Journal of the Operational Research Society (1997) 48, 1113–1122

Control of arrivals in a finite buffered queue with setup costs

F Karaesmen[1] and SM Gupta[2]

[1] *Université Pierre, Paris, France and* [2] *Northeastern University, USA*

We consider finite buffered queues where the arrival process is controlled by shutting down and restarting the arrival stream. In the absence of holding costs for items in the queue, the optimal (s, S) policy can be characterised by relating the arrival control problem to a corresponding service control problem. With the inclusion of holding costs however, this characterisation is not valid and efficient numerical computation of the queue length probability distribution is necessary. We perform this computation by using a duality property which relates queue lengths in the controlled arrival system to a controlled service system. Numerical results which analyse the effect of setup and holding costs and the variability of the arrival process on the performance of the system are included.

Keywords: control; queueing

Introduction

Admission control problems for single queues and queueing networks are well studied. In typical admission control problems, queue lengths are controlled by rejecting the incoming arrivals. A closely related problem is that of input control, in which arrival streams are controlled by a router and the router can interrupt and restart the arrival stream. We consider such a routing control problem in this paper for a finite buffered queue.

Finite buffered queues lead to the blocking phenomenon in which arrivals that find the buffer full are considered lost to the system. In this paper, we treat a different kind of finite buffered queue; one in which items are not lost as the arrival stream is turned off when the buffer is full. Hence, instead of modeling arrivals as external to the system as usual, we consider them as internal processes such as the first stage of a two stage production line.

Buzacott *et al*[1] consider a version of the above *lossless* finite buffered queue. They are motivated by multistage flow lines (see Buzacott and Shantikumar[2] for more details) in which the upstream machine stops production when the downstream buffer is full and restarts production when downstream buffer space is available. This is the 'manufacturing type blocking' phenomenon that occurs in multistage production lines with limited buffer space between the production stages and is known as 'blocking before service'. In the case of two machine flow lines which are the basis of approximation algorithms of multistage flow lines, the downstream buffer behaves like a G/G/1/K queue with a stopped arrival process. Efficient approximation algorithms for this queue are given in Buzacott *et al.*[1] Our model is different. For the G/G/1/K queue with stopped arrivals, the arrival process is restarted when the queue length decreases to $K-1$ whereas in our model, the arrival process can be restarted at any buffer level between 0 and $K-1$.

A related problem to the arrival control problem considered here is that of service control introduced by Yadin and Naor[3] in the context of the M/G/1 queue. In the service control problem, the server is shut down when the queue length is small and restarted at a higher queue length. To optimise the system, Yadin and Naor suggested using the following operating policy: shut down the server when the server becomes idle and restart it as the queue length reaches a threshold level N. This operating policy was termed *N-policy*. Later, Heyman[4] proved that N-policy is the optimal policy for operating M/G/1 queues under various cost criteria.

Most of the results in this paper are based on a queue length relationship for two different (but related) finite buffered queues. This relationship is called queue length *duality*. An early example by Harris[5] considers the finite buffered M/G/1 queue and its dual. Other duality relations were obtained by Gupta[6–9] and by Gupta and Melachrinoudis[10] for uncontrolled Markovian queues. Queue length duality relations for non-Markovian queues can be found in Hlynka and Wang[11] and Yang.[12] For controlled queues, this relationship has been studied by Gupta[13] and Karaesmen and Gupta.[14]

Sparaggis *et al*[15] have introduced a different kind of duality. They have shown that the problems of dynamically routing customers into multiple buffers and dynamically

Correspondence: Dr SM Gupta, Department of Mechanical, Industrial and Manufacturing Engineering, 334 Snell Engineering Center, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115, USA.
E-mail: gupta@neu.edu

scheduling between different classes of customers are equivalent when the buffer sizes are finite. This is an important result as the solution of either dynamic optimisation problem also provides the solution to the other. Using a similar approach, Xu and Shantikumar[16] and Xu[17] related the admission control/scheduling for queues to expulsion/scheduling type counterparts to solve dynamic control problems. We employ a similar strategy here to relate a dynamic arrival control problem to a dynamic service control problem for a single class queue under a certain type of routing policy. In this sense, our results are less generic than those of Sparragis *et al*,[15] however we are not restricted to exponential service times and Poisson arrivals.

The paper is arranged as follows. In the next section, we give a detailed description of the problem and introduce some notation. Then, we concentrate on the optimisation problem. After discussing the specific class of control policies to study, we obtain some structural results using queue length duality. We then focus on computing the optimal revenue for a given set of policy parameters. Efficient computation requires a fast method to obtain the stationary queue length distribution for the given set of parameters and we provide this method by obtaining the queue length distribution in closed form. To gain further insight, we give numerical results of experiments with different cost and traffic parameters. Finally we summarise our results.

Definitions and notation

We consider a modified version of the GI/M/1/K queue. The arrivals occur according to a renewal process with mean rate λ when the arrival stream is not interrupted due to the control policy. The service times are exponentially distributed with mean $1/\mu$ and the buffer size is K (including the item that is in service). The buffer size restriction states that the queue lengths can be at most K regardless of the arrival control policy employed, however the particular control policy can impose further restrictions on the maximum queue length. For example, the particular control policy may not allow for more than S (where $S \leqslant K$) items in the buffer thereby decreasing the effective buffer size to S. One important characteristic of the system we study is that arrivals are never lost regardless of the effective buffer size. This is achieved in the following way: an arrival process only takes place if there is (effective) buffer space available. Otherwise, the arrival process will be delayed until this space becomes available. To avoid any confusion, whenever we refer to queue lengths in this paper, this should be understood as the number of items in the system including the item that is currently in service.

The cost structure is as follows: the system earns rewards at rate R_1 for each processed job. In addition, a reward of R_2 dollars per unit time is earned when the arrival stream is off. Each time the arrival stream is turned off and on, setup costs of c_1 and c_2 dollars are incurred respectively. Finally, holding costs are incurred at rate h dollars per item for the items in the system (including the item that is in service).

