On completion times in a two-class priority queue with impatience

Ioane Muni Toke

ERIM,
University of New Caledonia,
BP R4, 98851 Noumea Cedex, New Caledonia
and
MAS Laboratory,
Ecole Centrale Paris,
Grande Voie des Vignes,
92290 Chatenay-Malabry, France
E-mail: ioane.muni-toke@univ-nc.nc

Abstract: In this note, we consider a two-class priority queueing system with Poisson arrivals, general service time distribution and one server, in which customers that are not currently being served may leave the queue according to exponentially distributed patience times, i.e., a M_1, $M_2/G/1 + M$ system using a generalised Kendall's notation. We expand the classic methodology to derive analytical formulas for the completion times in such a system, using preemptive repeat different and preemptive repeat identical disciplines. Known average completion times for priority queues without impatience are retrieved as limit cases.

Keywords: priority queues; queues with impatience; reneging; completion times; preemptive disciplines.

Reference to this paper should be made as follows: Toke, I.M. (2014) 'On completion times in a two-class priority queue with impatience', *Int. J. Mathematics in Operational Research*, Vol. 6, No. 3, pp.377–392.

Biographical notes: Ioane Muni Toke is a faculty member of the Science and Technology Department and a member of ERIM (research team in Mathematics and Computer Science) at the University of New Caledonia. He is also associated with the Chair of Quantitative Finance of the MAS Laboratory (Applied Mathematics), Ecole Centrale Paris, France. His research interests include applied probability models used in quantitative finance and queueing applications, with an emphasis on the empirical study and stochastic modelling of electronic limit order books.

1 Introduction

Queueing systems able to deal with several types of prioritised customers and/or their impatience are well known to be applicable to a wide range of fields. Among recent contributions, let us mention the use of priority queueing systems in inventory management (Zhao and Lian, 2011), IT support management (Zeltyn et al., 2009; Bartolini et al., 2012), or cognitive radio networks (Kim, 2012). Impatience, i.e., the

possibility that a customer leaves the queue without having received a full service, is a key feature in call centres management [see e.g., Aksin et al. (2007) for a multi-disciplinary review of the field] or healthcare management (Wang, 2004) for example. Queueing models involving both priority and impatience raise complex challenges. Some of them are tackled in the literature of design, control and scheduling of queueing systems [see e.g., Ata and Tongarlak (2012) and Kim (2012) for recent contributions]. Some of these results may be applicable to call centres or healthcare management [e.g., Jouini et al. (2010) among others].

This note deals with the exact analysis of a priority queue with impatient customers and a preemptive discipline. We consider a general two-class priority queueing system with Poisson arrivals, general service time distribution and one server. We assume that class-1 customers are the highest-priority customers, class-2 customers the lowest-priority ones and that a preemptive discipline is used. Each customer in the queue, i.e., *all customers in the system except the one currently being served*, may leave the system at any time. This is the impatience/reneging phenomenon. Patience times of customers that leave the queue without having completed service are assumed to be exponentially distributed. Such a system might therefore be noted M_1, $M_2/G/1 + M$ using a generalised Kendall's notation.

Explicit analysis of the stationary properties of such a system has not been fully completed yet. In this paper, we contribute to this study by deriving an analytical form of the Laplace-Stieltjes transform of the completion times in this system. The notion of completion time in such a system refers to the length of the time interval between the moment a customer starts service and the moment he leaves the system, which may be upon service completion or out of impatience. Results are obtained for both preemptive repeat different and preemptive repeat identical disciplines. Classic results of completion times in priority queues without impatience may be retrieved as limit cases.

1.1 Notations

Throughout the paper we will use the following notations. $\mathbb{N}$ is the set of natural integers, and $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For any complex number $p \in \mathbb{C}$, $\mathfrak{Re}(p)$ is its real part. For any positive probability measure μ on the positive half-line $\hat{\mu}$ denotes its Laplace-Stieltjes transform (sometimes abbreviated LST from now on), i.e.

$$\hat{\mu}(p) = \int_{\mathbb{R}+} e^{-pt} d\mu(t), \quad \forall p \in \mathbb{C},\ \mathfrak{Re}(p) \geq 0. \tag{1}$$

$\overline{\mu}$ denotes the (eventually infinite) expectation of the probability distribution μ. Finally, all random variables are assumed to be defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

2 Literature review

Priority disciplines have been largely addressed in queueing theory, as well as impatience/reneging/balking in a single queue, but few studies address the exact analysis of both difficulties at the same time.

As for priority queues *without impatience*, pioneering works of Gaver (1962) and Chang (1965) introduce the notion of completion times, which is defined as the length of

the time interval between the time a service starts and the time it is completed. Classic analysis goes on with the analysis in stationary state of queues sizes and waiting times, defined as the length of the time interval between the arrival of a customer in the queue and the moment he enters service. Standard results on priority queues may be found in textbook form in, e.g., Jaiswal (1968) or Takagi (1991).

There is a large body of literature on the impatience phenomenon in a queue of customers *without priority* [note that we focus in this paper on single-server queues, but impatience/reneging is obviously also of interest in a multi-server setting, see e.g., Choudhury and Medhi (2011)]. Baccelli et al. (1984) derives stability conditions for general queues with impatience. Impatience in these queues may refer to the waiting time (in which case a customer stays until service completion once its service has started) or to the sojourn time (in which case an impatient customer may leave even while being served). Boxma et al. (2011) recently proposed to write $G/G/1 + G^w$ in the first case and $G/G/1 + G^s$ in the later one. Daley (1965), building on Kovalenko (1961), derives an integral equation for the waiting time of a customer, defined as the length of the time interval between its arrival in the queue and the moment he enters service or the moment he leaves out of impatience, whichever happens first. In particular, waiting times in the cases of deterministic ($M/G/1 + D^w$) and exponential ($M/G/1 + M^w$) impatience distributions are studied. The busy period of such systems, i.e., the time interval between the arrival of a customer in an empty system and the moment the system is empty again, will be of particular importance in this paper. Bae et al. (2001) provides the expected busy period of the $M/G/1 + D^w$ queue. Also using deterministic impatience times, Xiong et al. (2008) provides an exact analysis of the special case in which service times are two-stage hyper-exponentially distributed ($M/H_2/1 + D^w$). Rao (1967) is a pioneer paper in which the Laplace-Stieltjes transform of the busy period of the $M/G/1 + M^w$ queue is derived. Very recently, Brandt and Brandt (2011) retrieves this result as a special case of a joint analysis of the workloads and busy times of such systems. The 'balking' case, or 'restricted accessibility' case, in which the customer does not enter the queue if the workload seen upon its arrival exceeds his patience, is investigated in Perry et al. (2000) with a deterministic patience. Boxma et al. (2010) extends this study, and among other results derives the Laplace-Stieltjes transform of the busy period in an $M/G/1 + M^w$ queue in the restricted accessibility case. Very recently, Boxma et al. (2011) sheds some new light on the $M/G/1 + G^w$ and $M/G/1 + G^s$ queue, analysing it with a modified system in which customers that ran out of impatience stay in the system with an arbitrarily small deterministic service time. This analysis provides, among other results, the distribution of the number of customers in the queue and the distribution of the busy period of the system. Finally, let us mention that Ward (2012) reviews asymptotics of such systems in a recent survey.

