Adv. Appl. Prob. **48,** 13–31 (2016)
*doi:*10.1017/apr.2015.4
© *Applied Probability Trust* 2016

QUEUES WITH ADVANCED RESERVATIONS: AN INFINITE-SERVER PROXY FOR THE BOOKINGS DIARY

R. J. MAILLARDET * ** AND
P. G. TAYLOR,* *** *University of Melbourne*

Abstract

Queues with advanced reservations are endemic in the real world. In such a queue, the 'arrival' process is an incoming stream of customer 'booking requests', rather than actual customers requiring immediate service. We consider a model with a Poisson booking request process with rate λ. Associated with each request is a pair of independent random variables (R_i, S_i) constituting a request for service over a period S_i, starting at a time R_i into the future. Our interest is in the probability that a customer will be rejected due to capacity constraints. We present a simulation of a finite-capacity queue in which we record the proportion of rejected customers, and then move to an analysis of a queue with infinitely-many servers. Obviously no customers are rejected in the latter case. However, the event that the arrival of the extra customer will cause the number of customers in the queue to exceed C at some point during its service can be used as a proxy for the event that the customer would have been rejected in a system with finite capacity C. We start by calculating the transient and stationary distributions for some performance measures for the infinite-server queue. By observing that the stationary measure for the bookings diary (that is, the list of customers currently on hand, together with their start times and service times) is the same as the law for the entire sample path of an infinite server queue with a specified nonhomogenous Poisson input process, which we call the *bookings queue*, we are able to write down expressions for the abovementioned probability that, at some time during a requested service, the number of customers exceeds C. This measure serves as a bound for the probability that an incoming arrival would be refused admission in a system with C servers and, for a well-dimensioned system, it is to be hoped that it is a good approximation. We test the quality of this approximation by comparing our analytical results for the infinite-server case against simulation results for the finite-server case.

Keywords: Advanced reservations; infinite-server queue; blocking probability

2010 Mathematics Subject Classification: Primary 60K25
Secondary 68M20; 90B22

1. Introduction

Reservations are an inherent feature of many real-world queuing systems. They are used to manage hotel and restaurant bookings, medical appointments, and operating theatre schedules, where there is a need for certainty of service at some period in the future. They have also been proposed to facilitate the operation of various telecommunications systems, especially optical burst networks [3], [4], [11], [14], [15], [23]. Over the years, various authors have looked at

Received 23 May 2014; revision received 18 December 2014.
* Postal address: Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia.
** Email address: rjmail@unimelb.edu.au
*** Email address: taylorpg@unimelb.edu.au

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

queueing systems with advanced reservation. However, despite their ubiquity, these systems have not received as much attention as might be expected.

Motivated by telecommunications applications, Liang *et al*. [17] proposed a slotted-time model for a queue with advanced reservations, and used it to derive some approximate and simulated results. Later, Kaheel *et al*. [14], [15] applied a similar analysis to optical burst switching networks, in which a header packet precedes a burst and reserves capacity for it. By assuming that calls making advanced reservations do so far ahead of time, Greenberg *et al*. [13] used a separation of timescales approach to approximate the blocking probability of a stream of calls making advanced reservations, as well as a stream of calls requesting immediate service.

Virtamo [22] and Coffman *et al*. [7]–[9] adopted a different approach. They analysed an interval-packing model, in which reservations of varying duration arrive to fill up space in an interval on the real line. The main measure of interest in these papers is the 'reservation probability' that a particular point is covered by a reservation, either in the transient case, or in the limiting case when all possible space is filled.

In the context of optical burst switching, a number of authors have analysed systems with a finite number of classes, each with a fixed reservation offset. Dolzer and Gauger [11] used a 'conservation law' to give some approximate basic formulae for the rejection probabilities of each class, while Barakat and Sargent [3], [4] defined the concept of a 'contention window' to obtain an exact expression for the blocking probability of the class with the largest reservation offset, and Vu and Zukerman [23] proposed an approximating M/G/k/k model to derive blocking probabilities for each class.

A feature of the work of [3] and [4] was that the authors studied an infinite-server system, and approximated the blocking probability of a call arriving to a system with finite-capacity C by the probability that it would cause the occupancy of the infinite-server system to exceed C at some point in time during its duration. We adopt a similar infinite-server approximation for the blocking probability in this paper.

For a simple system with a single server, van de Vrugt *et al*. [21] made a number of observations. In particular, the authors identified classes of queues where the advanced reservation increases the blocking probabilities and other classes where it decreases them.

Recently, in two heavy-traffic regimes: the critically-loaded and Halfin–Whitt regimes, Levi and Shi [16] studied methods of revenue management in queues with advanced reservation that could involve rejecting customers who are likely to tie up resources that will be required by more profitable customers. In order to bound and approximate the blocking probabilities, the authors also assumed that the queue has infinite capacity and calculated the probability that the number of customers exceeds C at some point during the service of the arriving customer.

In this paper we shall consider a model for a reservation queue that has the following characteristics.

1. The booking process is Poisson with rate λ.

2. Associated with each booking is a pair of random variables (R_i, S_i) constituting a request for service over a period S_i starting at a time R_i into the future. The random variables $\{R_i\}$ are independent and identically distributed, and independent of the sequence $\{S_i\}$, which are also independent and identically distributed. We denote the distribution functions of R_i and S_i by F_R and F_S, and assume that they have finite means $\mathbb{E}(R) = \xi$ and $\mathbb{E}(S) = \eta$, respectively.

3. There is a fixed number C of servers, which may be finite or infinite. If C is finite, a customer arriving at time t and requesting service over the time interval $[t + r, t + r + s)$

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

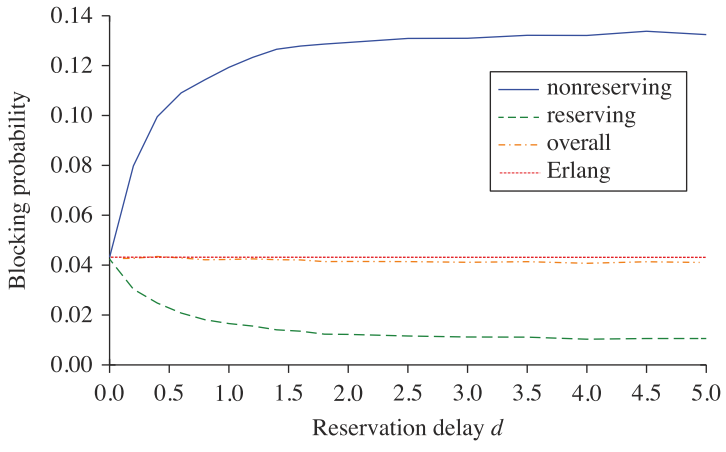

FIGURE 1: Simulated blocking probabilities for the reservation queue as a function of d. Service times are exponential with $\eta = 1$. Other parameters are $C = 10$, $\lambda = 6$, and $\gamma = \frac{1}{4}$.

is admitted if and only if the number of customers booked at time t is less than or equal to $C - 1$ for the entire time interval.

With respect to the third item above, our real interest is in calculating the blocking probability in systems where the number C of servers is finite. However, it appears that analysis of such queues is extremely difficult, if not intractable. As a consequence, apart from our initial simulation reported below, we shall follow the lead of Barakat and Sargent [3], [4] and Levi and Shi [16], and analyse an infinite-server model, calculating the probability that admission of an arriving customer would result in a given capacity C being exceeded at some point during its service time. A coupling argument can be used to establish that this is an upper bound for the blocking probability in an actual system and, for a well-dimensioned system in which the blocking probability is low, we would expect that it will constitute a reasonable approximation.

In our model, requested booking start times occur according to a Poisson process with rate λ (since they are independent and identically distributed translations of the request process using the reservation distribution) but, due to the different nature of the blocking mechanism, the blocking probability is not given by the well-known Erlang loss formula.