We want to find an operating policy that will maximise the expected revenues per unit time. We restrict our search to stationary policies. This allows us to express the expected revenue per unit time as a function of the stationary queue length distribution. Let f be a stationary control policy for this problem, π_{off} be the long run proportion of time that the arrival stream is off and π_0 be the long run fraction of the time that the server is idle. Also, let L denote the average queue length. Note that the stationary queue length distribution of the system is induced by the stationary policy f and that π_{off}, π_0 and L are all functions of f. (We do not express this dependence explicitly for the sake of keeping the notation simple.) Finally, let $C(f, t)$ be the total switching cost incurred until time t when an operating policy f is used. Hence we can write our objective as:

$$\max_j \left\{ R_1(1-\pi_0)\mu + R_2\pi_{\text{off}} - hL - \lim_{T\to\infty} \frac{C(f,T)}{T} \right\} \quad (1)$$

In the next section, we discuss the above optimisation problem in detail.

Optimisation

(s, S) policies for arrival control

As stated in the previous section, our objective is to find an operating policy that will maximise the revenue function. We start by restricting our search to the following policy: turn the arrival stream off when the queue length is S and turn it back on when the queue length decreases to s, where $0 \leqslant s < S \leqslant K$. We refer to this policy as the (s, S) policy for arrival control. Frequently used for controlling inventory systems (s, S) policies are attractive as they are easy to understand and easy to implement. Furthermore, for uncapacitated production systems, (s, S) policies are optimal. On the other hand, for capacitated production systems, Sobel[18] has shown that the expected revenue per unit time for any stationary policy is equal to that of an (s, S) policy. By following the same argument, we can argue that (s, S) policies are as good as any other stationary policy for our problem.

In the absence of setup costs, optimal admission control policies are shown to be of threshold type where the incoming arrivals are rejected when the queue length is S and are accepted when the queue length is less than S. For our purposes, these single parameter threshold policies can be treated as special cases of the (s, S) policy with $s = S - 1$.

The restriction to the (s, S) policy reduces our problem to optimisation with respect to the parameters s and S. More-

over, under the assumption of an (s, S) policy, it is possible to write an alternative expression for the expected cost per unit time, instead of (1), using the regenerative behaviour of the queue length process. Consider the regenerative cycle which starts when the queue length increases to S (and the arrival process is stopped) and lasts until the next time the queue length goes up to S. Note that, each regenerative cycle consists of an *off period* (starting at S and ending at s) where the arrival process is inactive and an *on period* (starting at s and ending at S) where the arrival process is active. Let T_{off} denote the expected time of an *off period*. (Note that, as mentioned in the previous section T_{off}, π_{off}, and L are functions of the control policy. In this case, these quantities are implicitly dependent on s and S).

Then we have:

$$\lim_{T\to\infty} \frac{C(f,T)}{T} = (c_1 + c_2)\frac{\pi_{\text{off}}}{T_{\text{off}}} \tag{2}$$

Noting that T_{off} is the expected time for $S - s$ service completions, we can write:

$$T_{\text{off}} = \frac{S-s}{\mu} \tag{3}$$

Hence, we can express the objective function in (1) in terms of the stationary queue length distribution as:

$$\max_{(s,S)} \left\{ R_1(1-\pi_0)\mu + R_2\pi_{\text{off}} - hL - (c_1+c_2)\frac{\mu\pi_{\text{off}}}{S-s} \right\}. \tag{4}$$

Note that in the above expression, if the value of S is less than or equal to K, any change in K has no effect on the total revenue.

Arrival and service control duality

Queue length duality is a relationship between the stationary distributions of two finite buffered queueing systems. Consider two queueing systems with buffer size K and stationary queue length distributions π_i^P and π_i^D $(i = 0, 1, 2, \ldots, K)$ respectively. Duality implies the following relationship:

$$\pi_i^P = \pi_{K-i}^D \quad \text{for } i = 0, 1, 2, \ldots, K \tag{5}$$

In this section, we establish a queue length duality relationship as in equation (5) between a queueing system with controlled arrivals through (s, S) policies and another queueing system with controlled service. To this end we state the following theorem.

Theorem 1 *Consider the* GI/M/1/K *queue with* (s, S) *arrival control. Let* $\pi_i^A (i = 0, 1, 2, \ldots, K)$ *be its stationary queue length distribution. The dual system is an* M/G/1/K *queue with* $(K - S, K - s)$ *service control and arrival and service processes interchanged, namely*

$$\pi_i^A = \pi_{K-i}^S \tag{6}$$

where π_i^S $(i = 0, 1, 2, \ldots, K)$ *is the stationary queue length distribution of the* M/G/1/K *queue with* $(K - S, K - s)$ *service control.*

Proof Consider a (closed) cyclic network of two nodes with K customers. The service times at node 1 are exponentially distributed and the service times at node 2 have an arbitrary probability distribution function F_A. Assume that node 1 of the above network operates using (s, S) arrival control. Thus, when the queue length at node 1 increases to S, the service process at the second node is stopped. The service process (of node 2) will restart the next time the queue length at node 1 falls down to s. First note that, node 1 of the above network, is identical to GI/M/1/K queue with stopped arrivals under (s, S) type arrival control. Furthermore, the departure times from node 1 are the arrival times to node 2 and the departure times from node 2 are the arrival times to node 1. Due to the arrival control policy, the arrival process to node 1 becomes inactive when the queue length at node 1 becomes S which corresponds to a departure from node 2 that leaves $K - S$ customers behind. The next time the arrival process becomes active is when the queue length at node 1 falls down to s. At the same instance the queue length at node 2 will increase to $K - s$. Hence, the second node of the cyclic network is an M/G/1/K queue operating under $(K - S, K - s)$ service control policy. To finalise the proof, let $L_1(t)$ and $L_2(t)$ denote the queue length processes at nodes 1 and 2 of the above network respectively. By the cyclic nature of the network, we have:

$$L_1(t) = K - L_2(t) \tag{7}$$

To summarise the argument, consider a sequence of interarrival times, $\{\omega_n\}$, sampled from F_A and a sequence of service times, $\{\eta_n\}$, sampled from an exponential distribution with rate μ. One can construct a sample path of the queue length process for the GI/M/1/K queue with (s, S) arrival control from the above sequences. Moreover, this sample path is symmetric around K to the sample path of M/G/1/K queue with $(K - S, K - s)$ service control that is constructed by using $\{\eta_n\}$ as the interarrival time sequence and $\{\omega_n\}$ as the service time sequence (provided that symmetry holds at time 0). In other words, this symmetry can be achieved under certain initial conditions with probability one. However, if a unique stationary queue length distribution exists, it is independent of the initial conditions. This implies the duality of the stationary queue length distributions for the two systems. □

