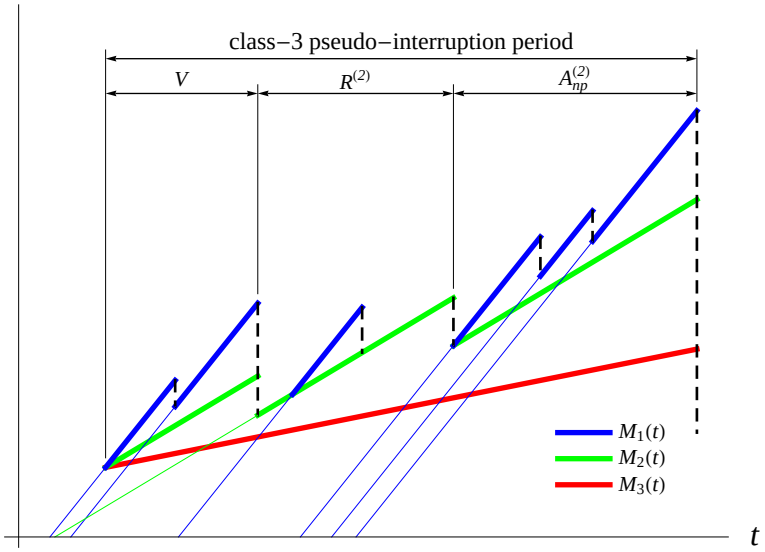


Figure 4: General structure of a class-3 pseudo-interruption period

For illustrative purposes, Fig. 4 depicts the general structure of a class-3 pseudo-interruption period as described above. Also, expressions for both the first and second moments of each of the pseudo-interruption periods can be found in the Appendix.

We next present three useful identities pertaining to the first moments of each of the pseudo-interruption periods. The proofs of these identities are omitted, but are readily verified by

induction. Let $\overline{U}_j = \sum_{i=1}^{j} \rho_i$ and $\overline{U}_j^{(k+1)} = \sum_{i=1}^{j} \lambda_i^{(k+1)} \mathbb{E}(G^{(i)})$. For $k = 1, 2, \ldots, N-1$, we therefore have

$$\gamma_k \mathbb{E}(A_{np}^{(k+1)}) = \frac{\sum_{i=1}^{k} \lambda_i (b_k/b_i) \mathbb{E}(G^{(i)})}{1 - \overline{U}_k^{(k+1)}} = \frac{\overline{U}_k - \overline{U}_{k-1}^{(k)}}{1 - \overline{U}_k^{(k+1)}}, \tag{18}$$

and

$$\Lambda_k \mathbb{E}(A_{p_0}^{(k+1)}) = \frac{\overline{U}_k}{1 - \overline{U}_k^{(k+1)}}. \tag{19}$$

Also, for each above value of k, we have

$$\Lambda_m^{(k+1)} \mathbb{E}(A_{p_{k+1}}^{(m+1)}) = \frac{\overline{U}_m^{(k+1)}}{1 - \overline{U}_m^{(m+1)}}\ , \quad m = 1, 2, \ldots, k. \tag{20}$$

We end this section with a remark on the existence of pseudo-interruption periods in the classical static preemptive priority queue.

Remark 4.1 *The pseudo-interruption periods, $A_{p_0}^{(k)}$ and $A_{p_j}^{(k)}$ for all $j > k$, are also inherent in the classical static preemptive priority queue. However, since priority is assigned via Eq.* (1) *in this model, these pseudo-interruption periods are equivalently distributed to an actual interruption period $A^{(k)}$.*

5 Residence periods and gross service times

In this section, we derive the LSTs of $R^{(k)}$ and $G^{(k)}$. We begin with a general observation concerning the composition of a class-k residence period in the PAPQ. Specifically, it is possible that a $\mathcal{C}_k$ may suffer from several iid interruption periods (each having LST $\widetilde{A}^{(k)}(s)$) between the moment of its first entry into service up until its eventual departure from the system. It is important to note that this general observation also holds true for the class-k residence period in the classical static preemptive priority queue. In fact, the only difference in the general compositions of the class-k residence period in the PAPQ and that in the classical static preemptive priority queue is the preemption rate during a class-k service. Thus, in order to obtain the LSTs

of $R^{(k)}$ and $G^{(k)}$ for the PAPQ, we simply apply the same analysis used in Conway et al. (1967) except here we use the preemption rate supplied by Lemma 3.3.

As a result, the LSTs of $R^{(k)}$ and $G^{(k)}$ for each of the three preemption disciplines are as follows:

Resume:

$$\widetilde{R}^{(k)}(s) = \widetilde{B}^{(k)}\big(s + \Lambda_{k-1}^{(k)}(1 - \widetilde{A}^{(k)}(s))\big) \tag{21}$$

$$\widetilde{G}^{(k)}(s) = \widetilde{B}^{(k)}(s). \tag{22}$$

Repeat-different:

$$\widetilde{R}^{(k)}(s) = \frac{\left(s + \Lambda_{k-1}^{(k)}\right)\widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)})}{s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(s)\left(1 - \widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)})\right)} \tag{23}$$

$$\widetilde{G}^{(k)}(s) = \frac{\left(s + \Lambda_{k-1}^{(k)}\right)\widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)})}{s + \Lambda_{k-1}^{(k)}\widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)})}. \tag{24}$$

Repeat-identical:

$$\begin{aligned}\widetilde{R}^{(k)}(s) &= \mathbb{E}[\mathbb{E}(e^{-sR^{(k)}}|X^{(k)})] \\ &= \int_{x=0}^{\infty} \frac{(s + \Lambda_{k-1}^{(k)})e^{-(s+\Lambda_{k-1}^{(k)})x}}{s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(s)(1 - e^{-(s+\Lambda_{k-1}^{(k)})x})}\,\mathrm{d}B^{(k)}(x)\end{aligned} \tag{25}$$

$$\begin{aligned}\widetilde{G}^{(k)}(s) &= \mathbb{E}[\mathbb{E}(e^{-sG^{(k)}}|X^{(k)})] \\ &= \int_{x=0}^{\infty} \frac{(s + \Lambda_{k-1}^{(k)})e^{-(s+\Lambda_{k-1}^{(k)})x}}{s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}(1 - e^{-(s+\Lambda_{k-1}^{(k)})x})}\,\mathrm{d}B^{(k)}(x).\end{aligned} \tag{26}$$

The first two moments of $R^{(k)}$ and $G^{(k)}$ for each preemption discipline can be found in the Appendix. We next present a similar result to Lemma 3.4. Suppose that a class-$(k+1)$

residence period begins with an initial priority level of u_0. Then, as similarly done for level-k accreditation intervals, we define the *arrivals-to-the-residence-period* to be the time epochs (during a class-$(k+1)$ residence period) for which a $\mathcal{C}_i$ for $i \in \{1, 2, \ldots, k\}$ accumulates a priority level equal to the initial level u_0.

Lemma 5.1 *In the long run, the proportion of $\mathcal{C}_i$s for $i \in \{1, 2, \ldots, k\}$ who arrive-to-the-residence-period and become level-k accredited is $1 - b_{k+1}/b_i$.*

Proof. We omit the details, but state that one can use similar arguments as those in the proof of Lemma 3.4 to prove this particular result. □

6 Waiting time distributions

In this section, we derive the marginal waiting time LSTs. It is clear that $\mathcal{C}_k$s who arrive to the system during an idle period enter into service immediately, and thus do not incur any amount of wait. Let $W_{BP}^{(k)}$ be the waiting time incurred by a $\mathcal{C}_k$ who arrives to the system during a busy period. Therefore, we have

$$\widetilde{W}^{(k)}(s) = \pi_0 + (1 - \pi_0)\widetilde{W}_{BP}^{(k)}(s), \tag{27}$$

where $\pi_0 = 1 - \overline{U}$ is the steady-state probability of the system being empty. We next define $P_{BP}^{(k)}$ to be the accumulated priority (immediately prior to entering service for the first time) of a $\mathcal{C}_k$ who arrives to the system during a busy period. Given that priority is assigned via Eq. (6), it follows that

$$\widetilde{W}_{BP}^{(k)}(s) = \widetilde{P}_{BP}^{(k)}(s/b_k). \tag{28}$$

Hence, to find $\widetilde{W}^{(k)}(s)$, we first find $\widetilde{P}_{BP}^{(k)}(s)$ and subsequently apply Eqs. (27) and (28).

