Oper Res Int J
DOI 10.1007/s12351-015-0211-z

ORIGINAL PAPER

An $M/E_k/1$ queues with emergency non-preemptive priority of a diagnostic resource

Jie Zhou[1,2] · Jun Li[1]

Received: 19 March 2015 / Revised: 1 November 2015 / Accepted: 3 November 2015

Abstract Each patient is assigned to a specific scanner in CT department of a large-size hospital. Emergency patients have non-preemptive priority access to service. The service time of each CT scanner follows an Erlang distribution by data analysis from this hospital. We develop an $M/E_k/1$ queueing model with emergency non-preemptive priority. Firstly, the expected waiting time of the jth phase regular patient in the waiting queue is given by Laplace transform. Using this and generating function of the steady-state of phase distribution, the expected waiting time of an arbitrary regular patient is obtained. A total cost function which includes the penalty cost for unutilized medical resources and waiting cost of regular patients is constructed. The optimal arrival rate of regular patients so as to minimize the total cost is given by Kuhn–Tucker condition. Some numerical examples which are based on real data are given.

Keywords $M/E_k/1$ · Healthcare operations management · Emergency non-preemptive · Kuhn–Tucker condition

Mathematics Subject Classification 90B22 · 68M20

✉ Jun Li
lijun@home.swjtu.edu.cn

Jie Zhou
jiezhou198208@126.com

[1] School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China

[2] Business School, Sichuan Normal University, Chengdu 610101, China

Published online: 24 November 2015

1 Introduction

Wherever queues exist, customer cutting in line may occur (Allon and Hanany 2012; He and Chavoushi 2013). Queueing system with priority clients has been widely studied. Afeche and Mendelson (2004) develop a queueing model which positions of customers in the waiting queue are determined by how much they pay to the system. Israeli Queue model is introduced in Boxma et al. (2008). Perel and Yechiali (2013) consider the Israeli Queue with priority. They provide an extensive probabilistic analysis and calculate the key performance measures of this queueing system. Queueing models with self generation of priorities are developed in the works of Wang (2004), Gómez-Corral et al. (2005), Krishnamoorthy et al. (2005, 2008, 2009). Self generation of priorities is described as follows. All the customers are homogeneous (in terms of priority) when they arrive in. Their conditions subsequently change while waiting in the queue. Then waiting customers generate into priority. Only one priority generated customer can wait at a time and a customer generating into priority at that time will have to leave the system in search for emergency service elsewhere (Gómez-Corral et al. 2005; Krishnamoorthy et al. 2005, 2009). Krishnamoorthy et al. (2008) consider a queueing system with a waiting space of capacity c (as many as the number of servers) which is provided exclusively for the priority generated customers. The matrix analytic method is employed to analyze these queueing models. Zhang and Shi (2010) consider an $M/M/1$ preemptive priority queueing model with two classes of customers. The stationary queue length distribution is given by quasi-birth-and-death (QBD) process with infinitely many phases. He and Chavoushi (2013) develop a queueing model with customer interjections, where customers are distinguished into normal and interjecting. All customers join a single queue. A normal customer joins the queue at its end, while an interjecting customer tries to join the queue as close to the head of the queue as possible. They use two parameters to describe the interjection behavior: the percentage of customers interjecting and the tolerance level of interjection by individual customers. The waiting times of normal customers and of interjecting customers are given. Moreover, Wang and Huang (1995) deal with the economic behavior of a removable server in the N policy $M/E_k/1$ queueing system with finite capacity. A cost function is given. But no analytical solution of the optimal N policy to minimize the cost function is given since the structure of the cost function is complex.

This work was originally motivated by a problem faced by Sichuan Provincial People's Hospital (SCPH), which is a large-size hospital in China. After interviewing the medical staff and manager in SCPH, CT scan is always a bottleneck for providing patients timely service. Demands for CT scan come from both regular patients, who make appointments in advanced, and emergency patients, who come without appointments and get higher priority (Green et al. 2006; Luo et al. 2012). CT scan items include plain scan, enhancement scan and other types of scan. Enhancement scan and other types of scan will be assigned to the specific scanner because of the different preparation work and operation process. Plain scan

can be carried out on each scanner. Emergency patients only need plain scan according to historical data.

One feature in this setting is the non-preemptive priority of emergency patients. Emergency patients get non-preemptive priority over regular patients at the time of their arrivals. All the emergency patients should be accommodated. The ability to match fluctuating demand will increase the reputation of the hospital (Eric and Thomas 2009). The random demand of emergency patient makes it necessary to consider queue problem with emergency non-preemptive.

The hospital currently allocates one separate scanner for the exclusive use for emergency patients. The utilization rate of the dedicated CT scanner is low. Two scanners for regular patients are overcrowded, even though regular patients and emergency patients are isolated. Therefore, there are two tasks that need to be done. First, the dedicated scanner should be shared by regular patients (Green et al. 2006; Luo et al. 2012). Meanwhile, emergency patients are given non-preemptive priority access to service (Anderson et al. 2010). In other words, after completion of the current patient, server must necessarily be attending to emergency patient if there is any emergency patient waiting in the queue. Regular patients have to accept this queue jumping. Second, the arrival rate of regular patients should be determined to reduce overcrowding. This could be done by appointment scheduling since all the patients except emergency patients should make appointments in advance.

The service time is not exponentially distributed but Erlang distributed since the various items of scanning. We first model this queueing system. The expected waiting time of an arbitrary regular patient is given. An optimal arrival rate of regular patients is obtained by Kuhn–Tucker condition to achieve tradeoff between the waiting cost of regular patients and the penalty cost for unutilized medical resources. The penalty cost for unutilized medical resources in this paper is similar to that described in Gupta and Wang (2008).

The rest of this paper is organized as follows. Section 2: model description. In Sect. 3, we give performance measures of the queueing system. In Sect. 4, numerical study is conducted on basis of real data in SCPH. A conclusion is given in Sect. 5 with a discussion of our results and potential direction for future research.

2 Model description

A single queue is developed at each CT scanner since each patient will be assigned to a specific scanner. This work is done by triage nurses who are in charge of the main service desk. So we consider a single server queue with two types of independent arrivals, regular patients and emergency patients. Regular patients join the queue at its end. The new emergency arrival cuts in line in front of the first regular patient in the waiting queue. Namely, emergency patients in the waiting queue are served in a first-come, first-served (FCFS) manner.

