Queueing Syst (2011) 67: 251–273
DOI 10.1007/s11134-010-9207-9

Analysis of a dynamic assignment of impatient customers to parallel queues

A. Movaghar

Received: 18 July 2009 / Revised: 24 November 2010 / Published online: 3 February 2011
© Springer Science+Business Media, LLC 2011

Abstract Consider a number of parallel queues, each with an arbitrary capacity and multiple identical exponential servers. The service discipline in each queue is first-come-first-served (FCFS). Customers arrive according to a state-dependent Poisson process. Upon arrival, a customer joins a queue according to a state-dependent policy or leaves the system immediately if it is full. No jockeying among queues is allowed. An incoming customer to a parallel queue has a general patience time dependent on that queue after which he/she must depart from the system immediately. Parallel queues are of two types: type 1, wherein the impatience mechanism acts on the waiting time; or type 2, a single server queue wherein the impatience acts on the sojourn time. We prove a key result, namely, that the state process of the system in the long run converges in distribution to a well-defined Markov process. Closed-form solutions for the probability density function of the virtual waiting time of a queue of type 1 or the offered sojourn time of a queue of type 2 in a given state are derived which are, interestingly, found to depend only on the local state of the queue. The efficacy of the approach is illustrated by some numerical examples.

Keywords Analytical models · Dynamic policy · Impatient customers · Parallel queues

Mathematics Subject Classification (2000) 60K25 · 68M20 · 90B22

1 Introduction

We consider s parallel queues where the ith queue has a capacity $K_i \leq \infty$ and m_i identical exponential servers with service rate μ_i, $1 \leq i \leq s$. Customers arrive accord-

A. Movaghar (✉)
Computer Engineering Department, Sharif University of Technology, P.O. Box 11155-9517, Azadi Ave., Tehran 1998717869, Iran
e-mail: movaghar@sharif.edu

ing to a state-dependent Poisson process with rate λ, where λ is a function of the number of customers in each queue in the system. Upon arrival in a state $\mathbf{n} = (n_1, \ldots, n_s)$, where n_i is the number of customers in the ith queue, $1 \leq i \leq s$, a customer joins the jth queue, $1 \leq j \leq s$, with a probability $r_{\mathbf{n}}(j)$ (where $\sum_{j=1}^{s} r_{\mathbf{n}}(j) = 1$). No jockeying among queues is allowed. Customers in each queue are served in the order of their arrival and leave the system after they finish their service requirements. Each incoming customer to the ith queue has a deadline. The difference between the deadline of this customer and his/her arrival time, referred to as a *relative deadline*, is a random variable with a general probability distribution function $G_i(.)$, where $G_i(0) = 0$. Customers leave the system and are considered lost as soon as they miss their deadlines. Each parallel queue belongs to one of two types: type 1 or type 2. In a type 1 queue a customer keeps his/her deadline only until he/she begins service, so that once he/she begins service he/she will complete his/her service. A type 2 queue has only a single server. In such a queue a customer keeps his/her deadline until he/she ends service, so that he/she may miss his/her deadline during his/her service. Customer service times and relative deadlines form sequences of mutually independent i.i.d. random variables. Given the number of customers in the system at any time, the future arrival process is conditionally independent of the past history of the system.

The above system is an instance of a queueing system with impatient customers where customers have deadlines and may not stay in the system indefinitely [3, 4, 6–10, 12]. Recently, such queues have also been studied in Markovian random environments [1, 2, 14, 16]. Moreover, due to the difficult nature of the analysis for such models, almost all of them are assumed to be single-queues. Apart from an earlier work by the author [13], no other analysis of the dynamic assignment of impatient customers to parallel queues has been reported. The current paper is in fact an extension to this latter work in the following respects. First, it considers an arbitrary combination of parallel queues of type 1 or type 2 instead of only parallel queues of type 1. Secondly, the arrival process is more general, i.e., it is a state-dependent Poisson process with a generally state-dependent rate, not necessarily dependent on the total number of customers in the system as was the case in the previous work. Thirdly, the customer impatience is more general here where the relative deadline of an incoming customer to a parallel queue may depend on that queue. Fourthly, important new performance variables along with some interesting closed-form solutions for the probability density functions of these variables are introduced that were not considered in the previous work. Fifthly, the proof of the main results are presented differently and more rigorously here. Finally, the model properties and numerical results in the current paper are more enriched. Dynamic routing of impatient customers among some parallel queues often occurs in practice and has many important applications. In most of today's high-speed packet switching networks, for example, individual packets usually have some real-time constraints and must dynamically be routed among some high-speed links.

An important application domain of interest in this paper is a class of real-time systems called *firm real-time* (FRT) [5]. Contrary to *hard real-time* (HRT) systems, FRT systems are not required to meet all their deadlines. Deadlines in such systems can be met statistically with an upper bound on the fraction of allowed deadline misses, where the ability to respect this bound is very much affected by the scheduling mech-

anism being used. Some examples are multimedia-related applications in mobile devices or target tracking applications in sensor networks. Moreover, contrary to *soft real-time* (SRT) systems wherein jobs (customers) missing their deadlines can continue their execution with degraded values, such jobs (customers) in FRT systems are of no value and are usually thrown away. Two models of job (customer) behavior are considered in such systems: deadlines until the beginning of service and deadlines until the end of systems. Jobs (customers) in the former model are usually assigned to type 1 queues while those in the latter model are usually scheduled to type 2 queues (see the previous paragraphs) for further processing. Often, it is essential to find a dynamic policy that assigns incoming jobs (customers) to such parallel queues appropriately.

This paper presents an analytical modeling method for a dynamic assignment of customers with general impatience to a number of parallel queues of type 1 or type 2 as mentioned earlier. The method is novel and in fact, to the best of our knowledge, no other analytical method for a similar problem exists. The paper is organized as follows. Section 2 identifies some performance measures of interest. Section 3 introduces some important parameters and performance variables. Closed-form solutions for these parameters and the probability density function of these variables are derived. The latter results are used in an analytical model of the system in Sect. 4. Finally, Sect. 5 presents some simple numerical examples to illustrate the efficacy of our method.

2 Performance measures

In this section, we identify some performance measures of interest. These measures depend on the types of parallel queues. Recall that each parallel queue may belong to one of two types: type 1 or type 2. Different performance variables will be of interest for each type. Throughout this paper, we will assume statistical equilibrium and use τ to denote a variable with values in the set of non-negative real numbers. Let us first assume that the ith parallel queue is type 1, i.e., the deadlines of customers are effective until the beginning of their service. An important performance variable for this type of queue may be defined as

$$U^i \equiv \text{the time an incoming customer to the } i\text{th queue with no deadline must wait before he/she begins his/her service in the long run.} \tag{2.1}$$

We assume $U^i = \infty$ if the arriving customer to the ith queue is blocked due to the queue's being full. U^i is called the *virtual waiting time* of the ith queue. We will be interested in finding the probability distribution function of U^i, denoted as $F_{U^i}(.)$, or equivalently, its probability density function $f_{U^i}(.)$.

More specific measures of performance may also be defined. Let θ^i be a random variable representing the relative deadline of a customer, i.e., the difference between the deadline of a customer and his/her arrival time in the ith queue, and ρ_i the time

average fraction of customers joining the ith queue in the long run. We will be interested in the *probability of missing deadline* of the ith queue, defined as

$$\alpha_d^i = \rho_i P\big(\theta^i < U^i < \infty\big) = \rho_i \int_0^\infty G_i(\tau)\, dF_{U^i}(\tau). \tag{2.2}$$

α_d^i represents the steady-state probability that a customer misses his/her deadline in the ith queue. When the ith queue has a finite capacity (i.e., $K_i < \infty$), where blocking is allowed, an important measure of performance is the *probability of blocking* α_b^i, defined as

$$\alpha_b^i = \rho_i P\big(U^i = \infty\big) = \rho_i\big(1 - F_{U^i}(\infty - 0)\big). \tag{2.3}$$

α_b^i is interpreted as the steady-state probability that an arriving customer is rejected due to full capacity at the ith queue.

Next, we assume that the ith queue is type 2, i.e., the deadlines of customers are effective until the end of their service. Our principal performance variable for this type of queue may be defined as

$$\begin{aligned} V^i \equiv{}& \text{the time an incoming customer to the } i\text{th queue with no} \\ & \text{deadline must wait before he/she completes his/her service in} \\ & \text{the long run.} \end{aligned} \tag{2.4}$$

We assume $V^i = \infty$ if the arriving customer to the ith queue is blocked due to the queue's being full. V^i is called the *offered sojourn time* of the ith queue. We will also be interested in finding the probability distribution function of V^i, denoted as $F_{V^i}(.)$, or equivalently, its probability density function $f_{V^i}(.)$.

