Queueing Syst
DOI 10.1007/s11134-014-9402-1

Appointments in care pathways: the $Geo^x/D/1$ queue with slot reservations

Maartje E. Zonderland · Richard J. Boucherie · Ahmad Al Hanbali

Received: 1 April 2013 / Revised: 8 March 2014

Abstract Motivated by the increasing popularity of care pathways in outpatient clinics, where some patients complete a significant part of the path in one day, hospitals aim to optimize the flow of these patients by prioritizing them in the appointment planning process. This paper considers the $Geo^x/D/1$ queue with slot reservations that serves regular patients and priority patients. Priority patients reserve a time slot in a reservation window and are blocked when all slots in the reservation window are occupied by other priority patients. The reservation window models the advance reservation of service slots by patients on a care pathway. We model the $Geo^x/D/1$ queue as a $M/G/1$-type queue and apply a matrix-analytic approach, which simplifies to a matrix-geometric solution. We use the vector generating functions to derive the patients' waiting times. Numerical experiments illustrate the influence of the reservation window on the number of regular and priority patients present and the blocking probabilities for priority patients.

Keywords Matrix-geometric analysis · Probability generating functions · Healthcare · Appointment planning · Care pathways

Mathematics Subject Classification 60K25 · 90B22

M. E. Zonderland (✉)· R. J. Boucherie
Stochastic Operations Research & Center for Healthcare Operations Improvement and Research, University of Twente, Postbox 217, 7500 AE Enschede, The Netherlands
e-mail: maartje@zonderlandzorglogistiek.nl

A. Al Hanbali
Industrial Engineering and Business Information Systems, University of Twente, Postbox 217, 7500 AE Enschede, The Netherlands

Published online: 16 April 2014

1 Introduction

Care pathways have gained popularity in healthcare over the last two decades [1]. A care pathway is a management tool to organize multidisciplinary care for patients with similar characteristics (e.g., disease symptoms, diagnosis, age, etcetera). The care pathway specifies the steps in the care process [2] and routes patients along a predefined path of care providers and diagnostic facilities. Patients may complete a significant part of the path in one day (one-stop shop). Due to the large number of hospital facilities that may be incorporated in the path, planning these patients is quite involved. Hospitals therefore tend to prioritize the care pathway patients in the planning process. It is thus not uncommon to reserve slots for care pathway patients in a walk-in clinic, where the main flow of patients arrives without appointments. Examples include diagnostic services, such as Radiology outpatient clinics (X-ray, CT) and blood withdrawal facilities. When these clinics are highly utilized, reserving even a few slots for care pathway patients may lead to a significant increase of walk-in patients' waiting times.

This paper considers the $Geo^x/D/1$ queue with slot reservations that serves regular patients and priority patients. Service is slotted, which represents the common appointment system in outpatient clinics. The number of regular patients arriving in a time slot has a geometric distribution. A regular patient takes the first available position in the queue. The number of priority patients arriving in a time slot also has a geometric distribution. A priority patient reserves a time slot in a reservation window between L and H slots in the future. This mimics the advance reservation of a time slot by a priority patient. If time slot L is occupied by a regular patient, the priority patient takes the position of the regular patient, and all regular patients from position L onward are shifted to the next time slot. If slot L is occupied by a priority patient, then the new priority patient moves to the first available slot in the interval $L+1, \dots, H$ that is not occupied by a priority patient, and takes that position. All regular patients from that position onward shift to the next available slot. If all slots in the interval $L, \dots, H$ are occupied by priority patients, the new priority patient is blocked, which models when this priority patient cannot be served in a one-stop shop fashion.

We model the $Geo^x/D/1$ queue as a Matrix-Geometric queue of the $M/G/1$-type, where the levels correspond to the number of regular patients, and the phases are the slots in the range $1, \dots, H$ that are occupied by priority patients. Due to the special structure of the state space, we may factorize the sub matrices of the transition probability matrix into the probability of j regular arrivals in a slot, $j = 0, 1, 2, \dots$, and the behavior of priority patients. This factorization allows for a detailed analysis of the equilibrium distribution similar to that of the $M/G/1$ queue, but with states replaced by levels, and allows us to mimic the generating function approach for the $M/G/1$ queue; see [7]. Using a matrix-analytical approach, we derive the vector generating function of the equilibrium distribution of the $Geo^x/D/1$ queue with slot reservations. Numerical experiments illustrate the influence of the reservation window on the number of regular and priority patients present and the blocking probabilities for priority patients.

The service and hospitality industry is quite familiar with policies where part of an unscheduled customer stream is diverted and scheduled to a later moment on the day. This concept is also known as virtual queuing [3,4]. Probably the most famous organization that employs virtual queuing is Walt Disney, using the FastPass system for the most popular attractions in its theme parks [5]. Park guests decide upon arrival at an attraction whether they want to join the waiting line, or get a ticket (the "FastPass"), which gives them a time-frame to return and subsequently enter the attraction without waiting. To avoid a large number of no-shows and long waiting time for the non-FastPass guests, it is only allowed to possess a FastPass ticket for a single attraction. The queuing system behind FastPass is analyzed in [6]. In the FastPass system park guests are supported by information on the state of both the regular and FastPass queue (e.g., the waiting time in the regular queue and the given return time for the FastPass ticket) and decide upon arrival at the attraction which queue they want to join. In this paper, however, the two patient types originate from separate arrival processes (walk-in or care pathway) which determine their type and thus the queuing discipline.