By way of motivation, we start with some simulation results for a finite capacity ($C = 10$) model with a simple discrete reservation distribution: a proportion $\gamma = \frac{1}{4}$ of arriving customers request immediate service and a proportion $1 - \gamma = \frac{3}{4}$ request service commencing at exactly d time units into the future. We call these customers *nonreserving* and *reserving*, respectively. Arrivals occur in a Poisson process with $\lambda = 6$, and the service time is taken to be either exponential or deterministic, in both cases with mean $\eta = 1$.

In Figures 1 and 2 we illustrate how the simulated blocking probabilities vary with d for the cases of exponential service and deterministic service, respectively. Plotted are the proportions of blocked nonreserving customers and reserving customers, together with the overall proportion of blocked customers as a function of d. The blocking probability given by the Erlang loss formula is shown as a reference line. As d approaches 0, the blocking probability of both nonreserving and reserving customers approaches this value. For large d, the blocking probabilities for both the nonreserving and reserving customers become constant with respect to d.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

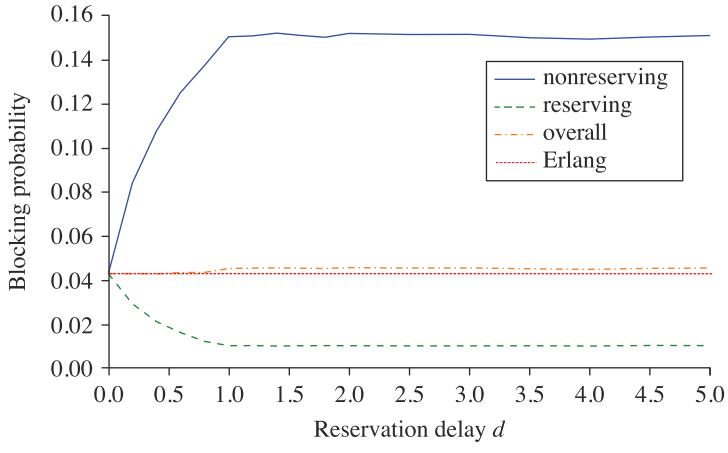

FIGURE 2: Simulated blocking probabilities for the reservation queue as a function of d. Service times are deterministic with $\eta = 1$. Other parameters are $C = 10$, $\lambda = 6$, and $\gamma = \frac{1}{4}$.

Furthermore, there is a degradation of overall system performance when the service times are deterministic, but a slight improvement when service times are exponential. This verifies the observation of [21] that there are circumstances where reservation improves average blocking performance, and circumstances where it does not. For both types of service time distribution, the blocking probabilities for reserving requests are similar for large enough d, and the difference in performance is largely explained by a significantly lower blocking probability for nonreserving requests with exponential service times compared with deterministic service times. As suggested by the authors of [21], we believe that this can be explained by the fact that there are more frequent requests for a short service that could fit in before an existing reservation in the exponential case than in the deterministic case.

For the rest of this paper we shall assume that $C = \infty$. This will enable us to exploit independence properties that are inherent in infinite-server queueing systems. However, it is still nontrivial to derive expressions for the performance measure that we are interested in, the probability that addition of an arriving customer will cause the occupancy of the queue to exceed C at some point.

In Section 2 we derive stationary and transient distributions for some simple performance measures of the infinite-capacity model. This is followed in Section 3 by an analysis of the 'bookings diary' generated by this model, which retains full information on the start and service times for all confirmed bookings. In Section 4 we derive the stationary measure for the 'bookings profile' which represents the variation of total booking load with time, sacrificing unique identification of every booking. Section 5 contains the main observation of this paper, that the stationary distribution of the bookings diary is the same as the law of sample paths of an associated M(t)/G/∞ queue, which we shall call the *bookings queue*. In Section 6 we carry out the analysis for the two examples that were simulated above. We conclude with a conclusion and a discussion of future research directions in Section 7.

2. Distributions of numbers of customers in the $C = \infty$ case

In the case where $C = \infty$, we can obtain transient and stationary distributions for certain performance measures by adapting the approach to analysing an M/G/∞ queue that was used

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

by Foley [12]. Let

- $N_{\mathrm{D}}(t)$ be the total number of diary bookings at time t,
- $N_{\mathrm{A}}(t)$ be the number of active (that is, 'in service') bookings at time t,
- $N_{\mathrm{P}}(t)$ be the number of pending (that is, reserved but not yet active) bookings at time t,
- $N_{\mathrm{A}}(t, x)$ be the number of bookings at time t that will be active at time $t + x$.

All of these quantities are of the general form '$N_\Gamma(t)$ is the number of bookings meeting condition Γ from requests in $[0, t]$'.

The number $N(t)$ of booking requests received by time t is distributed according to a Poisson distribution with parameter λt. If $N(t) = n$ then it is well known (see, for example, [5, Theorem 5.2]) that the n arrival times are uniformly and independently distributed on $[0, t]$. So each of the n booking requests has probability

$$p(t) = \frac{1}{t}\int_0^t \mathbb{P}(\text{the booking meets condition } \Gamma \mid \text{it arrives at } u)\,\mathrm{d}u$$

of meeting condition Γ, independently of the other requests. Hence,

$$N_\Gamma(t) \stackrel{\mathrm{D}}{=} \text{Binomial}(N(t), p(t)),$$

where $N(t) \stackrel{\mathrm{D}}{=} \text{Poisson}(\lambda t)$. It follows that

$$N_\Gamma(t) \stackrel{\mathrm{D}}{=} \text{Poisson}(\nu(t)),$$

where $\nu(t) = \lambda p(t)$.

With F_{R+S} the convolution of F_R and F_S, the specific processes that we defined earlier have functions $\nu(t)$ given in Table 1. Letting $t \to \infty$, we obtain the fact that, for each of the random variables discussed above,

$$\lim_{t\to\infty} \mathbb{P}(N_\Gamma(t) \le n) = \mathbb{P}(N_\Gamma \le n),$$

where N_Γ has a Poisson distribution with parameter ν given in Table 2. Note that the limiting distributions for $N_{\mathrm{D}}(t)$, $N_{\mathrm{A}}(t)$, and $N_{\mathrm{P}}(t)$ are insensitive to the form of the distributions F_R and F_S.

TABLE 1: Parameters for $N_\Gamma(t)$.

Name	Type	$\nu(t)$
$N_{\mathrm{D}}(t)$	Diary	$\lambda\int_0^t (1 - F_{R+S}(t-u))\,\mathrm{d}u$
$N_{\mathrm{A}}(t)$	Active	$\lambda\int_0^t\int_0^{t-u} (1 - F_S(t-u-v))\,\mathrm{d}F_R(v)\,\mathrm{d}u$
$N_{\mathrm{P}}(t)$	Pending	$\lambda\int_0^t (1 - F_R(t-u))\,\mathrm{d}u$
$N_{\mathrm{A}}(t, x)$	Hits $t + x$	$\lambda\int_0^t\int_0^{t+x-u} (1 - F_S(t+x-u-v))\,\mathrm{d}F_R(v)\,\mathrm{d}u$

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

TABLE 2: Parameters for N_Γ.

