Stochastic Models, 21:745–765, 2005
Copyright © Taylor & Francis, Inc.
ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1081/STM-200056212

RESPONSE TIME DISTRIBUTION IN A D-MAP/PH/1 QUEUE WITH GENERAL CUSTOMER IMPATIENCE

J. Van Velthoven, B. Van Houdt, and C. Blondia □ *University of Antwerp, Antwerpen, Belgium*

□ *This paper presents two methods to calculate the response time distribution of impatient customers in a discrete-time queue with Markovian arrivals and phase-type services, in which the customers' patience is generally distributed (i.e., the D-MAP/PH/1 queue). The first approach uses a GI/M/1 type Markov chain and may be regarded as a generalization of the procedure presented in Van Houdt*[14] *for the D-MAP/PH/1 queue, where every customer has the same amount of patience. The key construction in order to obtain the response time distribution is to set up a Markov chain based on the age of the customer being served, together with the state of the D-MAP process immediately after the arrival of this customer. As a by-product, we can also easily obtain the queue length distribution from the steady state of this Markov chain.*

We consider three different situations: (i) customers leave the system due to impatience regardless of whether they are being served or not, possibly wasting some service capacity, (ii) a customer is only allowed to enter the server if he is able to complete his service before reaching his critical age and (iii) customers become patient as soon as they are allowed to enter the server. In the second part of the paper, we reduce the GI/M/1 type Markov chain to a Quasi-Birth-Death (QBD) process. As a result, the time needed, in general, to calculate the response time distribution is reduced significantly, while only a relatively small amount of additional memory is needed in comparison with the GI/M/1 approach. We also include some numerical examples in which we apply the procedures being discussed.

Keywords Age process; D-MAP/PH/1 queue; Impatient customers; Matrix analytic method; QBD process.

Mathematics Subject Classification Primary 60K25; Secondary 60M20, 90B22.

Received September 2004; Accepted February 2005

Address correspondence to B. Van Houdt, University of Antwerp, Middelheimlaan 1, Antwerpen B-2020, Belgium; E-mail: benny.vanhoudt@ua.ac.be

1. INTRODUCTION

In this paper, we discuss two algorithms to calculate the response time distribution in a D-MAP/PH/1 queue with general customer impatience (see section 2 for definitions). Let Z be the patience distribution. We allow Z to be generally distributed, except that we assume that there exists some $r \geq 0$ sufficiently large, such that $P[Z > r] = 0$, i.e., the maximum amount of patience that a customer can have is bounded above by some constant r. We consider three different systems: (i) a customer is impatient during his sojourn time (waiting + service) and may thus be partially served, (ii) customers are aware of their service time and only enter the service facility if their amount of patience is sufficient to complete service, (iii) a customer can only run out of patience while waiting, i.e., he becomes patient as soon as he enters the service facility. We shall denote each of these three D-MAP/PH/1 queues as D-MAP/PH/1 + GI, to reflect the general nature of the patience distribution. It should be noted that in the literature this notation is mostly used for systems where the customers have a limited waiting time (case (iii)).

The first approach presented in this paper uses a GI/M/1-type Markov chain and is similar to the procedure introduced in Van Houdt[14], where a system is being discussed in which every customer has the same amount of patience. In order to obtain the response time distribution, we set up a Markov chain by keeping track of the age of the customer being served, while remembering the state of the D-MAP process immediately after the arrival of this customer. The disadvantage of this algorithm is that the time needed to calculate the response time distribution is a square function of the maximum amount of patience r that a customer can have. A partial solution to this problem is given by the introduction of the second algorithm which is based on a QBD reduction. As a result, the time complexity becomes a linear function of the maximum amount of patience r. On the other hand, an implementation of the QBD based algorithm requires somewhat more memory in comparison with the GI/M/1 approach.

Queueing systems with impatient customers have many applications in telecommunications, for instance in telephone systems, where people have to wait for a dial tone (Zhao[15]) or call centers (Garnet[8]), where the customers are only willing to wait a certain amount of time before they can be served. Other examples include systems with real time constraints, where packets have to arrive before a given deadline in order to be useful, data communication networks with a time-out protocol, etc. Other applications of queues with impatient customers can be found in manufacturing and service industries, e.g., inventory systems with perishable goods.

The study of single-server queues with impatient customers has a long history. It seems Palm[12] was the first to consider customer impatience.

Barrar[2] analyzed the M/M/1 + D system. A key reference for the general GI/GI/1 + GI is Baccelli et al.[1]. In this work a stability condition was established for the general case, while for the M/GI/1 + GI queue the virtual waiting time was studied. Markovian arrivals were considered by Combé[7], who studied the MAP/G/1 + M queue and derived an expression for the transform of the virtual waiting time and rejection probability of a customer. Most of these studies assume that a customer becomes patient when entering the server and set up a Markov process using the virtual (offered) waiting time. Van Houdt et al.[14] developed an algorithm to compute the response time distribution in a D-MAP/PH/1 + D queue, by setting up a finite GI/M/1 type Markov chain. In this paper, we generalize the work in Van Houdt[14] in a number of ways. First, we allow the impatience distribution to be general, as opposed to deterministic. Second, we also consider the system where customers only enter the server if they are able to complete their service before becoming impatient. Third, a QBD reduction procedure is included to improve the time complexity of the algorithms and finally, we also show that apart from the response time distribution and the rejection probability, the queue length distribution can easily be obtained as a by-product.

The next section gives a description of the queueing system under consideration, whereas in sections 3, 4, and 5 we discuss the first approach, using a GI/M/1 type Markov chain. In section 3, customers leave the system due to impatience regardless of whether they are being served or not, possibly wasting some service capacity. In section 4 such capacity losses are avoided by only allowing a customer to enter the server if his service will be completed before he reaches his critical age. A system where the customers are only impatient while waiting, is discussed in section 5. Section 6 introduces an algorithm to calculate the response time distribution using a QBD and finally, in section 7, we apply both algorithms to some numerical examples, which provide, among others, insight on the influence of the patience distribution Z.

2. THE DISCRETE-TIME D-MAP/PH/1 + GI QUEUE

The arrival process of the queueing system of interest is the D-MAP, a discrete time Markovian arrival process (Blondia[3,4]). This process allows us to work with correlated arrivals and is a special case of the D-BMAP arrival process which allows batch arrivals. It is characterized by two $m \times m$ matrices D_0 and D_1, with m a positive integer. These matrices contain the transition probabilities of the underlying Markov chain when either a customer arrives (covered by D_1) or not (D_0). For example, the entry (j_1, j_2) of D_1 represents the probability that there is an arrival and the underlying

Markov chain makes a transition from state j_1 to state j_2. The matrix D_0 covers the case in which there is no arrival.

The matrix D represents the stochastic $m \times m$ transition matrix of the underlying Markov chain and is defined by $D = D_0 + D_1$. In this paper, we assume that D is irreducible. Denote θ as the solution to $\theta D = \theta$ and $\theta e = 1$, where e is a vector with all its entries equal to one of the appropriate dimension. That is, θ is the stationary probability vector of D. Then, the stationary arrival rate is given by $\lambda = \theta D_1 e$. The arriving customers are being served in a first-come first-served (FCFS) order, except for those leaving the waiting room when they reach their critical age.

