A Generalization of $M/G/1$ Priority Models via Accumulating Priority

by

Val Andrei Fajardo

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctoral of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2015

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Abstract

Priority queueing systems are oftentimes set up so that arriving customers are placed into one of N distinct priority classes. Moreover, to determine the order of service, each customer (upon arriving to the system) is assigned a priority level that is unique to the class to which it belongs. In static priority queues, the priority level of a class-k ($k = 1, 2, \ldots, N$) customer is assumed to be constant with respect to time. This simple prioritization structure is easy to implement in practice, and as such, various types of static priority queues have been analyzed and subsequently applied to real-life queueing systems. However, the assumption of constant priority levels for the customers may not always be appropriate. Furthermore, static priority queues can often display poor system performance as their design does not provide systems managers the means to balance the classical trade-off inherent in all priority queues, that is: reducing wait times of higher priority customers consequently increases the wait times for those of lower priority.

An alternative to static priority queues are accumulating priority queues, where the priority level of a class-k customer is assumed to accumulate linearly at rate $b_k > 0$ throughout the class-k customer's time in the system. The main benefit of accumulating priority queues is the ability, through the specification of the accumulating priority rates $\{b_k\}_{k=1}^{N}$, to control the waiting times of each class. In the past, due to the complex nature of the accumulating prioritization structure, the control of waiting times in accumulating priority queues was limited — being administered only through their first moments. Nowadays, with the advent of a very useful tool called the maximal priority process, it is possible to characterize the waiting time distributions of several types of accumulating priority queues.

In this thesis, we incorporate the concept of accumulating priority to several previously analyzed static priority queues, and use the maximal priority process to establish the corresponding steady-state waiting time distributions. In addition, since static priority queues may be captured from accumulating priority queues, useful

comparisons between the considered accumulating priority queues and their static priority counterparts are made throughout this thesis. Thus, in the end, this thesis results in a set of extensive analyses on these highly flexible accumulating priority queueing models that provide a better understanding of their overall behaviour, as well as exemplify their many advantages over their static priority equivalents.

Acknowledgements

I would like to extend thanks to the many people who made this thesis possible.

Specific mention goes to my PhD supervisor, Dr. Steve Drekic. With his enthusiasm, his guidance, his wisdom, and his unrelenting support, he helped me overcome many of the research-related difficulties I faced in this thesis. I have certainly benefitted from his exceptional writing abilities and his unparalleled fine attention to detail throughout my PhD. He has been an invaluable mentor and I thank him wholeheartedly for all that he has done for me throughout this process.

I would like to thank Dr. Douglas Down, Dr. Qi-Ming He, Dr. Gordon Willmot, and Dr. Martin Lysy for agreeing to serve on my thesis committee and for reading my thesis. My sincere thanks also goes to Dr. Percy Brill who taught me the fundamentals of the level-crossing methodology during a mini workshop that was held at Brock University. I must also thank my masters supervisor Dr. Mei Ling Huang who encouraged me to pursue doctoral studies in the first place.

I am also very grateful to the support staff of the Department of Statistics and Actuarial Science, with a special mention to Mary Lou Dufton, for assisting me in many different ways.

I also wish to thank all of my friends and family - inlcuding my many aunts and uncles, my two lovely grandmothers, and my two brothers and sister - for always supporting me and inspiring me to be the best that I can be. For always believing in me even when I doubt myself, I wish to thank my soon-to-be wife Amy. Last, but certainly not least, I wish to thank my parents who have cared for, supported, taught, and unconditionally loved me from the day I was born. Thank you so much.

Dedication

To my wonderful parents, Carlos and Emilina.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background information and preliminaries

1.1.1 Queueing theory: the mathematical study of queueing systems

Queueing systems are comprised of the following two fundamental elements: (i) entities commonly referred to as customers that arrive to the system and require a particular servicing before departing, and (ii) the system's server(s) that fulfill the service requirements of these customers. The mathematical study of these systems is called *queueing theory*. For obvious reasons, queueing theorists are concerned with the study of queueing systems that are limited in resources, for which the likelihood of congestion or the formation of large queues (of the customers) is great.

In 1909, A.K. Erlang introduced and analyzed the first mathematical queueing model for the purpose of studying the congestion within telephone networks. Over one hundred years have since passed, and a survey of the current literature on the subject would easily verify the wide applicability of queueing theory. In particular, countless mathematical queueing models have been analyzed by researchers studying queueing systems inherent in various areas such as telecommunications (e.g., see Giambene (2005)), vehicular traffic control (e.g., see Boon (2011)), and health care scheduling (e.g., see Lakshmi and Iyer (2013)). Nevertheless, the rapid advancement of technology necessitates the further advancement of queueing theory and the

continued pursuit of the mathematical study of complicated queueing systems.

This thesis focuses on the study of a particular kind of queueing system, namely, *priority queueing systems*. These systems are particularly useful for situations when certain kinds of customers should (or need to) be given faster access times to the server(s). An obvious example of such a situation deals with the classification and overall care of patients arriving to an emergency room of a hospital. Another health care application involves the scheduling of patients requiring a specific surgery or transplant, for which a key factor in a patient's position on the wait list is its current health relative to that of the other patients. Several other examples of these sorts of situations also arise from call center applications and the scheduling of computer jobs.

While it is true that priority systems reduce the waiting times of the higher priority customers, they also necessarily increase the waiting times of the lower priority ones. This is the trade-off that a systems manager is faced with when deciding to incorporate a prioritization structure. For the classical static priority queue, this trade-off cannot at all be controlled or lessened, and thus, at times, leads to poor system performance. Hence, in an effort to provide a systems manager the ability to control the waiting times (amongst other performance measures), this thesis focuses on the analysis of accumulating priority queues.

The rest of the thesis is organized as follows. For the remainder of this chapter, we provide the necessary background information on mathematical queueing models as well as provide a literature review on priority queueing systems. In Chapter 2, we analyze a certain single-server queueing model that is without a prioritization structure. Nonetheless, this model and its analysis serves as a building block for the two priority queueing models presented later in Chapters 3 and 4. Finally, we offer some final remarks including several possible extensions for the models which we investigate in this thesis.

1.1.2 Mathematical setup of a queueing system

In developing a mathematical queueing model, we must always specify the characteristics of the two fundamental elements of the queueing system. In addition, there may be other characteristics of this queueing system which we may want to incorporate into the mathematical model. A very convenient notation used for cataloguing mathematical queueing models is the so-called *Kendall's notation*, which was first introduced in Kendall (1951).

In using Kendall's notation, a queueing system is labelled as $A/B/m/c$, where each individual letter specifies the characteristics of a certain element of the queueing system. Specifically,

(i) A specifies the arrival process of the customers,

(ii) B specifies the service requirements of the customers,

(iii) m specifies the number of servers that the queueing system has,

(iv) c specifies the capacity of the queueing system (i.e., the maximum number of customers that can occupy the system at any point in time).

Note that in queueing theory, there are commonly used symbols which can occupy both the first and second positions of Kendall's notation. These symbols, more often than not, are used to specify the distribution of inter-arrival times of customers (i.e., for A) and/or the distribution of the service times of the customers (i.e., for B). In the next subsection, we present two examples of such commonly used symbols. Furthermore, if the last symbol c is omitted, then the queueing system under consideration is assumed to have infinity capacity for customers.

In addition to the characteristics being specified through Kendall's notation, another very important characteristic of any queueing system is the so-called *service discipline*, which governs the order of service of the customers. Examples of some well-known service disciplines include the *first-come-first-served* (FCFS) and the *last-come-first-served* (LCFS) disciplines, which stipulate the order of service as their names suggest. Priority queueing systems employ priority service disciplines which

dictate the order of service on the basis of the priority levels of the customers present in the system.

Beyond these fundamental characteristics, several other assumptions can be made such as those pertaining to customer behaviour (e.g., the so-called *jockeying*, *balking*, and *reneging* of customers). In this thesis, we do not consider such assumptions. However, for a review of such queueing systems, as well as numerous others, we refer the reader to the notable queueing theory texts of Asmussen (2008), Bhat (2008), Cohen (1982), Gross et al. (2008), Kleinrock (1975, 1976), Prabhu (1997), and Takács (1962).

1.1.3 The $M/G/1$ queueing system and some of its fundamental results

In this subsection, we introduce the well-known $M/G/1$ queueing system and provide some of the key results related to it. First of all, in using Kendall's notation, the symbol M, standing for *Markovian* or *memoryless*, implies that the characteristic for which it is describing has an exponential distribution. Since M appears in the first position of Kendall's notation, this implies that an $M/G/1$ queue has exponential inter-arrival times. In other words, the customer arrivals to this system form a Poisson process. Another commonly used symbol within the Kendall notation framework is the symbol G, standing for *general*, which is used to imply that the characteristic for which it is describing follows a general distribution. Hence, an $M/G/1$ queue is a single-server queueing system in which customer service times are generally distributed.

We next present some distributional results within the $M/G/1$ framework for two of the most fundamental performance measures of any queueing system, namely the *busy period duration* and the *waiting time*. To do this, we first need to introduce some parameters for our $M/G/1$ queueing system. Hence, let λ denote the customer arrival rate, thereby implying that the distribution function (df) of the inter-arrival times is given by

$$F(t) = 1 - e^{-\lambda t}, \quad t \geq 0. \tag{1.1}$$

Next, we let X represent the generally distributed service time random variable, whose df and corresponding Laplace-Stieltjes transform (LST) we denote by

$$B(x) = \mathbb{P}(X \leq x) \quad \text{and} \quad \widetilde{B}(s) = \mathbb{E}(e^{-sX}), \tag{1.2}$$

respectively. Furthermore, note that we say that the server is *idle* whenever the server is not servicing a customer (i.e., simply because there are no customers present in the system). Conversely, when the server is not idle, it must mean that a customer is being served, and so, at those times, we simply say that the server is *busy*.

Now, if T denotes the duration of a typical busy period, then T represents the interval of time from the instant that the server first becomes busy to the next moment in time that the server becomes idle. Furthermore, it has been shown that the LST of T, $\widetilde{\Gamma}(s) = \mathbb{E}(e^{-sT})$, is the solution to the functional equation

$$\widetilde{\Gamma}(s) \equiv \widetilde{\Gamma}(s; \lambda, X) = \widetilde{B}(s + \lambda - \lambda\widetilde{\Gamma}(s)) \tag{1.3}$$

(e.g., see Conway et al. (1967, Section 8-3)). Moreover, it is straightforward to obtain the first two moments of T via differentiation of the above LST:

$$\mathbb{E}(T) = \frac{\mathbb{E}(X)}{1 - \rho} \tag{1.4}$$

and

$$\mathbb{E}(T^2) = \frac{\mathbb{E}(X^2)}{(1 - \rho)^3}, \tag{1.5}$$

where $\rho = \lambda\mathbb{E}(X)$ is known as the *traffic intensity*. It can be shown (e.g., see Takács (1962, Theorem 3, p. 58)) that if $\rho < 1$, then busy periods have finite lengths with probability 1 (i.e., $\mathbb{P}(T < \infty) = 1$). Conversely, if $\rho > 1$, then T has an improper df (i.e., $\mathbb{P}(T < \infty) < 1$).

Remark 1.1 *The distribution of the $M/G/1$ busy period is equivalent for all service disciplines which do not add work or insert idleness (e.g., both the FCFS and the LCFS disciplines).*

Although the above results are essential to this thesis, we typically use the corresponding results in a slight variant of the $M/G/1$ busy period. In particular, consider

an $M/G/1$ system whose *zero-wait* customers (i.e., those customers who initiate the busy period) have exceptional service so that their service time distributions differ from the distribution of the other subsequent service times (pertaining to customers who incur positive wait times). In general, we refer to such resulting busy periods as *delay* busy periods. Let T_d represent the complete duration of such a delay busy period. Furthermore, suppose that X_0, the zero-wait service time (or the initial delay), has df $B_0(x)$ and corresponding LST $\widetilde{B}_0(s)$. Then, the LST of T_d is given by

$$\widetilde{\Gamma}_0(s) \equiv \widetilde{\Gamma}_0(s; \lambda, X, X_0) = \widetilde{B}_0(s + \lambda - \lambda\widetilde{\Gamma}(s)), \tag{1.6}$$

where $\widetilde{\Gamma}(s)$ is the solution to Eq. (1.3). The associated first two moments are

$$\mathbb{E}(T_d) = \frac{\mathbb{E}(X_0)}{1-\rho} \tag{1.7}$$

and

$$\mathbb{E}(T_d^2) = \frac{\lambda\mathbb{E}(X^2)}{(1-\rho)^3}\mathbb{E}(X_0) + \frac{\mathbb{E}(X_0^2)}{(1-\rho)^2}. \tag{1.8}$$

When $\rho < 1$, we say that the system is *stable* or *stationary*. That is, limiting distributions of certain random variables are known to exist. For example, under such conditions, the limiting distribution for the waiting time of the n-th arriving customer (denoted by W_n) exists (e.g., see Takács (1962, Theorem 10, p. 69)). The associated LST is given by the *Pollaczek-Khinchin* formula for the $M/G/1$ system:

$$\lim_{n\to\infty} \widetilde{W}_n(s) = \widetilde{W}(s) = \frac{s(1-\rho)}{s - \lambda + \lambda\widetilde{B}(s)}. \tag{1.9}$$

For stationary queueing systems, ρ can be interpreted as the long-run fraction of time that the server is busy. Hence, an alternate representation of the stationary waiting time LST is

$$\widetilde{W}(s) = (1-\rho) + \rho\widetilde{W}_{BP}(s), \tag{1.10}$$

where $\widetilde{W}_{BP}(s)$ is the waiting time LST for customers who arrive during busy periods. From Eq. (1.9), it immediately follows that

$$\widetilde{W}_{BP}(s) = \frac{(1-\rho)(1-\widetilde{B}(s))}{\mathbb{E}(X)(s - \lambda + \lambda\widetilde{B}(s))}. \tag{1.11}$$

The first and second moments associated with the waiting time are

$$\mathbb{E}(W) = \frac{\lambda \mathbb{E}(X^2)}{2(1-\rho)} \tag{1.12}$$

and

$$\mathbb{E}(W^2) = \frac{\lambda \mathbb{E}(X^3)}{3(1-\rho)} + \frac{\left(\lambda \mathbb{E}(X^2)\right)^2}{2(1-\rho)^2}. \tag{1.13}$$

In the $M/G/1$ variant with zero-wait customers having exceptional service, the LST for the waiting time of a customer serviced within a delay busy period (with the same inputs as above) is given by

$$\widetilde{W}_{BP}(s) = \frac{(1-\rho)(1-\widetilde{B}_0(s))}{\mathbb{E}(X_0)(s-\lambda+\lambda\widetilde{B}(s))}. \tag{1.14}$$

Note that if $\widetilde{B}_0(s) = \widetilde{B}(s)$, then Eq. (1.14) is equivalent to Eq. (1.11). The first moment associated with the above LST is

$$\mathbb{E}(W_{BP}) = \frac{\lambda \mathbb{E}(X^2)}{2(1-\rho)} + \frac{\mathbb{E}(X_0^2)}{2\mathbb{E}(X_0)}. \tag{1.15}$$

1.2 Priority queueing systems: an introduction and a brief review of the literature

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

Oftentimes, the customers of a priority queueing system are categorized into a fixed number of distinct priority classes labelled with class indices $1, 2, \ldots, N$. Throughout the thesis, we use the symbol $\mathcal{C}_i$ which is to be read as "class-i customer". In general, we say that $\mathcal{C}_i$s are prioritized over $\mathcal{C}_j$s whenever $i < j$. With this setup,

one can assign priorities to customers quantitatively by using the so-called *priority functions*, which are generally class-dependent. We denote the priority function for the $\mathcal{C}_k$s by $q_k(t)$, where the argument t represents time.

A priority queueing system such that $q_k(t)$ is constant with respect to t for all $k = 1, 2, \ldots, N$ is known as a *static* priority queue, satisfying

$$q_k(t) = a_k, \quad k = 1, 2, \ldots, N, \tag{1.16}$$

where the set of constants $\{a_i\}_{i=1}^N$ are arranged so that $a_1 > a_2 > \cdots > a_N$. Furthermore, amongst all of the customers belonging to the same class, it is assumed that the oldest such customer is the one with the greatest priority. In other words, the service amongst the $\mathcal{C}_k$s is administered via the FCFS discipline.

Priority queues for which $q_k(t)$ is dependent on t have been more or less termed in the literature as dynamic priority queues. 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

$$q_k(t) = \phi_k(t - \tau_k), \quad t \geq \tau_k, \quad k = 1, 2, \ldots, N, \tag{1.17}$$

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

$$\phi_1(0) \geq \phi_2(0) \geq \cdots \geq \phi_N(0) \tag{1.18}$$

and

$$\phi_1'(x) \geq \phi_2'(x) \geq \cdots \geq \phi_N'(x) \geq 0 \qquad \text{for all } x > 0. \tag{1.19}$$

For $i < j$, note that Eq. (1.18) 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. (1.19) implies that a $\mathcal{C}_i$ earns priority at least as fast as a $\mathcal{C}_j$ does. Hence, for dynamic priority queues employing the general Priority Service Guideline, Eq. (1.18) and Eq. (1.19) imply that, within a given class, service is administered based on the order of arrival (as in the case of the static priority discipline).

Another very important distinction of priority queues is based on the decision of whether or not to interrupt the servicing of a customer for another higher priority

customer present in the system. In this regard, there are generally three types of priority queues:

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

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

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

For the preemptive and mixed types of priority queueing systems, the rule governing the servicing of an interrupted customer, upon its re-entry into service, must be specified, and can be performed via any one of the following three traditional disciplines:

(i) *Resume*: service of the interrupted customer continues from where it was interrupted,

(ii) *Repeat-different*: all previous work is lost and a new service time is independently sampled from the corresponding service time distribution,

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

In addition to these required specifications of a priority queueing system, additional features pertaining to customer behaviour may also be incorporated. For example, priority queueing systems may also include customer reneging (i.e., customers who abandon the queue while waiting for the server), jockeying (i.e., customers vying for better position while in the queue), and balking (i.e., customers who arrive to the system and decide not to enter the queue at all). We note, however, that this thesis focuses on priority queueing systems that do not incorporate such customer behaviours.

We now provide a review of the literature on priority queueing models, beginning with those queueing systems for which the assignment of priority to the customers is static. The first static non-preemptive priority queue was analyzed by Cobham (1954), while the idea of preemption seemed to originate in the paper by White and Christie (1958). Nowadays, these priority models are coined as being the "classical" priority queueing systems, which have been rigorously analyzed by numerous queueing theorists. For a detailed analysis on both static non-preemptive and preemptive priority queues, we refer the reader to the texts by Conway et al. (1967), Jaiswal (1968), and Takagi (1991).

With regards to mixed priority queues, several researchers have previously considered various guidelines and discretion rules to dictate the interruptions of service. A well-known guideline for prescribing interruptions based solely on the class indices is the so-called *preemption distance* (PD) rule. The PD rule allows for preemption only if the difference in the class indices of the two customers under consideration exceeds a specified value. Adiri and Domb (1982, 1984) and Paterok and Ettl (1994) have analyzed static priority queues implementing the PD rule. Mixed priority queues for which the discretion rules are based on the service time of the customer currently in service have also been previously considered. For example, three such discretion rules are:

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

(ii) *Front-end time-based* (FETB) *policy:* Once $\mathcal{T}$ time units of service have been successfully rendered, further preemptions are prevented,

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

The above "threshold-based" discretion rules were first studied by Cho and Un (1993). Later, Drekic and Stanford (2000) considered a generalized version of these discretion rules by allowing the threshold parameters to be class-dependent.

Shifting the focus of our discussion now to dynamic priority queues, we remark that 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

$$q_k(t) = a_k + (t - \tau_k), \quad t \geq \tau_k, \tag{1.20}$$

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 notably 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 inter-arrival and service times (i.e., an $M/M/1$-type priority queue) using priority functions of the form

$$q_k(t) = b_k \cdot (t - \tau_k), \quad t \geq \tau_k, \tag{1.21}$$

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 service discipline as the *delay dependent priority* discipline. Kleinrock and Finkelstein (1967) subsequently extended this work by considering the same $M/M/1$-type priority system but 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. (1.20) for which he derived both upper and lower bounds for the marginal expected waiting times of each class.

Netterman and Adiri (1979) followed up and 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. In addition, 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 also 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. (1.21) 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. (1.21) 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. It is perhaps the case that the overall complexity of these models is what deterred researchers from determining the distributions of the 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. (1.21)) and applied it to an $M/G/1$-type priority system. With a newly defined stochastic process, called the *maximal priority process*, Stanford et al. (2014) shed new light on the specific structuralization of such a dynamic priority queue. Ultimately, by virtue of the maximal priority process, these authors derived the LST

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

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.

Unlike its counterpart (i.e., static priority queues), however, 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 preemptive resume discipline in Hsu's (1970) decreasing priority model. Once again, the analysis therein focused on finding the steady-state expected waiting times of each class.

In Chapter 3 of the thesis, we consider the preemptive priority queue with priority functions of the form given by Eq. (1.21), whereas in Chapter 4, we analyze a mixed priority queueing system using a generalization of the threshold-based discretion rules introduced earlier. The analysis of both of these models borrows results from the analysis of the $M/G/1$-type queueing system considered in Chapter 2, which incorporates a new blocking policy called the q-policy.

1.3 Main contributions

As a whole, this thesis advances the study of accumulating priority queues. Specifically, it furthers the knowledge of the maximal priority process, providing a better understanding of how it can be used as tool in the analysis of accumulating priority queues. As a specific example, the maximal priority process is used to obtain the waiting time distributions in the fully preemptive accumulating priority queue under all three of the traditional preemption disciplines: resume, repeat-different, and repeat-identical. In deriving the class-k waiting time LSTs for this accumulat-

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

ing priority queue and for the others considered in this thesis, we empower systems managers to control, through the specification of the accumulating priority rates $\{b_k\}_{k=1}^{N}$, several aspects of the class-k waiting time distribution, including its moments and quantiles.

Our analyses of accumulating priority queues is indeed quite exhaustive in that, in addition to establishing the waiting time distributions, formulas for several other important quantities that provide further insight into the characteristics of these queueing models (and which can provide alternate measures of overall system performance) are obtained. It also bears mentioning that the models considered in this thesis are quite general, capturing a wide variety of previously analyzed static priority queues as special cases and allowing for useful comparisons between old and new models to be made with ease. Finally, we remark that this thesis provides the first-ever (to the best of our knowledge) analysis on a dynamic preemptive priority queue under the two preemptive repeat service disciplines (repeat-different and repeat-identical), as well as the first-ever analysis on a mixed dynamic priority queueing system.

Chapter 2

$M/G/1$ queue under the q-policy

2.1 Introduction

In this chapter, we study an $M/G/1$-type queueing model in which the arrival process is controlled by a systems manager so as to decrease the lengths of the general busy period. In some applications, for example, a systems manager may be more inclined to regularly decrease the overall length of the busy period if it is the case that the server/machine becomes highly susceptible to expensive breakdowns after operating for extended periods of time. These breakdowns can be costly both in terms of the repair costs and the opportunity costs due to closures of the system. To alleviate the risk of incurring an expensive breakdown, a systems manager may choose to rest the server/machine during *closedown* periods on a regular basis. In addition, cost-effective *maintenance checks* can be performed during these rest periods to ensure the long-run functionality of the machine.

In what follows, we present one such policy which would allow a systems manager to control busy period lengths. Specifically, during each busy period, the control is exercised by closing the system to potential customers over a constant proportion of the overall busy period. The flexibility to disallow (or to block) customers from entering the system may be desirable if, for instance, a *holding cost* for customers during their sojourn in the system exists. Our aim here is to study the effect of the new policy, which we refer to as the *q-policy*, on various performance measures of

interest such as the length of busy periods and the wait of serviceable customers.

The literature on the optimal design and control of queueing systems is quite extensive. In regards to the arrival control of queueing systems, the usual goal is to find the optimal policy which maximizes (or minimizes) a specific objective function. In the seminal paper by Naor (1969), an $M/M/1$-type queueing system is studied where the arrival process is controlled by the administration of a toll charge for arriving customers. In particular, customers receive a fixed reward K upon successful service but also incur a holding cost h per unit time spent in the system. Naor studies the optimal policies from two perspectives, namely:

(i) individual optimization, where the objective function is the individual expected net benefit rate function, and

(ii) social optimization, where the objective function is the expected overall net benefit rate function.

Naor assumes that the optimal policies for both problems is of the critical number form (i.e., customers are accepted for service if the number of customers currently occupying the system is less than the critical number), and this form of optimal policy can be validated through the use of Markov decision processes (see Stidham (2002) and references therein). Under this framework, Naor establishes a key result which states that an individually optimal policy admits more customers than its counterpart, the socially optimal policy.

Naor's work inspired several other researchers to consider various generalizations for both the model and the net benefit rate structure. Rue and Rosenshine (1981) considered Naor's model and studied the effect of the arrival rate on the parameters for both kinds of optimal policies. Yechiali (1971) extended Naor's work by relaxing the assumption of the arrival process to be merely a renewal process. The $M/M/s$ variant was considered by Knudsen (1972) where Naor's main result was shown to still hold true. Doshi (1977) considered the continuous-time arrival control of an $M/G/1$ queueing system which operated under a policy that opened and closed the system to potential arrivals depending on the level of the workload. Johansen and Stidham (1980) showed that Naor's main result actually holds true under a

set of fairly general conditions (e.g., dependent arrivals, batch arrivals, and random rewards). For excellent surveys of the literature, we refer the interested reader to Stidham (1985, 2002). To the best of our knowledge, the q-policy presented in this chapter has not been previously studied.

The optimal policies found by these researchers has usually resulted in the formulation of threshold-form policies (i.e., thresholds for the number of customers in the system or for the residual workload). We emphasize, however, that our focus is not one that searches for an optimal policy which maximizes a specific objective function, but instead analyzes the effects of a given policy which aims to lessen the workload of a system. Nevertheless, we do formulate an optimization problem in Section 2.6 which illustrates that, in certain situations, the reduction of the busy period lengths via the q-policy can result in increased profits.

The rest of the chapter is organized as follows. In the next section, we introduce the queueing model and the q-policy. Section 2.3 is devoted to the study of the busy period as well as some fundamental steady-state probabilities associated with the system. The steady-state waiting time distribution of serviceable customers is analyzed in Section 2.4 by virtue of the level-crossing methodology. In Section 2.5, we present a queueing model which enables a systems manager to block customers during busy periods similar to the q-policy, but has the property that it does not require knowledge of the service times upon arrival. A numerical example is provided in Section 2.6. We remark that most of the work presented in this chapter is found in Fajardo and Drekic (2015a).

2.2 The model and the q-policy

We assume that the Poisson arrival rate of customers to the system is $\lambda > 0$. If the system is *open* (i.e., accepting of new customers) when a customer arrives, then this customer joins the queue (which is assumed to have infinite capacity). Otherwise, the customer is lost and unrecoverable. Let $\{X_i\}_{i=1}^{\infty}$ denote the sequence of independent and identically distributed (iid) customer service times having common mean $\mu = \mathbb{E}(X_i)$ and common second moment $\gamma = \mathbb{E}(X_i^2)$. Similar to the model studied by

Johansen and Stidham (1980), the customer service times are assumed to be known to the server (or systems manager) immediately upon a customer's entry to the system. We denote the corresponding df and LST by

$$B(x) = \mathbb{P}(X_i \leq x) \quad \text{and} \quad \widetilde{B}(s) = \int_0^\infty e^{-sx} \mathrm{d}B(x), \tag{2.1}$$

respectively. The FCFS service discipline is used to govern the order of service for the admitted customers. We denote the traffic intensity of the classical (i.e., unblocked) $M/G/1$ queue, as usual, by $\rho = \lambda\mu$. Note that we reserve the notation $\bar{B}(x) = \mathbb{P}(X_i > x)$ for the complementary df of X_i.

Before formally introducing the q-policy, we recall that for an arbitrary busy period of the classical (work-conserving) $M/G/1$ queue, any customer who arrives during this busy period will always be admitted for service (i.e., they will eventually be served in this busy period). However, suppose that a systems manager would like to restrict (or control) the arrival process during a busy period, so that the system is not obligated to serve all customers who arrive during the busy period. In such a situation, a systems manager could, for intervals of time within the busy period, close the system to potential arrivals. A blocking policy provides a set of guidelines which allows a systems manager to administrate the openings and closures of the system. We denote such a policy in general by $\pi(t)$, where $\pi(t) = 1$ implies that the system is open at time t, and similarly $\pi(t) = 0$ implies that the system is closed at time t. An example of such a blocking policy is the q-policy, denoted by $\pi_q(\cdot)$, which we define next.

Definition 2.1 (The q-policy) *Without loss of generality, assume that a customer arrives to an empty queue at time $\tau_1 = 0$, thereby initiating the start of a busy period. For all $t \geq 0$ during this busy period, we define the process $\{R(t), t \geq 0\}$, which is similar to the workload process. In particular, for $0 \leq q \leq 1$:*

1. *$R(0) = (1-q)X_1$, where X_1 is initial customer's service time.*

2. *$R(t)$ decreases at unit rate unless the process is at level 0.*

3. For the sequence of customer arrival epochs, $\{\tau_i\}_{i=2}^{\infty}$, during this busy period,

$$R(\tau_i) = \begin{cases} R(\tau_i^-) + (1-q)X_i & \text{if } R(\tau_i^-) > 0 \\ 0 & \text{if } R(\tau_i^-) = 0 \end{cases}, \tag{2.2}$$

where $R(t^-) = \lim_{\epsilon \to 0} R(t-\epsilon)$.

Then, for all $t \geq 0$ *during this busy period,*

$$\pi_q(t) = \begin{cases} 1 & \text{if } R(t) > 0 \\ 0 & \text{if } R(t) = 0 \end{cases}. \tag{2.3}$$

Remark 2.2 *The process* $\{R(t), t \geq 0\}$ *acts as a timer for the busy period. That is,* $R(t)$ *represents the time remaining, at time* t*, before the system is closed to potential arrivals.*

Figure 2.1 illustrates a busy period under the q-policy. Here, at some point during the servicing of the third customer (denoted by C_3), the timer becomes drained (i.e., $R(\cdot)$ hits level 0), and this results in the system becoming closed to potential arrivals. Hence, both customers C_5 and C_6 are blocked from entering the system. It is important to note that, although the system is closed at this point, the server must still complete the servicing of C_3 and C_4. In other words, the busy period terminates when all admitted customers have been fully served. Moreover, the end of the busy period signals the reopening of the system and the commencement of the ensuing idle period which ends at the next customer arrival instant. The busy period and the subsequent idle period together form a *busy cycle*.

Clearly, under the q-policy, the resulting busy periods are stochastically smaller than those corresponding to a system not implementing any sort of blocking policy. It is also apparent that if we set $q = 0$, then $\{R(t), t \geq 0\}$ exactly becomes the so-called *workload process* during a busy period in the classical $M/G/1$ queue. In fact, a blocking proportion equal to zero simply implies that no customers are blocked from service, and thus the resulting model is equivalent to the classical $M/G/1$ queue. On the other hand, with $q = 1$, the system is closed to potential customers throughout the entire busy period (i.e., $R(t) = 0$ for all t), implying that only the customers that arrive to an empty system are accepted for service. As a result, we obtain the $M/G/1/1$ queue as a special case when $q = 1$.

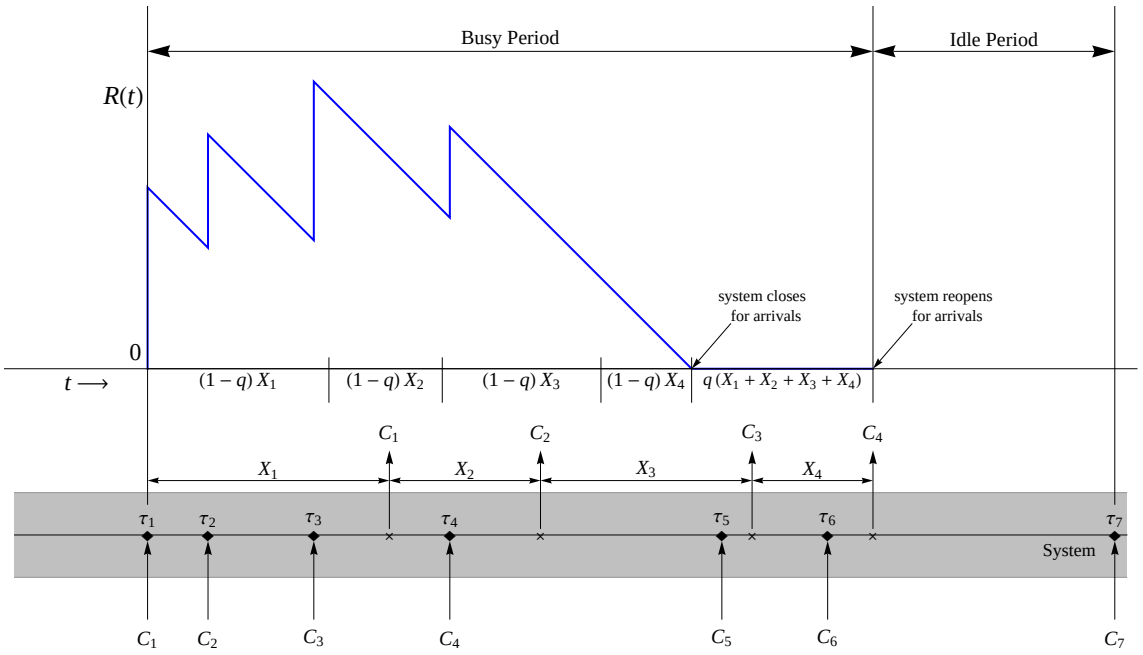


Figure 2.1: A typical busy period under the q-policy

2.3 The busy period and some steady-state probabilities

In this section, we first establish a functional equation for the LST corresponding to the distribution of the busy period duration operating under the q-policy. Let T be the length of such a busy period, whose df and LST are denoted by $G(x)$ and $\widetilde{G}(s)$, respectively.

To derive the LST of T, we note that the order in which serviceable customers are served does not, in any way, affect the duration of the busy period. As in the classical case, this important observation leads to the derivation of a functional equation for $\widetilde{G}(s)$. We now introduce a new service discipline which we refer to as the *q-restricted last-come-first-served (q-restricted LCFS for short)* discipline. First of all, recall that $\{R(t), t \geq 0\}$ consists of up-jumps at the arrival epochs of each serviceable customer, and further that the magnitude of the jump is equal to the service time

of the customer multiplied by $(1-q)$. Let us refer to these entities simply as the unblocked portions of the service times. Now, the order of service determined by the q-restricted LCFS discipline is precisely the order of service obtained by applying the usual LCFS discipline to a system in which the unblocked portions are effectively considered as the actual service times (i.e., $(1-q)X_i$ instead of X_i).