Research mixing both a fixed priority discipline and impatience in the same system is much less abundant, especially when it comes to the exact analysis of the queueing system. Choi et al. (2001) addresses a two-class problem, in which the first class has higher-priority and deterministic impatience, whereas the lower-priority class has no impatience. All inter-arrival times distributions are exponential, including service. Studied systems are thus M_1, $M_2/M/1 + D^w$ and M_1, $M_2/M/1 + D^s$, in which the lower priority class has no impatience. Joint distribution of queue sizes and Laplace-Stieltjes transforms of total response times for both classes of customers are obtained. Brandt and Brandt (2004) generalises the previous results to a general distribution of patience time,

i.e., in the $M_1, M_2/M/1 + G^w$ case. Note however that this is still assuming an exponential service time and no impatience for the lower-priority customers. The system studied by Iravani and Balcioglu (2008) is closer to the one under scrutiny here: a two-class problem with general service time distribution is studied, in which both classes exhibit exponentially distributed impatience. Using a level-crossing method, waiting time distributions are obtained, assuming a preemptive-resume policy in which a customer that started service may not leave the system.

The system we analyse here differs from the previous ones in several key points. Contrary to Choi et al. (2001) or Brandt and Brandt (2004), service time is general, and an important feature is that all customers, including lower priority customers, exhibit impatience. Moreover, as opposed to Iravani and Balcioglu (2008), our system allows that a class-2 customer may leave the system even if its service started: if such a customer, who started service, is left by the server called by a class-1 priority customer (preemptive discipline), then it does not wait indefinitely but may leave the system out of impatience. In other words, in our system, impatience of the class-1 customers applies to their waiting time, and impatience of the class-2 customers applies to their sojourn time, with the restriction that they cannot leave while being served. This feature is very realistic in many standard queueing applications such as call centres, and is not taken into account in cited works, which assume that any customer starting service will stay in the system until its service is completed, even during a preemption, however long that may be. Thus, our system has two types of reneging customers: those that leave before they even start service, and those that leave out of impatience after their service started but was interrupted. From a customer satisfaction point of view, the latter case might be the worst (most frustrating) case. Thus in what follows, we compute 'completion' times (in a generalised sense) for all class-2 customers that have started service, whichever way they left the system.

Note that, as usual with two-class priority queues with preemptive discipline, high-priority customers are not influenced by lower-priority customers. Therefore, all class-1 customers that start service stay in the system until completion, since they cannot be preempted. For these customers, there cannot be any reneging during a preemption, the phenomenon that we will study for lower priority customers. Therefore, their completion time distribution is the service time distribution. Moreover, the queueing system formed by the sole customers of class-1 is obviously a standard $M/G/1$ queue with exponential impatience of the customers waiting in line. Such a system is exactly the one studied by Rao (1967), where the Laplace-Stieltjes transform of the busy period of the system is derived. We use this result to derive the completion times of class-2 customers.

The remainder of the note is organised as follows. Section 3 defines notations describing the model and states some useful results for the subsequent computations. Section 4 proves analytical formulas for the Laplace-Stieltjes transforms of completion times of class-2 customers when a preemptive repeat different discipline is used. Our approach follows and generalises the classic study of completion times in priority queues used by Gaver (1962), Chang (1965), Jaiswal (1968) and Takagi (1991). We show that the classic results without impatience are retrieved as limit cases of the formulas we obtain. Section 5 states similar results for the preemptive repeat identical discipline. Since it is very close to the previous one, the proof is only sketched. Finally, Section 6 provides specific formulas in the particular case where service times are exponentially distributed.

3 Model and preliminary results

We consider a single-server two-class priority queueing system. Arrivals of class-k customers, $k \in \{1, 2\}$, follow independent Poisson processes with parameter λ_k. A unique server deals with the queue, and service is based on a first-come-first-served discipline. Service times are assumed to be a set of mutually independent random variables, identically distributed with probability distribution B with positive support. Furthermore, class-1 customers are the highest-priority customers, class-2 customers the lowest-priority ones, and a preemptive discipline is used. Consequently, class-2 customers cannot be served as long as there are customers of class-1 in the system; and if a class-1 customer enters the system while the server is attending a class-2 customer, then he is immediately attended by the server, and service for the class-2 customer is interrupted. Finally, customers of any class that are not currently being served and that have not yet completed service may leave the queue at any time. If a customer leaves before its service is completed, the time interval between its arrival and its departure is assumed to be distributed according to an exponential distribution with parameter θ strictly positive. Let $\nu_k = \frac{\lambda_k}{\theta}$, $k \in \{1,2\}$ denote the arrival rate of customers of class k normalised by the impatience rate 0.