There are total c minutes in one service session. We assume that the arrivals of both emergency patients and regular patients are homogeneous Poisson process with rate λ_1 and λ_2 per minute, respectively. $\lambda = \lambda_1 + \lambda_2$. The service time is the time from the patient going in to come out from CT room. The service time follows an

Erlang distribution with mean $\frac{1}{\mu}$ and stage parameter k from data analysis of SCPH. The Erlang type k distribution is made up of k independent and identical exponential stages, each with mean $\frac{1}{k\mu}$. A patient goes into the first stage of service (say stage k), then progresses through the remaining stages and must complete the last stage (say stage 1) before the next patient enters the first stage k.

The queueing system could be formulated as a continuous time Markov chain. The number of phases increases k if there is a new arrival. The number of phases decreases 1 if the patient in service finishes one stage. Let $N(t)$ denote the number of phases in the system at time t and define p_j $(j \geqslant 0)$ by $p_j = \lim_{t\to\infty} P\{N(t) = j\}$, where we assume the preceding limit exists. The steady-state of phase distribution exists if $\rho = \frac{\lambda}{\mu} < 1$. Meng (1989) gives the steady-state of phase distribution of $M/E_k/1$ as follows.

$$\begin{cases} p_0 = 1 - \rho, \\ p_j = (1-\rho)\sum_{i=1}^{k} A_i \left(\dfrac{1}{s_i}\right)^j, \quad j = 1, 2, \ldots, \end{cases} \tag{1}$$

where $A_i = \prod_{l=1, l\neq i}^{k} \frac{1}{1-\frac{s_i}{s_l}}$. s_i $(1 \leqslant i \leqslant k)$ is the different real root of polynomial equation $\lambda(s + s^2 + \cdots + s^k) - k\mu = 0$ of degree k and $|s_i| > 1$. The generating function $\Phi(s)$ of the phase distribution (1) is

$$\Phi(s) = \frac{k\mu(1-\rho)(1-s)}{k\mu + \lambda s^{k+1} - (\lambda + k\mu)s}. \tag{2}$$

3 Performance measures of the model

Emergency patient in the waiting queue will not be jumped since emergency patients are served in a FCFS manner. It means that the position of each emergency patient in the waiting queue is monotone decreasing. However, this is not the case for regular patients. We focus on the waiting time of regular patients. Let the position of a new arrival regular patient in the waiting queue be the jth phase $(j \geqslant 1)$. $j = 0$ means waiting has ended. Let W_j $(j \geqslant 1)$ denote the waiting time of the jth phase regular patient in the waiting queue. Obviously, $W_0 = 0$. The first thing is to give the waiting time of the jth phase regular patient in the waiting queue at time t by Laplace transform. Then we give the expected waiting time of an arbitrary regular patient based on the Poisson arrivals see time averages principle (PASTA) and the fact that the steady-state of phase distribution is not influenced by emergency cutting in line. At last, we obtain the optimal arrival rate of regular patients so as to minimize the total cost function in Sect. 3.4.

3.1 Waiting time of the jth phase regular patient in the waiting queue

For the jth phase regular patient in the waiting queue at time t, his/her position will be changed if one of the next two random events occurs.

1. Remaining arriving time T_1 of the next emergency patient is longer than departure time T_2 of the current stage in service, then the position of this regular patient will goes from j to $j-1$.
2. Remaining arriving time T_1 of the next emergency patient is shorter than departure time T_2 of the current stage in service, then the position of this regular patient will goes from j to $j+k$.

The cumulative distribution of T_1 and T_2 are expressed by $P(T_1 \leqslant t) = 1 - e^{-\lambda_1 t}$ and $P(T_2 \leqslant t) = 1 - e^{-k\mu t}$, respectively. Obviously, $P(T_1 \leqslant T_2) = \frac{\lambda_1}{\lambda_1 + k\mu}$. So T_1 and T_2 are independent and exponentially distributed random variables with parameters λ_1 and $k\mu$, respectively. Let Z_j be the position's changed time of the jth phase regular patient in the waiting queue. Then $Z_j = \min(T_1, T_2)$. The density function of Z_j is given by $f_{Z_j}(t) = (\lambda_1 + k\mu)e^{-(\lambda_1 + k\mu)t}$.

Let $w_j^*(s) = E[e^{-sW_j}] = \int_0^\infty e^{-st} f_{W_j}(t)dt$ be the Laplace transform of W_j, where $f_{W_j}(t)$ is the density function of W_j. Then we give Theorem 1 as follows.

Theorem 1 *For the jth* $(j \geqslant 1)$ *phase regular patient in the waiting queue, the Laplace transform* $w_j^*(s)$ *of waiting time* W_j *satisfies*

$$w_j^*(s) = 1 - \frac{s}{k\mu}\sum_{m=1}^{j}\left\{\sum_{l=1}^{\infty}\rho_1^l\left[\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\cdots\sum_{i_l=1}^{k} w^*_{m+i_1+i_2+\cdots+i_l}(s)\right]\right\} - \frac{s}{k\mu}\sum_{m=1}^{j} w_m^*(s), \tag{3}$$

where $\rho_1 = \frac{\lambda_1}{k\mu}$.

Proof From the above discussion and property of conditional expectation, $w_j^*(s)$ could be written as follows.

$$\begin{aligned} w_j^*(s) &= E[e^{-sW_j}|T_1 < T_2]P(T_1 < T_2) + E[e^{-sW_j}|T_1 > T_2]P(T_1 > T_2) \\ &= E[e^{-s(Z_j + W_{j+k})}]P(T_1 < T_2) + E[e^{-s(Z_j + W_{j-1})}]P(T_1 > T_2) \\ &= E[e^{-sZ_j}]\left[\frac{\lambda_1}{\lambda_1 + k\mu}\cdot w^*_{j+k}(s) + \frac{k\mu}{\lambda_1 + k\mu}\cdot w^*_{j-1}(s)\right] \\ &= \frac{\lambda_1 + k\mu}{\lambda_1 + k\mu + s}\left[\frac{\lambda_1}{\lambda_1 + k\mu}\cdot w^*_{j+k}(s) + \frac{k\mu}{\lambda_1 + k\mu}\cdot w^*_{j-1}(s)\right] \\ &= \frac{\lambda_1}{\lambda_1 + k\mu + s}\cdot w^*_{j+k}(s) + \frac{k\mu}{\lambda_1 + k\mu + s}\cdot w^*_{j-1}(s). \end{aligned} \tag{4}$$