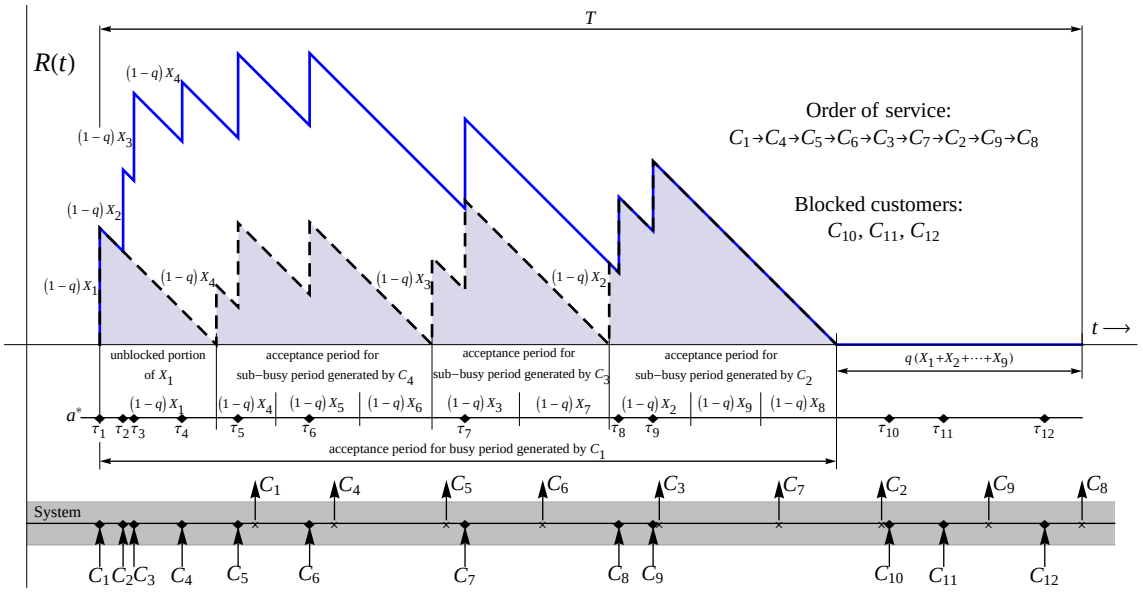


Figure 2.2: A busy period under the q-restricted LCFS discipline

Figure 2.2 demonstrates the q-restricted LCFS discipline in a typical busy period. Again, we determine the order of service under this discipline by effectively considering the unblocked portions as the actual service times. Specifically, in Figure 2.2, one can determine the order of service by projecting the arrival epochs to the a^*-axis and applying the usual LCFS discipline. Moreover, under the q-restricted LCFS discipline, we see that the interval of time during which $R(t)$ is positive (i.e., the system is open to accepting new customers) can be decomposed into smaller, well-understood subintervals of time. Indeed, these subintervals are merely the acceptance periods of their corresponding sub-busy periods. For example, in Figure 2.2, C_4 generates a sub-busy period in which C_5 and C_6 both are serviced; the length of the acceptance period for this sub-busy period is equal to $(1-q)\times(X_4+X_5+X_6)$.

It is clear that these sub-busy periods are identically distributed to the overall busy period (generated by C_1). However, we do note that in the intermediate sub-busy periods (i.e., sub-busy periods generated by C_4 and C_3 in Figure 2.2), customers who fail to arrive in their acceptance periods are not blocked from the system, but instead are serviced in the next sub-busy period.

Theorem 2.3 *If $\lambda^{(q)} = \lambda(1-q)$ and $\rho^{(q)} = \lambda^{(q)}\mu < 1$, then T has a proper (i.e., non-defective) distribution and its corresponding LST satisfies the functional equation*

$$\widetilde{G}(s) = \widetilde{B}(s + \lambda^{(q)}(1 - \widetilde{G}(s))). \tag{2.4}$$

Proof. Similar to the LST derivation of the busy period duration in the classical $M/G/1$ queue (e.g., see Kleinrock (1975, Section 5.8)), we invoke the fact that T is independent of the service discipline, so long as it is a work-conserving one. Kleinrock's derivation involves the usual LCFS discipline, but here, we employ the q-restricted LCFS discipline. Define N to be the number of customers who arrive during the unblocked portion of the initial customer's service time. As discussed above, each of the N customers generates a sub-busy period of their own which is identically distributed to the overall busy period and, moreover, is mutually independent from the others.

Conditioning on both $N = n$ and the first service time $X_1 = x$, we obtain

$$\mathbb{E}(e^{-sT}|X_1 = x, N = n) = e^{-sx}\big(\widetilde{G}(s)\big)^n. \tag{2.5}$$

Given $X_1 = x$, N is Poisson distributed with rate $\lambda^{(q)}x$, and this leads to

$$\mathbb{E}(e^{-sT}|X_1 = x) = e^{-sx}e^{-\lambda^{(q)}x}\sum_{n=0}^{\infty}\frac{\big(\lambda^{(q)}x\widetilde{G}(s)\big)^n}{n!} = e^{-x(s+\lambda^{(q)}-\lambda^{(q)}\widetilde{G}(s))}. \tag{2.6}$$

Lastly, removing the condition on X_1 immediately yields

$$\widetilde{G}(s) = \mathbb{E}(e^{-sT}) = \widetilde{B}(s + \lambda^{(q)}(1 - \widetilde{G}(s))), \tag{2.7}$$

and the result is proven. □

As in the classical case, we are left with an implicit expression for the LST of T. Nonetheless, we are still able to obtain the moments of T through successive differentiation. In particular, the first two moments of T are:

$$\mathbb{E}(T) = \frac{\mu}{1-\rho^{(q)}}, \tag{2.8}$$

$$\mathbb{E}(T^2) = \frac{\gamma}{(1-\rho^{(q)})^3}. \tag{2.9}$$

Remark 2.4 *Theorem 2.3 implies that the busy period under the q-policy is distributed equivalently to the busy period of a classical $M/G/1$ queue with arrival rate $\lambda^{(q)}$ and service time distribution $B(\cdot)$ (i.e., $\widetilde{G}(s) = \widetilde{\Gamma}(s; \lambda^{(q)}, X_i)$ as defined by Eq. (1.3)). Furthermore, the busy period is also equivalently distributed to the busy period of an $M/G/1$ queue with the following Bernoulli-type blocking policy:*

(i) customers arrive according to a Poisson process with rate $\lambda > 0$;

(ii) at each customer arrival epoch, the server conducts a Bernoulli experiment, where with probability $(1-q)$ the customer is admitted for service, and with probability q the customer is blocked.

A common feature of this model with the system under the q-policy is that during busy periods, the probability that an arriving customer is blocked from entering the system is precisely q.

We next establish the form of the probability generating function (pgf) for N_{bp}, the number of customers served in a busy period. We define $m(z) = \mathbb{E}(z^{N_{bp}})$ to be the pgf of N_{bp}. Like the duration of the busy period T, the number served in a busy period is unaffected by the order of service. Hence, by implementing the q-restricted LCFS discipline, we obtain

$$\mathbb{E}(z^{N_{bp}}|N=n) = \mathbb{E}(z^{1+M_1+M_2+\cdots+M_n}), \tag{2.10}$$

where N is the number of customers in the initial queue (i.e., those customers arriving during the unblocked portion of the initial customer's service time) and M_i

denotes the number of customers served in the i-th customer's sub-busy period. By independence, we have

$$\mathbb{E}(z^{N_{bp}}|N=n) = z\big(m(z)\big)^n. \tag{2.11}$$

It immediately follows, by removing the condition on N, that

$$m(z) = z\widetilde{B}\big(\lambda^{(q)}(1-m(z))\big), \tag{2.12}$$

from which the first moment of N_{bp} is readily given by

$$\mathbb{E}(N_{bp}) = \frac{1}{1-\rho^{(q)}}. \tag{2.13}$$

To conclude this section, we shift our focus to the derivation of some key steady-state probabilities of the system, namely:

$P_I \equiv$ steady-state probability the server is idle;

$P_B \equiv$ steady-state probability the server is busy;

$P_{B,0} \equiv$ steady-state probability the server is busy and the system is closed;

$P_{B,1} \equiv$ steady-state probability the server is busy and the system is open.

To obtain these probabilities, we apply the theory of regenerative processes (e.g., see Kao (1996, Section 3.6)). Define a busy cycle, D, to consist of a busy period T and the ensuing idle period I (i.e., $D = T + I$). Clearly, the set of regeneration points associated with D are the epochs defined by busy period commencements. Thus, from elementary renewal theory, we readily obtain:

$$P_I = \frac{\mathbb{E}(I)}{\mathbb{E}(D)} = \frac{1-\rho^{(q)}}{1+\rho q}, \tag{2.14}$$

$$P_B = \frac{\mathbb{E}(T)}{\mathbb{E}(D)} = \frac{\rho}{1+\rho q}, \tag{2.15}$$

$$P_{B,0} = qP_B = \frac{\rho q}{1+\rho q}, \tag{2.16}$$

$$P_{B,1} = (1-q)P_B = \frac{\rho^{(q)}}{1+\rho q}. \tag{2.17}$$

2.4 Steady-state wait of serviceable customers

2.4.1 The workload and virtual wait processes

The motivation for our study of the virtual wait process stems from the well-known fact that for $M/G/1$-type queues, the distributions of virtual wait and actual wait are equivalent in steady-state. In what follows, we denote the (unfinished) workload process under a q-policy by $\{U_q(t), t \geq 0\}$, whereas the virtual wait process is denoted by $\{W_q(t), t \geq 0\}$.

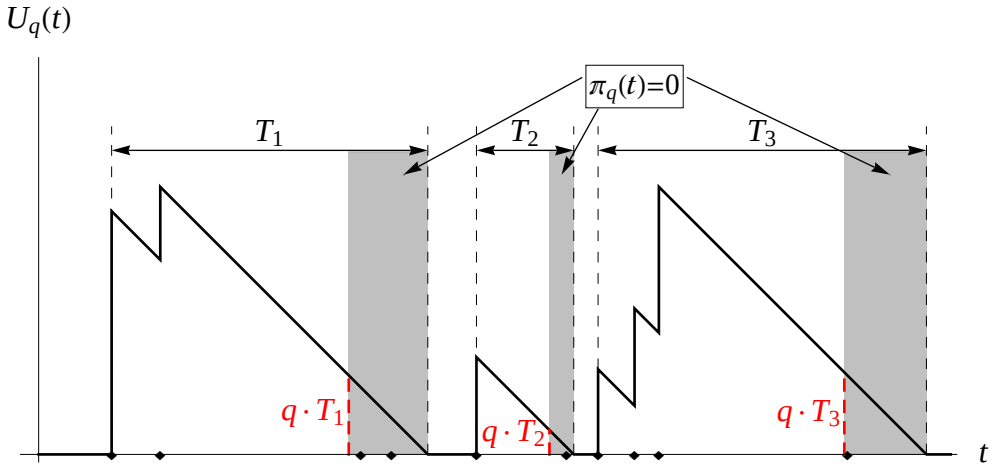
Obviously, $\{U_0(t), t \geq 0\}$ and $\{W_0(t), t \geq 0\}$ are the corresponding workload and virtual wait processes for the classical $M/G/1$ system. Now, for times $t > 0$ when the system is open (i.e., $\pi_q(t) = 1$), one notes that $U_q(t)$ behaves in the same manner as the $U_0(t)$ in that:

(i) $U_q(t)$ decreases at unit rate, except during times of idleness,

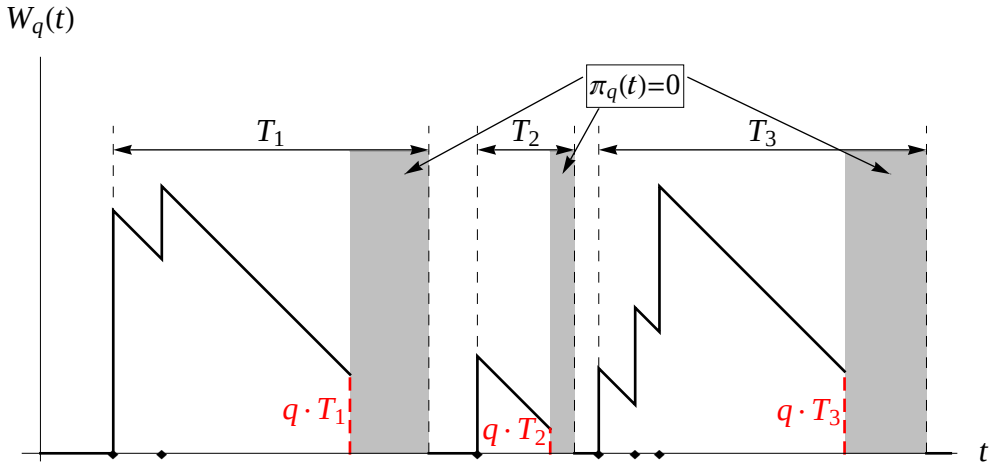
(ii) $U_q(t)$ up-jumps at customer arrival epochs, with the magnitude of the jumps being equal to the arriving customer's service time.

On the other hand, for times $t > 0$ when $\pi_q(t) = 0$, we have that $U_q(t)$ decreases at unit rate. In particular, if $t_* > 0$ is such that $\pi_q(t_*) = 0$ and $\pi_q(t_*^-) = 1$, then starting from time t_*, the workload depletes at unit rate until it hits level 0. Now, similar to how $\{U_0(t), t \geq 0\}$ and $\{W_0(t), t \geq 0\}$ are equivalent processes, during times t when the system is open, the processes $\{W_q(t), t \geq 0\}$ and $\{U_q(t), t \geq 0\}$ are also equivalent. However, the virtual wait process is further complicated by the fact that during a closure period for the system, the process is essentially undefined (i.e., does not exist).

Figure 2.3 depicts the sample paths of both processes for three consecutive busy periods of the system. The grey-shaded regions correspond to the times during which the system is closed (i.e., $\pi_q(t) = 0$), and thus, also represent the times when $W_q(t)$ is undefined. Customer arrival epochs are marked on the time axis with diamond symbols, and observe that both processes up-jump at arrivals occurring only during times when the system is open. As is also evident from Figure 2.3, the instant in time



(a) A typical sample path of the workload process



(b) Corresponding sample path of the virtual wait process

Figure 2.3: Typical sample paths of the processes $\{U_q(t), t \geq 0\}$ and $\{W_q(t), t \geq 0\}$

at which the system becomes closed during a busy period is exactly the same instant in time that $W_q(t)$ (or equivalently $U_q(t)$) hits level qT_i, where T_i is the duration of the i-th busy period. In what follows, we define $G_q(x) = 1 - \bar{G}_q(x) = \mathbb{P}(qT \leq x) = G(x/q)$ as well as $\widetilde{G}_q(s) = \mathbb{E}(e^{-s(qT)}) = \widetilde{G}(sq)$.

In order to study the wait of admitted customers, it is clear that we must analyze the virtual wait process only during times of its existence. Hence, we introduce the *censored* virtual wait process $\{\mathcal{W}_q(t), t \geq 0\}$, as illustrated in Figure 2.4. This process can be considered as $\{W_q(t), t \geq 0\}$ with the censorship (or removal) of the periods of non-existence. Indeed, by simply removing these periods, the resulting censored process will have a different time clock than the non-censored version. However, due to the memoryless property of the Poisson arrival process, the analysis of $\{W_q(t), t \geq 0\}$ during its times of existence is equivalent to the analysis of $\{\mathcal{W}_q(t), t \geq 0\}$.

As is evident in Figure 2.4, the sample path never continuously hits level 0 (unless $q = 0$), but instead always down-jumps to level 0. Furthermore, the magnitude of these down-jumps have distribution $G_q(\cdot)$. This simple observation allows us to derive the steady-state integral equation for the probability density function (pdf) of the virtual wait (during times of its existence).

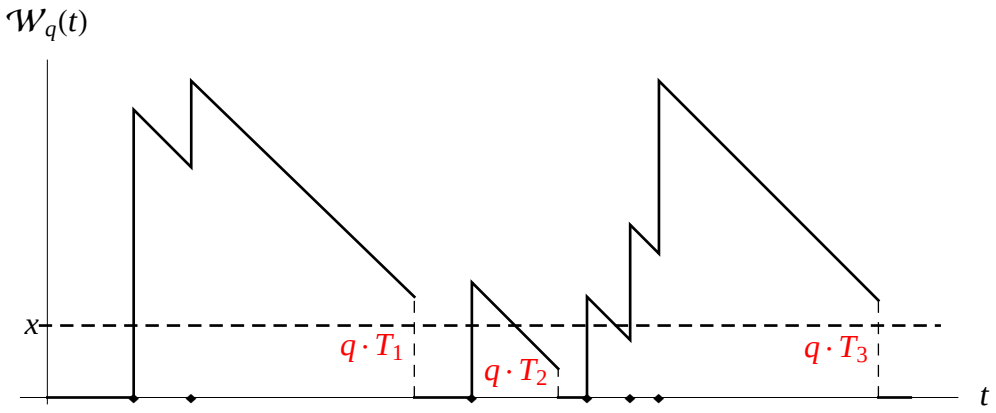


Figure 2.4: Sample path up- and down-crossings of level x for $\{\mathcal{W}_q(t), t \geq 0\}$

2.4.2 Steady-state integral equation for the pdf of the virtual wait

We characterize the transient distribution of the censored virtual wait by the functions

$$\left.\begin{array}{l} F_t(x) = \mathbb{P}(\mathcal{W}_q(t) \leq x), \quad x \geq 0,\ t \geq 0; \\ f_t(x) = \frac{\partial}{\partial x} F_t(x), \quad x > 0,\ t \geq 0; \\ P_0(t) = \mathbb{P}(\mathcal{W}_q(t) = 0), \quad t \geq 0. \end{array}\right\} \tag{2.18}$$

The steady-state distribution is obtained by letting $t \to \infty$ in the functions of Eq. (2.18), resulting in

$$F(x) = \lim_{t\to\infty} F_t(x), \quad f(x) = \lim_{t\to\infty} f_t(x), \quad \text{and} \quad P_0 = \lim_{t\to\infty} P_0(t). \tag{2.19}$$

When appropriate, we will use $f(x;q)$ equivalently as $f(x)$ to specify the value of q being used in the blocking policy. Also, in what follows, we extend the definition of $P_0(t)$ by defining $P_0(t) = 0$ for all $t < 0$.

Considering the censored virtual wait process, let $\mathcal{U}_t(x)$ and $\mathcal{D}_t(x)$ denote the number of sample path up- and down-crossings of level x, respectively, during the time interval $(0,t)$. Moreover, let $\mathcal{D}_t^c(x)$ (and $\mathcal{D}_t^j(x)$) denote the number of continuous down-crossings (jump down-crossings) of level x in the time interval $(0,t)$. Clearly,

$$\mathcal{D}_t(x) = \mathcal{D}_t^c(x) + \mathcal{D}_t^j(x). \tag{2.20}$$

Correspondingly, we remark that $\mathcal{U}_t^j(x) = \mathcal{U}_t(x)$ for all $x \geq 0$. The ingenuity of the level-crossing methodology lies in the principle of set balance (e.g., see Brill (2008, Section 2.4.6)). That is, in steady-state, the up-crossing and down-crossing rates of level x are equal:

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t(x))}{t} = \lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{U}_t(x))}{t}, \tag{2.21}$$

$$\lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t} \overset{a.s.}{=} \lim_{t\to\infty} \frac{\mathcal{U}_t(x)}{t}, \tag{2.22}$$

where "$a.s.$" means almost surely, or with probability 1. Thus, to develop an integral equation for the steady-state pdf of the virtual wait (provided it exists), we must establish both the up- and down-crossing rates of level x. The next theorem provides the means to do so.

Theorem 2.5 *The up- and down-crossing rates of level x are given by*

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{U}_t(x))}{t} = \lambda \bar{B}(x) P_0 + \lambda \int_{y=0}^{x} \bar{B}(x-y) f(y)\,\mathrm{d}y, \quad x > 0, \tag{2.23}$$

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t^c(x))}{t} = f(x), \quad x > 0, \tag{2.24}$$

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t^j(x))}{t} = \lambda P_0 \bar{G}_q(x), \quad x > 0. \tag{2.25}$$

Proof. The proof for both the up-crossing rate and the continuous down-crossing rate (i.e., Eq. (2.23) and Eq. (2.24)) can be derived in the exact same manner as for the classical $M/G/1$ virtual wait process (e.g., see Brill (2008, Theorems 3.3 and 3.4)). Thus, we omit their proofs and only prove Eq. (2.25).

To establish Eq. (2.25), we consider $\mathbb{E}(\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x))$ for very small h. Clearly, $\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x)$ represents the number of jump down-crossings of level x in a small interval of size h. Thus, $\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x)$ can take values in the set of non-negative integers. Concerning the expectation of this quantity, we can obviously omit the case of it being equal to 0. In addition, it is not difficult to see that $\mathbb{P}(\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x) \geq 2) = o(h)$.

Therefore, the only event we must really consider is when $\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x) = 1$. This event implies that a busy period initiates before time t, and also that sometime within the time interval $(t, t+h)$, the server finishes processing all but the last q-th proportion of the workload of this busy period (assume again that the system is empty at time 0). Conditioning on the length of this busy period leads to

$$\mathbb{P}(\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x) = 1) = \int_{y=x/q}^{\infty} \lambda h P_0(t - (1-q)y)\,\mathrm{d}G(y) + o(h). \tag{2.26}$$

The above result is obtained by recalling that the sample path immediately jumps down to level 0 as soon as the censored virtual wait process hits level qy. In particular, a jump down-crossing of level x will occur only if the busy period duration y is such that $qy > x$. Thus,

$$\mathbb{E}(\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x)) = \int_{y=x/q}^{\infty} \lambda h P_0(t - (1-q)y)\,\mathrm{d}G(y) + o(h). \tag{2.27}$$

Dividing the above equality by h and letting $h \to 0$, we subsequently obtain

$$\frac{\partial}{\partial t}\mathbb{E}(\mathcal{D}_t^j(x)) = \lambda \int_{y=x/q}^{\infty} P_0(t-(1-q)y)\,\mathrm{d}G(y). \tag{2.28}$$

It then follows (since $\mathbb{E}(\mathcal{D}_0^j(x)) = 0$) that

$$\mathbb{E}(\mathcal{D}_t^j(x)) = \lambda \int_{s=0}^{t} \int_{y=x/q}^{\infty} P_0(s-(1-q)y)\,\mathrm{d}G(y)\mathrm{d}s. \tag{2.29}$$

Finally, Eq. (2.25) follows since $\lim_{s\to\infty} \int_{y=x/q}^{\infty} P_0(s-(1-q)y)\,\mathrm{d}G(y) = P_0\bar{G}_q(x)$ via the dominated convergence theorem (e.g., see Parzen (1962, Section 6-10)). □

Corollary 2.6 *If* $\rho^{(q)} < 1$, *then*

$$\lim_{t\to\infty} \frac{\mathcal{D}_t^c(x)}{t} \overset{a.s.}{=} f(x), \quad x \geq 0 \quad \textit{and} \quad \lim_{t\to\infty} \frac{\mathcal{D}_t^j(x)}{t} \overset{a.s.}{=} \lambda P_0 \bar{G}_q(x), \quad x \geq 0. \tag{2.30}$$

Proof. By the memoryless property of Poisson arrivals, both $\{\mathcal{D}_t^j(x), t \geq 0\}$ and $\{\mathcal{D}_t^c(x), t \geq 0\}$ are (delayed) renewal processes. The desired result then follows from a well-known limiting theorem from renewal theory (e.g., see Parzen (1962, Section 5-3, Theorem 3A)). □

From Theorem 2.5, we can obtain an integral equation for the steady-state pdf of the virtual wait (provided it exists). Specifically, by using Eq. (2.23) through Eq. (2.25) along with the balance rate equation given by Eq. (2.21), we end up with

$$f(x) + \lambda P_0 \bar{G}_q(x) = \lambda \bar{B}(x) P_0 + \lambda \int_{y=0}^{x} \bar{B}(x-y) f(y)\mathrm{d}y. \tag{2.31}$$

Remark 2.7 *An attractive feature of the level-crossing technique is that we are able to intuitively explain each of the individual algebraic components of the resulting integral equation, which is indeed a renewal-type equation (e.g., see Kao (1996, Section 3.2)). We note that Eq.* (2.31) *is almost identical to the integral equation corresponding to the classical* $M/G/1$ *virtual wait, with the only addition being the second term on the left-hand side of the equality sign. This term (the jump down-crossing rate of level* x*) can be explained as follows: the rate that a busy period initiates is* λP_0, *where*

the proportion of these busy periods that result in a jump down-crossing of level x is $\bar{G}_q(x) = \mathbb{P}(qT > x)$. The other terms are interpreted in the same manner as for the classical $M/G/1$ virtual wait.

Remark 2.8 *Letting $x \to 0$ in Eq. (2.31) results in $f(0^+) = 0$ where, in general, $f(z^+) = \lim_{\epsilon \to 0} f(z + \epsilon)$. This result is as expected since $f(x)$ represents the continuous down-crossing rate of level x, and under the q-policy, any sample path of $\{\mathcal{W}_q(t), t \geq 0\}$ never down-crosses level 0 continuously — it always jumps down to level 0.*

To find P_0, we use the normalizing condition $\int_0^\infty f(x)\,\mathrm{d}x + P_0 = 1$. Now,

$$\int_0^\infty f(x)\,\mathrm{d}x = \lambda P_0(\mu - \mathbb{E}(qT)) + \lambda \int_{y=0}^\infty \int_{x=y}^\infty \bar{B}(x-y) f(y)\,\mathrm{d}x\,\mathrm{d}y, \tag{2.32}$$

which implies that $\int_0^\infty f(x)\mathrm{d}x(1 - \lambda\mu) = \lambda P_0(\mu - q\mathbb{E}(T))$. Using Eq. (2.8), we get

$$\begin{aligned}\int_0^\infty f(x)\,\mathrm{d}x &= P_0 \frac{\rho(1 - \rho^{(q)} - q)}{(1-\rho)(1-\rho^{(q)})} \\ &= P_0 \frac{\rho(1-q)(1-\rho)}{(1-\rho)(1-\rho^{(q)})} \\ &= P_0 \frac{\rho^{(q)}}{1-\rho^{(q)}}. \end{aligned} \tag{2.33}$$

Therefore, $P_0 = 1 - \rho^{(q)}$. This result too is as expected, since P_0 represents the long-run proportion of time that the server is idle conditional on the system being open for arrivals (i.e., conditional on the existence of the censored virtual wait process). From Eq. (2.14) and Eq. (2.17), the long-run fraction of time the system accepts new customers is $P_I + P_{B,1} = (1 + \rho q)^{-1}$. Thus, $P_0 = P_I/(1 + \rho q)^{-1}$.

From Eq. (2.31), we can readily obtain the LST of the steady-state actual wait of serviceable customers.

Theorem 2.9 *The LST of W, the steady-state waiting time of serviceable customers, is*

$$\widetilde{W}(s) \equiv \mathbb{E}(e^{-sW}) = \frac{(1-\rho^{(q)})(s - \lambda + \lambda\widetilde{G}(qs))}{s - \lambda + \lambda\widetilde{B}(s)}. \tag{2.34}$$

Proof. Clearly, $\widetilde{W}(s) = \int_0^\infty e^{-sx}\mathrm{d}F(x) = P_0 + \int_0^\infty e^{-sx}f(x)\mathrm{d}x$. Thus, the desired result is readily obtained by first multiplying both sides of Eq. (2.31) by e^{-sx} and then integrating over $x \in (0, \infty)$. □

Alternatively, we can express the above LST as

$$\widetilde{W}(s) = (1 - \rho^{(q)}) + \rho^{(q)}\widetilde{W}_+(s), \tag{2.35}$$

where W_+ represents the stationary waiting time for those customers who are admitted for service upon their arrival but incur a positive wait time prior to entering service. We refer to W_+ as the *delayed* waiting time whose LST $\widetilde{W}_+(s)$ is given by

$$\widetilde{W}_+(s) = \frac{(1 - \rho^{(q)})(\widetilde{G}(qs) - \widetilde{B}(s))}{\mu(1 - q)(s - \lambda + \lambda\widetilde{B}(s))}. \tag{2.36}$$

One can obtain the first moment of waiting time by differentiating $\widetilde{W}(s)$ and twice applying L'Hôpital's rule. After some algebra, we acquire the following illuminating form of the mean waiting time:

$$\mathbb{E}(W) = \frac{\lambda^{(q)}\gamma}{2(1 - \rho^{(q)})} \times (1 + \sigma(q)), \tag{2.37}$$

where $\sigma(q) = q/(1 - \rho^{(q)})$. We observe that the first term of Eq. (2.37) is equal to the average waiting time in the classical $M/G/1$ queue with arrival rate $\lambda^{(q)}$ and service time distribution $B(\cdot)$. Clearly, $\sigma(q) \geq 0$ since $0 \leq q \leq 1$, which implies that a system under the q-policy has a greater average waiting time than a classical $M/G/1$ queue with the aforementioned parameters.

In addition, the first moment of waiting time can be re-written as

$$\mathbb{E}(W) = \frac{\lambda\gamma}{2} \times \kappa(q), \quad 0 \leq q \leq 1, \tag{2.38}$$

where

$$\kappa(q) = \frac{1 - q}{1 - \rho^{(q)}}(1 + \sigma(q)) = \frac{(1 - q)(1 - \rho^{(q)} + q)}{(1 - \rho^{(q)})^2}. \tag{2.39}$$

Differentiating $\kappa(q)$ with respect to q yields

$$\kappa'(q) = -\frac{2q}{(1 - \rho^{(q)})^3}. \tag{2.40}$$

Therefore, for $\rho^{(q)} < 1$, $\mathbb{E}(W)$ is a decreasing function of q. Considering $\mathbb{E}(W)$ at the extreme values of q, we see that for $q = 0$, $\mathbb{E}(W) = \lambda\gamma(1-\rho)^{-1}/2$ which is the classical $M/G/1$ average waiting time without a blocking policy, and for $q = 1$, $\kappa(1) = 0$ so that $\mathbb{E}(W) = 0$. The latter result is due to the fact that during busy periods, the system is closed to all potential arrivals, and above that, only customers who arrive to an idle server will be served (and these customers experience zero wait).

Finally, we close this analysis by considering the first moment of delayed waiting time, namely:

$$\mathbb{E}(W_+) = \frac{\mathbb{E}(W)}{\rho^{(q)}} = \frac{\gamma}{2\mu} \times \frac{1-\rho^{(q)}+q}{(1-\rho^{(q)})^2}, \quad 0 \leq q \leq 1. \tag{2.41}$$

It is indeed true that for $q = 1$, there is zero probability that an arbitrary customer will experience positive wait; however, as $q \to 1$, we see that $\mathbb{E}(W_+)$ becomes

$$\mathbb{E}(W_+)\big|_{q=1} = \frac{\gamma}{\mu}. \tag{2.42}$$

We recognize Eq. (2.42) as the mean of the limiting *total-life* random variable of a renewal process with $B(\cdot)$ serving as the inter-arrival time df (e.g., see Kao (1996, Section 3.3)).

2.4.3 $M/G/1$ queue under a q-policy with closedown periods

We now consider a slight variant of the $M/G/1$ queue operating under the q-policy. Specifically, we incorporate a closedown period, S, after each busy period. It is assumed that the sequence of successive closedown periods are iid with df $A(x) = \mathbb{P}(S \leq x)$. The facility is closed to all potential arrivals during a closedown period. Thus, the incorporation of a closedown period will increase the proportion of customers that are blocked from the system. In addition, it is obvious that the closedown periods do not affect the waiting time distributions for serviceable customers, and so our analysis of waiting time in the previous subsections is still applicable.

We view the total idle period as the durations of time when the server is not busy. Hence, similar to the partitioning of the steady-state probability of the system

being busy, we define the following:

$P_{I,0} \equiv$ steady-state probability the server is idle and the system is closed;

$P_{I,1} \equiv$ steady-state probability the server is idle and the system is open.

In this variation, the busy cycle remains $D = T + I$ (note though that the closedown period is contained in I). Again, applying elementary renewal theory arguments, we obtain:

$$P_I = \frac{\mathbb{E}(I)}{\mathbb{E}(D)} = \frac{(1-\rho^{(q)})(1+\lambda\mathbb{E}(S))}{(1+\lambda\mathbb{E}(S))(1+\rho q) - \lambda\mathbb{E}(S)\rho}, \tag{2.43}$$

$$P_{I,0} = \frac{\mathbb{E}(S)}{\mathbb{E}(I)} P_I = \frac{(1-\rho^{(q)})\lambda\mathbb{E}(S)}{(1+\lambda\mathbb{E}(S))(1+\rho q) - \lambda\mathbb{E}(S)\rho}, \tag{2.44}$$

$$P_{I,1} = \frac{\lambda^{-1}}{\mathbb{E}(I)} P_I = \frac{1-\rho^{(q)}}{(1+\lambda\mathbb{E}(S))(1+\rho q) - \lambda\mathbb{E}(S)\rho}, \tag{2.45}$$

$$P_B = \frac{\mathbb{E}(T)}{\mathbb{E}(D)} = \frac{\rho}{(1+\lambda\mathbb{E}(S))(1+\rho q) - \lambda\mathbb{E}(S)\rho}, \tag{2.46}$$

$$P_{B,0} = qP_B = \frac{\rho q}{(1+\lambda\mathbb{E}(S))(1+\rho q) - \lambda\mathbb{E}(S)\rho}, \tag{2.47}$$

$$P_{B,1} = (1-q)P_B = \frac{\rho^{(q)}}{(1+\lambda\mathbb{E}(S))(1+\rho q) - \lambda\mathbb{E}(S)\rho}. \tag{2.48}$$

Thus, the long-run fraction of time the system is accepting of new customers is $P_{I,1} + P_{B,1} = [(1+\lambda\mathbb{E}(S))(1+\rho q) - \lambda\mathbb{E}(S)\rho]^{-1}$.

2.4.4 Zero-wait customers having exceptional service

In this subsection, we consider yet another variant of the $M/G/1$ queue operating under a q-policy by assuming that the service time distribution of those customers who arrive to an idle system is given by $V(\cdot)$, possibly differing from $B(\cdot)$ (i.e., the service time distribution of customers arriving to the system during busy periods). In other words, in this queueing system, zero-wait customers have exceptional service times. In what follows, we define the random variable V whose df and LST are given by $V(x)$ and $\widetilde{V}(s)$, respectively.