Similarly, the *probability of missing deadline* of the ith queue may be defined as

$$\alpha_d^i = \rho_i P\big(\theta < V^i < \infty\big) = \rho_i \int_0^\infty G(\tau)\, dF_{V^i}(\tau). \tag{2.5}$$

The *probability of blocking* of the ith queue, α_b^i, may also be defined as

$$\alpha_b^i = \rho_i P\big(V^i = \infty\big) = \rho_i\big(1 - F_{V^i}(\infty - 0)\big). \tag{2.6}$$

More general measures of performance may also be considered. In particular, we may define the *probability of missing deadline* of the system as

$$\alpha_d = \sum_{i=1}^{s} \alpha_d^i. \tag{2.7}$$

α_d represents the steady-state probability that a customer misses his/her deadline in the system. Similarly, we may define the *probability of blocking* of the system as

$$\alpha_b = \sum_{i=1}^{s} \alpha_b^i. \tag{2.8}$$

α_b represents the steady-state probability that an incoming customer to the system is rejected due to the queue's being full.

Combining the last two measures, we may define the *probability of loss* of the system as

$$\alpha = \alpha_d + \alpha_b. \tag{2.9}$$

α is viewed as the steady-state probability that a customer is lost due to either missing his/her deadline or being rejected due to full queue.

When all parallel queues are type 1 (i.e., all customers in the system have deadlines until the beginning of their service), we may define a more relevant performance variable as follows. Let

$$\begin{aligned} U \equiv & \text{ the time an incoming customer with no deadline to the} \\ & \text{system must wait before he/she begins his/her service} \\ & \text{in the long run.} \end{aligned} \tag{2.10}$$

U is referred to as the *virtual waiting time* of the system. We will be interested in finding the probability distribution function of U, denoted as $F_U(.)$, or equivalently, its probability density function $f_U(.)$. Similarly, when all parallel queues are type 2 (i.e., all customers in the system have deadlines until the end of their service), we can define the performance variable

$$\begin{aligned} V \equiv & \text{ the time an incoming customer with no deadline to the} \\ & \text{system must wait before he/she completes his/her service} \\ & \text{in the long run.} \end{aligned} \tag{2.11}$$

V is called the *offered sojourn time* of the system. We will be interested in finding the probability distribution function of V, denoted as $F_V(.)$, or equivalently, its probability density function $f_V(.)$.

3 Loss rate functions

This section presents notions of some important parameters and performance variables. Closed-form solutions for these parameters and the probability density function of these variables are derived. These results will be used in an analytical model of the system in the next section.

Let $\mathbf{n} = (n_1, \ldots, n_s)$ be a s-tuple of natural numbers. Denote $\psi^i(t, \mathbf{n}, \epsilon)$ to be the probability that a customer in the ith queue, $1 \le i \le s$, misses his/her deadline during $[t, t+\epsilon)$, given there are n_j customers in the jth queue, $1 \le j \le s$, at time t. Define

$$\Gamma_i(t, \mathbf{n}) = \lim_{\epsilon \to 0} \frac{\psi^i(t, \mathbf{n}, \epsilon)}{\epsilon}. \tag{3.1}$$

Assuming statistical equilibrium, we have

$$\Gamma_i(\mathbf{n}) = \lim_{t \to \infty} \Gamma_i(t, \mathbf{n}), \tag{3.2}$$

where $\Gamma_i(.)$ is said to be the *loss rate* function of the ith queue.

Barrer [4] first introduced a similar function for a single queue with Poisson arrival process and deterministic customer impatience. He found some closed-form solutions for this function in terms of service rate (μ) and mean relative deadline ($\overline{\theta}$) for both queue types mentioned earlier. Barrer's results have been extended to a single queue with a state-dependent Poisson arrival process and a generally distributed customer impatience [7, 12]. We further extend these results to a larger class of models, namely, the class of parallel queues considered in this paper. In particular, we demonstrate a very interesting result, namely, that the loss rate for the ith queue is a function only of n_i.

Let us first assume that the ith queue in the system is type 1 and in a busy period (i.e., the number of customers in the ith queue is at least m_i). We may define the following random variable:

$$U^i_{\mathbf{n}}(t) \equiv \text{the time a virtual customer with no deadline and arriving at the } i\text{th queue at time } t \text{ must wait before he/she begins service, given there are } n_j \text{ customers in the } j\text{th queue, } j = 1, \ldots, s, \text{ at time } t. \tag{3.3}$$

$$A^i_{\mathbf{n}}(t) \equiv \text{the time a customer being served by one of the } i\text{th queue's servers at time } t \text{ has spent in the system, given there are } n_j \text{ customers in the } j\text{th queue, } j = 1, \ldots, s, \text{ at time } t. \tag{3.4}$$

$U^i_{\mathbf{n}}(t)$ above is referred to as the conditional *virtual waiting time* of the ith queue at time t. The second random variable above is also well-defined when there are multiple servers at the ith queue, i.e., $m_i > 1$. This is because all such multiple severs are supposed to be busy and have similar exponentially distributed service times. Thus, this random variable must have similar distribution for any given server of the ith queue and hence it is uniquely defined up to distribution. It is called the conditional *attained waiting time* of the ith queue at time t. Let T^i_k be the time of the kth arrival at the ith queue and S^i_k the time of the kth arrival of a customer who will successfully be served at the ith queue, $k = 1, 2, \ldots, i = 1, \ldots, s$, given the arrival observes that the number of customers in each queue is represented by a vector $\mathbf{n}$. For any time t, we also use $t+$ and $t-$ to denote a time immediately after and before t, respectively, and $=^d$ to denote the equality in distribution among random variables. Assuming statistical equilibrium, we may define the following random variables:

$$U^i_{\mathbf{n}} =^d \lim_{k\to\infty} U^i_{\mathbf{n}}\big(T^i_k-\big), \tag{3.5}$$

$$\hat{U}^i_{\mathbf{n}} =^d \lim_{k\to\infty} U^i_{\mathbf{n}}\big(S^i_k+\big), \tag{3.6}$$

$$\tilde{U}^i_{\mathbf{n}} =^d \lim_{t\to\infty} U^i_{\mathbf{n}}(t), \tag{3.7}$$

$$A^i_{\mathbf{n}} =^d \lim_{k\to\infty} A^i_{\mathbf{n}}\big(T^i_k-\big), \tag{3.8}$$

$$\hat{A}^i_{\mathbf{n}} =^d \lim_{k\to\infty} A^i_{\mathbf{n}}\big(S^i_k+\big). \tag{3.9}$$

$$\tilde{A}^i_{\mathbf{n}} =^d \lim_{t\to\infty} A^i_{\mathbf{n}}(t), \tag{3.10}$$

$U^i_{\mathbf{n}}$ and $\hat{U}^i_{\mathbf{n}}$ above represent the conditional steady-state virtual waiting time in the ith queue immediately before the arrival of a new customer and immediately after the arrival of a new successful customer at the ith queue, respectively. $\tilde{U}^i_{\mathbf{n}}$ represents the conditional steady-state time average of the virtual waiting time in the ith queue. Similarly, $A^i_{\mathbf{n}}$ and $\hat{A}^i_{\mathbf{n}}$ above represent the conditional steady-state attained waiting time in the ith queue immediately after the departure of customer and immediately before the departure of a successful customer from the ith queue, respectively. $\tilde{A}^i_{\mathbf{n}}$ represents the conditional steady-state time average of the attained waiting time in the ith queue.

