Pergamon

PII: S0305-0548(96)00072-X

Computers Ops Res. Vol. 24, No. 6, pp. 529–538, 1997

0305-0548/97 $17.00+0.00

DUALITY RELATIONS FOR QUEUES WITH ARRIVAL AND SERVICE CONTROL

Fikri Karaesmen‡ and Surendra M. Gupta†§

Department of Mechanical, Industrial and Manufacturing Engineering, 334 Snell Engineering Center, Northeastern University, Boston, MA 02115, U.S.A.

(Received November 1995; in revised form August 1996)

Scope and Purpose—We present queue length duality results for queues with finite buffer space. Our results apply to queueing systems in which either arrivals or services are subject to control. These queues are common in manufacturing systems where the processing or the routing of parts can be stopped and restarted. Our results are free of the common restrictions (i.e. exponential distributions and independence) for arrival and service time distributions. The duality relationship enables us to obtain the stationary queue length distributions for two systems at once. We demonstrate the usefulness of this feature through examples.

Abstract—We consider finite buffered queues with service or arrival control. In the case of service control, service may be stopped and restarted depending on the queue length. In the case of arrival control, the arrival stream can be turned off and on or arrivals may be rejected depending on the queue length. We give duality relations for various systems with arrival and service control that enables us to relate their stationary queue length distributions. We use physical coupling arguments which imply the stochastic coupling necessary to relate the queue lengths. We also discuss special cases for which queue length relationships can be obtained by analyzing the underlying Markov process. Two examples are provided to demonstrate the application of the duality property. The first example is a case where the existing queue length distribution for a given model can be used to obtain the queue length distribution of another model. In the second example, we obtain the previously unknown queue length distributions for two related models at once. © 1997 Elsevier Science Ltd

INTRODUCTION

In many instances, queueing systems are controlled by turning off the service or the arrival process. The service control problem was introduced by Yadin and Naor [1] in the context of the M/G/1 queue. Yadin and Naor's problem includes holding costs for customers waiting in line and set up costs for starting and stopping the server. To optimize the system, they suggest using the following operating policy: shut down the server when the server becomes idle and restart it when the queue length reaches a threshold level N. This operating policy was termed '*N-policy*'. Heyman [2] proves that N-policy is the optimal policy for operating M/G/1 queues under various cost criteria.

Now, consider the case where the service process cannot be interrupted but instead arrivals can be controlled by turning off the arrival stream. For this arrival control policy, the arrivals can be stopped when the queue is full and restarted when the queue length decreases to a certain threshold level F. This policy is called '*F-policy*' [3].

In this article, we study the relationships between the queue lengths for finite capacity queues operating under N-policy [4] and F-policy. In particular we study duality relationships for the queue lengths as previously studied by Gupta [5–8] and by Gupta and Melachrinoudis [9] for uncontrolled Markovian queues and Gupta [3] for controlled Markovian queues. The queue length duality for finite buffered queues is a relationship between two queueing systems of equal buffer capacity K. Consider, two such queueing systems referred to as the '*primal*' (P) and '*dual*' (D) systems respectively. Let $\pi_i^P \wedge \pi_i^D$ (for $i=0, 1, 2..., K$) denote the stationary probability of having i customers in the primal and dual systems respectively. The duality relationship is given by:

† To whom all correspondence should be addressed (e-mail: gupta@neu.edu).

‡ Fikri Karaesmen obtained a B.S. degree in Industrial Engineering from Middle East Technical University in Turkey and an M.S. and Ph.D. degrees in Industrial Engineering from Northeastern University. His research interests are in the area of applied probability and optimization. His work has been published in several journals and conference proceedings.

§ Surendra M. Gupta is an Associate Professor of Industrial Engineering at Northeastern University in Boston. He received his MBA from Bryant College and MSIE and Ph.D. in Industrial Engineering from Purdue University. His research interests are in the area of Operations Research and Production Systems. He has contributed to both national and international conferences. In addition, his publications have appeared in a variety of journals including the *Computers and Industrial Engineering*, *Computers and Mathematics with Applications*, *Computers and Operations Research*, *Journal of Operational Research Society*, *Microelectronics and Reliability*, *Performance Evaluation* and *Transactions of Operational Research*. He is a registered Professional Engineer in the State of Massachusetts.

$$\pi_i^{\mathrm{P}} = \pi_{K-i}^{\mathrm{D}} \text{ for } i=0,1,2,\ldots,K. \quad (1)$$

Gupta [3], established the duality relationship in Eq. (1) for finite buffered Markovian queues under arrival and service control. Here we consider a more general setup that allows us to establish similar relationships for non Markovian systems. In related work, duality relations of finite buffer size M/G/1 type queueing systems were previously studied by Harris [10], Hlynka and Wang [11] and Yang [12]. In Hlynka and Wang [11], duality relations were introduced for finite buffered G/G/1 queues without arrival or service control. Yang [12] studied the case of queue length dependent arrivals and service and allowed multiple service (or arrival) rates. However, the inter-departure (or the interarrival) times were restricted to be exponential. Note that, our definition of 'state dependency' is stronger than in [12] as the status of the service or (arrival) process is also part of the state description.