The remainder of this paper is organized as follows. In the next section we describe our queuing model, followed by the analysis in Sect. 3. In Sect. 4 we include numerical experiments, and we conclude with the discussion in Sect. 5.

2 Model

For ease of notation we refer to the care pathway patients as priority, or type 1 patients, and to the walk-in patients as regular, or type 2 patients.

2.1 Assumptions

We consider a hospital clinic that serves regular and priority patients. Both patient types have a deterministic service time requirement of one unit. We consider the state of the system at discrete times 0, 1, 2, . . ., and start with an empty system at time 0. Service may start at discrete times 0, 1, 2, . . . only. A precise definition of the system state shall be given later on.

The number of priority and regular arrivals per slot follow two, independent, geometric distributions with success probabilities q_1 and q_2, respectively. The service discipline among patients of the same type is according to FCFS order. In case of multiple arrivals per slot the first arrival is served first then the second and so forth. An arriving regular patient takes the first available slot, i.e., the slot closest to the server. An arriving priority patient reserves the first available slot h in the slot reservation window, $L \leq h \leq H$, where $1 \leq L \leq H < \infty$. When all slots in the reservation window are booked by priority patients, the newly arrived priority patient is blocked and cleared. When the slot selected by the priority patient is occupied by a regular patient, then the regular patient is shifted one slot backward, i.e., one slot to the left in Fig. 1. If that slot was occupied by another regular patient, then also this patient is shifted one slot backward, etcetera. Using this reservation scheme, the priority patients are to a large extent protected from the additional waiting time due to an increasing flow of regular patients during some periods of time. On the other hand, the regular patients may experience significant delay when one or more priority patients join the queue.

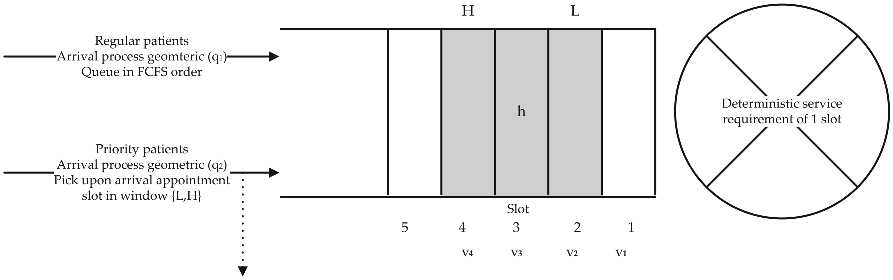

Fig. 1 The $Geo^x/D/1$ queue with slot reservations, slot reservation window $(L,\ldots,H) = (2, 3, 4)$, and $h = 3$

Upon a service completion, if there is a patient in the first slot, e.g., slot 1 in Fig. 1, this patient is served. In the case of no regular patients waiting in the queue, the server may be idle, even though there may be a priority patient on a slot position higher up in the queue. This implies that we are considering a non-work-conserving queuing system.

As a convention, we assume the following sequence of events. At time i, $i \geq 1$, (i) the patient in service, if any, departs from the system, (ii) the patient that occupies the first slot will be admitted in service, and the position of all patients in the queue is shifted one slot forward, i.e., one slot to the right in Fig. 1, (iii) the newly arrived regular and priority patients in the interval at time i are added to the queue according to the reservation scheme explained above, finally (iv) we record the system state.

We model our queuing system as a discrete-time Markov chain with infinite state space. To do so, we need to keep track of the following stochastic processes: (i) $N_2(i)$, $i \geq 1$, total number of regular patients waiting in the queue at the end of each time slot, and (ii) $V(i)$, $i \geq 1$, a state vector of length H with entries denoting which slot in the set $\{1,\ldots,H\}$ contains a priority patient. More specifically, the n-entry, $n \in \{1,\ldots,H\}$, of $V(i)$ is equal to one if slot n is occupied by a priority patient and zero otherwise. The system state at time i is defined as the two-dimensional process $\{(N_2(i), V(i)) : i = 0, 1, \ldots\}$. We shall refer to $N_2(i)$ as the level process and $V(i)$ as the phase process. Since the jumps $N_2(i) - N_2(i-1) \in \{-1, 0, 1, \ldots\}$, $i \geq 1$, we conclude that the transition probability matrix P of $(N_2(i), V(i))$ is an $M/G/1$-type matrix, also known as skip-free to the left, i.e., it has the following canonical form, see [7],

$$P = \begin{pmatrix} C_0 & C_1 & C_2 & \cdots & C_s & \cdots \\ B_0 & A_0 & A_1 & \cdots & A_{s-1} & \cdots \\ 0 & B_0 & A_0 & \cdots & A_{s-2} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \cdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

where the matrices A_i, B_0, and C_i shall be defined below based on elementary matrices D, A^*, B^*, and C^*. Here D represents the one step transition probability matrix of the reservation priority vector from $V(i)$ to $V(i+1)$, A^* and B^* are used for the representation of the transition of the number of regular patients from $N_2(i)$ to

$N_2(i+1)$ to the right and left, respectively, and C^* is used in the special case to represent the transition from $N_2(i) = 0$ to $N_2(i+1)$ to the right. We emphasize that the one step transitions of the priority vector $V(i)$ are independent of the number of regular patients present. However, the one step transitions of the number of regular patients $N_2(i)$ depend on the priority vector $V(i)$, see Sect. 2.3.