Let θ^i_n and E^i_n represent the relative deadline of the nth customer in the ith queue and the time between when this customer begins service, if ever, and the next service completion, respectively, in the long run. Thus, θ^i_n is a random variable with a probability distribution function $G_i(.)$ and E^i_n is a random variable with an exponential probability distribution function with rate $m_i\mu_i$. Moreover, $\{\theta^i_n; n \geq 0\}$ and $\{E^i_n; n \geq 0\}$ form independent sets of i.i.d. random variables. Also, let $\mathbf{n} = (n_1, \ldots, n_i, \ldots, n_s)$ be a s-tuple of natural numbers and $\mathbf{e}_i$ one such s-tuple with value of 0 at each coordinate except for coordinate i at which it has a value of 1, i.e., $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)$. Denote $\mathbf{n} + \mathbf{e}_i = (n_1, \ldots, n_i + 1, \ldots, n_s)$ and $\mathbf{n} - k\mathbf{e}_i = (n_1, \ldots, n_i - k, \ldots, n_s)$, where k is a natural number no greater than n_i. We show:

Lemma 3.1 *For $n_i \geq m_i$,*

$$P\big(U^i_{\mathbf{n}} \leq \tau\big) = P\big(U^i_{\mathbf{n}-\mathbf{e}_i} + E^i_{n_i} \leq \tau | U^i_{\mathbf{n}-\mathbf{e}_i} \leq \theta^i_{n_i}\big), \tag{3.11}$$

where $E^i_{n_i}$ is a random variable with an exponential probability distribution function with rate $m_i\mu_i$ which is independent of $U^i_{\mathbf{n}-\mathbf{e}_i}$ and $\theta^i_{n_i}$.

Proof Consider now that the system is in equilibrium run where the ith queue is in a busy period and the number of customers in each queue is represented by the vector $\mathbf{n} - \mathbf{e}_i$. Suppose, a new customer arrives at the ith queue which will successfully be served. Clearly, the virtual waiting time in the ith queue before and after this new arrival can be represented by $U^i_{\mathbf{n}-\mathbf{e}_i}$, conditioned by the event $\{U^i_{\mathbf{n}-\mathbf{e}_i} \leq \theta^i_{n_i}\}$, and $\hat{U}^i_{\mathbf{n}}$, respectively. Moreover, the virtual waiting time will increase immediately after the new arrival by exactly the same value as the time between when this new arrival begins service and the next service completion. Thus, for $n_i \geq m$, we have

$$P\big(\hat{U}^i_{\mathbf{n}} \leq \tau\big) = P\big(U^i_{\mathbf{n}-\mathbf{e}_i} + E^i_{n_i} \leq \tau | U^i_{\mathbf{n}-\mathbf{e}_i} \leq \theta^i_{n_i}\big), \tag{3.12}$$

where $E^i_{n_i}$ is an exponentially distributed random variable with rate $m_i\mu_i$ which is independent of $U_{\mathbf{n}-\mathbf{e}_i}$ and $\theta^i_{n_i}$. We need to show

$$P\big(\hat{U}^i_{\mathbf{n}} \leq \tau\big) = P\big(U^i_{\mathbf{n}} \leq \tau\big). \tag{3.13}$$

Let $SA^i_{\mathbf{n}}(t-\epsilon, t)$ be the event that a customer who will successfully be served arrives at the ith queue during $[t-\epsilon, t)$ given there are n_j customers in the jth queue,

$1 \le j \le s$, at time t. Clearly, we have

$$P\big(\hat{U}^i_{\mathbf{n}} \le \tau\big) = \lim_{t\to\infty} \lim_{\epsilon\to 0} P\big[U^i_{\mathbf{n}}(t) \le \tau \,|\, SA^i_{\mathbf{n}}(t-\epsilon, t)\big]. \tag{3.14}$$

Equation (3.13) can now be proved by induction on n_i. For $n_i < m_i$, (3.13) is obviously true because we simply have $\hat{U}^i_{\mathbf{n}} = U^i_{\mathbf{n}} = 0$. For the induction step, suppose (3.13) is true for all $n_i < k_i$ where $k_i \ge m_i$. We need to prove (3.13) is also valid for $n_i = k_i$. From (3.12) and the induction assumption, it can easily be shown that $P(\hat{U}^i_{\mathbf{n}} \le \tau)$ is independent of the overall system arrival rate $\lambda(\mathbf{n})$ when $n_i = k_i$. (For more clarification, please see Lemma 3.3.) On the other hand, for any $t > 0$ and a sufficiently small $\epsilon > 0$, we have

$$P\big(SA^i_{\mathbf{n}}(t-\epsilon, t)\big) = \lambda(\mathbf{n}) r_{\mathbf{n}}(i) P\big(U^i_{\mathbf{n}-\mathbf{e}_i}(t-\epsilon) \le \theta^i_{n_i}\big)\epsilon + o(\epsilon), \tag{3.15}$$

where $\lim_{\epsilon\to 0} \frac{o(\epsilon)}{\epsilon} = 0$. This means that by varying $\lambda(\mathbf{n})$ arbitrarily, the value of $P(\hat{U}^i_{\mathbf{n}} \le \tau)$ must remain unchanged when $n_i = k_i$. We may choose $\lambda(\mathbf{n})$ large enough so that $P(SA^i_{\mathbf{n}}(t-\epsilon, t)) = 1$. Then, from (3.9), we must have

$$P\big(\hat{U}^i_{\mathbf{n}} \le \tau\big) = \lim_{t\to\infty} P\big(U^i_{\mathbf{n}}(t) \le \tau\big) = P\big(\tilde{U}^i_{\mathbf{n}} \le \tau\big). \tag{3.16}$$

Using conditional PASTA [15], we get

$$P\big(\tilde{U}^i_{\mathbf{n}} \le \tau\big) = P\big(U^i_{\mathbf{n}} \le \tau\big), \tag{3.17}$$

which completes the proof. □

Let θ^i_1 and E^i_1 represent the relative deadline of a customer who is to depart next from the ith queue and the time between when this customer completes service and the previous service completion at the ith queue, respectively, in the long run. Thus, θ^i_1 is a random variable with a probability distribution function $G_i(.)$ and E^i_1 is a random variable with an exponential probability distribution function with rate $m_i\mu_i$. We have

Lemma 3.2 *For $n_i \ge m_i$,*

$$P\big(A^i_{\mathbf{n}} \le \tau\big) = P\big(A^i_{\mathbf{n}-\mathbf{e}_i} + E^i_1 \le \tau \,|\, A^i_{\mathbf{n}-\mathbf{e}_i} \le \theta^i_1\big), \tag{3.18}$$

where E^i_1 is an exponentially distributed random variable with rate $m_i\mu_i$ which is independent of $A^i_{\mathbf{n}-\mathbf{e}_i}$ and θ^i_1.

Proof Consider that the system is in equilibrium where the ith queue is in a busy period and the state of the system is represented by a vector $\mathbf{n}$. Suppose a customer completes his/her service successfully at the ith queue. Clearly, the attained waiting time in the ith queue right before this service completion can be represented by $\hat{A}^i_{\mathbf{n}}$ which, by definition, is equal to $A^i_{\mathbf{n}-\mathbf{e}_i} + E^i_1$, conditioned by the event $\{A^i_{\mathbf{n}-\mathbf{e}_i} \le \theta^i_1\}$,

where E_1^i is an exponentially distributed random variable with rate $m_i\mu_i$ which is independent of $A^i_{\mathbf{n}-\mathbf{e}_i}$ and θ_1^i. Thus, for $n_i \geq m_i$, we can write

$$P\big(\hat{A}^i_{\mathbf{n}} \leq \tau\big) = P\big(A^i_{\mathbf{n}-\mathbf{e}_i} + E_1^i \leq \tau | A^i_{\mathbf{n}-\mathbf{e}_i} \leq \theta_1^i\big). \tag{3.19}$$

Denote by $SD^i_{\mathbf{n}}(t, t+\epsilon)$ the event that a customer in the ith queue completes his/her service successfully during $(t, t+\epsilon]$, given there are n_j customers in the jth queue, $1 \leq j \leq s$, at time t. Obviously, we have

$$P\big(\hat{A}^i_{\mathbf{n}} \leq \tau\big) = \lim_{t\to\infty}\lim_{\epsilon\to 0} P\big[A^i_{\mathbf{n}}(t) \leq \tau \,|\, SD^i_{\mathbf{n}}(t, t+\epsilon)\big]. \tag{3.20}$$

We note that for any $t > 0$, $n_i \geq m$ and a sufficiently small ϵ, the successful customer departures from the ith queue will form a Poisson process with the rate $m_i\mu_i$. Moreover, the event $\{A^i_{\mathbf{n}}(t) \leq \tau\}$ is independent of the event $SD^i_{\mathbf{n}}(t, t+\epsilon)$, and we can write

$$\lim_{\epsilon\to 0} P\big[A^i_{\mathbf{n}}(t) \leq \tau \,|\, SD^i_{\mathbf{n}}(t, t+\epsilon)\big] = P\big[A^i_{\mathbf{n}}(t) \leq \tau\big], \tag{3.21}$$

or

$$P\big(\hat{A}^i_{\mathbf{n}} \leq \tau\big) = \lim_{t\to\infty} P\big(A^i_{\mathbf{n}}(t) \leq \tau\big) = P\big(\tilde{A}^i_{\mathbf{n}} \leq \tau\big). \tag{3.22}$$

Using conditional reverse ASTA [11], we get

$$P\big(\tilde{A}^i_{\mathbf{n}} \leq \tau\big) = P\big(A^i_{\mathbf{n}} \leq \tau\big), \tag{3.23}$$

which completes the proof. □

Let $F_{U^i_{\mathbf{n}}}(.)$ ($F_{A^i_{\mathbf{n}}}(.)$) and $f_{U^i_{\mathbf{n}}}(.)$ ($f_{A^i_{\mathbf{n}}}(.)$) denote the probability distribution function and the probability density function of $U^i_{\mathbf{n}}$ ($A^i_{\mathbf{n}}$), respectively.