Multiplying both sides of (4) by $\frac{\lambda_1 + k\mu + s}{k\mu}$, (4) becomes

$$\begin{aligned} w_j^*(s) - w_{j-1}^*(s) &= \rho_1 \left[w_{j+k}^*(s) - w_j^*(s) \right] - \frac{s}{k\mu} w_j^*(s) \\ &= \rho_1 \left\{ \sum_{i=1}^{k} \left[w_{j+i}^*(s) - w_{j+i-1}^*(s) \right] \right\} - \frac{s}{k\mu} w_j^*(s). \end{aligned} \tag{5}$$

Iterating the above equality by itself, we have

$$\begin{aligned} & w_j^*(s) - w_{j-1}^*(s) \\ &= \rho_1 \sum_{i_1=1}^{k} \left[w_{j+i_1}^*(s) - w_{j+i_1-1}^*(s) \right] - \frac{s}{k\mu} w_j^*(s) \\ &= \rho_1 \sum_{i_1=1}^{k} \left\{ \rho_1 \sum_{i_2=1}^{k} \left[w_{j+i_1+i_2}^*(s) - w_{j+i_1+i_2-1}^*(s) \right] - \frac{s}{k\mu} w_{j+i_1}^*(s) \right\} - \frac{s}{k\mu} w_j^*(s) \\ &= \rho_1^2 \sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \left[w_{j+i_1+i_2}^*(s) - w_{j+i_1+i_2-1}^*(s) \right] - \rho_1 \sum_{i_1=1}^{k} \frac{s}{k\mu} w_{j+i_1}^*(s) - \frac{s}{k\mu} w_j^*(s) \\ &= \rho_1^n \sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \cdots \sum_{i_n=1}^{k} \left[w_{j+i_1+i_2+\cdots+i_n}^*(s) - w_{j+i_1+i_2+\cdots+i_n-1}^*(s) \right] \\ &\quad - \frac{s}{k\mu} \sum_{l=1}^{n-1} \rho_1^l \left[\sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \cdots \sum_{i_l=1}^{k} w_{j+i_1+i_2+\cdots+i_l}^*(s) \right] - \frac{s}{k\mu} w_j^*(s) \\ &= \rho_1^n \sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \cdots \sum_{i_{n-1}=1}^{k} \left[w_{j+i_1+i_2+\cdots+i_{n-1}+k}^*(s) - w_{j+i_1+i_2+\cdots+i_{n-1}}^*(s) \right] \\ &\quad - \frac{s}{k\mu} \sum_{l=1}^{n-1} \rho_1^l \left[\sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \cdots \sum_{i_l=1}^{k} w_{j+i_1+i_2+\cdots+i_l}^*(s) \right] - \frac{s}{k\mu} w_j^*(s). \end{aligned} \tag{6}$$

The last equality of (6) holds since

$$\sum_{i_n=1}^{k} \left[w_{j+i_1+i_2+\cdots+i_n}^*(s) - w_{j+i_1+i_2+\cdots+i_n-1}^*(s) \right] = w_{j+i_1+i_2+\cdots+i_{n-1}+k}^*(s) - w_{j+i_1+i_2+\cdots+i_{n-1}}^*(s).$$

The Laplace transform $w_j^*(s)$ is always between 0 and 1 since $W_j \geqslant 0$ and $s \geqslant 0$ (Ross 2007). That is, $-1 \leqslant w_{j_1}^*(s) - w_{j_2}^*(s) \leqslant 1$ $(\forall\ j_1,\ j_2)$. Let $M = \max\{w_{j_1}^*(s) - w_{j_2}^*(s)\}$. Then $w_{j_1}^*(s) - w_{j_2}^*(s) \leqslant M$ holds, $\forall\ j_1, j_2$. There are at most $k^{n-1}M$ from

$$\sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \cdots \sum_{i_{n-1}=1}^{k} \left[w_{j+i_1+i_2+\cdots+i_{n-1}+k}^*(s) - w_{j+i_1+i_2+\cdots+i_{n-1}}^*(s) \right].$$

So

$$\rho_1^n \sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\cdots\sum_{i_{n-1}=1}^{k}\left[w^*_{j+i_1+i_2+\cdots+i_{n-1}+k}(s) - w^*_{j+i_1+i_2+\cdots+i_{n-1}}(s)\right]$$
$$\leqslant \rho_1^n k^{n-1} M = \rho^n \cdot \frac{1}{k^n}\cdot k^{n-1}M = \rho^n \cdot \frac{1}{k}\cdot M.$$

$\lim_{n\to\infty}\rho^n \cdot \frac{M}{k} = 0$ since $\rho < 1$. As $n \to \infty$, (6) becomes

$$w^*_j(s) - w^*_{j-1}(s) = -\frac{s}{k\mu}\sum_{l=1}^{\infty}\rho_1^l\left[\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\cdots\sum_{i_l=1}^{k} w^*_{j+i_1+i_2+\cdots+i_l}(s)\right] - \frac{s}{k\mu}w^*_j(s). \quad (7)$$

Taking summation on both sides of the above equality from $j = 1$ to $j = j$ and using the fact that $w^*_0(s) = 1$, we have

$$w^*_j(s) = 1 - \frac{s}{k\mu}\sum_{m=1}^{j}\left\{\sum_{l=1}^{\infty}\rho_1^l\left[\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\cdots\sum_{i_l=1}^{k} w^*_{m+i_1+i_2+\cdots+i_l}(s)\right]\right\} - \frac{s}{k\mu}\sum_{m=1}^{j} w^*_m(s).$$

□

Theorem 2 *The expected waiting time of the jth* $(j \geqslant 1)$ *phase regular patient in the waiting queue* $E[W_j]$ *satisfies*

$$E[W_j] = \frac{j}{k(\mu - \lambda_1)}. \quad (8)$$

Proof By (3) and $E[W_j] = -(w^*_j(s))' \mid_{s=0}$, we could obtain $E[W_j]$.