2.2 Priority vector transition probability matrix D

Recall that the reservation priority vector $V(i)$ is of length H, and its entries specify which slots in the set $\{1, \ldots, H\}$ contain priority patient slot reservations. Let $V(i) := \big(V_1(i), \ldots, V_H(i)\big)$, where $V_h(i)$ is a binary random variable, which is equal to 1 when slot h is reserved by a priority patient and 0 otherwise. Note that the slots $\{1, \ldots, L-1\}$ cannot be selected directly by priority patients, they possibly contain slot reservations of previously arrived priority patients and thus should be taken into account in the analysis. We shall refer to a realization of the random vector $V(i)$ by $\big(v_1(i), \ldots, v_H(i)\big)$, $i = 1, 2, 3, \ldots$.

At the end of each time slot $V(i)$ is updated; existing slot reservations are moved forward one slot, and new slot reservations, if any, are added. Let Ω denote the state space of $V(i)$. The set Ω contains $2^{L-1}(H-L+2)$ vectors, each with length H. We order the states of Ω lexicographically, so that the states with $v_1(.)$ equal to zero come first. For example, when $L = 2$ and $H = 3$, $\Omega = \{(000), (010), (011), (100), (110), (111)\}$. Note that positions subsequent to a zero position in the slot reservation window are always equal to zero as well, since priority patients occupy the first available slot in the reservation window. It follows immediately that the one step transition probability matrix, D, which defines the transitions for $V(i) \to V(i+1)$, has size $2^{2(L-1)}(H-L+2)^2$. We emphasize that Ω is not necessarily of size 2^H. In the following, we shall first analyze these cases and afterward report the non-zero transition probabilities. The transitions for $V(i) \to V(i+1)$ where $v_L(i+1) = 0$ and $v_h(i+1) = 1$ for some $h \in \{L+1, \ldots, H\}$ have zero probability due to the definition of the priority patients reservation scheme. Moreover, according to the sequence of events the slots are first shift forward by one slot then new priority patients are added. Therefore, the transition from $\big(v_1(i), \ldots, v_H(i)\big)$ to $\big(v_1(i+1), \ldots, v_H(i+1)\big)$ has a positive probability if the following condition is satisfied: $\big(v_1(i+1), \ldots, v_{L-1}(i+1)\big) = \big(v_2(i), \ldots, v_L(i)\big)$ and $\big(v_L(i+1), \ldots, v_{H-1}(i+1)\big) \geq \big(v_{L+1}(i), \ldots, v_H(i)\big)$, otherwise this transition has a zero probability. The non-zero transition probabilities of D are given by

$$\mathbb{P}\big[V(i+1) = \big(v_1(i+1), \ldots, v_H(i+1)\big) \mid V(i)\big] = \begin{cases} (1-q_1)q_1^n, & v_H(i+1) = 0, \\ q_1^n \quad, & v_H(i+1) = 1, \end{cases}$$

where n denotes the number of accepted priority arrivals at time i, given by

$$n = \sum_{h=L}^{H} v_h(i+1) - \sum_{h=L+1}^{H} v_h(i). \tag{1}$$

Note that D is aperiodic and irreducible. We shall denote by π^∞ its unique steady state probability. In the following, we are interested in the subset of Ω with vectors that have zero as a first entry, i.e., $v_1(.) = 0$. Let Ω_0 denote this subset with cardinality $|\Omega_0|$. Note that, $|\Omega_0| = 1$ for $L = 1$ and $|\Omega_0| = |\Omega|/2$ otherwise. This is because according to our reservation priority scheme when $L = 1$, the subset $\Omega_0 = \{(0, \ldots, 0)\}$.

2.3 Level transition probability matrices A_j, B_0, and C_j

Recall that the levels represent the number of regular patients present in the queue. The regular patient transition probability matrices, which contain the probabilities for transitions of $N_2(i)$, depend on the slot reservation vector $V(i)$. Since we consider one step transitions, only $v_1(i)$ is of interest to determine whether or not a regular patient is served. Let us first introduce some notation. Let **u** and **w** denote row vectors of length $2^{L-1}(H - L + 2)$. The first $|\Omega_0|$ entries of **u** equal q_2, and the rest equal 1. Similarly, the first $|\Omega_0|$ entries of **w** equal 1, and the rest equal 0. Furthermore, define **e** as the row vector with all elements equal to 1 and of length $2^{L-1}(H-L+2)$. Let $A^* = \mathbf{u}^T \times \mathbf{e}$, $B^* = \mathbf{w}^T \times \mathbf{e}$, and $C^* = \mathbf{e}^T \times \mathbf{e}$. We define the following elementary matrices that are used to find the level transition matrices A_j, B_0, and C_j, $j = 0, 1, 2, \ldots$,