The service time of a customer has a common discrete-time phase-type (PH) distribution with a matrix representation (m_{ser}, α, T), where m_{ser} is a positive integer. The vector α contains the probabilities that the service of a customer starts in a given phase. The probability that a customer continues his service in phase j at time $n+1$, given the phase at time n equals i, is represented by the (i, j)th entry of the $m_{ser} \times m_{ser}$ substochastic matrix T. Finally, define $t = e - Te$, that is, t is a vector which contains the probabilities that a customer completes his service, given the current phase of the service process. Notice, the service time of a customer equals k time units with probability $\alpha T^{k-1} t$ and the mean service time is given by $\frac{1}{\mu} = \alpha(I - T)^{-1} e$. Let $m_{tot} = m_{ser} m$. It is well known that PH distributions are well suited for representing most of the types of services encountered in communication systems (Lang[10]).

Each customer has a finite amount of patience that is i.i.d. and arbitrarily distributed according to some random variable Z. The patience distribution Z is characterized by the stochastic vector $\tilde{a} = (a_1, a_2, \ldots, a_r)$, where a_i is the probability $P[Z = i]$ that the amount of patience of a customer equals i time units, while r is the maximum amount of patience a customer can have. Let p_i be the probability $P[Z \leq i]$ that the patience of a customer is at most i time units, hence, $p_i = \sum_{j=1}^{i} a_j$. In order to simplify the notation, we define $p_0 = 0$, which we may regard as the probability that an impatient customer has a critical age of zero time units. Because such customers are never served, we do not consider them in this paper.[a] The amount of patience of a customer is also referred to as his critical age.

Finally, if there is a departure and an arrival at the same time, we assume that the departure occurs first. Also, when a customer arrives while the server is idle, his service starts immediately. In each of the models presented, we will consider the system just prior to possible arrivals, departures or phase changes. Thus, if we refer to the system state at time n, such events happening at time n are not yet taken into account. For instance, when a customer starts his service at time n, the probability that

[a] Such customers can easily be taken into account by modifying the characteristics of the arrival process.

the phase of the service process equals i at time $n+1$ is given by the ith component of the stochastic vector α. Also, the minimum age of a customer in the service facility is one time unit.

3. IMPATIENT CUSTOMERS IN THE SYSTEM

In this section we consider the D-MAP/PH/1+GI queue with customers who are impatient in the system, that is, all customers are impatient irrespective of whether they are being served or not. A customer reaching his critical age will leave the queue without starting or completing his service. This system is also referred to as the "limitation on sojourn time with unaware customers" (Baccelli[11]), as customers are impatient during their entire sojourn time and are not aware, when entering the queue, whether their total sojourn time will be larger than their patience.

Consider the following Markov chain (MC) with $rm_{tot} + m$ states. Denote level zero of the MC as the set of states $\{1, \ldots, m\}$ and level i as the set of states $\{(i-1)m_{tot} + m + 1, \ldots, im_{tot} + m\}$, for $0 < i \leq r$. The states of level $i > 0$ are labeled as (s, j), with $1 \leq s \leq m_{ser}$ and $1 \leq j \leq m$. Let n be the current time instant and let state (s, j) of level i (with $0 < i \leq r$) correspond to the situation in which the age of the customer in service equals i, the service process is currently in phase s and the D-MAP was in state j at time $n - i + 1$. We say that the age of a customer equals i when he arrived in the system i time units ago. Also, let state j of level zero correspond to the situation in which the server is idle and the current state of the arrival process is j. Then, the system can be described by a transition matrix P with the following structure:

$$P = \begin{bmatrix} B_1 & B_0 & 0 & 0 & \ldots & 0 & 0 \\ B_2 & A_1^1 & A_0^1 & 0 & \ldots & 0 & 0 \\ B_3 & A_2^2 & A_1^2 & A_0^2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ B_r & A_{r-1}^{r-1} & A_{r-2}^{r-1} & A_{r-3}^{r-1} & \ldots & A_1^{r-1} & A_0^{r-1} \\ E & C_r & C_{r-1} & C_{r-2} & \ldots & C_2 & C_1 \end{bmatrix}.$$

Here, B_1 is an $m \times m$ matrix, B_0 is an $m \times m_{tot}$ matrix, $B_i (i > 1)$ and E are $m_{tot} \times m$ matrices and A_k^i and C_i are $m_{tot} \times m_{tot}$ matrices. We will now derive an expression for each of these matrices.

Let us start with the first level of P, level zero. If the MC is in a state of this level, the server is idle and only two events can take place: either a customer arrives or not. If there is no arrival, the MC remains at level zero and makes a transition to some state of this level, according to the transition of the underlying Markov chain of the arrival process.

Hence, $B_1 = D_0$. If a customer does arrive at the current time instant n, a transition from level zero to level one occurs, this new arrival immediately enters the service facility and starts his service in a phase determined by the vector α, so $B_0 = \alpha \otimes D_1$.

If the MC is in state (s_1, j_1) of level i at time n, there is a transition to level zero if the customer in service leaves the system and there is no customer present in the waiting room at time $n+1$. There are two possible causes for the customer in service to leave the service facility: either he completes his service or he reaches his critical age. Therefore, the customer in service leaves the system with a probability $t_{s_1} + \frac{a_i}{1-p_{i-1}}(1 - t_{s_1})$.

Let us now derive the probability that the system will be empty at time $n+1$, provided that an age i customer leaves the system at time n. Unlike the case in which all of the customers have the same amount of patience (Van Houdt[14]), it is possible that a customer arrival occurred during some of the time instants $n-i+1, n-i+2, \ldots, n-1$, who left before or at time n due to impatience. Notice, a customer arriving at time n is always present in the system at time $n+1$ as the minimum amount of patience is one time unit. If the critical age of a customer arriving at time $n-i+k$, for $k = 1, \ldots, i-1$, is at most $i-k$, which is the case with probability p_{i-k}, this customer is no longer in the queue at time $n+1$. Obviously, there could be multiple customers of this kind. When we combine these probabilities, we find

$$B_{i+1} = \left(t + \frac{a_i}{1-p_{i-1}}(e-t)\right) \otimes \prod_{k=1}^{i}(D_0 + p_{i-k}D_1), \tag{1}$$

for $0 < i < r$ (recall, $p_0 = 0$). A transition from level i to level $i+1$ occurs when the customer in service has not reached his critical age nor completed his service. In this case the customers stays in the system and the state of the D-MAP remains the same. Hence,

$$A_0^i = \frac{1-p_i}{1-p_{i-1}} T \otimes I_m, \tag{2}$$

where I_m denotes the $m \times m$ unity matrix. The only remaining transitions from level i are those to some lower level, $i-l$ $(0 < l < i)$. They occur when the customer in service leaves the system and the first customer in the waiting room arrived at time $n-i+l+1$. Notice, this customer is still patient, otherwise he would have left the system before time $n+1$. This yields,

$$A_{l+1}^i = \left(t + \frac{a_i}{1-p_{i-1}}(e-t)\right)\alpha \otimes \left(\prod_{k=1}^{l}(D_0 + p_{i-k}D_1)\right)(1 - p_{i-l-1})D_1. \tag{3}$$