As observed above, the queuing system formed by the sole customers of type 1 (referred to as the '1-queue') is a well known $M/G/1$ queue with exponential impatience of the customers waiting in line. Following Rao, (1967, Section 5), one can derive for such a queue the Laplace-Stieltjes transform of the distribution of the length of the busy period starting with one customer, i.e., the time interval between the arrival of a class-1 customer arriving in an empty 1-queue and the first time this 1-queue becomes empty again (idle server). Let Ξ_1 denote the probability distribution of such a busy period. Rao [1967, equation (19)] states that the Laplace transform $\widehat{\Xi}_1$ is written using notations introduced above:

$$\widehat{\Xi}_1(p) = \frac{\widehat{B}(p) + \sum_{r=1}^{\infty} \frac{\nu_1^r}{r!} \psi_{r-1}(p) \widehat{B}(p + r\theta)}{1 + \sum_{r=1}^{\infty} \frac{\nu_1^r}{r!} \psi_{r-1}(p)} \tag{2}$$

where we have defined:

$$\psi_r(p) = \prod_{j=0}^{r} \left(1 - \widehat{B}(p + j\theta)\right). \tag{3}$$

Note that a new derivation of this result has recently been proposed by Brandt and Brandt (2011). By differentiation at the origin, Rao [1967, equation (22)] is able to compute the mean length $\overline{\Xi}_1$ of this busy period:

$$\overline{\Xi}_1 = \overline{B} \sum_{r=0}^{\infty} \frac{\nu_1^r}{r!} \psi_{r-1}(\theta). \tag{4}$$

We now give results concerning conditional distribution of random variables compared to exponentially distributed random variables. Although these are straighforward results, we choose to state them as a lemma since we will explicitly rely on them for the computations in the proofs to follow:

Lemma 1: Let T be a random variable with exponential distribution with parameter X. Let Z be a random variable with general probability distribution F_Z with positive support. Z and T are assumed to be independent. Then:

1 conditionally on being lower than Z, the distribution of T admits the conditional LST:

$$\widehat{F}_{T|T\leq Z}(p)=\frac{\lambda}{p+\lambda}\frac{1-\widehat{F_Z}(p+\lambda)}{1-\widehat{F_Z}(\lambda)} \tag{5}$$

2 conditionally on being lower than T, the distribution of Z admits the conditional LST:

$$\widehat{F}_{Z|Z\leq T}(p)=\frac{\widehat{F_Z}(p+\lambda)}{\widehat{F_Z}(\lambda)} \tag{6}$$

3 conditionally on being greater that T, the distribution of Z admits the conditional

$$\widehat{F}_{Z|Z\geq T}(p)=\frac{\widehat{F_Z}(p)-\widehat{F_Z}(p+\lambda)}{1-\widehat{F_Z}(\lambda)}. \tag{7}$$

Proof: Direct computations. □

4 Completion times of class-2 customers in the case of a preemptive repeat different discipline

In this section, we assume that a preemptive repeat different discipline is used, i.e., when a class-2 customer that started service earlier prior to be preempted resumes service, it is assumed its remaining service time is chosen anew following probability distribution *B* and does not depend on previous time spent being served before. Completion time is defined as the time interval between the first time a customer enters service and the time it leaves the system. Let us recall that in all the cited references, the latter event may only be service completion. In our case however, a customer that started service, but was denied service completion because of the arrival of higher priority customers, may leave the system out of impatience. We will thus distinguish three completion times (for class-2 customers):

- C_2^S denotes the probability distribution of the time interval between the first time a customer enters service and the time of its service completion, conditionally on this service completion

- C_2^R denotes the probability distribution of the time interval between the first time a customer enters service and the time of its departure out of impatience while waiting for the end of a busy period in the virtual 1-queue, conditionally on this reneging
- C_2 denotes the probability distribution of the time interval between the first time a customer enters service and the time it leaves the system, unconditionally.

Let us consider one customer of class-2 entering service. Obviously, if A denotes the event that this customer leaves after having received full service, then:

$$C_2 = C_2^S \mathbf{P}(A) + C_2^R (1 - \mathbf{P}(A)), \tag{8}$$

which is immediately translated by linearity in terms of LST. Let A_n, $n \in \mathbb{N}$, be the event 'the customer is preempted n times and then completes service' and B_n, $n \in \mathbb{N}$, be the event 'the customer is preempted n times and then leaves out of impatience during the n^{th} preemption'. After $i - 1$ preemptions, at the time of the i^{th} service re-entrance, if S_i denotes the service time (distributed according to B), T_i the time to the (virtual) next preemption (exponentially distributed with parameter λ_i), then there is a $i + 1^{\text{th}}$ preemption with probability

$$\mathbf{P}\left(T_i \le S_i\right) = 1 - \hat{B}\left(\lambda_1\right). \tag{9}$$

Similarly, upon the i^{th} preemption, the probability that the customer reneges during this i^{th} preemption is $1 - \hat{\Xi}_1(\theta)$. Because of the independence of the service times and reneging times, we straightforwardly have a geometric framework:

$$\forall n \in \mathbb{N}, \mathbf{P}\left(A_n\right) = \left(1 - \hat{B}\left(\lambda_1\right)\right)^n \left(\hat{\Xi}_1(\theta)\right)^n \hat{B}\left(\lambda_1\right), \tag{10}$$

$$\forall n \in \mathbb{N}^*, \mathbf{P}\left(B_n\right) = \left(1 - \hat{B}\left(\lambda_1\right)\right)^n \left(\hat{\Xi}_1(\theta)\right)^{n-1} \left(1 - \hat{\Xi}_1(\theta)\right). \tag{11}$$

Therefore, the probability that the class-2 customer considered here receives full service is

$$\mathbf{P}(A) = \sum_{n=0}^{\infty} \mathbf{P}\left(A_n\right) = \frac{\hat{B}\left(\lambda_1\right)}{1 - \hat{\Xi}_1(\theta)\left(1 - \hat{B}\left(\lambda_1\right)\right)}, \tag{12}$$

where $\hat{\Xi}_1(\theta)$ is given by equation (2). The probability that it leaves out of impatience during preemption is

$$\mathbf{P}(B) = \sum_{n=1}^{\infty} \mathbf{P}\left(B_n\right) = 1 - \mathbf{P}(A).$$