Structure of the optimal arrival control policy in the absence of holding costs

In the previous section, we studied the relationship between the queue lengths of GI/M/1/K and M/G/1/K operating under (s, S) arrival and $(K - S, K - s)$ service policies respectively. In this section, we discuss the structure of

the optimal arrival control policy when holding costs are zero, using the dual service control problem.

For negligible holding costs, the arrival control problem is:

$$\max_{(s,S)}\left\{R_1(1-\pi_0)\mu + R_2\pi_{\text{off}} - (c_1+c_2)\frac{\mu\pi_{\text{off}}}{S-s}\right\} \quad (8)$$

As a corrollary to Theorem 1 we can state an equivalent problem to (8).

Corollary 1 *Define* $Q=K-S$ *and* $q=K-s$, *and consider the optimisation problem in* (8). *An equivalent problem is to find the optimal* (q,Q) *policy that will maximise the total revenue minus the setup costs for an* M/G/1/K *queue with arrival rate* μ *and service rate* λ (*and identical cost parameters*). *The optimal revenues for both problems are equal and the optimal threshold levels are related as follows:* $s^*=K-Q^*$ *and* $S^*=K-q^*$ (*where* (s^*,S^*) *and* (q^*,Q^*) *are the optimal policies for the arrival and service control problems respectively*).

Proof In the corresponding service problem, R_1 dollars are earned per item processed. However, this time a fraction of the arriving items are lost. Hence, the throughput of the system is: $(1-\pi_K)\mu$ and the objective function can be written as:

$$\max_{(q,Q)}\left\{R_1(1-\pi_K)\mu + R_2\pi_{\text{off}} - (c_1+c_2)\frac{\mu\pi_{\text{off}}}{Q-q}\right\} \quad (9)$$

By duality, when $s=K-Q$ and $S=K-q$, π_0 in (8) is equal to π_K in (9), furthermore the proportion of off times in both systems are also equal. Hence, the problems are identical and yield the same optimal value. That is, if (q^*,Q^*) is the optimal pair for (9) then the dual pair $(K-Q^*,K-q^*)$ must be optimal for (8).

The service control problem given by (9) has been studied in detail by Hersh and Brosh[19] and Teghem.[20] The following important result is reported in Teghem[20]: the optimal (s,S) policy for (9) is of the form $(0,S)$ where $0\leqslant S\leqslant K+1$. The $(0,0)$ policy corresponds to never turning the server off and the $(0,K+1)$ policy corresponds to keeping the server closed at all times. We can immediately transform this result into an arrival control result using duality and Corollary 1.

Corollary 2 *For the arrival control problem in* (8), *the optimal* (s,S) *policy is of the form* (s,K) *where* $-1\leqslant s\leqslant K$. *The* (K,K) *policy corresponds to never interrupting the arrival process and restarting it as soon as space opens up and the* $(-1,K)$ *policy corresponds to always keeping the arrival process off.*

In summary, for negligible holding costs, we have shown that, the optimal (s,S) pairs are characterised by $S=K$. If a search procedure is required to solve (8), then one can keep S fixed at K and perform a one-dimensional search for s which saves a lot of computational time.

The case with holding costs

In the previous section, we have seen that the absence of holding costs leads to a simplification in the structure of the optimal policy. In this section, we discuss some of the implications of relaxing this requirement.

We begin with the dual service control problem with holding costs to gain insight into its arrival control counterpart. The problem is:

$$\max_{(q,Q)}\left\{R_1(1-\pi_K)\mu + R_2\pi_{\text{off}} - hL - (c_1+c_2)\frac{\mu\pi_{\text{off}}}{Q-q}\right\} \quad (10)$$

First note that, if the buffer size were not restricted (namely as $K\to\infty$ and the mean traffic load, ρ, is less than 1) in (10), the optimal (s,S) policy would have $s=0$ as shown by Heyman.[4] When the buffer size is finite, this property seems hard to prove, but numerical results suggest that for most problems, $s=0$ continues to hold (Karaesmen and Gupta[21]). As a heuristic argument, $(0,S)$ policies perform better than other policies when holding costs are not considered but the fact that they continue to perform well when holding costs are added implies that $(0,S)$ policies are good with respect to holding costs as well. This argument can be made concrete in the case of infinite buffers as it is possible to prove that $(0,S)$ policies lead to shorter expected queue lengths than (s,S) policies with $s=1,2,\ldots,S$ (Karaesmen and Gupta[21]).

Going back to the arrival control problem, the above argument suggests that as $(0,S)$ service control policies lead to shorter expected queue lengths, $(K-S,K)$ type policies will lead to longer expected queue lengths due to duality. Therefore we may expect that the queue lengths will decrease by decreasing the upper threshold in the arrival control problem. In terms of optimisation, this decrease brings complications, as setup costs and revenues favor large values of the upper threshold but holding costs favor lower thresholds. The potential optimality of any (s,S) pair prevents the reduction of optimisation problem to a single dimensional search. However, for an exhaustive search over all (s,S) pairs efficient computation of the stationary queue lengths is necessary. This will be handled in the next section.