$$(w^*_j(s))' \mid_{s=0} = -\frac{1}{k\mu}\sum_{m=1}^{j}\left\{\sum_{l=1}^{\infty}\rho_1^l\left[\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\cdots\sum_{i_l=1}^{k} w^*_{m+i_1+i_2+\cdots+i_l}(0)\right]\right\} - \frac{1}{k\mu}\sum_{m=1}^{j} w^*_m(0)$$
$$= -\frac{1}{k\mu}\sum_{m=1}^{j}\left\{\sum_{l=1}^{\infty}\rho_1^l \cdot k^l\right\} - \frac{j}{k\mu} = -\frac{1}{k\mu}\sum_{m=1}^{j}\left\{\sum_{l=1}^{\infty}\left(\frac{\lambda_1}{\mu}\right)^l\right\} - \frac{j}{k\mu}$$
$$= -\frac{j}{k\mu}\cdot\frac{\lambda_1}{\mu-\lambda_1} - \frac{j}{k\mu} = -\frac{j}{k\mu}\cdot\frac{\mu}{\mu-\lambda_1}.$$

Then $E[W_j] = \frac{j}{k\mu}\cdot\frac{\mu}{\mu-\lambda_1} = \frac{j}{k(\mu-\lambda_1)}$. □

Theorem 2 shows that the average waiting time of the jth phase regular patient in the waiting queue is closely related to the arrival rate of emergency patients λ_1 and independent of the arrival rate of regular patients λ_2 since an new arrival regular patient joins the queue at its end. So those regular patients who have been in the waiting queue will not be influenced by the new arrival regular patients.

3.2 Expected waiting time of an arbitrary regular patient

An new arrival regular patient will occupy the $(x+1)$th to $(x+k)$th phases in the system if he/she sees x phases in the system at time t. His/her position in the waiting queue is the xth phase. Let W_R be the waiting time of an arbitrary regular patient. $w_R^*(s)$ is the Laplace transform of W_R. Theorem 3 could be obtained by PASTA principle and steady-state of phase distribution (1)–(2).

Theorem 3 *The expected waiting time of an arbitrary regular patient* $E[W_R]$ *satisfies*

$$E[W_R] = \frac{1}{k(\mu-\lambda_1)}\left[\frac{\lambda(k+1)}{2(\mu-\lambda)}\right]. \tag{9}$$

Proof The cumulative distribution function $W_R(t)$ of W_R satisfies $W_R(t) = p_0 + \sum_{j=1}^{\infty} p_j \Pr(0 < W_j \leqslant t)$. Then $w_R^*(s) = \sum_{j=1}^{\infty} p_j w_j^*(s)$. According to (2) and (8), we have

$$\begin{aligned} E[W_R] &= -(w_R^*(s))' \mid_{s=0} = \sum_{j=1}^{\infty} p_j E[W_j] = \sum_{j=1}^{\infty} p_j \frac{j}{k(\mu-\lambda_1)} = \frac{1}{k(\mu-\lambda_1)} \sum_{j=1}^{\infty} p_j \cdot j \\ &= \frac{1}{k(\mu-\lambda_1)}\left[\frac{d\Phi(s)}{ds} \mid_{s=1}\right]. \end{aligned} \tag{10}$$

Using L' Hospital Law twice on $\frac{d\Phi(s)}{ds} \mid_{s=1}$, we have

$$E[W_R] = \frac{1}{k(\mu-\lambda_1)}\left[\frac{\mu(1-\rho)\lambda(k+1)}{2(\lambda-\mu)^2}\right] = \frac{1}{k(\mu-\lambda_1)}\left[\frac{\lambda(k+1)}{2(\mu-\lambda)}\right].$$

□

3.3 Expected idle time of the service system

p_0 is the long-run probability that there will be exactly 0 patient in the system. It usually turns out that p_0 equals to the long-run proportion of time that the system contains exactly 0 patient. So the expected idle time $E(I)$ of medical resources is $p_0 \cdot c$.

3.4 Optimal arrival rate of regular patients

All the emergency patients should be accommodated. Emergency arrival rate is independent of the arrival rate of regular patients. CT department couldn't further decrease the waiting time of emergency patients except buying new scanners since emergency patients get non-preemptive priority over regular patients. However, adding new devices is impossible in the short-term because of the limitation of fund and space. Emergency arrival rate is not influenced by schedulers. How to reduce overcrowded in CT department? One way is to modulate the arrival rate of regular patients. This could be done by appointment scheduling in advance.

Intuitively, CT department wants to decrease regular patients' waiting time and idle time of medical resources. The expected waiting time of regular patients $E(W)$ is $\lambda_2 c \cdot E(W_R)$. The expected idle time $E(I)$ of medical resources is $p_0 \cdot c = (1 - \frac{\lambda_1+\lambda_2}{\mu}) \cdot c$. Obviously, $E(W)$ increases as λ_2 increases. $E(I)$ decreases as λ_2 increases. So a proper arrival rate of regular patients should be determined to reduce overcrowding of regular patients, as well as to avoid too much idle time of medical resources.

In what follows, a nonlinear programming is constructed to obtain the optimal arrival rate of regular patients.

$$\begin{cases} \min G(\lambda_2) = \lambda_2 c \cdot \dfrac{1}{k(\mu-\lambda_1)} \left[\dfrac{(\lambda_1+\lambda_2)(k+1)}{2(\mu-\lambda_1-\lambda_2)} \right] + \theta c \cdot \left(1 - \dfrac{\lambda_1+\lambda_2}{\mu}\right) \\ \qquad 0 \leqslant \lambda_2 < \mu - \lambda_1. \end{cases} \tag{11}$$

The objective function is to minimize the total cost function $G(\lambda_2)$ which includes penalty cost for unutilized medical resources and waiting cost of regular patients, where the parameter θ $(\theta > 0)$ is the unit penalty cost for unutilized medical resources supposing that the unit waiting cost of a regular patient is 1. The constraint condition is obtained from $\rho = \frac{\lambda}{\mu} < 1$. Waiting time of emergency patients is not considered in (11) since emergency patients have been given non-preemptive priority. This is an appropriate way to decrease their waiting time.

$\mu - \lambda_1$ is not the minimum point of $G(\lambda_2)$ since $G(\lambda_2)$ is a continuous function with respect to λ_2 on $[0, \mu - \lambda_1)$ and $G(\lambda_2) \to \infty$ as $\lambda_2 \to \mu - \lambda_1$.