We now extend the classic computations of Laplace-Stieltjes transforms of completion times to our case with impatience. Let us assume that the class-2 customer entering service is preempted exactly n times by higher-priority customers before leaving, either

after service completion (event A_n, $n \geq 0$) or during the n^{th} preemption (event B_n, $n \geq 1$). Its completion time can be written conditionally on A_n:

$$C_2 = \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \xi_i + \epsilon_n, \tag{13}$$

and conditionally on B_n:

$$C_2 = \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n-1} \xi_i + \gamma_n, \tag{14}$$

where α_i is the time spent in service before the i^{th} preemption, ξ_i the length of the i^{th} preemption, ϵ_n is the non-preempted service time of the customer on the event A_n, and γ_n is the time interval between the n^{th} preemption and the departure of the customer out of impatience on the event B_n. $\{\alpha_i\}_i$ and $\{\xi_i\}_i$ are both families of independent and identically distributed random variables. Therefore, we have:

$$\mathbf{E}\left[e^{-pC_2} \mid A_n\right] = \left(\mathbf{E}\left[e^{-p\alpha_1} \mid A_n\right]\right)^n \left(\mathbf{E}\left[e^{-p\xi_1} \mid A_n\right]\right)^n \mathbf{E}\left[e^{-p\epsilon_n} \mid A_n\right], \tag{15}$$

and

$$\mathbf{E}\left[e^{-pC_2} \mid B_n\right] = \left(\mathbf{E}\left[e^{-p\alpha_1} \mid B_n\right]\right)^n \left(\mathbf{E}\left[e^{-p\xi_1} \mid B_n\right]\right)^{n-1} \mathbf{E}\left[e^{-p\gamma_n} \mid B_n\right]. \tag{16}$$

Now, $\{\alpha_i\}_i$, $\xi_i\}_i$, ϵ_n and γ_n all have a distribution that belongs to one of the three cases of Lemma 1, which leads to:

$$\mathbf{E}\left[e^{-p\alpha_1} \mid A_n\right] = \mathbf{E}\left[e^{-p\alpha_1} \mid B_n\right] = \frac{\lambda_1}{p+\lambda_1} \frac{1-\widehat{B}\left(p+\lambda_1\right)}{1-\widehat{B}\left(\lambda_1\right)} \tag{17}$$

$$\mathbf{E}\left[e^{-p\xi_1} \mid A_n\right] = \mathbf{E}\left[e^{-p\xi_1} \mid B_n\right] = \frac{\widehat{\Xi}_1(p+\theta)}{\widehat{\Xi}_1(\theta)}, \tag{18}$$

$$\mathbf{E}\left[e^{-p\epsilon_n} \mid A_n\right] = \frac{\widehat{B}\left(p+\lambda_1\right)}{\widehat{B}\left(\lambda_1\right)}, \tag{19}$$

$$\mathbf{E}\left[e^{-p\gamma_n} \mid B_n\right] = \frac{\theta}{p+\theta} \frac{1-\widehat{\Xi}_1(p+\theta)}{1-\widehat{\Xi}_1(\theta)}. \tag{20}$$

Finally, taking the LST of equations (13) and (14) and combining the result with equations (8), (10) and (11) yields the result that we state in the following proposition.

Proposition 1: In the M_1, M_2/G/1 + M queue with preemptive repeat different discipline, the Laplace transform of the distribution C_2^S is given by:

$$\widehat{C}_2^S(p) = \frac{\widehat{B}(p+\lambda_1)}{\widehat{B}(\lambda_1)} \frac{1-\widehat{\Xi}_1(\theta)\left(1-\widehat{B}(\lambda_1)\right)}{1-\dfrac{\lambda_1}{p+\lambda_1}\widehat{\Xi}_1(p+\theta)\left(1-\widehat{B}(p+\lambda_1)\right)}. \tag{21}$$

The Laplace transform of the distribution C_2^R *is given by:*

$$\begin{aligned}\widehat{C}_2^R(p) = {} & \frac{\theta}{p+\theta}\frac{\lambda_1}{p+\lambda_1}\frac{1-\widehat{B}(p+\lambda_1)}{1-\widehat{B}(\lambda_1)}\frac{1-\widehat{\Xi}_1(p+\theta)}{1-\widehat{\Xi}_1(\theta)} \\ & \times \frac{1-\widehat{\Xi}_1(\theta)\left(1-\widehat{B}(\lambda_1)\right)}{1-\dfrac{\lambda_1}{p+\lambda_1}\widehat{\Xi}_1(p+\theta)\left(1-\widehat{B}(p+\lambda_1)\right)}.\end{aligned} \tag{22}$$

Finally, the Laplace transform of the unconditional distribution C_2 *is given by:*

$$\begin{aligned}\widehat{C}_2(p) = {} & \frac{\widehat{B}(p+\lambda_1)}{1-\dfrac{\lambda_1}{p+\lambda_1}\widehat{\Xi}_1(p+\theta)\left(1-\widehat{B}(p+\lambda_1)\right)} \\ & + \frac{\dfrac{\theta}{p+\theta}\dfrac{\lambda_1}{p+\lambda_1}\left(1-\widehat{B}(p+\lambda_1)\right)\left(1-\widehat{\Xi}_1(p+\theta)\right)}{1-\dfrac{\lambda_1}{p+\lambda_1}\widehat{\Xi}_1(p+\theta)\left(1-\widehat{B}(p+\lambda_1)\right)}.\end{aligned} \tag{23}$$

Note that equations (21) and (23) are actually still valid for a class-1 customer, since letting 'λ_0' $\to 0$ yields $\widehat{C}_1^S = \widehat{C}_1 = \widehat{B}$.