Notice that there could have been arrivals on some of the time instants $n - i + k$, with $0 < k \leq l$, who reached there critical age before time $n + 1$.[b]

Finally, consider the situation where the age of the customer in service equals r. This customer will leave the system, whether he finishes his service or not. If there are no customers in the waiting room at time $n + 1$, the MC will make a transition to level zero. In view of Eqn. (1) it follows

$$E = e \otimes \prod_{k=1}^{r} (D_0 + p_{r-k} D_1). \tag{4}$$

Otherwise the first customer in the waiting room, who has clearly not yet reached his critical age, will enter the server. Implying, for $0 \leq i < r$,

$$C_{i+1} = e\alpha \otimes \left(\prod_{k=1}^{i} (D_0 + p_{r-k} D_1) \right) (1 - p_{r-i-1}) D_1. \tag{5}$$

The transition matrix P is a finite level dependent GI/M/1 type Markov chain, the steady state vector of which can be computed efficiently by the Latouche-Jacobs-Gaver (LJG) algorithm (Latouche[11]). Because this algorithm is less efficient as the one we will discuss in section 6, we do not include a detailed description of the LJG algorithm.

Having calculated the steady state vector $\pi = (\pi_0, \pi_1, \ldots)$ of the transition matrix P, we want to determine the probability $P[X = i]$ that a customer receives a complete service and that his response time equals i time units. This probability is given by the expected number of customers who complete their service i time units after they entered the queue, divided by the expected number of customers who leave the system at an arbitrary time instant. Denote the probability vector π_i that the MC is at level i at an arbitrary time as $\pi_i = ((\pi_i)_{(1,1)}, (\pi_i)_{(1,2)}, \ldots, (\pi_i)_{(1,m)}, (\pi_i)_{(2,1)}, \ldots, (\pi_i)_{(m_{ser},m)})$. Thus, $\sum_{j=1}^{m} (\pi_i)_{(s,j)}$ is the probability that at an arbitrary time instant an age i customer is in the service facility, with the phase of the service process equaling s. Also, $(t)_s$ is the probability that such a customer completes his service. Hence, we obtain $P[X = i] = \frac{1}{\lambda} \sum_{s=1}^{m_{ser}} (t)_s \sum_{j=1}^{m} (\pi_i)_{(s,j)}$, for $0 < i \leq r$.

The rejection probability P_{out}, being the probability that a customer leaves the system without entering or completing his service, reads $P_{out} = 1 - \sum_{i=1}^{r} P[X = i]$. Although we focus mainly on the response time distribution, we can also obtain the queue length distribution as a by-product. To do so, we define the $m \times 1$ vector $h_{k,d}$ as the probability that in an interval of length d, k customers arrive, who are still in the queue at the end of this interval. The j^{th} entry of this vector $h_{k,d}(j)$ represents the

[b] When the next customer arrival occurs at time $n - i + 1$, the MC remains at the same level. In this case we define the subexpression $\prod_{k=1}^{0}(\ldots)$ equal to one.

case in which the state of the D-MAP process equals j at the start of the interval. The following relation can be used to calculate these probability vectors:

$$h_{k,0} = 1_{\{k=0\}}e,$$

$$h_{0,d} = \left(\prod_{i=1}^{d}(D_0 + p_{d-i}D_1)\right)e,$$

$$h_{k,d} = (1 - p_{d-1})D_1 h_{k-1,d-1} + (D_0 + p_{d-1}D_1)h_{k,d-1}.$$

Denote $P[Q = q]$ as the probability that there are q customers in the waiting room at an arbitrary time instant. This situation is only possible when a customer with an age of at least $q + 1$ time units is in the service facility. Therefore, we have

$$P[Q = q] = \sum_{i=q+1}^{r}\sum_{j=1}^{m} h_{q,i-1}(j)\sum_{s=1}^{m_{ser}}(\pi_i)_{(s,j)}, \quad \text{for } 0 < q < r, \text{ and} \tag{6}$$

$$P[Q = 0] = \sum_{j=1}^{m}(\pi_0)_j + \sum_{i=1}^{r}\sum_{j=1}^{m} h_{0,i-1}(j)\sum_{s=1}^{m_{ser}}(\pi_i)_{(s,j)}. \tag{7}$$

A similar method can be used to obtain the queue length distribution of the systems discussed in sections 4 and 5.

Remark. In the special case that the service time distribution is geometric with parameter μ, we have $(1 - P_{out}) = (\mu/\lambda)(1 - P_{idle})$, where P_{idle} is the probability that the server is idle. We can rewrite this as $P_{busy} = \lambda(1 - P_{out})/\mu$. This formula is an unexpected result, because one may, incorrectly, get the impression that the server is only busy serving customers who are successful, that is, manage to complete their service before running out of patience. This is however not the case, as some customers get partially served. The reason that we still get this expression can be explained by noticing that the mean service time $1/\mu_{suc}$ of a customer who is successful is less than $1/\mu$. Actually, if we denote Z_{rem} as the stationary distribution of the remaining patience when a customer enters the server, one can easily prove that the mean time $1/\mu_{in}$ that a customer spends in the server (irrespective of whether he is successful) equals

$$1/\mu_{in} = 1/\mu\left(1 - \sum_{j=1}^{\infty}P[Z_{rem} = j](1 - \mu)^j\right) = 1/\mu P_{suc|in}, \tag{8}$$

where $P_{suc|in}$ is the probability that a customer is successful provided that he entered the server. Clearly, $P_{busy} = \lambda P_{in}/\mu_{in}$ where P_{in} is the probability that a customer enters the server. Thus, we get $P_{busy} = \lambda(1 - P_{out})/\mu$, as required. This equality is not valid for all phase-type distributions, for instance if the service is deterministic then $1/\mu_{suc} = 1/\mu$ and $P_{busy} \geq \lambda(1 - P_{out})/\mu$.

4. SERVICE TIME AWARE IMPATIENT CUSTOMERS

In the system studied in the previous section a customer who does not complete his service before reaching his critical age, is still allowed to enter the server. In this way, such a customer wastes some service capacity which could have been used by other customers. In this section we consider a system where customers are only allowed to enter the server, if they manage to compete service before reaching their critical age. As such, we might refer to this model as the "limitation on sojourn time with service time aware customers," as customers are still impatient during the entire sojourn time, but are aware of their service time. Therefore, they can decide to leave the system, when the service facility becomes available to them, if they notice that their remaining amount of patience does not suffice to complete service. As long as we are concerned with the rejection probability or the response time distribution (of successful customers), this system is equivalent to the "limitation on sojourn time with aware customers" (Baccelli[11]). In such a system, customers are assumed to be aware of their required sojourn time upon arrival to the queue and immediately leave if their amount of patience does not suffice. The queue length distribution is, however, somewhat different as aware customers will often not enter the waiting line at all, whereas in our case some unsuccessful customers still spend some time waiting in the queue.