Recall that a $\mathcal{C}_k$ who arrives to the system during a busy period can only either be a $\mathcal{C}^{(acc:\ell)}$ for some $\ell \geq k$ or a $\mathcal{C}^{(int:\ell)}$ for some $\ell > k$. Let us denote a $\mathcal{C}_k$ of the former kind by $\mathcal{C}_k^{(acc)}$,

and that of the latter kind by $\mathcal{C}_k^{(int)}$. Furthermore, let $\widetilde{P}_{acc}^{(k)}(s)$ and $\widetilde{P}_{int}^{(k)}(s)$ denote the LSTs of the accumulated priority of a $\mathcal{C}_k^{(acc)}$ and $\mathcal{C}_k^{(int)}$, respectively. Then,

$$\widetilde{P}_{BP}^{(k)}(s) = \frac{1}{1-\pi_0}\left[\pi_k^{(acc)}\widetilde{P}_{acc}^{(k)}(s) + \alpha_k^{(int)}\widetilde{P}_{int}^{(k)}(s)\right], \tag{29}$$

where $\pi_k^{(acc)}$ and $\alpha_k^{(int)}$ represent the steady-state probabilities that a $\mathcal{C}_k$ arrives during a busy period and is a $\mathcal{C}_k^{(acc)}$ or $\mathcal{C}_k^{(int)}$, respectively.

6.1 The distribution of accumulated priority of a $\mathcal{C}_k^{(acc)}$

We present here a recursion for $\widetilde{P}_{acc}^{(k)}(s)$. First of all, let $P_{acc:k}^{(k)}$ denote the accumulated priority of a $\mathcal{C}_k^{(acc:k)}$. Let $P_{unacc:k}^{(k)}$ denote the accumulated priority of a $\mathcal{C}_k^{(acc:\ell)}$ for some $\ell > k$. Then,

$$\widetilde{P}_{acc}^{(k)}(s) = \frac{1}{\pi_k^{(acc)}}\left[\pi_k^{(k)}\widetilde{P}_{acc:k}^{(k)}(s) + \sum_{\ell=k+1}^{N}\pi_k^{(\ell)}\widetilde{P}_{unacc:k}^{(k)}(s)\right], \tag{30}$$

where $\pi_k^{(j)}$ is the steady-state probability that a $\mathcal{C}_k$ arrives to a busy period and is serviced at level-j accreditation. Now, it follows from Lemma 3.1 and Remark 3.2 that the distribution of accumulated priority of a $\mathcal{C}^{(acc:\ell)}$ is the same regardless of the specific class to which the customer belongs. This previous argument, coupled with the fact that $\pi_k^{(\ell)} = (b_{k+1}/b_k)\pi_{k+1}^{(\ell)}$ for $\ell > k$ (as shown in Section 6.3), ultimately leads to the following recursive scheme for the desired LST:

$$\widetilde{P}_{acc}^{(k)}(s) = \frac{1}{\pi_k^{(acc)}}\left[\pi_k^{(k)}\widetilde{P}_{acc:k}^{(k)}(s) + \frac{b_{k+1}}{b_k}\pi_{k+1}^{(acc)}\widetilde{P}_{acc}^{(k+1)}(s)\right]. \tag{31}$$

In order to find $\widetilde{P}_{acc:k}^{(k)}(s)$, we first note that a $\mathcal{C}^{(acc:k)}$ (for any $1 \leq k \leq N$) is always served in a level-k accreditation interval. Now, suppose that a level-k accreditation interval starts with an initial priority level of u_0. Then, the accumulated priorities of all $\mathcal{C}^{(acc:k)}$s serviced in this interval must have an accumulated priority which is greater than u_0. In other words, the accumulated priority of a $\mathcal{C}^{(acc:k)}$ is decomposed into two parts: u_0 and the additional accumulated priority after having accumulated priority level u_0, which we denote by $\mathcal{P}^{(acc:k)}$. It is important to note

that the distribution of $\mathcal{P}^{(acc:k)}$ is independent of the specific value of u_0 (i.e., this independence is similar to the one between $W^{(k)}$ and $R^{(k)}$).

We next make our second use of the connection between the PAPQ and the $M/G/1$ queue with accumulating priority and blocking, as outlined in Important Observation 2. In particular, it follows from Important Observation 2 that the distribution of $\mathcal{P}^{(acc:k)}$, associated with an initial delay V (i.e., the initial delay of the level-k accreditation interval), is given by (see Fajardo and Drekic, 2015, Eq. (58))

$$\widetilde{\mathcal{P}}^{(acc:k)}(s) \equiv \widetilde{\mathcal{P}}^{(acc:k)}(s;V) = \frac{\left(1-\gamma_k^{(k+1)}\mu_{k,1}\right)\left(\widetilde{\mathcal{A}}_k(b_{k+1}s)-\widetilde{V}(b_k s)\right)}{\mathbb{E}(V)\left(1-b_{k+1}/b_k\right)\left(b_k s-\gamma_k+\gamma_k\Phi_k(b_k s)\right)}, \tag{32}$$

where $\widetilde{\mathcal{A}}_k(s)$ is given by Eq. (11) and $\mu_{k,i}$ is the i-th moment of the random variable whose distribution has LST $\Phi_k(s)$. Note that the result in Eq. (32) was first derived by Stanford et al. (2014) for the NPAPQ. Upon differentiation and after some algebra, we obtain the first moment of $\mathcal{P}^{(acc:k)}$ as

$$\mathbb{E}(\mathcal{P}^{(acc:k)}) = b_k\left(\frac{\mathbb{E}(V^2)}{2\mathbb{E}(V)}\cdot\left[1+\frac{b_{k+1}/b_k}{1-\gamma_k^{(k+1)}\mu_{k,1}}\right] + \frac{\gamma_k\mu_{k,2}}{2(1-\gamma_k\mu_{k,1})}\cdot\left[1-\left(\frac{b_{k+1}/b_k}{1-\gamma_k^{(k+1)}\mu_{k,1}}\right)^2\right]\right). \tag{33}$$

We must consider all of the level-k accreditation intervals in which a $\mathcal{C}^{(acc:k)}$ can be serviced. From the previous sections, we know that there are only three types of level-k accreditation intervals, all of which correspond to a specific kind of pseudo-interruption period. In particular, a $\mathcal{C}^{(acc:k)}$ must be serviced within $A_{p_0}^{(k+1)}$, $A_{np}^{(k+1)}$, or $A_{p_j}^{(k+1)}$ for some $j > k$. Now, it follows from independence that the LST of the accumulated priorities of $\mathcal{C}^{(acc:k)}$s serviced in each of these pseudo-interruption periods is simply a product of the LST of the initial priority level and the LST of the additional accumulated priority $\mathcal{P}^{(acc:k)}$.

The initial priority level for a level-k accreditation interval of type (i) is clearly zero. Therefore, the accumulated priority of a $\mathcal{C}^{(acc:k)}$ serviced in $A_{p_0}^{(k+1)}$ simply has LST $\widetilde{\mathcal{P}}^{(acc:k)}(s;V_{p_0}^{(k)})$.

A pseudo-interruption period $A_{np}^{(k+1)}$ is initiated whenever a $\mathcal{C}_a^{(acc:\ell)}$ or a $\mathcal{C}_k^{(acc:\ell)}$ for $\ell > k$ enters into service. Hence, the accumulated priority of a $\mathcal{C}^{(acc:k)}$ serviced in $A_{np}^{(k+1)}$ and initiated by a $\mathcal{C}_a^{(acc:\ell)}$ or a $\mathcal{C}_k^{(acc:\ell)}$ has LST $\widetilde{P}_{acc:\ell}^{(\ell)}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{np}^{(k)})$ for all $\ell > k$. Lastly, recall that the pseudo-interruption period $A_{p_\ell}^{(k+1)}$ for $\ell > k$ initiates whenever a $\mathcal{C}_a$ or a $\mathcal{C}_k$ preempts a $\mathcal{C}_\ell$ out of service. Letting $P_{int:\ell}$ be the accumulated priority of a customer who preempts a $\mathcal{C}_\ell$ out of service, the accumulated priority of a $\mathcal{C}^{(acc:k)}$ serviced in $A_{p_\ell}^{(k+1)}$ and initiated by a $\mathcal{C}_a^{(int:\ell)}$ or a $\mathcal{C}_k^{(int:\ell)}$ has LST $\widetilde{P}_{int:\ell}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{p_\ell}^{(k)})$ for all $\ell > k$.