We begin our analysis of the current system with the distribution of its busy periods. In particular, if we define the random variable T_d (whose df and LST we denote by $G_d(\cdot)$ and $\widetilde{G}_d(s)$, respectively) as the duration of a busy period in the current system, then it easily follows from similar arguments to those made in the proof of Theorem 2.3 that

$$\widetilde{G}_d(s) = \widetilde{V}(s + \lambda^{(q)}(1 - \widetilde{G}(s))), \tag{2.49}$$

where $\widetilde{G}(s)$ satisfies the functional equation given by Eq. (2.4). Furthermore, the first two moments of T_d are simply given by

$$\mathbb{E}(T_d) = \frac{\mathbb{E}(V)}{1 - \rho^{(q)}} \tag{2.50}$$

and

$$\mathbb{E}(T_d^2) = \frac{\lambda^{(q)}\gamma}{(1 - \rho^{(q)})^3}\mathbb{E}(V) + \frac{\mathbb{E}(V^2)}{(1 - \rho^{(q)})^2}. \tag{2.51}$$

The steady-state probabilities of the current system are obtained by first realizing that the busy cycle of this system is now given by $D = I + T_d$, and then subsequently applying the same renewal theory arguments as before. The result of this leads to

$$P_I = \frac{\mathbb{E}(I)}{\mathbb{E}(D)} = \frac{1 - \rho^{(q)}}{1 - \rho^{(q)} + \lambda\mathbb{E}(V)}, \tag{2.52}$$

$$P_B = \frac{\mathbb{E}(T_d)}{\mathbb{E}(D)} = \frac{\lambda\mathbb{E}(V)}{1 - \rho^{(q)} + \lambda\mathbb{E}(V)}, \tag{2.53}$$

$$P_{B,0} = qP_B = \frac{\lambda\mathbb{E}(V)q}{1 - \rho^{(q)} + \lambda\mathbb{E}(V)}, \tag{2.54}$$

$$P_{B,1} = (1 - q)P_B = \frac{\lambda^{(q)}\mathbb{E}(V)}{1 - \rho^{(q)} + \lambda\mathbb{E}(V)}. \tag{2.55}$$

Shifting our focus now to the wait of serviceable customers, we remark that an integral equation for the pdf of virtual wait can be obtained via similar level-crossing techniques as to those used in Section 2.4.2. Specifically, if we let $G_{d,q}(x) = 1 - \bar{G}_{d,q}(x) = \mathbb{P}(qT_d \leq x) = G_d(x/q)$, then an integral equation for the pdf of the virtual wait $f(x)$ is given by

$$f(x) + \lambda P_0 \bar{G}_{d,q}(x) = \lambda P_0 \bar{V}(x) + \lambda \int_{y=0}^{x} \bar{B}(x - y)f(y)\mathrm{d}y. \tag{2.56}$$

It is obvious that Eq. (2.56) is equivalent to Eq. (2.31) if $V(x) = B(x)$. Furthermore, it follows from the normalizing condition $\int_0^\infty f(x)\,\mathrm{d}x + P_0 = 1$ that

$$P_0 = \frac{1-\rho^{(q)}}{1-\rho^{(q)}+\lambda^{(q)}\mathbb{E}(V)}. \tag{2.57}$$

Adding $\int_0^\infty e^{-sx} f(x)\,\mathrm{d}x$ to the previous expression for P_0 ultimately yields the following expression for the LST of the steady-state waiting time of serviceable customers:

$$\widetilde{W}(s) = \left(\frac{1-\rho^{(q)}}{1-\rho^{(q)}+\lambda^{(q)}\mathbb{E}(V)}\right) \times \left[\frac{s-\lambda(1-\widetilde{B}(s))+\lambda(\widetilde{G}_d(qs)-\widetilde{V}(s))}{s-\lambda+\lambda\widetilde{B}(s)}\right]. \tag{2.58}$$

Similarly, the delayed waiting time LST is given by

$$\widetilde{W}_+(s) = \frac{(1-\rho^{(q)})(\widetilde{G}_d(qs)-\widetilde{V}(s))}{\mathbb{E}(V)(1-q)(s-\lambda+\lambda\widetilde{B}(s))}. \tag{2.59}$$

The first moment of W can be obtained from Eq. (2.56) and from the fact that $\mathbb{E}(W) = \int_0^\infty x f(x)\,\mathrm{d}x$. Hence,

$$\mathbb{E}(W) = \frac{\lambda P_0(\mathbb{E}(V^2)-q^2\mathbb{E}(T_0^2)-\gamma)}{2(1-\rho)} + \frac{\lambda\gamma}{2(1-\rho)}. \tag{2.60}$$

Finally, a simple expression for the first moment of W_+ can be obtained from the relation $\mathbb{E}(W) = (1-P_0)\mathbb{E}(W_+)$. After some straightforward but tedious algebra, we obtain

$$\mathbb{E}(W_+) = \frac{\mathbb{E}(V^2)}{2\mathbb{E}(V)} \times \left(1+\frac{q}{1-\rho^{(q)}}\right) + \frac{\lambda\gamma}{2(1-\rho)} \times \left(1-\frac{q^2}{(1-\rho^{(q)})^2}\right). \tag{2.61}$$

2.5 $M/G/1$ queue with accumulating priority

In order to implement the q-policy, a systems manager must know the service times of the customers upon their arrival to the system. However, such knowledge may not always be available. In this section, we introduce another $M/G/1$-type queueing model which enables a systems manager to reduce the length of busy periods, in a similar fashion as the q-policy, without the knowledge of service times upon

arrival. In addition to maintaining the reduction in the busy period lengths, this blocking mechanism also results in the same waiting time distribution for serviceable customers as the q-policy; however, it also results in some waiting (or holding) times experienced by the unserviceable customers. Nevertheless, the same main benefits of the q-policy are captured. We remark that this system is a variant of the $M/G/1$ queue with accumulating priority, which was recently studied by Stanford et al. (2014).

The first key aspect of the $M/G/1$ queue with accumulating priority has to do with how priority is accumulated for customers. Specifically, customers arrive to the system with zero initial priority, and throughout their sojourn in the system, earn priority linearly at rate $\xi_1 > 0$. At service completion epochs, the customer with the greatest accumulated priority is serviced next. The second key feature of this model lies in the concept of an accreditation threshold, which increases linearly at rate ξ_2 where $0 \leq \xi_2 \leq \xi_1$. In fact, the accreditation threshold is a stochastic process which we denote as $\{\Theta(t), t \geq 0\}$. It is important to note that the accreditation threshold and its implementation does not, in any way, affect the order of service for customers. Hence, the way in which the $M/G/1$ queue with accumulating priority operates is actually equivalent to the classical $M/G/1$ queue under the FCFS discipline. However, the incorporation of the accreditation threshold does shed new light on the structuralization of the general busy period, providing a useful classification of those customers who arrive during busy periods.

The above basic model was introduced by Stanford et al. (2014) in their analysis of the non-preemptive accumulating priority queue. In order to analyze the $M/G/1$ queue with accumulating priority, these authors introduced something known as the maximal priority process. To incorporate a blocking policy into this system, we require a slight modification to their definition of the maximal priority process. Following that, we establish the connection between our modified maximal priority process and the censored virtual wait process of the previous section. We exploit this connection to obtain the steady-state integral equation of the *accumulated priority* of serviceable customers.

2.5.1 The maximal priority process

Upon arrival to the system, customers begin to accumulate priority at a linear rate. During busy periods, a customer will be admitted for service only if its priority overtakes (i.e., becomes greater than) the accreditation threshold, governed by $\{\Theta(t), t \geq 0\}$. At a service completion instant, if there are any admitted customers present in the system, the one with the greatest accumulated priority is selected next for service. The busy period ends at a service completion instant which leaves no more admitted customers in the system. Note that the busy period may end while there are still customers present in the system. In this situation, these customers depart the system without ever entering into service.

Let τ_k denote the arrival epoch of customer C_k, so that we may define $\Phi_k(t)$ to be this customer's priority function (i.e., the amount of accumulated priority C_k has at time t), namely:

$$\Phi_k(t) = \xi_1 \cdot (t - \tau_k), \quad t > \tau_k. \tag{2.62}$$

Furthermore, let $n(k)$ denote the arrival position of the k-th customer to be serviced. The definition of the maximal priority process now follows.

Definition 2.10 *The maximal priority process is a two-dimensional stochastic process $\mathcal{M}(t) = \{(M(t), \Theta(t)), t \geq 0\}$, satisfying the following conditions:*

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

2. *For all t not corresponding to service commencement/completion instants, we have*

$$\frac{\mathrm{d}M(t)}{\mathrm{d}t} = \xi_1 \quad \textit{and} \quad \frac{\mathrm{d}\Theta(t)}{\mathrm{d}t} = \xi_2, \tag{2.63}$$

   where $0 \leq \xi_2 \leq \xi_1$.

3. *At the sequence of service completion times $\{\delta_k\}_{k=1}^{\infty}$,*

$$M(\delta_k) = 1\big\{\Phi_\vee(\delta_k^-) > \Theta(\delta_k^-)\big\} \cdot \Phi_\vee(\delta_k^-), \tag{2.64}$$

$$\Theta(\delta_k^+) = \min\{M(\delta_k), \Theta(\delta_k^-)\}, \tag{2.65}$$

where

$$\Phi_{\vee}(\delta_k^-) = \max_{m\in\{n(k)+1,n(k)+2,\ldots\}} \Phi_m(\delta_k^-) \tag{2.66}$$

and $1\{A\}$ *is the indicator function of the event* A*.*

The above definition shows that $\{M(t), t \geq 0\}$ is closely related to the well-known age process (i.e., when $\xi_1 = 1$, $M(t)$ represents the age of the oldest admitted customer at time t). Furthermore, the accreditation threshold process increases linearly at rate ξ_2 during busy periods. Stanford et al. (2014) referred to those customers who arrive during busy periods and whose priority overtakes the accreditation threshold as *accredited customers*.

With this definition in place, we can now introduce the blocking scheme for our modified $M/G/1$ queue with accumulating priority. In particular, serviceable customers consist of accredited customers and customers who arrive during idle times. On the other hand, those customers whose priority fails to overtake the accreditation threshold during a busy period are blocked, thereby departing the system without ever entering into service. We refer to such customers as non-accredited customers.

Figure 2.5 depicts a typical sample path of $\{\mathcal{M}(t), t \geq 0\}$. Note that customers C_4, C_5, and C_9 are of the non-accredited type and thus end up being blocked from service. Moreover, a notable difference between the current model and the one considered in the previous sections is that with the current system, blocked customers experience some wait before being forced to depart the system.

Suppose now at the end of an arbitrary busy period, we wish to find the latest time by which a customer would have to arrive in order to be admitted for service. This can be done by simply dividing the height of the accreditation threshold at time t_* (i.e., the time at which the busy period completes) by ξ_1 and subsequently subtracting this quantity from t_*. For a sample path such as the one shown in Figure 2.5, this is equivalent to determining the t-intercept of a line with slope ξ_1 which crosses the point $(t_*, \Theta(t_*^-))$.

For each busy period, we define the accreditation interval as the duration of time within which customers must arrive in order to be admitted for service. An important observation is that the ratio of the accreditation interval to the busy period is always

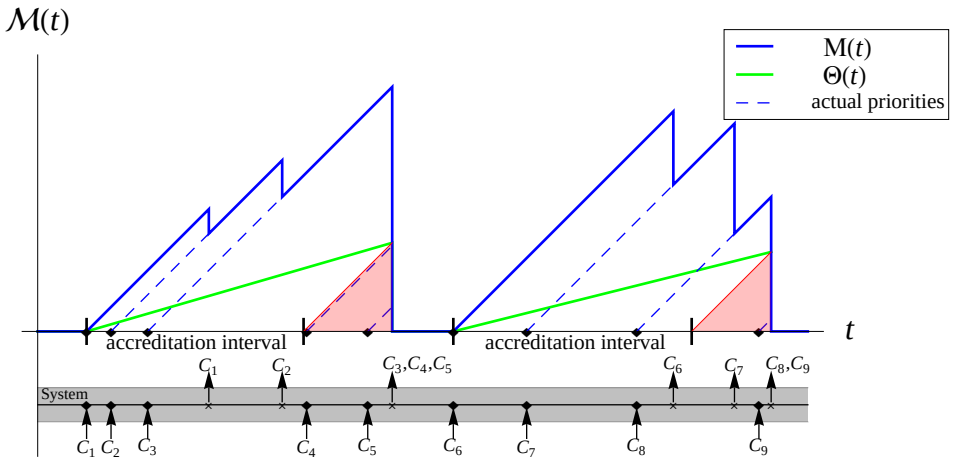


Figure 2.5: A typical sample path of $\{\mathcal{M}(t), t \geq 0\}$

$(1-\xi_2/\xi_1)$. Therefore, this model is similar to the one of Section 2.2 in that admitted customers must arrive within the first $(1-q)$-th proportion of the busy period with $q = \xi_2/\xi_1$. In fact, it can be shown that the LST of the busy period is the solution to Eq. (2.4) with $q = \xi_2/\xi_1$ (see Stanford et al. (2014) and their discussion on accredited busy periods). In addition, using the same argument as in Brill (1988), we can show that the steady-state distribution of $\{M(t), t \geq 0\}$ when $\xi_1 = 1$ is equivalent to the steady-state distribution of the workload process $\{U_{\xi_2}(t), t \geq 0\}$ of Section 2.4.1.

2.5.2 The distribution of accumulated priority for accredited customers

Recall that serviceable customers represent those customers who either arrive to the system during idle periods or arrive to the system during busy periods and become accredited. It is obvious that customers that arrive to the system during idle periods experience zero wait, and thus have no accumulated priority immediately before entering into service. On the other hand, accredited customers do experience positive wait, and hence will have accumulated a positive amount of priority immediately prior to entering into service. If we let $\mathcal{P}^{(acc)}$ be the accumulated priority of an

arbitrary accredited customer, then it must be that

$$\mathcal{P}^{(acc)} = \xi_1 \times W_+, \tag{2.67}$$

where W_+ is the wait of accredited customers.

It is straightforward to understand that the waiting time distribution of serviceable customers under the current blocking mechanism is equivalent to that of the serviceable customers under a q-policy with $q = \xi_2/\xi_1$. Moreover, the waiting time distribution of accredited customers W_+ exactly follows the same distribution as the delayed waiting time random variable introduced in Section 2.4.2. Therefore, it readily follows from Eq. (2.36) that the LST of $\mathcal{P}^{(acc)}$ is given by

$$\widetilde{\mathcal{P}}^{(acc)}(s) = \widetilde{W}_+(\xi_1 s) = \frac{(1-\rho^{(\xi_2/\xi_1)})(\widetilde{G}(\xi_2 s) - \widetilde{B}(\xi_1 s))}{\mu(1-\xi_2/\xi_1)(\xi_1 s - \lambda + \lambda\widetilde{B}(\xi_1 s))}. \tag{2.68}$$

Similarly, in the case of zero-wait customers having exceptional service, Eq. (2.59) leads to

$$\widetilde{\mathcal{P}}^{(acc)}(s;V) \equiv \widetilde{\mathcal{P}}^{(acc)}(s) = \frac{(1-\rho^{(\xi_2/\xi_1)})(\widetilde{G}_d(\xi_2 s) - \widetilde{V}(\xi_1 s))}{\mathbb{E}(V)(1-\xi_2/\xi_1)(\xi_1 s - \lambda + \lambda\widetilde{B}(\xi_1 s))}. \tag{2.69}$$

Clearly, if $\widetilde{V}(s) = \widetilde{B}(s)$, then Eq. (2.69) becomes identical to Eq. (2.68). Note that the notation $\widetilde{\mathcal{P}}^{(acc)}(s;V)$ above symbolizes the LST of accumulated priority of an arbitrary accredited customer serviced during a delay busy period with an initial delay of V.

We remark that both Eqs. (2.68) and (2.69) were first presented by Stanford et al. (2014). However, their result was obtained under a different setting, as they studied a particular multi-class non-preemptive priority queueing system and obtained the steady-state marginal waiting time distributions of each class. We emphasize that in their model, there is no concept of customer blocking. The authors obtained their result for a random variable which they called the *additional accumulated priority*. We direct readers to their paper for more details. Moreover, the authors' method of analysis differs from ours in that their proofs of Eqs. (2.68) and (2.69) are inspired by the Conway et al. (1967, Chapter 8-4) derivation of the flow time LST in a classical FCFS $M/G/1$ system.

In summary, our level-crossing analysis provides an alternate proof of Stanford et al.'s (2014) main results (i.e., Eqs. (2.68) and (2.69)) and also yields the steady-state integral equation for the pdf of $\mathcal{P}^{(acc)}$. In particular, if we let $g_{\xi_1}(x)$ denote the steady-state pdf of $\mathcal{P}^{(acc)}$, then

$$g_{\xi_1}(x) = \frac{f(x/\xi_1; \xi_2/\xi_1)}{\xi_1}, \quad x > 0, \tag{2.70}$$

where $f(x; q)$ was defined in Section 2.4.2 (i.e., the steady-state pdf of virtual wait of serviceable customers in a q-policy). Therefore, from Eq. (2.31), we ultimately get

$$g_{\xi_1}(x) = \frac{\lambda \bar{B}(x/\xi_1)P_0 - \lambda P_0 \bar{G}_{\xi_2/\xi_1}(x/\xi_1)}{\xi_1} + \frac{\lambda}{\xi_1} \int_{y=0}^{x} \bar{B}((x-y)/\xi_1) g_{\xi_1}(y) \mathrm{d}y. \tag{2.71}$$

Remark 2.11 *The integral equation of $g_{\xi_1}(x)$ for the case of zero-wait customers having exceptional service can be similarly obtained from Eq.* (2.56)*.*

2.5.3 The overall distribution of wait

We next establish the distribution of the overall waiting time random variable. First of all, let W_0 and W_1 represent the waiting times of unserviceable and serviceable customers, respectively. Clearly, by design of the model, customers who are blocked from service will experience a (steady-state) waiting time (or total time in the system) which follows the limiting distribution of the forward recurrence time of qT. Hence, it must be that

$$\widetilde{W}_0(s) = \frac{1 - \widetilde{G}_{\xi_2/\xi_1}(s)}{\mathbb{E}(T) s \xi_2/\xi_1}. \tag{2.72}$$

Now, since the wait of serviceable customers under the current blocking mechanism is equivalent to the wait of serviceable customers under a q-policy with $q = \xi_2/\xi_1$, it immediately follows from Eq. (2.34) that

$$\widetilde{W}_1(s) = \frac{(1 - \rho^{(\xi_2/\xi_1)})(s - \lambda + \lambda \widetilde{G}(\xi_2 s/\xi_1))}{s - \lambda + \lambda \widetilde{B}(s)}. \tag{2.73}$$

Using the steady-state probabilities given by Eq. (2.14) through Eq. (2.17), we derive the overall LST of waiting time as

$$\widetilde{W}(s) = \frac{1}{1+\rho\xi_2/\xi_1}\widetilde{W}_1(s) + \frac{\rho\xi_2/\xi_1}{1+\rho\xi_2/\xi_1}\widetilde{W}_0(s). \tag{2.74}$$

After some elementary algebra, we obtain

$$\widetilde{W}(s) = \left(\frac{1-\rho^{(\xi_2/\xi_1)}}{1+\rho(\xi_2/\xi_1)}\right) \times \left(\frac{s-\lambda+\lambda\widetilde{G}(s\xi_2/\xi_1)}{s-\lambda+\lambda\widetilde{B}(s)} + \frac{\lambda(1-\widetilde{G}(s\xi_2/\xi_1))}{s}\right). \tag{2.75}$$

Remark 2.12 *Similar arguments combined with the results of Section 2.4.4 can be applied to obtain the overall distribution of wait for the case of zero-wait customers having exceptional service.*

2.6 Numerical examples

In this section, we formulate a numerical study to demonstrate a potential usage of the q-policy. We remark that the inspiration for this study originates from a similar study considered by Kao (1996, Example 3.6.4). In what follows, we consider a queueing system with closedown periods as described in Section 2.4.3. For this system, suppose we have the following monetary parameters:

$K \equiv$ the cost of each closedown period;
$h \equiv$ the cost of holding one customer per unit time;
$R \equiv$ the toll fee paid by each serviced customer.

The objective function which we seek to optimize is the long-run expected profit per unit time. Clearly, the instants of busy period commencements define a set of regeneration points. Thus, our objective function is

$$P(q) = \frac{R\cdot\mathbb{E}(N_{bp}) - K - \mathbb{E}(C_{bp})}{\mathbb{E}(D)}, \tag{2.76}$$

where $\mathbb{E}(C_{bp})$ is the expected holding cost incurred during a busy period. We remark that $\mathbb{E}(N_{bp})$ is given by Eq. (2.13) and $\mathbb{E}(D) = \mathbb{E}(T) + \mathbb{E}(S) + \lambda^{-1}$. Moreover, it can be shown, following a similar line of reasoning to Kao (1996, pp. 139–140), that for

all work-conserving service disciplines (e.g., both FCFS and the q-restricted LCFS disciplines),

$$\mathbb{E}(C_{bp}) = h\mathbb{E}(N_{bp})(\mu + \mathbb{E}(W)). \tag{2.77}$$

Note that the quantity $\mu + \mathbb{E}(W)$ represents the long-run average flow time.

Recalling the form of $\mathbb{E}(W)$ in Eq. (2.37), it is immediately clear that $\mathbb{E}(C_{bp})$ depends only on the first two moments of the service time distribution. Consequently, the expected profit function $P(q)$ is also affected by the variability of the service time distribution. We use the coefficient of variation of the service time distribution, denoted by $CV = \sqrt{\gamma - \mu^2}/\mu$, to assess the effect of the variability of the service time distribution on the profit function. In particular, we present five numerical examples of nearly identical models, differing only in their respective coefficients of variation of the service time distribution. In Examples 1 through 5, we consider five service time distributions with common mean $\mu = 1$, but with coefficients of variation 0, 0.5, 1, 1.5, and 2, respectively. Furthermore, we set $\lambda = 0.95$, $\mathbb{E}(S) = 1$, $h = 1$, $K = 5$, and $R = 50$.

Figure 2.6 displays the profit functions corresponding to the five examples. With the exception of the profit functions for Examples 1 and 2, we observe that the expected profit per unit time can be maximized by implementing the q-policy. Letting q^* denote the optimal blocking proportion which maximizes $P(q)$, we find q^* (to 4 decimal places of accuracy) for Examples 1 through 5 to be 0, 0, 0.1000, 0.1710, and 0.2538, respectively. In Table 2.1, we calculate the expected profit function and several other quantities of interest corresponding to various values of the blocking proportion q for Examples 1 through 5. We note that since $\mu = 1$, Eq. (2.8) and Eq. (2.13) together imply that $\mathbb{E}(T) = \mathbb{E}(N_{bp})$ for all values of q.

Although it is indeed true that the maximum long-run expected profit per unit time is obtained without the usage of a q-policy (i.e., $q^* = 0$) for both Examples 1 and 2, there are other viable reasons for the implementation of a q-policy. In regard to Example 2, let us define q_r^* to be the relative maxima of $P(q)$. By using standard calculus-based methods, we find that $q_r^* = 0.0406$. From Table 2.1 (and the rows corresponding to Example 2), we see that the resulting expected profits with $q = q^* = 0$ and $q = q_r^*$ differ only by a small amount. However, the advantage of implementing

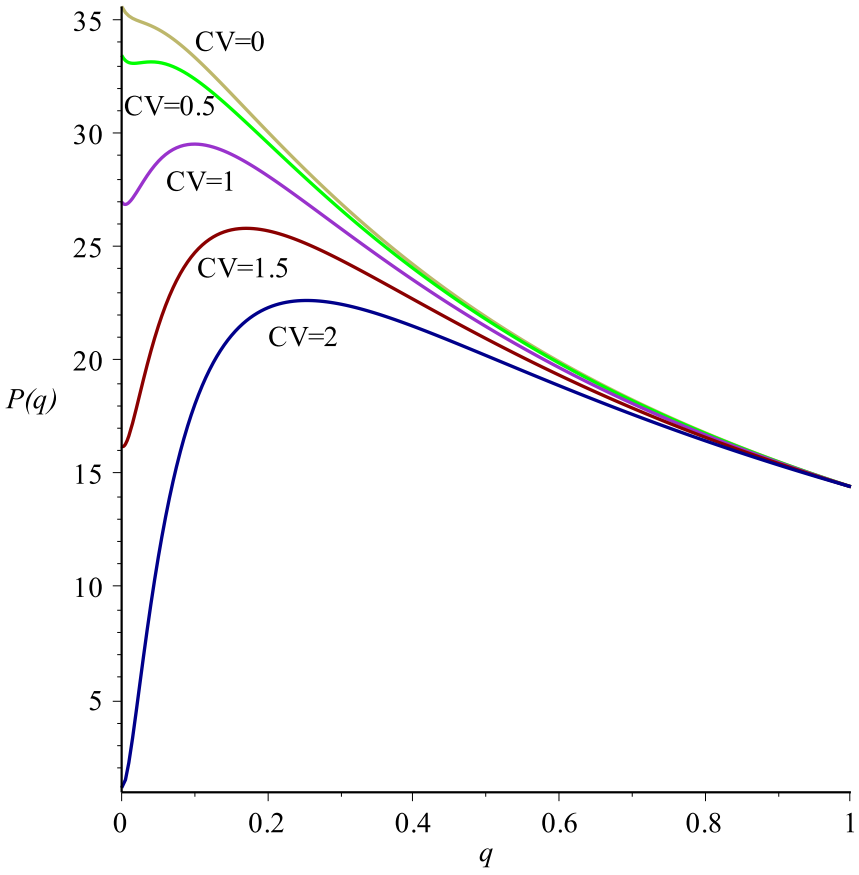


Figure 2.6: Expected profit per unit time for Examples 1 through 5

Table 2.1: Expected profit per unit time and other quantities of interest against various q-values for Examples 1 through 5

$M/G/1$ queue with $\lambda = 0.95$; $\mu = 1$; $\mathbb{E}(S) = 1$; $h = 1$; $K = 5$; $R = 50$								
Example 1: $CV = 0$; $q^* = 0.00$								
Quantity	———	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	———	35.5967	34.5886	33.3634	30.0792	24.2058	20.8748	16.1395
$\mathbb{E}(D)$	———	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	———	210.0000	82.0681	41.2522	16.2616	5.3008	3.0254	1.3591
$\mathbb{E}(N_{bp})$	———	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	———	9.5000	7.0016	4.9816	2.9028	1.2793	0.7321	0.1655
Example 2: $CV = 0.5$; $q^* = 0.00$; $q_r^* = 0.0406$								
Quantity	$q = q_r^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	33.1506	33.4427	33.1301	32.4037	29.5931	24.0359	20.7907	16.1245
$\mathbb{E}(D)$	13.3439	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	117.2050	257.5000	100.0211	49.8411	19.2853	6.0446	3.3451	1.4074
$\mathbb{E}(N_{bp})$	11.2912	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	9.3802	11.8750	8.7521	6.2270	3.6285	1.5992	0.9151	0.2068
Example 3: $CV = 1$; $q^* = 0.1000$								
Quantity	$q = q^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	29.5245	26.9809	28.7545	29.5245	28.1345	23.5263	20.5383	16.0795
$\mathbb{E}(D)$	8.9491	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	75.6071	400.0000	153.8799	75.6079	28.3565	8.276	4.3041	1.5521
$\mathbb{E}(N_{bp})$	6.8965	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	9.9631	19.0000	14.0033	9.9631	5.8056	2.5587	1.4641	0.3309
Example 4: $CV = 1.5$; $q^* = 0.1710$								
Quantity	$q = q^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	25.8100	16.2112	21.4619	24.7257	25.7036	22.6768	20.1176	16.0046
$\mathbb{E}(D)$	6.7606	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	55.9079	637.5000	243.6445	118.5524	43.4751	11.995	5.9025	1.7933
$\mathbb{E}(N_{bp})$	4.7080	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	10.8751	30.8750	22.7553	16.1901	9.4340	4.1579	2.3792	0.5377
Example 5: $CV = 2$; $q^* = 0.2538$								
Quantity	$q = q^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	22.6279	1.1337	11.2523	18.0075	22.3003	21.4876	19.5286	15.8997
$\mathbb{E}(D)$	5.4881	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	42.5880	970.0000	369.3151	178.6748	64.6412	17.2016	8.1402	2.1309
$\mathbb{E}(N_{bp})$	3.4354	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	11.3967	47.5000	35.0082	24.9078	14.5139	6.3967	3.6603	0.8273

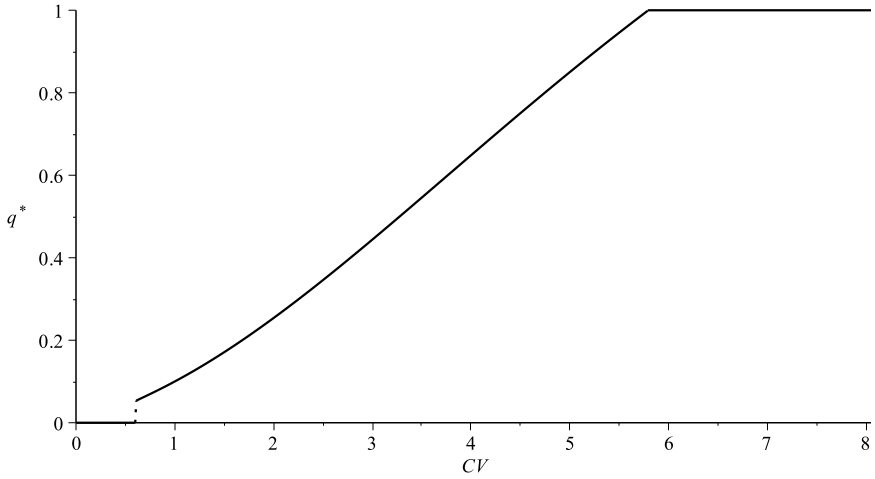


Figure 2.7: Behaviour of the optimal blocking proportion q^* as a function of CV

a q-policy still lies in the fact that both the cycle and busy period lengths are smaller when compared to the system without a q-policy in place. Ultimately, with $q = q_r^*$, the system is essentially earning the same expected profit as for the case with $q = 0$, but at the same time allowing for more frequent maintenance checks on the server/machine. Similar remarks can be made for Example 1.

In these numerical examples, we showed that by reducing the cycle lengths, a system manager can significantly decrease the incurred costs and thus capture the potential profit (or, as in both Examples 1 and 2, obtain nearly maximal expected profit). It is also apparent that as CV increases, so too does the optimal blocking proportion q^*, as evidenced in Figure 2.7. It is interesting to note the presence of a discontinuity point in Figure 2.7, which occurs for a certain value of CV residing in the interval $(0.6014, 0.6015)$. This particular value of CV corresponds to the first instance in which a non-zero blocking proportion yields a higher expected profit.

Chapter 3

The preemptive accumulating priority queue

3.1 Introduction

This chapter has to do with the analysis of a certain dynamic preemptive priority queueing system. In particular, we consider the preemptive variant of the model considered by Stanford et al. (2014), which they referred to as the *Accumulating Priority Queue*. For convenience, we instead refer to their priority queueing model as the *Non-Preemptive Accumulating Priority Queue* (NPAPQ), and refer to the preemptive priority model of this chapter as the *Preemptive Accumulating Priority Queue* (PAPQ).

Similar to researchers who have previously studied dynamic priority queues, our primary motivation for studying the PAPQ is the ability to control waiting times. While this control has mainly been administered through the expected waiting times, our analysis in this chapter also enables a systems manager to control waiting times via other performance measures such as their quantiles. In particular, the main objective of this chapter is to characterize the LSTs of the steady-state waiting time distributions for each priority class in the PAPQ for all three preemption disciplines: resume, repeat-different, and repeat-identical.

The rest of the chapter is organized as follows. In the next section, we provide the model specifications of the PAPQ, as well as other preliminaries, including the in-

troduction of several key random variables of interest. The maximal priority process for the PAPQ is defined in Section 3.3, while Sections 3.4 and 3.5 are devoted to the derivation of the LSTs corresponding to the service-structure elements. In Section 3.6, we provide a general recursive scheme for obtaining the marginal steady-state waiting time LSTs. In Section 3.7, we investigate the PAPQ under a hybrid-based preemption discipline comprised of a random mixture of the three traditional preemption disciplines. Lastly, in Section 3.8, we provide several numerical examples to illustrate the versatility of the PAPQ. We remark that most of the work presented in this chapter is found in Fajardo and Drekic (2015c).

3.2 The model

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 Poisson 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 df $B^{(k)}(x) = \mathbb{P}(X^{(k)} \leq x)$ and corresponding LST $\widetilde{B}^{(k)}(s) = \int_0^\infty e^{-sx} \mathrm{d}B^{(k)}(x)$. For each $k = 1, 2, \ldots, N$, the class-k priority function is given by

$$q_k(t) = b_k \cdot (t - \tau_k), \quad t \geq \tau_k, \tag{3.1}$$

with $b_1 \geq b_2 \geq \cdots \geq b_N \geq 0$. In other words, upon arrival to the system, a customer begins to accumulate priority linearly at a rate that is distinct to the class to which it belongs. At a service selection instant (i.e., a departure instant of a customer), the system employs the general Priority Service Guideline.

In addition, the 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. (3.1), 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). Furthermore, 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 $\Lambda_{k-1} = \sum_{i=1}^{k-1} \lambda_i$. Finally, it is also important to realize that a preemption instant is not considered to be a service selection instant.

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 chapter is to establish the LST corresponding to the steady-state distribution of $W^{(k)}$. 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 solely 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 the PAPQ is given by

$$\overline{U} = \sum_{i=1}^{N} \rho_i = \sum_{i=1}^{N} \lambda_i \mathbb{E}(G^{(i)}) < 1, \tag{3.2}$$

where $\overline{U}$ is known as the *utilization factor*. The stability condition given by Eq. (3.2) is assumed throughout the chapter. At this juncture of the chapter, we are not in position to provide the expression to calculate $\mathbb{E}(G^{(i)})$, which itself depends

on the class-i preemption rate as well as on the specific preemption discipline in place. Hence, the formulas for $\mathbb{E}(G^{(i)})$, $i = 1, 2, \ldots, N$, are provided later in Section 3.5. Nonetheless, it is important to note that Eq. (3.2) can always be checked first as the expression for $\mathbb{E}(G^{(i)})$ is comprised only of the fundamental elements of the system (i.e., the service time distributions and the arrival rates). 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 iid interruption periods $A^{(k)}$. Furthermore, due to independence, the LST of $F^{(k)}$ can be expressed as

$$\widetilde{F}^{(k)}(s) = \widetilde{W}^{(k)}(s)\widetilde{R}^{(k)}(s). \tag{3.3}$$

Figure 3.1 illustrates the fundamental relationships between the service-structure elements.

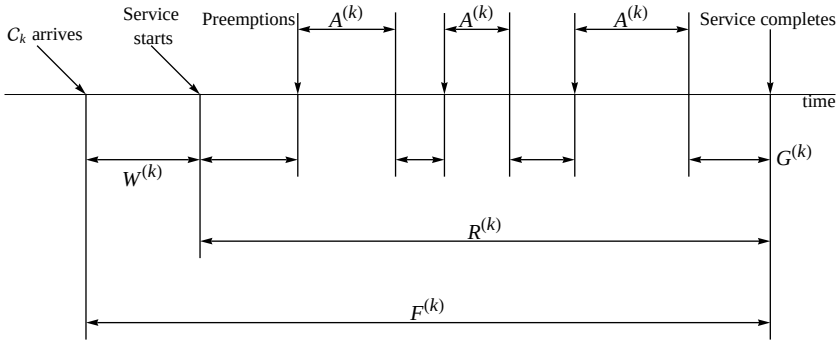


Figure 3.1: Depiction of the service-structure elements for a preemptive priority queue

3.3 The maximal priority process

In this section, we define an upper bound $M_k(t)$ for the accumulated priority 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 priorities 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 steady-state 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 $\chi(t) = \arg\max_{1\le i\le N}\{Q_i(t)\}$ and $Q_\vee(t) = \max_{1\le i\le N}\{Q_i(t)\}$. For any t during an idle period, we further define $\chi(t) = Q_\vee(t) = 0$. Our formal definition of the maximal priority process for the PAPQ now follows.

Definition 3.1 *The maximal priority process is an 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, we 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 smaller of b_k and $b_{\chi(t)}$, and down-jumps at some of the service selection instants (i.e., customer departure instants). Figure 3.2 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.

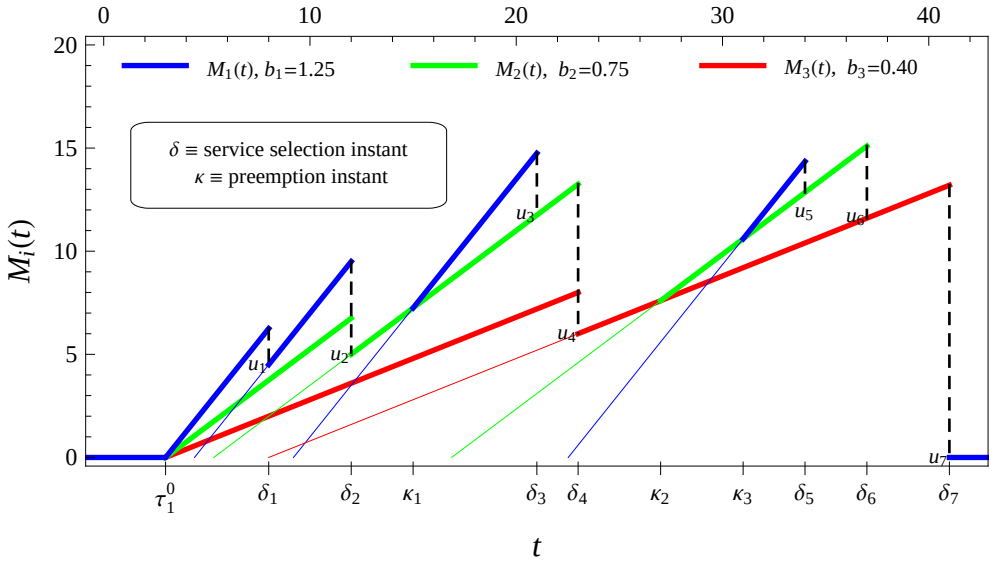


Figure 3.2: $\mathcal{M}(t)$ in a typical busy period of the PAPQ for $N = 3$

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.