$$A_j^* = a_j A^*, \; B_0^* = a_0 B^*, \; C_j^* = a_j C^*, \tag{2}$$

where $a_j = (1 - q_2)q_2^j$. The intuition behind the definitions of elementary matrices A_j^*, B_0^*, and C_j^* is as follows. When $N_2(i) > 0$ and $V_1(i) = 0$, i.e., slot one is reserved for a regular patient, exactly $j + 1$ regular patients should join the queue such that the level process jumps from l at time i to $l + j$ at time $i + 1$. However, when $N_2(i) > 0$ and $V_1(i) = 1$, i.e., slot one is reserved for a priority patient, exactly j regular patients should join the queue so that the level process jumps from l at time i to $l + j$ at time $i + 1$. This justifies the definition of **u** and explain why it is multiplied by a_j to get A_j^*. Similarly, when $N_2(i) > 0$ and $V_1(i) = 0$, exactly zero regular patients should join the queue so that the level process jumps from l at time i to $l - 1$ at time $i + 1$. However, when $N_2(i) > 0$ and $V_1(i) = 1$, it is not possible that the level process jumps from l at time i to $l - 1$ at time $i + 1$. This justifies the definition of **w** and explain why it is multiplied by a_0 to get B_0^*. Note that when $N_2(i) = 0$, regardless of $V_1(i)$, exactly j regular patients should join the queue such that the level process jumps from 0 at time i to j at time $i + 1$. The priority and regular patient arrival processes are independent. Therefore, the level transition matrices A_j, B_0, and C_j are found by multiplying A_j^*, B_0^*, and C_j^* element-wise with D, i.e., every (s, n)-entry of D is multiplied with the (s, n)-entry of A_j^*, B_0^*, and C_j^*. This completes the description of the transition probability matrix P.

Note that A_j can be rewritten as $a_j \bar{A} D$, where $\bar{A}$ is the diagonal matrix with the elements of **u** on the diagonal, B_0 can be written as $a_0 \bar{B} D$, where $\bar{B}$ is the diagonal matrix with the elements of **w** on the diagonal, and C_j can be written as $a_j D$.

3 Analysis

The structure of the matrix P is equal to that of the transition probability matrix of $M/G/1$-type queues, see [7]. An overview of discrete time queuing systems can be found in [9]. Several priority disciplines have been studied for discrete time queuing models, but these are usually related to the non-preemptive [10] or preemptive resume priority disciplines [11]. In [12] a different, but related, service discipline is considered, where a slot is reserved for regular patients at the end of the queue. Priority patients are placed within a fixed reservation window, not necessarily at the front of the queue. In the case of high load traffic from priority patients, it is then guaranteed that regular patients receive service as well.

We assume that the queue is stable, see Sect. 3.1. A key issue in computing the steady state probabilities Π, $\Pi P = \Pi$, is to find the minimal non-negative solution X of the following matrix equation (see, e.g., [8, Sect. 13.1] for further details):

$$X = B_0 + \sum_{i=1}^{\infty} A_{i-1} X^i. \tag{3}$$

By inserting $A_{i-1} = a_{i-1}\bar{A}D$ we find that X is the minimal non-negative solution of

$$B_0 + EX + q_2 X^2 = 0, \tag{4}$$

where $E = A_0 - q_2 B_0 - I$. State of the art algorithms to solve this quadratic equation is the logarithmic reduction and quadratically convergent algorithm, see [8, Sect. 8.2, 8.4],

$$X_{(n+1)} = -(E + q_2 X_{(n)})^{-1} B_0, \text{ (logarithmic)}, \tag{5}$$

$$X_{(n+1)} = -E^{-1}(B_0 + q_2 X_{(n)}^2), \text{ (quadratic)}, \tag{6}$$

with $X_{(0)} = 0$. Note that E is nonsingular since $(A_0 - q_2 B_0)^n \to 0$ for $n \to \infty$. The latter gives that $E - q_2 X$ is nonsingular.

Given our priority reservation scheme the size of D, equivalently X, is equal to $2^{2(L-1)}(H - L + 2)^2$. This yields that the logarithmic reduction and quadratically convergent algorithms have a complexity of order $2^{3(L-1)}(H - L + 2)^3$. Moreover, given that X has been computed it is necessary to apply a rather complicated recursive scheme to find the steady state probability vector of levels, see, for example [8, Theorem 13.1.5]. Applying this scheme to our model we find that π_l satisfies the following matrix-geometric solution:

$$\pi_l = -\pi_0 S\big(q_2 I - \bar{A}S\big)^{l-1}, l \geq 1, \tag{7}$$

where $S = C_1(A_0 - I + q_2 X)^{-1}$. The previous simplified result is mainly due to the special structure of A_j, B_0, and C_j block matrices. By solving the linear system of equations $\pi_0(C_0 - I - SB_0) = 0$ together with the normalization condition which reduces to $\pi_0\big(I - S((1 - q_2)I + \bar{A}S)^{-1}\big)e = 1$, π_0 is computed exactly.

Remark We conclude that it is computationally prohibitive to attempt to use the previous algorithms for large values of L and H. For example, for $L = H = 10$ the complexity is of order 2^{30}; this is equivalent to one second run time with a processor of 1Ghz speed. However, when $L = H = 20$, the complexity is of order 2^{60} which is equivalent to a billion seconds run time with a processor of 1Ghz speed. Another problem with large L is due to limitation of the computer active memory (RAM). For example, for $L = H = 10$ the matrices will have a total number of entries equal to 2^{20}. Given that an entry occupies one byte of memory space, therefore, we need 1 Gbyte to store a full matrix. This already consumes a large part of the active memory. We note that the matrices that we have are not full matrices. Moreover, the extent to which these matrices are sparse depends on L given a certain value of H. Therefore, some memory space could be saved by, for example, a factor of 3 or 4 on average. But these savings are incomparable to the exponential growth of the matrix size. By taking into account this complexity, in Sect. 3.2 we shall use an approach based on generating functions to find the moments of the number of regular waiting patients in the queue.