We can construct a similar MC as in the previous section and denote its $m + rm_{tot} \times m + rm_{tot}$ transition matrix as

$$\widehat{P} = \begin{bmatrix} \widehat{B}_1 & \widehat{B}_0 & 0 & 0 & \dots & 0 & 0 \\ \widehat{B}_2 & \widehat{A}_1^1 & \widehat{A}_0^1 & 0 & \dots & 0 & 0 \\ \widehat{B}_3 & \widehat{A}_2^2 & \widehat{A}_1^2 & \widehat{A}_0^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \widehat{B}_r & \widehat{A}_{r-1}^{r-1} & \widehat{A}_{r-2}^{r-1} & \widehat{A}_{r-3}^{r-1} & \dots & \widehat{A}_1^{r-1} & \widehat{A}_0^{r-1} \\ \widehat{E} & \widehat{C}_r & \widehat{C}_{r-1} & \widehat{C}_{r-2} & \dots & \widehat{C}_2 & \widehat{C}_1 \end{bmatrix}.$$

Let n be the current time instant and assume some customer c leaves the server at time n. Then, the first customer that arrived after customer c at some time $n - k$ $(0 \leq k < r)$ does not enter the server at time n if he has already reached his critical age or if his remaining patience is not sufficient to be served entirely. This probability is given by $q_k = p_k + \sum_{i=k+1}^{r} a_i \alpha T^{i-k} e$.

Assume a customer c of age k is in phase s of the service process at time n. Denote the probability that the sum of his age and his remaining service time does not exceed his critical age as the sth component of the $m_{ser} \times 1$ vector v^k. Then, $(v^k)_s = \sum_{i=k}^{r} a_i \sum_{j=0}^{i-k} (T^j t)_s$, which equals $1 - p_{k-1} - \sum_{i=k}^{r} a_i (T^{i-k+1} e)_s$, using the relation $t = e - Te$. Hence,

$$v^k = e - p_{k-1} e - \sum_{i=k}^{r} a_i T^{i-k+1} e.$$

Let u^k, a column vector of dimension m_{ser}, represent the probability that a customer of age k completes his service. Since this customer was admitted by the server, we know that his critical age is at least as much as the sum of his age (k) and his remaining service time. Thus, for $1 \leq k \leq r$ and for $1 \leq s \leq m_{ser}$ we have,

$$(u^k)_s = \frac{a_k(t)_s + a_{k+1}(t)_s + \cdots + a_r(t)_s}{(v^k)_s} = \frac{(1 - p_{k-1})(t)_s}{(v^k)_s}.$$

Consider an age k customer in phase i of the service process at time n. The probability that this customer continues his service in phase j is represented by the (i,j)th entry of the $m_{ser} \times m_{ser}$ matrix U^k. Notice, since the customer was admitted in the server, he only leaves the system when his service finishes. Hence, this probability is given by

$$(U^k)_{ij} = \frac{\sum_{h=k+1}^{r} a_h (T)_{ij} \sum_{l=1}^{h-k} (T^{l-1} t)_j}{(v^k)_i} = \frac{(T)_{ij} (v^{k+1})_j}{(v^k)_i},$$

for $1 \leq k \leq r$ and $1 \leq i, j \leq m_{ser}$.

We are now in a position to set up an equation for each of the transition blocks appearing in the matrix $\widehat{P}$. For the matrices on level zero, the only difference with the system discussed in section 3 is that an arriving customer who finds the server idle, does not necessarily enter the service facility. If he has not enough patience to stay in the system until his service has completed, he abandons the system. This leads to $\widehat{B}_1 = D_0 + q_0 D_1$ and $\widehat{B}_0 = (1 - q_0) D_1$.

The MC makes a transition from level i to level zero if the customer in service[c] completes his service and there is no customer who enters the server. For $0 < i < r$, the expression of the $m_{tot} \times m$ matrix $\widehat{B}_{i+1}$ is given by the following equation:

$$\widehat{B}_{i+1} = u^i \otimes \prod_{k=1}^{i} (D_0 + q_{i-k} D_1). \tag{9}$$

[c] Due to the definition of the levels of the MC, the age of this customer equals i time units.

Remember, the continuation and completion of the service of a customer with an age equal to k time units, is given by the matrix U^k and the vector u^k, respectively. Arguments similar to those presented in section 3 yield the following expressions:

$$\widehat{A}_0^i = U^i \otimes I_m, \tag{10}$$

$$\widehat{A}_{l+1}^i = u^i \alpha \otimes \left(\prod_{k=1}^{l} (D_0 + q_{i-k} D_1) \right) (1 - q_{i-l-1}) D_1, \tag{11}$$

$$\widehat{E} = e \otimes \prod_{k=1}^{r} (D_0 + q_{r-k} D_1), \text{ and} \tag{12}$$

$$\widehat{C}_{i+1} = e\alpha \otimes \prod_{k=1}^{i} (D_0 + q_{r-k} D_1)(1 - q_{r-i-1}) D_1. \tag{13}$$

As the transition matrices P and $\widehat{P}$ have the same form, we can make use of the LJG algorithm to obtain the steady state vector $\hat{\pi}$ of $\widehat{P}$. An expression for the probability that a customer receives a complete service and has a response time of i time units is then established as $P[\widehat{X} = i] = \frac{1}{\lambda} \sum_{s=1}^{m_{ser}} (u^i)_s \sum_{j=1}^{m} (\hat{\pi}_i)_{(s,j)}$, for $0 < i \leq r$. The rejection probability $\widehat{P}_{out}$, being the probability that a customer abandons the system without entering the service facility, is given by $\widehat{P}_{out} = 1 - \sum_{i=1}^{r} P[\widehat{X} = i]$.

5. IMPATIENT CUSTOMERS IN THE WAITING ROOM

In sections 3 and 4, we considered customers who remain impatient even when they have already entered the service facility. In this section we assume that a customer is no longer impatient during his sojourn time, but only while waiting. Once a customer enters the server, he remains there until his service is completed. When focusing on the rejection probabilities or on the response time distribution (of successful customers), this system coincides with both the "limitation on waiting time with aware or unaware customers" (Baccelli[1]) as in neither system abandoning customers use any service capacity. We assume that a customer who reaches his critical age at the exact moment that the server becomes available to him, leaves the system.