Next, we define the following steady-state probabilities:

$$\begin{aligned}
\pi_k^{(k:i)} &\equiv \text{probability that a } \mathcal{C}_k \text{ is serviced at level-}k \text{ accreditation in an } A_{p_0}^{(k+1)};\\
\pi_k^{(k:ii:\ell)} &\equiv \text{probability that a } \mathcal{C}_k \text{ is serviced at level-}k \text{ accreditation in an } A_{np}^{(k+1)} \text{ which is initiated by a } \mathcal{C}_a^{(acc:\ell)} \text{ or a } \mathcal{C}_k^{(acc:\ell)} \text{ for } \ell > k;\\
\pi_k^{(k:iii:\ell)} &\equiv \text{probability that a } \mathcal{C}_k \text{ is serviced at level-}k \text{ accreditation in an } A_{p_\ell}^{(k+1)} \text{ which is initiated by a } \mathcal{C}_a^{(int:\ell)} \text{ or a } \mathcal{C}_k^{(int:\ell)} \text{ for } \ell > k.
\end{aligned}$$

Therefore, we have that

$$\widetilde{P}_{acc:k}^{(k)}(s) = \frac{1}{\pi_k^{(k)}}\Bigg[\pi_k^{(k:i)}\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{p_0}^{(k)}) + \sum_{\ell=k+1}^{N} \pi_k^{(k:ii:\ell)}\widetilde{P}_{acc:\ell}^{(\ell)}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{np}^{(k)}) + \sum_{\ell=k+1}^{N} \pi_k^{(k:iii:\ell)}\widetilde{P}_{int:\ell}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{p_\ell}^{(k)})\Bigg]. \quad (34)$$

6.2 The distribution of accumulated priority of a $\mathcal{C}_k^{(int)}$

Let $P_{int:\ell}^{(k)}$ be the accumulated priority of a $\mathcal{C}_k^{(int:\ell)}$ for $\ell > k$. Similar to the decomposition in the previous subsection, we have $P_{int:\ell}^{(k)} = u_0 + \mathcal{P}^{(int:\ell)}$ where u_0 is the initial priority level of the class-ℓ residence period $R^{(\ell)}$ and $\mathcal{P}^{(int:\ell)}$ is the additional accumulated priority earned by the interrupting customer after having accumulated priority level u_0. It is important to note that the

distribution of $\mathcal{P}^{(int:\ell)}$ is independent of the value u_0, which is equal to zero if the interrupted $\mathcal{C}_\ell$ arrived to an empty system and is greater than zero otherwise (i.e., assuming that $b_\ell > 0$). Clearly, u_0 represents the accumulated priority of the $\mathcal{C}_\ell$ immediately prior to the first time it enters service, so that

$$\widetilde{P}^{(k)}_{int:\ell}(s) = \frac{\alpha_k^{(0:\ell)}\widetilde{\mathcal{P}}^{(int:\ell)}(s) + \alpha_k^{(1:\ell)}\widetilde{P}^{(\ell)}_{BP}(s)\widetilde{\mathcal{P}}^{(int:\ell)}(s)}{\alpha_k^{(\ell)}}, \tag{35}$$

where:

$$\begin{aligned}
\alpha_k^{(\ell)} \equiv &\quad \text{probability that a } \mathcal{C}_k \text{ interrupts a } \mathcal{C}_\ell \text{ (for } \ell > k\text{) out of service;}\\
\alpha_k^{(0:\ell)} \equiv &\quad \text{probability that a } \mathcal{C}_k \text{ interrupts a } \mathcal{C}_\ell \text{ (for } \ell > k\text{), who arrived to an empty system, out of service;}\\
\alpha_k^{(1:\ell)} \equiv &\quad \text{probability that a } \mathcal{C}_k \text{ interrupts a } \mathcal{C}_\ell \text{ (for } \ell > k\text{), who arrived to the system during a busy period, out of service.}
\end{aligned}$$

We show in the next subsection that $\alpha_i^{(0:\ell)}/\alpha_i^{(\ell)} = \pi_0$ and $\alpha_i^{(1:\ell)}/\alpha_i^{(\ell)} = 1 - \pi_0$ for all $i \in \{1, \ldots, k, \ldots, \ell - 1\}$. This implies that the distribution of the accumulated priority of an interrupting customer is independent of the actual class to which the interrupting customer belongs. Therefore, we can re-write Eq. (35) as

$$\widetilde{P}^{(k)}_{int:\ell}(s) = \widetilde{P}_{int:\ell}(s) = \pi_0\widetilde{\mathcal{P}}^{(int:\ell)}(s) + (1 - \pi_0)\widetilde{P}^{(\ell)}_{BP}(s)\widetilde{\mathcal{P}}^{(int:\ell)}(s). \tag{36}$$

Note that in the second equality above, we drop the superscript in the notation to indicate that this distribution does not depend on the class of the interrupting customer. Furthermore, Eq. (36) is used in Eq. (34). It is also clear that a $\mathcal{C}_k$ can interrupt any $\mathcal{C}_i$ for $i \in \{k+1, k+2, \ldots, N\}$. Therefore,

$$\widetilde{P}^{(k)}_{int}(s) = \frac{1}{\alpha_k^{(int)}} \sum_{\ell=k+1}^{N} \alpha_k^{(\ell)}\widetilde{P}_{int:\ell}(s). \tag{37}$$

To conclude this subsection, we establish $\widetilde{\mathcal{P}}^{(int:k)}$ for each of three preemption disciplines.

Resume: Under this strategy, we can find $\widetilde{\mathcal{P}}^{(int:k)}(s)$ by conditioning on the partially completed service time, $X_{past}^{(k)}$, and the number of preemptions $\mathcal{N}$ encountered during that time. In particular,

$$\mathbb{E}(e^{-s\mathcal{P}^{(int:k)}}|X_{past}^{(k)}=x,\mathcal{N}=n)=e^{-sb_kx}\left[\widetilde{A}^{(k)}(b_ks)\right]^n.$$

By Lemma 3.3, given that $X_{past}^{(k)}=x$, $\mathcal{N}$ is Poisson distributed with rate $\Lambda_{k-1}^{(k)}x$. Removing the conditional statements, we readily obtain

$$\widetilde{\mathcal{P}}^{(int:k)}(s)=\frac{1-\widetilde{B}^{(k)}(sb_k+\Lambda_{k-1}^{(k)}-\Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_ks))}{\mathbb{E}(X^{(k)})(sb_k+\Lambda_{k-1}^{(k)}-\Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_ks))}, \tag{38}$$

with corresponding first moment

$$\mathbb{E}(\mathcal{P}^{(int:k)})=b_k\left(\frac{\mathbb{E}[(X^{(k)})^2]}{2\mathbb{E}(X^{(k)})}(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\right). \tag{39}$$

Repeat-different: Under this strategy, we can view each time a $\mathcal{C}_k$ enters into service as a Bernoulli experiment, where a successful outcome is defined as service going to completion, which happens with probability $\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})$. Following the convention of Conway et al. (1967, pp. 171-172), we denote the wasted service time random variable as $X_w^{(k)}$ (i.e., an interrupted service attempt) whose LST is given by

$$\widetilde{X}_w^{(k)}(s)=\frac{\Lambda_{k-1}^{(k)}\big(1-\widetilde{B}^{(k)}(s+\Lambda_{k-1}^{(k)})\big)}{\big(s+\Lambda_{k-1}^{(k)}\big)\big(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})\big)}.$$

Considering only the times when a class-k residence period is in progress, define the system to be in state m at a particular instant if the number of previous interruptions (not including the current interruption period, if applicable) suffered by the oldest $\mathcal{C}_k$ is m. Suppose now that a $\mathcal{C}_a$ preempts a $\mathcal{C}_k$ when the system is in state m. This implies that, at the time our marked $\mathcal{C}_a$ begins service, the ongoing residence period is already comprised of m independent pairs of $X_w^{(k)}+A^{(k)}$, followed by another independent $X_w^{(k)}$. Note that these $2m+1$ random variables are all independent, and so

$$\mathbb{E}(e^{-s\mathcal{P}^{(int:k)}}|\text{state } m)=\left[\widetilde{X}_w^{(k)}(b_ks)\right]^{m+1}\left[\widetilde{A}^{(k)}(b_ks)\right]^m.$$

If we define P_m to be the steady-state probability that the system is in state m (i.e., $P_m = \mathbb{P}(\text{state } m \,|R^{(k)} \text{ in progress})$), then the probability of a $\mathcal{C}_a$ becoming accredited during a class-k residence period while the system is in state m is also P_m by virtue of the PASTA property (e.g., see Wolff, 1982). Therefore,

$$\mathbb{E}(e^{-s\mathcal{P}^{(int:k)}}) = \sum_{m=0}^{\infty} P_m \left[\widetilde{X}_w^{(k)}(b_k s)\right]^{m+1} \left[\widetilde{A}^{(k)}(b_k s)\right]^m.$$

Using results from semi-Markov theory and discrete-time Markov chains, it can be shown that $P_m = \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})[1 - \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})]^m$, thereby leading to