From Lemma 3.1, we have

$$F_{U^i_{\mathbf{n}}}(\tau) = 1, \quad \text{if } n_i < m_i,$$

$$F_{U^i_{\mathbf{n}}}(\tau) = \frac{1}{P(U^i_{\mathbf{n}-\mathbf{e}_i} \leq \theta^i_{n_i})} \int_0^\tau \big(1 - e^{-m_i\mu_i(\tau-x)}\big)\big(1 - G_i(x)\big)\, dF_{U^i_{\mathbf{n}-\mathbf{e}_i}}(x), \tag{3.24}$$

$$\text{if } n_i \geq m_i,$$

or, equivalently,

$$f_{U^i_{\mathbf{n}}}(\tau) = 0, \quad \text{if } n_i < m_i,$$

$$f_{U^i_{\mathbf{n}}}(\tau) = \frac{m_i\mu_i e^{-m_i\mu_i\tau}}{P(U^i_{\mathbf{n}-\mathbf{e}_i} \leq \theta^i_{n_i})} \int_0^\tau f_{U^i_{\mathbf{n}-\mathbf{e}_i}}(x) e^{m_i\mu_i x}\big(1 - G_i(x)\big)dx, \tag{3.25}$$

$$\text{if } n_i \geq m_i.$$

A solution for (3.25) may be found as

$$f_{U^i_{\mathbf{n}}}(\tau) = 0, \quad \text{if } n_i < m_i,$$

$$f_{U^i_{\mathbf{n}}}(\tau) = \frac{(m_i\mu_i)^{n_i-m_i+1}}{(n_i-m_i)!\prod_{k=1}^{n_i-m_i} P(U^i_{\mathbf{n}-k\mathbf{e}_i} \le \theta^i_{(n_i-k+1)})} \times \left[\int_0^\tau \big(1-G_i(x)\big)\,dx\right]^{n_i-m_i} e^{-m_i\mu_i\tau}, \quad \text{if } n_i \ge m_i. \tag{3.26}$$

Similarly, using Lemma 3.2, one can show

$$f_{A^i_{\mathbf{n}}}(\tau) = 0, \quad \text{if } n_i < m_i,$$

$$f_{A^i_{\mathbf{n}}}(\tau) = \frac{(m_i\mu_i)^{n_i-m_i+1}}{(n_i-m_i)!\prod_{k=1}^{n_i-m_i} P(A^i_{\mathbf{n}-k\mathbf{e}_i} \le \theta^i_{(n_i-k+1)})} \times \left[\int_0^\tau \big(1-G_i(x)\big)\,dx\right]^{n_i-m_i} e^{-m_i\mu_i\tau}, \quad \text{if } n_i \ge m_i. \tag{3.27}$$

Define $\Phi^i_n(s)$ to be the Laplace transform of $[\int_0^\tau (1-G_i(x))\,dx]^n$, i.e.,

$$\Phi^i_n(s) = \int_0^\infty \left[\int_0^\tau \big(1-G_i(x)\big)\,dx\right]^n e^{-s\tau}\,d\tau. \tag{3.28}$$

We have

Lemma 3.3

$$f_{U^i_{\mathbf{n}}}(\tau) = f_{A^i_{\mathbf{n}}}(\tau) = 0, \quad \text{if } n_i < m_i,$$

$$f_{U^i_{\mathbf{n}}}(\tau) = f_{A^i_{\mathbf{n}}}(\tau) = \frac{1}{\Phi_{n_i-m_i}(m_i\mu_i)}\left[\int_0^\tau \big(1-G_i(x)\big)dx\right]^{n_i-m_i} e^{-m_i\mu_i\tau}, \tag{3.29}$$

$$\text{if } n_i \ge m_i.$$

Proof The proof is simple by noting that $f_{U^i_{\mathbf{n}}}(\tau)$ and $f_{A^i_{\mathbf{n}}}(\tau)$ are probability density functions which can also be derived as in (3.27) and (3.28), respectively. □

Let θ^i_n be the relative deadline of the nth customer in the ith queue and θ^i_1 the relative deadline of a customer who is to depart next from the same queue. Clearly, $P(U^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i})$ and $P(A^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_1)$ represent the ratio of the incoming customers to and the ratio of the departing customers from the ith queue who meet their deadlines in the long run, respectively, given the state of the system seen by these customers is represented by the vector $\mathbf{n}-\mathbf{e_i}$. We have

Lemma 3.4

$$P\left(U^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i}\right) = P\left(A^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_1\right) = 1, \quad \text{if } 0 < n_i \le m_i,$$
$$P\left(U^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i}\right) = P\left(A^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_1\right) = \frac{(m_i\mu_i)\Phi^i_{n_i-m_i}(m_i\mu_i)}{(n_i-1)\Phi^i_{n_i-m_i-1}(m_i\mu_i)}, \quad \text{if } n_i > m_i. \tag{3.30}$$

Proof Comparing (3.27) and (3.28) with (3.30), the proof is immediate. □

The above lemma indicates that the proportion of the incoming customers to and the proportion of the departing customers from the ith queue of type 1 who meet their deadlines in the long run, given the state of the system seen by these customers is represented by a vector $\mathbf{n}$, are the same and only depend on the number of customers in the ith queue, i.e., n_i, and do not depend on the number of customers in the other queues, i.e., n_j, $j \neq i$, $j = 1, \ldots, s$. This is an important result which can help us derive a closed-form solution for the loss rate function for the case of a queue of type 1, i.e., when customers have deadlines until the beginning of their service. More specifically, let

$$\gamma^1_i(n) = \begin{cases} (n-m_i)\dfrac{\Phi^i_{n-m_i-1}(m_i\mu_i)}{\Phi^i_{n-m_i}(m_i\mu_i)} - m_i\mu_i, & \text{if } n > m_i, \\ 0, & \text{if } n \le m_i. \end{cases} \tag{3.31}$$

We have

Theorem 3.1 *Let* $\Gamma^1_i(.)$ *be defined as in* (3.2), *where the* i*th queue is type* 1, *and* $\gamma^1_i(.)$ *is defined as in* (3.31). *Then* $\Gamma^1_i(\mathbf{n}) = \gamma^1_i(n_i)$.

Proof Consider that the system is in equilibrium and in a state represented by a vector $\mathbf{n}$ where $n_i \ge m_i$. Suppose that a customer departs from the ith queue. Using the definition of $\Gamma^1_i(\mathbf{n})$ as in (3.2), the probability that this latter departure is successful is simply

$$\frac{m_i\mu_i}{m_i\mu_i + \Gamma^1_i(\mathbf{n})}.$$

On the other hand, by definition, the same probability may also be written as

$$P\left(A^i_{\mathbf{n}-\mathbf{e}_i} \le \theta^i_1\right),$$

where θ^i_1 represents the relative deadline of the departing customer from the ith queue. Thus, we have

$$P\left(A^i_{\mathbf{n}-\mathbf{e}_i} \le \theta^i_1\right) = \frac{m_i\mu_i}{m_i\mu_i + \Gamma^1_i(\mathbf{n})}. \tag{3.32}$$

Using Lemma 3.4, we finally get

$$\begin{aligned} \Gamma_i^1(\mathbf{n}) &= 0, \quad \text{if } n_i \le m_i, \\ \Gamma_i^1(\mathbf{n}) &= (n_i - m_i)\frac{\Phi^i_{n_i-m_i-1}(m_i\mu_i)}{\Phi^i_{n_i-m_i}(m_i\mu_i)} - m_i\mu_i, \quad \text{if } n_i > m_i, \end{aligned} \tag{3.33}$$

which completes the proof. □

Next, we assume that the ith queue is type 2. Let

$$\begin{aligned} V^i_{\mathbf{n}}(t) \equiv{}& \text{the time a virtual customer with no deadline and arriving at the} \\ & i\text{th queue at time } t \text{ must wait before he/she completes service,} \\ & \text{given there are } n_j \text{ customers in the } j\text{th queue, } j = 1, \ldots, s, \\ & \text{at time } t. \end{aligned} \tag{3.34}$$