If we define a MC analogue to the previous sections, the corresponding transition matrix P' is either finite or infinite, depending on the service time distribution. For unbounded service times, we get an infinite matrix P' as there is no limit on the age of the customer in service. In order to find the steady state vector of P', we introduce a new MC with an $m + rm_{tot} \times m + rm_{tot}$ transition matrix $\overline{P}$, on which we can apply the LJG algorithm to find its steady state vector. The transition matrix $\overline{P}$ corresponds to the MC that we get when censoring the MC P' on the set of states for which the age of the customer in service is at most r:

$$\overline{P} = \begin{bmatrix} B_1 & B_0 & 0 & 0 & \dots & 0 & 0 \\ \overline{B}_2 & \overline{A}_1^1 & \overline{A}_0^1 & 0 & \dots & 0 & 0 \\ \overline{B}_3 & \overline{A}_2^2 & \overline{A}_1^2 & \overline{A}_0^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \overline{B}_r & \overline{A}_{r-1}^{r-1} & \overline{A}_{r-2}^{r-1} & \overline{A}_{r-3}^{r-1} & \dots & \overline{A}_1^{r-1} & \overline{A}_0^{r-1} \\ \overline{E} & \overline{C}_r & \overline{C}_{r-1} & \overline{C}_{r-2} & \dots & \overline{C}_2 & \overline{C}_1 \end{bmatrix}.$$

Notice, the matrices corresponding to level zero are identical to those in section 3, as the distinction between this model and the one in section 3 lies only in the customer behavior inside the service facility. The expressions for the matrices, representing a transition from level i $(0 < i < r)$ are altered, because the probability that a customer leaves (remains in) the server no longer depends upon the age of this customer. However, apart from the vector and the matrix holding the probabilities that a service completion and continuation occurs, respectively, the system evolves identical to the one in section 3:

$$\overline{B}_{i+1} = t \otimes \prod_{k=1}^{i} (D_0 + p_{i-k} D_1), \tag{14}$$

$$\overline{A}_0^i = T \otimes I_m, \text{ and} \tag{15}$$

$$\overline{A}_{l+1}^i = t\alpha \otimes \left(\prod_{k=1}^{l} (D_0 + p_{i-k} D_1) \right) (1 - p_{i-l-1}) D_1, \tag{16}$$

for $1 \le i < r$ and $0 \le l < i$. Next, let us derive an expression for the matrices at level r. The matrix $\overline{E}$ contains the probabilities that the MC makes a transition from state (s_1, j_1) at level r to state j_2 at level zero. Consider a customer c of age r who is being served at time n. Customers present in the service facility are patient and remain in the system until their service is completed. Thus, with probability $(T^k t)_{s_1}$, customer c will remain in the server for another k time units, meaning the next time instant observed by the censored chain $\overline{P}$ is time $n + k + 1$. When customer c leaves the server at time $n + k$, the next customer if present can start his service. Because no customer has a critical age of more than r time units, customers who arrived at some time $n - r + i$, with $0 < i \le k$, will leave the waiting room before or at time $n + k$. In order to have a transition to level zero, some arrivals may occur at one of the time instants $n-r+k+1,\ldots,n+k-1$, as long as these customers reach their critical age before or at time $n+k$.[d] This probability is given by $\prod_{h=1}^{r-1}(D_0+p_{r-h}D_1)$ and therefore the probability of having an empty waiting room at time $n+k+1$ equals $D^k\prod_{h=1}^{r}(D_0+p_{r-h}D_1)$. This leads to

$$\overline{E}=\left(\sum_{k=0}^{\infty}(T^k t)\otimes D^k\right)\prod_{h=1}^{r}(D_0+p_{r-h}D_1). \tag{17}$$

The MC makes a transition to level $r-i$ from level r if there are no arrivals at time $n-r+k+1,\ldots,n-r+k+i$ of customers who are still in the waiting room at time $n+k$ and a customer c' with a critical age larger than $r-i-1$ arrives at time $n-r+k+i+1$. If the customer in service leaves the system, customer c' enters the server and will remain there until his service is completed. For $0\le i<r$, we get:

$$\overline{C}_{i+1}=\left(\sum_{k=0}^{\infty}(T^k t)\otimes D^k\right)\left(\alpha\otimes\prod_{h=1}^{i}(D_0+p_{r-h}D_1)(1-p_{r-i-1})D_1\right). \tag{18}$$

If the service time is unbounded, an infinite sum occurs in both Eqn. (17) and (18). This sum can be truncated after k' terms if $\sum_{k=k'}^{\infty}(T^k t)\otimes D^k<\epsilon$, for some ϵ small. Because T, resp. D, is a substochastic, resp. stochastic matrix, we can always find such a k'.

The LJG algorithm can be used to find the steady state vector $\bar{\pi}=(\bar{\pi}_0,\bar{\pi}_1,\ldots,\bar{\pi}_r)$ of $\overline{P}$. The steady state vector $\pi'=(\pi'_0,\pi'_1,\ldots)$ of P' obeys the following relation:

$$\pi'_i=\bar{\pi}_i/c \text{ and } \pi'_{r+j}=\bar{\pi}_r(T^j\otimes I_m)/c,$$

for $0\le i<r$ and $j\ge 0$. The normalization constant c can be calculated by $c=\sum_{i=0}^{r-1}\bar{\pi}_i e+\bar{\pi}_r\big((I_{m_{ser}}-T)^{-1}\otimes I_m\big)e$. As before, the probability that a customer receives a complete service and has a response time of i time units is found as $P[X'=i]=\frac{1}{\lambda}\sum_{s=1}^{m_{ser}}(t)_s\sum_{j=1}^{m}(\pi'_i)_{(s,j)}$, and the rejection probability $P'_{out}=1-\sum_{i>0}P[X'=i]$.

Remark. In this case one easily obtains the relation $1-P'_{idle}=\lambda(1-P'_{out})/\mu$, as $\sum_{i,j}(\pi'_i)_{(s,j)}=\beta_s$, where $\beta(T+t\alpha)=\beta$, $\beta e=1$ and $\beta t=1/\mu$. This result is also intuitively clear as the server only serves successful customers and the mean service time of such a customer is $1/\mu$.

[d] Recall, a customer who reaches his critical age on the exact moment that the server becomes available, leaves the system nevertheless.

6. QUASI-BIRTH-DEATH REDUCTION

In the previous sections, we constructed a GI/M/1-type Markov chain to compute the response time distribution of the D-MAP/PH/$1+\mathrm{GI}$ queue. A limitation of this method lies in the fact that the time complexity of the algorithm is a square function of r, the maximum tolerable customer patience. In this section, we introduce a different approach, using QBDs, which speeds up the computational process. The idea behind this reduction was also applied in Van Houdt[13] to calculate the response time distribution in a queue with multiple types of customers (i.e., the MMAP[K]/PH[K]/1 queue).

In order to construct the QBD we add m states to each level i, for $0 < i < r$, to the state space of the transition matrix of interest (P, $\widehat{P}$ or $\overline{P}$). These additional states are referred to as artificial states. The basic idea behind this construction is to replace a transition from level i to $i-k$, for $k \geq 0$, by $k+1$ transitions, where for each of the first k transitions we decrease the level by one, while for the $(k+1)$st transition the level will remain identical.[e] Thus, instead of making a transition from level i to $i-k$ at once, the new MC will visit k intermediate states, which shall all be artificial states. We will only need m artificial states due to the specific form of the matrices that correspond to transitions causing the level to decrease.