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = \frac{1 - \widetilde{B}^{(k)}(b_k s + \Lambda_{k-1}^{(k)})}{\mathbb{E}(G^{(k)})\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_k s)(1 - \widetilde{B}^{(k)}(b_k s + \Lambda_{k-1}^{(k)}))\big)}, \tag{40}$$

with corresponding first moment

$$\mathbb{E}(\mathcal{P}^{(int:k)}) = b_k \left(\frac{\mathbb{E}[(G^{(k)})^2]}{2\mathbb{E}(G^{(k)})} + \Lambda_{k-1}^{(k)} \mathbb{E}(A^{(k)}) \mathbb{E}(G^{(k)}) \right). \tag{41}$$

Repeat-identical: The derivation of $\widetilde{\mathcal{P}}^{(int:k)}(s)$ under the repeat-identical strategy is similar to the repeat-different case; however, it is now necessary to condition on the originally drawn service time of the $\mathcal{C}_k$. Specifically, the desired LST under this discipline works out to be

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = \int_{x=0}^{\infty} \frac{\big(1 - e^{-(sb_k + \Lambda_{k-1}^{(k)})x}\big)}{\mathbb{E}(G^{(k)})\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_k s)(1 - e^{-(sb_k + \Lambda_{k-1}^{(k)})x})\big)} \, \mathrm{d}B^{(k)}(x). \tag{42}$$

In addition, we can express the corresponding first moment as

$$\mathbb{E}(\mathcal{P}^{(int:k)}) = b_k \int_{x=0}^{\infty} \left\{ \frac{\mathbb{E}[(G^{(k)})^2 | X^{(k)} = x]}{2\mathbb{E}(G^{(k)})} + \frac{\Lambda_{k-1}^{(k)} \mathbb{E}(A^{(k)}) (\mathbb{E}[G^{(k)} | X^{(k)} = x])^2}{\mathbb{E}(G^{(k)})} \right\} \mathrm{d}B^{(k)}(x). \tag{43}$$

The first and second conditional moments of $G^{(k)}$ found in the integrand of Eq. (43) are given in the Appendix.

6.3 Steady-state probabilities

We next derive formulas for the steady-state probabilities introduced in the previous subsections. Clearly, $\pi_k^{(acc)} = \sum_{\ell=k}^{N} \pi_k^{(\ell)}$ and $\alpha_k^{(int)} = \sum_{\ell=k+1}^{N} \alpha_k^{(\ell)}$. The following proposition provides the forms of the steady-state probabilities $\pi_k^{(\ell)}$ and $\alpha_k^{(\ell)}$.

Proposition 6.1 *The probability that a $\mathcal{C}_k$ arrives to a busy period and is serviced at level-ℓ accreditation is*

$$\pi_k^{(\ell)} = \overline{U}_\ell (b_\ell - b_{\ell+1})/b_k \quad \textit{for } \ell \geq k. \tag{44}$$

Furthermore, the probability that a $\mathcal{C}_k$ arrives to a busy period and preempts a $\mathcal{C}_\ell$ out of service is

$$\alpha_k^{(\ell)} = \rho_\ell (1 - b_\ell/b_k) \quad \textit{for } \ell > k. \tag{45}$$

Proof. We consider first the case for $\ell = N$. Note that a busy period is a level-N accreditation interval. Thus, from our previous arguments, we observe that a busy period is partitioned by subperiods which can only either be level-$(N-1)$ accreditation intervals (i.e., class-N pseudo-interruption periods) or class-N residence periods. Following the logic used in the proofs of Lemmas 3.4 and 5.1, the proportion of a busy period which would lead to an eventual level-$(N-1)$ accreditation of a $\mathcal{C}_k$ is always $1 - b_N/b_k$. Therefore, by virtue of the PASTA property, we have that $\pi_k^{(N)} = \overline{U} b_N/b_k$. Now, some of those $\mathcal{C}_k$s who earn level-$(N-1)$ accreditation will enter into service by preempting a $\mathcal{C}_N$ out of service. In other words, these are the $\mathcal{C}_k$s who become level-$(N-1)$ accredited during the servicing of a $\mathcal{C}_N$. The long-run proportion of the busy period dedicated to the servicing of a $\mathcal{C}_N$ is $\rho_N/\overline{U}_N$. It therefore follows that $\alpha_k^{(N)} = \rho_N(1 - b_N/b_k)$.

The remaining proportion of $\mathcal{C}_k$s who become level-$(N-1)$ accredited will do so during the servicing of a $\mathcal{C}_i$ for $i < N$. This implies that these $\mathcal{C}_k$s are serviced in a class-N pseudo-interruption period (or equivalently, in a level-$(N-1)$ accreditation interval). Recall that a

level-$(N-1)$ accreditation interval is again decomposed into subperiods which can only either be a level-$(N-2)$ accreditation interval or a class-$(N-1)$ residence period. Once again, the same logic applied above establishes that the proportion of level-$(N-1)$ accredited $\mathcal{C}_k$s who also become level-$(N-2)$ accredited is $(1-b_{N-1}/b_k)/(1-b_N/b_k)$. Therefore, we have that $\pi_k^{(N-1)} = \overline{U}_{N-1}(b_{N-1}-b_N)/b_k$. Furthermore, since $\rho_{N-1}/\overline{U}_{N-1}$ represents the conditional probability that a $\mathcal{C}_{N-1}$ is in service given that some customer belonging to one of classes $\{1,2,\ldots,N-1\}$ is in service, it follows that $\alpha_k^{(N-1)} = \rho_{N-1}(1-b_{N-1}/b_k)$.

By continuing along in this fashion, we eventually establish the remaining probabilities. □

To find $\pi_k^{(k:i)}$, $\pi_k^{(k:ii:\ell)}$, and $\pi_k^{(k:iii:\ell)}$ for $\ell > k$, we first need to find the long-run proportion of time that all of these level-k accreditation intervals are in progress. It follows from Lemma 3 that the desired probabilities are found by multiplying the previous proportions by $(b_k - b_{k+1})/b_k$. In particular, the long-run proportion of time that an $A_{p_0}^{(k+1)}$ is in progress is given by

$$\pi_0 \Lambda_k \mathbb{E}(A_{p_0}^{(k+1)}) = \pi_0 \frac{\overline{U}_k}{1-\overline{U}_k^{(k+1)}},$$

where the equality holds by Eq. (19). Therefore, we have that

$$\pi_k^{(k:i)} = \pi_0 \frac{\overline{U}_k}{1-\overline{U}_k^{(k+1)}} \left(\frac{b_k - b_{k+1}}{b_k}\right). \tag{46}$$

We similarly obtain the following results for $\ell > k$:

$$\pi_k^{(k:ii:\ell)} = \left[\frac{\overline{U}_\ell \sum_{i=1}^k \rho_i\big((b_\ell - b_{\ell+1})/b_i\big)}{1-\overline{U}_k^{(k+1)}}\right] \left(\frac{b_k - b_{k+1}}{b_k}\right), \tag{47}$$

and

$$\pi_k^{(k:iii:\ell)} = \left[\frac{\rho_\ell \overline{U}_k^{(\ell)}}{1-\overline{U}_k^{(k+1)}}\right] \left(\frac{b_k - b_{k+1}}{b_k}\right). \tag{48}$$

One can easily verify that $\pi_k^{(k)} = \pi_k^{(k:i)} + \sum_{\ell=k+1}^N \pi_k^{(k:ii:\ell)} + \sum_{\ell=k+1}^N \pi_k^{(k:iii:\ell)}$. In addition, we readily obtain from Lemma 4 that

$$\alpha_k^{(0:\ell)} = \pi_0 \rho_\ell (1 - b_\ell/b_k), \tag{49}$$

and

$$\alpha_k^{(1:\ell)} = (1-\pi_0)\rho_\ell(1-b_\ell/b_k). \tag{50}$$

6.4 Connections between the PAPQ and other queueing models

We begin with a remark concerning the LST of the waiting time distribution of the lowest priority class, $\widetilde{W}^{(N)}(s)$. Note that since $b_{N+1}=0$, it follows that $\pi_N^{(acc)} = \overline{U}$. Furthermore, it is clear that $\mathcal{C}_N$s can never preempt another customer out of service, and thus it is readily observed from Eqs. (29) and (34) that

$$\widetilde{P}_{BP}^{(N)}(s) = \widetilde{\mathcal{P}}^{(acc:N)}(s; V_{p_0}^{(N)}) = \frac{\big(1-\gamma_N^{(N+1)}\mu_{N,1}\big)\big(1-\widetilde{V}_{p_0}^{(N)}(b_N s)\big)}{\mathbb{E}(V_{p_0}^{(N)})\big(b_N s - \gamma_N + \gamma_N \Phi_N(b_N s)\big)}. \tag{51}$$

The waiting time LST of the lowest priority class is readily obtained via Eqs. (27) and (28). Moreover, Eq. (51) serves as the starting point for the recursive scheme to establish the remaining LSTs $\widetilde{P}_{PB}^{(N-1)}(s), \widetilde{P}_{PB}^{(N-2)}(s), \ldots, \widetilde{P}_{PB}^{(1)}(s)$ given in Eqs. (29), (31), (34), (35), and (37).