3.1 Stability of the queue

In order for the queue to be stable, the mean load, ρ, should be less than one. Given that we have an M/G/1-type process, the stability condition is given, see [7, Sect. 3.6], by

$$\rho = \pi\beta^* < 1, \tag{8}$$

where π is the steady state probability of the matrix $B_0 + \sum_{n=0}^{\infty} A_n$, and β^* is defined by $\sum_{n=1}^{\infty} nA_{n-1}\mathbf{e}^T$. Inserting B_0 and A_n by $a_0\bar{B}D$ and $a_n\bar{A}D$, respectively, we find that

$$B_0 + \sum_{n=0}^{\infty} A_n = D, \quad \text{and} \quad \sum_{n=1}^{\infty} nA_{n-1}\mathbf{e}^T = \frac{1}{1-q_2}\mathbf{u}^T. \tag{9}$$

The first equality gives that π is equal to π^∞, the steady state probability of D. Inserting π and β^* by their values in the stability condition gives

$$\rho = \frac{1}{1-q_2}\pi^\infty\mathbf{u}^T = \frac{1}{1-q_2} - \sum_{h=1}^{|\Omega_0|}\pi_h^\infty < 1. \tag{10}$$

Recall that $|\Omega_0|$ is the cardinality of the set Ω_0 with all elements which have $v_1(.) = 0$. Note that a necessary condition for stability is $q_2 < \frac{1}{2}$. This condition is not sufficient, since the number of accepted priority patients per slot also depends on the blocking probability for priority patients.

3.2 Vector generating function of equilibrium probabilities π_s

We derive the vector generating function of the equilibrium probability π_s for $N_2(\infty)$, the number of regular patients present in the queue, where $s = 0, 1, \ldots$, jointly with $V(\infty)$, the priority reservation vector. Denote $L_1 = N_1(\infty)$, the number of priority patients in steady state. Note that we use a vector generating function approach instead of the more common, exact matrix-geometric approach. Applying the latter method would not be manageable, see the previous discussion on the complexity in the computation of X. From $\Pi P = \Pi$, we obtain

$$\pi_s = \pi_0 C_s + \sum_{i=1}^{s} \pi_i A_{s-i} + \pi_{s+1} B_0 \quad \text{for} \quad s \geq 1, \quad \text{and} \tag{11}$$

$$\pi_0 = \pi_0 C_0 + \pi_1 B_0, \quad \text{where} \quad \sum_{s=0}^{\infty} \pi_s \mathbf{e}^T = 1. \tag{12}$$

Define the vector generating function for π_s, $P_\Pi(z)$, as

$$P_\Pi(z) = \sum_{s=0}^{\infty} \pi_s z^s. \tag{13}$$

Furthermore, define

$$A(z) = \sum_{s=0}^{\infty} A_s z^s, \quad \text{and} \quad C(z) = \sum_{s=0}^{\infty} C_s z^s. \tag{14}$$

Multiplying both sides of (11) with the scalar z^s, where $|z| \leq 1$, and summing the result for $s = 0, 1, \ldots$, we obtain

$$\sum_{s=0}^{\infty} \pi_s z^s = \sum_{s=0}^{\infty} \pi_0 C_s z^s + \sum_{s=1}^{\infty} \sum_{i=1}^{s} \pi_i A_{s-i} z^s + \sum_{s=0}^{\infty} \pi_{s+1} B_0 z^s, \tag{15}$$

and it follows that

$$P_\Pi(z) = \pi_0 C(z) + P_\Pi(z) A(z) - \pi_0 A(z) + B_0 z^{-1} P_\Pi(z) - \pi_0 B_0 z^{-1}. \tag{16}$$

Multiplication of (16) with z and rearranging terms gives

$$P_\Pi(z) \left[zI - zA(z) - B_0 \right] = \pi_0 \left[zC(z) - zA(z) - B_0 \right]. \tag{17}$$

3.3 Mean number of regular patients

We derive the mean number of regular patients in the queue, $\mathbb{E}[L_2]$, following the analysis of [7], pp. 143–148. Let $z = 1$. First, note that

$$\begin{aligned} \mathbb{E}[L_2] &= P'_\Pi(1) \mathbf{e}^T, \\ P_\Pi(1) &= \pi^\infty, \\ A(1) + B_0 &= C(1) = D. \end{aligned} \tag{18}$$

The first derivative of (17) with respect to z is

$$\begin{aligned}&P'_{\Pi}(z)\,[zI - zA(z) - B_0] + P_{\Pi}(z)\left[I - A(z) - zA'(z)\right]\\&= \pi_0\left[C(z) + zC'(z) - A(z) - zA'(z)\right].\end{aligned} \tag{19}$$

For $z = 1$, it follows that

$$P'_{\Pi}(1)\,[I - D] + \pi^{\infty}\left[I - A(1) - A'(1)\right] = \pi_0\left[C(1) + C'(1) - A(1) - A'(1)\right]. \tag{20}$$

Denote $U = I - D + \mathbf{e}^T\pi^{\infty}$ and $K = I - \frac{1}{1-q_2}\bar{A}D$. Furthermore, note that $\frac{1}{1-q_2}D - \frac{1}{1-q_2}\bar{A}D = \bar{B}D$. From (20), using $P'_{\Pi}(1)\mathbf{e}^T\pi^{\infty} = \mathbb{E}[L_2]\pi^{\infty}$, we obtain

$$P'_{\Pi}(1)U + \pi^{\infty}K = \mathbb{E}[L_2]\pi^{\infty} + \pi_0\left[\frac{1}{1-q_2}D - \frac{1}{1-q_2}\bar{A}D\right]. \tag{21}$$