Roughly speaking, we can explain the idea behind the QBD reduction as follows. A transition from level i to $i-k < i$ can be regarded as *scanning* the $k+1$ time instants $n-i+1, \ldots, n-i+k+1$ for a customer arrival who is still present in the system at time $n+1$ (where n is the current time instant). The first k of these scans results in a negative result, whereas the last one is successful (unless $i-k=0$). Whether the lth scan is successful, is solely determined by the D-MAP state at time $n-i+l$, say state j_l, and the amount of patience $i-l+1$ needed to be in the system at time $n+1$. Thus, instead of going directly from some state (s, j_1) of level i to a state of the form (s', j') of level $i-k$, we make a series of $k+1$ transitions[f]: the first one to artificial state j_2 of level $i-1$, the lth, for $l = 2, \ldots, k$, going from artificial state j_l of level $i-l+1$ to state j_{l+1} of level $i-l$ and the last one from artificial state j_{k+1} of level $i-k$ to state (s', j') of level $i-k$. We do not need to keep track the phase s' as this is determined by the vector α.

We only discuss the system presented in section 3 in detail, the QBD Markov chain for the other two systems can be set up in an analogue way. First, we add m additional states to every level, except the first and the last one (level zero and level r). As explained above, these states will correspond to those of the D-MAP arrival process. Define $\Omega_i = \{j | 1 \leq j \leq m\}$,

[e] Actually, for $i-k=0$ we split the transition into k steps instead of $k+1$.

[f] If $i-k=0$, we have only k transitions, the kth one going from artificial state j_k at level one to state j_{k+1} at level zero.

for $0 \leq i < r$ and $\Psi_i = \{(s,j)|1 \leq s \leq m_{ser}, 1 \leq j \leq m\}$, for $0 < i \leq r$. Using these definitions, the states of level zero are denoted as Ω_0, those of level i $(0 < i < r)$ as $\Omega_i \cup \Psi_i$ and the states of level r are denoted as Ψ_r. The states of Ω_0 and Ψ_i $(0 < i \leq a)$ correspond to those of the Markov chain presented in section 3. The states of Ω_i $(1 < i < r)$ are those that have been added, which we call the artificial states. For our convenience, let us call the other states, original states and let us define $m_{art} = m_{tot} + m$.

Define the QBD using the following $rm_{art} \times rm_{art}$ matrix P^*:

$$P^* = \begin{bmatrix} B_1^* & B_0^* & 0 & 0 & \dots & 0 & 0 \\ B_2^* & A_1^{1*} & A_0^{1*} & 0 & \dots & 0 & 0 \\ 0 & A_2^{2*} & A_1^{2*} & A_0^{2*} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_2^{r-1*} & A_1^{r-1*} & C_0^* \\ 0 & 0 & 0 & \dots & 0 & C_2^* & C_1^* \end{bmatrix},$$

where B_1^* is an $m \times m$ matrix, B_2^* an $m_{art} \times m$ matrix, B_0^* an $m \times m_{art}$ matrix and A_0^{i*}, A_1^{i*} $(0 < i < a)$ and A_2^{i*} are $m_{art} \times m_{art}$ matrices. Also, the matrix C_0^* is an $m_{art} \times m_{tot}$ matrix, C_1^* an $m_{tot} \times m_{tot}$ and C_2^* is an $m_{tot} \times m_{art}$ matrix.

The matrices B_1^* and B_0^* represent the situation in which the QBD is at level zero, that is, the system is idle. As the transitions from level 0 are unaltered, the only difference with the expressions for B_1 and B_0 is caused by the presence of the artificial states added to level one, therefore $B_1^* = B_1$ and $B_0^* = \begin{bmatrix} 0 & B_0 \end{bmatrix}$. The expression for the $m_{art} \times m_{tot}$ matrix B_2^* can be divided into two parts. The first m rows correspond to the situation in which the QBD is in an artificial state of level one at time n. If no customer arrives at this moment, the QBD makes a transition to level zero. These probabilities are, by definition, given by the matrix D_0. The other rows represent the situation in which the age of the customer in service equals one time unit.[g] In this case, the QBD makes a transition to level zero if this customer leaves the server and there is no arrival at the current time instant. Notice, the waiting room is empty because the customer in service arrived at time $n - 1$. Hence,

$$B_2^* = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} D_0 \\ (t + a_1(e - t)) \otimes D_0 \end{bmatrix}. \tag{19}$$

The same arguments can be used to determine an expression for the matrices A_2^{i*} $(1 < i < r)$, which contain the probabilities of a transition from level i to level $i - 1$. In order to have such a transition, there can be

[g] The QBD is in an original state of level one at time n.

no arrival $i-1$ time units ago of a customer with a critical age greater than or equal to i time units (that is, the *scan* of time instant $n-i+1$ ought to be unsuccessful). When the QBD makes a transition to a lower level, the resulting state is always an artificial state. Therefore, all the entries in the last m_{tot} columns are equal to zero and we get,

$$A_2^{i*} = \begin{bmatrix} D_0 + p_{i-1}D_1 & 0 \\ \left(t + \frac{a_i}{1-p_{i-1}}(e-t)\right) \otimes (D_0 + p_{i-1}D_1) & 0 \end{bmatrix}. \tag{20}$$

The transitions to a higher level are identical to those in section 3. Moreover, we never increase the level from an artificial state, implying

$$A_0^{i*} = \begin{bmatrix} 0 & 0 \\ 0 & A_0^i \end{bmatrix} \quad \text{and} \quad C_0^* = \begin{bmatrix} 0 \\ A_0^{r-1} \end{bmatrix}. \tag{21}$$

The probabilities of a transition between two original states of level i, are given by the matrix A_1^i, for $0 < i < r$ and by C_1 for $i = r$. Apart from the transitions between original states, a transition from an artificial to an original state at level i can occur when a customer with a critical age of at least i time units arrives (that is, the *scan* at time $n-i+1$ is a success). This yields,

$$A_1^{i*} = \begin{bmatrix} 0 & (1-p_{i-1})\alpha \otimes D_1 \\ 0 & \left(t + \frac{a_i}{1-p_{i-1}}(e-t)\right)\alpha \otimes (1-p_{i-1})D_1 \end{bmatrix} \tag{22}$$

and $C_1^* = e\alpha \otimes a_r D_1$.

Finally, the probabilities that the QBD makes a transition from level r to level $r-1$, can be found in the $m_{tot} \times m_{art}$ matrix C_2^*. The customer in service leaves the system, whether he finishes his service or not. There is a transition to level $r-1$ if no customer of age r will be in the queue at the next time instant. Hence, $\widetilde{C}_2^* = \left[\, e \otimes (D_0 + (1-a_r)D_1) \; 0 \,\right]$.