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 occurrence 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. (3.1), 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 priorities 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 Figure 3.2 given only the information found in Table 3.1. 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 3.1: Partial information (a)–(d) required to recreate $\mathcal{M}(t)$ of Figure 3.2

	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

3.3.1 Structuralization of the general busy period and its customers

Following the convention of Stanford et al. (2014), we introduce some important 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 chapter, 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 Figure 3.2 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 3.2 *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 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 3.3 *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 3.2 holds true regardless of the specific class to which the $\mathcal{C}^{(acc:m)}$ belongs.*

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

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

We refer to 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 3.4 *The rate of preemption during 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 denotes the aggregation of classes $\{1, 2, \ldots, k-1\}$ as defined earlier) with a priority level greater than or 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. Due to the memoryless property, 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 quite straightforward to show that $Y_i = T_i(1 - b_k/b_i)^{-1}$, and the result readily follows. □

We further this subsection with the introduction of a *level-k accreditation interval*, which starts in one of three possible 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 initial 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 and a remark on the difference between the maximal priority process presented here and the one given in Section 2.5.1.

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. Figure 3.3 illustrates a level-k accreditation interval with four class-i arrivals-to-the-interval.

Lemma 3.5 *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 $\{T_j\}_{j=1}^n$ denote the durations of these subperiods (e.g., see Figure 3.3). 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 Figure 3.3 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 Figure 3.3). 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.

□

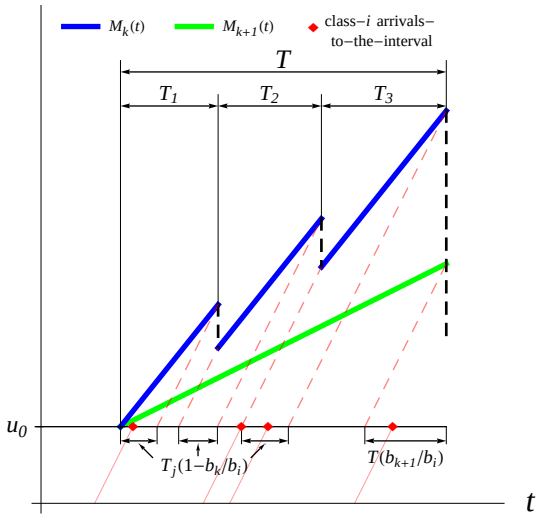


Figure 3.3: Supplemental illustration of a level-k accreditation interval for the proof of Lemma 3.5. Note that T_2 is initiated by a $\mathcal{C}^{(acc:k)}$ not belonging to class i.

Remark 3.6 *In Section 2.5.1, we defined the maximal priority process for an $M/G/1$-type queueing system with only a single class of arriving customers. This process consisted of two components, the second of which was called the accreditation threshold and was used to determine which of the arriving customers would eventually be serviced. In contrast, the maximal priority process of the PAPQ has N components, and with the exception of $M_1(t)$, each plays the role of an "accreditation threshold" in certain circumstances. Specifically, $M_{k+1}(t)$ serves as the "accreditation threshold" in a level-k accreditation interval.*

3.4 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 3.4 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 3.4 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*:

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

We stress that $A_{p_{k+1}}^{(m)}$ 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 $A_{p_{k+1}}^{(k)}$.

For the subsequent subperiods of $A^{(k+1)}$, we realize 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 3.2 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:

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

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

$$\widetilde{V}_{p_{k+1}}^{(k)}(s) = \sum_{i=1}^{k-1} \frac{\lambda_i^{(k+1)}}{\Lambda_k^{(k+1)}} \widetilde{A}_{p_{k+1}}^{(k)}(s) + \frac{\lambda_k^{(k+1)}}{\Lambda_k^{(k+1)}} \widetilde{R}^{(k)}(s), \tag{3.4}$$

and

$$\Phi_k(s) = \frac{\sum_{i=1}^{k-1} \lambda_i/b_i}{\sum_{i=1}^{k} \lambda_i/b_i} \widetilde{A}_{np}^{(k)}(s) + \frac{\lambda_k/b_k}{\sum_{i=1}^{k} \lambda_i/b_i} \widetilde{R}^{(k)}(s). \tag{3.5}$$

In order to verify this claim, 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 in Section 2.5. Recall that this latter 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 (introduced in Section 2.5.1) 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 3.7 *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 the 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*

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

Important Observation 3.8 *It follows from Important Observation 3.7 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) service time LST of $\widetilde{V}(s)$ for zero-wait customers,

(ii) service time LST of $\Phi_k(s)$ for customers arriving during busy periods,

(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 3.8 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 3.8 and Eq. (2.49) that the distribution of the duration of a level-k accreditation interval has corresponding LST

$$\widetilde{\mathcal{A}}_k(s) \equiv \widetilde{\mathcal{A}}_k(s; V) = \widetilde{V}\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big), \tag{3.6}$$

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

$$\eta_k(s) = \Phi_k\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big). \tag{3.7}$$

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. (3.4). Moreover, from Eq. (3.6), we observe that

$$\widetilde{A}^{(k+1)}(s) = \widetilde{A}_{p_{k+1}}^{(k+1)}(s) = \widetilde{\mathcal{A}}_k(s; V_{p_{k+1}}^{(k)}). \tag{3.8}$$

Eq. (3.8) also leads to the following recursive scheme which starts with $\widetilde{A}_{p_{k+1}}^{(1)}(s) = 1$ and holds for all $m = 1, 2, \ldots, k$:

$$\widetilde{A}_{p_{k+1}}^{(m+1)}(s) = \frac{\Lambda_{m-1}^{(k+1)}}{\Lambda_m^{(k+1)}} \widetilde{A}_{p_{k+1}}^{(m)}\big(s + \gamma_m^{(m+1)}(1 - \eta_m(s))\big) + \frac{\lambda_m^{(k+1)}}{\Lambda_m^{(k+1)}} \widetilde{R}^{(m)}\big(s + \gamma_m^{(m+1)}(1 - \eta_m(s))\big). \tag{3.9}$$

By taking the first and second derivatives of $\widetilde{A}_{p_{k+1}}^{(m+1)}(s)$, recursions for the first two moments of $A_{p_{k+1}}^{(m+1)}$ are obtained. In particular, for $m = 1, 2, \ldots, k$, we get

$$\mathbb{E}(A_{p_{k+1}}^{(m+1)}) = \frac{\Lambda_{m-1}^{(k+1)} \mathbb{E}(A_{p_{k+1}}^{(m)}) + \lambda_m^{(k+1)} \mathbb{E}(R^{(m)})}{\Lambda_m^{(k+1)} \left(1 - \sum_{i=1}^{m-1} \lambda_i \frac{b_m - b_{m+1}}{b_i} \mathbb{E}(A_{np}^{(m)}) - \lambda_m^{(m+1)} \mathbb{E}(R^{(m)})\right)} \tag{3.10}$$

and

$$\mathbb{E}\big((A_{p_{k+1}}^{(m+1)})^2\big) = \frac{\gamma_m^{(m+1)} \mu_{m,2} \left(\Lambda_{m-1}^{(k+1)} \mathbb{E}(A_{p_{k+1}}^{(m)}) + \lambda_m^{(k+1)} \mathbb{E}(R^{(m)})\right)}{\Lambda_m^{(k+1)} \left(1 - \sum_{i=1}^{m-1} \lambda_i \frac{b_m - b_{m+1}}{b_i} \mathbb{E}(A_{np}^{(m)}) - \lambda_m^{(m+1)} \mathbb{E}(R^{(m)})\right)^3} + \frac{(1 - \gamma_m^{(m+1)} \mu_{m,1}) \left(\Lambda_{m-1}^{(k+1)} \mathbb{E}\big((A_{p_{k+1}}^{(m)})^2\big) + \lambda_m^{(k+1)} \mathbb{E}\big((R^{(m)})^2\big)\right)}{\Lambda_m^{(k+1)} \left(1 - \sum_{i=1}^{m-1} \lambda_i \frac{b_m - b_{m+1}}{b_i} \mathbb{E}(A_{np}^{(m)}) - \lambda_m^{(m+1)} \mathbb{E}(R^{(m)})\right)^3}, \tag{3.11}$$

where $\mu_{k,i}$ is the i-th moment of the random variable whose distribution has LST $\Phi_k(s)$.

It is clear that the recursive scheme of Eq. (3.9) requires that both $\widetilde{R}^{(m)}(s)$ and $\widetilde{A}_{np}^{(m)}(s)$ for $m = 1, 2, \ldots, k$ be priorly established. The former is the subject of the next section. Consider $A_{np}^{(k+1)}$, which represents a level-k accreditation interval which begins with the service of a $\mathcal{C}_k^{(acc:\ell)}$ or $\mathcal{C}_a^{(acc:\ell)}$ for some $\ell > k$. It follows from Lemma 3.2 that the initial subperiod is either $A_{np}^{(k)}$ with probability $(\sum_{i=1}^{k-1} \lambda_i/b_i)/(\sum_{i=1}^{k} \lambda_i/b_i)$ or $R^{(k)}$ with probability $(\lambda_k/b_k)/(\sum_{i=1}^{k} \lambda_i/b_i)$. In other words, for the level-k accreditation interval $A_{np}^{(k+1)}$, V has LST $\widetilde{V}_{np}^{(k)}(s) = \Phi_k(s)$. Therefore, we have that

$$\widetilde{A}_{np}^{(k+1)}(s) = \widetilde{\mathcal{A}}_k(s; V_{np}^{(k)}) = \eta_k(s), \tag{3.12}$$

which again yields a recursive scheme starting with $\widetilde{A}_{np}^{(1)}(s) = 1$. Furthermore, the corresponding first two moments are:

$$\mathbb{E}(A_{np}^{(k+1)}) = \frac{\sum_{i=1}^{k-1} \lambda_i \frac{b_k}{b_i} \mathbb{E}(A_{np}^{(k)}) + \lambda_k \mathbb{E}(R^{(k)})}{\gamma_k \left(1 - \sum_{i=1}^{k-1} \lambda_i \frac{b_k - b_{k+1}}{b_i} \mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)} \mathbb{E}(R^{(k)})\right)} \tag{3.13}$$

and

$$\mathbb{E}\big((A_{np}^{(k+1)})^2\big) = \frac{\sum_{i=1}^{k-1} \lambda_i \frac{b_k}{b_i} \mathbb{E}\big((A_{np}^{(k)})^2\big) + \lambda_k \mathbb{E}\big((R^{(k)})^2\big)}{\gamma_k \left(1 - \sum_{i=1}^{k-1} \lambda_i \frac{b_k - b_{k+1}}{b_i} \mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)} \mathbb{E}(R^{(k)})\right)^3}. \tag{3.14}$$

The recursive schemes of Eqs. (3.8) and (3.12) 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:

$A_{p_0}^{(m)}$ (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 $A_{p_0}^{(k+1)}$ and remark that the initial subperiod is either $R^{(k)}$ with probability λ_k/Λ_k or $A_{p_0}^{(k)}$ 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,

$$\widetilde{A}_{p_0}^{(k+1)}(s) = \widetilde{\mathcal{A}}_k(s; V_{p_0}^{(k)}), \tag{3.15}$$

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

$$\widetilde{A}_{p_0}^{(k+1)}(s) = \frac{\Lambda_{k-1}}{\Lambda_k}\widetilde{A}_{p_0}^{(k)}\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big) + \frac{\lambda_k}{\Lambda_k}\widetilde{R}^{(k)}\big(s + \gamma_k^{(k+1)}(1 - \eta_k(s))\big). \tag{3.16}$$

Through differentiation again, the associated first two moments work out to be

$$\mathbb{E}(A_{p_0}^{(k+1)}) = \frac{\Lambda_{k-1}\mathbb{E}(A_{p_0}^{(k)}) + \lambda_k\mathbb{E}(R^{(k)})}{\Lambda_k\left(1 - \sum_{i=1}^{k-1}\lambda_i\frac{b_k - b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k)}\mathbb{E}(R^{(k)})\right)} \tag{3.17}$$

and

$$\mathbb{E}\big((A_{p_0}^{(k+1)})^2\big) = \frac{\gamma_k^{(k+1)}\mu_{k,2}\left(\Lambda_{k-1}\mathbb{E}(A_{p_0}^{(k)}) + \lambda_k\mathbb{E}(R^{(k)})\right)}{\Lambda_k\left(1 - \sum_{i=1}^{k-1}\lambda_i\frac{b_k - b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)}\mathbb{E}(R^{(k)})\right)^3} + \frac{(1 - \gamma_k^{(k+1)}\mu_{k,1})\Big(\Lambda_{k-1}\mathbb{E}\big((A_{p_0}^{(k)})^2\big) + \lambda_k\mathbb{E}\big((R^{(k)})^2\big)\Big)}{\Lambda_k\left(1 - \sum_{i=1}^{k-1}\lambda_i\frac{b_k - b_{k+1}}{b_i}\mathbb{E}(A_{np}^{(k)}) - \lambda_k^{(k+1)}\mathbb{E}(R^{(k)})\right)^3}. \tag{3.18}$$

For illustrative purposes, Figure 3.4 depicts the general structure of a class-3 pseudo-interruption period. We also remark that 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. (1.16) in this model, these pseudo-interruption periods are equivalent in distribution to an actual interruption period $A^{(k)}$.

We close this section with the following proposition which provides three useful identities involving the means of each of the pseudo-interruption periods.

Proposition 3.9 *Let* $\overline{U}_j = \sum_{i=1}^{j}\lambda_i\mathbb{E}(G^{(i)})$ *and* $\overline{U}_j^{(k+1)} = \sum_{i=1}^{j}\lambda_i^{(k+1)}\mathbb{E}(G^{(i)})$. *For* $k = 1, 2, \ldots, N-1$,

$$\gamma_k\mathbb{E}(A_{np}^{(k+1)}) = \frac{\sum_{i=1}^{k}\lambda_i(b_k/b_i)\mathbb{E}(G^{(i)})}{1 - \overline{U}_k^{(k+1)}} = \frac{\overline{U}_k - \overline{U}_{k-1}^{(k)}}{1 - \overline{U}_k^{(k+1)}}, \tag{3.19}$$

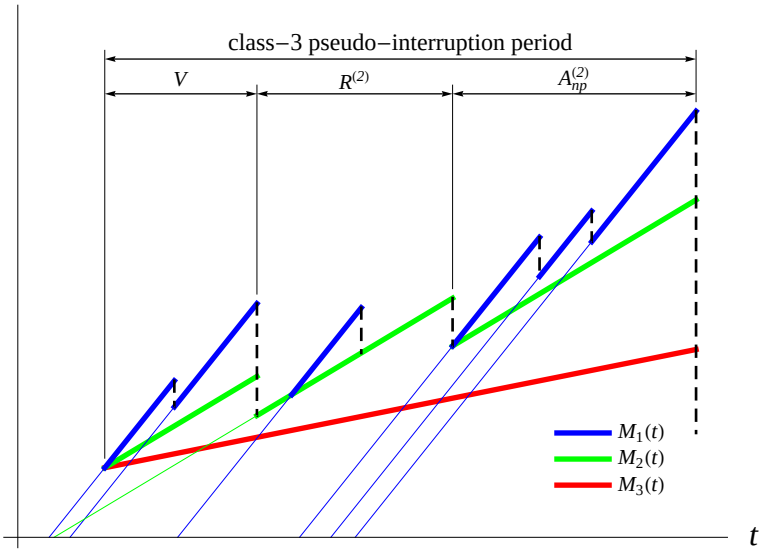


Figure 3.4: General structure of a class-3 pseudo-interruption period

and for each value of k*,*

$$\Lambda_m^{(k+1)}\mathbb{E}(A_{p_{k+1}}^{(m+1)}) = \frac{\overline{U}_m^{(k+1)}}{1-\overline{U}_m^{(m+1)}}, \quad m = 1, 2, \ldots, k. \tag{3.20}$$

Furthermore, for $k = 1, 2, \ldots, N-1$*,*

$$\Lambda_k\mathbb{E}(A_{p_0}^{(k+1)}) = \frac{\overline{U}_k}{1-\overline{U}_k^{(k+1)}}. \tag{3.21}$$

Proof. We prove Eqs. (3.19) and (3.20) by induction. For $k = 1$, it readily follows from Eq. (3.13) (since $X^{(1)} = R^{(1)} = G^{(1)}$) that

$$\mathbb{E}(A_{np}^{(2)}) = \frac{\lambda_1\mathbb{E}(G^{(1)})}{\gamma_1(1-\lambda_1^{(2)}\mathbb{E}(G^{(1)}))} = \frac{\lambda_1\mathbb{E}(G^{(1)})}{\gamma_1(1-\overline{U}_1^{(2)})}.$$

Similarly, from Eq. (3.10), we have that

$$\mathbb{E}(A_{p_2}^{(2)}) = \frac{\lambda_1^{(2)}\mathbb{E}(G^{(1)})}{\Lambda_1^{(2)}(1-\lambda_1^{(2)}\mathbb{E}(G^{(1)}))} = \frac{\overline{U}_1^{(2)}}{\Lambda_1^{(2)}(1-\overline{U}_1^{(2)})}.$$

Next, we assume that Eq. (3.19) holds for $k = 1, 2, \ldots, n-1$, and also that for each k, Eq. (3.20) holds for $m = 1, 2, \ldots, k$. Hence, Eq. (3.19) with $k = n-1$ yields

$$\gamma_{n-1}\mathbb{E}(A_{np}^{(n)}) = \frac{\sum_{i=1}^{n-1}\lambda_i(b_{n-1}/b_i)\mathbb{E}(G^{(i)})}{1-\overline{U}_{n-1}^{(n)}}, \tag{3.22}$$

and Eq. (3.20) with $m = k = n-1$ together with the result $\mathbb{E}(R^{(n)}) = \mathbb{E}(G^{(n)})(1 + \Lambda_{n-1}^{(n)}\mathbb{E}(A_{p_n}^{(n)}))$ (as indicated in the next section) yields

$$\mathbb{E}(R^{(n)}) = \frac{\mathbb{E}(G^{(n)})}{1-\overline{U}_{n-1}^{(n)}}. \tag{3.23}$$

On the other hand, Eq. (3.13) with $k = n$ results in

$$\mathbb{E}(A_{np}^{(n+1)}) = \frac{\sum_{i=1}^{n-1}\lambda_i(b_n/b_i)\mathbb{E}(A_{np}^{(n)}) + \lambda_n\mathbb{E}(R^{(n)})}{\gamma_n\left(1-\sum_{i=1}^{n-1}\lambda_i((b_n-b_{n+1})/b_i)\mathbb{E}(A_{np}^{(n)}) - \lambda_n^{(n+1)}\mathbb{E}(R^{(n)})\right)}. \tag{3.24}$$

Note that $\sum_{i=1}^{n-1}\lambda_i(b_n/b_i)\mathbb{E}(A_{np}^{(n)}) = (b_n/b_{n-1})\gamma_{n-1}\mathbb{E}(A_{np}^{(n)})$. Thus, after appropriate substitution of Eqs. (3.22) and (3.23), the numerator of Eq. (3.24) can be re-written as

$$\frac{\sum_{i=1}^{n-1}\lambda_i(b_n/b_i)\mathbb{E}(G^{(i)}) + \lambda_n\mathbb{E}(G^{(n)})}{1-\overline{U}_{n-1}^{(n)}}.$$

Upon observing that $\sum_{i=1}^{n-1}\lambda_i((b_n-b_{n+1})/b_i)\mathbb{E}(A_{np}^{(n)}) = ((b_n-b_{n+1})/b_{n-1})\gamma_{n-1}\mathbb{E}(A_{np}^{(n)})$, it similarly follows that the denominator of Eq. (3.24) can be re-expressed as

$$\frac{\gamma_n\big(1-\overline{U}_{n-1}^{(n)} - \sum_{i=1}^{n-1}\lambda_i((b_n-b_{n+1})/b_i)\mathbb{E}(G^{(i)}) - \lambda_n^{(n+1)}\mathbb{E}(G^{(n)})\big)}{1-\overline{U}_{n-1}^{(n)}}.$$

Therefore,

$$\mathbb{E}(A_{np}^{(n+1)}) = \frac{\sum_{i=1}^{n-1}\lambda_i(b_n/b_i)\mathbb{E}(G^{(i)}) + \lambda_n\mathbb{E}(G^{(n)})}{\gamma_n\left(1-\overline{U}_{n-1}^{(n)} - \sum_{i=1}^{n-1}\lambda_i((b_n-b_{n+1})/b_i)\mathbb{E}(G^{(i)}) - \lambda_n^{(n+1)}\mathbb{E}(G^{(n)})\right)},$$

which, after some straightforward algebra, becomes

$$\mathbb{E}(A_{np}^{(n+1)}) = \frac{\sum_{i=1}^{n}\lambda_i(b_n/b_i)\mathbb{E}(G^{(i)})}{\gamma_n(1-\overline{U}_{n}^{(n+1)})}.$$

All that remains to complete the proof is to show that Eq. (3.20) holds for $m = 1, 2, \ldots, k$ when $k = n$. To accomplish this, we again employ a proof by induction. Using Eq. (3.10) and the fact that $X^{(1)} = R^{(1)} = G^{(1)}$, it follows that

$$\mathbb{E}(A^{(2)}_{p_{n+1}}) = \frac{\lambda_1^{(n+1)}\mathbb{E}(G^{(1)})}{\Lambda_1^{(n+1)}(1 - \lambda_1^{(2)}\mathbb{E}(G^{(1)}))} = \frac{\overline{U}_1^{(n+1)}}{\Lambda_1^{(n+1)}(1 - \overline{U}_1^{(2)})}.$$

Next, we assume that when $k = n$, Eq. (3.20) holds for $m = 1, 2, \ldots, j-1$ where $j - 1 < n$. Under this assumption, we therefore have for $k = n$ and $m = j - 1$ that

$$\Lambda_{j-1}^{(n+1)}\mathbb{E}(A^{(j)}_{p_{n+1}}) = \frac{\overline{U}_{j-1}^{(n+1)}}{1 - \overline{U}_{j-1}^{(j)}}. \tag{3.25}$$

Moreover, it follows from our initial inductive hypothesis (since $j < n - 1$) that

$$\gamma_{j-1}\mathbb{E}(A^{(j)}_{np}) = \frac{\sum_{i=1}^{j-1}\lambda_i(b_{j-1}/b_i)\mathbb{E}(G^{(i)})}{1 - \overline{U}_{j-1}^{(j)}} \tag{3.26}$$

and

$$\mathbb{E}(R^{(j)}) = \mathbb{E}(G^{(j)})(1 + \Lambda_{j-1}^{(j)}\mathbb{E}(A^{(j)}_{p_j})) = \frac{\mathbb{E}(G^{(j)})}{1 - \overline{U}_{j-1}^{(j)}}. \tag{3.27}$$

On the other hand, Eq. (3.10) with $k = n$ and $m = j$ gives us

$$\mathbb{E}(A^{(j+1)}_{p_{n+1}}) = \frac{\Lambda_{j-1}^{(n+1)}\mathbb{E}(A^{(j)}_{p_{n+1}}) + \lambda_j^{(n+1)}\mathbb{E}(R^{(j)})}{\Lambda_j^{(n+1)}\left(1 - \sum_{i=1}^{j-1}\lambda_i((b_j - b_{j+1})/b_i)\mathbb{E}(A^{(j)}_{np}) - \lambda_j^{(j+1)}\mathbb{E}(R^{(j)})\right)}. \tag{3.28}$$

After using similar arguments to those made earlier in this proof, and following the appropriate substitution of Eqs. (3.25)–(3.27) into Eq. (3.28), we ultimately obtain

$$\begin{aligned}\mathbb{E}(A^{(j+1)}_{p_{n+1}}) &= \frac{\overline{U}_{j-1}^{(n+1)} + \lambda_j^{(n+1)}\mathbb{E}(G^{(j)})}{\Lambda_j^{(n+1)}\left(1 - \overline{U}_{j-1}^{(j)} - \sum_{i=1}^{j-1}\lambda_i((b_j - b_{j+1})/b_i)\mathbb{E}(G^{(i)}) - \lambda_j^{(j+1)}\mathbb{E}(G^{(j)})\right)} \\ &= \frac{\overline{U}_j^{(n+1)}}{\Lambda_j^{(n+1)}(1 - \overline{U}_j^{(j+1)})}.\end{aligned}$$

We omit the details, but similar inductive arguments can be used to prove Eq. (3.21).

□

3.5 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 experience 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 realize 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 3.4.

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

Resume:

$$\widetilde{R}^{(k)}(s) = \widetilde{B}^{(k)}\big(s + \Lambda_{k-1}^{(k)}(1 - \widetilde{A}^{(k)}(s))\big) \tag{3.29}$$

$$\mathbb{E}(R^{(k)}) = \big(1 + \Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}(X^{(k)}) \tag{3.30}$$

$$\mathbb{E}\big((R^{(k)})^2\big) = \big(1 + \Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)^2\mathbb{E}\big((X^{(k)})^2\big) + \Lambda_{k-1}^{(k)}\mathbb{E}(X^{(k)})\mathbb{E}\big((A^{(k)})^2\big) \tag{3.31}$$

$$\widetilde{G}^{(k)}(s) = \widetilde{B}^{(k)}(s) \tag{3.32}$$

$$\mathbb{E}(G^{(k)}) = \mathbb{E}(X^{(k)}) \tag{3.33}$$

$$\mathbb{E}\big((G^{(k)})^2\big) = \mathbb{E}\big((X^{(k)})^2\big) \tag{3.34}$$

Repeat-different:

$$\widetilde{R}^{(k)}(s) = \frac{\left(s + \Lambda_{k-1}^{(k)}\right)\widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)})}{s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(s)\left(1 - \widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)})\right)} \tag{3.35}$$

$$\mathbb{E}(R^{(k)}) = \big(1 + \Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}(G^{(k)}) \tag{3.36}$$

$$\mathbb{E}\big((R^{(k)})^2\big) = (1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\mathbb{E}\big((G^{(k)})^2\big) + \Lambda_{k-1}^{(k)}\mathbb{E}\big((A^{(k)})^2\big)\mathbb{E}(G^{(k)}) + 2\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\big(\mathbb{E}(G^{(k)})\big)^2 \quad (3.37)$$

$$\widetilde{G}^{(k)}(s) = \frac{\left(s+\Lambda_{k-1}^{(k)}\right)\widetilde{B}^{(k)}(s+\Lambda_{k-1}^{(k)})}{s+\Lambda_{k-1}^{(k)}\widetilde{B}^{(k)}(s+\Lambda_{k-1}^{(k)})} \quad (3.38)$$

$$\mathbb{E}(G^{(k)}) = \frac{1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})}{\Lambda_{k-1}^{(k)}\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})} \quad (3.39)$$

$$\mathbb{E}\big((G^{(k)})^2\big) = \frac{2}{\left(\Lambda_{k-1}^{(k)}\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})\right)^2}\left(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}) - \Lambda_{k-1}^{(k)}\mathbb{E}\Big(X^{(k)}e^{-\Lambda_{k-1}^{(k)}X^{(k)}}\Big)\right) \quad (3.40)$$

Repeat-identical:

$$\widetilde{R}^{(k)}(s) = \mathbb{E}[\mathbb{E}(e^{-sR^{(k)}}|X^{(k)})] = \int_{x=0}^{\infty}\frac{(s+\Lambda_{k-1}^{(k)})e^{-(s+\Lambda_{k-1}^{(k)})x}}{s+\Lambda_{k-1}^{(k)}-\Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(s)(1-e^{-(s+\Lambda_{k-1}^{(k)})x})}\,\mathrm{d}B^{(k)}(x) \quad (3.41)$$

$$\mathbb{E}(R^{(k)}) = \big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}(G^{(k)}) \quad (3.42)$$

$$\mathbb{E}\big((R^{(k)})^2\big) = \big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)\mathbb{E}\big((G^{(k)})^2\big) + \Lambda_{k-1}^{(k)}\mathbb{E}\big((A^{(k)})^2\big)\mathbb{E}(G^{(k)}) + \frac{2\mathbb{E}(A^{(k)})\big(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)})\big)}{\Lambda_{k-1}^{(k)}}\mathbb{E}\big((e^{\Lambda_{k-1}^{(k)}X^{(k)}}-1)^2\big) \quad (3.43)$$

$$\widetilde{G}^{(k)}(s) = \mathbb{E}[\mathbb{E}(e^{-sG^{(k)}}|X^{(k)})] = \int_{x=0}^{\infty}\frac{(s+\Lambda_{k-1}^{(k)})e^{-(s+\Lambda_{k-1}^{(k)})x}}{s+\Lambda_{k-1}^{(k)}-\Lambda_{k-1}^{(k)}(1-e^{-(s+\Lambda_{k-1}^{(k)})x})}\,\mathrm{d}B^{(k)}(x) \quad (3.44)$$

$$\mathbb{E}(G^{(k)}) = \mathbb{E}[\mathbb{E}(G^{(k)}|X^{(k)})] = \mathbb{E}\left(\frac{e^{\Lambda_{k-1}^{(k)}X^{(k)}}-1}{\Lambda_{k-1}^{(k)}}\right) = \frac{\widetilde{B}^{(k)}\left(-\Lambda_{k-1}^{(k)}\right)-1}{\Lambda_{k-1}^{(k)}} \quad (3.45)$$

$$\mathbb{E}\big((G^{(k)})^2\big) = \mathbb{E}\big[\mathbb{E}\big((G^{(k)})^2|X^{(k)}\big)\big] = \mathbb{E}\left[\frac{2}{(\Lambda_{k-1}^{(k)})^2}\left(e^{2\Lambda_{k-1}^{(k)}X^{(k)}} - e^{\Lambda_{k-1}^{(k)}X^{(k)}} - \Lambda_{k-1}^{(k)}X^{(k)}e^{\Lambda_{k-1}^{(k)}X^{(k)}}\right)\right] \quad (3.46)$$

$$= \frac{2}{(\Lambda_{k-1}^{(k)})^2} \left(\widetilde{B}^{(k)}(-2\Lambda_{k-1}^{(k)}) - \widetilde{B}^{(k)}(-\Lambda_{k-1}^{(k)}) - \Lambda_{k-1}^{(k)} \mathbb{E}(X^{(k)} e^{\Lambda_{k-1}^{(k)} X^{(k)}}) \right) \tag{3.47}$$

We next present a similar result to Lemma 3.5. 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) at which a $\mathcal{C}_i$ for $i \in \{1, 2, \ldots, k\}$ accumulates a priority level equal to the initial level u_0.

Lemma 3.10 *The steady-state 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. One can use similar arguments as those used in the proof of Lemma 3.5 to prove this particular result. Specifically, observe that for class i, $i \leq k$, there exists a subperiod during $R^{(k+1)}$ for which a class-i arrival-to-the-residence-period within it would eventually lead to level-k accreditation for the $\mathcal{C}_i$. Furthermore, the ratio of this subperiod to the entire $R^{(k+1)}$ is always $1 - b_{k+1}/b_i$. Thus, the result follows from the fact that the class-i arrivals-to-the-residence-period form a Poisson process with rate λ_i. □

3.6 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

$$\widetilde{W}^{(k)}(s) = \pi_0 + (1 - \pi_0)\widetilde{W}_{BP}^{(k)}(s), \tag{3.48}$$

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. (3.1), it follows that

$$\widetilde{W}_{BP}^{(k)}(s) = \widetilde{P}_{BP}^{(k)}(s/b_k). \tag{3.49}$$

Hence, to find $\widetilde{W}^{(k)}(s)$, we first find $\widetilde{P}_{BP}^{(k)}(s)$ and subsequently apply Eqs. (3.48) and (3.49).

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 a $\mathcal{C}_k$ 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. Therefore,

$$\widetilde{P}_{BP}^{(k)}(s) = \frac{1}{1-\pi_0}\left[\pi_k^{(acc)}\widetilde{P}_{acc}^{(k)}(s) + \alpha_k^{(int)}\widetilde{P}_{int}^{(k)}(s)\right], \tag{3.50}$$

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.

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

We present first a recursion for $\widetilde{P}_{acc}^{(k)}(s)$. To begin, 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, we have

$$\widetilde{P}_{acc}^{(k)}(s) = \frac{1}{\pi_k^{(acc)}}\left[\pi_k^{(k)}\widetilde{P}_{acc:k}^{(k)}(s) + \sum_{\ell=k+1}^{N}\pi_k^{(\ell)}\widetilde{P}_{unacc:k}^{(k)}(s)\right], \tag{3.51}$$

where $\pi_k^{(j)}$ represents the steady-state probability that a $\mathcal{C}_k$ arrives to a busy period and is serviced at level-j accreditation. It follows from Lemma 3.2 and Remark 3.3 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 later in Section 3.6.3), ultimately leads to the following recursive scheme for the desired LST:

$$\widetilde{P}_{acc}^{(k)}(s) = \frac{1}{\pi_k^{(acc)}}\left[\pi_k^{(k)}\widetilde{P}_{acc:k}^{(k)}(s) + \frac{b_{k+1}}{b_k}\pi_{k+1}^{(acc)}\widetilde{P}_{acc}^{(k+1)}(s)\right]. \tag{3.52}$$

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. Suppose now 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 that which exists 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 3.8. In particular, it readily follows from Important Observation 3.8 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), can be expressed as an application of Eq. (2.59) with $q = b_{k+1}/b_k$ and LST argument $b_k s$, namely:

$$\widetilde{\mathcal{P}}^{(acc:k)}(s) \equiv \widetilde{\mathcal{P}}^{(acc:k)}(s;V) = \frac{\left(1-\gamma_k^{(k+1)}\mu_{k,1}\right)\left(\widetilde{\mathcal{A}}_k(b_{k+1}s) - \widetilde{V}(b_k s)\right)}{\mathbb{E}(V)\left(1-b_{k+1}/b_k\right)\left(b_k s - \gamma_k + \gamma_k \Phi_k(b_k s)\right)}. \tag{3.53}$$

Note that in Eq. (3.53), $\widetilde{\mathcal{A}}_k(s)$ is given by Eq. (3.6) and $\mu_{k,i}$ is the i-th moment of the random variable whose distribution has LST $\Phi_k(s)$. The first moment of $\mathcal{P}^{(acc:k)}$ can be found by substituting the appropriate parameters into Eq. (2.61) and subsequently multiplying by b_k, thus yielding

$$\mathbb{E}(\mathcal{P}^{(acc:k)}) = b_k\left(\frac{\mathbb{E}(V^2)}{2\mathbb{E}(V)} \cdot \left[1 + \frac{b_{k+1}/b_k}{1-\gamma_k^{(k+1)}\mu_{k,1}}\right] + \frac{\gamma_k \mu_{k,2}}{2(1-\gamma_k\mu_{k,1})} \cdot \left[1 - \left(\frac{b_{k+1}/b_k}{1-\gamma_k^{(k+1)}\mu_{k,1}}\right)^2\right]\right). \tag{3.54}$$

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 either 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 either 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 $\ell > k$.

We next 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

$$\widetilde{P}_{acc:k}^{(k)}(s) = \frac{1}{\pi_k^{(k)}}\Bigg[\pi_k^{(k:i)}\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{p_0}^{(k)}) + \sum_{\ell=k+1}^{N} \pi_k^{(k:ii:\ell)}\widetilde{P}_{acc:\ell}^{(\ell)}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{np}^{(k)}) + \sum_{\ell=k+1}^{N} \pi_k^{(k:iii:\ell)}\widetilde{P}_{int:\ell}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; V_{p_\ell}^{(k)})\Bigg]. \quad (3.55)$$

Remark 3.11 *In Important Observation 3.8, we described a key relation between the maximal priority process of the PAPQ with the maximal priority process of the*

FCFS $M/G/1$ queue with accumulating priority and blocking (of Section 2.5), which led to our expression for $\widetilde{\mathcal{P}}^{(acc:k)}(s)$. We stress that the idea of relating processes of a priority queue to that of a related $M/G/1$ queue is commonly used in the analysis of priority queueing systems. For example, one possible method for analyzing the static non-preemptive/preemptive priority model is via level-crossing techniques, where the analysis is simplified by relating 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, Section 3.12)).