Proposition 1 *There exists an unique optimal solution of the nonlinear programming* (11) *on* $[0, \mu - \lambda_1)$.

Proof Taking second derivative of $G(\lambda_2)$ with respect to λ_2, the following inequality

$$\frac{d^2 G(\lambda_2)}{d\lambda_2^2} = \frac{c(k+1)}{k(\mu-\lambda_1)} \cdot \frac{(\mu-\lambda_1-\lambda_2)^2 + [(\lambda_1+2\lambda_2)(\mu-\lambda_1-\lambda_2) + \lambda_2(\lambda_1+\lambda_2)]}{(\mu-\lambda_1-\lambda_2)^3} > 0$$

holds for $0 \leqslant \lambda_2 < \mu - \lambda_1$. Nonlinear programming (11) is a strictly convex programming since the total cost function $G(\lambda_2)$ is a strictly convex function and the constraint conditions are linear functions. From continuity and strictly convexity with respect to λ_2, there exists an unique optimal solution of the programming (11) on $[0, \mu - \lambda_1)$. □

Obviously, $G(\lambda_2)$ is closely related to the parameter θ. θ is more subjective, compared with the parameters k, μ and λ_1. To compensate for the unreliability of θ as well as to represent a spectrum of possible actual operating situations across hospitals, we will consider the total cost function with different θ. The objective function is denoted by $G(\lambda_2, \theta)$ if necessary.

Theorem 4 *Let $\lambda_2^*(\theta)$ be the optimal regular patient's arrival rate for $\forall\ \theta > 0$, then*

(1) $\lambda_2^*(\theta) = 0$, *if* $\mu\lambda_1(k+1) - 2\theta k(\mu - \lambda_1)^2 \geqslant 0$ *holds*;

(2) $\lambda_2^*(\theta)$ *should satisfy* $0 < \lambda_2^*(\theta) < \mu - \lambda_1$ *and*

$$\mu(k+1)[(\lambda_1 + 2\lambda_2^*(\theta))\mu - (\lambda_1 + \lambda_2^*(\theta))^2] - 2k\theta(\mu - \lambda_1)(\mu - \lambda_1 - \lambda_2^*(\theta))^2 = 0, \tag{12}$$

if $\mu\lambda_1(k+1) - 2\theta k(\mu - \lambda_1)^2 < 0$ *holds.*

Proof There is an unique Kuhn–Tucker point of (11) on $[0, \mu - \lambda_1)$ since (11) is a strictly convex programming with respect to λ_2. Let the Kuhn–Tucker point of the programming (11) be $\lambda_2^*(\theta)$. The Kuhn–Tucker condition is listed below.

$$\begin{cases} \dfrac{dG(\lambda_2,\theta)}{d\lambda_2}\,|_{\lambda_2^*(\theta)}\ -\gamma_1^* + \gamma_2^* = 0, \\ \gamma_1^*\lambda_2^*(\theta) = 0, \\ \gamma_2^*(\mu - \lambda_1 - \lambda_2^*(\theta)) = 0, \\ \gamma_1^*, \gamma_2^* \geqslant 0, \end{cases} \tag{13}$$

where γ_1^*, γ_2^* are the generalized Lagrange multipliers and

$$\frac{dG(\lambda_2,\theta)}{d\lambda_2} = \frac{c(k+1)}{2k(\mu - \lambda_1)} \cdot \frac{(\lambda_1 + 2\lambda_2)\mu - (\lambda_1 + \lambda_2)^2}{(\mu - \lambda_1 - \lambda_2)^2} - \frac{c\theta}{\mu}.$$

We look at four separate cases to solving the equation set (13).

- If $\gamma_1^* \neq 0$, $\gamma_2^* \neq 0$, there is no solution.
- If $\gamma_1^* = 0$, $\gamma_2^* \neq 0$, then $\lambda_2^*(\theta) = \mu - \lambda_1$. It is impossible for our queueing system.
- If $\gamma_1^* \neq 0$, $\gamma_2^* = 0$, then $\lambda_2^*(\theta) = 0$. $\frac{dG(\lambda_2,\theta)}{d\lambda_2}\,|_{\lambda_2^*(\theta)}\ -\gamma_1^* + \gamma_2^* = \frac{c\lambda_1(k+1)}{2k(\mu-\lambda_1)^2} - \frac{c\theta}{\mu} - \gamma_1^* = 0.$

    1. If $\mu\lambda_1(k+1) - 2\theta k(\mu - \lambda_1)^2 > 0$, namely, $\gamma_1^* > 0$. $\lambda_2^*(\theta) = 0$ is the Kuhn–Tucker point. $\lambda_2^*(\theta) = 0$ is the unique optimal solution since (11) is a strictly convex programming.
    2. If $\mu\lambda_1(k+1) - 2\theta k(\mu - \lambda_1)^2 \leqslant 0$, namely, $\gamma_1^* < 0$ ($\gamma_1^* \neq 0$). $\lambda_2^*(\theta) = 0$ is not the Kuhn–Tucker point.

- If $\gamma_1^* = 0$, $\gamma_2^* = 0$, then $\frac{dG(\lambda_2,\theta)}{d\lambda_2}\,|_{\lambda_2^*(\theta)}\ -\gamma_1^* + \gamma_2^* = \frac{c(k+1)}{2k(\mu-\lambda_1)} \cdot \frac{(\lambda_1+2\lambda_2^*(\theta))\mu-(\lambda_1+\lambda_2^*(\theta))^2}{(\mu-\lambda_1-\lambda_2^*(\theta))^2}$ $-\frac{c\theta}{\mu} = 0.$

  If $\mu\lambda_1(k+1) - 2\theta k(\mu - \lambda_1)^2 = 0$, namely, $\lambda_2^*(\theta) = 0$ is the Kuhn–Tucker point. $\lambda_2^*(\theta) = 0$ is the unique optimal solution since (11) is a strictly convex programming. Otherwise, there is an unique Kuhn–Tucker point $\lambda_2^*(\theta)$ which satisfies $\mu(k+1)[(\lambda_1 + 2\lambda_2^*(\theta))\mu - (\lambda_1 + \lambda_2^*(\theta))^2] - 2k\theta(\mu - \lambda_1)(\mu - \lambda_1 - \lambda_2^*(\theta))^2 = 0$. The Kuhn–Tucker point $\lambda_2^*(\theta)$ also should satisfy the constraint condition.