The stationary queue length distribution

In this section, we study the stationary queue length distribution of the GI/M/1/K queue with (s,S) arrival control policy. Note that, by duality the stationary distribution of the GI/M/1/K queue with (s,S) arrival control can be obtained from that of the corresponding M/G/1/K queue with service control. The stationary distribution for the

service controlled M/G/1/K queue is reported by Teghem[20] and Takagi[22]. However, in both cases the computation of the distribution is not straightforward. To facilitate the computation we may follow an approach pioneered by Morse[23] and later systematically developed by Neuts[24] and consider phase-type service distributions. In fact, when the service distribution is of phase-type, the queue length distribution can be obtained in closed form (Karaesmen and Gupta[21]). Thus, the stationary queue length distribution of the GI/M/1/K queue with (s, S) arrival control policy is easily computable when the arrivals are of phase-type. As the effective buffer size of this queue will be determined by S, we can simplify notation by considering (s, K) type arrival control and eliminate S from the following discussion.

Let the arrival distribution be of phase-type with representation $(\boldsymbol{\beta}, \mathbf{T})$ where $\boldsymbol{\beta}$ is a vector of the initial distribution of the phases and $\mathbf{T}$ is a matrix which characterises the transition probabilities between phases (see Neuts[24]). If the distribution has n phases, then $\boldsymbol{\pi}_i = (\pi_{i,0}, \pi_{i,1}, \ldots, \pi_{i,n})$ with $\pi_{i,j}$ denoting the stationary probability that there are $i (i = 1, 2, \ldots, K)$ customers and the arrival phase is $j (j = 0, 1, 2, \ldots, n)$. Also, to specify the matrix-geometric solution, let:

$$\mathbf{R} = \mu(\mu\mathbf{I} - \mu\mathbf{B}^{\mathbf{00}} - \mathbf{T})^{-1} \tag{11}$$

$$\mathbf{B}^{\mathbf{00}} = \mathbf{e}\boldsymbol{\beta}. \tag{12}$$

and

$$\mathbf{D} = \mathbf{I} - \mathbf{R} \tag{13}$$

with $\mathbf{I}$ denoting the $n \times n$ identity matrix and $\mathbf{e}$ denoting a $1 \times n$ (or $n \times 1$) vector $(1, 1, 1, \ldots, 1)$.

Lemma 1 *The stationary queue length distribution for the* PH/M/1/K *queue with* (s, K) *arrival control is given by*:

$$\pi_0 = \frac{(1 - (1/\rho))\alpha_s(K)}{(1 - (1/\rho)\alpha_s(K))}$$

$$\pi_i = \frac{\kappa(1 - (1/\rho))}{K - s}(\boldsymbol{\beta}\mathbf{R}^{s-i+1}(\mathbf{I} - \mathbf{R}^{K-s})\mathbf{D}^{-1})\mathbf{e} \quad \text{for } i = 1, 2, \ldots, s$$

$$\pi_i = \frac{\kappa(1 - (1/\rho))}{K - s}(1 + ((\boldsymbol{\beta}(\mathbf{I} - \mathbf{R}^{K-i+1})\mathbf{D}^{-1} - \mathbf{I})\mathbf{e})) \quad \text{for } i = s + 1, s + 2, \ldots, K - 1$$

$$\pi_K = \kappa\frac{(1 - (1/\rho))}{K - s} \tag{14}$$

where

$$\alpha_s(K) = \frac{(1 - (1/\rho))}{K - s}\boldsymbol{\beta}(\mathbf{D}^{-1} - \mathbf{D}^s(\mathbf{I} - \mathbf{R}^{K-s})\mathbf{D}^{-1})\mathbf{e} \tag{15}$$

and

$$\kappa = \frac{1}{(1 - (1/\rho)\alpha_s(K))} \tag{16}$$

Proof Consider the dual system, namely the M/PH/1/K queue with $(0, Q)$ service control where $Q = K - s$ with mean arrival rate μ, mean service rate λ and traffic intensity $\rho = \mu/\lambda$. The stationary queue length distribution of this queue can be obtained by relating it to the identical queue with infinite buffer capacity (Karaesmen and Gupta[21]). This requires the computation of a normalisation constant, κ. Using the results of Keilson and Servi,[25] this normalisation constant can be written in terms of the random variable L^{∞}, which denotes the queue length in the infinite capacity queue (for the service control problem). In particular, κ turns out to be a function of the probability $\mathrm{P}\{L^{\infty} > K\}$, and is given by:

$$\kappa = \frac{1}{1 - \rho\mathrm{P}\{L^{\infty} > K\}} \tag{17}$$

Note that, since $\mathrm{P}\{L^{\infty} > K\}$ depends on the threshold Q as well as the buffer size, we denote it by $\alpha_Q(K)$. Moreover, $\alpha_Q(K)$ can be written in closed form as follows:

$$\alpha_Q(K) = \frac{(1 - \rho)}{Q}\boldsymbol{\beta}(\mathbf{D}^{-1} - \mathbf{D}^{K-Q}(\mathbf{I} - \mathbf{R}^Q)\mathbf{D}^{-1})\mathbf{e} \tag{18}$$

Now, letting p_i denote the stationary distribution for the M/PH/1/K queue with $(0, Q)$ service control, we obtain (Karaesmen and Gupta[21]):

$$p_0 = \kappa\frac{(1 - \rho)}{Q}$$

$$p_i = \frac{\kappa(1 - \rho)}{Q}(1 + ((\boldsymbol{\beta}(\mathbf{I} - \mathbf{R}^{i+1})\mathbf{D}^{-1} - \mathbf{I})\mathbf{e})) \quad \text{for } i = 1, 2, \ldots, Q - 1$$

$$p_i = \frac{\kappa(1 - \rho)}{Q}(\boldsymbol{\beta}\mathbf{R}^{i-Q+1}(\mathbf{I} - \mathbf{R}^Q)\mathbf{D}^{-1})\mathbf{e} \quad \text{for } i = Q, Q + 1, \ldots, K - 1$$

$$p_K = \frac{(1 - \rho)\alpha_Q(K)}{(1 - \rho\alpha_Q(K))} \tag{19}$$

Returning to the arrival control problem, by Theorem 1, we have:

$$\pi_i = p_{K-i} \quad \text{for } i = 0, 1, 2, \ldots, K \tag{20}$$

Therefore a relabeling of the states and the appropriate conversion of the required constants by interchanging λ and μ yield the stationary queue length distribution for the arrival control problem. □

Remark Note that the stationary distribution depends on the particular values of the parameters s and S $(=K)$. This dependence is exhibited in (14) as the stationary probabilities are obtained as functions of s and K (or $\alpha_s(K)$) as well as the traffic parameters.