Under a preemptive resume service discipline, Eq. (51) yields after some algebra the following expression for the class-N waiting time LST:

$$\widetilde{W}^{(N)}(s) = \frac{\big(s+\Lambda_{N-1}^{(N)}(1-\psi_{N-1}(s))\big)(1-\overline{U})}{s - \sum_{i=1}^{N}\lambda_i(b_N/b_i)\big(1-\widetilde{B}^{(i)}\big(s+\Lambda_{N-1}^{(N)}(1-\psi_{N-1}(s))\big)\big)}, \tag{52}$$

where

$$\psi_{N-1}(s) = \sum_{i=1}^{N-1}\frac{\lambda_i^{(N)}}{\Lambda_{N-1}^{(N)}}\widetilde{B}^{(i)}\big(s+\Lambda_{N-1}^{(N)}(1-\psi_{N-1}(s))\big). \tag{53}$$

We remark that Eq. (52) is identical to the waiting time LST of the lowest priority class in the NPAPQ (see Stanford et al., 2014, Eq. (65)). This relationship is well understood due to the fact that the non-preemptive and preemptive resume service disciplines are both work-conserving disciplines. We note that the same relationship holds in the case of the static non-preemptive and preemptive resume priority queueing models (e.g., see Takagi, 1991).

We end Section 6 with two limiting cases of the PAPQ involving the ratio b_{k+1}/b_k which must lie in the interval [0,1]. On the one hand, suppose that $b_{k+1}/b_k \approx 1$ for all $k = 1, 2, \ldots, N-$

1. Under this setting, it is quite difficult for customers of higher priority to preempt customers of lower priority. Hence, as the ratio b_{k+1}/b_k approaches one, the PAPQ approaches the FCFS M/G/1 queue whose arrival rate is Λ_N and service time LST is given by $\widetilde{B}(s) = (1/\Lambda_N)\sum_{i=1}^{N}\lambda_i\widetilde{B}^{(i)}(s)$.

On the other hand, suppose that $b_{k+1}/b_k \approx 0$ for all $k = 1, 2, \ldots, N-1$. In contrast to the previous situation, it is now easier for higher priority customers to preempt lower priority ones out of service (i.e., preemptions essentially occur at higher priority customer arrival instants). Therefore, as b_{k+1}/b_k gets closer to zero, the PAPQ approaches the static preemptive priority model. These limiting cases illustrate a potential benefit in that the PAPQ can be useful to system managers of FCFS queueing systems who wish to implement a static prioritization scheme, but feel that the resulting congestion would still be too great. In such situations, the PAPQ is a viable alternative as it could provide the desired balance between the two extremes of FCFS and static preemptive priority.

7 Numerical examples

In this section, we present two numerical examples which illustrate the versatility of the PAPQ. It is well understood that the main advantage of the PAPQ (and other dynamic priority queues of the like) is the ability to control waiting times through the selection of the accumulating priority rates $\{b_k\}_{k=1}^{N}$. For our first example, we consider a 3-class PAPQ with class arrival rates $\lambda_1 = 0.25$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.14$. Furthermore, we assume that $X^{(1)} \sim$ Gam(0.25,0.25), $X^{(2)} \sim$ Gam(2,1.6), and $X^{(3)} \sim$ Gam(3,2), where "Gam(α,β)" denotes the gamma distribution with LST $\widetilde{B}(s) = (1 + s/\beta)^{-\alpha}$. This example was first considered by Drekic (2003, p. 69) in which a static priority queue under a hybrid-based preemption discipline (called the preemptive resume with expiry time discipline) was analyzed. The accumulating

priority rates are arranged as follows:

$$b_1 = 1, \quad b_2 = e^{-x}, \quad \text{and} \quad b_3 = e^{-2x} \quad \text{for some } x \geq 0. \tag{54}$$

We conduct a mean value analysis for this particular PAPQ by tabulating, over a range of values for x, the expected values of $W^{(k)}$ and $F^{(k)}$, $k = 1, 2, 3$, under all three preemption disciplines. The results are reported to 4 decimal places of accuracy in Tables 2 and 3. Moreover, if we define $N^{(k)}$ as the steady-state number of $\mathcal{C}_k$s waiting in the queue, then it immediately follows via the distributional form of Little's Law (e.g., see Keilson and Servi, 1990) that the z-transform of $N^{(k)}$ is given by

$$\widehat{N}^{(k)}(z) = \mathbb{E}(z^{N^{(k)}}) = \widetilde{W}^{(k)}\big(\lambda_k(1-z)\big). \tag{55}$$

Table 4 reports to 4 decimal places of accuracy the expected values of $N^{(k)}$, $k = 1, 2, 3$, over the same range of values for x.

Note that as $x \to \infty$, Eq. (54) implies that $b_{k+1}/b_k \to 0$ for $k = 1, 2$, and the PAPQ becomes equivalent to the static preemptive priority model. Hence, when $x = 100$ (corresponding to the first row of Tables 2–4), we expect the results to be fairly close to the static preemptive priority model (see Drekic, 2003, Tables 1 and 2). This is indeed the case. Conversely, we observe that $b_{k+1}/b_k \to 1$ as $x \to 0$ for $k = 1, 2$. As we move down the rows in Tables 2–4, the results are approaching those of the limiting FCFS $M/G/1$ queue (as described in Section 6.4), and these results are consistent under all three preemption disciplines.

Our second example takes inspiration from the 2-class static priority queue analyzed in Conway et al. (1967, p. 177) for which both class-1 and class-2 service times are assumed to be exponentially distributed with mean one. Conway et al. (1967) analyzed the overall mean flow time (i.e., $(\lambda_1\mathbb{E}(F^{(1)}) + \lambda_2\mathbb{E}(F^{(2)}))/\Lambda_2$) of this system across several different values of λ_1 and λ_2. Their results illustrated the generally accepted assertion which states that the repeat-identical discipline suffers most from congestion than the other two preemption disciplines.

Table 2: Expected waiting times for three preemption disciplines in Example 1

x	Resume			Repeat-Different			Repeat-Identical		
	$\mathbb{E}(W^{(1)})$	$\mathbb{E}(W^{(2)})$	$\mathbb{E}(W^{(3)})$	$\mathbb{E}(W^{(1)})$	$\mathbb{E}(W^{(2)})$	$\mathbb{E}(W^{(3)})$	$\mathbb{E}(W^{(1)})$	$\mathbb{E}(W^{(2)})$	$\mathbb{E}(W^{(3)})$
100.0000	0.8333	2.2917	7.3750	0.8333	2.5798	12.8610	0.8333	4.1539	101.6713
10.0000	0.8334	2.2918	7.3748	0.8334	2.5802	12.8604	0.8335	4.1579	101.6498
7.5000	0.8340	2.2934	7.3730	0.8341	2.5841	12.8542	0.8350	4.2033	101.4103
5.0000	0.8414	2.3130	7.3501	0.8436	2.6311	12.7792	0.8578	4.7368	98.5466
2.5000	0.9531	2.5401	7.0614	1.0031	3.1496	11.8632	1.4872	9.0468	70.4359
1.0000	1.6460	3.1987	5.8924	2.0340	4.2534	8.5789	4.1032	9.8782	23.1396
0.7500	1.9670	3.3600	5.4721	2.4439	4.3695	7.5254	4.4425	8.6005	16.1676
0.5000	2.4029	3.5121	4.9590	2.9137	4.3541	6.3174	4.5066	6.9804	10.5447
0.2500	2.9742	3.6310	4.3570	3.3717	4.1415	5.0067	4.2389	5.2549	6.4186
0.1000	3.3856	3.6743	3.9613	3.5887	3.8988	4.2086	3.9343	4.2807	4.6284
0.0100	3.6564	3.6868	3.7151	3.6797	3.7103	3.7389	3.7134	3.7444	3.7733
0.0010	3.6844	3.6874	3.6903	3.6867	3.6898	3.6926	3.6901	3.6932	3.6960
0.0001	3.6872	3.6875	3.6878	3.6874	3.6877	3.6880	3.6878	3.6881	3.6884
0.0000	3.6875	3.6875	3.6875	3.6875	3.6875	3.6875	3.6875	3.6875	3.6875