3.6.2 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

$$\widetilde{P}_{int:\ell}^{(k)}(s) = \frac{\alpha_k^{(0:\ell)}\widetilde{\mathcal{P}}^{(int:\ell)}(s) + \alpha_k^{(1:\ell)}\widetilde{P}_{BP}^{(\ell)}(s)\widetilde{\mathcal{P}}^{(int:\ell)}(s)}{\alpha_k^{(\ell)}}, \tag{3.56}$$

where:

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

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. (3.56) as

$$\widetilde{P}_{int:\ell}^{(k)}(s) = \widetilde{P}_{int:\ell}(s) = \pi_0 \widetilde{\mathcal{P}}^{(int:\ell)}(s) + (1-\pi_0)\widetilde{P}_{BP}^{(\ell)}(s)\widetilde{\mathcal{P}}^{(int:\ell)}(s). \tag{3.57}$$

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. (3.57) is used in Eq. (3.55). 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,

$$\widetilde{P}_{int}^{(k)}(s) = \frac{1}{\alpha_k^{(int)}} \sum_{\ell=k+1}^{N} \alpha_k^{(\ell)} \widetilde{P}_{int:\ell}(s). \tag{3.58}$$

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

$$\mathbb{E}(e^{-s\mathcal{P}^{(int:k)}} | X_{past}^{(k)} = x, \mathcal{N} = n) = e^{-sb_k x}\left[\widetilde{A}^{(k)}(b_k s)\right]^n. \tag{3.59}$$

By Lemma 3.4, given that $X_{past}^{(k)} = x$, $\mathcal{N}$ is Poisson distributed with rate $\Lambda_{k-1}^{(k)} x$. On the other hand, the LST of $X_{past}^{(k)}$ is well-known, being given by (e.g., see Takagi (1991, Eq. (1.52a)))

$$\widetilde{X}_{past}^{(k)}(s) = \frac{1-\widetilde{B}^{(k)}(s)}{s\mathbb{E}(X^{(k)})}. \tag{3.60}$$

Thus, by removing the conditional statements on both $\mathcal{N}$ and $X_{past}^{(k)}$, we readily obtain

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = \frac{1-\widetilde{B}^{(k)}(sb_k + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_k s))}{\mathbb{E}(X^{(k)})(sb_k + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_k s))} \tag{3.61}$$

with corresponding first moment

$$\mathbb{E}(\mathcal{P}^{(int:k)}) = b_k\left(\frac{\mathbb{E}[(X^{(k)})^2]}{2\mathbb{E}(X^{(k)})}(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\right). \tag{3.62}$$

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 progressing 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, we 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

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

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,

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

Using results from semi-Markov theory (e.g., see Kao (1996, Section 6.2)) and discrete-time Markov chains, it follows that $P_m = \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})[1 - \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)})]^m$, thereby leading to

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = \frac{1 - \widetilde{B}^{(k)}(b_k s + \Lambda_{k-1}^{(k)})}{\mathbb{E}(G^{(k)})\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_k s)(1 - \widetilde{B}^{(k)}(b_k s + \Lambda_{k-1}^{(k)}))\big)} \qquad (3.63)$$

with corresponding first moment

$$\mathbb{E}(\mathcal{P}^{(int:k)}) = b_k \left(\frac{\mathbb{E}[(G^{(k)})^2]}{2\mathbb{E}(G^{(k)})} + \Lambda_{k-1}^{(k)} \mathbb{E}(A^{(k)}) \mathbb{E}(G^{(k)}) \right). \tag{3.64}$$

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 $\mathcal{C}_k$, which we denote as $X_*^{(k)}$. It can be shown that the LST corresponding to $X_*^{(k)}$ is given by

$$\mathbb{E}(e^{-sX_*^{(k)}}) = \frac{\widetilde{B}^{(k)}(s - \Lambda_{k-1}^{(k)}) - \widetilde{B}^{(k)}(s)}{\widetilde{B}^{(k)}(-\Lambda_{k-1}^{(k)}) - 1}. \tag{3.65}$$

From Eq. (3.65), we readily obtain that

$$\mathbb{P}(x < X_*^{(k)} < x + \mathrm{d}x) = \frac{\mathbb{E}[G^{(k)}|X^{(k)} = x]}{\mathbb{E}(G^{(k)})} \mathrm{d}B^{(k)}(x). \tag{3.66}$$

Following along the same line of reasoning as for the repeat-different case, we obtain

$$\mathbb{E}[e^{-s\mathcal{P}^{(int:k)}}|X_*^{(k)} = x] = \frac{1 - e^{-(sb_k + \Lambda_{k-1}^{(k)})x}}{\mathbb{E}[G^{(k)}|X^{(k)} = x]\big(b_k s + \Lambda_{k-1}^{(k)}(1 - \widetilde{A}^{(k)}(b_k s)(1 - e^{-(sb_k + \Lambda_{k-1}^{(k)})x}))\big)}, \tag{3.67}$$

which, after removing the condition $X_*^{(k)} = x$, yields

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = \int_{x=0}^{\infty} \frac{\big(1 - e^{-(sb_k + \Lambda_{k-1}^{(k)})x}\big)}{\mathbb{E}(G^{(k)})\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)} \widetilde{A}^{(k)}(b_k s)(1 - e^{-(sb_k + \Lambda_{k-1}^{(k)})x})\big)} \, \mathrm{d}B^{(k)}(x). \tag{3.68}$$

In addition, we can express the corresponding first moment as

$$\mathbb{E}(\mathcal{P}^{(int:k)}) = b_k \int_{x=0}^{\infty} \left\{ \frac{\mathbb{E}[(G^{(k)})^2|X^{(k)} = x]}{2\mathbb{E}(G^{(k)})} + \frac{\Lambda_{k-1}^{(k)} \mathbb{E}(A^{(k)})(\mathbb{E}[G^{(k)}|X^{(k)} = x])^2}{\mathbb{E}(G^{(k)})} \right\} \mathrm{d}B^{(k)}(x). \tag{3.69}$$

The first and second conditional moments of $G^{(k)}$ found in the integrand of Eq. (3.69) are given in Eqs. (3.45) and (3.46), respectively.

3.6.3 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 3.12 *The probability that a $\mathcal{C}_k$ arrives to a busy period and is serviced at level-ℓ accreditation is*

$$\pi_k^{(\ell)} = \overline{U}_\ell (b_\ell - b_{\ell+1})/b_k \quad \textit{for } \ell \geq k. \tag{3.70}$$

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

$$\alpha_k^{(\ell)} = \rho_\ell (1 - b_\ell/b_k) \quad \textit{for } \ell > k. \tag{3.71}$$

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 3.5 and 3.10, 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. $\square$

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 3.5 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. (3.21). Therefore, we have that

$$\pi_k^{(k:i)} = \pi_0\frac{\overline{U}_k}{1-\overline{U}_k^{(k+1)}}\left(\frac{b_k-b_{k+1}}{b_k}\right). \tag{3.72}$$

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

$$\pi_k^{(k:ii:\ell)} = \left[\frac{\overline{U}_\ell\sum_{i=1}^k \rho_i\big((b_\ell-b_{\ell+1})/b_i\big)}{1-\overline{U}_k^{(k+1)}}\right]\left(\frac{b_k-b_{k+1}}{b_k}\right) \tag{3.73}$$

and

$$\pi_k^{(k:iii:\ell)} = \left[\frac{\rho_\ell\overline{U}_k^{(\ell)}}{1-\overline{U}_k^{(k+1)}}\right]\left(\frac{b_k-b_{k+1}}{b_k}\right). \tag{3.74}$$

It is easy to 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 3.10 that

$$\alpha_k^{(0:\ell)} = \pi_0\rho_\ell(1-b_\ell/b_k) \tag{3.75}$$

and

$$\alpha_k^{(1:\ell)} = (1-\pi_0)\rho_\ell(1-b_\ell/b_k). \tag{3.76}$$

3.6.4 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 $b_{N+1} = 0$ (as defined on p. 53), 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. (3.50) and (3.55) that

$$\widetilde{P}_{BP}^{(N)}(s) = \widetilde{\mathcal{P}}^{(acc:N)}(s; V_{p_0}^{(N)}) = \frac{\big(1 - \gamma_N^{(N+1)}\mu_{N,1}\big)\big(1 - \widetilde{V}_{p_0}^{(N)}(b_N s)\big)}{\mathbb{E}(V_{p_0}^{(N)})\big(b_N s - \gamma_N + \gamma_N \Phi_N(b_N s)\big)}. \tag{3.77}$$

The waiting time LST of the lowest priority class is readily obtained via Eqs. (3.48) and (3.49). Moreover, Eq. (3.77) 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. (3.50), (3.52), (3.55), (3.56), and (3.58).

Under a preemptive resume service discipline, Eq. (3.77) yields after some algebra the following expression for the class-N waiting time LST:

$$\widetilde{W}^{(N)}(s) = \frac{\big(s + \Lambda_{N-1}^{(N)}(1 - \psi_{N-1}(s))\big)(1 - \overline{U})}{s - \sum_{i=1}^{N} \lambda_i (b_N/b_i)\big(1 - \widetilde{B}^{(i)}\big(s + \Lambda_{N-1}^{(N)}(1 - \psi_{N-1}(s))\big)\big)}, \tag{3.78}$$

where

$$\psi_{N-1}(s) = \sum_{i=1}^{N-1} \frac{\lambda_i^{(N)}}{\Lambda_{N-1}^{(N)}} \widetilde{B}^{(i)}\big(s + \Lambda_{N-1}^{(N)}(1 - \psi_{N-1}(s))\big). \tag{3.79}$$

We remark that Eq. (3.78) 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, p. 345)).

We end Section 3.6 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 systems 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.

3.7 The PAPQ under a Bernoulli-based decision rule for the resumption of service

In this section, we consider a hybrid-based preemption discipline, which we refer to as the *Bernoulli-based decision rule for the resumption of service* (BBD-resume for short) discipline, that involves a certain combination of the traditional preemptive resume and preemptive repeat disciplines. Specifically, the decision of whether to resume the service attempt of an interrupted $\mathcal{C}_k$ is made through a Bernoulli-type experiment, where the probability that the service is resumed is given by $1 - \nu_k$ for some $0 \leq \nu_k \leq 1$. If a $\mathcal{C}_k$'s service attempt is not resumed, then the entire previously rendered service for this $\mathcal{C}_k$ is lost (or wasted), and one of the two preemptive repeat disciplines are employed (i.e., repeat-different or repeat-identical) for the next service attempt.

In what follows, we assume that the decision to resume the service of an interrupted $\mathcal{C}_k$ is made at the precise moment that the preemption occurs (and the ensuing interruption period begins). It is important to note that the decisions made at every preemption instant are independent from one another (i.e., they are iid Bernoulli

trials). Furthermore, it is clear that the exact timing of a decision has absolutely no effect on the system, so long as one is being made for each preemption (and ensuing interruption period).

For the sake of clarity in our analysis, we say that an interruption period $A^{(k)}$ is *non-resumable*, denoted by $A_+^{(k)}$, if the previously rendered service is wasted. Conversely, we say that an interruption period is *resumable*, denoted by $A_-^{(k)}$, if the service is resumed after its completion. Furthermore, we use the symbol $\mathcal{C}_+^{(int:k)}$ to denote a $\mathcal{C}^{(int:k)}$ that causes an $A_+^{(k)}$, and similarly use $\mathcal{C}_-^{(int:k)}$ to denote a $\mathcal{C}^{(int:k)}$ that causes an $A_-^{(k)}$. It is important to realize that the decision to resume service is made independently of the ensuing interruption period. As a result, it is obvious that $\widetilde{A}^{(k)}(s) = \widetilde{A}_-^{(k)}(s) = \widetilde{A}_+^{(k)}(s)$.

Our main objective in this section is to establish the class-k steady-state waiting time LST for the PAPQ under the BBD-resume discipline. Accomplishing this task is actually quite simple, as we can borrow most of the results established in the previous sections pertaining to the PAPQ under the traditional preemption disciplines. This is due to the fact that the same structural dependence among the service-structure elements that was displayed for the PAPQ of the previous sections (i.e., under the three traditional preemption disciplines) is also inherent for the PAPQ under the BBD-resume discipline. The previous observation implies that nearly all of the results, including the general recursive procedure for obtaining the steady-state waiting time LSTs derived in the previous section, apply equally to the PAPQ under the BBD-resume discipline. In fact, to complete the current analysis for the BBD-resume discipline, we need only provide updated expressions for the LSTs of $R^{(k)}$, $G^{(k)}$, and $\mathcal{P}^{(int:k)}$. We present the required results below for each combination of BBD-resume with repeat-different and BBD-resume with repeat-identical.

Repeat-different: We begin by defining the following two random variables:

$$\textit{Wasted service time } X_w^{(k)} \equiv \text{ The total amount of rendered service for a class-}k\text{ service attempt before it is interrupted by a } \mathcal{C}_+^{(int:k)}.$$

Successful service time $X_{suc}^{(k)} \equiv$ The service time of a class-k service attempt that is not interrupted by a $\mathcal{C}_{+}^{(int:k)}$.

If we let Y represent the time from the start of a class-k service attempt to the next time that an $A_{+}^{(k)}$ occurs (i.e., the next time that a $\mathcal{C}_{+}^{(int:k)}$ enters into service), then it is obvious that Y is exponentially distributed with rate $\Lambda_{k-1}^{(k)}\nu_k$. Furthermore, we understand that $X_w^{(k)} = Y|(X^{(k)} > Y)$, which readily leads to

$$\widetilde{X}_w^{(k)}(s) = \frac{\Lambda_{k-1}^{(k)}\nu_k\big(1-\widetilde{B}^{(k)}(s+\Lambda_{k-1}^{(k)}\nu_k)\big)}{\big(s+\Lambda_{k-1}^{(k)}\nu_k\big)\big(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\big)}. \tag{3.80}$$

Similarly, $X_{suc}^{(k)} = X^{(k)}|(X^{(k)} < Y)$, so that

$$\widetilde{X}_{suc}^{(k)}(s) = \frac{\widetilde{B}^{(k)}\big(s+\Lambda_{k-1}^{(k)}\nu_k\big)}{\widetilde{B}^{(k)}\big(\Lambda_{k-1}^{(k)}\nu_k\big)}. \tag{3.81}$$

It is important to note that the servicing of $X_w^{(k)}$ and $X_{suc}^{(k)}$ can both be interrupted several times for the processing of resumable interruption periods that are initiated by $\mathcal{C}_{-}^{(int:k)}$s. As a result of this observation, we define the following random intervals of time:

$H_w^{(k)} \equiv$ The time interval from the start of a $X_w^{(k)}$ to the moment that the $\mathcal{C}_k$ returns to service following the completion of the associated non-resumable interruption period $A_{+}^{(k)}$.

$H_{suc}^{(k)} \equiv$ The time interval from the start of a $X_{suc}^{(k)}$ to the departure instant of the $\mathcal{C}_k$.

It is apparent that $H_w^{(k)} = X_w^{(k)} + \sum_{i=1}^{\mathcal{N}_{-}} A_{-i}^{(k)} + A_{+}^{(k)}$, where $\mathcal{N}_{-}$ represents the number of resumable interruption periods occurring within $X_w^{(k)}$ and $A_{-i}^{(k)}$ is the i-th resumable interruption period. Conditioning on $X_w^{(k)} = x$, $\mathcal{N}_{-}$ has a Poisson distribution with mean $\Lambda_{k-1}^{(k)}(1-\nu_k)x$, and this ultimately leads to

$$\widetilde{H}_w^{(k)}(s) = \widetilde{A}_{+}^{(k)}(s)\widetilde{X}_w^{(k)}\big(s+\Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_{-}^{(k)}(s))\big). \tag{3.82}$$

For similar reasons, we also obtain

$$\widetilde{H}_{suc}^{(k)}(s) = \widetilde{X}_{suc}^{(k)}\big(s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_{-}^{(k)}(s))\big). \tag{3.83}$$

To obtain the LST of $R^{(k)}$, let $\mathcal{N}$ be the number of wasted service attempts that a $\mathcal{C}_k$ experiences before departing the system. Observe that the probability that a class-k service attempt is wasted is given by $P(X^{(k)} < Y) = 1 - \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)$, which implies that $\mathcal{N}$ has a geometric distribution with mean $\big(1 - \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\big)/\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)$. Conditional on $\mathcal{N} = n$, it is clear that $R^{(k)}$ is comprised of n periods of time, each having LST $\widetilde{H}_{w}^{(k)}$, and one period of time having LST $\widetilde{H}_{suc}^{(k)}$. Furthermore, these $n+1$ intervals of time are all mutually independent, so that $\mathbb{E}(e^{-sR^{(k)}}|\mathcal{N} = n) = \big(\widetilde{H}_{w}^{(k)}(s)\big)^n \widetilde{H}_{suc}^{(k)}(s)$. Removing the condition on $\mathcal{N}$ and using Eqs. (3.82) and (3.83) yields

$$\widetilde{R}^{(k)}(s) = \frac{\widetilde{X}_{suc}^{(k)}\big(s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_{-}^{(k)}(s)\big)\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)}{1 - \widetilde{A}_{+}^{(k)}(s)\widetilde{X}_{w}^{(k)}\big(s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_{-}^{(k)}(s))\big)(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}))}. \tag{3.84}$$

Substituting Eqs. (3.80) and (3.81) into Eq. (3.84), and using the fact that $\widetilde{A}^{(k)}(s) = \widetilde{A}_{-}^{(k)}(s) = \widetilde{A}_{+}^{(k)}(s)$, leads to an alternate expression of $\widetilde{R}^{(k)}(s)$, namely

$$\widetilde{R}^{(k)}(s) = \frac{\omega_k(s)\widetilde{B}^{(k)}\big(\omega_k(s)\big)}{s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(s)\big(1 - \nu_k\widetilde{B}^{(k)}\big(\omega_k(s)\big)\big)}, \tag{3.85}$$

where $\omega_k(s) = s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}(1-\nu_k)\widetilde{A}^{(k)}(s)$. The first two moments of $R^{(k)}$ are obtained via differentiation of Eq. (3.85):

$$\mathbb{E}(R^{(k)}) = (1 + \Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\mathbb{E}(G^{(k)}) \tag{3.86}$$

and

$$\begin{aligned}\mathbb{E}\big((R^{(k)})^2\big) &= \Lambda_{k-1}^{(k)}\mathbb{E}\big((A^{(k)})^2\big)\mathbb{E}(G^{(k)}) \\ &\quad + (1 + \Lambda_{k-1}^{(k)}(1-\nu_k)\mathbb{E}(A^{(k)}))(1 + \Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\mathbb{E}\big((G^{(k)})^2\big) \\ &\quad + 2\Lambda_{k-1}^{(k)}\nu_k\mathbb{E}(A^{(k)})(1 + \Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\big(\mathbb{E}(G^{(k)})\big)^2. \end{aligned} \tag{3.87}$$

A simple expression for the LST of $G^{(k)}$ is obtained by substituting $\widetilde{A}^{(k)}(s) = 1$ into Eq. (3.85), thereby leading to

$$\widetilde{G}^{(k)}(s) = \frac{\widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)}\nu_k)(s + \Lambda_{k-1}^{(k)}\nu_k)}{s + \Lambda_{k-1}^{(k)}\nu_k\widetilde{B}^{(k)}(s + \Lambda_{k-1}^{(k)}\nu_k)}. \tag{3.88}$$

The first two moments of $G^{(k)}$ are

$$\mathbb{E}(G^{(k)}) = \frac{1 - \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)}{\Lambda_{k-1}^{(k)}\nu_k\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)} \tag{3.89}$$

and

$$\mathbb{E}\big((G^{(k)})^2\big) = \frac{2\big[1 - \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k) - \Lambda_{k-1}^{(k)}\nu_k\mathbb{E}\big(X^{(k)}e^{-\Lambda_{k-1}^{(k)}\nu_k X^{(k)}}\big)\big]}{\big(\Lambda_{k-1}^{(k)}\nu_k\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\big)^2}. \tag{3.90}$$

All that remains is the derivation of the LST of $\mathcal{P}^{(int:k)}$. For the current model, it is necessary to condition on the type of interrupting customer that we are dealing with. In particular, if we let $\mathcal{P}_-^{(int:k)}$ and $\mathcal{P}_+^{(int:k)}$ denote the additional accumulated priority (after having accumulated the intial priority level) for a $\mathcal{C}_-^{(int:k)}$ and $\mathcal{C}_+^{(int:k)}$, respectively, then obviously

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = (1 - \nu_k)\widetilde{\mathcal{P}}_-^{(int:k)}(s) + \nu_k\widetilde{\mathcal{P}}_+^{(int:k)}(s). \tag{3.91}$$

We derive the LSTs of $\mathcal{P}_-^{(int:k)}$ and $\mathcal{P}_+^{(int:k)}$ in a similar fashion to our derivation of $\mathcal{P}^{(int:k)}$ for the PAPQ under the preemptive repeat-different discipline (i.e., see Section 3.6.2). First of all, we consider only the times that an $R^{(k)}$ is in progress and say that the system is in state m if the oldest $\mathcal{C}_k$ has already experienced $m \geq 0$ failed service attempts. It can be shown, using the same techniques as before, that $P_m = \mathbb{P}(\text{state } m \,|R^{(k)} \text{ in progress}) = \big(1 - \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\big)^m\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)$.

Next, observe that if either a $\mathcal{C}_-^{(int:k)}$ or a $\mathcal{C}_+^{(int:k)}$ preempts a $\mathcal{C}_k$ while the system is in state m, then it implies that the ongoing residence period has already experienced m independent $H_w^{(k)}$ periods of time. Now, for a $\mathcal{C}_-^{(int:k)}$, we must also consider whether the service attempt which it is interrupting is a wasted one or a successful one. In particular, we have the following:

$$\mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}} \mid \text{state } m \text{ and interrupt a } X_w^{(k)}) = \big(\widetilde{H}_w^{(k)}(b_k s)\big)^m \widetilde{X}_{w,past}^{(k)}\big(b_k s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s))\big) \quad (3.92)$$

and

$$\mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}} \mid \text{state } m \text{ and interrupt a } X_{suc}^{(k)}) = \big(\widetilde{H}_w^{(k)}(b_k s)\big)^m \widetilde{X}_{suc,past}^{(k)}\big(b_k s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s))\big), \quad (3.93)$$

where, in general,

$$\widetilde{Z}_{past}(s) = \frac{1-\widetilde{Z}(s)}{s\mathbb{E}(Z)} \quad (3.94)$$

for a given random variable Z. If we let σ_k denote the probability that a $\mathcal{C}^{(int:k)}$ interrupts a wasted service time, then it must be that

$$\mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}} \mid \text{state } m) = \sigma_k \mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}} \mid \text{state } m \text{ and interrupt a } X_w^{(k)}) + (1-\sigma_k)\mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}} \mid \text{state } m \text{ and interrupt a } X_{suc}^{(k)}). \quad (3.95)$$

It can be shown, from semi-Markov theory (e.g., see Kao (1996, Section 6.2)) and the PASTA property, that

$$\sigma_k = \frac{(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k))\mathbb{E}(X_w^{(k)})}{(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k))\mathbb{E}(X_w^{(k)}) + \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\mathbb{E}(X_{suc}^{(k)})}. \quad (3.96)$$

Therefore, by substituting Eq. (3.96) along with Eqs. (3.92) and (3.93) into Eq. (3.95), we ultimately obtain

$$\mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}} \mid \text{state } m) = \big(\widetilde{H}_w^{(k)}(b_k s)\big)^m \widetilde{X}_{*,past}^{(k)}\big(b_k s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s))\big), \quad (3.97)$$

where $\widetilde{X}_*^{(k)}(s) = (1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k))\widetilde{X}_w^{(k)}(s) + \widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\widetilde{X}_{suc}^{(k)}(s)$. Finally, removing the condition of the system being in state m yields after some algebra

$$\widetilde{\mathcal{P}}_-^{(int:k)}(s) = \frac{\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\widetilde{X}_{*past}^{(k)}\big(b_k s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s))\big)}{1-\widetilde{H}_w^{(k)}(b_k s)\big(1-\widetilde{B}^{(k)}(\Lambda_{k-1}^{(k)}\nu_k)\big)}. \quad (3.98)$$

Furthermore, it is straightforward but tedious to show that

$$\widetilde{X}^{(k)}_{*past}\big(b_k s + \Lambda^{(k)}_{k-1}(1-\nu_k)(1-\widetilde{A}^{(k)}_{-}(b_k s))\big) = \frac{\big(1-\widetilde{B}^{(k)}\big(\omega_k(b_k s)\big)\big)\Lambda^{(k)}_{k-1}\nu_k}{\omega_k(b_k s)(1-\widetilde{B}^{(k)}(\Lambda^{(k)}_{k-1}\nu_k))}. \quad (3.99)$$

Substituting Eq. (3.99) into Eq. (3.98) ultimately yields the following simplified expression for $\widetilde{\mathcal{P}}^{(int:k)}_{-}(s)$:

$$\widetilde{\mathcal{P}}^{(int:k)}_{-}(s) = \frac{1-\widetilde{B}^{(k)}\big(\omega_k(b_k s)\big)}{\mathbb{E}(G^{(k)})\big(b_k s + \Lambda^{(k)}_{k-1} - \Lambda^{(k)}_{k-1}\widetilde{A}^{(k)}(b_k s)\big(1-\widetilde{B}^{(k)}\big(\omega_k(b_k s)\big)\big)\big)}. \quad (3.100)$$

We can similarly find the LST of $\mathcal{P}^{(int:k)}_{+}$. In particular, if a $\mathcal{C}^{(int:k)}_{+}$ causes an interruption while the system is in state m, then in addition to the m previously experienced $H^{(k)}_w$ periods of time, the ongoing residence period has also (most recently) experienced one full $X^{(k)}_w$ along with all of the resumable interruption periods $A^{(k)}_{-}$ occurring within it. Hence, it must be that

$$\mathbb{E}(e^{-s\mathcal{P}^{(int:k)}_{+}} \mid \text{state } m) = \big(\widetilde{H}^{(k)}_w(b_k s)\big)^m \widetilde{X}^{(k)}_w\big(b_k s + \Lambda^{(k)}_{k-1}(1-\nu_k)(1-\widetilde{A}^{(k)}_{-}(b_k s))\big). \quad (3.101)$$

Therefore,

$$\widetilde{\mathcal{P}}^{(int:k)}_{+}(s) = \frac{\widetilde{B}^{(k)}(\Lambda^{(k)}_{k-1}\nu_k)\widetilde{X}^{(k)}_w\big(b_k s + \Lambda^{(k)}_{k-1}(1-\nu_k)(1-\widetilde{A}^{(k)}_{-}(b_k s))\big)}{1-\widetilde{H}^{(k)}_w(b_k s)\big(1-\widetilde{B}^{(k)}(\Lambda^{(k)}_{k-1}\nu_k)\big)}. \quad (3.102)$$

It is quite straightforward to show that Eq. (3.102) can be simplified to become

$$\widetilde{\mathcal{P}}^{(int:k)}_{+}(s) = \frac{1-\widetilde{B}^{(k)}\big(\omega_k(b_k s)\big)}{\mathbb{E}(G^{(k)})\big(b_k s + \Lambda^{(k)}_{k-1} - \Lambda^{(k)}_{k-1}\widetilde{A}^{(k)}(b_k s)\big(1-\widetilde{B}^{(k)}\big(\omega_k(b_k s)\big)\big)\big)}, \quad (3.103)$$

which is equivalent to $\widetilde{\mathcal{P}}^{(int:k)}_{-}(s)$ as given by Eq. (3.100). Therefore, it must be that

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = \widetilde{\mathcal{P}}^{(int:k)}_{-}(s) = \widetilde{\mathcal{P}}^{(int:k)}_{+}(s), \quad (3.104)$$

as given by Eq. (3.100) or Eq. (3.103). We obtain the first moment of $\mathcal{P}^{(int:k)}$ through the differentiation of its LST, leading to

$$\mathbb{E}(\mathcal{P}^{(int:k)}) = b_k\left(\big(1+\Lambda^{(k)}_{k-1}(1-\nu_k)\mathbb{E}(A^{(k)}_{-})\big)\frac{\mathbb{E}\big((G^{(k)})^2\big)}{2\mathbb{E}(G^{(k)})} + \Lambda^{(k)}_{k-1}\nu_k\mathbb{E}(A^{(k)}_{+})\mathbb{E}(G^{(k)})\right). \quad (3.105)$$

Repeat-identical: The arguments used for establishing the LSTs of $R^{(k)}$ and $G^{(k)}$ for the repeat-different case are also applicable to the repeat-identical case, with the exception that we now must condition on the originally sampled service time (since this service time is simply restarted after each wasted service attempt). For instance, given that $X^{(k)} = x$, the conditional pdf of $X_w^{(k)}$ is expressible as

$$\mathbb{P}(p \leq X_w^{(k)} \leq p + \mathrm{d}p | X^{(k)} = x) = \frac{\Lambda_{k-1}^{(k)} \nu_k e^{-\Lambda_{k-1}^{(k)} \nu_k p}}{1 - e^{-\Lambda_{k-1}^{(k)} \nu_k x}} \mathrm{d}p, \quad p < x. \tag{3.106}$$

Therefore,

$$\mathbb{E}(e^{-sX_w^{(k)}} | X^{(k)} = x) = \frac{\Lambda_{k-1}^{(k)} \nu_k \big(1 - e^{-(s + \Lambda_{k-1}^{(k)} \nu_k)x}\big)}{\big(s + \Lambda_{k-1}^{(k)} \nu_k\big)\big(1 - e^{-\Lambda_{k-1}^{(k)} \nu_k x}\big)}. \tag{3.107}$$

On the other hand, since each $\mathcal{C}_k$ eventually departs the system, it must be that

$$\mathbb{E}(e^{-sX_{suc}^{(k)}} | X^{(k)} = x) = e^{-sx}. \tag{3.108}$$

From these results, it is also straightforward to show that

$$\mathbb{E}(e^{-sH_w^{(k)}} | X^{(k)} = x) = \frac{\Lambda_{k-1}^{(k)} \nu_k \big(1 - e^{-\omega_k(s)x}\big)}{\omega_k(s)\big(1 - e^{-\Lambda_{k-1}^{(k)} \nu_k x}\big)} \widetilde{A}_+^{(k)}(s) \tag{3.109}$$

and

$$\mathbb{E}(e^{-sH_{suc}^{(k)}} | X^{(k)} = x) = e^{-(s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(s)))x}. \tag{3.110}$$

Moreover, if the service time of a $\mathcal{C}_k$ is $X^{(k)} = x$, then the number of failed service attempts that this $\mathcal{C}_k$ experiences before departing the system has a geometric distribution with mean $(1 - e^{-\Lambda_{k-1}^{(k)} \nu_k x})/e^{-\Lambda_{k-1}^{(k)} \nu_k x}$. Applying similar arguments to those made in the repeat-different case ultimately yields

$$\widetilde{R}^{(k)}(s) = \int_{x=0}^{\infty} \frac{\omega_k(s) e^{-\omega_k(s)x}}{s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)} \widetilde{A}^{(k)}(s)(1 - \nu_k e^{-\omega_k(s)x})} \mathrm{d}B^{(k)}(x). \tag{3.111}$$

The first two moments of $R^{(k)}$ are

$$\mathbb{E}(R^{(k)}) = (1 + \Lambda_{k-1}^{(k)} \mathbb{E}(A^{(k)})) \mathbb{E}(G^{(k)}) \tag{3.112}$$

and

$$\begin{aligned}\mathbb{E}\big((R^{(k)})^2\big) &= \Lambda_{k-1}^{(k)}\mathbb{E}\big((A^{(k)})^2\big)\mathbb{E}(G^{(k)}) \\ &\quad + (1+\Lambda_{k-1}^{(k)}(1-\nu_k)\mathbb{E}(A^{(k)}))(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\mathbb{E}\big((G^{(k)})^2\big) \\ &\quad + \frac{2}{\Lambda_{k-1}^{(k)}\nu_k}\mathbb{E}(A^{(k)})(1+\Lambda_{k-1}^{(k)}\mathbb{E}(A^{(k)}))\mathbb{E}\big((e^{\Lambda_{k-1}^{(k)}\nu_k X^{(k)}}-1)^2\big). \quad (3.113)\end{aligned}$$

Also, substituting $\widetilde{A}^{(k)}(s) = 1$ into Eq. (3.111) yields

$$\widetilde{G}^{(k)}(s) = \int_{x=0}^{\infty} \frac{(s+\Lambda_{k-1}^{(k)}\nu_k)e^{-(s+\Lambda_{k-1}^{(k)}\nu_k)x}}{s+\Lambda_{k-1}^{(k)}\nu_k e^{-(s+\Lambda_{k-1}^{(k)}\nu_k)x}}\mathrm{d}B^{(k)}(x), \quad (3.114)$$

from which we obtain

$$\mathbb{E}(G^{(k)}) = \mathbb{E}\big(\mathbb{E}(G^{(k)}|X^{(k)}=x)\big) = \mathbb{E}\left(\frac{e^{\Lambda_{k-1}^{(k)}\nu_k X^{(k)}}-1}{\Lambda_{k-1}^{(k)}\nu_k}\right) = \frac{\widetilde{B}^{(k)}(-\Lambda_{k-1}^{(k)}\nu_k)-1}{\Lambda_{k-1}^{(k)}\nu_k} \quad (3.115)$$

and

$$\begin{aligned}\mathbb{E}\big((G^{(k)})^2\big) &= \mathbb{E}\big[\mathbb{E}\big((G^{(k)})^2|X^{(k)}=x\big)\big] \\ &= \mathbb{E}\left[\frac{2}{(\Lambda_{k-1}^{(k)}\nu_k)^2}\big(e^{2\Lambda_{k-1}^{(k)}\nu_k X^{(k)}} - e^{\Lambda_{k-1}^{(k)}\nu_k X^{(k)}} - \Lambda_{k-1}^{(k)}\nu_k X^{(k)}e^{\Lambda_{k-1}^{(k)}\nu_k X^{(k)}}\big)\right] \\ &= \frac{2[\widetilde{B}^{(k)}(-2\Lambda_{k-1}^{(k)}\nu_k) - \widetilde{B}^{(k)}(-\Lambda_{k-1}^{(k)}\nu_k) - \Lambda_{k-1}^{(k)}\nu_k\mathbb{E}(X^{(k)}e^{\Lambda_{k-1}^{(k)}\nu_k X^{(k)}})]}{(\Lambda_{k-1}^{(k)}\nu_k)^2}. \quad (3.116)\end{aligned}$$

To find our final required result, namely $\widetilde{\mathcal{P}}^{(int:k)}(s)$, we again must condition on the originally sampled service time of the interrupted $\mathcal{C}_k$. Let $X_*^{(k)}$ denote such a service time. Similar to the repeat-different case, we first find $\mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}}|X_*^{(k)}=x)$ and $\mathbb{E}(e^{-s\mathcal{P}_+^{(int:k)}}|X_*^{(k)}=x)$, which also end up being equivalent to one another. Note that if we consider only the times that a $R^{(k)}$ (with the service time of the associated $\mathcal{C}_k$ being equal to x) is in progress, then it can be shown, via similar methods as before (i.e., see Section 3.6.2), that the probability that the system is in state m (i.e., the oldest $\mathcal{C}_k$ has suffered m previous interruptions) is given by $(1-e^{-\Lambda_{k-1}^{(k)}\nu_k x})^m e^{-\Lambda_{k-1}^{(k)}\nu_k x}$.