From Theorem 5.1.3 in [13] it follows directly that the matrix U is invertible. We then have that $\pi^{\infty}U^{-1} = \pi^{\infty}$, and thus,

$$P'_{\Pi}(1) = \mathbb{E}[L_2]\pi^{\infty} + \pi_0\bar{B}DU^{-1} - \pi^{\infty}KU^{-1}. \tag{22}$$

Multiplying with $\mathbf{e}^T$ it follows that

$$\pi_0\bar{B}D\mathbf{e}^T = \pi^{\infty}K\mathbf{e}^T. \tag{23}$$

By taking the second derivative of (17) with respect to z, setting $z = 1$ and multiplying with $\mathbf{e}^T$ we obtain

$$P''_{\Pi}(1)\,[I - D]\,\mathbf{e}^T + 2P'_{\Pi}(1)K\mathbf{e}^T = \pi^{\infty}\left[\frac{2q_2}{(1-q_2)^2}\bar{A}\mathbf{e}^T\right] + \pi_0\left[\frac{2q_2}{1-q_2}\bar{B}D\mathbf{e}^T\right]. \tag{24}$$

Since $P''_{\Pi}(1)\,[I - D]\,\mathbf{e}^T = 0$, we get

$$P'_{\Pi}(1)K\mathbf{e}^T = \frac{q_2}{(1-q_2)^2}\pi^{\infty}\bar{A}\mathbf{e}^T + \frac{q_2}{1-q_2}\pi_0\bar{B}D\mathbf{e}^T. \tag{25}$$

Now we combine (22) and (25) to obtain an expression for $\mathbb{E}[L_2]$:

$$\begin{aligned}&\mathbb{E}[L_2]\pi^{\infty}K\mathbf{e}^T\\&\quad= \pi^{\infty}\left[\frac{q_2}{(1-q_2)^2}\bar{A} + KU^{-1}K\right]\mathbf{e}^T + \pi_0\bar{B}\left[\frac{3q_2-1}{1-q_2}\mathbf{e}^T + DU^{-1}(\mathbf{e}^T - \mathbf{w}^T)\right].\end{aligned} \tag{26}$$

Using (23) this simplifies to

$$\mathbb{E}[L_2] = \left[\pi^\infty \left[\frac{q_2}{(1-q_2)^2}\bar{A} + KU^{-1}K\right]\mathbf{e}^T - \pi_0 \bar{B}DU^{-1}\mathbf{w}^T\right]\left[\pi^\infty K\mathbf{e}^T\right]^{-1} + \frac{2q_2}{1-q_2}. \tag{27}$$

The second and higher moments of $\mathbb{E}[L_2]$ can be computed using the same approach. Note that the mean number of priority patients, $\mathbb{E}[L_1]$, is easy to compute using π^∞, the equilibrium probability vector of the stochastic matrix D. The blocking probability for priority patients, $\mathbb{P}_B$, can be computed using that priority arrivals are blocked when the number of arrivals exceeds the number of available positions in the slot reservation window $L, \ldots, H$:

$$\mathbb{P}_B = \pi^\infty \alpha^*, \tag{28}$$

with α^* the vector defining the probabilities that the number of arrivals is larger than n as given in equation (1).

4 Results

In this section we provide numerical results. In the practical case of the care pathways it is usually the case that the load for regular patients is much higher than that for priority patients. Therefore, we start with an extensive numerical experiment where we have a high load for regular patients and a low to moderate load for priority patients. We then show the effect of in- and decreasing patient arrivals.

4.1 Standard scenario

In this standard scenario we use a high load for regular patients, $q_2 = 0.45$, and a low to moderate load for priority patients, $q_1 = 0.10$. This results in on average 0.1111 priority patient arrivals per slot and 0.8182 regular patient arrivals per slot. In Table 1 we see the effect of the size and position of the slot reservation window on the number of priority ($\mathbb{E}[L_1]$) and regular patients ($\mathbb{E}[L_2]$). Also, the blocking probability of priority patients, $\mathbb{P}_\mathbb{B}$, and the load ρ are given.

The results show that a larger slot reservation window increases the number of regular and priority patients waiting and decreases the blocking probability for priority patients. Increasing H with fixed L first increases the number of regular and priority patients; for larger H compared to L they remain constant, while the blocking probability for priority patients further decreases. Furthermore, a slot reservation window with larger L increases the number of priority patients waiting. Since the blocking probability for priority patients is very small, even for a slot reservation window of size 1, using $L = H$ results in the least possible number of regular patients waiting, while priority patients have a high probability of acceptance.