Direct differentiation yields average completion times. The average value $\widehat{C}_2^S$ of the completion time of a class-2 customer leaving the system after receiving full service is:

$$\overline{C}_2^S = \frac{\widehat{B}'(\lambda_1)}{\widehat{B}(\lambda_1)}\frac{\widehat{\Xi}_1(\theta)-1}{1-\widehat{\Xi}_1(\theta)\left(1-\widehat{B}(\lambda_1)\right)} + \frac{\left(1-\widehat{B}(\lambda_1)\right)\left(\dfrac{1}{\lambda_1}\widehat{\Xi}_1(\theta)-\widehat{\Xi}'_1(\theta)\right)}{1-\widehat{\Xi}_1(\theta)\left(1-\widehat{B}(\lambda_1)\right)}. \tag{24}$$

This result is a generalisation of the known result without impatience, which we may retrieve as a limit case. By letting $\theta \to 0$ in equation (24), we obtain:

$$\lim_{\theta\to 0}\overline{C}_2^S = \frac{1-\widehat{B}(\lambda_1)}{\widehat{B}(\lambda_1)}\left(\frac{1}{\lambda_1}+\overline{\Xi}_1^0\right), \tag{25}$$

where $\overline{\Xi}_1^0 = \overline{B}/(1-\lambda_1\overline{B})$ is the mean busy period of the 1-queue without impatience, i.e., the limit of $\overline{\Xi}_0$ defined at equation (4) as $\theta \to 0$. Equation (25) is the known average completion time in a priority queue without impatience, see e.g., Jaiswal [1968, Chapter 4, equation (7.8)] or Takagi [1991, Chapter 3, equation (4.70a)].

The average value $\overline{C}_2^R$ of the completion time of a class-2 customer leaving the system out of impatience while its service is being preempted is:

$$\overline{C}_2^R = \frac{1}{\theta} + \frac{\dfrac{\hat{B}(\lambda_1)}{1-\hat{B}(\lambda_1)} + \dfrac{1}{\lambda_1} + \dfrac{\hat{B}(\lambda_1)\hat{\Xi}'_1(\theta)}{1-\hat{\Xi}_1(\theta)}}{1-\hat{\Xi}_1(\theta)\left(1-\hat{B}(\lambda_1)\right)}. \tag{26}$$

Letting $\theta \to +\infty$ in equation (26) obviously gives:

$$\lim_{\theta \to +\infty} \overline{C}_2^R = \frac{\hat{B}(\lambda_1)}{1-\hat{B}(\lambda_1)} + \frac{1}{\lambda_1}, \tag{27}$$

which is exactly the expectation of a random variable with density

$$\frac{\lambda_1 e^{-\lambda_1 x}\left(1-B(x)\right)}{1-\hat{B}(\lambda_1)}, \tag{28}$$

i.e., a random variable with exponential distribution with rate λ_1, conditionally on the fact that it is lower than an independent random variable with distribution B. In other words, if the impatience is high, a customer entering service will leave as soon as it is preempted.

Finally, the average (unconditional) completion time $\overline{C}_2$ is:

$$\overline{C}_2 = \frac{\left(1-\hat{B}(\lambda_1)\right)\left(1+\nu_1\left(1-\hat{\Xi}_1(\theta)\right)\right)}{\lambda_1\left(1-\hat{\Xi}_1(\theta)\left(1-\hat{B}(\lambda_1)\right)\right)}. \tag{29}$$

Contrary to the previous averages (24) and (26), this result does not depend on the derivatives of the LST of the service time and busy period distribution, which will lead to simple formulas in Section 6.

5 Completion times of class-2 customers in the case of a preemptive repeat identical discipline

We now compute the Laplace transform of the completion time distribution in the case of a preemptive repeat identical discipline. In this discipline, service time is determined the first time a customer enters service and will stay the same if the customer has to re-enter service after preemptions. Therefore, we obtain the Laplace transforms of completion times using the same proof as in the previous section, conditioning this time on the service time S_1. Result is stated in the following proposition.

Proposition 2: In the M_1, M_2/G/1 + M queue with preemptive repeat identical discipline, the Laplace transform of the distribution C_2^S is given by:

$$\widehat{C}_2^S(p)=\int_0^\infty \frac{1-\widehat{\Xi}_1(\theta)\left(1-e^{-\lambda_1 x}\right)}{1-\dfrac{\lambda_1}{p+\lambda_1}\widehat{\Xi}_1(p+\theta)\left(1-e^{-(p+\lambda_1)x}\right)}e^{-px}dB(x). \tag{30}$$

The Laplace transform of the distribution C_2^R *is given by:*

$$\begin{aligned}\widehat{C}_2^R(p)=&\frac{\theta}{p+\theta}\frac{\lambda_1}{p+\lambda_1}\frac{1-\widehat{\Xi}_1(p+\theta)}{1-\widehat{\Xi}_1(\theta)}\\ &\times\int_0^\infty \frac{1-e^{-(p+\lambda_1)x}}{1-e^{-\lambda_1 x}}\frac{1-\widehat{\Xi}_1(\theta)\left(1-e^{-\lambda_1 x}\right)}{1-\widehat{\Xi}_1(p+\theta)\left(1-e^{-(p+\lambda_1)x}\right)}dB(x).\end{aligned} \tag{31}$$

The Laplace transform of the distribution C_2 *is given by:*

$$\begin{aligned}\widehat{C}_2(p)=\int_0^\infty\Bigg\{&\frac{e^{-(p+\lambda_1)x}}{1-\dfrac{\lambda_1}{p+\lambda_1}\widehat{\Xi}_1(p+\theta)\left(1-e^{-(p+\lambda_1)x}\right)}\\ &+\frac{\dfrac{\theta}{p+\theta}\dfrac{\lambda_1}{p+\lambda_1}\left(1-e^{-(p+\lambda_1)x}\right)\left(1-\widehat{\Xi}_1(p+\theta)\right)}{1-\dfrac{\lambda_1}{p+\lambda_1}\widehat{\Xi}_1(p+\theta)\left(1-e^{-(p+\lambda_1)x}\right)}\Bigg\}dB(x).\end{aligned} \tag{32}$$