$V^i_{\mathbf{n}}(t)$ is called the conditional *offered sojourn time* of the ith queue at time t. Let T^i_k and S^i_k be defined as in the proof of Lemma 3.1. Assuming statistical equilibrium, we have

$$V^i_{\mathbf{n}} =^d \lim_{k\to\infty} V^i_{\mathbf{n}}(T_k-), \tag{3.35}$$

$$\hat{V}^i_{\mathbf{n}} =^d \lim_{k\to\infty} V^i_{\mathbf{n}}(S_k+). \tag{3.36}$$

$$\tilde{V}^i_{\mathbf{n}} =^d \lim_{t\to\infty} V_{\mathbf{n}}(t), \tag{3.37}$$

$V^i_{\mathbf{n}}$ and $\hat{V}^i_{\mathbf{n}}$ above represent the conditional steady-state sojourn time in the ith queue immediately before the arrival of a new customer and immediately after the arrival of a new successful customer to the system, respectively. $\tilde{V}^i_{\mathbf{n}}$ is simply the conditional steady-state time average of the offered sojourn time in the ith queue.

$V^i_{\mathbf{n}}$ is called the conditional *offered sojourn time* of the ith queue. A similar definition may be given for a departing customer from the ith queue as follows. Let

$$\begin{aligned} S^i_{\mathbf{n}} \equiv{}& \text{the offered sojourn time (previously) seen upon arrival by} \\ & \text{a departing customer from the } i\text{th queue in the long run} \\ & \text{who finds the number of customers left in each queue is} \\ & \text{represented by a vector } \mathbf{n}. \end{aligned} \tag{3.38}$$

$S^i_{\mathbf{n}}$ is called the conditional *offered sojourn time* seen by a departing customer from the ith queue. Later, we will prove that these two random variables have similar distributions. Let θ^i_n and E^i_n represent the relative deadline and service time of the nth customer in the ith queue, respectively, in the long run. We have

Lemma 3.5

$$\begin{aligned} &P\big(V^i_{\mathbf{n}} \le \tau\big) = 1 - e^{-\mu_i \tau}, \quad \text{if } n_i = 0, \\ &P\big(V^i_{\mathbf{n}} \le \tau\big) = P\big(V^i_{\mathbf{n}-\mathbf{e_i}} + E^i_{n_i+1} \le \tau | V^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i}\big), \quad \text{if } n_i > 0, \end{aligned} \tag{3.39}$$

where $E^i_{n_i+1}$ *is independent of* $V^i_{\mathbf{n}-\mathbf{e}_i}$ *and* $\theta^i_{n_i}$.

Proof Consider that the system is in equilibrium and in a state represented by the vector $\mathbf{n} - \mathbf{e}_i$ where $n_i > 0$. Suppose, a new customer arrives at the ith queue which will successfully be served. Clearly, the offered sojourn time in the ith queue before and after this new arrival may be represented by $V^i_{\mathbf{n}-\mathbf{e}_i}$, conditioned by the event $\{V^i_{\mathbf{n}-\mathbf{e}_i} \le \theta^i_{n_i}\}$, and $\hat{V}^i_{\mathbf{n}}$, respectively. Moreover, the offered sojourn time will increase immediately after the new arrival by the service time of the next (virtual) customer. Thus, we can write

$$P\big(\hat{V}^i_{\mathbf{n}} \le \tau\big) = P\big(V^i_{\mathbf{n}-\mathbf{e}_i} + E^i_{n_i+1} \le \tau | V^i_{\mathbf{n}-\mathbf{e}_i} \le \theta^i_{n_i}\big), \tag{3.40}$$

where $E^i_{n_i+1}$ is an exponentially distributed random variable with rate μ_i, representing the service time of the next (virtual) arrival, which is independent of $V^i_{\mathbf{n}-\mathbf{e}_i}$ and $\theta^i_{n_i}$. Let $SA^i_{\mathbf{n}}(t-\epsilon, t)$ be an event defined exactly in the same manner as in the proof of Lemma 3.1. We can write

$$P\big(\hat{V}^i_{\mathbf{n}} \le \tau\big) = \lim_{t\to\infty} \lim_{\epsilon\to 0} P\big[V^i_{\mathbf{n}}(t) \le \tau \,|\, SA^i_{\mathbf{n}}(t-\epsilon, t)\big]. \tag{3.41}$$

Similarly, by induction on n_i, we can show

$$P\big(\hat{V}^i_{\mathbf{n}} \le \tau\big) = \lim_{t\to\infty} P\big(V^i_{\mathbf{n}}(t) \le \tau\big) = P\big(\tilde{V}^i_{\mathbf{n}} \le \tau\big). \tag{3.42}$$

Using conditional PASTA [15], we get

$$P\big(\tilde{V}^i_{\mathbf{n}} \le \tau\big) = P\big(V^i_{\mathbf{n}} \le \tau\big), \tag{3.43}$$

which completes the proof. □

Let θ^i_1 and E^i_1 represent the relative deadline and service time of a customer who is to depart next from the ith queue in the long run. Thus, θ^i_1 is a random variable with a probability distribution function $G(.)$ and E^i_1 is a random variable with an exponential probability distribution function with rate μ_i. We have

Lemma 3.6

$$\begin{aligned} &P\big(S^i_{\mathbf{n}} \le \tau\big) = 1 - e^{-\mu_i \tau}, \quad \text{if } n_i = 0, \\ &P\big(S^i_{\mathbf{n}} \le \tau\big) = P\big(S^i_{\mathbf{n}-\mathbf{e_i}} + E^i_1 \le \tau | S^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_1\big), \quad \text{if } n_i > 0, \end{aligned} \tag{3.44}$$

where E^i_1 *is independent of* $S^i_{\mathbf{n}-\mathbf{e}_i}$ *and* θ^i_1.

Proof Recall $A^i_{\mathbf{n}}$ as defined in (3.8). By definition, we have

$$S^i_{\mathbf{n}} =^d A^i_{\mathbf{n}} + E^i_1, \tag{3.45}$$

where $S^i_{\mathbf{n}}$ is defined as in (3.38). Using similar reasoning as in the proof of Lemma 3.2, we show:

$$P\big(A^i_{\mathbf{n}} \le \tau\big) = P\big(S^i_{\mathbf{n}-\mathbf{e_i}} \le \tau \mid S^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_1\big). \tag{3.46}$$

Using (3.45) and (3.46), the proof is immediate. □

Let $F_{V^i_{\mathbf{n}}}(.)$ $(F_{S^i_{\mathbf{n}}}(.))$and $f_{V^i_{\mathbf{n}}}(.)$ $(f_{S^i_{\mathbf{n}}}(.))$ represent the probability distribution function and the probability density function of $V^i_{\mathbf{n}}$ $(X^i_{\mathbf{n}})$, respectively. From Lemma 3.5, we have

$$\begin{aligned} F_{V^i_{\mathbf{n}}}(\tau) &= 1 - e^{-\mu_i \tau}, \quad \text{if } n_i = 0, \\ F_{V^i_{\mathbf{n}}}(\tau) &= \frac{1}{P(V^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i})} \int_0^\tau \big(1 - e^{-\mu_i(\tau - x)}\big)\big(1 - G_i(x)\big)\, dF_{V^i_{\mathbf{n}-\mathbf{e_i}}}(x), \quad \text{if } n_i \ge 1, \end{aligned} \tag{3.47}$$

or, equivalently,

$$\begin{aligned} f_{V^i_{\mathbf{n}}}(\tau) &= \mu_i e^{-\mu_i \tau}, \quad \text{if } n_i = 0, \\ f_{V^i_{\mathbf{n}}}(\tau) &= \frac{\mu_i e^{-\mu_i \tau}}{P(V_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i})} \int_0^\tau f_{V^i_{\mathbf{n}-\mathbf{e_i}}}(x) e^{\mu_i x}\big(1 - G_i(x)\big)\, dx, \quad \text{if } n_i \ge 1. \end{aligned} \tag{3.48}$$