Table 3: Expected flow times for three preemption disciplines in Example 1

x	Resume			Repeat-Different			Repeat-Identical		
	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$
100.0000	1.8333	3.9583	10.3750	1.8333	4.3767	16.7380	1.8333	6.3121	107.6576
10.0000	1.8334	3.9585	10.3748	1.8334	4.3770	16.7374	1.8335	6.3161	107.6358
7.5000	1.8340	3.9598	10.3721	1.8341	4.3805	16.7298	1.8350	6.3607	107.3924
5.0000	1.8414	3.9759	10.3399	1.8436	4.4231	16.6381	1.8578	6.8860	104.4824
2.5000	1.9531	4.1624	9.9338	2.0031	4.8882	15.5157	2.4872	11.0992	75.8297
1.0000	2.6460	4.6833	8.2893	3.0340	5.8114	11.4456	5.1032	11.6157	26.8298
0.7500	2.9670	4.7999	7.6980	3.4439	5.8688	10.1226	5.4425	10.2404	19.3615
0.5000	3.4029	4.8985	6.9762	3.9137	5.7831	8.5921	5.5066	8.5060	13.1888
0.2500	3.9742	4.9542	6.1294	4.3717	5.4875	6.9106	5.2389	6.6500	8.4846
0.1000	4.3856	4.9547	5.5726	4.5887	5.1888	5.8727	4.9343	5.5903	6.3503
0.0100	4.6564	4.9399	5.2264	4.6797	4.9644	5.2554	4.7134	5.0004	5.2951
0.0010	4.6844	4.9377	5.1914	4.6867	4.9402	5.1943	4.6901	4.9438	5.1982
0.0001	4.6872	4.9375	5.1879	4.6874	4.9378	5.1882	4.6878	4.9381	5.1886
0.0000	4.6875	4.9375	5.1875	4.6875	4.9375	5.1875	4.6875	4.9375	5.1875

Table 4: Expected number of waiting customers for three preemption disciplines in Example 1

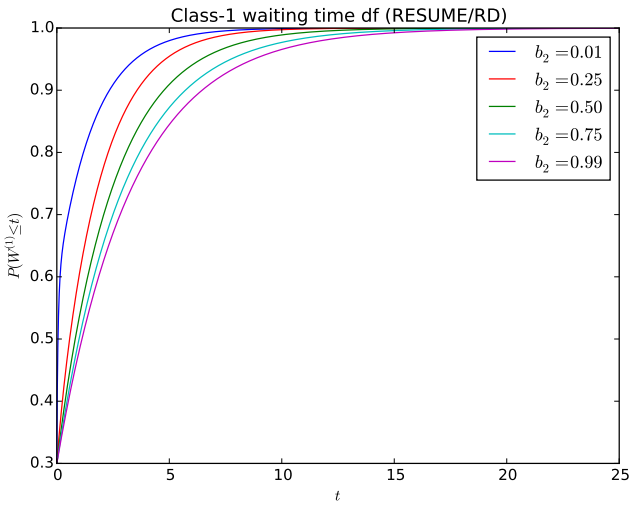
x	Resume			Repeat-Different			Repeat-Identical		
	$\mathbb{E}(N^{(1)})$	$\mathbb{E}(N^{(2)})$	$\mathbb{E}(N^{(3)})$	$\mathbb{E}(N^{(1)})$	$\mathbb{E}(N^{(2)})$	$\mathbb{E}(N^{(3)})$	$\mathbb{E}(N^{(1)})$	$\mathbb{E}(N^{(2)})$	$\mathbb{E}(N^{(3)})$
100.0000	0.2083	0.4583	1.0325	0.2083	0.5160	1.8005	0.2083	0.8308	14.2340
10.0000	0.2083	0.4584	1.0325	0.2084	0.5160	1.8005	0.2084	0.8316	14.2310
7.5000	0.2085	0.4587	1.0322	0.2085	0.5168	1.7996	0.2088	0.8407	14.1974
5.0000	0.2104	0.4626	1.0290	0.2109	0.5262	1.7891	0.2144	0.9474	13.7965
2.5000	0.2383	0.5080	0.9886	0.2508	0.6299	1.6608	0.3718	1.8094	9.8610
1.0000	0.4115	0.6397	0.8249	0.5085	0.8507	1.2011	1.0258	1.9756	3.2395
0.7500	0.4918	0.6720	0.7661	0.6110	0.8739	1.0536	1.1106	1.7201	2.2635
0.5000	0.6007	0.7024	0.6943	0.7284	0.8708	0.8844	1.1266	1.3961	1.4763
0.2500	0.7435	0.7262	0.6100	0.8429	0.8283	0.7009	1.0597	1.0510	0.8986
0.1000	0.8464	0.7349	0.5546	0.8972	0.7798	0.5892	0.9836	0.8561	0.6480
0.0100	0.9141	0.7374	0.5201	0.9199	0.7421	0.5234	0.9283	0.7489	0.5283
0.0010	0.9211	0.7375	0.5166	0.9217	0.7380	0.5170	0.9225	0.7386	0.5174
0.0001	0.9218	0.7375	0.5163	0.9219	0.7375	0.5163	0.9219	0.7376	0.5164
0.0000	0.9219	0.7375	0.5163	0.9219	0.7375	0.5163	0.9219	0.7375	0.5163

In our investigation, we consider the same model as Conway et al. (1967) with the exception that priority is assigned according to Eq. (6). The accumulating priority rates are such that $b_1 = 1$ and $0 \leq b_2 \leq 1$. Furthermore, we assume that $\lambda_1 = 0.4$ and $\lambda_2 = 0.3$. Our study focuses on the marginal waiting time distributions across several values of b_2. In particular, we compute waiting time probabilities for both classes via numerical inversion of the LST given by Eq. (27). To conduct the numerical inversion, we employ the two methods outlined in Abate and Whitt (1995). Both methods (referred to as EULER and POST-WIDDER) are used to confirm the accuracy of the overall numerical inversion. For our example, we employed the EULER and POST-WIDDER methods using the authors' suggested parameter settings (see Abate and Whitt, 1995, Section 3) and found that the two methods produced equivalent results.

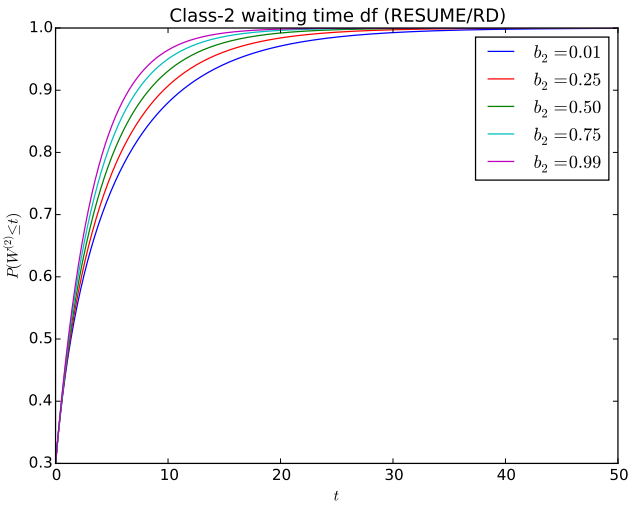
It is important to note that, in this example, the resume and repeat-different (RD) disciplines yield the exact same results. This is due to the memoryless property of the class-2 service time distribution. Figs. 5 and 6 plot the waiting time dfs of both classes (for various values

of b_2) under the resume/RD and repeat-identical (RI) disciplines, respectively. Furthermore, in Table 5, we calculate to 2 decimal places of accuracy several quantiles of the waiting time distributions under the resume/RD and RI disciplines, where $w_q^{(k)}$ denotes the q-th quantile of $W^{(k)}$ satisfying $\mathbb{P}(W^{(k)} \leq w_q^{(k)}) = q$. In addition, in Table 6 we compare the corresponding medians and expected values of $W^{(k)}$ for $k = 1, 2$.

We observe that the PAPQ approaches a FCFS queue as b_2 approaches one. However, the convergence appears to be slower under the RI discipline than it is in the resume/RD case. The benefit of the PAPQ here, as evidenced by Tables 5 and 6, is the ability to control waiting time distributions, allowing one to select the appropriate value of b_2 to satisfy a certain performance metric.