Recall that a $\mathcal{C}_-^{(int:k)}$ can interrupt either a wasted service time $X_w^{(k)}$ or a successful service time $X_{suc}^{(k)}$. Conditioning on $X_*^{(k)} = x$ and the system being in state m yields

$$\mathbb{E}\big(e^{-s\mathcal{P}_-^{(int:k)}}|X_*^{(k)} = x, \text{state } m, \text{ and interrupt a } X_w^{(k)}\big) = \big(\mathbb{E}(e^{-b_k s H_w^{(k)}}|X^{(k)} = x)\big)^m \mathbb{E}\big(e^{-(b_k s+\Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s)))X_{w,past}^{(k)}}|X^{(k)} = x\big) \quad (3.117)$$

and

$$\mathbb{E}\big(e^{-s\mathcal{P}_-^{(int:k)}}|X_*^{(k)} = x, \text{state } m, \text{ and interrupt a } X_{suc}^{(k)}\big) = \big(\mathbb{E}(e^{-b_k s H_w^{(k)}}|X^{(k)} = x)\big)^m \mathbb{E}\big(e^{-(b_k s+\Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s)))X_{suc,past}^{(k)}}|X^{(k)} = x\big). \quad (3.118)$$

By removing the condition of the system state and after some algebra, we obtain

$$\mathbb{E}\big(e^{-s\mathcal{P}_-^{(int:k)}}|X_*^{(k)} = x \text{ and interrupt a } X_w^{(k)}\big) = \frac{\omega_k(b_k s)(1-e^{-\Lambda_{k-1}^{(k)}\nu_k x}) - \Lambda_{k-1}^{(k)}\nu_k(1-e^{-\omega_k(b_k s)x})}{\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_k s)(1-\nu_k e^{-\omega_k(b_k s)x})\big)\mathbb{E}(X_w^{(k)}|X^{(k)} = x)} \times \frac{1}{(e^{\Lambda_{k-1}^{(k)}\nu_k x}-1)\big(b_k s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s))\big)} \quad (3.119)$$

and (since $X_{suc}^{(k)} = x$)

$$\mathbb{E}\big(e^{-s\mathcal{P}_-^{(int:k)}}|X_*^{(k)} = x \text{ and interrupt a } X_{suc}^{(k)}\big) = \frac{\omega_k(b_k s)\big(e^{-\Lambda_{k-1}^{(k)}\nu_k x} - e^{-\omega_k(b_k s)x}\big)}{x\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)}\widetilde{A}^{(k)}(b_k s)(1-\nu_k e^{-\omega_k(b_k s)x})\big)} \times \frac{1}{b_k s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1-\widetilde{A}_-^{(k)}(b_k s))}. \quad (3.120)$$

The probability that a $\mathcal{C}_-^{(int:k)}$ interrupts a wasted service time can be found using semi-Markov theory (e.g., see Kao (1996, Section 6.2)) as well as the PASTA property, and is given by

$$\sigma_{k,x} = \frac{(1-e^{-\Lambda_{k-1}^{(k)}\nu_k x})\mathbb{E}(X_w^{(k)}|X^{(k)} = x)}{(1-e^{-\Lambda_{k-1}^{(k)}\nu_k x})\mathbb{E}(X_w^{(k)}|X^{(k)} = x) + xe^{-\Lambda_{k-1}^{(k)}\nu_k x}}, \quad (3.121)$$

which simplifies to become $\sigma_{k,x} = \Lambda_{k-1}^{(k)} \nu_k \mathbb{E}(X_w^{(k)} | X^{(k)} = x)$. Furthermore, it is straightforward to show that $1 - \sigma_{k,x} = x / \mathbb{E}(G^{(k)} | X^{(k)} = x)$. Therefore,

$$\begin{aligned}\mathbb{E}(e^{-s\mathcal{P}_-^{(int:k)}} | X_*^{(k)} = x) &= \sigma_{k,x} \mathbb{E}\big(e^{-s\mathcal{P}_-^{(int:k)}} | X_*^{(k)} = x \text{ and interrupt a } X_w^{(k)}\big) \\ &\quad + (1 - \sigma_{k,x}) \mathbb{E}\big(e^{-s\mathcal{P}_-^{(int:k)}} | X_*^{(k)} = x \text{ and interrupt a } X_{suc}^{(k)}\big) \\ &= \frac{1 - e^{-\omega_k (b_k s) x}}{\mathbb{E}(G^{(k)} | X^{(k)} = x)\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)} \widetilde{A}^{(k)}(b_k s)(1 - \nu_k e^{-\omega_k (b_k s)})\big)}. \end{aligned} \tag{3.122}$$

For the additional accumulated priority of a $\mathcal{C}_+^{(int:k)}$, we understand that

$$\begin{aligned}\mathbb{E}(e^{-s\mathcal{P}_+^{(int:k)}} | X_*^{(k)} = x \text{ and state } m) &= \big((1 - e^{-\Lambda_{k-1}^{(k)} \nu_k x}) \mathbb{E}(e^{-b_k s H_w^{(k)}} | X^{(k)} = x)\big)^m \\ &\quad \times e^{-\Lambda_{k-1}^{(k)} \nu_k x} \mathbb{E}\big(e^{-(b_k s + \Lambda_{k-1}^{(k)}(1-\nu_k)(1 - \widetilde{A}_-^{(k)}(b_k s)))X_w^{(k)}} | X^{(k)} = x\big). \end{aligned} \tag{3.123}$$

Again, by removing the condition of the system being in state m and after some algebra, we obtain

$$\mathbb{E}(e^{-s\mathcal{P}_+^{(int:k)}} | X_*^{(k)} = x) = \frac{1 - e^{-\omega_k (b_k s) x}}{\mathbb{E}(G^{(k)} | X^{(k)} = x)\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)} \widetilde{A}^{(k)}(b_k s)(1 - \nu_k e^{-\omega_k (b_k s)})\big)}, \tag{3.124}$$

which is identical to Eq. (3.122). Therefore, it must be that $\widetilde{\mathcal{P}}^{(int:k)}(s) = \widetilde{\mathcal{P}}_-^{(int:k)}(s) = \widetilde{\mathcal{P}}_+^{(int:k)}(s)$. It is important to realize here that $X_*^{(k)}$ does not have df $B^{(k)}(x)$ since we are making the underlying assumption here that an interruption has occurred. As for the traditional repeat-identical case, it can be shown that

$$\mathbb{P}(x < X_*^{(k)} \leq x + \mathrm{d}x) = \frac{\mathbb{E}(G^{(k)} | X^{(k)} = x)}{\mathbb{E}(G^{(k)})} \mathrm{d}B^{(k)}(x). \tag{3.125}$$

The above result has the following intuitive interpretation: the probability that a $\mathcal{C}^{(int:k)}$ interrupts a $\mathcal{C}_k$ with service time x is proportional to $\mathbb{E}(G^{(k)} | X^{(k)} = x)$ as well as to the relative occurrence of such a service time given by $\mathrm{d}B^{(k)}(x)$. The denominator is simply the normalization factor. Nevertheless, we therefore obtain

$$\widetilde{\mathcal{P}}^{(int:k)}(s) = \int_{x=0}^{\infty} \frac{1 - e^{-\omega_k (b_k s) x}}{\mathbb{E}(G^{(k)})\big(b_k s + \Lambda_{k-1}^{(k)} - \Lambda_{k-1}^{(k)} \widetilde{A}^{(k)}(b_k s)(1 - \nu_k e^{-\omega_k (b_k s)})\big)} \mathrm{d}B^{(k)}(x). \tag{3.126}$$

The associated first moment, obtained through differentiation of Eq. (3.126), works out to be

$$\mathbb{E}(\mathcal{P}^{(int:k)}) = b_k \Bigg(\frac{\mathbb{E}\big((G^{(k)})^2\big)}{2\mathbb{E}(G^{(k)})} \big(1 + \Lambda_{k-1}^{(k)} \mathbb{E}(A^{(k)})\big) + \frac{\nu_k \Lambda_{k-1}^{(k)} \mathbb{E}(A^{(k)})}{\mathbb{E}(G^{(k)})} \int_{x=0}^{\infty} \frac{\Lambda_{k-1}^{(k)} \nu_k x e^{\Lambda_{k-1}^{(k)} \nu_k x} + 1 - e^{\Lambda_{k-1}^{(k)} \nu_k x}}{\big(\Lambda_{k-1}^{(k)} \nu_k\big)^2} \mathrm{d}B^{(k)}(x) \Bigg). \quad (3.127)$$

We close this subsection by stating that the BBD-resume discipline captures all three traditional preemption disciplines. In particular, when $\nu_k = 0$, every $A^{(k)}$ is resumable, and thus the BBD-resume discipline becomes the traditional preemptive resume discipline. On the other hand, when $\nu_k = 1$, every $A^{(k)}$ is non-resumable, implying that the BBD-resume with repeat-different (repeat-identical) discipline is equivalent to the traditional repeat-different (repeat-identical) discipline.

3.8 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* (PRWET) discipline was analyzed. The accumulating priority rates are arranged as follows:

$$b_1 = 1, \quad b_2 = e^{-x}, \quad \text{and} \quad b_3 = e^{-2x} \quad \text{for some } x \geq 0. \quad (3.128)$$

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 traditional preemption disciplines. The results are reported to 4 decimal places of accuracy in Tables 3.2 and 3.3. 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

$$\widehat{N}^{(k)}(z) = \mathbb{E}(z^{N^{(k)}}) = \widetilde{W}^{(k)}\big(\lambda_k(1-z)\big). \tag{3.129}$$

Table 3.4 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. (3.128) 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 3.2–3.4), 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 3.2–3.4, the results are approaching those of the limiting FCFS $M/G/1$ queue (as described in Section 3.6.4), and these results are consistent under all three preemption disciplines.

In the second part of this example, we analyze the same 3-class PAPQ model, but now under the BBD-resume discipline. Recall, from the previous section, that the BBD-resume discipline leads to system performance that is essentially a balance between the system performances of the PAPQ under the traditional preemptive resume and repeat (-different or -identical) disciplines. To illustrate this fact, we report to 4 decimal places of accuracy the mean flow times associated with the PAPQ under a BBD-resume with repeat-different and with repeat-identical disciplines in Tables 3.5 and 3.6, respectively.

Note that the mean flow times reported in Tables 3.5 and 3.6 correspond to a PAPQ under a BBD-resume discipline with $\nu_k = \mathbb{P}(A^{(k)} > T_k)$ for some $T_k \geq 0$. In doing this, the BBD-resume discipline can be viewed as an approximation to the PRWET discipline, for which a class-k interruption period is non-resumable if it is longer than T_k units of time. Hence, the probability that a class-k interruption period is non-resumable under the PRWET discipline is also given by ν_k. However, recall that the classification of an interruption period as being non-resumable under

Table 3.2: 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 3.3: 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 3.4: 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

the BBD-resume discipline is made completely at random, thus independent of the duration of that interruption period. Furthermore, under the BBD-resume discipline, the probability that an interruption period is both non-resumable and greater than T_k is equal to ν_k^2. As a result, we expect the approximation of the PRWET through the BBD-resume to be better for $\nu_k \approx 0$ or $\nu_k \approx 1$. In comparing the values found in row $x = 100$ of Tables 3.5 and 3.6 to the appropriate values reported in Drekic (2003, Tables 1 and 2), we see that, for this 3-class model, the BBD-resume discipline reasonably approximates the PRWET discipline.

It should be noted here that in order to compute the probabilities corresponding to ν_k, we implement the recursive-based method outlined in Abate and Whitt (1992) coupled with the two numerical inversion methods found 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. In all of our examples, we employed the EULER and POST-WIDDER methods using the authors' suggested parameter settings (see the Appendix for a brief overview of these methods) and found that

both methods produced equivalent results.

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

$$\frac{\lambda_1 \mathbb{E}(F^{(1)}) + \lambda_2 \mathbb{E}(F^{(2)})}{\Lambda_2}$$

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.

In our investigation, we consider the same model as Conway et al. (1967) with the exception that priority is assigned according to Eq. (3.1). 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 defined by Eq. (3.48). To conduct the numerical inversion, we again employ the EULER and POST-WIDDER methods of Abate and Whitt (1995) 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. Figures 3.5 and 3.6 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 3.7, 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, we compare in Table 3.8 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 3.7 and

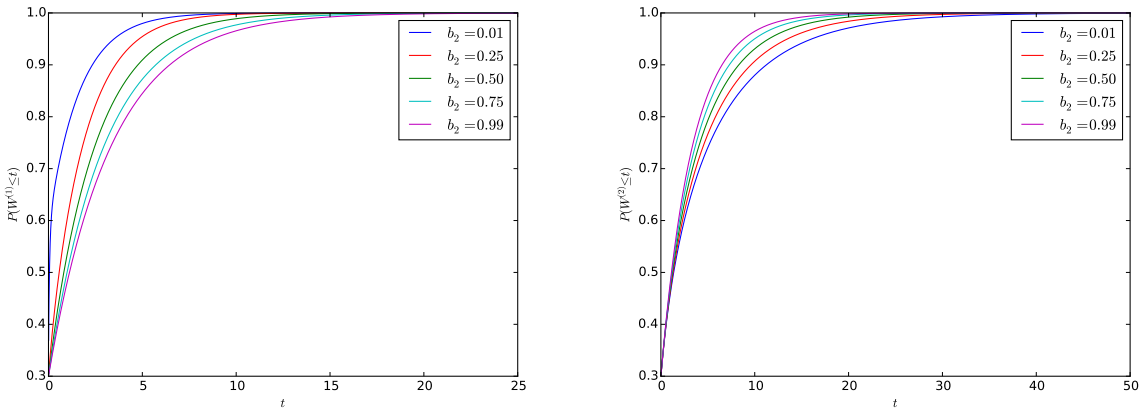
Table 3.5: Expected flow times for PAPQ in Example 1 under BBD-resume with repeat-different

x	$(T_2,T_3)=(6,6)$				$(T_2,T_3)=(2,2)$			
	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$
100.0000	1.8333	3.9813	10.7332	0.7153	1.8333	4.0290	11.5038	0.7259
10.0000	1.8334	3.9814	10.7329	0.7153	1.8334	4.0291	11.5034	0.7259
7.5000	1.8340	3.9829	10.7299	0.7153	1.8340	4.0308	11.4999	0.7259
5.0000	1.8416	4.0003	10.6939	0.7152	1.8418	4.0514	11.4571	0.7258
2.5000	1.9557	4.2006	10.2412	0.7148	1.9615	4.2849	10.9223	0.7248
1.0000	2.6641	4.7361	8.4434	0.7130	2.7112	4.8729	8.8413	0.7203
0.7500	2.9880	4.8471	7.8090	0.7123	3.0469	4.9788	8.1185	0.7186
0.5000	3.4235	4.9341	7.0429	0.7116	3.4879	5.0454	7.2524	0.7163
0.2500	3.9878	4.9726	6.1565	0.7107	4.0392	5.0413	6.2594	0.7135
0.1000	4.3916	4.9617	5.5815	0.7103	4.4182	4.9923	5.6212	0.7114
0.0100	4.6570	4.9406	5.2271	0.7100	4.6600	4.9438	5.2310	0.7101
0.0010	4.6844	4.9378	5.1915	0.7100	4.6847	4.9381	5.1918	0.7100
0.0001	4.6872	4.9375	5.1879	0.7100	4.6872	4.9376	5.1879	0.7100
0.0000	4.6875	4.9375	5.1875	0.7100	4.6875	4.9375	5.1875	0.7100
x	$(T_2,T_3)=(0.5,0.5)$				$(T_2,T_3)=(0.001,0.001)$			
	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$
100.0000	1.8333	4.1082	13.1145	0.7457	1.8333	4.3166	15.9215	0.7734
10.0000	1.8334	4.1084	13.1141	0.7457	1.8334	4.3169	15.9209	0.7734
7.5000	1.8340	4.1106	13.1094	0.7457	1.8341	4.3202	15.9141	0.7734
5.0000	1.8422	4.1382	13.0522	0.7455	1.8433	4.3593	15.8304	0.7732
2.5000	1.9729	4.4463	12.3451	0.7434	1.9962	4.7887	14.8050	0.7699
1.0000	2.8099	5.1558	9.6765	0.7345	2.9844	5.6663	11.0593	0.7552
0.7500	3.1714	5.2541	8.7723	0.7309	3.3843	5.7344	9.8314	0.7488
0.5000	3.6257	5.2812	7.7005	0.7259	3.8519	5.6754	8.4033	0.7400
0.2500	4.1518	5.1905	6.4854	0.7192	4.3259	5.4255	6.8229	0.7276
0.1000	4.4780	5.0604	5.7112	0.7140	4.5662	5.1626	5.8401	0.7178
0.0100	4.6671	4.9511	5.2399	0.7104	4.6772	4.9618	5.2523	0.7108
0.0010	4.6855	4.9389	5.1927	0.7100	4.6865	4.9399	5.1940	0.7101
0.0001	4.6873	4.9376	5.1880	0.7100	4.6874	4.9377	5.1881	0.7100
0.0000	4.6875	4.9375	5.1876	0.7100	4.6875	4.9375	5.1876	0.7100

Table 3.6: Expected flow times for PAPQ in Example 1 under BBD-resume with repeat-identical

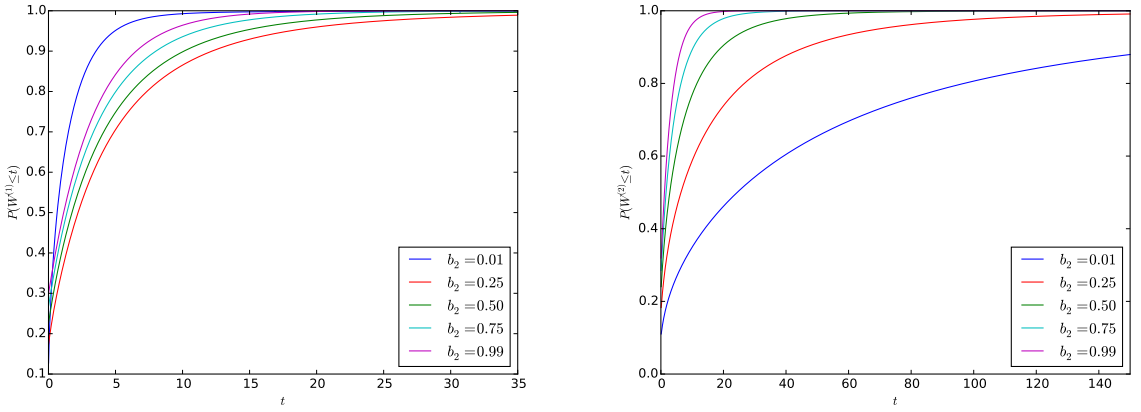
x	$(T_2,T_3)=(6,6)$				$(T_2,T_3)=(2,2)$			
	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$
100.0000	1.8333	4.0364	11.2813	0.7219	1.8333	4.2110	13.6522	0.7478
10.0000	1.8334	4.0366	11.2810	0.7219	1.8334	4.2112	13.6517	0.7478
7.5000	1.8340	4.0382	11.2775	0.7219	1.8341	4.2136	13.6461	0.7477
5.0000	1.8418	4.0574	11.2348	0.7218	1.8428	4.2424	13.5780	0.7475
2.5000	1.9607	4.2754	10.7036	0.7207	1.9804	4.5580	12.7459	0.7448
1.0000	2.6928	4.8211	8.6639	0.7166	2.8300	5.2210	9.7534	0.7337
0.7500	3.0203	4.9209	7.9655	0.7152	3.1850	5.2913	8.7864	0.7296
0.5000	3.4542	4.9883	7.1358	0.7135	3.6261	5.2874	7.6700	0.7243
0.2500	4.0079	5.0002	6.1941	0.7116	4.1380	5.1755	6.4447	0.7177
0.1000	4.4005	4.9722	5.5938	0.7106	4.4656	5.0476	5.6878	0.7132
0.0100	4.6579	4.9416	5.2282	0.7101	4.6652	4.9494	5.2371	0.7103
0.0010	4.6845	4.9379	5.1916	0.7100	4.6853	4.9387	5.1925	0.7100
0.0001	4.6872	4.9375	5.1879	0.7100	4.6873	4.9376	5.1880	0.7100
0.0000	4.6875	4.9375	5.1875	0.7100	4.6875	4.9375	5.1875	0.7100
x	$(T_2,T_3)=(0.5,0.5)$				$(T_2,T_3)=(0.001,0.001)$			
	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$	$\mathbb{E}(F^{(1)})$	$\mathbb{E}(F^{(2)})$	$\mathbb{E}(F^{(3)})$	$\overline{U}$
100.0000	1.8333	4.5440	21.5219	0.8026	1.8333	5.7975	66.2128	0.9019
10.0000	1.8334	4.5445	21.5206	0.8026	1.8335	5.7997	66.2032	0.9019
7.5000	1.8342	4.5499	21.5061	0.8026	1.8348	5.8248	66.0965	0.9018
5.0000	1.8447	4.6160	21.3296	0.8021	1.8532	6.1222	64.8114	0.9008
2.5000	2.0351	5.2995	19.2548	0.7957	2.3001	8.7490	51.2013	0.8878
1.0000	3.2135	6.3102	12.7877	0.7699	4.4891	9.9071	22.2998	0.8338
0.7500	3.6280	6.2667	10.9723	0.7601	4.9034	9.0685	16.8989	0.8130
0.5000	4.0650	6.0378	9.0200	0.7473	5.1224	7.8518	12.0941	0.7861
0.2500	4.4499	5.5902	7.0438	0.7308	5.0541	6.4024	8.1519	0.7519
0.1000	4.6159	5.2197	5.9060	0.7189	4.8660	5.5108	6.2560	0.7277
0.0100	4.6819	4.9668	5.2576	0.7109	4.7070	4.9936	5.2875	0.7118
0.0010	4.6870	4.9404	5.1945	0.7101	4.6895	4.9431	5.1974	0.7102
0.0001	4.6874	4.9378	5.1882	0.7100	4.6877	4.9381	5.1885	0.7100
0.0000	4.6875	4.9375	5.1876	0.7100	4.6875	4.9376	5.1876	0.7100

3.8, 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 3.5: 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 3.6: Marginal waiting time dfs for various values of b_2 (under RI) in Example 2

Table 3.7: 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 3.8: 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	$w^{(1)}_{0.50}$	$w^{(2)}_{0.50}$	$\mathbb{E}(W^{(1)})$	$\mathbb{E}(W^{(2)})$	$w^{(1)}_{0.50}$	$w^{(2)}_{0.50}$	$\mathbb{E}(W^{(1)})$	$\mathbb{E}(W^{(2)})$
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

Chapter 4

A general mixed priority queue

4.1 Introduction

In this chapter, we consider an $M/G/1$ mixed priority queue with N distinct priority classes of customers that are each designated as either *urgent* or *non-urgent*. Specifically, the urgent set of classes refers to those classes which have preemptive resume priority over at least one lower priority class, whereas those classes which only have non-preemptive priority amongst lower priority classes form the non-urgent set. Moreover, urgent customers are assigned static priority in accordance to Eq. (1.16), while the non-urgent customers are assigned dynamic priority as defined by Eq. (1.21).

The resulting priority queueing system is quite general and can be used to model several real world situations. For example, the main motivation of Stanford et al. (2014) was to study the effectiveness of triage policies in an emergency room of a hospital. Their model was universally non-preemptive; however, it is quite reasonable to assume that some arriving patients will be more urgent than others and should require a doctor's attention immediately. The priority queueing model of this chapter allows for the consideration of such types of patients with preemptive priority over those which are less urgent. Moreover, in some instances, a doctor may decide to continue the servicing of a lower priority patient even in the midst of an arrival of an urgent-type patient. The new model can also have potential use in computer job

scheduling applications, as well as other areas (such as those discussed in Drekic and Stanford (2000, 2001) and Paterok and Ettl (1994)).

The rest of the chapter is organized as follows. In the next section, we provide a more detailed description of the model and other preliminaries. Section 4.3 describes the general methodology which is employed for deriving the LSTs of the marginal waiting time distributions. In Section 4.4, we establish the LSTs for the auxiliary random variables used to obtain the waiting time distributions. Finally, two numerical examples, comparing our new priority system to previously analyzed priority models of a similar nature, are given in Section 4.5. We remark that most of the work presented in this chapter is taken from Fajardo and Drekic (2015b).

4.2 The model

Similar to the priority queueing model of Chapter 3, we consider a single-server queueing system featuring N distinct priority classes of customers. The arrival processes for each class of customers form individual and independent Poisson processes, where λ_i denotes the arrival rate for class i, $i = 1, 2, \ldots, N$. We also let $\Lambda_i = \sum_{j=1}^{i} \lambda_j$ for $i = 1, 2, \ldots, N$. The service requirements for each customer are assumed to be class-dependent and independent of the arrival streams. As before, let $X^{(i)}$ represent the class-i service time random variable whose df and LST are denoted by

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

respectively. The *utilization factor* associated with the current priority queueing model is given by

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

which we assume satisfies the stability condition $\rho < 1$. Note that, in general, we let $Y(x) = 1 - \bar{Y}(x) = \mathbb{P}(Y \leq x)$ and $\widetilde{Y}(s) = \mathbb{E}(e^{-sY})$ represent the df and LST, respectively, of a random variable Y.

In addition to the assumption that $\mathcal{C}_i$s have priority over $\mathcal{C}_j$s whenever $i < j$, the N classes of customers are further categorized into two distinct types:

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

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

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

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

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

$$q_k(t) = a_k, \tag{4.1}$$

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

- For $k \in \mathscr{N}$: if τ_k is the arrival time of a $\mathcal{C}_k$, then

$$q_k(t) = b_k \cdot (t - \tau_k), \quad t \geq \tau_k, \tag{4.2}$$

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

It is further assumed that

$$a_m >> b_{m+1}, \tag{4.3}$$

which guarantees that at no point in time could a $\mathcal{C}_{\mathscr{N}}$ ever have greater priority than a $\mathcal{C}_{\mathscr{U}}$. Moreover, we assume that a $\mathcal{C}_i$ has preemptive resume priority over a $\mathcal{C}_j$ whenever $i < j$ and only if $i \in \mathscr{U}$; otherwise, if $i \in \mathscr{N}$, then the $\mathcal{C}_i$ has only non-preemptive priority over the $\mathcal{C}_j$. To illustrate, Table 4.1 represents the *priority relations matrix* (similar to those found in Adiri and Domb (1982)) for a 7-class mixed priority queue with $m = 3$. The (i, j)-th element of this matrix indicates the type of priority that class i has over class j for $i \leq j$, where p and np denote preemptive and non-preemptive priority, respectively.

Table 4.1: The priority relations matrix of a 7-class mixed priority queue with $m = 3$

i	j						
	1	2	3	4	5	6	7
1	FCFS	p	p	p	p	p	p
2		FCFS	p	p	p	p	p
3			FCFS	p	p	p	p
4				FCFS	np	np	np
5					FCFS	np	np
6						FCFS	np
7							FCFS

We next describe, in careful detail, the service discipline of this priority queueing model. First of all, recall that when we speak of a service selection instant, we are referring to an instant in time when a customer departs the system (i.e., after being completely serviced) and the server must subsequently select, from all the remaining customers in the system, the next customer to be serviced. It is important to realize that we do not consider a preemption instant to be a service selection instant. In general, mixed priority queues, such as the one considered in this chapter, employ the general Priority Service Guideline (as defined earlier in Chapter 1) at service selection instants; however, certain policies may further be put into place so as to override this guideline at a special kind of service selection instant. We provide the details to these exceptions later on in this section.

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

$b \equiv$ The set of classes which class k has priority over,

$a_{np} \equiv$ The set of classes which have non-preemptive priority over class k,

$a_p \equiv$ The set of classes which have preemptive priority over class k,

$a = a_{np} \cup a_p \equiv$ The set of classes which have priority over class k.

To begin, suppose that a $\mathcal{C}_k$ enters into service for the first time. For systems with at least one urgent class (i.e., $m > 0$), a_p must be a non-empty set if $k > 1$, and hence, it is possible for the service of this $\mathcal{C}_k$ to be interrupted by a $\mathcal{C}_{a_p}$. An interruption may take place if there exists a $\mathcal{C}_{a_p}$ with greater priority than the $\mathcal{C}_k$ currently in service. Since $a_p \subset \mathscr{U}$, it follows as a consequence of Eqs. (4.1) and (4.3) that any interruption period must commence immediately upon the arrival of the interrupting $\mathcal{C}_{a_p}$ to the system.

Although it is true that the set of classes in a_p have preemptive priority over class k, the ultimate decision on whether to interrupt the current servicing of the $\mathcal{C}_k$ is made according to the three threshold-based discretion rules which were first mentioned in Chapter 1. For convenience, we restate these discretion rules here:

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

(ii) *Front-end time-based* (FETB) *policy:* Once $\mathcal{T}$ time units of service have been successfully rendered, further preemptions are prevented,

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

As previously noted, Drekic and Stanford (2000) investigated the class-dependent case by allowing α_k, $\mathcal{T}_k$, and t_k to represent the corresponding class-k threshold parameters. We extend this idea one step further by allowing these threshold parameters to also depend on the class of the customer causing the interruption. Thus, we introduce $\alpha_{i,k} \in (0,1)$, $\mathcal{T}_{i,k} \geq 0$, and $t_{i,k} \geq 0$ as the corresponding class-k threshold parameters pertaining to a newly arriving high priority $\mathcal{C}_i$, $i \in a_p$. Furthermore, for any $k > 1$ and $i < j \in a_p$, we assume that

$$\alpha_{i,k} \geq \alpha_{j,k}, \qquad \mathcal{T}_{i,k} \geq \mathcal{T}_{j,k}, \qquad \text{and} \qquad t_{i,k} \leq t_{j,k}. \tag{4.4}$$

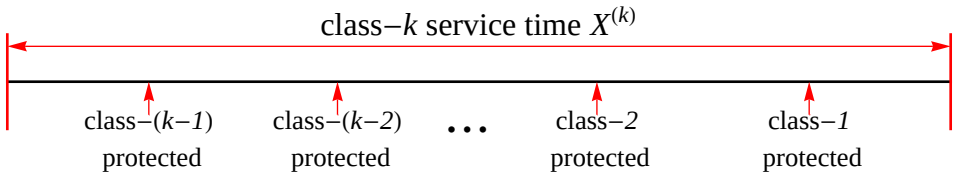


Figure 4.1: The protection of a class-k service time

We say that a class-k service becomes *class-i protected* the moment that the service of the $\mathcal{C}_k$ can no longer be preempted by a $\mathcal{C}_i$, $i \in a_p$. Hence, the consequences of Eq. (4.4) are that a class-k service becomes class-j protected before it becomes class-i protected for $i < j \in a_p$. For illustrative purposes, Figure 4.1 depicts the general sequence of protection for a class-k service time.

Remark 4.1 *Under various parameter settings, our mixed priority queueing model includes a number of previously analyzed priority queueing models as special cases. For example, by setting $m = 0$, our priority model exactly becomes the one considered by Stanford et al. (2014). By setting $m = N$ and assigning threshold parameters to be $\alpha_{i,k} = \alpha_k$, $\mathcal{T}_{i,k} = \mathcal{T}_k$, and $t_{i,k} = t_k$, our priority model is equivalent to the one considered by Drekic and Stanford (2000). Moreover, by setting $m = N$ and using threshold parameters of the form*

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

$$\mathcal{T}_{i,k} = \begin{cases} \infty & \text{if } k - i \geq d \\ 0 & \text{otherwise} \end{cases},$$

and

$$t_{i,k} = \begin{cases} 0 & \text{if } k - i \geq d \\ \infty & \text{otherwise} \end{cases},$$

our priority queueing model is equivalent to the one using the PD rule (resume-IPF case, where IPF denotes "interrupted processing first") as analyzed by Paterok and Ettl (1994, p. 1148), where d is the so-called preemption distance parameter (see Section 1.2 of the thesis when the PD rule was first introduced). Finally, it is also evident that the classical non-preemptive and preemptive priority queues, as well as the $\sum_{i=1}^{N} M_i/G_i/1$ FCFS queue, are all special cases of our general model.

Whenever a $\mathcal{C}_k$ is preempted out of service, the server returns to the interrupted $\mathcal{C}_k$ once the work associated with the following two items are completed:

(i) the complete servicing of the interrupting customer (which itself may also be interrupted), and

(ii) the complete servicing of all those $\mathcal{C}_{a_p}$s that the interrupting customer leaves behind whom, if they had arrived to the system at the time of the preemption, would have also caused an interruption.

Hence, at the end of an interruption period, the $\mathcal{C}_k$ re-enters service despite the fact that there may be customers of higher priority in the system (i.e., these are the higher priority customers who either never could, or can no longer cause an interruption to the $\mathcal{C}_k$). As an example, a class-k interruption period that occurs at some point after a class-k service time (as illustrated in Figure 4.1) becomes class-$(k-1)$ protected but before it becomes class-$(k-2)$ protected can only consist of the servicing of $\mathcal{C}_i$s for $i = 1, 2, \ldots, k-2$.

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

- For type-1 service selection instants, the general Priority Service Guideline is used to select the next customer for service.
- For type-2 service selection instants, the most recently interrupted customer re-enters into service.
- Preemption instants within the service of a $\mathcal{C}_k$ $(k > 1)$ occur at the arrivals of $\mathcal{C}_{a_p}$s in accordance with the threshold-based discretion rules of PB, FETB, and TETB.

We close this section with the mention of several key random variables of interest. In addition to the class-k waiting time $W^{(k)}$, residence period $R^{(k)}$, and flow time

$F^{(k)}$ (which were all previously defined in Chapter 3), we require yet another service-structure element, namely, the class-k *completion period* which we define as follows:

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

To find the LST of the class-k flow time $F^{(k)}$, we use the relation

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

which readily follows from the independence of $W^{(k)}$ and $R^{(k)}$. Furthermore, in order to derive the LST of $W^{(k)}$ (which is the focus of the next section), we require the LSTs of the following two auxiliary random variables:

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

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

The derivations of the LSTs of $C^{(k)}$, $R^{(k)}$, $\Upsilon_i^{(k)}$, and $\Phi_i^{(k)}$ are carried out in Section 4.4.

Remark 4.2 *For $k \in \mathscr{U}$, the first time that the server is ready to select a $\mathcal{C}_k$ after any one of these time intervals have started represents the first time that the system is clear of all $\mathcal{C}_a$s. However, for the case of $k \in \mathscr{N}$, the first time that the server is ready to select a $\mathcal{C}_k$ represents the first time that the system is clear of all $\mathcal{C}_{\mathscr{U}}$s and all those $\mathcal{C}_{\mathscr{N}}$s which are level-$(k-1)$ accredited.*

Remark 4.3 *Throughout the remainder of this chapter, we extend the definition of $\Upsilon_i^{(k)}$ to include the case when $i = k$, with the understanding that $\Upsilon_k^{(k)} = C^{(k)}$.*

4.3 Derivation of the waiting time LST

To derive an expression for $\widetilde{W}^{(k)}(s)$, we employ two analytical approaches: one for each of the cases $k \in \mathscr{U}$ and $k \in \mathscr{N}$. The reason for the two different approaches lies in the fact that the assignment of priority for a $\mathcal{C}_{\mathscr{U}}$ (which is via Eq. (4.1)) differs from that for a $\mathcal{C}_{\mathscr{N}}$ (which is via Eq. (4.2)). For the case $k \in \mathscr{U}$, we apply a similar level-crossing argument to the one used in Paterok and Ettl (1994). As evidenced in their work, the level-crossing method provides a straightforward approach to obtain the integral equation for the pdf of the steady-state class-k virtual wait. For dynamic priority queues, it is quite difficult to define the class-k virtual wait. Hence, we apply the same approach to the one used in Chapter 3 to establish $\widetilde{W}^{(k)}(s)$ for $k \in \mathscr{N}$.