Numerical examples

In this section we compute the optimal (s, S) values for various traffic parameters, cost parameters and arrival distribution types using the stationary distribution obtained in the previous section. Arrival processes may be highly variable and it is important to understand how the optimal

Table 1 Some properties of the arrival distributions used

Type	*CV*	*Properties*	*Mixing probabilities*
D1	0.44721	E_6 with a rate of 6 in each phase	
D2	0.70711	E_2 with a rate of 2 in each phase	
D3	1	Exponential distribution with rate 1	
D4	1.5	H_2 with rates 2.82085228 and 0.50806659	0.6, 0.4
D5	2.0	H_2 with rates 0.22540333 and 1.77459677	0.11270167, 0.88729833

Notation: CV repesents the coefficient of variation, E_k is the k-stage Erlang distribution and H_2 is the 2-stage hyperexponential distribution.

policy and the optimal revenue change as a function of this variability. For this purpose, we experiment with arrival distributions with coefficients of variation (standard deviation/mean) ranging from 0.44–2.0. The distributions used here are a subset of distributions also used in Neuts and Rao.[26] The detailed properties of these distributions are given in Table 1.

The general experimental setup is as follows. The arrival rate λ is set to 1. We then tabulate the optimal values of the thresholds s and S and the optimal revenue z^* as the service rate μ changes between 0.1 and 1.9. For most of the experiments, the buffer size, K is fixed at 10 so that the effect of finite buffers is not negligible. However, we also report a case with larger buffers to analyse the effect of the buffer size restriction. The results of the experiments are suummarised in Tables 2–9.

Tables 2, 3 and 4 display the effect of increasing setup cost. As expected, the adjustment in policy for a setup cost increase is an increase in the difference $S - s$ forcing less setups per unit time. The difference in optimal revenue is negligible for very high and very low values of the service rate but can cause a difference of about 3% for a 4-fold increase in the setup costs. Figure 1 is a summary taken from Tables 2, 3 and 4 (with $\mu = 1.1$ and D3).

Table 2 Summary results of data set 1

Cost parameters: $R_1 = 20$, $R_2 = 10$, $c_1 + c_2 = 5$, $h = 0.5$, *Buffer Size* $= 10$

	D1			*D2*			*D3*			*D4*			*D5*		
μ	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
0.1	0	2	10.02	0	2	10.02	0	2	10.02	0	2	10.02	0	2	10.02
0.3	0	3	11.51	0	3	11.49	0	3	11.48	0	3	11.45	0	3	11.42
0.5	1	4	13.13	1	4	13.08	1	4	13.00	0	4	12.88	0	4	12.80
0.7	1	5	14.91	1	5	14.79	1	5	14.62	1	6	14.34	1	6	14.15
0.9	2	6	16.63	2	6	16.41	2	7	16.16	2	7	15.69	1	7	15.41
1.1	4	8	18.08	4	8	17.81	4	9	17.48	3	9	16.89	2	9	16.54
1.3	6	10	18.98	6	10	18.75	6	10	18.44	5	10	17.84	4	10	17.49
1.5	8	10	19.36	8	10	19.21	7	10	18.98	6	10	18.47	5	10	18.21
1.7	8	10	19.53	8	10	19.43	8	10	19.27	7	10	18.88	6	10	18.70
1.9	9	10	19.62	8	10	19.55	8	10	19.44	7	10	19.15	7	10	19.04

Table 3 Summary results of data set 2

Cost parameters: $R_1 = 20$, $R_2 = 10$, $c_1 + c_2 = 10$, $h = 0.5$, *Buffer Size* $= 10$

	D1			*D2*			*D3*			*D4*			*D5*		
μ	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
0.1	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00
0.3	6	10	11.15	0	4	11.15	0	4	11.13	0	4	11.11	0	4	11.08
0.5	5	10	12.78	0	5	12.75	0	5	12.70	0	5	12.58	0	5	12.49
0.7	4	9	14.61	1	6	14.51	1	6	14.36	1	7	14.07	1	7	13.86
0.9	3	8	16.44	2	7	16.24	2	8	15.98	2	8	15.50	1	8	15.21
1.1	1	7	18.02	3	9	17.74	3	10	17.40	3	10	16.79	2	10	16.41
1.3	0	4	18.97	5	10	18.73	5	10	18.40	4	10	17.75	3	10	17.40
1.5	11	11	19.35	7	10	19.20	6	10	18.95	6	10	18.41	4	10	18.13
1.7	11	11	19.53	11	11	19.43	11	11	19.26	6	10	18.84	5	10	18.65
1.9	11	11	19.62	11	11	19.55	11	11	19.43	11	11	19.12	6	10	19.00

Table 4 Summary results of data set 3

Cost parameters: $R_1 = 20$, $R_2 = 10$, $c_1 + c_2 = 20$, $h = 0.5$, *Buffer Size* = 10

	D1			*D2*			*D3*			*D4*			*D5*		
μ	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
0.1	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00
0.3	0	5	10.63	0	5	10.62	0	5	10.62	0	5	10.60	0	5	10.57
0.5	0	6	12.27	0	6	12.24	0	6	12.19	0	6	12.09	0	6	11.99
0.7	1	7	14.15	1	7	14.05	1	7	13.91	0	8	13.66	0	8	13.46
0.9	1	8	16.16	1	8	15.95	1	9	15.70	1	10	15.23	1	10	14.89
1.1	3	10	17.94	3	10	17.64	3	10	17.27	2	10	16.59	2	10	16.17
1.3	11	11	18.96	5	10	18.70	4	10	18.33	4	10	17.60	3	10	17.21
1.5	11	11	19.35	11	11	19.20	11	11	18.95	11	11	18.33	4	10	18.00
1.7	11	11	19.53	11	11	19.42	11	11	19.26	11	11	18.82	11	11	18.60
1.9	11	11	19.62	11	11	19.55	11	11	19.43	11	11	19.12	11	11	19.00