Table 1 Results for standard scenario, $q_1 = 0.10$, $q_2 = 0.45$

L	H	ρ	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{P}_B$
1	1	0.9182	0.1000	10.0000	$1e^{-2}$
1	2	0.9281	0.1209	11.4985	$1e^{-3}$
1	3	0.9292	0.1244	11.7038	$1e^{-4}$
1	5	0.9293	0.1250	11.7317	$2e^{-6}$
1	8	0.9293	0.1250	11.7321	$2e^{-9}$
1	9	0.9293	0.1250	11.7321	$2e^{-10}$
1	10	0.9293	0.1250	11.7321	$3e^{-11}$
1	11	0.9293	0.1250	11.7321	$3e^{-12}$
2	2	0.9182	0.2000	10.0000	$1e^{-2}$
2	3	0.9281	0.2308	11.4985	$1e^{-3}$
2	5	0.9293	0.2360	11.7287	$1e^{-5}$
2	8	0.9293	0.2361	11.7321	$2e^{-8}$
2	9	0.9293	0.2361	11.7321	$2e^{-9}$
2	10	0.9293	0.2361	11.7321	$2e^{-10}$
2	11	0.9293	0.2361	11.7321	$3e^{-11}$
3	3	0.9182	0.3000	10.0000	$1e^{-2}$
3	5	0.9292	0.3463	11.7038	$1e^{-4}$
3	8	0.9293	0.3472	11.7320	$2e^{-7}$
3	9	0.9293	0.3472	11.7321	$2e^{-8}$
3	10	0.9293	0.3472	11.7321	$2e^{-9}$
3	11	0.9293	0.3472	11.7321	$2e^{-10}$
5	5	0.9182	0.5000	10.0000	$1e^{-2}$
5	8	0.9293	0.5693	11.7287	$1e^{-5}$
5	9	0.9293	0.5694	11.7321	$2e^{-6}$
5	10	0.9293	0.5694	11.7321	$2e^{-7}$
5	11	0.9293	0.5696	11.7321	$2e^{-8}$
8	8	0.9182	0.8000	10.0000	$1e^{-2}$
8	9	0.9281	0.8901	11.4986	$1e^{-3}$
8	10	0.9292	0.9012	11.7038	$1e^{-4}$
8	11	0.9293	0.9026	11.7287	$1e^{-5}$
9	9	0.9182	0.9000	10.0000	$1e^{-2}$
9	10	0.9281	1.0000	11.4986	$1e^{-3}$
9	11	0.9292	1.0122	11.7038	$1e^{-4}$
10	10	0.9182	1.0000	10.0000	$1e^{-2}$
10	11	0.9281	1.1099	11.4986	$1e^{-3}$
11	11	0.9182	1.1000	10.0000	$1e^{-2}$

Table 2 Results for in- and decreasing patient arrivals

q_1	q_2	L	H	ρ	$\mathbb{E}[L_1]$	$\mathbb{E}[L_2]$	$\mathbb{P}_B$
0.10	0.45	1	1	0.9182	0.1000	10.0000	$1e^{-2}$
0.10	0.45	1	5	0.9293	0.1250	11.7317	$2e^{-6}$
0.10	0.45	1	10	0.9293	0.1250	11.7321	$3e^{-11}$
0.10	0.45	5	5	0.9182	0.5000	10.0000	$1e^{-2}$
0.10	0.45	5	10	0.9293	0.5694	11.7321	$2e^{-7}$
0.10	0.45	10	10	0.9182	1.0000	10.0000	$1e^{-2}$
0.15	0.45	1	1	0.9682	0.1500	25.6989	$2e^{-2}$
0.15	0.45	1	5	0.9946	0.2141	157.1787	$2e^{-5}$
0.15	0.45	1	10	0.9947	0.2143	157.9901	$4e^{-9}$
0.15	0.45	5	5	0.9682	0.7500	25.7010	$2e^{-2}$
0.15	0.45	5	10	0.9946	0.9201	157.8476	$4e^{-6}$
0.15	0.45	10	10	0.9682	1.5000	25.7035	$2e^{-2}$
0.10	0.40	1	1	0.7667	0.1000	2.8571	$1e^{-2}$
0.10	0.40	1	5	0.7778	0.1250	3.0416	$2e^{-6}$
0.10	0.40	1	10	0.7778	0.1250	3.0417	$3e^{-11}$
0.10	0.40	5	5	0.7667	0.5000	2.8571	$1e^{-2}$
0.10	0.40	5	10	0.7778	0.5694	3.0417	$2e^{-7}$
0.10	0.40	10	10	0.7667	1.0000	2.8571	$1e^{-2}$
0.15	0.40	1	1	0.8167	0.1500	3.6364	$2e^{-2}$
0.15	0.40	1	5	0.8431	0.2141	4.4093	$2e^{-5}$
0.15	0.40	1	10	0.8167	0.2143	4.4107	$4e^{-9}$
0.15	0.40	5	5	0.8167	0.7500	3.6364	$2e^{-2}$
0.15	0.40	5	10	0.8431	0.9201	4.4104	$4e^{-6}$
0.15	0.40	10	10	0.8167	1.5000	3.6364	$2e^{-2}$

4.2 In- and decreasing patient arrivals

In this section we compare the results from the standard scenario, where the expected number of patient arrivals per slot = 0.9293, with three scenarios where we in- and decrease the number of patient arrivals:

- Scenario 1: increase priority patient arrivals: $q_1 := 0.15$. This results in 0.1765 priority patient arrivals per slot and 0.9947 arbitrary patient arrivals per slot.
- Scenario 2: decrease regular patient arrivals: $q_2 := 0.40$. This results in 0.6667 regular patient arrivals per slot and 0.7778 arbitrary patient arrivals per slot.
- Scenario 3: increase priority patient arrivals and decrease regular patient arrivals: $q_1 := 0.15$ and $q_2 := 0.40$. This results in 0.8432 arbitrary patient arrivals per slot.

In Table 2 the results for the three scenarios are given. For convenience, the results of the standard scenario are also given for the studied instances.