Name	Type	ν
$N_{\mathrm{D}}(\cdot)$	Diary	$\lambda(\eta+\xi)$
$N_{\mathrm{A}}(\cdot)$	Active	$\lambda\eta$
$N_{\mathrm{P}}(\cdot)$	Pending	$\lambda\xi$
$N_{\mathrm{A}}(\cdot, x)$	Hits future time x	$\lambda\int_x^\infty\int_0^u (1-F_S(u-v))\,\mathrm{d}F_R(v)\,\mathrm{d}u$

Using arguments similar to those above, we can derive the fact that the distribution at time t of the number of customers already booked at time t whose service intersects with an interval of the form $[t+r, t+r+s)$ is also Poisson. The parameter of this distribution is

$$\psi(t,r,s) = \lambda\int_0^t\int_0^{t+r-u}(1-F_S(t+r-u-v))\,\mathrm{d}F_R(v)\,\mathrm{d}u + \lambda\int_0^t [F_R(t+r+s-u)-F_R(t+r-u)]\,\mathrm{d}u.$$

The first term in this expansion takes into account customers who arrive at time $u \in [0,t]$, whose reservation time expires at time $u+v \in [0, t+r]$ and who are still being served at time $t+r$, while the second term takes into account customers who arrive at time $u \in [0,t]$ and whose reservation time expires during the interval $[t+r, t+r+s]$. Making the substitution $w = t+r-u$ in both integrals, we can write

$$\psi(t,r,s) = \lambda\int_r^{t+r}\int_0^w (1-F_S(w-v))\,\mathrm{d}F_R(v)\,\mathrm{d}w + \lambda\int_r^{t+r}[F_R(w+s)-F_R(w)]\,\mathrm{d}w.$$

Now letting $t\to\infty$, we see that the parameter of the corresponding limiting distribution is

$$\psi(r,s) = \lambda\int_r^\infty\int_0^w (1-F_S(w-v))\,\mathrm{d}F_R(v)\,\mathrm{d}w + \lambda\int_r^\infty [F_R(w+s)-F_R(w)]\,\mathrm{d}w.$$

If there is a point in the interval $[t+r, t+r+s)$ where C customers are simultaneously present, then there must be at least C customers covering this interval. However, the converse does not hold: there can be more than C customers covering the interval while the maximum at any given time during the interval is strictly less than C. So, the above result gives an upper bound for our infinite-server bound for the blocking probability of a customer arriving at time t with a reservation time r and a service time s in the finite-capacity system. However, this bound is unlikely to be tight, and we would really like a method for calculating the infinite-server bound itself. The rest of this paper is devoted to this derivation.

3. The bookings diary

In this section we analyse a process whose states give a complete characterisation of the 'current' bookings diary. This description includes the start times and service times of active and pending bookings. To define such a state, we use the description

$$(N_{\mathrm{D}}(t), \boldsymbol{Y}(t), \boldsymbol{X}(t)) = (N_{\mathrm{D}}(t), Y_1(t), \ldots, Y_{N_{\mathrm{D}}(t)}(t), X_1(t), \ldots, X_{N_{\mathrm{D}}(t)}(t)), \qquad (1)$$

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

where, as above, $N_{\mathrm{D}}(t)$ is the total number of diary bookings at time t and the individual bookings are allocated labels $j = 1, \dots, N_{\mathrm{D}}(t)$. The random variable $Y_j(t)$ is the requested service time of the customer with label j and $X_j(t)$ is its *residual reservation time*, that is, the time difference between the customer's commencement of service and t. Note that $X_j(t)$ will be negative if customer j has already commenced service, but it must be greater than $-Y_j(t)$, because the customer will depart the system when $X_j(t) = -Y_j(t)$. So, the state space is

$$\{(n, \boldsymbol{y}, \boldsymbol{x}) \colon n \in \mathbb{Z}_+,\, y_j > 0 \text{ and } x_j > -y_j,\ j = 1, \dots, n\}.$$

Consider the situation where the ith customer arrives at time τ_i to find the system in state $(N_{\mathrm{D}}(\tau_i), \boldsymbol{Y}(\tau_i), \boldsymbol{X}(\tau_i))$ with $N_{\mathrm{D}}(\tau_i) = n$ and he/she samples a service time S_i and a reservation time R_i independently from their respective distributions. We allocate the customer a label j chosen uniformly from the numbers 1 to $n+1$ and put its service time $Y_j(\tau_i) = S_i$. We also put its residual reservation time $X_j(\tau_i) = R_i$. The customers that previously had labels $j, \dots, n$ each have their label increased by one, so that their labels are now $j+1, \dots, n+1$, and their service and residual reservation times are relabelled in accordance with this.

Notwithstanding changes of labels, the customer's service time remains constant throughout its stay in the system. However, for $t > \tau_i$, its residual reservation time $X_j(t) = R_i + \tau_i - t$ decreases linearly at unit rate until it is equal to $-Y_j$, at which time the customer has completed service and we remove it from the current bookings diary. When this happens, the labels of customers $j+1, \dots, n$ are each decreased by one, so that their new labels are $j, \dots, n-1$, again with their service and residual reservation times relabelled in accordance.

Assuming that the queue starts empty, the following theorem gives an expression for the law of the bookings diary at time t. In it, we use $\prod_{j=1}^n (a_j, b_j]$ to denote the Cartesian product of intervals in $\mathbb{R}^n$.

Theorem 1. *Assume that the bookings diary starts empty at time* 0. *For time $t > 0$, $n = 0, 1, \dots$, $\mathcal{Y} = \prod_{j=1}^n (0, y_j]$, and $\mathcal{X} = \prod_{j=1}^n (-y_j, x_j]$, let $\pi(n, \mathcal{Y}, \mathcal{X}, t)$ be the probability that there are n customers in the bookings diary at time t with $Y_j(t) \in (0, y_j]$ and $X_j(t) \in (-y_j, x_j]$. Then*

$$\pi(n, \mathcal{Y}, \mathcal{X}, t) = \pi(0, t) \frac{\lambda^n}{n!} \prod_{j=1}^n \left[\int_0^{y_j} \int_{\max(-w_j, -t)}^{x_j} \int_{\max(0, v_j)}^{v_j + t} F_R(\mathrm{d}u_j)\, \mathrm{d}v_j\, F_S(\mathrm{d}w_j) \right], \tag{2}$$

where

$$\pi(0, t) = \exp\left(-\lambda \int_0^t (1 - F_{R+S}(t-u))\, \mathrm{d}u \right). \tag{3}$$

Proof. As in Section 2 we adapt the approach of Foley [12], using the fact that the number $N(t)$ of booking requests received by time t is distributed according to a Poisson distribution with parameter λt, and that, if $N(t) = n$, the n arrival times are uniformly and independently distributed on $[0, t]$.

A customer who arrived at time τ with a requested service time Y and reservation time R has a residual reservation time $X(t)$ at time t equal to $R + \tau - t$ if $R + \tau - t > -Y$, and is no longer recorded in the bookings diary otherwise. Such a customer could have arrived at any time during the interval $[0, t]$ if $X(t)$ is nonnegative, but if $X(t)$ is negative the customer must have arrived at least $|X(t)|$ before time t. We conclude that a customer can arrive at any time in the interval $[0, t + X(t)]$ if $X(t)$ is negative. Note that this means that $X(t) \geq -t$.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

So, conditional on $Y = w$,

$$\mathbb{P}(X(t) \in (-w, x]) = \int_{\max(-w,-t)}^{x} \int_{0}^{t+\min(0,v)} \frac{1}{t}\, \mathrm{d}F_R(t + v - \tau)\, \mathrm{d}v,$$

independently of the service and residual reservation times of other customers. Making the substitution $u = t + v - \tau$ in the inner integral, we see that this expression reduces to

$$\mathbb{P}(X(t) \in (-w, x]) = \int_{\max(-w,-t)}^{x} \int_{\max(0,v)}^{v+t} \frac{1}{t}\, \mathrm{d}F_R(u)\, \mathrm{d}v.$$

Integrating with respect to the distribution of the requested service time Y, we see that the probability that a given customer with label j who arrived in the interval $[0, t]$ has $Y_j \in (0, y_j]$ and $X_j(t) \in (-y_j, x_j]$ is

$$p(t) = \frac{1}{t} \int_{0}^{y_j} \int_{\max(-w_j,-t)}^{x_j} \int_{\max(0,v_j)}^{v_j+t} F_R(\mathrm{d}u_j)\, \mathrm{d}v_j\, F_S(\mathrm{d}w_j).$$

The independence of the arrival times, given that $N(t) = n$, the reservation times R_i and the service times S_i and the random allocation of labels means that the probability that $Y_j \in (0, y_j]$ and $X_j(t) \in (-y_j, x_j]$ for all $j = 1, \ldots, n$ is

$$\frac{1}{t^n} \prod_{j=1}^{n} \int_{0}^{y_j} \int_{\max(-w_j,-t)}^{x_j} \int_{\max(0,v_j)}^{v_j+t} F_R(\mathrm{d}u_j)\, \mathrm{d}v_j\, F_S(\mathrm{d}w_j).$$

Removing the conditioning on n by observing that the number of arrivals in $[0, t]$ is a Poisson random variable with parameter λt gives the result.