The effect of increased holding costs on the optimal policy can be viewed in Tables 2, 5 and 6. The first notable point here is that a change in holding costs lead to significant changes in the values of the parameters s and S. The striking difference caused by the change in holding costs is not only the difference $S - s$ but the positions of s and S as well. For $\mu = 1.1$ and D5 for example, $s = 2$ and $S = 9$ for a holding cost of \$0.5 per item per unit time and $s = 0$ and $S = 4$ for a holding cost of \$2 per item per unit time. Intuitively, for large values of the holding cost the system reduces the effect of holding cost by keeping the maximum number of items that can be buffered at a low level. Note that, this supports our argument in the section on the stationary queue length distribution, where it was suggested that expected queue lengths and hence the holding costs decrease by decreasing the upper threshold. Figure 2, taken from Tables 2, 5 and 6 (with $\mu = 1.1$ and D = 3) displays this decrease in S as holding costs increase.

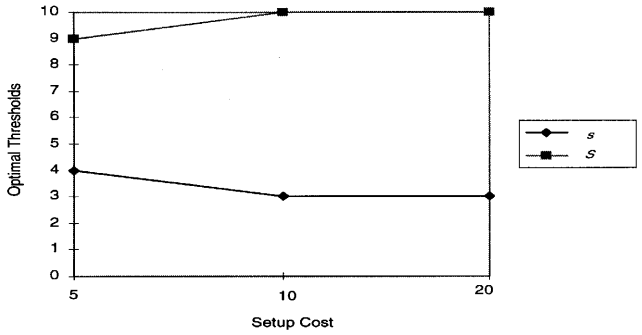

Figure 1 Change in optimal thresholds with respect to setup cost ($\lambda = 1$, $\mu = 1.1$, distribution = D3, $R_1 = 20$, $R_2 = 10$, $h = 0.5$, buffer size = 10)

Tables 7, 2, 8 display the effects of an increase in the off-time revenues. For moderate values of the service rate, an increase in the secondary revenue does not affect $S - s$ significantly, however pushes the optimal values of s lower. For $\mu = 1.1$ and D3 the values of the optimal pair are (5,10) and (2,7) respectively for $R_2 = 5$ and $R_2 = 20$. Figure 3, extracted from Tables 7, 2 and 8 (with $\mu = 1.1$ and D3), displays this effect.

The buffer size effect can be viewed in Tables 2 and 9 in which the buffer sizes are 10 and 20 respectively. Increased buffer size improves the performance only slightly, for the given set of cost parameters. The values of s and S vary, but the optimal revenue does not change significantly. For

Table 5 Summary results of data set 4

Cost parameters: $R_1 = 20$, $R_2 = 10$, $c_1 + c_2 = 5$, $h = 1$, *Buffer Size* = 10

	D1			*D2*			*D3*			*D4*			*D5*		
μ	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
0.1	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00
0.3	0	2	10.81	0	2	10.80	0	2	10.78	0	2	10.75	0	2	10.74
0.5	0	3	12.25	0	3	12.21	0	3	12.15	0	3	12.04	0	3	11.98
0.7	1	4	13.84	1	4	13.70	0	4	13.53	0	4	13.29	0	4	13.17
0.9	1	4	15.48	1	5	15.24	1	5	14.96	1	5	14.50	1	5	14.29
1.1	2	6	16.96	2	6	16.63	2	6	16.24	1	6	15.61	1	5	15.36
1.3	3	7	18.07	3	7	17.73	3	7	17.30	2	7	16.58	2	6	16.29
1.5	5	9	18.72	5	9	18.45	4	9	18.08	3	9	17.38	2	8	17.09
1.7	7	10	19.06	6	10	18.86	6	10	18.58	4	10	17.98	3	9	17.74
1.9	8	10	19.25	7	10	19.11	7	10	18.89	6	10	18.41	4	10	18.24

Table 6 Summary results of data set 5

Cost parameters: $R_1 = 20$, $R_2 = 10$, $c_1 + c_2 = 5$, $h = 2$, *Buffer Size* = 10

	D1			*D2*			*D3*			*D4*			*D5*		
μ	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
0.1	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00
0.3	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00	0	11	10.00
0.5	0	2	11.01	0	2	10.97	0	2	10.92	0	2	10.85	0	2	10.82
0.7	0	3	12.35	0	3	12.24	0	3	12.11	0	3	11.91	0	2	11.87
0.9	1	3	13.83	0	3	13.59	0	3	13.35	0	3	13.02	0	3	12.94
1.1	1	4	15.32	1	4	14.96	1	4	14.55	0	4	14.00	0	4	13.86
1.3	2	5	16.57	1	5	16.13	1	5	15.63	0	4	14.91	0	4	14.76
1.5	2	6	17.51	2	6	17.08	2	6	16.54	1	5	15.75	1	5	15.58
1.7	3	7	18.12	3	7	17.76	2	7	17.27	1	6	16.45	1	5	16.29
1.9	4	8	18.49	4	8	18.22	3	8	17.80	2	7	17.05	2	6	16.89

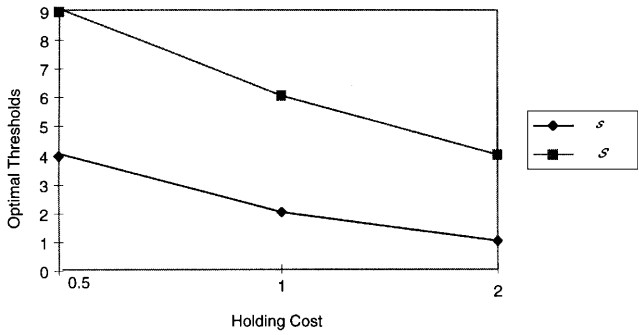

Figure 2 Change in optimal thresholds with respect to holding cost ($\lambda = 1$, $\mu = 1.1$, distribution = D3, $R_1 = 20$, $R_2 = 10$, $c_1 + c_2 = 5$, buffer size = 10)