A solution for (3.48) may be given as

$$\begin{aligned} f_{V^i_{\mathbf{n}}}(\tau) &= \mu_i e^{-\mu_i \tau}, \quad \text{if } n_i = 0, \\ f_{V^i_{\mathbf{n}}}(\tau) &= \frac{\mu^{n_i+1}}{n_i! \prod_{k=1}^{n_i} P(V_{\mathbf{n}-k\mathbf{e_i}} \le \theta^i_{(n_i-k+1)})} \left[\int_0^\tau \big(1 - G_i(x)\big)\, dx\right]^{n_i} e^{-\mu_i \tau}, \quad \text{if } n_i \ge 1. \end{aligned} \tag{3.49}$$

Similarly, using Lemma 3.6, one can show

$$\begin{aligned} f_{S^i_{\mathbf{n}}}(\tau) &= \mu_i e^{-\mu_i \tau}, \quad \text{if } n_i = 0, \\ f_{S^i_{\mathbf{n}}}(\tau) &= \frac{\mu^{n_i+1}}{n_i! \prod_{k=1}^{n_i} P(S_{\mathbf{n}-k\mathbf{e_i}} \le \theta^i_1)} \left[\int_0^\tau \big(1 - G_i(x)\big)\, dx\right]^{n_i} e^{-\mu_i \tau}, \quad \text{if } n_i \ge 1. \end{aligned} \tag{3.50}$$

We have

Lemma 3.7

$$f_{V^i_{\mathbf{n}}}(\tau) = f_{S^i_{\mathbf{n}}}(\tau) = \frac{1}{\Phi^i_{n_i}(\mu_i)} \left[\int_0^\tau \big(1 - G_i(x)\big)\, dx\right]^{n_i} e^{-\mu_i \tau}, \tag{3.51}$$

where $\Phi^i_{n_i}(\mu_i)$ *is defined as in* (3.28).

Proof The proof is simple by noting that $f_{V^i_{\mathbf{n}}}(\tau)$ and $f_{S^i_{\mathbf{n}}}(\tau)$ are probability density functions which are derived as in (3.49) and (3.50), respectively. □

Let θ^i_n be the relative deadline of the nth customer in the ith queue and θ^i_1 the relative deadline of a customer who is to depart next from the same queue. Clearly, $P(V^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i})$ and $P(S^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_1)$ represent the proportion of the incoming customers to and the proportion of the departing customers from the ith queue who meet their deadlines in the long run, respectively, given the state of the system seen by these customers is represented by the vector $\mathbf{n} - \mathbf{e_i}$. We have

Lemma 3.8

$$P\big(V^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_{n_i}\big) = P\big(S^i_{\mathbf{n}-\mathbf{e_i}} \le \theta^i_1\big) = \frac{\mu_i \Phi^i_{n_i}(\mu_i)}{n_i \Phi^i_{(n_i-1)}(\mu_i)}. \tag{3.52}$$

Proof Comparing (3.49) and (3.50) with (3.51), the proof is immediate. □

The above lemma indicates that the proportion of the incoming customers to and the proportion of the departing customers from the ith queue of type 2 who meet their deadlines in the long run, given the state of the system seen by these customers is represented by a vector $\mathbf{n}$, are the same and only depend on the number of customers in the ith queue, i.e., n_i, and do not depend on the number of customers in the other queues, i.e., n_j, $j \neq i$, $j = 1, \ldots, s$. This is similar to an earlier result for the case of a queue of type 1 as implied by Lemma 3.3. Similarly, this result will help us derive a closed-form solution for the loss rate function for the case of a queue of type 2, i.e., when customers have deadlines until the end of their service. More specifically, let

$$\gamma^2_i(n) = \begin{cases} n\frac{\Phi^i_{n-1}(\mu_i)}{\Phi^i_n(\mu_i)} - \mu_i, & \text{if } n > 0, \\ 0, & \text{if } n = 0. \end{cases} \tag{3.53}$$

We have

Theorem 3.2 *Let* $\Gamma^2_i(.)$ *be defined as in* (3.2), *when the* i*th queue is type* 2, *and* $\gamma^2_i(.)$ *is defined as in* (3.53). *Then* $\Gamma^2_i(\mathbf{n}) = \gamma^2_i(n_i)$.

Proof Consider a scenario where the system is in the long run and the state is represented by a vector $\mathbf{n}$ for $n_i > 0$. Suppose that a customer departs from the ith queue in this system. Using the definition of $\Gamma^2_i(\mathbf{n})$ as in (3.2), the probability that this latter departure is successful is simply

$$\frac{\mu_i}{\mu_i + \Gamma^2_i(\mathbf{n})}.$$

On the other hand, the same probability may also be derived as

$$P\big(S^i_{\mathbf{n}-\mathbf{e}_i} \le \theta^i_1\big),$$

where θ_1^i represents the relative deadline of the departing customer from the ith queue. Thus, we have

$$P\big(S_{\mathbf{n}-\mathbf{e}_i} \le \theta_1^i\big) = \frac{\mu_i}{\mu_i + \Gamma_i^2(\mathbf{n})}. \tag{3.54}$$

Using Lemma 3.8, we finally get

$$\begin{aligned} &\Gamma_i^2(\mathbf{n}) = 0, \quad \text{if } n_i = 0, \\ &\Gamma_i^2(\mathbf{n}) = n_i \frac{\Phi_{n_i-1}^i(\mu_i)}{\Phi_{n_i}^i(\mu_i)} - \mu_i, \quad \text{if } n_i > 0, \end{aligned} \tag{3.55}$$

which completes the proof. □

4 Analytical models

Consider the system described in Sect. 1 where queues of both type 1 and type 2 may be used. We now prove a key result, namely, that the state process of the system in the long run converges in distribution to a well-defined Markov process. Let $\mathbf{n} = (n_1, \ldots, n_s)$ be a s-tuple of natural numbers. Define the following notations:

$$\begin{aligned} \Omega_{\mathbf{n}} &= \big\{(j_1, \ldots, j_s) : j_i = 0, 1, \ldots, n_i, i = 1, \ldots, s\big\}; \\ q(\mathbf{n}) &= \{i : n_i > 0, i = 1, \ldots, s\}; \\ |\mathbf{n}| &= \sum_{i=1}^{s} n_i; \\ \mathbf{0} &= (0, \ldots, 0); \\ \mathbf{K} &= (K_1, \ldots, K_s). \end{aligned}$$

Let

$$\begin{aligned} p(t; \mathbf{n}) \equiv{}& \text{the probability that there are } n_i \text{ customers in the } i\text{th queue, } i = 1, \ldots, s, \\ & \text{at time } t. \end{aligned} \tag{4.1}$$

Considering the limiting state behavior of the system during $[t, t + \Delta]$ as $\Delta \to 0$, we can write

$$\begin{aligned} \frac{dp(t; \mathbf{0})}{dt} &= -\lambda(\mathbf{0})p(t; \mathbf{0}) + \sum_{i=1}^{s}\big(\mu_i + \Gamma_i(t, \mathbf{e}_i)\big)p(t; \mathbf{e}_i), \\ \frac{dp(t; \mathbf{n})}{dt} &= \sum_{i \in q(\mathbf{n})} \lambda(\mathbf{n} - \mathbf{e}_i) r_{\mathbf{n}-\mathbf{e}_i}(i)\, p(t; \mathbf{n} - \mathbf{e}_i) \\ &\quad - \left(\lambda(\mathbf{n}) + \sum_{i=1}^{s}\big(\min(m_i, n_i)\mu_i + \Gamma_i(t, \mathbf{n})\big)\right)p(t; \mathbf{n}) \\ &\quad + \sum_{i=1}^{s}\big(\min(m_i, n_i + 1)\mu_i + \Gamma_i(t, \mathbf{n} + \mathbf{e}_i)\big)p(t; \mathbf{n} + \mathbf{e}_i), \\ &\quad \text{if } |\mathbf{n}| > 0 \text{ and } \mathbf{n} \in \Omega_{\mathbf{K}} - \{\mathbf{K}\}, \end{aligned} \tag{4.2}$$

where $\min(m,n)$ represents the minimum of m and n, $\Gamma_i(t,\mathbf{n})$ is defined as in (3.1), and $r_{\mathbf{n}}(i)$ is the probability that an incoming customer in state $\mathbf{n}$ is assigned to the ith queue. A sufficient condition for the statistical stability of the above system of equations in the long run may be given as