Proof: The proof follows closely the one of the previous section, while keeping intermediary results conditional on S_1, the service time S_1 that is determined the first time the customer enters service. Because of this similarity, we only give the main computational steps of the proof. Using notations introduced in Section 4, we have:

$$\mathbf{P}\left(A_n\middle|S_1\right)=\left(1-e^{-\lambda_1 S_1}\right)^n\left(\widehat{\Xi}_1(\theta)\right)^n e^{-\lambda_1 S_1},\quad n\geq 0, \tag{33}$$

$$\mathbf{P}\left(B_n\middle|S_1\right)=\left(1-e^{-\lambda_1 S_1}\right)^n\left(\widehat{\Xi}_1(\theta)\right)^{n-1}\left(1-\widehat{\Xi}_1(\theta)\right),\quad n\geq 1, \tag{34}$$

and using Lemma 1 we obtain:

$$\mathbf{E}\left[e^{-p\alpha_1}\middle|A_n,S_1\right]=\mathbf{E}\left[e^{-p\alpha_1}\middle|B_n,S_1\right]=\frac{\lambda_1}{p+\lambda_1}\frac{1-e^{-(p+\lambda_1)S_1}}{1-e^{-\lambda_1 S_1}}, \tag{35}$$

$$\mathbf{E}\left[e^{-p\xi_1}\middle|A_n,S_1\right]=\mathbf{E}\left[e^{-p\xi_1}\middle|B_n,S_1\right]=\frac{\widehat{\Xi}_1(p+\theta)}{\widehat{\Xi}_1(\theta)}, \tag{36}$$

$$\mathbf{E}\left[e^{-p\epsilon_n}\middle|A_n,S_1\right]=e^{-pS_1}, \tag{37}$$

$$\mathbf{E}\left[e^{-p\gamma_n}\middle|B_n,S_1\right]=\frac{\theta}{p+\theta}\frac{1-\widehat{\Xi}_1(p+\theta)}{1-\widehat{\Xi}_1(\theta)}. \tag{38}$$

Therefore,

$$\mathbf{E}\left[e^{-pC_2}\,\middle|\,S_1\right]=\frac{e^{-(p+\lambda_1)S_1}+\dfrac{\theta\lambda_1\left(1-e^{-(p+\lambda_1)S_1}\right)\left(1-\hat{\Xi}_1(p+\theta)\right)}{(p+\theta)\left(p+\lambda_1\right)}}{1-\dfrac{\lambda_1}{p+\lambda_1}\hat{\Xi}_1(p+\theta)\left(1-e^{-(p+\lambda_1)S_1}\right)}, \tag{39}$$

which then gives equation (32). Equations (30) and (31) follow by computations similar to the ones already described in the proof of Proposition 1. □

By differentiation, we obtain the average completion times for the preemptive repeat identical discipline. The average value $\overline{C}_2^S$ of the completion time of a class-2 customer leaving the system after receiving full service is:

$$\begin{aligned}\overline{C}_2^S=&\left(\frac{1}{\lambda_1}\hat{\Xi}_1(\theta)-\hat{\Xi}'_1(\theta)\right)\mathbf{E}\left[\frac{1-e^{-\lambda_1 S_1}}{1-\hat{\Xi}_1(\theta)\left(1-e^{-\lambda_1 S_1}\right)}\right]\\&+\left(1-\hat{\Xi}_1(\theta)\right)\mathbf{E}\left[\frac{S_1}{1-\hat{\Xi}_1(\theta)\left(1-e^{-\lambda_1 S_1}\right)}\right].\end{aligned} \tag{40}$$

The average value $\overline{C}_2^R$ of the completion time of a customer of type k leaving the system out of impatience while its service is being preempted is:

$$\begin{aligned}\overline{C}_2^R=&\frac{1}{\theta}+\frac{1}{\lambda_1}-\frac{\hat{\overline{\Xi}}'_1(\theta)}{\hat{\Xi}_1(\theta)-1}\mathbf{E}\left[\frac{e^{-\lambda_1 S_1}}{1-\hat{\Xi}_1(\theta)\left(1-e^{-\lambda_1 S_1}\right)}\right]\\&-\mathbf{E}\left[\frac{S_1}{\left(e^{\lambda_1 S_1}-1\right)\left(1-\hat{\Xi}_1(\theta)\left(1-e^{-\lambda_1 S_1}\right)\right)}\right].\end{aligned} \tag{41}$$

Note that by letting $\theta\to 0$ in equation (40), we retrieve results known in the case of priority queues without impatience:

$$\lim_{\theta\to 0}\overline{C}_2^S=\left(\frac{1}{\lambda_1}+\overline{\Xi}_1^0\right)\mathbf{E}\left[e^{\lambda_1 S_1}-1\right], \tag{42}$$

which is for example found in Jaiswal [1968, Chapter 4 equation (7.5)] or Takagi [1991, Chapter 3 equation (4.67a)].

6 An application to the M_1,..., M_K/M/1 + M queue

In this section, we consider the case in which service times are exponentially distributed with parameter using the preemptive repeat different discipline. We adapt some of the previous results to this special case. It is important to remark that in the exponential service case, the model can easily be generalised to K-class priority queue with impatience, $K \geq 2$. Indeed, for any integer $k \geq 2$, all customers of class in {1,..., $k-1$} form a (virtual) '$1 \to k-1$'-queue, using the terminology previously introduced, which,

thanks to the memoryless properties of the exponential distribution, is a $M/M/1 + M$ single queue with impatience. This would not be true with a general distribution. We introduce the scaled parameters:

$$\delta = \frac{\mu}{\theta}, \quad \nu_k = \frac{\lambda_k}{\theta}, \quad \nu_{1\to k-1} = \sum_{i=1}^{k-1} \nu_i. \tag{43}$$

The LST of the service times distribution is thus:

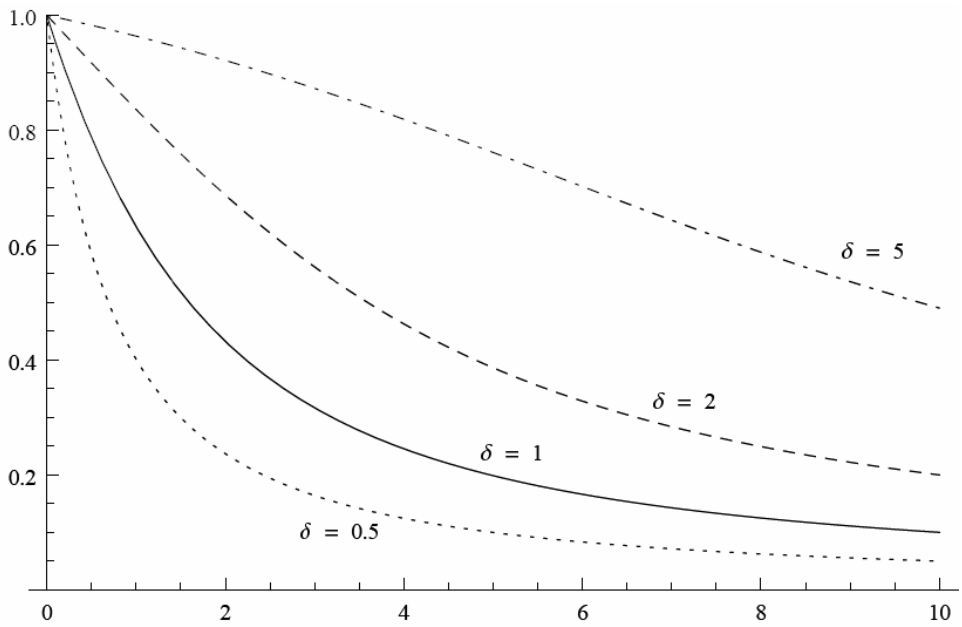
$$\hat{B}(p) = \frac{\delta}{\frac{p}{\theta} + \delta}. \tag{44}$$

Inserting this into equations (2) and (12), we obtain after some simplifications the probability that a class-k customer entering service will receive full service (event A^k):

$$\mathbf{P}\left(A^k\right) = \frac{\delta}{(\delta - 1)\nu_{1\to k-1}} \frac{\Gamma_{\nu_{1\to k-1}}(\delta)}{\Gamma_{\nu_{1\to k-1}}(\delta - 1)}, \tag{45}$$

where Γ denotes the Euler Gamma function and $\Gamma_y(x)$ its incomplete lower version. Figure 1 shows the probability that a class-k customer entering service will receive full service, for different values of δ.

Figure 1 Probability $\mathbf{P}(A^k)$ that a class-k customer entering service in the $M_1,..., M_K/M/1 + M$ queue completes its service, as a function of $\nu_{1\to k-1}$ and for several values of δ ($\theta = 1$)



$\mathbf{P}(A^k)$ obviously increases with δ: the larger δ, the smaller the service time, the smaller the probability his service is preempted, the larger the probability he completes service,

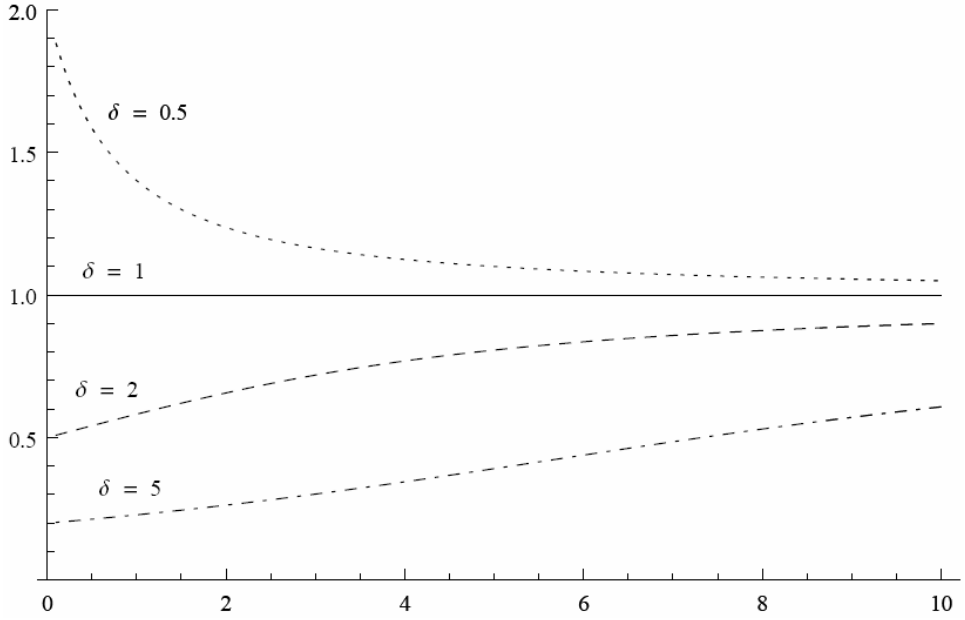
all other things being equal. $\mathbf{P}(A^k)$ obviously decreases with $\nu_{1\to k-1}$: the larger $\nu_{1\to k-1}$, the larger the number of preemptive customers with higher priority, the lower the probability to complete service, all other things being equal. In the special case $\delta = 1$, we have $\mathbf{P}(A^k) = \frac{1-e^{\nu_{1\to k-1}}}{\nu_{1\to k-1}}$ (full line).

Average completion times $\bar{C}_k^S$ and $\bar{C}_k^R$ in this special case do not yield strikingly simple results, although writable in terms of special hypergeometric functions. However, we obtain afters some rearrangements of terms a simple form for the (unconditional) average completion time of a class-k customer entering service:

$$\mathbf{P}(A^k) = \frac{\delta}{(\delta-1)\nu_{1\to k-1}} \frac{\Gamma_{\nu_{1\to k-1}}(\delta)}{\Gamma_{\nu_{1\to k-1}}(\delta-1)}, \tag{46}$$

Figure 2 plots this average time spent until service completion or reneging by class-k customer entering service, as functions of $\nu_{1\to k-1}$.

Figure 2 Average completion time $\bar{C}_k$ for a class-k customer entering service in the $M_1,\ldots, M_K/M/1 + M$ queue as a function of $\nu_{1\to k-1}$ and for several values of illustrating regimes of service times $\delta(\theta = 1)$



Obviously, $\bar{C}_k$ decreases with δ. If $\delta = 1$, then service and patience times have the same average value, which as a consequence is also the average completion time: $\bar{C}_2 = 1/\theta$ (full line). If $\delta < 1$, then the average service time is larger than the average patience time of the customers. Therefore, $\bar{C}_k$ is actually larger for customers with smaller $\nu_{1\to k-1}$, i.e., higher-priority customers (dotted line). Indeed, since customers do not abandon while being served, higher-priority customers, which have a larger probability to receive full

service, also have a larger unconditional completion time. If $\delta > 1$, then the situation is reversed: the average service time is smaller than the average patience time of the customers, and therefore lower-priority customers have a larger average completion time that higher-priority ones. In all case, $\overline{C}_k$ tends to $1/\theta$ as $\nu_{1\to k-1} \to +\infty$, since the probability $\mathbf{P}(A^k)$ of service completion goes to 0 as $\nu_{1\to k-1} \to +\infty$.