Table 3 Maximum values of ρ, $\mathbb{E}[L_1]$, $\mathbb{E}[L_2]$ and $\mathbb{P}_B$

(q_1, q_2)	(0.10,0.40)	(0.10,0.45)	(0.15,0.40)	(0.15,0.45)
$\rho_{(max)}$	0.7778	0.9293	0.8431	0.9947
$\mathbb{E}[L_1]_{(max)}$	1.0000	1.0000	1.5000	1.5000
$\mathbb{E}[L_2]_{(max)}$	3.0417	11.7321	4.4107	157.9901
$\mathbb{P}_{B(max)}$	$1e^{-2}$	$1e^{-2}$	$2e^{-2}$	$2e^{-2}$

In Table 3 the maximum values of ρ, $\mathbb{E}[L_1]$, $\mathbb{E}[L_2]$, and $\mathbb{P}_B$ are given for the standard scenario and the three scenarios. We see that, in the case of a high regular patient load, even minimal priority patient arrivals have a significant influence on the number of regular patients waiting.

5 Discussion and conclusion

In this paper we have studied the $Geo^x/D/1$ queue with slot reservations that serves regular patients and priority patients. Priority patients reserve a time slot in a reservation window and are blocked when all slots in the reservation window are occupied by other priority patients. The reservation window models the advance reservation of service slots by patients on a care pathway. We have modeled the $Geo^x/D/1$ queue as a $M/G/1$-type queue and have applied a matrix-analytical approach, which simplified to a matrix-geometric solution to characterize the equilibrium probabilities. A numerical experiment has illustrated the influence of the reservation window on the number of regular and priority patients present and the blocking probabilities for priority patients. A second numerical experiment showed the effects of in- or decreasing patient arrivals. From the experiments it follows that, in case the load for the regular patients is much higher than that for priority patients, the size of the slot reservation window should be kept as small as possible, in order to keep the waiting time for regular patients minimal.

Until $L = H = 12$ it is possible to follow a generic approach, which does not use the structure of matrix D. For $L, H > 12$ numerical problems arise, since the memory of a regular computer is not sufficient to store all data entries. However, the methodology is not limited to $L, H \leq 12$, but extension for $L, H > 12$ requires a specific implementation using the structure of matrix D, which is sparse. Furthermore, the computation of π_0 is very time consuming for larger values of $L = H$. In the practical case of the care pathways it is usually the case that the load for regular patients is much higher than that for priority patients. Therefore the probability that the server is idle, when there are priority patients in the queue, is low. An approximation of π_0 is thus obtained by setting $\pi_0 = (1-\rho)\pi^{\infty}$. Note that $(1-\rho)$ is the probability that the server is idle, but then in the related queue without priorities. However, in the healthcare case we studied in this article, values for $L, H > 12$ do not have practical value. Typical appointment slots have a length of 10–15–20–30–45–60–90 min. Even in the case of appointment slots with length 10 min, a reservation window with $H = 12$ would lead to postponing priority patients for more than two hours, which is quite long. Studying longer time intervals do not have practical relevance for this specific healthcare example.

The methodology we have developed is mainly meant as a capacity planning tool, so that managers can study for instance the effect of the size of the slot reservation window. In reality, a steady state situation, especially in an environment that does not offer 24/7 service such as an outpatient clinic, will maybe not be reached. However, given the managerial insights that the methodology gives, we still feel it can be very valuable in these cases.

In future research we plan to analyze queuing networks consisting of these types of queues.

Acknowledgments The authors would like to thank Ivo Adan, Vidyadhar Kulkarni, and Maartje van der Vrugt for their valuable comments.

References

1. Allen, D., Rixson, L.: How has the impact of 'care pathway technologies' on service integration in stroke care been measured and what is the strength of the evidence to support their effectiveness in this respect? Int. J. Evid. Based Healthc. **6**(1), 78–110 (2008)
2. Allen, D.: From boundary concept to boundary object: the practice and politics of care pathway development. Soc. Sci. Med. **69**(3), 354–361 (2009)
3. Dickson, D., Ford, R.C., Laval, B.: Managing real and virtual waits in hospitality and service organizations. Corneel Hotel Restaur. Admin. Q. **46**(1), 52–68 (2005)
4. Pullman, M., Rodgers, S.: Capacity management for hospitality and tourism: a review of current approaches. Int. J. Hosp. Manag. **29**, 177–187 (2010)
5. Disney's Fastpass: Wikipedia the free encyclopedia. Retrieved from http://en.wikipedia.org/w/index.php?title=Disney%27s_Fastpass&oldid=505956546 on August 28, (2012)
6. Kostami, V., Ward, A.R.: Managing service systems with an offline waiting option and customer abandonment. Manuf. Serv. Oper. Manag. **11**(4), 644–656 (2009)
7. Neuts, M.F.: Structured stochastic matrices of $M/G/1$ type and their applications. Marcel Dekker, New York (1989)
8. Latouche, G., Ramaswami, V.: Introduction to matrix analytic methods in stochastic modeling. Siam, Philadelphia (1999)
9. Bruneel, H.: Performance of discrete-time queueing systems. Comp. Oper. Res. **20**(3), 303–320 (1993)
10. Walraevens, J., Steyaert, B., Bruneel, H.: Delay characteristics in discrete-time $GI/G/1$ queues with non-preemptive priority queueing discipline. Perform. Eval. **50**, 53–75 (2002)
11. Takahashi, Y., Hashida, O.: Delay analysis of discrete-time priority queue with structured inputs. Queueing Syst. **8**, 149–164 (1991)
12. de Vuyst, S., Wittevrongel, S., Bruneel, H.: Delay differentiation by reserving space in queue. Electr. Lett. **41**(9), 564–565 (2005)
13. Kemeny, J.G., Snell, J.L.: Finite Markov Chains, 2nd edn. Springer, New York (1976)