$$\exists \mathbf{k} \text{ such as } \forall \mathbf{n} > \mathbf{k} \left[\frac{\lambda(\mathbf{n})}{\sum_{i=1}^{s} (m_i \mu_i + \gamma_i(n_i+1))} < 1 \right]. \tag{4.3}$$

where $\mathbf{k}, \mathbf{n} \in \Omega_{\mathbf{K}}$ and $\gamma_i(.)$ is defined as in (3.31) or (3.53), depending on whether the ith queue is type 1 or type 2, respectively. It is interesting to note that (4.3) will be satisfied if the system is finite; i.e., $|\mathbf{K}| < \infty$, or if the overall arrival process is Poisson, i.e., $\lambda(\mathbf{n})$ is independent of $\mathbf{n}$. Assuming statistical equilibrium, let

$$p(\mathbf{n}) = \lim_{t \to \infty} p(t;\mathbf{n}). \tag{4.4}$$

Using Theorems 3.1 and 3.2, we have

$$\begin{aligned}
0 &= -\lambda(\mathbf{0}) p(\mathbf{0}) + \sum_{i=1}^{s} \big(\mu_i + \gamma_i(1)\big) p(\mathbf{e}_i), \\
0 &= \sum_{i \in q(\mathbf{n})} \lambda(\mathbf{n} - \mathbf{e}_i) r_{\mathbf{n}-\mathbf{e}_i}(i)\, p(\mathbf{n} - \mathbf{e}_i) \\
&\quad - \left(\lambda(\mathbf{n}) + \sum_{i=1}^{s} \big(\min(m_i, n_i)\mu_i + \gamma_i(n_i)\big) \right) p(\mathbf{n}) \\
&\quad + \sum_{i=1}^{s} \big(\min(m_i, n_i+1)\mu_i + \gamma_i(n_i+1)\big) p(\mathbf{n} + \mathbf{e}_i), \\
&\quad \text{if } |\mathbf{n}| > 0 \text{ and } \mathbf{n} \in \Omega_{\mathbf{K}} - \{\mathbf{K}\}.
\end{aligned} \tag{4.5}$$

We also note that

$$\sum_{\mathbf{n} \in \Omega_{\mathbf{K}}} p(\mathbf{n}) = 1. \tag{4.6}$$

The system of equations in (4.5) and (4.6) can uniquely be solved for $p(\mathbf{n})$ using standard solution techniques. Accordingly, the probability of missing deadline at the ith queue (α_d^i) can be derived as

$$\alpha_d^i = \frac{\sum_{\mathbf{n} \in \Omega_{\mathbf{K}}} \gamma_i(n_i)\, p(\mathbf{n})}{\sum_{\mathbf{n} \in \Omega_{\mathbf{K}}} \lambda(\mathbf{n})\, p(\mathbf{n})}, \tag{4.7}$$

where $\gamma_i(.)$ is defined as in (3.31) or (3.53), depending on whether the ith queue is type 1 or type 2, respectively. The probability of missing deadline in the system (α_d)

may also be obtained as

$$\alpha_d = \sum_{i=1}^{s} \alpha_d^i. \tag{4.8}$$

Let $\pi(\mathbf{n})$ represent the steady-state probability that an incoming customer finds the state of the system to be $\mathbf{n}$ in the long run. Thus, we have

$$\pi(\mathbf{n}) = \frac{\lambda(\mathbf{n})p(\mathbf{n})}{\sum_{\mathbf{j}\in\Omega_\mathbf{K}} \lambda(\mathbf{j})p(\mathbf{j})}. \tag{4.9}$$

When the ith queue has a finite capacity (i.e., $K_i < \infty$), where blocking is allowed, the probability of blocking at the ith queue (α_b^i) may be derived as

$$\alpha_b^i = \sum_{\substack{\mathbf{n}\in\Omega_\mathbf{K} \\ n_i = K_i}} \pi(\mathbf{n}) r_\mathbf{n}(i). \tag{4.10}$$

The probability of blocking in the system (α_b) can be given as

$$\alpha_b = \sum_{i=1}^{s} \alpha_b^i. \tag{4.11}$$

The time average fraction of customers joining the ith queue in the long run (ρ_i) can be written as

$$\rho_i = \sum_{\mathbf{n}\in\Omega_\mathbf{K}} \pi(\mathbf{n}) r_\mathbf{n}(i). \tag{4.12}$$

Let us first assume that the ith queue is type 1 (i.e., the customers in the ith queue have deadlines until the beginning of their service). The probability density function of the virtual waiting time of the ith queue (as defined in (2.1)) may be found as

$$f_{U^i}(\tau) = \frac{1}{\rho_i} \sum_{\mathbf{n}\in\Omega_\mathbf{K}} \pi(\mathbf{n}) r_\mathbf{n}(i) f_{U_\mathbf{n}^i}(\tau), \tag{4.13}$$

where $f_{U_\mathbf{n}^i}(\tau)$ is obtained from Lemma 3.3. Suppose next that the ith queue is type 2 (i.e., the customers in the ith queue have deadlines until the end of their service). Then, the probability density function of the offered sojourn time of the ith queue (as defined in (2.4)) may be derived as

$$f_{V^i}(\tau) = \frac{1}{\rho_i} \sum_{\mathbf{n}\in\Omega_\mathbf{K}} \pi(\mathbf{n}) r_\mathbf{n}(i) f_{V_\mathbf{n}^i}(\tau), \tag{4.14}$$

where $f_{V_\mathbf{n}^i}(\tau)$ is obtained from Lemma 3.7.

We now consider the case where all parallel queues are of the same type (i.e., all customers in the system have either deadlines until the beginning of the service or

deadlines until the end of service). When all queues are type 1, we have

$$f_U(\tau) = \sum_{i=1}^{s} \sum_{\mathbf{n}\in\Omega_{\mathbf{K}}} \pi(\mathbf{n}) r_{\mathbf{n}}(i) f_{U_{\mathbf{n}}^i}(\tau), \tag{4.15}$$

where $f_U(\tau)$ is the probability density function of the virtual waiting time of the system (as defined in (2.10)) and $f_{U_{\mathbf{n}}^i}(\tau)$ is obtained from Lemma 3.3. Finally, when all queues are type 2, we get

$$f_V(\tau) = \sum_{i=1}^{s} \sum_{\mathbf{n}\in\Omega_{\mathbf{K}}} \pi(\mathbf{n}) r_{\mathbf{n}}(i) f_{V_{\mathbf{n}}^i}(\tau), \tag{4.16}$$

where $f_V(\tau)$ is the probability density function of the offered sojourn time of the system (as defined in (2.11)) and $f_{V_{\mathbf{n}}^i}(\tau)$ is obtained from Lemma 3.7.

For the case of a finite capacity model (i.e., $K_i < \infty, i = 1, \ldots, s$) where incoming customers are blocked only when all the queues are full, the probability of blocking (α_b), defined (in (2.8)) as the probability that an incoming customer is denied entering the system in the long run due to full queue, can be given as

$$\alpha_b = \frac{\lambda(\mathbf{K}) p(\mathbf{K})}{\sum_{\mathbf{n}\in\Omega_{\mathbf{K}}} \lambda(\mathbf{n}) p(\mathbf{n})}. \tag{4.17}$$

Similarly, the probability of loss (α), defined (in (2.9)) as the probability that an incoming customer is lost due to missing deadline or blocking in the long run, may be obtained as

$$\alpha = \alpha_d + \alpha_b, \tag{4.18}$$

where α_d and α_b are given as in (4.8) and (4.17), respectively.