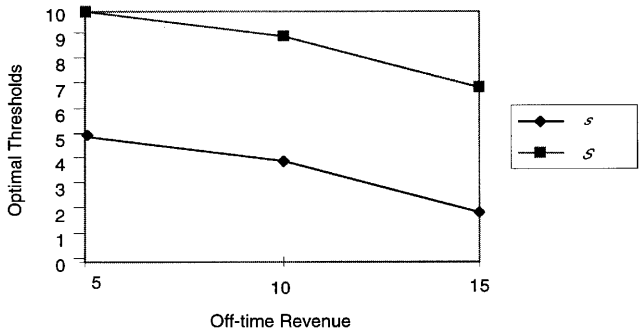

Figure 3 Change in optimal thresholds with respect to off-time revenues ($\lambda = 1$, $\mu = 1.1$, distribution = D3, $R_1 = 20$, $h = 0.5$, $c_1 + c_2 = 5$, buffer size = 10)

example, for $\mu = 1.9$ and for D5, the values of (s, S) are (7,10) versus (12,20) in Tables 2 and 9 respectively. However, the corresponding change in optimal revenue is from 19.04 to 19.10 only. In many cases, the effect of the increase on the buffer size on the total revenue is not even visible in the tables as the change occurs only in the 8th or 9th decimal place (and the tables only display the first two decimal places).

The degrading effect of the variability of the arrival process in the performance of the system is apparent from all of the Tables 2–9. In all of the experiments reported, optimal revenue of the system decreases as the coefficient of variation of arrival process increases. The difference in the optimal revenue between D1 and D5 can be as large as 9.3% as in the row corresponding to $\mu = 1.1$ of Table 2. Even so, variability does not seem to affect the optimal

Table 7 Summary results of data set 6

Cost parameters: $R_1 = 20$, $R_2 = 5$, $c_1 + c_2 = 5$, $h = 0.5$, *Buffer Size* = 10

	D1			*D2*			*D3*			*D4*			*D5*		
μ	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
0.1	0	2	5.50	0	2	5.50	0	2	5.50	0	2	5.50	0	2	5.49
0.3	0	3	7.87	0	3	7.85	0	3	7.83	0	3	7.77	0	3	7.74
0.5	1	4	10.51	1	4	10.42	1	5	10.31	1	5	10.11	1	5	9.95
0.7	2	5	13.18	2	6	13.01	2	6	12.81	2	7	12.38	1	7	12.06
0.9	3	7	15.76	3	7	15.49	3	8	15.14	3	9	14.49	3	9	14.01
1.1	5	9	17.85	5	10	17.52	5	10	17.10	5	10	16.29	4	10	15.73
1.3	8	10	18.96	8	10	18.70	7	10	18.32	7	10	17.54	6	10	17.06
1.5	11	11	19.35	11	11	19.20	11	11	18.95	11	11	18.33	7	10	17.98
1.7	11	11	19.53	11	11	19.43	11	11	19.26	11	11	18.82	11	11	18.60
1.9	11	11	19.66	11	11	19.55	11	11	19.43	11	11	19.12	11	11	18.99

Table 8 Summary results of data set 7

Cost parameters: $R_1 = 20$, $R_2 = 15$, $c_1 + c_2 = 5$, $h = 0.5$, *Buffer Size* = 10

	D1			*D2*			*D3*			*D4*			*D5*		
μ	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
0.1	0	11	15.00	0	11	15.00	0	11	15.00	0	11	15.00	0	11	15.00
0.3	0	3	15.14	0	3	15.14	0	3	15.13	0	3	15.11	0	3	15.11
0.5	0	3	15.88	0	3	15.86	0	4	15.83	0	4	15.78	0	4	15.74
0.7	0	4	16.72	0	4	16.65	0	4	16.57	0	5	16.45	0	4	16.38
0.9	1	5	17.61	1	5	17.48	1	6	17.33	0	6	17.09	0	5	16.98
1.1	2	6	18.42	2	6	18.24	2	7	18.03	1	7	17.70	1	6	17.55
1.3	3	8	19.02	3	8	18.84	3	8	18.62	2	8	18.23	1	7	18.06
1.5	5	10	19.36	4	10	19.22	4	10	19.03	3	10	18.66	2	9	18.50
1.7	7	10	19.53	6	10	19.43	5	10	19.28	4	10	18.97	3	10	18.85
1.9	8	10	19.62	7	10	19.55	6	10	19.44	5	10	19.19	4	10	19.11

Table 9 Summary results of data set 8

Cost parameters: $R_1 = 20$, $R_2 = 10$, $c_1 + c_2 = 5$, $h = 0.5$, *Buffer Size* = 20

	D1			*D2*			*D3*			*D4*			*D5*		
μ[a]	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*	s	S	z^*
1.3	6	11	18.98	6	11	18.75	6	12	18.45	5	12	17.86	4	11	17.50
1.5	10	15	19.36	9	15	19.22	9	15	19.00	8	15	18.54	6	14	18.25
1.7	14	20	19.53	13	20	19.43	13	20	19.29	11	19	18.95	9	17	18.77
1.9	17	20	19.62	17	20	19.55	16	20	19.44	14	20	19.20	12	20	19.10

[a] For $0.1 \leqslant \mu \leqslant 1.1$, the values of s, S, and z^* are identical to those in the corresponding rows of Table 2.

(s, S) pairs significantly. In general, an increase in the coefficient of variation increases $S - s$ but the increase in S is not that significant. Figure 4, taken from Table 2 (with $\mu = 1.1$) displays the change in revenue for three levels of the traffic load.