The optimal arrival rate $\lambda_2^*(\theta)$ of regular patients is obtained. □

It can be seen from Theorem 4 that $\lambda_2^*(\theta) = 0$ if $\theta \leqslant \frac{\mu\lambda_1(k+1)}{2k(\mu-\lambda_1)^2}$. Namely, it is better not to schedule any regular patient if the unit penalty cost for unutilized medical resources is below a specified level. We will show this property of the total cost function with respect to θ and λ_2 in the next section.

4 Numerical analysis

Most of the regular patients in CT department make appointments in advance. The scheduler will assign each patient to a specific scanner and a specific time on the work time (the work time is 8:00–12:00, 14:00–18:00 except holidays). Scheduled patient arrives according to the appointment time. It is always overcrowded on the workday. The scheduler could modulate regular patient flow by appointment. We will give the proper arrival rate of regular patients for different scenarios by theoretical results in Sect. 3. Emergency arrival rate is objective and all the emergency patients should be accommodated. So the first thing we should do is to find out the emergency arrival rate λ_1.

We collect historical data from April 2011 to March 2012 in SCPH. The most busy month is March 2012. Emergency patients in the morning session (8:00–12:00) in March 2012 expect holidays (9 days) are selected. There are 840 emergency patients in this period. The emergency arrival rate is $\frac{840}{22\cdot 4\cdot 60} = 0.15909$ per minute. So we take $\lambda_1 = \frac{0.15909}{3} = 0.05303$ in each scanner since emergency patients are equally assigned to each scanner and they can be carried out on each device.

Next, we find out the service time distribution. On the work time, No. 1 scanner (in Room 1) carries out enhancement scanning and plain scanning. No. 2 scanner (in Room 2) carries out other types of scanning and plain scanning. No. 3 scanner (in Room 3) carries out emergency scanning (plain scanning). The entry time of each patient going into the room will be recorded by computer. But the departure time is not recorded. So we select the most busy month's most busy period (9:00–10:00, March 2012) to find out the service time because the scanner is busy in serving in this period. The service time is between two entry time. Kolmogorov–Smirnov (KS) test statistic and Anderson–Darling (AD) test statistic are used to conduct the test of goodness of fit. The significance level was set to 0.05. The service time distribution is presented in Table 1.

Reasons for the difference between two CT scanners' service time are listed below. The number of scanning parts per one patient and scanning items have an impact on the service time. The more parts per one patient operates, the more time he/she needs. Enhancement scanning and other types of scanning need more time than plain scanning.

There are 4 h in one service session ($c = 240\,\text{min}$). We first give the expected waiting time of an arbitrary regular patient $E[W_R]$ for Room 1 and Room 2, respectively. They are showed by Fig. 1.

Both of the expected waiting time in Room 1 and 2 increase as λ_2 increases. The expected waiting time of a regular patient will increase significantly if $\lambda_2 \to \mu - \lambda_1$. The difference between two rooms is narrow when $\lambda_2 < 0.08$.

Theorem 4 tells us that $\lambda_2^*(\theta) = 0$, if $\theta \leqslant \frac{\mu\lambda_1(k+1)}{2k(\mu-\lambda_1)^2}$. We show this aspect by numerical calculation according to the parameters given above. For Room 1, $\lambda_2^*(\theta) = 0$ if $\theta \leqslant 0.4115$. For Room 2, $\lambda_2^*(\theta) = 0$ if $\theta \leqslant 0.5548$. We list the results of the total cost in Tables 2 and 3. Datum in Tables 2 and 3 show that $\lambda_2^*(\theta)$ is consistent with the theoretical result in Theorem 4 as well as the convexity of total cost function with respect to λ_2. Moreover, $\lambda_2^*(\theta) \to 0$ if $\theta \to (\frac{\mu\lambda_1(k+1)}{2k(\mu-\lambda_1)^2})^+$ since the total cost function $G(\lambda_2)$ is continuous with respect to λ_2. For example, for Room 1, if $\theta = 0.5 > 0.4115$, $\lambda_2^*(0.5) = 0.004 \to 0$.

Next, we calculate $G(\lambda_2, \theta)$ and a rate $\frac{E(W)}{G(\lambda_2,\theta)}$. In what follows, $\theta \geqslant 1$ always holds. Reasons are listed below. Unutilized of scarce capacity is specially unreasonable. CT department would rather make a regular patient wait 1 min than make medical resources idle 1 min. That is, managers will try to avoid resources' idling under a reasonable waiting time of regular patients. So the value of θ goes from 1 to 20. The results are given in Figs. 2 and 3.

Both of the two Rooms show the same tendency on total cost. If $\lambda_2 \to \mu - \lambda_1$, the total cost $G(\lambda_2, \theta)$ will increase significantly because of the sharp increase of regular patients' waiting time. $G(\lambda_2, \theta)$ monotonically increases as θ increases. $G(\lambda_2, \theta)$ decreases first and then increases as λ_2 increases since $G(\lambda_2, \theta)$ is a strictly convex function with respect to λ_2. Convexity is apparent with a large θ. Reasons can be seen from Fig. 3. The major factor in total cost is penalty cost for unutilized medical resources if λ_2 is small. So the total cost is greater with a larger θ. In the initial stage of growth process in λ_2, the idle time of medical resources decreases while the waiting time of regular patients will not significantly increase. So the total cost decreases more significantly with a large θ. The waiting cost of regular patients becomes the major factor in total cost as λ_2 increases. Combining with the results in Tables 2, 3 and Fig. 2, $G(\lambda_2, \theta)$ is convex with respect to λ_2 for $\theta > 0$.