The matrix P^* is a finite level-dependent QBD matrix, therefore, its steady vector $\pi^* = (\pi_0^*, \pi_1^*, \ldots, \pi_r^*)$ can be computed by a variant on the LJG algorithm, described in Gaver[9]. For reasons of completeness, the details of this algorithm are presented in Appendix A. In this appendix, we also indicate how to obtain the steady state vector π of P from π^*. Having found π we can apply the formulas given at the end of section 3 to obtain the performance measures of interest.

7. NUMERICAL EXAMPLES

In this section, we discuss some fairly arbitrary numerical examples of systems with impatient customers. The D-MAP arrival process is characterized by the following two matrices:

$$D_0 = \begin{bmatrix} 0.76 & 0.19 \\ 0.09 & 0.81 \end{bmatrix} \quad \text{and} \quad D_1 = \begin{bmatrix} 0.04 & 0.01 \\ 0.01 & 0.09 \end{bmatrix}.$$

This arrival process has a mean arrival rate $\lambda = 1/12$. The service time of a customer consists of two components. Every customer needs an initial service time of 3 time units. On top of that, with probability 0.4, 0.4, and 0.2 a customers needs some extra, geometrically distributed, service with an average of 5, 10, and 20 time units, respectively. We have, $\alpha = (1, 0, 0, 0, 0, 0)$ and

$$T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.95 \end{bmatrix}.$$

The mean service time $1/\mu$ equals 13 time units, meaning without the customer impatience we would have an unstable system as $\lambda/\mu = 13/12 > 1$.

We consider 4 possible patience distributions Z, each having the same mean $E[Z]$ of 275 time units:

Z_a: One half of the customers runs out of patience after 50 time units, whereas the other half will be patient for 500 time units, i.e., $a_{50} = a_{500} = 0.5$.

Z_b: The critical age of a customer equals 50, 175, 275, 375 or 500 time units, each with a probability 0.2.

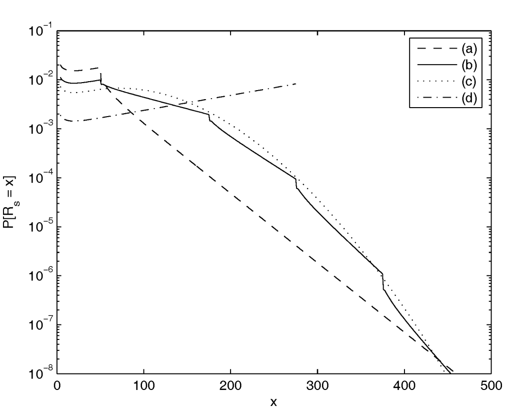
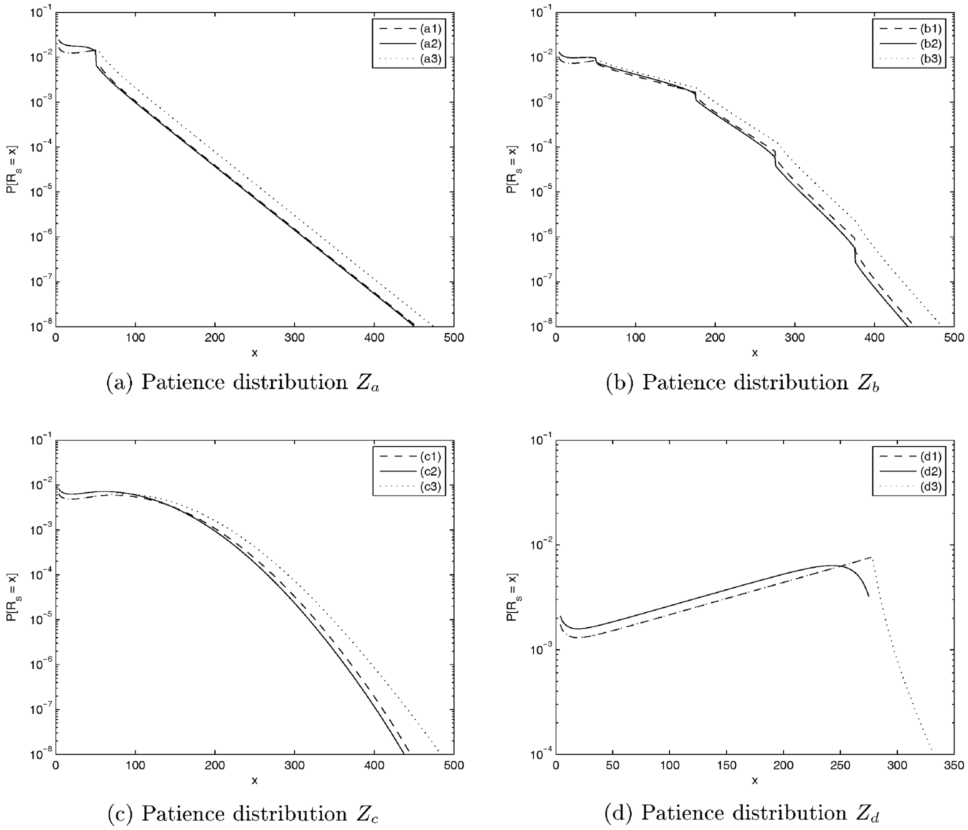


FIGURE 1 Response time distribution of successful impatient customers.



(a) Patience distribution Z_a

(b) Patience distribution Z_b

(c) Patience distribution Z_c

(d) Patience distribution Z_d

FIGURE 2 Response time distributions of successful impatient customers in the entire system/waiting room.

Z_c: The patience distribution of the customers is uniformly distributed between 50 and 500 time units.

Z_d: All customers have the same amount of patience, being 275 time units.

For each of the patience distributions considered, Figure 1 shows us the response time distribution R_s of a successful customer, who remains impatient in the entire system (i.e., section 3). The behavior of these curves can be understood as follows: when considering the probability that a successful customer has a response time of x time units, one might intuitively say that the load of the system $\rho_x \approx \lambda P[Z \geq x]/\mu$. Notice, ρ_x decreases as a function of x and equals 13/12 for $x = 1$. As such, we expect that the curves in Figure 1 start to decrease as soon as ρ_x drops below 1. Moreover, the smaller ρ_x becomes the sharper the decrease. For the distributions Z_a and Z_b, the value $P[Z \geq x]$ decreases only in a few steps, which correspond to the different "stairs" in those curves. The curve corresponding to Z_c reaches a maximum at $x = 72$, even though

TABLE 1 Expected number of lost customers

Distribution	Impatient in server	Server time aware	Patient in server
Z_a	0.1885	0.1437	0.1673
Z_b	0.1415	0.1286	0.1324
Z_c	0.1192	0.1047	0.1144
Z_d	0.0918	0.0334	0.0873

$\rho_{72} \approx 1.03 > 1$. This might be explained by the fact that the mean time that a customer spends in the server is less than $1/\mu$, due to the impatience (i.e., ρ_x slightly overestimates the load).