The fact that $\pi(0, t)$ is given by (3) follows immediately from the first line of Table 1. However, it can also be established by showing that

$$\int_{0}^{\infty} \int_{\max(-w,-t)}^{\infty} \int_{\max(0,v)}^{v+t} F_R(\mathrm{d}u)\, \mathrm{d}v F_S(\mathrm{d}w) = \int_{0}^{t} (1 - F_{R+S}(t - u))\, \mathrm{d}u.$$

This can be achieved via a somewhat tedious series of integral substitutions and uses of Fubini's theorem that we choose not to detail here. □

Letting $t \to \infty$ in the transient measure (2), we can derive a limiting measure for the bookings diary.

Corollary 1. *Assume that the bookings diary starts empty at time* 0. *For* $n = 0, 1, \ldots$, $\mathcal{Y} = \prod_{j=1}^{n}(0, y_j]$, *and* $\mathcal{X} = \prod_{j=1}^{n}(-y_j, x_j]$, *let* $\pi(n, \mathcal{Y}, \mathcal{X}) = \lim_{t\to\infty} \pi(n, \mathcal{Y}, \mathcal{X}, t)$. *Then*

$$\pi(n, \mathcal{Y}, \mathcal{X}) = \pi(0) \frac{\lambda^n}{n!} \prod_{j=1}^{n} \left[\int_{0}^{y_j} \int_{-w_j}^{x_j} \int_{\max(0,v_j)}^{\infty} F_R(\mathrm{d}u_j)\, \mathrm{d}v_j\, F_S(\mathrm{d}w_j) \right], \tag{4}$$

where $\pi(0) = \exp(-\lambda(\eta + \xi))$.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

4. The bookings profile

The state description (1) that we used to characterise the bookings diary in Section 3 is more detailed than we need to decide on whether the number of customers will exceed C during the requested service interval of an arriving customer. To make this decision, we just need to know the service commencement times and departure times of the customers that are present in the bookings diary. We call such a description the *bookings profile*. An example of a bookings profile is depicted in Figure 3.

Given that $N_{\mathrm{D}}(t) = n$, let $i(j)$ be the position in the arrival sequence of the customer that is labelled j at time t. Then, $\tau_{i(j)}$, $S_{i(j)}$, and $R_{i(j)}$, are the arrival time, service request, and reservation request of the customer that is labelled j at time t. So, as in Section 3, $X_j(t) = \tau_{i(j)} + R_{i(j)} - t$ is the residual reservation time of customer j and $D_j(t) = \tau_{i(j)} + R_{i(j)} + S_{i(j)} - t$ is the remaining time until it departs. Now let $X_{(\ell)}(t)$ be the ℓth order statistic of $(X_1(t), \ldots, X_n(t))$ and $D_{(\ell)}(t)$ be the ℓth order statistic of $(D_1(t), \ldots, D_n(t))$. Then the $X_{(\ell)}(t)$ are the ordered residual reservation times, which we can think of as the service commencement times relative to time t, and $D_{(\ell)}(t)$ the ordered departure times relative to time t. A knowledge of the bookings profile defined by the $X_{(\ell)}(t)$ and $D_{(\ell)}(t)$ is sufficient to decide whether the acceptance of an arriving customer will cause the limit C to be exceeded at some time during its service. It is thus of interest to derive the law of $(N_{\mathrm{D}}(t), X_{(\ell)}(t), D_{(\ell)}(t), \ell = 1, \ldots, N_{\mathrm{D}}(t))$.

An immediate observation is that there can be more than one bookings diary that has the same bookings profile. First, the labelling of customers in the distribution for the bookings diary is arbitrary and each labelling leads to the same bookings profile. Thus, we need to sum the distributions (2) and (4) over all $n!$ possible labellings when $N_{\mathrm{D}}(t) = n$.

The second thing to note is that the bookings profile $(N_{\mathrm{D}}(t), X_{(\ell)}(t), D_{(\ell)}(t))$ defines the starting points and ending points of services, but it does not specify which customer departs at each departure time. That is, it does not match a departure point $D_{(\ell)}(t)$ with the service commencement time of the departing customer. Consider, for example, Figure 4. In the first bookings diary the customer who departs at time 1.2 commenced service at time 0.5 and the customer who departs at time 1.5 commenced service at time 0.6, while these commencement times are interchanged in the second bookings diary. Both bookings diaries have the same bookings profile.

In general, bookings diaries, such as those in Figure 4, that lead to the same bookings profile have different distributions as given by (2) and (4). However, in some cases, this problem does

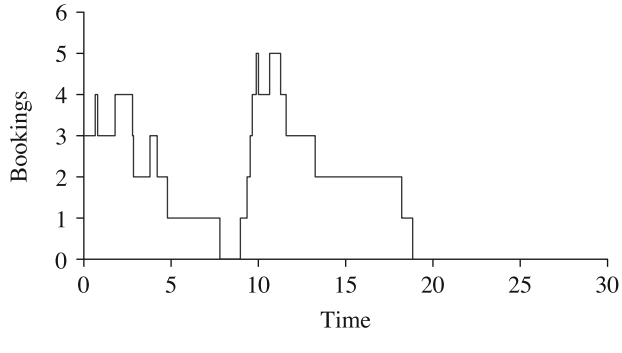

FIGURE 3: A bookings profile.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

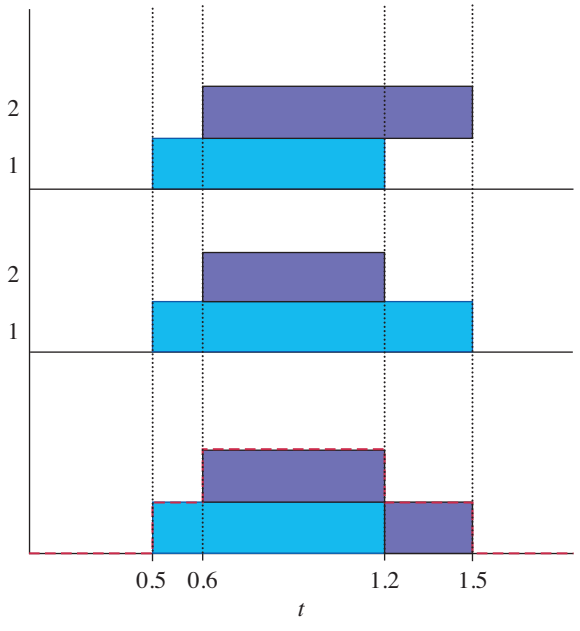

FIGURE 4: Two bookings diaries that give the same bookings profile.

not arise. When the service time distribution is deterministic, the departure time does indeed define the service commencement time, and the phenomenon described in Figure 4 cannot occur.