At last, the optimal arrival rate $\lambda_2^*(\theta)$ ($1 \leqslant \theta \leqslant 20$) for two Rooms are obtained by Theorem 4. Results are given in Fig. 4. The optimal $\lambda_2^*(\theta)$ increases as θ increases. Namely, if the hospital is concerned about the penalty cost of unutilized

Table 1 The service time distribution of two rooms

Room no.	Distribution (min.)	Sample size	*P* value of KS	*P* value of AD
1	Erlang (4, 5.88)	134	0.249	0.255
2	Erlang (1, 5.36)	115	0.282	0.562

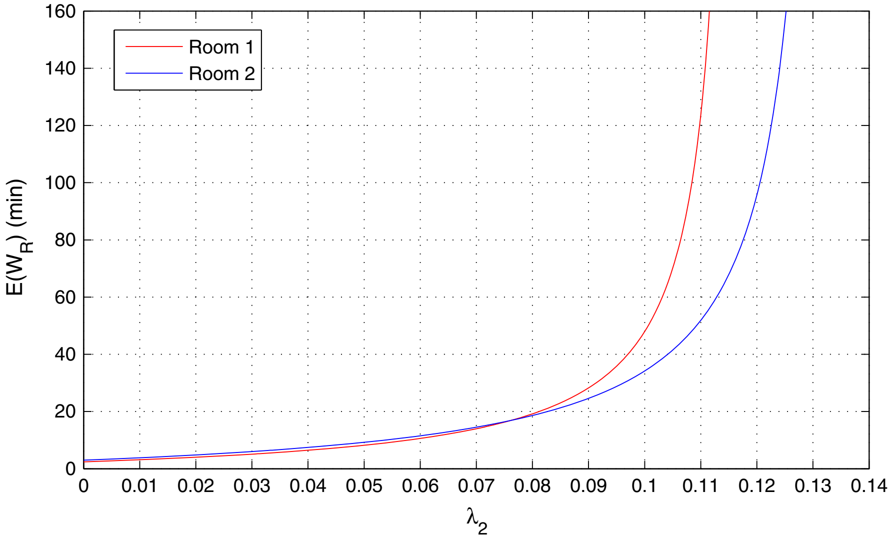

Fig. 1 The expected waiting time of a regular patient

Table 2 The trend of the total cost for Room 1

θ	λ_2								
	0.000	0.004	0.008	0.011	0.014	0.017	0.020	0.060	0.100
0.0	0.0	2.6	5.7	8.5	11.7	15.3	19.3	152.4	1151.1
0.1	16.5	18.5	21.1	23.5	26.2	29.4	33.0	160.4	1153.5
0.2	33.0	34.5	36.5	38.4	40.8	43.5	46.7	168.5	1155.9
0.3	49.5	50.4	51.9	53.4	55.3	57.6	60.4	176.5	1158.3
0.4	66.1	66.4	67.3	68.4	69.8	71.7	74.1	184.6	1160.7
0.5	82.6	82.3	82.7	83.3	84.4	85.8	87.8	192.6	1163.1
0.6	99.1	98.3	98.1	98.3	98.9	100.0	101.5	200.7	1165.5
0.7	115.6	114.2	113.5	113.3	113.5	114.1	115.1	208.7	1168.0
0.8	132.1	130.2	128.8	128.2	128.0	128.2	128.8	216.8	1170.4
0.9	148.6	146.2	144.2	143.2	142.5	142.3	142.5	224.8	1172.8
1.0	165.2	162.1	159.6	158.2	157.1	156.4	156.2	232.9	1175.2

medical resources more than the waiting time of regular patients, the scheduler should accept more scheduled patients in advance to increase the regular arrival rate so as to avoid idling of medical resources. It can be seen from Figs. 2 and 4 that $\lambda_2^*(\theta)$ achieves tradeoff between idle time of medical resources and waiting time of regular patients.

Table 3 The trend of the total cost for Room 2

θ	λ_2								
	0.000	0.002	0.005	0.008	0.010	0.013	0.020	0.060	0.100
0.0	0.0	1.5	4.1	7.0	9.2	12.8	23.1	165.7	820.1
0.1	17.2	18.4	20.6	23.1	25.1	28.3	37.7	175.2	824.4
0.2	34.4	35.3	37.1	39.3	41.0	43.8	52.3	184.7	828.7
0.3	51.5	52.3	53.7	55.4	56.8	59.3	66.9	194.1	833.0
0.4	68.7	69.2	70.2	71.6	72.7	74.8	81.5	203.6	837.3
0.5	85.9	86.1	86.7	87.7	88.6	90.3	96.1	213.0	841.7
0.6	103.1	103.0	103.3	103.9	104.5	105.8	110.8	222.5	846.0
0.7	120.2	120.0	119.8	120.0	120.4	121.3	125.4	232.0	850.3
0.8	137.4	136.9	136.3	136.2	136.3	136.8	140.0	241.4	854.6
0.9	154.6	153.8	152.9	152.3	152.2	152.4	154.6	250.9	858.9
1.0	171.8	170.7	169.4	168.5	168.1	167.9	169.2	260.3	863.2

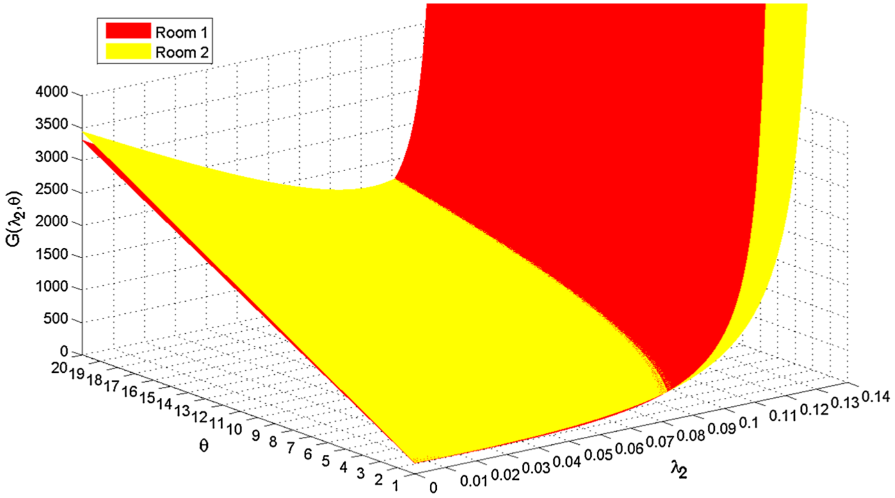

Fig. 2 The total cost of two rooms

5 Conclusion

In this paper, we first describe a problem faced by CT department of SCPH. Two ways to solve this problem are given. Let the dedicated scanner be shared by regular patients and give emergency patients non-preemptive priority for access to service. Then we analyze the queueing system to give the waiting time of regular patients. The optimal arrival rate of regular patients which achieves tradeoff between the