In Figure 2, we compare the distribution R_s of the three systems discussed in sections 3-5, for each of the patience distributions considered. The system with impatient customers in the entire system is represented by a dashed line (section 3), service time aware customers correspond to a full line (section 4), whereas the situation in which the customers are only impatient in the waiting room is given by a dotted line (section 5). The computation times for the curves in this figure, using an AMD Athlon 2.0 GHz processor with 512 Mb of memory, are about 173 seconds for the method using the GI/M/1 type Markov chains and about 21 seconds when applying the QBD approach (see Section 6). Because the maximum critical age of a customer equals either 275 or 500 time units and the number of states of the arrival process is relatively small, the method using QBDs takes remarkably less time while the memory requirements only increase with a factor 1.36.

Table 1 represents the probabilities that a customer abandons the system without receiving a complete service. If we compare the four patience distributions, the probability that a customer leaves the system early decreases together with the variance of the patience distribution Z, while the average response time of a customer (given by Table 2) increases. Whether becoming patient while entering the server reduces the rejection probability (as is the case in our numerical example) depends to a great extent on the variation of the service time distribution (see Van Houdt[14]). An interesting open problem related to this result is whether, given a

TABLE 2 Expected response time of a successful impatient customer

Distribution	Impatient in server	Rejected by server	Patient in server
Z_a	38.9042	34.4974	46.1523
Z_b	73.4010	68.1397	80.1541
Z_c	90.4167	83.9073	98.0788
Z_d	179.8562	171.7786	189.5891

mean m_u, the deterministic patience distribution would result in the lowest rejection probability P_{out} of all patience distributions Z with a mean $E[Z] = m_u$. For the system with customers who are only impatient while waiting, one can prove that this is the case for the $M/M/1 + GI$ queue by making use of Theorem 3.1 from Brandt[6] or Boxma[5].

APPENDIX A: COMPUTING THE STEADY STATE VECTOR π OF P VIA THE QBD MC P*

The matrix $Q^* = P^* - I$ is a tridiagonal infinitesimal generator block matrix. Thus, we can apply the algorithm presented in Gaver[9], the time and space complexity of which equal $O(rm_{art}^3)$ and $O(rm_{art}^2)$, respectively:

1. Input: D_0 and D_1 of the arrival process, the PH distribution, characterized by α and T and the stochastic vector $\tilde{a}$.
2. Calculate the matrices G_i, for $0 \leq i \leq r$, by means of the equations:

   - $G_0 = B_1^* - I_m$
   - $G_1 = A_1^{1^*} - I_{m_{art}} + B_2^*(-G_0^{-1})B_0^*$
   - $G_i = A_1^{i^*} - I_{m_{art}} + A_2^{i^*}(-G_{i-1}^{-1})A_0^{i-1^*}$ (for $1 < i < r$)
   - $G_r = C_1^* - I_{m_{tot}} + C_2^*(-G_{r-1}^{-1})C_0^*$

3. The steady state probabilities are found using the following expressions: $\pi_r^* G_r = 0$, with $\pi_r^* e = 1$, $\pi_{r-1}^* = \pi_r^* C_2^*(-G_{r-1}^{-1})$, $\pi_i^* = \pi_{i+1}^* A_2^{i+1^*}(-G_i^{-1})$, for $i = r-2, \ldots, 1$, $\pi_0^* = \pi_1^* B_2^*(-G_0^{-1})$ and $\sum_{i=0}^{r} \pi_i^* e = 1$.
4. Finally, compute the steady state vector π of the original transition matrix P as follows. Denote π_i^* $(0 < i < r)$ as $(\pi_i^*(m), \pi_i^*(m_{tot}))$, where the probabilities of the row vector $\pi_i^*(m)$ of size m correspond to the artificial states and those of the $1 \times m_{tot}$ vector $\pi_i^*(m_{tot})$ to the original states. Then, the steady state vector π can be computed by $\pi_0 = \pi_0^*/(1-d)$, $\pi_i = \pi_i^*(m_{tot})/(1-d)$, for $1 \leq i < r$ and $\pi_r = \pi_r^*/(1-d)$, where $d = \sum_{i=1}^{r-1} \pi_i^*(m)e$.

The advantage of this algorithm in comparison with the LJG algorithm lies in the difference in the time complexity. The time needed by the algorithm using QBDs is only linear in the maximum critical age r of a customer, whereas with the GI/M/1 type MC it is a square function of r.

REFERENCES

1. Baccelli, F.; Boyer, P.; Hebuterne, G. Single-server queues with impatient customers. Adv. in Appl. Prob. **1984**, *16*, 887–905.
2. Barrar, D.Y. Queueing with impatient customers and ordered services. Operations Research **1957**, *5*, 650–656.

3. Blondia, C. A discrete-time batch markovian arrival process as B-ISDN traffic model. Belgian Journal of Operations Research, Statistics and Computer Science **1993**, *32* (3,4), 3–23.
4. Blondia, C.; Casals, O. Statistical multiplexing of VBR sources: A matrix-analytical approach. Performance Evaluation **1992**, *16*, 5–20.
5. Boxma, O.J.; de Waal, P.R. Multiserver queues with impatient customers. In *Proceedings of ITC-14*; Labetoulle, J., Roberts, J. Eds.; Elsevier Science B.V.: Amsterdam (North-Holland), 1994; 743–756.
6. Brandt, A.; Brandt, M. One the M(n)/M(n)/s queue with impatient calls. Performance Evaluation **1999**, *35*, 1–18.
7. Combé, M.B. Impatient customers in the MAP/G/1 queue. Technical Report BS-R9413, CWI; Amsterdam, April 1994.
8. Garnet, O.; Mandelbaum, A.; Reiman, M. Designing a call center with impatient customers. Manufacturing & Service Operations Management **2002**, *4*, 208–227.
9. Gaver, D.P.; Jacobs, P.A.; Latouche, G. Finite birth-and-death models in randomly changing environments. Adv. Appl. Prob. **1984**, *16*, 715–731.
10. Lang, A.; Arthur, J.L. Parameter approximation for phase-type distributions. In *Matrix-Analytic Methods in Stochastic Models*; Chakravarthy, S.R., Alfa, A.S., Eds.; Marcel-Dekker, Inc.: New York, 1996; 151–206.
11. Latouche, G.; Jacobs, P.A.; Gaver, D.P. Finite markov chain models skip-free in one direction. Naval Research Logistics Quarterly **1984**, *31*, 571–588.
12. Palm, C. Methods of judging the annoyance caused by congestion. Tele. (English Edition). **1953**, *2*, 1–20.
13. Van Houdt, B.; Blondia, C. The waiting time distribution of a type k customer in a MMAP[K]/PH[K]/c $(c = 1, 2)$ queue using QBDs. Stochastic Models **2004**, *20* (1), 55–69.
14. Van Houdt, B.; Lenin, R.B.; Blondia, C. Delay distribution of (im)patient customers in a discrete time D-MAP/PH/1 queue with age dependent service times. Queueing Systems and Applications **2003**, *45* (1), 59–73.
15. Zhao, Y.Q.; Alfa. A.S. Performance analysis of a telephone system with both patient and impatient customers. Telecommunication Systems **1995**, *4*, 201–215.