Furthermore, if the service times are exponential with parameter $\mu = 1/\eta$, then the measures defined by (2) and (4) have densities

$$\phi(n, \boldsymbol{x}, \boldsymbol{y}, t) = \pi(0, t)\frac{(\lambda\mu)^n}{n!}\exp\left(-\mu\sum_{j=1}^{n} y_j\right)\prod_{j=1}^{n}[F_R(x_j + t) - F_R(\max(0, x_j))]$$

and

$$\phi(n, \boldsymbol{x}, \boldsymbol{y}) = \pi(0)\frac{(\lambda\mu)^n}{n!}\exp\left(-\mu\sum_{j=1}^{n} y_j\right)\prod_{j=1}^{n}[1 - F_R(\max(0, x_j))]$$

over the set $E = \{\boldsymbol{y} > \boldsymbol{0}, \boldsymbol{x} > -\boldsymbol{y}\}$. These densities are invariant over states that lead to the same bookings profile, since the total service time is identical for such states.

By realising that whenever a departure occurs, the assumption that service times are exponential means that the departing customer is chosen uniformly from those present just before the departure, we see that the number of bookings diaries that corresponds to a given bookings profile is $\prod_{\ell=1}^{N_D(t)} Q(D_{(\ell)}(t)^-)$, where $Q(D_{(\ell)}(t)^-)$ is the number of customers present just before the ℓth departure in the bookings profile. We arrive at the following corollary.

Corollary 2. *Assume that the reservation distribution is general, but that the service time distribution is exponential with parameter μ, and that the queue starts empty at time 0.*

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

1. *The bookings profile* $(N_{\mathrm{D}}(t), X_{(\ell)}(t), D_{(\ell)}(t))$ *at time* t *has a density*

$$\psi(n, \boldsymbol{x}, \boldsymbol{d}, t) = \pi(0,t)(\lambda\mu)^n \mathrm{e}^{(-\mu\sum_{\ell=1}^n (d_\ell - x_\ell))} \times \prod_{\ell=1}^n Q(d_\ell^-)[F_R(x_\ell + t) - F_R(\max(0, x_\ell))].$$

2. *The limiting bookings profile* $(N_{\mathrm{D}}, X_{(\ell)}, D_{(\ell)})$ *has a density*

$$\begin{aligned}\psi(n, \boldsymbol{x}, \boldsymbol{d}) &= \pi(0)(\lambda\mu)^n \mathrm{e}^{(-\mu\sum_{\ell=1}^n (d_\ell - x_\ell))} \prod_{\ell=1}^n Q(d_\ell^-)[1 - F_R(\max(0, x_\ell))] \\ &= \pi(0)(\lambda\mu)^n \mathrm{e}^{(-\mu\sum_{\ell=1}^n (d_\ell - x_\ell))} \left[\prod_{\ell=1}^n Q(d_\ell^-)\right]\left[\prod_{\ell:\, x_\ell > 0} \int_{x_j}^\infty \mathrm{d}F_R(u_j)\right].\end{aligned}$$

5. The bookings queue

Both Theorem 1 and Corollary 1 are product-form results. They establish that the distribution of the service times and residual reservation times in the bookings diary can be decomposed into a product of the distribution of the service times and residual reservation times of the individual customers.

Considering a single customer, we see that, for $y > 0$ and $x > -y$, the limiting joint distribution of the service time of a customer present in the bookings diary and its residual reservation time is given by

$$F_{X,Y}(x, y) = A \int_0^y \int_{-w}^x \int_{\max(0,v)}^\infty F_R(\mathrm{d}u)\, \mathrm{d}v F_S(\mathrm{d}w). \tag{5}$$

A normalisation argument can be used to show that the constant A in (5) is equal to $(\eta + \xi)^{-1}$.

We can decompose the right-hand side of (5), so that

$$\begin{aligned}F_{X,Y}(x, y) &= A \int_{-y}^x \int_{\max(-v,0)}^y \int_{\max(0,v)}^\infty F_R(\mathrm{d}u) F_S(\mathrm{d}w)\, \mathrm{d}v \\ &= A\left[\int_{-y}^0 \int_{-v}^y \int_0^\infty F_R(\mathrm{d}u) F_S(\mathrm{d}w)\, \mathrm{d}v + \int_0^x \int_0^y \int_v^\infty F_R(\mathrm{d}u) F_S(\mathrm{d}w)\, \mathrm{d}v\right] \\ &= A\left[\int_{-y}^0 \int_{-v}^y F_S(\mathrm{d}w)\, \mathrm{d}v + \int_0^x \int_0^y [1 - F_R(v)] F_S(\mathrm{d}w)\, \mathrm{d}v\right] \\ &= A\left[\eta \int_0^y \frac{[F_S(y) - F_S(v)]}{\eta}\, \mathrm{d}v + \xi F_S(y) \int_0^x \frac{[1 - F_R(v)]}{\xi}\, \mathrm{d}v\right].\end{aligned} \tag{6}$$

Noting that a customer in the bookings diary is currently in service if and only if its residual reservation time is negative, we can think of the first term as characterising the distribution of the service time of an active customer, and the second term as giving the joint distribution of the service time and residual reservation time of a pending customer. From Table 2 we know that the numbers of active and pending customers have Poisson distributions with parameters $\lambda\eta$ and $\lambda\xi$, respectively, a fact that we can also derive from (4). We conclude that a customer in the bookings diary is active with probability $\eta/(\eta + \xi)$ and pending with probability $\xi/(\eta + \xi)$.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

We recognise the integral in the first term in (6) as the *limiting distribution of the spread* of a renewal process with inter-event time distribution F_S (see, for example, [24, p. 67]). It follows that the limiting distribution of an active customer's remaining service time in the bookings diary is

$$F_S^{\mathrm{e}}(w) = \int_0^w \frac{[1 - F_S(u)]}{\eta}\,\mathrm{d}u. \tag{7}$$

So we can conclude that the number of active customers has a Poisson distribution with parameter $\lambda\eta$ and that their remaining service times are chosen independently according to the distribution (7).

Now focussing on the second term of (6), we see that the time until the commencement of service of a pending customer has limiting distribution

$$F_X(x) = \int_0^x \frac{[1 - F_R(v)]}{\xi}\,\mathrm{d}v. \tag{8}$$

Furthermore, the service commencement times of the N_P pending customers are chosen independently according to the distribution (8). It follows from Daley and Vere-Jones [10, Exercise 2.1.6(a)] that the order statistics of these commencement times have the same distribution as the points of a nonhomogeneous Poisson process with intensity

$$\alpha(x) = \lambda[1 - F_R(x)], \tag{9}$$

conditional on there being N_P points in total. Since N_P has a Poisson distribution, the process of service commencement times follows a nonhomogeneous Poisson process with parameter $\alpha(x)$ as defined in (9).

We arrive at the observation that the limiting distribution of the bookings diary is identical to the law of whole sample paths of an *associated M(t)/G/∞ queue*, which we shall call the *bookings queue*. As with any queue, the law of this queue can be specified by giving

- the distribution of the number of customers initially in the queue,
- the distribution of the remaining service times of each of these customers,
- a characterisation of the arrival process, and
- the distribution of the service times of the customers.

Specifically, for the bookings queue,

- the number of customers initially present has a Poisson distribution with parameter $\lambda\eta$,
- the remaining service times of these customers are chosen independently according to F_S^{e} in (7),
- the arrival process is a nonhomogeneous Poisson process with intensity $\alpha(x)$ at time x, as in (9), and
- service times are selected independently from F_S.

Note that the bookings queue has almost surely finitely-many customers in total. In fact, the total number of customers that are ever served in this queue has a Poisson distribution with parameter $\lambda(\eta + \xi)$.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

6. Examples

In the situation where the bookings diary has reached stationarity, we are interested in determining the probability that the addition of a customer who arrives at time t with a reservation request r and service request s would result in the number of customers in the diary exceeding C at some point in the interval $[t+r, t+r+s)$. We approach this by considering the probability that the number of customers in the bookings queue defined in Section 5 is greater than or equal to C at some point in the interval $[r, r+s)$. If this is the case, then the addition of the extra customer will cause the occupancy to exceed C.