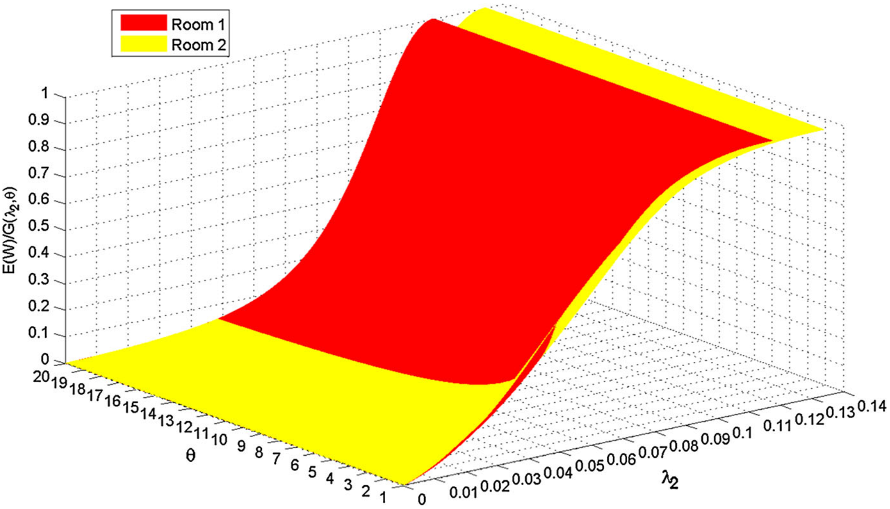

Fig. 3 The ratio of waiting cost to total cost of two rooms

waiting cost of regular patients and the penalty cost for unutilized medical resources is obtained.

The optimal arrival rate of regular patients is relatively small in this work. One way to increase the number of regular patients been served in one session is overtime working. In the future research, a queueing model with finite capacity could be constructed to achieve tradeoff among overtime cost of medical resources, waiting cost of regular patient and idling cost of medical resources.

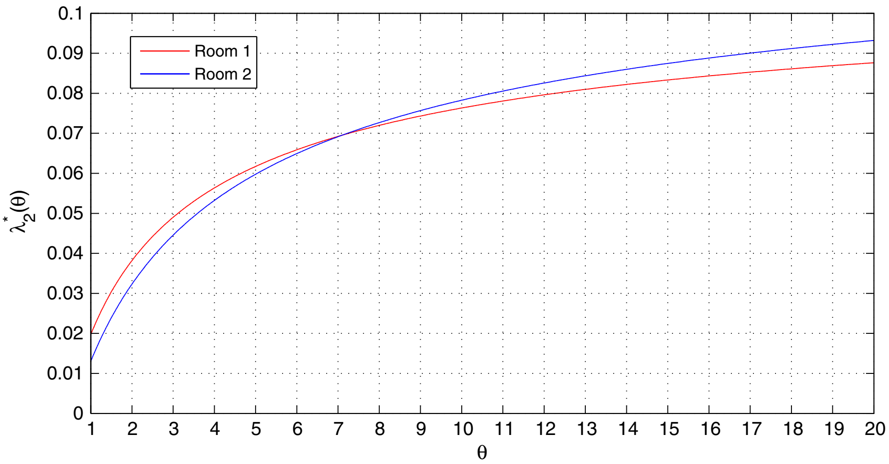

Fig. 4 The optimal arrival rate of regular patients

Acknowledgments This research is partially supported by National Natural Science Foundation of P.R. China (71490722, 71490725, 71301111), Research Project of Education Department of Sichuan Province (13ZB0140), National Social Science Foundation of P.R. China (15BGL198). We would like to express our great thankfulness to the anonymous reviewers for their constructive suggestions and encouragement. We thank the administrators, medical staff at Sichuan Provincial People's Hospital for their contribution to this work.

References

Afeche P, Mendelson H (2004) Pricing and priority auctions in queueing systems with a generalized delay cost structure. Manag Sci 50:869–882

Allon G, Hanany E (2012) Cutting in line: social norms in queues. Manag Sci 58:493–506

Anderson C, Butcher C, Moreno A (2010) Emergency department patient flow simulation at HealthAlliance. Project proposal, Worcester Polytechnic Institute, Worcester: MA, (Chapter 1)

Boxma OJ, van der Wal J, Yechiali U (2008) Polling with batch service. Stoch Models 24(4):604–625

Eric PJ, Thomas LP (2009) A review and synthesis of demand management, capacity management and performance in health-care services. Int J Manag Rev 11:149–174

Gómez-Corral A, Krishnamoorthy A, Narayanan VC (2005) The impact of self-generation of priorities on multi-server queues with finite capacity. Stoch Models 21:427–447

Green L, Savin S, Wang B (2006) Managing patient demand in a diagnostic medical facility. Oper Res 54:11–25

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 56:576–592

He QM, Chavoushi AA (2013) Analysis of queueing systems with customer interjections. Queueing Syst 73:79–104

Krishnamoorthy A, Babu S, Narayanan VC (2008) $MAP/(PH/PH)/c$ queue with self-generation of priorities and non-preemptive service. Stoch Ana Appl 26:1250–1266

Krishnamoorthy A, Babu S, Narayanan VC (2009) The $MAP/(PH/PH)/1$ queue with self-generation of priorities and non-preemptive service. Eur J Oper Res 195:174–185

Krishnamoorthy A, Narayanan VC, Deepak TG (2005) On a queueing system with self generation of priorities. Neural Parallel Sci Comput 13:119–130

Luo J, Kulkarni V, Ziya S (2012) Appointment scheduling under patient no-shows and service interruptions. Manuf Serv Oper Manag 14:670–684

Meng YK (1989) The foundation of queueing theory and application. Tongji University Press, Shanghai (Chapter 5, in Chinese)

Perel N, Yechiali U (2013) The Israeli queue with priorities. Stochas Models 29:353–379

Ross M (2007) Introduction to probability models, 9th edn. Academic Press, Inc., Orlando (Chapter 2)

Wang HK, Huang HM (1995) Optimal control of a removable server in an $M/E_k/1$ queueing system with finite capacity. Microelectron Reliab 35:1023–1030

Wang Q (2004) Modeling and analysis of high risk patient queues. Eur J Oper Res 155:502–515

Zhang HB, Shi DH (2010) Explicit solution for $M/M/1$ preemptive priority queue. Int J Inf Manag Sci 21:197–208