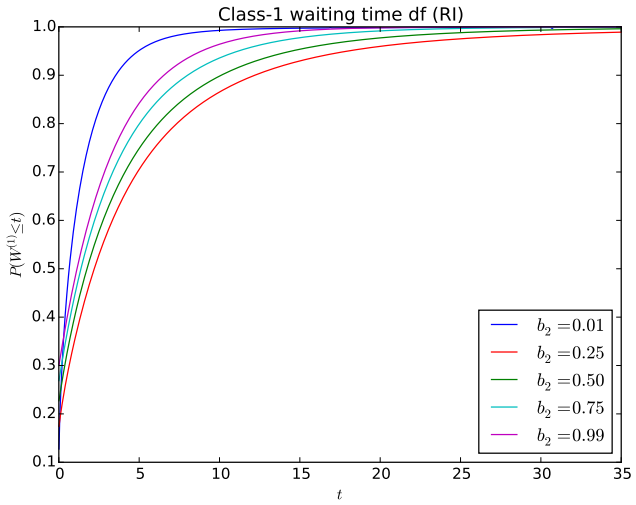


(a) Class-1 waiting time df

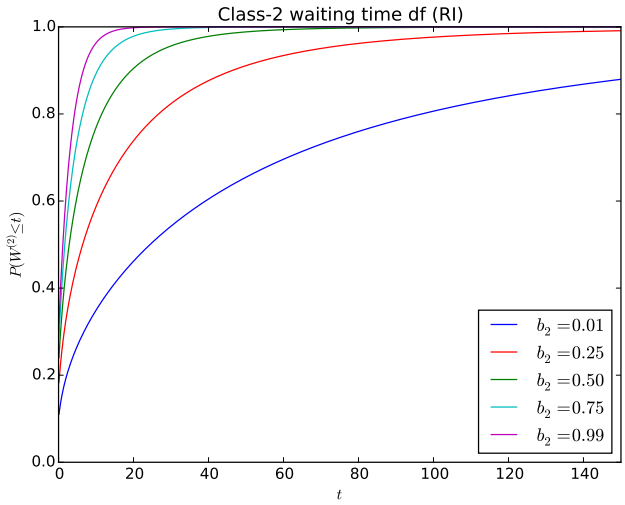


(b) Class-2 waiting time df

Figure 5: Marginal waiting time dfs for various values of b_2 (under RESUME/RD) in Example 2



(a) Class-1 waiting time df



(b) Class-2 waiting time df

Figure 6: Marginal waiting time dfs for various values of b_2 (under RI) in Example 2

Table 5: Some quantiles of $W^{(k)}$ $(k = 1, 2)$ for various values of b_2 in Example 2

Resume/Repeat-Different										
b_2	$w^{(1)}_{0.70}$	$w^{(2)}_{0.70}$	$w^{(1)}_{0.80}$	$w^{(2)}_{0.80}$	$w^{(1)}_{0.90}$	$w^{(2)}_{0.90}$	$w^{(1)}_{0.95}$	$w^{(2)}_{0.95}$	$w^{(1)}_{0.99}$	$w^{(2)}_{0.99}$
0.01	0.51	4.12	1.18	6.63	2.34	11.24	3.50	16.08	6.18	27.75
0.25	1.52	3.71	2.28	5.82	3.57	9.60	4.83	13.49	7.70	22.69
0.50	2.08	3.36	3.08	5.16	4.76	8.30	6.43	11.49	10.25	18.97
0.75	2.49	3.07	3.69	4.62	5.72	7.29	7.74	9.98	12.44	16.23
0.99	2.81	2.83	4.16	4.19	6.46	6.52	8.76	8.84	14.10	14.23
Repeat-Identical										
b_2	$w^{(1)}_{0.70}$	$w^{(2)}_{0.70}$	$w^{(1)}_{0.80}$	$w^{(2)}_{0.80}$	$w^{(1)}_{0.90}$	$w^{(2)}_{0.90}$	$w^{(1)}_{0.95}$	$w^{(2)}_{0.95}$	$w^{(1)}_{0.99}$	$w^{(2)}_{0.99}$
0.01	1.49	61.06	2.22	96.87	3.54	172.01	4.95	268.94	8.92	654.36
0.25	4.90	16.79	7.34	26.67	12.19	46.36	17.98	69.65	36.24	142.91
0.50	4.12	7.37	6.19	11.51	10.10	19.41	14.48	28.21	26.50	52.32
0.75	3.36	4.21	5.02	6.42	8.01	10.41	11.16	14.62	19.02	25.12
0.99	2.84	2.86	4.20	4.24	6.54	6.60	8.88	8.96	14.32	14.46

Table 6: Comparison of the median and mean of $W^{(k)}$ ($k = 1, 2$) for various values of b_2 in Example 2

	Resume/Repeat-Different				Repeat-Identical			
b_2	$w_{0.50}^{(1)}$	$w_{0.50}^{(2)}$	$\mathbb{E}(W^{(1)})$	$\mathbb{E}(W^{(2)})$	$w_{0.50}^{(1)}$	$w_{0.50}^{(2)}$	$\mathbb{E}(W^{(1)})$	$\mathbb{E}(W^{(2)})$
0.01	0.05	1.36	0.69	3.86	0.65	24.39	1.41	74.99
0.25	0.57	1.29	1.26	3.33	2.20	6.42	4.83	17.62
0.50	0.82	1.23	1.71	2.92	1.75	2.83	3.79	7.08
0.75	0.99	1.17	2.06	2.59	1.38	1.66	2.92	3.75
0.99	1.12	1.12	2.32	2.34	1.13	1.14	2.35	2.37

8 Conclusions

In this paper, we presented the PAPQ and obtained the steady-state waiting time distributions of each class. Our method of analysis used the maximal priority process of the PAPQ and related it to the maximal priority process of the FCFS $M/G/1$ queue with accumulating priority and blocking, as introduced by Fajardo and Drekic (2015). We stress that this approach mimics that used for the analysis of static preemptive priority models where one relates the virtual wait process in those models to the virtual wait process of the classical FCFS $M/G/1$ queue (e.g., see Brill, 2008).

As evidenced by our two numerical examples, the main benefit of incorporating the PAPQ is the ability to control waiting times. The ability to control waiting times has served as the primary motivation for researchers studying dynamic priority queues in the past. While this control has mainly been administered through the expected waiting times, our paper also enables one to control waiting times via other performance measures such as their quantiles. By appropriately selecting the parameters $\{b_k\}_{k=1}^N$, a system manager can fine-tune its system so as to satisfy a wide variety of performance metrics. We have also demonstrated that the static priority queue is a limiting case of the PAPQ. We further believe that several other previously analyzed static priority queueing models can be generalized by incorporating a similar accumulating prioritization structure.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. In particular, Steve Drekic acknowledges the financial support provided via the agencys Discovery Grants program (#238675-2010-RGPIN). The authors are very grateful to the anonymous referees for their careful reading of the paper and their useful suggestions that have improved the overall presentation of the paper.

References

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

Bagchi, U. (1984). A note on linearly decreasing, delay-dependent non-preemptive queue disciplines. *Operations Research*, 32:952–957.

Bagchi, U. and Sullivan, R. S. (1985). Dynamic, non-preemptive priority queues with general, linearly increasing priority. *Operations Research*, 33:1278–1298.

Brill, P. H. (2008). *Level Crossing Methods in Stochastic Models*. Springer, New York.

Conway, R. W., Maxwell, W. L., and Miller, L. W. (1967). *Theory of Scheduling*. Addison-Wesley, Reading.

Drekic, S. (2003). A preemptive resume queue with an expiry time for retained service. *Performance Evaluation*, 54:59–74.

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

Holtzman, J. M. (1971). Bounds for a dynamic-priority queue. *Operations Research*, 19:461–468.

Hsu, J. (1970). A continuation of delay-dependent queue disciplines. *Operations Research*, 18:733–738.

Jackson, J. R. (1960). Some problems in queueing with dynamic priorities. *Naval Research Logistics Quarterly*, 7:235–249.

Jackson, J. R. (1961). Queues with dynamic priority discipline. *Management Science*, 8:18–34.

Jackson, J. R. (1962). Waiting-time distributions for queues with dynamic priorities. *Naval Research Logistics Quarterly*, 9:31–36.

Jaiswal, N. K. (1968). *Priority Queues*. Academic Press, New York.

Kanet, J. (1982). A mixed delay dependent queue discipline. *Operations Research*, 30:93–96.

Keilson, J. and Servi, L. D. (1990). The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *Operations Research Letters*, 9:239–247.

Kleinrock, L. (1964). A delay dependent queue discipline. *Naval Research Logistics Quarterly*, 11:329–341.

Kleinrock, L. and Finkelstein, R. P. (1967). Time dependent priority queues. *Operations Research*, 15:104–116.

Netterman, A. and Adiri, I. (1979). A dynamic priority queue with general concave priority functions. *Operations Research*, 27:1088–1100.

Sharma, K. C. and Sharma, G. C. (1994). A delay dependent queue without preemption with general linearly increasing priority function. *Journal of the Operational Research Society*, 45:948–953.

Stanford, D. A., Taylor, P., and Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77:297–330.