Evaluating this probability for general reservation and service time distributions can still be a difficult calculation. Here, we shall carry out the analysis for the two examples that we presented in Section 1: when the reservation distribution is a two point distribution with mass γ at 0 and mass $1-\gamma$ at d, and the service time distribution is either exponential or deterministic with mean η. Note that the mean ξ of the reservation distribution in this case is equal to $(1-\gamma)d$.

In both of the abovementioned cases, the nonhomogeneous Poisson arrival process to the bookings queue, defined in (9), has constant rate $\tilde{\lambda} \equiv \lambda(1-\gamma)$ on the interval $[0, d)$ and is equal to 0 on the interval $[d, \infty)$.

6.1. Exponential service times

Customers who arrive at the reservation queue at time t with a reservation time equal to d require service during the interval $[t+d, t+d+s)$ for some s. To derive our infinite-server bound for the blocking probability of these customers, we consider the probability that the number of customers in the bookings queue exceeds C during the interval $[d, d+s)$. Since the bookings queue has no arrivals subsequent to time d, this is the same as the probability that the number of customers in the bookings queue exceeds C at time d. With $\mu = 1/\eta$, we know from the final line of Table 2 that the number of such customers follows a Poisson distribution with parameter

$$\nu(d) = \lambda \int_d^\infty (\gamma \exp(-\mu u) + (1-\gamma)\exp(-\mu(u-d)))\,\mathrm{d}u = \frac{\lambda}{\mu}[\gamma \mathrm{e}^{-\mu d} + (1-\gamma)],$$

and our infinite-server bound for the blocking probability of customers who have a reservation time of d is then

$$B_d = \sum_{\ell=C}^{\infty} \frac{(\nu(d))^\ell \mathrm{e}^{-\nu(d)}}{\ell!}. \tag{10}$$

Note that the dependence of $\nu(d)$ on d decays rapidly for large d, which explains the limiting behaviour that we observed in Figure 1.

To calculate the infinite-server bound for the blocking probability for customers who arrive to the queue with a reservation time of 0 and a service request S equal to s, we need to calculate the probability that the bookings queue reaches a capacity C at some time in the interval $[0, s)$. This queue is an infinite-server queue with a Poisson (λ/μ) initial number of customers, arrival rate $\tilde{\lambda}$ on the time interval $[0, d)$, and 0 thereafter, and per-customer service rate μ.

The Laplace transform of the probability of such a queue reaching a capacity C in time $[0, s)$ can be derived using techniques of transient analysis of infinite-server queues, as described in [19, Chapter 5], [20] or [2]. This result is stated explicitly in [18, Equation (28)].

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

Let $L_{m_0}(t)$ be the probability that the capacity of the bookings queue reaches C in the interval $[0, t)$ given that there were $m_0 < C$ customers initially present, and $\widetilde{L}_{m_0}(\sigma)$ be its Laplace transform. Then

$$\widetilde{L}_{m_0}(\sigma) = \frac{H_{m_0}(\sigma/\tilde{\lambda})}{\sigma H_C(\sigma/\tilde{\lambda})},$$

where $H_k(\sigma/\tilde{\lambda}) = (-\mu/\tilde{\lambda})^k C_k^{(\tilde{\lambda}/\mu)}(-\sigma/\mu)$, and $C_k^{(\tilde{\lambda}/\mu)}(\cdot)$ is a Charlier polynomial; see [6].

To derive the bound for the blocking probability of a customer who requests immediate service of length s when there are m_0 customers present, we need to invert $\widetilde{L}_{m_0}(\sigma)$ and then evaluate $L_{m_0}(\min(s, d))$. This could be done using the techniques described in [1]. However, we did it by defining the Laplace transform symbolically and using the symbolic computation package in MATLAB®.

Integrating with respect to the service request and summing over possible initial distributions, we arrive at the conclusion that the probability that the number of customers in the bookings queue will exceed C during the newly-arriving customer's service is

$$B_0 = \sum_{m_0=C}^{\infty} \frac{(\lambda/\mu)^{m_0} \mathrm{e}^{-\lambda/\mu}}{m_0!} + \sum_{m_0=0}^{C-1} \frac{(\lambda/\mu)^{m_0} \mathrm{e}^{-\lambda/\mu}}{m_0!} \left[\int_0^d \mu \mathrm{e}^{-\mu u} L_{m_0}(u)\, \mathrm{d}u + \mathrm{e}^{-\mu d} L_{m_0}(d) \right]. \tag{11}$$

This serves as our infinite-server bound for the blocking probability of nonreserving customers.

The results of our analytical calculations for an example with the same parameters as that simulated in Figure 1 are shown in Figure 5. There, we have plotted the value of the analytic upper bounds (10) and (11), as well as the simulation for both reserving and nonreserving customers, against the value of d. We observe that the upper bounds exceed the simulated values by an amount in the region of 0.01 to 0.04 for the reserving customers and consistently by about 0.04 for nonreserving customers. The upper bounds do, however, capture the shapes of the simulated curves very well.

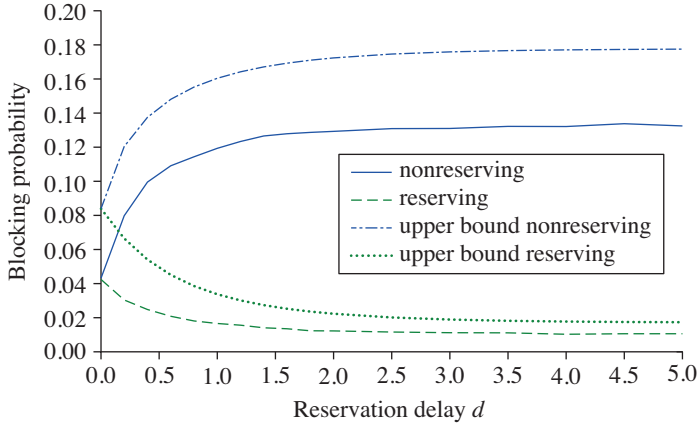

FIGURE 5: Blocking probabilities derived from (10) and (11), together with simulated results for the finite capacity system, plotted as a function of d. Parameters are $C = 10$, $\lambda = 6$, and $\gamma = \frac{1}{4}$.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

6.2. Deterministic service times

Now consider the case where the service time distribution is deterministic, with requested services equal to η with probability 1. Then the number of customers who are initially in service in the bookings queue has a Poisson distribution with mean $\lambda\eta$ and the remaining service times of these customers are chosen independently according to the distribution $F_S^{\mathrm{e}}(w)$ in (7), which is uniform on $[0, \eta]$. On the other hand, as in the exponential example in Section 6.1, the inhomogeneous Poisson process of arrival times in the bookings queue has rate $\tilde{\lambda}$ on the interval $[0, d)$ and 0 on the interval $[d, \infty)$.

As in the previous section, since the bookings queue has no arrivals subsequent to d, we can derive our bound for the blocking probability by calculating the probability that the number of customers present in the reservation queue at a time d into the future is greater than C. By the final line of Table 2, the number of such customers follows a Poisson distribution with parameter

$$\nu(d) = \lambda \int_d^\infty (\gamma[1 - F_s(u)] + (1 - \gamma)[1 - F_s(u - d)])\,\mathrm{d}u, \tag{12}$$

where F_s is the distribution of a deterministic random variable with mass concentrated at η.

If $d > \eta$ then (12) reduces to $\nu(d) = \lambda(1-\gamma)\eta$, and if $d \le \eta$ it reduces to $\nu(d) = \lambda(\eta - \gamma d)$. We conclude that our infinite-server bound for the blocking probabilities of reserving customers in the case where service times are deterministic is

$$B_d = \exp(-\nu(d)) \sum_{k=C}^{\infty} \frac{(\nu(d))^k}{k!}.$$

Note again that, provided that $d > \eta$, this expression does not depend on d, verifying our observations about Figure 2.