7 Conclusions

In this note, we provide explicit analytical formulas for the Laplace-Stieltjes transforms of the completion times of customers entering a two-class priority queue with preemptive discipline and exponential impatience. An interesting feature of the system considered is the possibility for a preempted customer to leave out of impatience during preemption. To our knowledge, this feature has not been previously considered in the exact analysis of priority queues with impatience. These results may have consequences in two directions. Firstly, analysis of completion times greatly helped the analysis of queueing systems without impatience. In a similar way, our results will hopefully prove helpful in further exact analysis of the stationary state of priority queues with impatience. Such results are still rare and analysis of these systems often resorts to simulation or asymptotic approximations for now. Secondly, the results on completion time may have consequences in solving strategy problems in queueing systems from a customer point of view. Using our results, a low priority customer knows the distribution of the time between the moment its (same priority) predecessor starts service and the moment he will actually be the first-in-line within his priority queue. Future work will show how such knowledge may be useful in strategic queueing problems.

References

Aksin, Z., Armony, M. and Mehrotra, V. (2007) 'The modern call center: a multi-disciplinary perspective on operations management research', *Production and Operations Management*, Vol. 16, No. 6, pp.665–688.

Ata, B. and Tongarlak, M.H. (2012) 'On scheduling a multiclass queue with abandonments under general delay costs', *Queueing Systems*, forthcoming.

Baccelli, F., Boyer, P. and Hebuterne, G. (1984) 'Single-server queues with impatient customers', *Advances in Applied Probability*, Vol. 16, No. 4, pp.887–905.

Bae, J., Kim, S. and Lee, E.Y. (2001) 'The virtual waiting time of the *M*/*G*/1 queue with impatient customers', *Queueing Systems*, Vol. 38, No. 4, pp.485–494.

Bartolini, C., Stefanelli, C. and Tortonesi, M. (2012) 'Modeling IT support organizations using multiple-priority queues', in *Proceedings of the 2012 IEEE Network Operations and Management Symposium (NOMS)*, pp.377–384.

Boxma, O., Perry, D. and Stadje, W. (2011) 'The M/G/1+G queue revisited', *Queueing Systems*, Vol. 67, No. 3, pp.207–220.

Boxma, O., Perry, D., Stadje, W. and Zacks, S. (2010) 'The busy period of an *M*/*G*/1 queue with customer impatience', *Journal of Applied Probability*, Vol. 47, No. 1, pp.130–145.

Brandt, A. and Brandt, M. (2004) 'On the two-class *M*/*M*/1 system under preemptive resume and impatience of the prioritized customers', *Queueing Systems*, Vol. 47, Nos. 1–2, pp.147–168.

Brandt, A. and Brandt, M. (2011) *Workload and Busy Period for M/GI/1 with a General Impatience Mechanism*, Technical report, Konrad-Zuse-Zentrum fur Informationstechnik Berlin.

Chang, W. (1965) 'Preemptive priority queues', *Operations Research*, Vol. 13, No. 5, pp.820–827.

Choi, B.D., Kim, B. and Chung, J. (2001) '*M*/*M*/1 queue with impatient customers of higher priority', *Queueing Systems*, Vol. 38, No. 1, pp.49–66.

Choudhury, A. and Medhi, P. (2011) 'Balking and reneging in multiserver Markovian queuing system', *International Journal of Mathematics in Operational Research*, Vol. 3, No. 4, pp.377–394.

Daley, D.J. (1965) 'General customer impatience in the queue *GI*/*G*/1', *Journal of Applied Probability*, Vol. 2, No. 1, pp.186–205.

Gaver, D.P. (1962) 'A waiting line with interrupted service, including priorities', *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 24, No. 1, pp.73–90.

Iravani, F. and Balcioglu, B. (2008) 'On priority queues with impatient customers', *Queueing Systems*, Vol. 58, No. 4, pp.239–260.

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

Jouini, O., Pot, A., Koole, G. and Dallery, Y. (2010) 'Online scheduling policies for multiclass call centers with impatient customers', *European Journal of Operational Research*, Vol. 207, No. 1, pp.258–268.

Kim, K. (2012) 'T-preemptive priority queue and its application to the analysis of an opportunistic spectrum access in cognitive radio networks', *Computers & Operations Research*, Vol. 39, No. 7, pp.1394–1401.

Kovalenko, I.N. (1961) 'Some queuing problems with restrictions', *Theory of Probability & its Applications*, Vol. 6, No. 2, pp.204–208.

Perry, D., Stadje, W. and Zacks, S. (2000) 'Busy period analysis for *M*/*G*/1 and *G*/*M*/1 type queues with restricted accessibility', *Operations Research Letters*, Vol. 27, No. 4, pp.163–174.

Rao, S.S. (1967) 'Queuing with balking and reneging in *M*|*G*|1 systems', *Metrika*, Vol. 12, No. 1, pp.173–188.

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

Wang, Q. (2004) 'Modeling and analysis of high risk patient queues', *European Journal of Operational Research*, Vol. 155, No. 2, pp.502–515.

Ward, A.R. (2012) 'Asymptotic analysis of queueing systems with reneging: a survey of results for FIFO, single class models', *Surveys in Operations Research and Management Science*, Vol. 16, No. 1, pp.1–14.

Xiong, W., Jagerman, D. and Altiok, T. (2008) '*M*/*G*/1 queue with deterministic reneging times', *Performance Evaluation*, Vol. 65, Nos. 3–4, pp.308–316.

Zeltyn, S., Feldman, Z. and Wasserkrug, S. (2009) 'Waiting and sojourn times in a multi-server queue with mixed priorities', *Queueing Systems*, Vol. 61, No. 4, pp.305–328.

Zhao, N. and Lian, Z. (2011) 'A queueing-inventory system with two classes of customers', *International Journal of Production Economics*, Vol. 129, No. 1, pp.225–231.