4.3.1 Waiting time LST for $k \in \mathscr{U}$

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

$$F_k(x) = \lim_{t\to\infty} \mathbb{P}(V_k(t) \leq x),\ f_k(x) = \lim_{t\to\infty} \tfrac{\partial}{\partial x}\mathbb{P}(V_k(t) \leq x),\ \text{and } P_{0,k} = \lim_{t\to\infty} \mathbb{P}(V_k(t) = 0),$$

subject to the normalizing condition

$$P_{0,k} + \int_0^\infty f_k(x)\mathrm{d}x = 1. \tag{4.5}$$

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

$$\widetilde{W}^{(k)}(s) = \int_{x=0}^\infty e^{-sx}\mathrm{d}F_k(x) = P_{0,k} + \int_0^\infty e^{-sx} f_k(x)\mathrm{d}x. \tag{4.6}$$

To obtain the desired LST, we apply a level-crossing approach to establish an integral equation for $f_k(x)$. Let $\mathcal{U}_t(x)$ and $\mathcal{D}_t(x)$ denote the respective number of up- and down-crossings of level x of the class-k virtual wait process during the time

interval $(0, t)$. Recall the principle of set balance (e.g., see Brill (2008, Section 2.4.6)) which states that

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t(x))}{t} = \lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{U}_t(x))}{t}.$$

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

To find the up-crossing rate of level x of $\{V_k(t), t \geq 0\}$, we observe that a sample path of $\{V_k(t), t \geq 0\}$ up-jumps in three instances of time: (i) whenever a $\mathcal{C}_k$ arrives to the system, (ii) when a newly arriving $\mathcal{C}_a$ finds the system in the virtually idle state, and (iii) the moment when a $\mathcal{C}_b$'s service becomes class-k protected. A typical sample path of $\{V_k(t), t \geq 0\}$ is illustrated in Figure 4.2. It is important to note that depending on the specification of the threshold-based discretion parameters, the service of a $\mathcal{C}_b$ may either be entirely, partially, or not at all class-k protected. In Figure 4.2, both the first and third waiting $\mathcal{C}_b$s have service times which are entirely class-k protected, whereas the second waiting $\mathcal{C}_b$ has a service time that is only partially class-k protected.

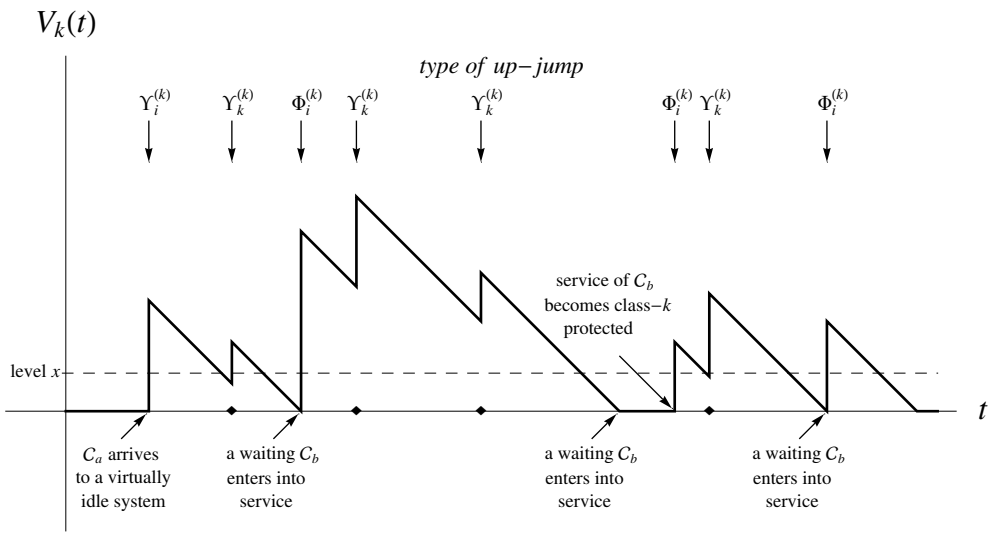


Figure 4.2: A typical sample path of $\{V_k(t), t \geq 0\}$

Let $\kappa_{k,i}$ denote the probability that the service of a $\mathcal{C}_i$ ($i \in b$) ever becomes class-k protected. Under the PB rule, $\kappa_{k,i} = 1$ as long as $\alpha_{k,i} < 1$ and is zero otherwise.

Similarly, under the TETB rule, $\kappa_{k,i} = 1$ if $t_{k,i} > 0$ and is zero otherwise. However, under the FETB rule, a class-i service becomes class-k protected only if the service time is greater than $\mathcal{T}_{k,i}$, and so $\kappa_{k,i} = \bar{B}^{(i)}(\mathcal{T}_{k,i})$ under this rule. The next theorem establishes the up- and down-crossing rates of level x.

Theorem 4.4 *The up- and down-crossing rates of level x are given by*

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{U}_t(x))}{t} = P_{0,k} \sum_{i=1}^{k} \lambda_i \bar{\Upsilon}_i^{(k)}(x) + \sum_{i=k+1}^{N} \kappa_{k,i} \lambda_i \bar{\Phi}_i^{(k)}(x) + \lambda_k \int_{y=0}^{x} \bar{\Upsilon}_k^{(k)}(x-y) f_k(y)\,\mathrm{d}y, \quad x > 0 \tag{4.7}$$

and

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t(x))}{t} = f_k(x), \quad x > 0. \tag{4.8}$$

Proof. We present intuitive explanations for each term of Eq. (4.7). For $i \in a$ or $i = k$, the rate of up-jumps caused by a $\mathcal{C}_i$ arriving to a virtually idle system is simply $\lambda_i P_{0,k}$. Furthermore, only the proportion $\bar{\Upsilon}_i^{(k)}(x)$ of these up-jumps lead to an up-crossing of level x. The rate at which a $\mathcal{C}_i$ ($i \in b$) arrives to the system that eventually induces a delay to the $\mathcal{C}_k$s is $\lambda_i \kappa_{k,i}$. Such arrivals eventually result in up-jumps of $\{V_k(t), t \geq 0\}$ which cross level x with probability $\bar{\Phi}_i^{(k)}(x)$. Finally, the long-run probability of an up-jump occurring from level y is $f_k(y)dy$, and the probability that an up-crossing of level x occurs from level y is $\bar{\Upsilon}_k^{(k)}(x-y)$. The justification of Eq. (4.8) is similar to that for the down-crossing rate of the virtual wait process in an $M/G/1$ queue (e.g., see Brill (2008, Theorem 3.3 and Corollary 3.2)). □

From the principle of set balance, we equate Eqs. (4.7) and (4.8) to yield the following integral equation for $f_k(x)$:

$$f_k(x) = P_{0,k} \sum_{i=1}^{k} \lambda_i \bar{\Upsilon}_i^{(k)}(x) + \sum_{i=k+1}^{N} \lambda_i \kappa_{k,i} \bar{\Phi}_i^{(k)}(x) + \lambda_k \int_{y=0}^{x} \bar{\Upsilon}_k^{(k)}(x-y) f_k(y)\,\mathrm{d}y, \quad x > 0. \tag{4.9}$$

By multiplying Eq. (4.9) by e^{-sx} and integrating with respect to x over $(0,\infty)$, we obtain

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

It follows from Eq. (4.6) that for $k\in\mathscr{U}$,

$$\widetilde{W}^{(k)}(s) = \frac{P_{0,k}\big(s+\sum_{i=1}^{k-1}\lambda_i(1-\widetilde{\Upsilon}_i^{(k)}(s))\big)+\sum_{i=k+1}^{N}\lambda_i\kappa_{k,i}(1-\widetilde{\Phi}_i^{(k)}(s))}{s-\lambda_k+\lambda_k\widetilde{C}^{(k)}(s)}. \tag{4.10}$$

An expression for $\mathbb{E}(W^{(k)})$ can be obtained by multiplying Eq. (4.9) by x and integrating with respect to x over $(0,\infty)$, leading to

$$\mathbb{E}(W^{(k)}) = \frac{P_{0,k}\sum_{i=1}^{k-1}\lambda_i\mathbb{E}\big((\Upsilon_i^{(k)})^2\big)+\lambda_k\mathbb{E}\big((C^{(k)})^2\big)+\sum_{i=k+1}^{N}\lambda_i\kappa_{k,i}\mathbb{E}\big((\Phi_i^{(k)})^2\big)}{2\big(1-\lambda_k\mathbb{E}(C^{(k)})\big)}. \tag{4.11}$$

The LST of $W_{BP}^{(k)}$ (i.e., the wait of a $\mathcal{C}_k$ arriving to the system during a busy period) can easily be obtained from Eq. (4.6) and the fact that $\widetilde{W}^{(k)}(s) = P_{0,k} + (1-P_{0,k})\widetilde{W}_{BP}^{(k)}(s)$. In particular, we have that

$$\widetilde{W}_{BP}^{(k)}(s) = \int_{x=0}^{\infty} e^{-sx} f_k(x)\mathrm{d}x/(1-P_{0,k}). \tag{4.12}$$

Moreover, we establish a formula for $P_{0,k}$ by first observing that

$$\int_{0}^{\infty} f_k(x)\,\mathrm{d}x = \frac{P_{0,k}\sum_{i=1}^{k}\lambda_i\mathbb{E}(\Upsilon_i^{(k)})+\sum_{i=k+1}^{N}\lambda_i\kappa_{k,i}\mathbb{E}(\Phi_i^{(k)})}{1-\lambda_k\mathbb{E}(C^{(k)})}.$$

It then follows, from the normalizing condition Eq. (4.5), that

$$P_{0,k} = \frac{1-\lambda_k\mathbb{E}(C^{(k)})-\sum_{i=k+1}^{N}\lambda_i\kappa_{k,i}\mathbb{E}(\Phi_i^{(k)})}{1+\sum_{i=1}^{k-1}\lambda_i\mathbb{E}(\Upsilon_i^{(k)})}. \tag{4.13}$$

We end the current subsection with a remark on the level-crossing approach used here and the one employed by Paterok and Ettl (1994).

Remark 4.5 *The level-crossing analysis of* $\{V_k(t), t \geq 0\}$ *carried out by Paterok and Ettl (1994) differs slightly from the one we use here. While their approach compares the expected number of up- and down-crossings of level* x *of* $\{V_k(t), t \geq 0\}$ *within a single regeneration cycle, our level-crossing analysis compares the long-run up- and down-crossing rates of level* x*. The latter level-crossing approach was first introduced by Brill (1975), whereas the former approach was independently developed by Cohen (1977).*

4.3.2 Waiting time LST for $k \in \mathscr{N}$

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

$$\widetilde{W}^{(k)}(s) = (1-\rho) + \rho\widetilde{W}_{BP}^{(k)}(s), \quad k \in \mathscr{N}. \tag{4.14}$$

Similar to our derivation of $\widetilde{W}_{BP}^{(k)}(s)$ for the PAPQ in Section 3.6, we first derive the LST of $P_{BP}^{(k)}$ and then apply the relation

$$\widetilde{W}_{BP}^{(k)}(s) = \widetilde{P}_{BP}^{(k)}(s/b_k). \tag{4.15}$$

In order to determine $\widetilde{P}_{BP}^{(k)}(s)$, we again make use of the maximal priority process, which must be defined for the current priority queueing model. For the NPAPQ, Stanford et al. (2014) defined the maximal priority process in terms of the service commencement times and departure instants of the system. Since the current priority queueing model allows for a $\mathcal{C}_{\mathscr{N}}$ to be preempted out of service, we require a slightly more general definition of the maximal priority process. Our definition of the maximal priority process follows below.

Definition 4.6 *The maximal priority process is an* $(N-m)$*-dimensional stochastic process* $\mathcal{M}(t) = \{(M_{m+1}(t), M_{m+2}(t), \ldots, M_N(t)), t \geq 0\}$*, satisfying the following conditions:*

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

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

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

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

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

$$M_k(\delta_i^+) = \begin{cases} \min\{M_k(\delta_i^-), q_\vee(\delta_i^+)\} & \text{if } \delta_i \text{ is of type 1} \\ M_k(\delta_i^-) & \text{if } \delta_i \text{ is of type 2} \end{cases}, \tag{4.16}$$

where $q_\vee(t)$ represents the greatest (accumulated) priority amongst all the customers present at time t, which is zero during idle periods. In Eq. (4.16)*, note that*

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

In what follows, we (artificially) set $b_{N+1} = 0$ (which correspondingly implies that $M_{N+1}(t) = 0$ for all $t > 0$). Definition 4.6 simply implies that during busy periods, $M_k(t)$ increases linearly at rate b_k and down-jumps at some of the service selection instants. Figure 4.3 illustrates a typical sample path of the maximal priority process for a 5-class mixed priority queue with $m = 2$. In Figure 4.3, the actual accumulated priorities of the customers present in the system are given by the thin lines.

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

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

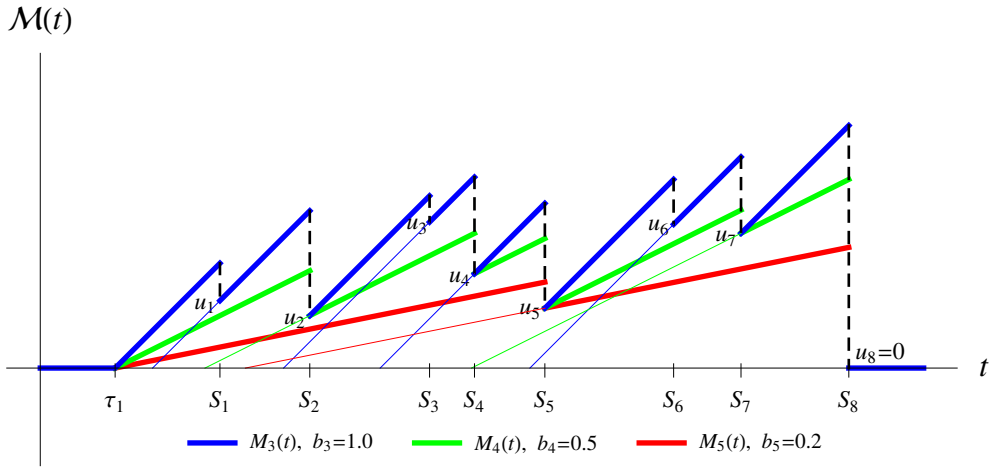


Figure 4.3: A typical sample path of $\{\mathcal{M}(t), t \geq 0\}$ for a 5-class mixed priority queue with $m = 2$ (i.e., $\mathscr{N} = \{3, 4, 5\}$)

(i) and (ii) for δ above. We refer to S_i as the i-th service selection instant for a $\mathcal{C}_{\mathscr{N}}$. Furthermore, let $\mathbf{S} = \{S_i\}_{i=1}^{\infty}$ be the sequence of service selection instants for the $\mathcal{C}_{\mathscr{N}}$s. It is important to note that S_i represents the service commencement of a $\mathcal{C}_{\mathscr{N}}$ only if there are still customers remaining in the system at S_i. Otherwise, S_i represents the end of a busy period, which is signalled by a down-jump of $M_{m+1}(t)$ to level 0 (e.g., see S_8 in Figure 4.3).

The main reason for defining $\mathbf{S}$, however, is stated in the next observation. The maximal priority process defined for the non-urgent classes in our new priority queue behaves identically to the maximal priority process for the NPAPQ (i.e., see Stanford et al. (2014)). In other words, we can similarly analyze the waiting times for a $\mathcal{C}_{\mathscr{N}}$ of the new priority queue as we would for a customer in the NPAPQ. In this equivalent non-preemptive priority queue, $\mathbf{S}$ would play the role of the sequence of departure instants of the customers, while $C^{(k)}$, $\Upsilon_i^{(k)}$, and $\Phi_i^{(k)}$ would serve as the effective service times.

Remark 4.7 *Similar to the interpretation of the upper bounds that the maximal priority process provides for the PAPQ and the NPAPQ, $M_k(t)$ is the least upper bound of class-k accumulated priorities which would not result in a violation of the*

service discipline. Furthermore, one can think of $\mathcal{M}(t)$ as the collection of these least upper bounds for accumulated priorities that one would sketch when given only the following three pieces of information:

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

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

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

To sketch $\mathcal{M}(t)$, one must also bear in mind some of the fundamental characteristics of the priority queueing system, namely that $\mathcal{C}_k$s accumulate priority via Eq. (4.2), $\mathcal{C}_{\mathscr{N}}$s arrive to the system with initial priority levels of zero, and $\mathcal{C}_{\mathscr{N}}$s cannot preempt service. For example, one can reproduce the sample path of $\mathcal{M}(t)$ in Figure 4.3 given only τ_1 and the pairs (S_i, u_i) for $i = 1, 2, \ldots, 8$.

We next provide some fundamental concepts and results pertaining to the current priority queueing model. First of all, recall from Chapter 3 that a $\mathcal{C}_j$ ($j \leq k$, $j \in \mathscr{N}$) is served at level-k accreditation if

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

where δ represents the time at which this $\mathcal{C}_j$ first enters into service and $q_{\vee}(\delta^+)$ is its priority level at that time. An important result pertaining to the proportion of $\mathcal{C}_k$s arriving during busy periods and that are $\mathcal{C}^{(acc:k)}$s is provided in the next lemma.

Lemma 4.8 *The steady-state probability that a $\mathcal{C}_k$ who arrives during a busy period and is serviced at level-k accreditation (i.e., is also a $\mathcal{C}^{(acc:k)}$) is given by $1 - b_{k+1}/b_k$ for any $k \in \mathscr{N}$.*

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

result then follows from the fact that $\mathcal{C}_k$s arrive to the system according to a Poisson process. $\square$

Level-k accreditation intervals, similar to those of the PAPQ and the NPAPQ, are also inherent in the current priority queueing model. Specifically, similar to the PAPQ and the NPAPQ, a level-k accreditation interval is a period of time that either starts at the beginning of a busy period, or when a $\mathcal{C}^{(acc:\ell)}$ for $\ell > k$ enters into service for the first time. However, for the current priority queueing model, a level-k accreditation interval ends once the system becomes clear of both the initial customer and all $\mathcal{C}^{(acc:i)}$s for $i = m+1, m+2, \ldots, k$ (i.e., all customers that have become at least level-k accredited).

Note that if δ represents the service selection instant for a $\mathcal{C}^{(acc:\ell)}$ where $\ell > k$, then this implies that $M_{k+1}(t)$ must have down-jumped at time δ (i.e., $q_{\mathrm{V}}(\delta^+) < M_{k+1}(\delta^-)$). In addition, if there are still customers present at the end of the ensuing level-k accreditation interval, then clearly, at this same instant, another $\mathcal{C}^{(acc:\ell)}$ for $\ell > k$ will commence service. Therefore, we observe that during busy periods, the commencement/termination instants of level-k accreditation intervals coincide with the service selection instants $\mathbf{S}$ for which $M_{k+1}(t)$ down-jumps. In other words, during busy periods, the level-k accreditation intervals are the time periods between successive down-jumps of $M_{k+1}(t)$. It is also obvious that a termination instant of a level-k accreditation interval which clears the system of all customers does not also represent a commencement instant of the next level-k accreditation interval, but rather signals the end of the busy period. Figure 4.4 illustrates the general structure of a level-4 accreditation interval for a 6-class mixed priority queue with $m = 2$.

Within a level-k accreditation interval, we note further that $M_k(t)$ down-jumps at instants corresponding to the service selection instants of all the $\mathcal{C}^{(acc:k)}$s. However, a down-jump of $M_k(t)$ also marks the commencement/termination of a level-$(k-1)$ accreditation interval. Therefore, a level-k accreditation interval is partitioned by a sequence of level-$(k-1)$ accreditation intervals. This suggests that it may be possible to view a level-k accreditation interval as a delay busy period of $\mathcal{C}^{(acc:k)}$s, whose effective service times are level-$(k-1)$ accreditation intervals. We show that this is precisely the case in Section 4.4.

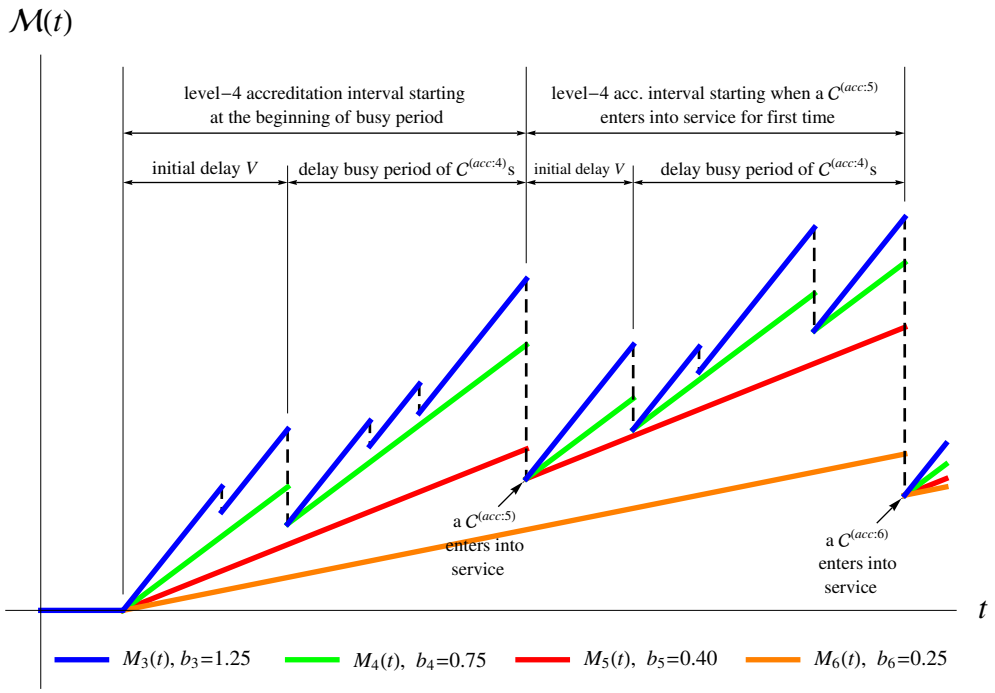


Figure 4.4: Level-4 accreditation intervals in a 6-class mixed priority queue with $m = 2$ (i.e., $\mathcal{N} = \{3, 4, 5, 6\}$)

We next proceed to establish the relation between level-k accreditation intervals and the previously introduced auxiliary variables (including the completion periods). First of all, observe that of the service selection instants $\mathbf{S}$, only those resulting in a down-jump of $M_{k+1}(t)$ represent the possible selection instants for a $\mathcal{C}_{k+1}$. As a result, the end of a level-k accreditation interval also represents the instant in time that the server is ready to select a $\mathcal{C}_{k+1}$ for service. Hence, the distribution of the level-k accreditation interval depends on the class of the initial customer and is given by the corresponding auxiliary random variable. Table 4.2 summarizes the distributions of the types of level-k accreditation intervals, including the distribution of the initiating level-$(k-1)$ accreditation interval, which we denote by V and refer to as the initial delay of the interval.

Table 4.2: Distributions of the level-k accreditation intervals

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

Remark 4.9 *The resulting structuralization of the busy period for this mixed priority queueing system is similar to that of the NPAPQ in that the entire busy period is partitioned by level-k accreditation intervals. Recall that for the PAPQ, the busy period is partitioned by subperiods that are either level-k accreditation intervals or class-$(k+1)$ residence periods.*

In order to obtain our recursive procedure for $\widetilde{P}_{BP}^{(k)}(s)$ pertaining to the PAPQ, we exploited in the previous chapter the decomposition of the accumulated priority of a $\mathcal{C}_k$ who arrives during a busy period. In particular, we decomposed the accumulated priority earned (immediately prior to entering service for the first time) by a $\mathcal{C}^{(acc:k)}$ into two independent parts: (i) the initiating priority level u_0, and (ii) $\mathcal{P}^{(acc:k)}$, the additional priority accumulated during the accreditation interval after having accumulated priority level u_0. As a result, our recursive procedure in Section 3.6 for

$\widetilde{P}_{BP}^{(k)}(s)$ relied heavily on the LST of $\mathcal{P}^{(acc:k)}$. Note that the recursive procedure of Section 3.6 also relied heavily on $\mathcal{P}^{(int:k)}$.

The decomposition of priority earned by a $\mathcal{C}^{(acc:k)}$ in the current mixed priority queueing model is similar to that of a $\mathcal{C}^{(acc:k)}$ in the PAPQ and NPAPQ. Figure 4.5 illustrates such a decomposition of the accumulated priority for a $\mathcal{C}^{(acc:4)}$ in a 5-class mixed priority queue with $m = 2$. In addition, since there is no preemption between the $\mathcal{C}_{\mathcal{N}}$s in the current priority queueing model, $\mathcal{P}^{(int:k)}$ is non-existent and the corresponding recursive scheme for $\widetilde{P}_{BP}^{(k)}(s)$ only depends on the LST of $\mathcal{P}^{(acc:k)}$.

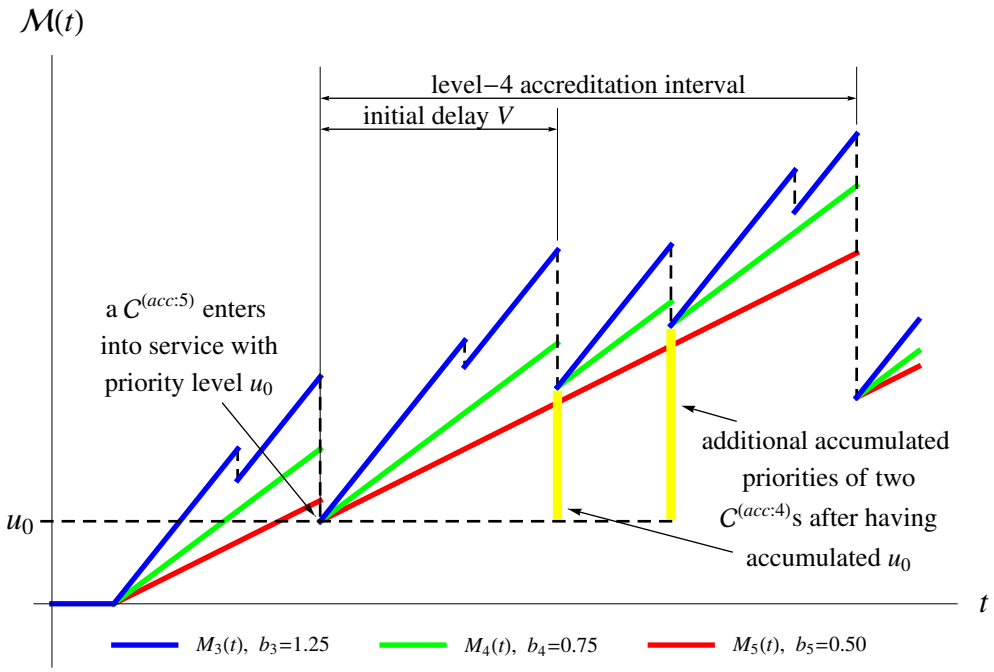


Figure 4.5: Decomposition of the accumulated priority for a $\mathcal{C}^{(acc:4)}$ in a 5-class mixed priority queue with $m = 2$ (i.e., $\mathcal{N} = \{3, 4, 5\}$)

An expression for $\widetilde{\mathcal{P}}^{(acc:k)}(s)$ can be readily obtained after observing the connection between the current priority queueing model and the $M/G/1$ queue with accumulating priority of Section 2.5. Before making this necessary observation, we state four important properties of the maximal priority process. We remark that these properties were first derived by Stanford et al. (2014). We do not provide the proofs of these properties but instead direct interested readers to Stanford et al.

(2014, Theorems 3.2 and 7.2) for their proofs. The four properties are as follows:

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

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

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

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

Important Observation 4.10 *Observe that from properties (P.2) and (P.4), it must be that the down-jumps of $M_k(t)$ during the level-k accreditation interval are exponentially distributed with parameter $\sum_{j=m+1}^{k} \lambda_j/b_j$. Moreover, during a level-k accreditation interval, the k-th and $(k+1)$-th components of the maximal priority process $(M_{k+1}(t), M_k(t))$ behave like the maximal priority process (during busy periods) of the FCFS $M/G/1$ queue with accumulating priority and blocking having the following characteristics:*

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

(ii) service time LST of $\widetilde{\beta}^{(k)}(s) = \sum_{i=m+1}^{k}(\lambda_i(b_k/b_i))/\gamma_k)\widetilde{\Upsilon}_i^{(k)}(s)$ for customers arriving during busy periods,

(iii) service time LST of $\widetilde{V}(s)$ for zero-wait customers,

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

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

From Important Observation 4.10 and after an application of Eq. (2.59) with $q = b_{k+1}/b_k$ and LST argument $b_k s$, it follows that an expression for the LST of $\mathcal{P}^{(acc:k)}$ (associated with an initial delay V) is given by

$$\widetilde{\mathcal{P}}^{(acc:k)}(s) \equiv \widetilde{\mathcal{P}}^{(acc:k)}(s;V) = \frac{\big(1-\gamma_k^{(k+1)}\mu_{k,1}\big)\big(\widetilde{\mathcal{A}}(b_{k+1}s)-\widetilde{V}(b_k s)\big)}{\mathbb{E}(V)\big(1-\frac{b_{k+1}}{b_k}\big)\big(b_k s-\gamma_k\big(1-\widetilde{\beta}^{(k)}(b_k s)\big)\big)}, \tag{4.17}$$

where $\widetilde{\mathcal{A}}(s) = \widetilde{\Gamma}_0(s;\gamma_k^{(k+1)},\beta^{(k)},V)$ from Eq. (1.6), $\gamma_k^{(k+1)} = \gamma_k(1-b_{k+1}/b_k)$, and $\mu_{k,i}$ represents the i-th moment of the random variable (to be denoted by $\beta^{(k)}$) whose LST is $\widetilde{\beta}^{(k)}(s)$ above. Furthermore, upon substitution of the appropriate parameters into Eq. (2.61), the first moment of $\mathcal{P}^{(acc:k)}$ works out to be

$$\mathbb{E}(\mathcal{P}^{(acc:k)}) = b_k\Bigg(\frac{\mathbb{E}(V^2)}{2\mathbb{E}(V)}\cdot\Bigg[1+\frac{b_{k+1}/b_k}{1-\gamma_k^{(k+1)}\mu_{k,1}}\Bigg] + \frac{\gamma_k\mu_{k,2}}{2(1-\gamma_k\mu_{k,1})}\cdot\Bigg[1-\Bigg(\frac{b_{k+1}/b_k}{1-\gamma_k^{(k+1)}\mu_{k,1}}\Bigg)^2\Bigg]\Bigg). \tag{4.18}$$

Remark 4.11 *Note the fact that a $\mathcal{C}^{(acc:k)}$ must belong to one of the classes in $\{m+1, m+2, \ldots, k\}$. This of course implies that one $\mathcal{C}^{(acc:k)}$ may accumulate priority linearly at a different rate from another $\mathcal{C}^{(acc:k)}$ (i.e., if they each belong to two different classes). Nevertheless, the distribution of $\mathcal{P}^{(acc:k)}$ remains the same regardless of the specific class to which the $\mathcal{C}^{(acc:k)}$ belongs.*

We are now ready to present the recursive procedure for obtaining $\widetilde{P}_{BP}^{(k)}(s)$, $k \in \mathcal{N}$. Let $P_{acc}^{(k)}$ be the accumulated priority of a $\mathcal{C}_k^{(acc:k)}$. Similarly, we define $P_{unacc}^{(k)}$ as the accumulated priority of a $\mathcal{C}_k^{(acc:\ell)}$ for some $\ell > k$. For convenience, let $\mathcal{C}_k^{(acc:>k)}$ denote a $\mathcal{C}_k^{(acc:\ell)}$ for some $\ell > k$. It therefore follows from Lemma 4.8 that

$$\widetilde{P}_{BP}^{(k)}(s) = \frac{b_k - b_{k+1}}{b_k}\widetilde{P}_{acc}^{(k)}(s) + \frac{b_{k+1}}{b_k}\widetilde{P}_{unacc}^{(k)}(s). \tag{4.19}$$

To develop a recursion for (4.19), Remark 4.11 implies that $\mathcal{C}_k^{(acc:>k)}$s have an accumulated priority that is identically distributed to that of a $\mathcal{C}_{k+1}$ who arrives during a busy period, so that $\widetilde{P}_{unacc}^{(k)}(s) = \widetilde{P}_{BP}^{(k+1)}(s)$. This result is an intuitive one as

both types of customers possess the property that their accumulated priorities are always bounded above by $M_{k+1}(t)$. We may now re-write Eq. (4.19) as

$$\widetilde{P}_{BP}^{(k)}(s) = \frac{b_k - b_{k+1}}{b_k}\widetilde{P}_{acc}^{(k)}(s) + \frac{b_{k+1}}{b_k}\widetilde{P}_{BP}^{(k+1)}(s), \tag{4.20}$$

thereby achieving a recursive relation.

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

$$\widetilde{P}_{acc,0}^{(k)}(s) = \widetilde{\mathcal{P}}^{(acc:k)}(s; V_0^{(k)}), \tag{4.21}$$

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

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

To understand Eq. (4.21), note that the initial priority level of a level-k accreditation interval which starts at the beginning of a busy period is zero. Therefore, the accumulated priority of a $\mathcal{C}_k^{(acc:k)}$ serviced within these kinds of level-k accreditation intervals is simply equal to the priority accumulated during the interval. Furthermore, the initial delay V_0 is a level-$(k-1)$ accreditation interval which can be initiated by any customer arriving to an empty system.

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

$$\widetilde{P}_{acc,1}^{(k)}(s) = \frac{\sum_{j=m+1}^{k} \pi_j^{(k)} \widetilde{P}_{BP}^{(k+1)}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; \Upsilon_j^{(k)}) + \sum_{j=k+1}^{N} \pi_j^{(k)} \widetilde{P}_{BP}^{(j)}(s)\widetilde{\mathcal{P}}^{(acc:k)}(s; \Phi_j^{(k)})}{\sum_{j=m+1}^{N} \pi_j^{(k)}}, \tag{4.22}$$

where $\pi_j^{(k)}$ is the long-run fraction of time that the system processes a level-k accreditation interval initiated by a $\mathcal{C}_j$ ($j \in \mathscr{N}$) arriving to the system during a busy

period. To understand Eq. (4.22), recall that the priority level of a $\mathcal{C}_k^{(acc:k)}$ serviced within a level-k accreditation interval starting at the service commencement of a $\mathcal{C}^{(acc:\ell)}$ for some $\ell > k$ can be decomposed into two independent components: (i) u_0, the accumulated priority of the initiating $\mathcal{C}^{(acc:\ell)}$, and (ii) $\mathcal{P}^{(acc:k)}$, the additional priority accumulated after having accumulated the initial priority level u_0. Hence, the accumulated priority of such a $\mathcal{C}_k^{(acc:k)}$ has LST which takes on the general form

$$\widetilde{P}_{acc,1}^{(k)}(s;V) = \widetilde{u}_0(s)\widetilde{\mathcal{P}}^{(acc:k)}(s;V),$$

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

The distributions of both u_0 and V depend solely on the class of the initial customer. In particular, if the initial customer is of class j for $m < j \leq k$, then $\widetilde{u}_0(s) = \widetilde{P}_{BP}^{(k+1)}(s)$ and $\widetilde{V}(s) = \widetilde{\Upsilon}_j^{(k)}(s)$. Otherwise, for $j > k$, $\widetilde{u}_0(s) = \widetilde{P}_{BP}^{(j)}(s)$ and $\widetilde{V}(s) = \widetilde{\Phi}_j^{(k)}(s)$. If we define $\pi_0^{(k)}$ as the long-run fraction of time that the system spends processing a level-k accreditation interval initiated by a customer who arrived to an empty queue, then it must be that

$$\widetilde{P}_{acc}^{(k)}(s) = \frac{1}{\rho}\big(\pi_0^{(k)}\widetilde{P}_{acc,0}^{(k)}(s) + (\rho - \pi_0^{(k)})\widetilde{P}_{acc,1}^{(k)}(s)\big). \tag{4.23}$$

Eqs. (4.20)–(4.23) together form our recursive procedure to obtain $\widetilde{P}_{BP}^{(k)}(s)$.