Calculating the infinite-server bound for the blocking probability of nonreserving customers is a little more complicated, because these customers can be blocked by customers who are in the bookings diary but are yet to commence service. We have to calculate the probability that the occupancy of the bookings queue is greater than or equal to C during the interval $[0, \eta)$. There are two classes of customers present in this interval:

- customers who were initially present in the queue, the number of which follows a Poisson distribution with parameter $\lambda\eta$ and whose remaining service times are independently and uniformly distributed on $[0, \eta]$;
- customers who arrive during the interval $[0, \eta)$ in a Poisson process with parameter $\tilde{\lambda}$: since they request service time equal to η, such customers must necessarily remain in the queue for the remainder of this interval.

Assume $d > \eta$ and that there are $m_0 < C$ customers initially present in the bookings queue. Then, the departure times $V_1, \ldots, V_{m_0}$ of these customers are distributed as the order statistics of m_0 independent uniform random variables on $[0, \eta]$. Thus, for $k = 1, \ldots, m_0 - 1$, the conditional density function of V_{k+1} given $V_k = v_k$ is

$$f_{V_{k+1} \mid V_k}(v_{k+1} \mid v_k) = \frac{(m_0 - k)(\eta - v_{k+1})^{m_0 - k - 1}}{(\eta - v_k)^{m_0 - k}}, \tag{13}$$

where we can take $v_0 = 0$. Putting (13) together with the fact that the arrival process of the bookings queue is Poisson with parameter $\tilde{\lambda}$ on the interval $[0, d)$, which contains $[0, \eta)$, we can generate a recursive expression for the probability that the queue size ever hits C in $[0, \eta)$.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

It turns out that it is more convenient to do this by writing (13) in terms of $u_k = \eta - v_k$. So, for $0 \le k \le m_0$, $0 \le \ell_k < C - m_0 + k$, and $u_k \in [0, \eta)$, let $\Gamma_{m_0,k}(\ell_k, u_k)$ be the probability that the queue reaches capacity C in $[\eta - u_k, \eta)$ given that $V_k = \eta - u_k$ and ℓ_k new arrivals have occurred in $[0, \eta - u_k)$. Then, conditioning on V_{k+1},

$$\begin{aligned}\Gamma_{m_0,k}(\ell_k, u_k) &= \sum_{\ell_{k+1}=\ell_k}^{C-m_0+k-1} \int_0^{u_k} \left[\frac{(m_0-k)(u_{k+1})^{m_0-k-1} \mathrm{e}^{-\tilde{\lambda}(u_k-u_{k+1})} (\tilde{\lambda}(u_k-u_{k+1}))^{\ell_{k+1}-\ell_k}}{(u_k)^{m_0-k}(\ell_{k+1}-\ell_k)!} \right] \\ &\qquad \times \Gamma_{m_0,k+1}(\ell_{k+1}, u_{k+1})\, \mathrm{d}u_{k+1} \\ &\quad + \sum_{\ell_{k+1}=C-m_0+k}^{\infty} \int_0^{u_k} \frac{(m_0-k)(u_{k+1})^{m_0-k-1} \mathrm{e}^{-\tilde{\lambda}(u_k-u_{k+1})} (\tilde{\lambda}(u_k-u_{k+1}))^{\ell_{k+1}-\ell_k}}{(u_k)^{m_0-k}(\ell_{k+1}-\ell_k)!}\, \mathrm{d}u_{k+1}. \end{aligned} \tag{14}$$

The integrand in the first term of (14) contains the probability that there are $\ell_{k+1} < C - m_0 + k$ arrivals to the bookings queue in the time interval $(0, \eta - u_{k+1}]$ and then the queue subsequently reaches capacity C, while the integrand in the second term contains the probability that the number of arrivals to the bookings queue reaches $C - m_0 + k$, and so the total number of customers in the queue reaches C, in the time interval $(\eta - u_k, \eta - u_{k+1}]$.

If there are $\ell_{m_0} < C$ new customers when the final initially-present customer departs at time $v_{m_0} = \eta - u_{m_0}$, then the probability that the bookings queue will fill up in the interval $[\eta - u_{m_0}, \eta)$ is

$$\Gamma_{m_0,m_0}(\ell_{m_0}, u_{m_0}) = \sum_{\ell_{m_0+1}=C}^{\infty} \exp(-\tilde{\lambda}(u_{m_0})) \frac{(\tilde{\lambda}(u_{m_0}))^{\ell_{m_0+1}-\ell_{m_0}}}{(\ell_{m_0+1}-\ell_{m_0})!},$$

which serves as a starting point for the backward recursion (14).

Given that there are $m_0 < C$ customers present at time 0, the probability that the number of customers in the bookings queue reaches C at some point in the interval $[0, \eta)$ is given, in the above notation, by the function $\Gamma_{m_0,0}(0, \eta)$. With this in hand, our infinite-server bound for the blocking probability is given by

$$B_0 = \sum_{m_0=0}^{C-1} \frac{\mathrm{e}^{-\lambda\eta}(\lambda\eta)^{m_0}}{m_0!} \Gamma_{m_0,0}(0, \eta) + \sum_{m_0=C}^{\infty} \frac{\mathrm{e}^{-\lambda\eta}(\lambda\eta)^{m_0}}{m_0!}.$$

Note once more that, provided that $d > \eta$, all the calculations that lead to this expression do not depend on d, again verifying our observation about Figure 2.

When $d \le \eta$, (13) still holds for the conditional departure points of the customers initially present in the bookings queue. However, there are no more arrivals to the bookings queue after time d, and so if the occupancy of the bookings queue has not reached C before time d, then it

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

will not do so in the interval $[d, \eta)$. In this case, (14) takes the form

$$\begin{aligned}\Gamma_{m_0,k}(\ell_k, u_k) &= \sum_{\ell_{k+1}=\ell_k}^{C-m_0+k-1} \int_{\eta-d}^{u_k} \left[\frac{(m_0-k)(u_{k+1})^{m_0-k-1} \mathrm{e}^{-\tilde{\lambda}(u_k-u_{k+1})} (\tilde{\lambda}(u_k-u_{k+1}))^{\ell_{k+1}-\ell_k}}{(u_k)^{m_0-k}(\ell_{k+1}-\ell_k)!} \right] \\ &\qquad \times \Gamma_{m_0,k+1}(\ell_{k+1}, u_{k+1})\, \mathrm{d}u_{k+1} \\ &\quad + \sum_{\ell_{k+1}=C-m_0+k}^{\infty} \int_{\eta-d}^{u_k} \frac{(m_0-k)(u_{k+1})^{m_0-k-1} \mathrm{e}^{-\tilde{\lambda}(u_k-u_{k+1})} (\tilde{\lambda}(u_k-u_{k+1}))^{\ell_{k+1}-\ell_k}}{(u_k)^{m_0-k}(\ell_{k+1}-\ell_k)!}\, \mathrm{d}u_{k+1} \\ &\quad + \sum_{\ell_{k+1}=C-m_0+k}^{\infty} \frac{\mathrm{e}^{-\tilde{\lambda}(u_k-\eta+d)} (\tilde{\lambda}(u_k-\eta+d))^{\ell_{k+1}-\ell_k}}{(\ell_{k+1}-\ell_k)!} \left[\frac{(\eta-d)}{(u_k)} \right]^{m_0-k}. \end{aligned} \tag{15}$$

The first two terms in this recursion are analogous to the two terms on the right-hand side of (14), while the third covers the possibility that the $(k+1)$th departure of an initial customer occurs after time d, and so there is a time period of only $u_k - \eta + d$ after the kth departure of an initial customer in which further arrivals can occur. Consequently, $\tilde{\lambda}(u_k - \eta + d)$ is the parameter of the Poisson number of customers that arrive after the kth departure of an initial customer.