Conclusion

In this paper, we analysed the arrival control problem for a *lossless* finite buffered queue. Our model allows setup costs for interrupting and restarting the arrivals, holding costs for customers, and revenues that depend on the status (on/off) of the arrival stream. To gain insight into the structure of the optimal policy, we related the the queue lengths in the arrival control problem to the queue lengths in a service control problem. This queue length relationship also led to a relationship in terms of the optimal policies through which properties on the optimal control parameter values can be obtained in the case of negligible holding costs.

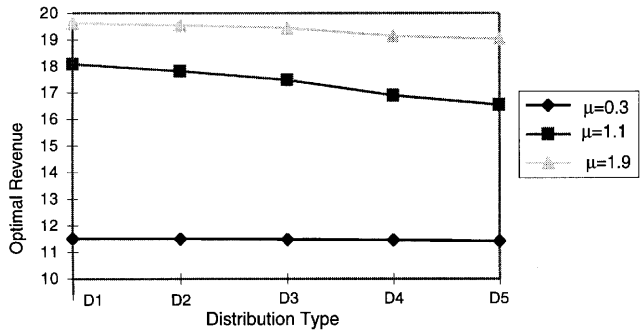

Figure 4 Change in optimal revenue with respect to arrival distribution ($\lambda = 1$, $R_1 = 20$, $R_2 = 10$, $h = 0.5$, $c_1 + c_2 = 5$, buffer size = 10)

When holding costs are not negligible, the properties of the parameters of the optimal policy are not evident from the related service control problem. For this case, we presented a fast method to compute the optimal revenue for a given set of policy parameters. The method permits phase-type arrival distribution that can model a wide range of arrival distributions with varying coefficients of variation. Using this method, we experimented with different traffic and cost structures to study the effects of setup and holding costs, off-time revenues, buffer size and arrival process variability on the optimal revenue and the parameters of the optimal policy. As a result of the numerical examples, the considerable effects of the cost structure on the optimisation problem was revealed. It was also seen that the variability of the arrival process has a significant effect on the optimal revenue. Based on these results, it is likely that the variability of the service process also has deteriorating effect on the optimal revenue. This motivates

future research where both arrival and service time distributions are modeled by general distributions. If the effects of variability in both service and arrival times can be quantified, it may be possible to allocate marginal effort on demand/supply management in a more effective way.

References

1 Buzacott JA, Liu XG and Shantikumar JG (1995). Multistage flow line analysis with the stopped arrival queue model. *IIE Trans* **27**: 444–455.
2 Buzacott JA and Shantikumar JG (1993). *Stochastic Models of Manufacturing Systems*. Prentice Hall: New Jersey.
3 Yadin M and Naor P (1963). Queueing systems with a removable service station. *Opl Res Q* **14**: 393–405.
4 Heyman D (1968). Optimal operating policies for M/G/1 queueing systems. *Opns Res* **16**: 362–382.
5 Harris T (1967). Duality of finite Markovian queues. *Opns Res* **15**: 575–576.
6 Gupta SM (1993). Duality in truncated steady state Erlang distribution based queueing processes. *J Opl Res Soc* **44**: 253–257.
7 Gupta SM (1994). Finite source Erlang based queueing systems: complementarity, equivalence and their implications. *Computers and Math with Applic* **28**: 57–74.
8 Gupta SM (1994). Interrelationship between queueing models with balking and reneging and machine repair problem with warm spares. *Microelect and Reliability* **34**: 201–209.
9 Gupta SM (1995). Queueing model with state dependent balking and reneging: its complementary and equivalence. *Perf Eval Rev* **22**: 63–72.
10 Gupta SM and Melachrinoudis E (1994). Complementarity and equivalence in finite source queueing models with spares. *Computers Opns Res* **21**: 289–296.
11 Hlynka M and Wang T (1993). Comments on duality of queues with finite buffer size. *Opns Res Lett* **14**: 29–33.
12 Yang P (1994). A unified algorithm for computing the stationary queue length distributions in M(k)/G/1/N and GI/M(k)/1/N queues. *Que Sys* **17**: 383–401.
13 Gupta SM (1995). Interrelationships between controlling arrival and service in queueing systems. *Computers Opns Res* **22**: 1005–1014.
14 Karaesmen F and Gupta SM (1997). Duality relations for queues with arrival and service control. *Computers Opns Res* **24**: 529–538.
15 Sparaggis P, Cassandras C and Towsley D (1993). On the duality between routing and scheduling systems with finite buffer space. *IEEE Trans Autom Control* **38**: 1440–1446.
16 Xu SH and Shantikumar JG (1993). Optimal expulsion control—a dual approach to admission control of an ordered entry system. *Opns Res* **41**: 1137–1152.
17 Xu SH (1994). A duality approach to admission and scheduling controls of queues. *Que Sys* **18**: 273–300.
18 Sobel M (1969). Optimal average cost policy for a queue with start-up and shut-down costs. *Opns Res* **17**: 145–162.
19 Hersh M and Brosh I (1980). The optimal strategy structure of an intermittently operated service channel. *Eur J Opl Res* **5**: 133–141.
20 Teghem Jr J (1987). Optimal control of a removable server in an M/G/1 queue with finite capacity. *Eur J Opl Res* **31**: 358–367.
21 Karaesmen F and Gupta SM (1996). Service control in a finite buffered queue with holding and setup costs. Technical Report, Dept. of Mech., Ind. and Manuf. Eng., Northeastern University.
22 Takagi H (1993). M/G/1/K queues with N-policy and setup times. *Que Sys* **14**: 79–98.
23 Morse PM (1958). *Queues*, *Inventories and Maintenance*. John Wiley and Sons: New York.
24 Neuts MF (1981). *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press: Baltimore.
25 Keilson J and Servi L (1989). Blocking probability for M/G/1 vacation systems with occupancy level dependent schedules. *Opns Res* **37**: 134–140.
26 Neuts MF and Rao BM (1992). On the design of a finite-capacity queue with phase-type service times and hysteretic control. *Eur J Opl Res* **62**: 221–240.

Received May 1996;
accepted July 1997 after two revisions