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

$$\pi_j^{(k)} = \rho\frac{\lambda_j\mathbb{E}(\Phi_j^{(k)})}{1-\gamma_k^{(k+1)}\mu_{k,1}}, \qquad j > k. \tag{4.24}$$

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

$$\pi_j^{(k)} = \rho\frac{\lambda_j(b_{k+1}/b_j)\mathbb{E}(\Upsilon_j^{(k)})}{1-\gamma_k^{(k+1)}\mu_{k,1}}, \qquad m < j \leq k. \tag{4.25}$$

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

$$\pi_0^{(k)} = \frac{1-\rho}{1-\gamma_k^{(k+1)}\mu_{k,1}} \Bigg[\sum_{j=1}^{k} \lambda_j \mathbb{E}(\Upsilon_j^{(k)}) + \sum_{j=k+1}^{N} \lambda_j \mathbb{E}(\Phi_j^{(k)}) \Bigg]. \tag{4.26}$$

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

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

In this section, we derive expressions for the LSTs of class-k completion periods, residence periods, and the auxiliary random variables introduced earlier in Section 4.2. Since the preemptive resume service discipline is a work-conserving one, it is straightforward to show that the LSTs of the class-k ($k \in \mathscr{U}$) auxiliary random variables are given by

$$\widetilde{\Upsilon}_i^{(k)}(s) = \widetilde{B}^{(i)}\big(s + \Lambda_{k-1}(1 - \widetilde{\Upsilon}_{1:k-1}^{(k)}(s))\big), \qquad i \in a \tag{4.27}$$

and

$$\widetilde{\Phi}_i^{(k)}(s) = \widetilde{Z}_k^{(i)}\big(s + \Lambda_{k-1}(1 - \widetilde{\Upsilon}_{1:k-1}^{(k)}(s))\big), \qquad i \in b, \tag{4.28}$$

where, from Eq. (1.3), $\widetilde{\Upsilon}_{1:k-1}^{(k)} = \widetilde{\Gamma}\big(s; \Lambda_{k-1}, \sum_{i=1}^{k-1}(\lambda_i/\Lambda_{k-1})X^{(i)}\big)$ is the busy period LST of the $\mathcal{C}_a$s and $Z_k^{(i)}$ represents the class-k protected portion of a class-i service. Table 4.3 summarizes the various forms of $Z_k^{(i)}$ and $\widetilde{Z}_k^{(i)}(s)$ under each of the three threshold-based discretion rules. Moreover, the class-k completion period LST is simply given by

$$\widetilde{C}^{(k)}(s) = \widetilde{\Upsilon}_k^{(k)}(s) = \widetilde{B}^{(k)}\big(s + \Lambda_{k-1}(1 - \widetilde{\Upsilon}_{1:k-1}^{(k)}(s))\big). \tag{4.29}$$

For the case $k \in \mathscr{N}$, both $\widetilde{\Upsilon}_i^{(k)}(s)$ and $\widetilde{\Phi}_i^{(k)}(s)$ are obtained recursively. Specifically, it immediately follows from Table 4.2, Important Observation 4.10, and Eq. (2.49) that for each $k \geq m+1$:

$$\widetilde{\Upsilon}_i^{(k+1)}(s) = \widetilde{\Upsilon}_i^{(k)}\big(s + \gamma_k^{(k+1)}(1 - \widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s))\big), \qquad i \leq k \tag{4.30}$$

Table 4.3: Various forms of $Z_k^{(i)}$ and its corresponding LST

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

and

$$\widetilde{\Phi}_i^{(k+1)}(s)=\widetilde{\Phi}_i^{(k)}\big(s+\gamma_k^{(k+1)}(1-\widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s))\big),\qquad i>k+1, \tag{4.31}$$

where $\widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s)=\widetilde{\Gamma}(s;\gamma_k^{(k+1)},\beta^{(k)})$ from Eq. (1.3). Furthermore, the class-$(k+1)$ completion period LST is given by

$$\widetilde{C}^{(k+1)}(s)=\widetilde{\Upsilon}_{k+1}^{(k+1)}(s)=\widetilde{\Phi}_{k+1}^{(k)}\big(s+\gamma_k^{(k+1)}(1-\widetilde{\Upsilon}_{m+1:k}^{(k+1)}(s))\big). \tag{4.32}$$

The respective starting points for the recursive expressions given in Eqs. (4.30)–(4.32) are $\widetilde{\Upsilon}_i^{(m+1)}(s)$ for all $i\le m+1$, $\widetilde{\Phi}_i^{(m+1)}(s)$ for all $i>m+2$, and $\widetilde{\Phi}_{m+2}^{(m+1)}(s)$. Since $\mathscr{U}$ also represents the set of classes which have priority over class $m+1$, it turns out that the formulas for $\widetilde{\Upsilon}_i^{(k)}(s)$, $\widetilde{\Phi}_i^{(k)}(s)$, and $\widetilde{C}^{(k)}(s)$ given by Eqs. (4.27)–(4.29) also hold true when $k=m+1$. Note that in using Eq. (4.28) with $k=m+1$, it is necessary to define the threshold parameters $\alpha_{m+1,i}=0$, $\mathcal{T}_{m+1,i}=0$, and $t_{m+1,i}=\infty$ for all $i>m+1$.

Remark 4.12 *The above formulas illustrate the fact that a level-k accreditation interval is merely a delay busy period of $\mathcal{C}^{(acc:k)}$s whose service times are level-$(k-1)$ accreditation intervals, corresponding to $\Upsilon_i^{(k)}$ for $i=m+1,m+2,\ldots,k$.*

Remark 4.13 *With $k=N$, Eq. (4.30) yields a recursive procedure for calculating $\widetilde{\Upsilon}_i^{(N+1)}(s)$, $i=1,2,\ldots,N$. We remark that $\Upsilon_i^{(N+1)}$ represents the duration of a busy period which is initiated by a $\mathcal{C}_i$.*

To obtain $\widetilde{R}^{(k)}(s)$, we require the joint transform of the preemptible and non-preemptible periods of a class-k service time. In particular, similar to the analysis

conducted by Drekic and Stanford (2000), we segment the class-k service time $X^{(k)}$ into its preemptible portion $X_p^{(k)}$ and its non-preemptible (or protected) portion $X_{p_0}^{(k)}$. For the current priority queueing model, however, we must further partition the preemptible portion $X_p^{(k)}$ as follows:

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

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

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

$$\Theta^{(k)}(\mathbf{s}) = \mathbb{E}(e^{-s_1 X_{p1}^{(k)} - s_2 X_{p2}^{(k)} - \cdots - s_{k-1} X_{p_{k-1}}^{(k)} - s_0 X_{p0}^{(k)}}).$$

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

$$\begin{aligned}\text{(PB)} \quad \Theta^{(k)}(\mathbf{s}) &= \int_{x=0}^{\infty} e^{-(\sum_{i=1}^{k-1} s_i(\alpha_{i,k} - \alpha_{i+1,k}) + s_0(1-\alpha_{1,k}))x} \mathrm{d}B^{(k)}(x) \\ &= \widetilde{B}^{(k)}\big(\textstyle\sum_{i=1}^{k-1} s_i(\alpha_{i,k} - \alpha_{i+1,k}) + s_0(1-\alpha_{1,k})\big), \end{aligned} \tag{4.33}$$

$$\begin{aligned}\text{(FETB)} \quad \Theta^{(k)}(\mathbf{s}) = \sum_{i=1}^{k-1} e^{-\sum_{j=i+1}^{k-1}(s_j - s_{j-1})\mathcal{T}_{j,k}} \int_{x=\mathcal{T}_{i+1,k}}^{\mathcal{T}_{i,k}} e^{-s_i x} \mathrm{d}B^{(k)}(x) \\ + e^{-(\sum_{j=2}^{k-1}(s_j - s_{j-1})\mathcal{T}_{j,k} + (s_1 - s_0)\mathcal{T}_{1,k})} \int_{x=\mathcal{T}_{1,k}}^{\infty} e^{-s_0 x} \mathrm{d}B^{(k)}(x), \end{aligned} \tag{4.34}$$

and

$$\text{(TETB)}\quad \Theta^{(k)}(\mathbf{s}) = \sum_{i=1}^{k-1} e^{-(\sum_{j=2}^{i}(s_{j-1}-s_j)t_{j,k}+(s_0-s_1)t_{1,k})} \int_{x=t_{i,k}}^{t_{i+1,k}} e^{-s_i x}\mathrm{d}B^{(k)}(x) + \int_{x=0}^{t_{1,k}} e^{-s_0 x}\mathrm{d}B^{(k)}(x). \quad (4.35)$$

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

$$\widetilde{R}^{(k)}(s) = \Theta^{(k)}\big(\textstyle\sum_{i=1}^{k-1} 1_i(s + \Lambda_i(1 - \widetilde{A}_{p_i}^{(k)}(s))) + s1_k\big), \quad (4.36)$$

where 1_i denotes a k-dimensional row vector whose i-th entry is one and all other entries are zero, and $A_{p_i}^{(k)}$ represents an interruption period occurring within the $X_{p_i}^{(k)}$ portion of the class-k service time (i.e., an interruption period in which only $\mathcal{C}_j$s for $j \le i$ can participate). From Eq. (1.3), we ultimately have

$$\widetilde{A}_{p_i}^{(k)}(s) = \widetilde{\Gamma}\big(s; \Lambda_i, \textstyle\sum_{j=1}^{i}(\lambda_j/\Lambda_i)X^{(j)}\big). \quad (4.37)$$

The first two moments of the auxiliary random variables can be obtained in a straightforward fashion by either differentiating their corresponding LSTs, or by applying the well-known formulas for the first two moments of an $M/G/1$ delay busy period (e.g., Eqs. (1.7) and (1.8)) with the appropriate parameters. Letting $\overline{U}_k = \sum_{i=1}^{k} \lambda_i \mathbb{E}(X^{(i)})$, we obtain for $k = 1, 2, \ldots, m+1$:

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

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

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

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

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

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

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

FETB rule

$$\mathbb{E}(Z_k^{(i)}) = \bigg(\int_{x=\mathcal{T}_{k,i}}^{\infty}(x-\mathcal{T}_{k,i})\,\mathrm{d}B^{(i)}(x)\bigg)/\bar{B}^{(i)}(\mathcal{T}_{k,i}), \qquad i > k,$$

$$\mathbb{E}\big((Z_k^{(i)})^2\big) = \bigg(\int_{x=\mathcal{T}_{k,i}}^{\infty} (x-\mathcal{T}_{k,i})^2\,\mathrm{d}B^{(i)}(x)\bigg)/\bar{B}^{(i)}(\mathcal{T}_{k,i}), \qquad i>k,$$

$$\mathbb{E}(R^{(k)}) = \mathbb{E}(X^{(k)}) + \sum_{i=1}^{k-1}\bigg[\big(B^{(k)}(\mathcal{T}_{i,k}) - B^{(k)}(\mathcal{T}_{i+1,k})\big)\sum_{j=i+1}^{k-1}\Lambda_j\mathbb{E}(A_{p_j}^{(k)})\cdot(\mathcal{T}_{j,k}-\mathcal{T}_{j+1,k}) + \Lambda_i\mathbb{E}(A_{p_i}^{(k)})\bigg((\mathcal{T}_{i,k}-\mathcal{T}_{i+1,k})\bar{B}^{(k)}(\mathcal{T}_{1,k}) + \int_{x=\mathcal{T}_{i+1,k}}^{\mathcal{T}_{i,k}}(x-\mathcal{T}_{i+1,k})\,\mathrm{d}B^{(k)}(x)\bigg)\bigg].$$

TETB rule

$$\mathbb{E}(Z_k^{(i)}) = \int_{x=0}^{t_{k,i}} x\,\mathrm{d}B^{(i)}(x) + t_{k,i}\bar{B}^{(i)}(t_{k,i}), \qquad i>k,$$

$$\mathbb{E}\big((Z_k^{(i)})^2\big) = \int_{x=0}^{t_{k,i}} x^2\,\mathrm{d}B^{(i)}(x) + t_{k,i}^2\bar{B}^{(i)}(t_{k,i}), \qquad i>k,$$

$$\mathbb{E}(R^{(k)}) = \mathbb{E}(X^{(k)}) + \sum_{i=1}^{k-1}\bigg[\Lambda_i\mathbb{E}(A_{p_i}^{(k)})\int_{x=t_{i,k}}^{t_{i+1,k}}(x-t_{i,k})\,\mathrm{d}B^{(k)}(x) + \big(B^{(k)}(t_{i+1,k}) - B^{(k)}(t_{i,k})\big)\sum_{j=1}^{i-1}\Lambda_j\mathbb{E}(A_{p_j}^{(k)})(t_{j+1,k}-t_{j,k})\bigg].$$

4.5 Numerical examples

We now present two numerical examples which illustrate the potential use of our mixed priority queueing model. Our first example takes inspiration from the example found in Stanford et al. (2014). The Canadian Triage and Acuity Scale (CTAS) provides five priority classifications for the triage assessment of patients arriving to a hospital emergency room. Furthermore, each class is given a "time to assessment" standard and an accompanying compliance target, which specifies the desired proportion of that class's patients to meet the standard. Table 4.4 reports these time to assessment standards along with their compliance targets, as taken from Stanford et al. (2014, p. 299).

As an attempt to meet these standards, we model an emergency room whose 5 classes of patients are defined by the CTAS and invoke a mixed priority queueing scheme with $m = 3$ (i.e., $\mathscr{U} = \{1,2,3\}$ and $\mathscr{N} = \{4,5\}$). The service times corresponding to each patient class are assumed to be exponentially distributed with

Table 4.4: CTAS key performance indicators

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

mean times of 30 minutes for class 1, 20 minutes for classes 2 and 3, and 10 minutes for classes 4 and 5. We assume further that the server (or doctor) implements a PB rule to govern how preemptions to patients take place. For the Resuscitation class, we assume that $\alpha_{1,i} = 1$ for $i = 2, 3, 4, 5$ (i.e., $\mathcal{C}_1$s always preempt lower priority customers). We consider several different values for the other threshold parameters such as $\alpha_{2,i}$ for $i = 3, 4, 5$ and $\alpha_{3,i}$ for $i = 4, 5$. The remaining parameters of the system correspond to the accumulating priority rates of the $\mathcal{C}_{\mathcal{N}}$s for which we assume $b_4 = 1$ and $0 \leq b_5 \leq 1$.

For each $k = 1, 2, \ldots, 5$, we are interested in calculating $P(W^{(k)} \leq t_k^*)$, where t_k^* denotes the class-k time to assessment standard given in Table 4.4. To do this, we numerically invert $\widetilde{W}^{(k)}(s)$ by employing the EULER and POST-WIDDER algorithms of Abate and Whitt (1995) with their suggested parameter settings (and found that the two methods produced equivalent results). We remark that in conducting the numerical inversions, there were several instances for which implicit functionals of LSTs (resembling those of an $M/G/1$ busy period) had to be evaluated at complex arguments. This was performed following the iterative procedure outlined in Abate and Whitt (1992). The main details associated with the use of these numerical inversion algorithms are provided in the Appendix. In addition to reporting the desired probabilities, we provide the mean class-k waiting times and flow times for $k = 1, 2, \ldots, 5$. The results under three separate settings are tabulated to 4 decimal places of accuracy in Table 4.5. Note also that the reported values are given in scaled multiples of 10 minutes.

In their example, Stanford et al. (2014) analyzed a 2-class NPAPQ, modelling

Table 4.5: Performance measures in Example 1 under various settings

Class k	λ_k	$P(W^{(k)} \leq t_k^*)$	$\mathbb{E}(W^{(k)})$	$\mathbb{E}(F^{(k)})$
Setting 1 ($\rho = 0.863$)				
$\alpha_{2,3} = 0.9$, $\alpha_{2,4} = 1$, $\alpha_{2,5} = 1$, $\alpha_{3,4} = 0.5$, $\alpha_{3,5} = 0.75$, and $b_5 = 0.10$				
1	0.001	0.9970	0.0090	3.0090
2	0.01	0.9885	0.0511	2.0571
3	0.02	0.9815	0.2775	2.3204
4	0.4	0.8873	2.7217	3.7671
5	0.4	0.6590	11.7522	12.8085
Setting 2 ($\rho = 0.833$)				
$\alpha_{2,3} = 0.75$, $\alpha_{2,4} = 0.9$, $\alpha_{2,5} = 1$, $\alpha_{3,4} = 0.25$, $\alpha_{3,5} = 0.5$, and $b_5 = 0.30$				
1	0.001	0.9970	0.0090	3.0090
2	0.005	0.9931	0.0361	2.0421
3	0.01	0.9832	0.4128	2.4341
4	0.4	0.8308	3.1880	4.2054
5	0.4	0.7781	7.5744	8.5980
Setting 3 ($\rho = 0.815$)				
$\alpha_{2,3} = 0.5$, $\alpha_{2,4} = 0.75$, $\alpha_{2,5} = 1$, $\alpha_{3,4} = 0.25$, $\alpha_{3,5} = 0.5$, and $b_5 = 0.275$				
1	0.001	0.9970	0.0090	3.0090
2	0.001	0.9958	0.0433	2.0494
3	0.005	0.9891	0.3652	2.3733
4	0.4	0.8795	2.6638	3.6709
5	0.4	0.8175	6.4787	7.4888

only CTAS classes 4 and 5. In our treatment, we utilized the same arrival rates and service rates for the two lowest priority classes as in their example. Moreover, they determined that without the presence of the three highest priority classes, the CTAS 4 and 5 compliance targets were both met as long as the accumulating priority rate of the lowest class did not exceed 0.5. As evidenced by the results in Table 4.5, this is not the case for our 5-class priority model. In fact, of the three settings considered, only in Setting 3, where the arrival rates of the 3 highest priority classes are the smallest, were all the CTAS compliance targets satisfied. It is also interesting to observe the changes in the mean flow times under the various settings.

In our second example, we consider the 9-class mixed priority queue studied by Paterok and Ettl (1994, pp. 1157–1159). The arrival rates and service time distributions, including the *priority group* of each class, are given in Table 4.6. Priority groups are used to specify the type of priority that the higher priority customers have over lower priority ones. In particular, a $\mathcal{C}_i$ has preemptive priority over a $\mathcal{C}_j$ $(i < j)$ if they belong to different priority groups; otherwise, the $\mathcal{C}_i$ has only non-preemptive priority over the $\mathcal{C}_j$. It is straightforward to obtain these specific priority relations using our mixed priority model. For example, if we define $\alpha_{(r,s)}$, $\mathcal{T}_{(r,s)}$, and $t_{(r,s)}$ for all $1 \leq r < s \leq 3$ as the threshold-based discretion parameters between priority groups (e.g., $t_{i,j} = t_{(r,s)}$ whenever a $\mathcal{C}_i$ belongs to priority group r and a $\mathcal{C}_j$ belongs to priority group s), then the desired priority relations are achieved by considering a 9-class mixed priority model with $m = 6$ and the following threshold parameters: $\alpha_{(r,s)} = 1$, $\mathcal{T}_{(r,s)} = \infty$, and $t_{(r,s)} = 0$ for all $r < s$. We note that in their analysis, Paterok and Ettl (1994) used a 15-class priority queue for which the arrival rates of six of the classes were set equal to zero in order to obtain the desired priority relations.

We define the weighted average flow time as $\overline{F} = \sum_{i=1}^{9}(\lambda_i/\Lambda_9)\mathbb{E}(F^{(i)})$, and similarly let $\overline{F}_i$ represent the weighted average flow time of classes belonging to priority group i, $i = 1, 2, 3$. In our numerical study, we report the expected flow times of each class, as well as the weighted average flow times under various settings for each of the threshold-based discretion rules. The results for the original Paterok and Ettl (1994) setting (denoted as the resume-IPF case) are tabulated to 3 decimal places of

Table 4.6: Parameters of the Paterok and Ettl (1994) example

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

accuracy in Table 4.7. The results for the PB, FETB, and TETB rules are provided in Tables 4.8, 4.9, and 4.10, respectively.

For the $\mathcal{C}_{\mathscr{N}}$s, we implement accumulating priority rates of the form $b_7 = 1$, $b_8 = e^{-x}$, and $b_9 = e^{-2x}$ for some $x \geq 0$. We note that as $x \to \infty$, the resulting accumulating prioritization becomes equivalent to that of the static non-preemptive priority service discipline. Conversely, with $x = 0$, the $\mathcal{C}_{\mathscr{N}}$s are serviced according to their order of arrival (i.e., regardless of the specific class to which they belong). As a consequence of having $x = 0$, the mean waiting times for each class belonging to the lowest priority level would all be identical — a potentially desirable setting. In Tables 4.7–4.10, we compute mean flow times for each of the non-urgent classes using $x = 0.1, 1, 10$. Similar to the PAPQ and NPAPQ, we emphasize that for this mixed priority queueing model, a systems manager is able to achieve a desired balance between the two extremes of FCFS and static non-preemptive priority between the $\mathcal{C}_{\mathscr{N}}$s by simply fine-tuning the parameter x. We also note that the mean flow times of the $\mathcal{C}_{\mathscr{U}}$s are unaffected by the choice of x.

It is evident from the results in Tables 4.8–4.10 that the new priority model is quite flexible. In testing several different parameter values for each of the threshold-based discretion rules, we are, in some instances, able to achieve a lower overall weighted average flow time $\overline{F}$. Furthermore, if instead a systems manager is more concerned with reducing the average flow time of the lowest priority level $\overline{F}_3$, and

is less concerned with minimizing $\overline{F}$, then it is clear that our priority model can achieve this objective while still maintaining reasonable weighted average flow times for both $\overline{F}_1$ and $\overline{F}_2$.

Table 4.7: Mean flow times in Example 2 under the original Paterok and Ettl (1994) setting

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

Table 4.8: Mean flow times in Example 2 under PB rule

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

Table 4.9: Mean flow times in Example 2 under FETB rule

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

Table 4.10: Mean flow times in Example 2 under TETB rule

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

Chapter 5

Conclusions

In many real-life queueing systems, it is often necessary and/or desirable to provide certain types of customers prompt access to the server(s). For such queueing systems, priority service disciplines are appropriate. The goal of priority queueing systems is to provide shorter wait times for those customers of higher priority. However, an obvious consequence of reducing wait times for the higher priority customers is the increase of that for the lower priority ones. This is the trade-off that systems managers are faced with when designing priority queueing systems. Unfortunately, this trade-off cannot at all be controlled in static priority queues, and so, these systems can oftentimes display poor performance. In an effort to remedy such issues, the central theme of this thesis is the generalization of static priority queueing systems via the concept of accumulating priority.

As evidenced by the research of this thesis, the main benefit in assigning priority via the accumulating priority mechanism is the ability, through the selection of the accumulating priority rates (e.g., the set of parameters $\{b_k\}_{k=1}^{N}$ for the PAPQ), to control the waiting time distributions of each class. Moreover, through the appropriate selection of the accumulating priority rates, both the FCFS and classical static priority service disciplines can be captured. Therefore, by characterizing the waiting time distributions of these dynamic priority queues, we provide systems managers the flexibility to design highly efficient queueing systems that are capable of satisfying a wide variety of system performance goals.

An important tool in the analysis of accumulating priority queues is the maximal priority process. Essentially, the maximal priority process provides a useful structuralization of a queueing system's busy period and the customers serviced within it. It is ultimately with this structuralization that we are able to decompose a customer's waiting time (or equivalently, its accumulated priority level prior to entering service) into several independent components. To obtain the LSTs of these individual components, the methodology employed combined several classical applied probability techniques such as those found in renewal theory, semi-Markov theory, and level-crossing analysis. Therefore, a major contribution of this thesis is that it sheds new light on the maximal priority process, providing a clearer understanding on how it can be used as a tool in the analysis of accumulating priority queues.

In Chapter 2, we analyzed an $M/G/1$ queue under a new blocking policy which we referred to as the q-policy and also highlighted a key connection between the virtual wait process of this system and the maximal priority process of a related $M/G/1$ queue with accumulating priority. This connection, along with the waiting time results established in this chapter, served as the foundation for our subsequent analyses of the accumulating priority queues considered in Chapters 3 and 4. In Chapter 3, we analyzed the fully preemptive accumulating priority queue. In Chapter 4, we analyzed a general mixed priority queueing system in which some classes of customers are assigned priority levels via accumulating priority while others are assigned static priority levels. In both of these chapters, we exploited the relationship between their respective maximal priority processes and the maximal priority process of Chapter 2 to obtain the LSTs of several key random variables of interest (such as $\mathcal{P}^{(acc:k)}$, the pseudo-interruption periods of Chapter 3, and the auxiliary random variables of Chapter 4), required for the overall recursive procedure to obtain the steady-state class-k waiting time LST.

In addition to characterizing the steady-state waiting time distributions, we have established mathematical expressions for the LSTs of several other important random variables of interest such as the service-structure elements (i.e., residence periods, flow times, and gross service times) and the newly defined additional accumulated priority of a class-k interrupting customer $\mathcal{P}^{(int:k)}$. By acquiring probabilistic knowl-

edge of these random variables, we gain more insight into the nuances and technical details of accumulating priority queues (including their advantages over static priority counterparts). Thus, the extensive analyses of these accumulating priority queueing models represent another main contribution of the thesis.

The recent success in characterizing waiting time distributions in accumulating priority queues has given rise to numerous viable future research problems. First of all, if the goal, as queueing theorists believe, is truly to construct efficient queueing systems, then there needs to be a strong sense of responsibility and desire to generalizing and converging currently existing designs. Hence, one notable avenue for future research deals with the continued improvement of previously analyzed static priority queues through the implementation of an accumulating priority mechanism. For example, of particular interest is the analysis of a priority queueing system similar to the one studied in Chapter 4, but with the additional assumption that the priority levels of the $\mathcal{C}_{\mathscr{U}}$s, as well as that of the $\mathcal{C}_{\mathscr{N}}$s, accumulate linearly throughout time. The resulting priority queueing system would be quite flexible, serving as a generalization of the PAPQ and the NPAPQ, as well as those static priority queues mentioned in Chapter 4.

As another example, one can consider the implementation of an accumulating priority mechanism in *polling-type* queues (i.e., systems in which the server serves multiple streams of customers in cyclical fashion). An exceptional source for recent research developments in the implementation of a prioritization structure within polling-type queues is the doctoral thesis by Boon (2011). There, the author illustrates another trade-off which cannot be controlled through static prioritization, namely that the reduction in mean waiting times for higher priority customers leads to greater variability in their waiting times (i.e., increase in the associated coefficient of variation). However, as one might expect, this trade-off can be controlled if an accumulating priority mechanism is implemented. Other future research ideas involve optimization problems, where, for example, one searches for the optimal set of accumulating priority rates $\{b_k\}_{k=1}^{N}$ under a specified objective function, and analyses of dynamic preemptive priority models in which priority levels accumulate in a non-linear fashion.

Before the advent of the maximal priority process, controlling the class-k waiting time of accumulating priority queueing systems (such as the NPAPQ) was limited and essentially administered only through its first moment. Nowadays, as evidenced in the study by Stanford et al. (2014) and the research of this thesis, it is possible to establish the class-k waiting time LSTs for the NPAPQ and several other types of accumulating priority queues. In doing so, it is possible to control several other important aspects of the class-k waiting time distribution, including its higher moments and, perhaps more importantly, its quantiles. Therefore, the recent advancements in the study of accumulating priority queues necessitates their consideration in the endeavour to attain optimal design and functionality of real-life priority queueing systems.

Bibliography

Abate, J. and Whitt, W. 1992. Solving probability transform functional equations for numerical inversion. *Operations Research Letters* 12:275–281.

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

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

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

Asmussen, S. 2008. Applied Probability and Queues, 2nd edition. Springer, New York.

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.

Bhat, U. N. 2008. An Introduction to Queueing Theory: Modeling and Analysis in Applications. Springer, New York.

Boon, M. 2011. From theory to traffic intersections. PhD thesis, Eindhoven University of Technology, Eindhoven, Netherlands.

Brill, P. H. 1975. System-point theory in exponential queues. PhD thesis, University of Toronto, Toronto, Canada.

Brill, P. H. 1988. Single-server queues with delay dependent arrival streams. *Probability in the Engineering and Informational Sciences* 2:231–247.

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

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

Cobham, A. 1954. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America* 2:70–76.

Cohen, J. W. 1977. On up- and downcrossings. *Journal of Applied Probability* 14:405–410.

Cohen, J. W. 1982. The Single Server Queue, 2nd edition. North-Holland, Amsterdam.

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

Doshi, B. T. 1977. Continuous time control of the arrival process in an *M/G/1* queue. *Stochastic Processes and Their Applications* 5:265–284.

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

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

Drekic, S. and Stanford, D. A. 2001. Reducing delay in preemptive repeat priority queues. *Operations Research* 49:145–156.

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

Fajardo, V. A. and Drekic, S. 2015b. On a general mixed priority queue with server discretion. *Submitted to: Stochastic Models* .

Fajardo, V. A. and Drekic, S. 2015c. Waiting time distributions in the preemptive accumulating priority queue. *Methodology in Computing and Applied Probability (in press)* .

Giambene, G. 2005. Queuing Theory and Telecommunications, 2nd edition. Springer, New York.

Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M. 2008. Fundamentals of Queueing Theory, 4th edition. Wiley, New Jersey.

Hass, J., Weir, M. D., Thomas, G. B., and Brinton, G. 2007. University Calculus. Pearson Addison-Wesley, Boston.

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.

Johansen, S. G. and Stidham, S. 1980. Control of arrivals to a stochastic input-output system. *Advances in Applied Probability* 12:972–999.

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

Kao, E. 1996. An Introduction to Stochastic Processes. Duxbury Press, Belmont.

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.

Kendall, D. G. 1951. Some problems in the theory of queues. *Journal of the Royal Statistical Society, Series B (Methodological)* 13:151–185.

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

Kleinrock, L. 1975. Queueing Systems, Volume I: Theory. Wiley, New York.

Kleinrock, L. 1976. Queueing Systems, Volume II: Computer Applications. Wiley, New York.

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

Knudsen, N. C. 1972. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* 40:515–28.

Lakshmi, C. and Iyer, S. A. 2013. Application of queueing theory in health care: a literature review. *Operations Research for Health Care* 2:25–39.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* 37:15–24.

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

Parzen, E. 1962. Stochastic Processes. Holden-Day, San Francisco.

Paterok, M. and Ettl, M. 1994. Sojourn time and waiting time distributions for *M/GI/1* queues with preemption-distance priorities. *Operations Research* 42:1146–1161.

Prabhu, N. 1997. Foundations of Queueing Theory. Kluwer Academic Publishers, Boston.

Rue, R. C. and Rosenshine, M. 1981. Some properties of optimal control policies for entry to an *M/M/1* queue. *Naval Research Logistics Quarterly* 28:525–532.

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.

Stidham, S. 1985. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* 30:705–713.

Stidham, S. 2002. Analysis, design, and control of queueing systems. *Operations Research* 50:197–216.

Takács, L. 1962. Introduction to the Theory of Queues. Oxford University Press, New York.

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.

White, H. and Christie, L. S. 1958. Queuing with preemptive priorities or with breakdown. *Operations Research* 6:79–95.

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

Yechiali, U. 1971. On optimal balking rules and toll charges in the *GI/M/1* queuing process. *Operations Research* 19:349–370.

The Appendix

To conduct the numerical inversions of the LSTs derived in this thesis, we implement the EULER and POST-WIDDER algorithms developed by Abate and Whitt (1995). Specifically, each of these algorithms provide the means to compute, for various values of $t > 0$, $f(t)$ from its LST

$$\widetilde{f}(s) = \int_0^\infty e^{-st} f(t) \mathrm{d}t, \tag{A.1}$$

where s is a complex number with a non-negative real part. Since the EULER and POST-WIDDER methods are based on two different mathematical approaches to invert Eq. (A.1), they can be used together to confirm the overall accuracy of the numerical inversion (i.e., the computations resulting from the EULER and POST-WIDDER methods should agree within a desired precision). We next describe in brief these two numerical inversion methods.

The EULER method provides an approximation to the Bromwich contour inversion integral, which can be expressed as

$$f(t) = \frac{2e^{at}}{\pi} \int_0^\infty \Re(\widetilde{f}(a + iu)) \cos(ut) \, \mathrm{d}u, \tag{A.2}$$

where $\Re(s)$ is the real part of s and a is chosen so that the vertical line $s = a$ is such that $\widetilde{f}(s)$ has no singularities on or to the right of it. In particular, the EULER method computes an approximation of the right-hand side of Eq. (A.2) via the following two steps: (i) apply the well-known *trapezoidal rule* (e.g., see Hass et al. (2007, p. 479)) with $h = \pi/2t$ to the right-hand side of Eq. (A.2) and (ii) use Euler summation to accelerate the convergence of the infinite sum involved in the

approximation employed in (i). The final approximation computed via the EULER method is given by

$$f(t) \approx E(m,n,t) = \sum_{k=0}^{m} \binom{m}{k} 2^{-m} s_{n+k}(t), \tag{A.3}$$

where

$$s_\ell(t) = \frac{e^{A/2}}{2t}\Re\Big(\widetilde{f}\Big(\frac{A}{2t}\Big)\Big) + \frac{e^{A/2}}{t}\sum_{j=1}^{\ell}(-1)^j \Re\Big(\widetilde{f}\Big(\frac{A+2j\pi i}{2t}\Big)\Big), \quad \ell > 0. \tag{A.4}$$

Note the re-parametrization of $a = A/2t$. The parameter A controls the discretization error of the approximation in step (i) above. Abate and Whitt (1995) suggest using $A = 18.4$, $m = 11$, and $n = 15$ to achieve a discretization error of 10^{-8}.

The POST-WIDDER method is based on the so-called POST-WIDDER Theorem which provides a sequence of functions $\{f_n(t)\}_{n=1}^{\infty}$ that converge to $f(t)$ as $n \to \infty$, namely

$$f_n(t) = \frac{(-1)^n}{n!}\left(\frac{n+1}{t}\right)^{n+1} \widetilde{f}^{(n)}((n+1)/t), \tag{A.5}$$

where $\widetilde{f}^{(n)}(s)$ is the n-th derivative of $\widetilde{f}(s)$. Note that it is possible to re-express Eq. (A.5) so as to involve an integral over a finite interval of real values. Hence, by subsequently applying the trapezoidal rule (with $h = \pi/n$) to this alternate expression, the following approximation to Eq. (A.5) is obtained:

$$f_n(t) \approx \frac{n+1}{2tnr^n}\Bigg(\widetilde{f}((n+1)(1-r)/t) + (-1)^n\widetilde{f}((n+1)(1+r)/t) + 2\sum_{k=1}^{n-1}(-1)^k \Re\Big(\widetilde{f}\Big(\frac{n+1}{t}(1-re^{\pi ik/n})\Big)\Big)\Bigg), \quad r > 0. \tag{A.6}$$

We remark that the parameter r in Eq. (A.6) controls the discretization error of this approximation to Eq. (A.5). To enhance the accuracy of the inversion, the POST-WIDDER method utilizes a linear combination of $f_n(t)$ for various values of $n > 0$. Therefore, the final approximation of $f(t)$ provided by the POST-WIDDER method consists of three parameters (r, j, m) and is given by

$$f(t) \approx \sum_{k=1}^{m}(-1)^{m-k}\frac{k^m}{k!(m-k)!} f_{j\cdot k}(t). \tag{A.7}$$

To achieve a discretization error around 10^{-8}, Abate and Whitt (1995) suggest using $r = 10^{-4}$, $j = 10$, and $m = 6$.

It is important to realize that in order to extract $f(t)$ from $\widetilde{f}(s)$ via either inversion algorithm, one must be able to evaluate the real part of $\widetilde{f}(s)$ for specific values of complex s. This task is straightforward when $\widetilde{f}(s)$ is given in an explicit form. However, in queueing related problems, it is often the case that $\widetilde{f}(s)$ is defined via an implicit function. For example, evaluating $\widetilde{f}(s)$ could involve the evaluation of the implicit functional corresponding to the LST of an $M/G/1$ busy period $\widetilde{\Gamma}(s)$, as defined by Eq. (1.3). To evaluate $\widetilde{\Gamma}(s)$ for complex s, we employ the iterative procedure described by Abate and Whitt (1992): starting with $D_0(s) = 0$ (or with $D_0(s) = 1$), recursively compute $D_n(s)$ via

$$D_n(s) = \widetilde{B}(s + \lambda - \lambda D_{n-1}(s)), \quad n > 0. \tag{A.8}$$

The fact that $D_n(s)$ converges to $\widetilde{\Gamma}(s)$ as $n \to \infty$ was proven by Abate and Whitt (1992, Section 2, Theorem 3). Furthermore, they showed that satisfactory accuracy can be obtained after performing a modest number of iterations. In this thesis, we used 30 such iterations.