Similar reasoning leads to the initial term of the backwards recursion. If the final initial customer departs before time d, that is, $u_{m_0} > \eta - d$, with $\ell_{m_0} < C$ new customers present, we can express the probability that the queue length exceeds C before time d as

$$\Gamma_{m_0,m_0}(\ell_{m_0}, u_{m_0}) = \sum_{\ell_{m_0+1}=C}^{\infty} \exp(-\tilde{\lambda}(u_{m_0}-\eta+d)) \frac{(\tilde{\lambda}(u_{m_0}-\eta+d))^{\ell_{m_0+1}-\ell_{m_0}}}{(\ell_{m_0+1}-\ell_{m_0})!}.$$

We implemented the two recursions (14) and (15) using the symbolic mathematics toolbox in MATLAB and produced Figure 6, in which the analytic bounds and the results of the finite capacity simulation are plotted against d for both reserving and nonreserving customers, against

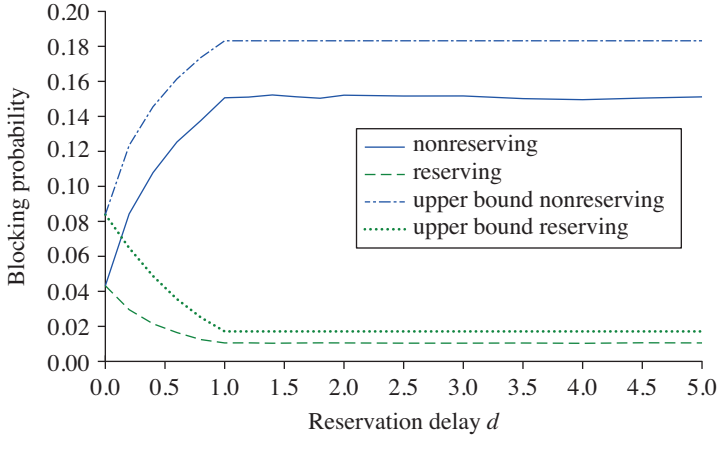

FIGURE 6: Blocking probabilities derived from (10) and (11), together with simulated results for the finite capacity system, plotted as a function of d. Parameters are $C = 10$, $\lambda = 6$, and $\gamma = \frac{1}{4}$.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

the value of d. We can make similar observations to those that we made about Figure 5. Again we see that the distance between the upper bound and the simulated values varies between 0.01 and 0.04 for reserving customers and is relatively consistent at about 0.04 for nonreserving customers. Again the shapes of the simulated curves are captured very well by the infinite-server bounds.

7. Conclusion

We have presented an infinite-server model for a continuous-time queueing system with advanced reservations. Analysis of this model provides an upper bound for the rejection probabilities experienced by customers arriving to a finite-capacity queue. Our major observation is that the stationary measure of the bookings diary is identical to the law of the $\mathrm{M}(t)/\mathrm{G}/\infty$ queue that we have called the *bookings queue*, which has almost surely finitely-many customers in total. This observation opens the possibility of deriving approximations to the blocking probability of reservation queues by analysing the time dependent behaviour of such queues, even if it is just via simulation.

In Section 6 we carried out the necessary analytical calculations for two examples. It would be of interest to extend these calculations to more general situations. Indeed, we think that the recursive procedure that we used to analyse the case with deterministic service times can be used more generally. Furthermore, the link between the bookings queue and queues with advanced reservations provides a good motivation for a general study of the properties of queues with inhomogeneous Poisson arrival processes and almost surely finitely-many customers.

Acknowledgements

The authors would like to thank an anonymous referee for carefully reading an earlier version and making a number of constructive suggestions that improved the paper considerably. Peter Taylor would like to thank the Australian Research Council for supporting this work through Laureate Fellowship FL130100039 and through the ARC Centre of Excellence for Mathematical and Statistical Frontiers.

References

[1] Abate, J. and Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comput.* **7,** 36–43.

[2] Abate, J. and Whitt, W. (1998). Calculating transient characteristics of the Erlang loss model by numerical transform inversion. *Commun. Statist. Stoch. Models* **14,** 663–680.

[3] Barakat, N. and Sargent, E. H. (2004). An accurate model for evaluating blocking probabilities in multi-class OBS systems. *IEEE Commun. Lett.* **8,** 119–121.

[4] Barakat, N. and Sargent, E. H. (2005). Analytical modelling of offset-induced priority in multiclass OBS networks. *IEEE Trans. Commun.* **53,** 1343–1352.

[5] Borovkov, K. (2003). *Elements of Stochastic Modelling*. World Scientific, River Edge, NJ.

[6] Chihara, T. S. (1978). *An Introduction to Orthogonal Polynomials*. Gordon and Breach, New York.

[7] Coffman, E. G. Jr., Flatto, L. and Jelenković, P. (2000). Interval packing: the vacant interval distribution. *Ann. Appl. Prob.* **10,** 240–257.

[8] Coffman, E. G. Jr., Jelenković, P. and Poonen, B. (1999). Reservation probabilities. *Adv. Performance Anal.* **2,** 129–158.

[9] Coffman, E. G. Jr., Flatto, L., Jelenković, P. and Poonen, B. (1998). Packing random intervals on-line. *Algorithmica* **22,** 448–476.

[10] Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, Vol. 1, *Elementary Theory and Methods*. Springer, New York.

[11] Dolzer, K. and Gauger, C. (2001). On burst assembly in optical burst switching networks—a performance evaluation of just-enough-time. In *Proceedings of the 17th International Teletraffic Congress*, pp. 149–160.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press

[12] Foley, R. D. (1982). The nonhomogeneous $M/G/\infty$ queue. *Opsearch* **19,** 40–48.
[13] Greenberg, A. G., Srikant, R. and Whitt, W. (1999). Resource sharing for book-ahead and instantaneous-request calls. *IEEE/ACM Trans. Networking* **7,** 10–22.
[14] Kaheel, A., Alnuweiri, H. and Gebali, F. (2004). Analytical evaluation of blocking probability in optical burst switching networks. In *Proc. IEEE Internat. Conf. Commun.*, Vol. 3, pp. 1548–1553.
[15] Kaheel, A. M., Alnuweiri, H. and Gebali, F. (2006). A new analytical model for computing blocking probability in optical burst switching networks. *IEEE J. Selected Areas Commun.* **24,** 120–128.
[16] Levi, R. and Shi, C. (2014). Revenue management of reusable resources with advanced reservations. Submitted.
[17] Liang, Y., Liao, K., Roberts, J. W. and Simonian, A. (1988). Queueing models for reserved set up telecommunications services. In *Proc. Teletraffic Science for New Cost-Effective Systems, Networks and Services*, Session 4.4B, 1.1–1.7.
[18] Ramakrishnan, M., Sier, D. and Taylor, P. G. (2005). A two-time-scale model for hospital patient flow. *IMA J. Manag. Math.* **16,** 197–215.
[19] Riordan, J. (1962). *Stochastic Service Systems.* John Wiley, New York.
[20] Takács, L. (1962). *Introduction to the Theory of Queues.* Oxford University Press.
[21] Van de Vrugt, M., Litvak, N. and Boucherie, R. J. (2014). Blocking probabilities in Erlang loss queues with advance reservation. *Stoch. Models* **30,** 187–196.
[22] Virtamo, J. T. (1992). A model of reservation systems. *IEEE Trans. Commun.* **40,** 109–118.
[23] Vu, H. L. and Zukerman, M. (2002). Blocking probability for priority classes in optical burst switching networks. *IEEE Commun. Lett.* **6,** 214–216.
[24] Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.

https://doi.org/10.1017/apr.2015.4 Published online by Cambridge University Press