5 Numerical results

We now study simple examples to illustrate the efficacy of the modeling approach proposed in this paper. We consider two parallel single-sever queues (i.e., $m_1 = m_2 = 1$) with finite capacities $K_1 = 5$ and $K_2 = 4$, and a Poisson arrival process of constant rate λ. Two types of customer impatience are considered: deterministic and exponentially distributed. They are referred to as type I and II customer impatience, respectively. We assume that the mean relative deadline $\overline{\theta} = 1$ for both types of customer impatience, where $\overline{\theta}$ is normalized with respect to $1/\mu_2$. We also consider two kinds of stationary policies for assigning incoming customers to parallel queues. One is the policy of joining shortest non-full queue (SNQ), which is a dynamic (state-dependent) policy that assigns an incoming customer to the shortest non-full queue. The other is the policy of joining each parallel queue with equal probability (RANDOM), which is simply a static (state-independent) policy. Figures 1 and 2 represent the probability of loss (α) for RANDOM and SNQ policies when we have deterministic (type I) customer impatience, $\mu_1 = 2$ and $\mu_2 = 1$, for the cases of deadlines until

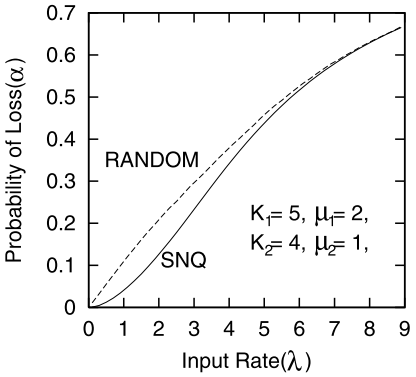

Fig. 1 Probability of loss for deterministic customer impatience and the case of deadlines until the beginning of service

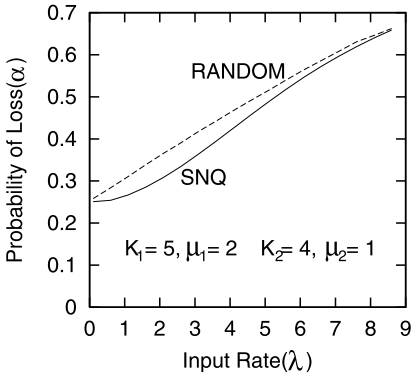

Fig. 2 Probability of loss for deterministic customer impatience and the case of deadlines until the end of service

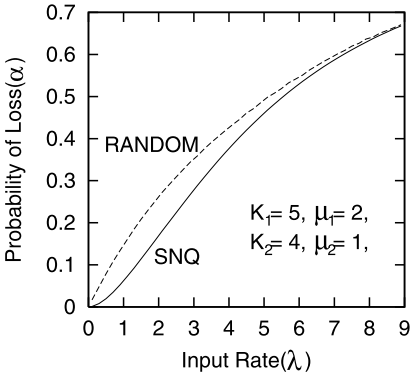

Fig. 3 Probability of loss for exponential customer impatience and the case of deadlines until the beginning of service

the beginning of service (i.e., both parallel queues are type 1) and deadlines until the end of service (i.e., both parallel queues are type 2), respectively. Figures 3 and 4 depict similar results for the case of exponential (type II) customer impatience. It is shown that the probability of loss for SNQ policy is always smaller than RANDOM policy. Figures 5 and 6 represent the probability of missing deadline (α_d) for deterministic (type I) and exponential (type II) customer impatience when we have SNQ policy and $\mu_1 = \mu_2 = 1$, for the cases of deadlines until the beginning of service (i.e., both parallel queues are type 1) and deadlines until the end of service (i.e., both

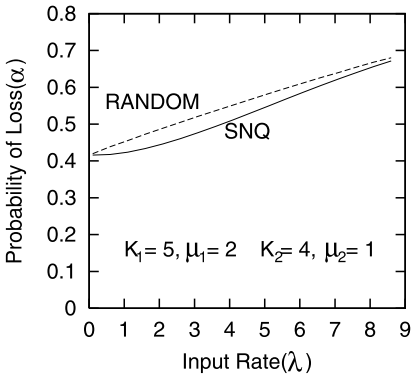

Fig. 4 Probability of loss for exponential customer impatience and the case of deadlines until the end of service

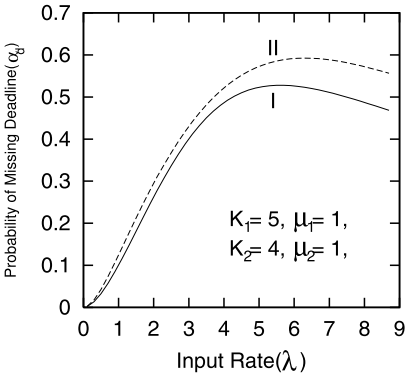

Fig. 5 Probability of missing deadline for SNQ policy and the case of deadlines until the beginning of service

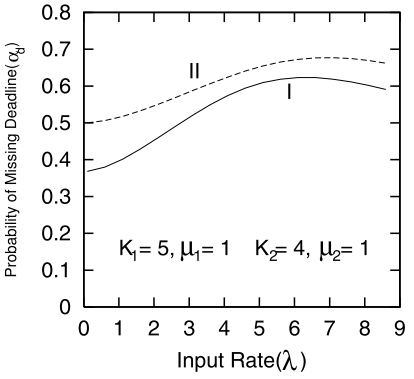

Fig. 6 Probability of missing deadline for SNQ policy and the case of deadlines until the end of service

parallel queues are Type 2), respectively. Figures 7 and 8 depict similar results when the performance measure of interest is the probability of blocking of a customer in the system (α_b). It is shown that the probability of missing deadline for deterministic (type I) customer impatience is always smaller than exponential (type II) customer impatience. However, for the probability of blocking, the opposite is true; that is, the probability of blocking for exponential (type II) customer impatience is always

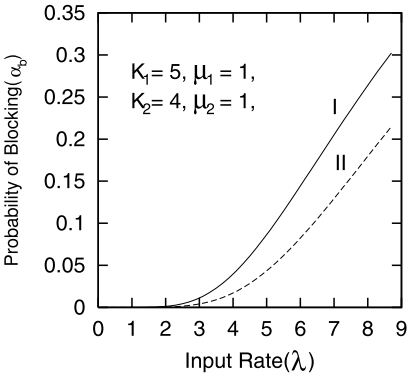

Fig. 7 Probability of blocking for SNQ policy and the case of deadlines until the beginning of service

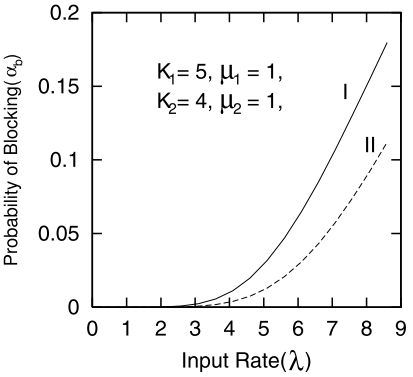

Fig. 8 Probability of blocking for SNQ policy and the case of deadlines until the end of service

smaller than deterministic (type I) customer impatience. Figures with odd numbers have also been used in [13].

References

1. Altman, E., Yechiali, U.: Analysis of customers' impatience in queues with server vacations. Queueing Syst. **52**, 261–279 (2006)
2. Altman, E., Yechiali, U.: Infinite-server queues with system's additional tasks and impatient customers. Probab. Eng. Inf. Sci. **22**, 477–493 (2008)
3. Baccelli, F., Boyer, P., Hebuterne, G.: Single-server queues with impatient customers. Adv. Appl. Probab. **16**, 887–905 (1984)
4. Barrer, D.Y.: Queueing with impatient customers and ordered service. Oper. Res. **5**, 650–656 (1957)
5. Bernat, G., Burns, A., Llamosi, A.: Weakly hard real-time systems. IEEE Trans. Comput. **50**(4), 308–321 (2001)
6. Brandt, A., Brandt, M.: On the $M(n)/M(n)/s$ queue with impatient calls. Perform. Eval. **35**, 1–18 (1999)
7. Brandt, A., Brandt, M.: Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + \mathrm{GI}$ system. Queueing Syst. **41**, 73–94 (2002)
8. Cohen, W.: Single server queue with uniformly bounded virtual waiting time. J. Appl. Probab. **5**, 93–122 (1968)
9. Daley, D.: General customer impatience in queue GI/G/1. J. Appl. Probab. **2**, 186–205 (1965)
10. Gnedenko, B., Kovalenko, I.: Introduction to Queueing Theory. Israel Program for Scientific Translations, Jerusalem (1968)

11. Melamed, B., Whitt, W.: On arrivals that see time averages. Oper. Res. **38**(1), 156–172 (1990)
12. Movaghar, A.: On queueing with customer impatience until the beginning of service. Queueing Syst. **29**, 337–350 (1998)
13. Movaghar, A.: On dynamic assignment of impatient customers to parallel queues. In: IEEE International Computer Performance and Dependability Symposium, vol. 29, San Fransisco, CA, June 2003, pp. 751–759 (2003)
14. Perel, N., Yechiali, U.: Queues with slow servers and impatient customers. Eur. J. Oper. Res. **201**, 247–258 (2010)
15. van Doorn, E., Regterschot, J.: Conditional PASTA. Oper. Res. Lett. **7**, 229–232 (1988)
16. Yechiali, U.: Queues with system disasters and impatient customers when system is down. Queueing Syst. **56**, 195–202 (2007)