Takagi, H. (1991). *Queueing Analysis, Volume 1, Vacation and Priority Systems, Part 1*. North Holland, Amsterdam.

Trivedi, S. K., Jain, M., and Sharma, G. C. (1984). A delay dependent queue with preemption. *Indian Journal of Pure and Applied Mathematics*, 15:1296–1301.

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operations Research*, 30:223–231.

Appendix

It is well-known that the first two moments of a random variable can be obtained from evaluating the first and second derivatives of the corresponding LST at $s = 0$. We use the LSTs given in Eqs. (13), (15), (16), and (21)–(26) to obtain the first two moments of the service-structure elements. Omitting the straightforward but tedious algebraic details, we simply present these formulas below:

$$\mathbb{E}(A_{p_{k+1}}^{(m+1)}) = \frac{\Lambda_{m-1}^{(k+1)}\mathbb{E}(A_{p_{k+1}}^{(m)}) + \lambda_m^{(k+1)}\mathbb{E}(R^{(m)})}{\Lambda_m^{(k+1)}\left(1 - \sum_{i=1}^{m-1}\lambda_i\frac{b_m - b_{m+1}}{b_i}\mathbb{E}(A_{np}^{(m)}) - \lambda_m^{(m+1)}\mathbb{E}(R^{(m)})\right)},\ m = 1, 2, \ldots, k$$

$$\mathbb{E}\big((A_{p_{k+1}}^{(m+1)})^2\big) = \frac{\gamma_m^{(m+1)}\mu_{m,2}\left(\Lambda_{m-1}^{(k+1)}\mathbb{E}(A_{p_{k+1}}^{(m)}) + \lambda_m^{(k+1)}\mathbb{E}(R^{(m)})\right)}{\Lambda_m^{(k+1)}\left(1-\sum_{i=1}^{m-1}\lambda_i\frac{b_m-b_{m+1}}{b_i}\mathbb{E}(A_{np}^{(m)}) - \lambda_m^{(m+1)}\mathbb{E}(R^{(m)})\right)^3}$$
$$+ \frac{(1-\gamma_m^{(m+1)}\mu_{m,1})\Big(\Lambda_{m-1}^{(k+1)}\mathbb{E}\big((A_{p_{k+1}}^{(m)})^2\big) + \lambda_m^{(k+1)}\mathbb{E}\big((R^{(m)})^2\big)\Big)}{\Lambda_m^{(k+1)}\left(1-\sum_{i=1}^{m-1}\lambda_i\frac{b_m-b_{m+1}}{b_i}\mathbb{E}(A_{np}^{(m)}) - \lambda_m^{(m+1)}\mathbb{E}(R^{(m)})\right)^3},\ m=1,2,\ldots,k$$

$$\mathbb{E}(A_{np}^{(k+1)}) = \frac{\sum_{i=1}^{k-1}\lambda_i\frac{b_k}{b_i}\mathbb{E}(A_{np}^{(k)}) + \lambda_k\mathbb{E}(R^{(k)})}{\gamma_k\left(1-\sum_{i=1}^{k-1}\lambda_i\frac{b_k-b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)}\mathbb{E}(R^{(k)})\right)}$$

$$\mathbb{E}\big((A_{np}^{(k+1)})^2\big) = \frac{\sum_{i=1}^{k-1}\lambda_i\frac{b_k}{b_i}\mathbb{E}\big((A_p^{(k)})^2\big) + \lambda_k\mathbb{E}\big((R^{(k)})^2\big)}{\gamma_k\left(1-\sum_{i=1}^{k-1}\lambda_i\frac{b_k-b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)}\mathbb{E}(R^{(k)})\right)^3}$$

$$\mathbb{E}(A_{p_0}^{(k+1)}) = \frac{\Lambda_{k-1}\mathbb{E}(A_{p_0}^{(k)}) + \lambda_k\mathbb{E}(R^{(k)})}{\Lambda_k\left(1-\sum_{i=1}^{k-1}\lambda_i\frac{b_k-b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k)}\mathbb{E}(R^{(k)})\right)}$$

$$\mathbb{E}\big((A_{p_0}^{(k+1)})^2\big) = \frac{\gamma_k^{(k+1)}\mu_{k,2}\left(\Lambda_{k-1}\mathbb{E}(A_{p_0}^{(k)}) + \lambda_k\mathbb{E}(R^{(k)})\right)}{\Lambda_k\left(1-\sum_{i=1}^{k-1}\lambda_i\frac{b_k-b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)}\mathbb{E}(R^{(k)})\right)^3}$$
$$+ \frac{(1-\gamma_k^{(k+1)}\mu_{k,1})\Big(\Lambda_{k-1}\mathbb{E}\big((A_{p_0}^{(k)})^2\big) + \lambda_k\mathbb{E}\big((R^{(k)})^2\big)\Big)}{\Lambda_k\left(1-\sum_{i=1}^{k-1}\lambda_i\frac{b_k-b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)}\mathbb{E}(R^{(k)})\right)^3}$$

Resume:

$$\mathbb{E}(R^{(k)}) = \big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}(X^{(k)})$$
$$\mathbb{E}\big((R^{(k)})^2\big) = \big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)^2\mathbb{E}\big((X^{(k)})^2\big) + \Lambda_{k-1}^{(k)}\mathbb{E}(X^{(k)})\mathbb{E}\big((A^{(k)})^2\big)$$
$$\mathbb{E}(G^{(k)}) = \mathbb{E}(X^{(k)})$$
$$\mathbb{E}\big((G^{(k)})^2\big) = \mathbb{E}\big((X^{(k)})^2\big)$$

Repeat-different:

$$\begin{aligned}
\mathbb{E}(R^{(k)}) &= \big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}(G^{(k)}) \\
\mathbb{E}((R^{(k)})^2) &= (1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\mathbb{E}\big((G^{(k)})^2\big) + \Lambda_{k-1}^{(k)}\mathbb{E}\big((A^{(k)})^2\big)\mathbb{E}(G^{(k)}) \\
&\qquad + 2\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\big(\mathbb{E}(G^{(k)})\big)^2 \\
\mathbb{E}(G^{(k)}) &= \frac{1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})}{\Lambda_{k-1}^{(k)}\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})} \\
\mathbb{E}\big((G^{(k)})^2\big) &= \frac{2}{\Big(\Lambda_{k-1}^{(k)}\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})\Big)^2}\left(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}) - \Lambda_{k-1}^{(k)}\mathbb{E}\Big(X^{(k)}e^{-\Lambda_{k-1}^{(k)}X^{(k)}}\Big)\right)
\end{aligned}$$

Repeat-identical:

$$\begin{aligned}
\mathbb{E}(R^{(k)}) &= \big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}(G^{(k)}) \\
\mathbb{E}\big((R^{(k)})^2\big) &= \big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}\big((G^{(k)})^2\big) + \Lambda_{k-1}^{(k)}\mathbb{E}\big((A^{(k)})^2\big)\mathbb{E}(G^{(k)}) \\
&\qquad + \frac{2\mathbb{E}(A^{(k)})\big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)}{\Lambda_{k-1}^{(k)}}\mathbb{E}\big((e^{\Lambda_{k-1}^{(k)}X^{(k)}}-1)^2\big) \\
\mathbb{E}(G^{(k)}) &= \mathbb{E}[\mathbb{E}(G^{(k)}|X^{(k)})] = \mathbb{E}\left(\frac{e^{\Lambda_{k-1}^{(k)}X^{(k)}}-1}{\Lambda_{k-1}^{(k)}}\right) = \frac{\widetilde{B}^{(k)}\left(-\Lambda_{k-1}^{(k)}\right)-1}{\Lambda_{k-1}^{(k)}} \\
\mathbb{E}\big((G^{(k)})^2\big) &= \mathbb{E}\big[\mathbb{E}\big((G^{(k)})^2|X^{(k)}\big)\big] \\
&= \mathbb{E}\left[\frac{2}{\Lambda_{k-1}^{(k)^2}}\left(e^{2\Lambda_{k-1}^{(k)}X^{(k)}} - e^{\Lambda_{k-1}^{(k)}X^{(k)}} - \Lambda_{k-1}^{(k)}X^{(k)}e^{\Lambda_{k-1}^{(k)}X^{(k)}}\right)\right] \\
&= \frac{2}{\Lambda_{k-1}^{(k)^2}}\left(\widetilde{B}^{(k)}(-2\Lambda_{k-1}^{(k)}) - \widetilde{B}^{(k)}(-\Lambda_{k-1}^{(k)}) - \Lambda_{k-1}^{(k)}\mathbb{E}(X^{(k)}e^{\Lambda_{k-1}^{(k)}X^{(k)}})\right)
\end{aligned}$$