The service control (N-policy) or arrival control (F-policy) problems studied in this article frequently arise in capacitated production systems. For systems with infinite production capacity that are frequently studied in inventory theory, a classical problem is the determination of the economic order quantity. The service control problem for a queueing system is analogous to production systems with finite production capacity and random demands and production times. The arrival control problem also deals with determination of the optimal production batch sizes for a two stage capacitated production system where control can be applied in the first stage only.

The notion of duality has recently been used to tackle some queueing control problems. Sparragis *et al.* [13] note the duality between routing and scheduling problems for finite buffered multiple class queues. Xu and Shantikumar [14] and Xu [15] use the duality between admission control and expulsion control for a queueing system to obtain new structural results on these problems. The identification of arrival and service control problems as duals of each other is also an initial step in this direction.

The main utility of the duality relationships studied in this article is in that, queue length distributions for two systems can be obtained at once. This is particularly helpful when one of the systems has a known queue length probability distribution and the other does not (though they are equally difficult in structure, one may seem less intuitive than the other). To establish the duality relationships, we introduce *physically* dual systems. This distinguishes our approach from the previous research in similar problems. In the sections to follow, we first discuss the case of controllable arrivals. To demonstrate the approach, we study the M/G/1/K queue operating under N-policy in detail. Next, we consider queue length duality for the case of uncontrollable or uninterruptable arrivals. Finally, we present two examples that highlight the use of the duality relations and provide some conclusions.

THE CASE OF CONTROLLABLE ARRIVALS

We first study a model where the arrival stream can be turned on and off at the controller's discretion. We derive duality relationships for queues with controllable arrivals and controllable service which generalize some of the results obtained thus far.

In the rest of the article, the terms '*controlled*' arrival stream and '*interrupted*' arrival stream will be used frequently. In particular, controlled arrival streams refer to F-policy queues where the arrival stream is stopped when the queue is full and restarted when the queue length drops to F. In the context of this article '*interrupted*' arrivals refer to queue operating under N-policy, where the arrival stream is turned off when the buffer is full and turned on when the buffer space becomes available. To summarize, arrivals can be controlled by stopping and restarting the arrival stream for F-policy queues and can be interrupted when the queue is full for N-policy queues.

Our general setup is as follows: we consider G/G/1/K queues for which either the service process or the arrival process can be shut down and restarted. Interarrival times, A and service times, S are random variables with mean $1/\lambda$ and $1/\mu$ respectively. When the service process is controlled the server can be stopped (removed) depending on the length of the queue, however, the removal of the server itself may take a random time denoted by α (with distribution F_α). The server can then be turned on depending on the queue length after a random delay denoted by β (with distribution F_β). Similarly, the arrival process can be stopped and restarted depending on the queue length with random delays at stopping and restarting times. For a concise description, we use the shorthand notation GI/G/1/K ($N/\alpha_N/\beta_N$) for the N-policy system and GI/G/1/K ($F/\alpha_F/\beta_F$) for the F-policy system operating under service and arrival control with interruption delay times α_i and restart delay times β_i respectively ($i=N$ or F).

The duality relationships that we seek are relationships between random variables. The relationships between the variables are difficult to establish. Rather than trying to characterize these random variables

explicitly, we consider coupled physical systems that have the desired duality property. In the case of controllable arrivals, the duality phenomenon can be explained by viewing the finite capacity queue as a node in a cyclic network of two tandem queues. Lavenberg [16] used this fact to obtain the waiting times in an M/G/1 queue with finite capacity.

We state the following theorem for the case of controlled arrivals or service.

Theorem 1: Let $\pi_i^N,(i=0,1,2,...,K)$ be the steady state probability of having i customers in a $G_1/G_2/1/K$ $(N/\alpha_N/\beta_N)$ queue with interruptable arrivals. Also, let $\pi_i^F,(i=0,1,2,...,K)$ be the steady state queue length probability in a $G_2/G_1/1/K$ $(F/\alpha_F/\beta_F)$ queue with controllable arrival stream. If $N=K-F$, $\alpha_N=\alpha_F$ and $\beta_N=\beta_F$ then:

$$\pi_i^N=\pi_{K-i}^F \text{ for } i=0,1,2,...,K \tag{2}$$

Proof: To prove the above relationship, we only need to argue that there is a physically coupled dynamic system for which the above relationship holds. If such a dynamic system can be found, then the existence of a stationary queue length distribution is sufficient for Eq. (2) to hold. Consider a cyclic network of two tandem nodes and K customers and let $\{\omega_n\},\{\eta_n\},\{\alpha_n\}$ and $\{\beta_n\}$ be the sequences of service times at node 1, service times at node 2, shut-down delay times and startup delay times respectively. Without loss of generality, assume that node 1 of the tandem network operates under N-policy (where $1\leq N\leq K$), i.e. the dynamics of the first node of this network is identical to the dynamics of a $G_1/G_2/1/K$ $(N/\alpha_N/\beta_N)$ queue with interrupted arrivals when the same input sequences are used. Note that, $\{\omega_n\}$ is sampled from G_2, $\{\eta_n\}$ from G_1, $\{\alpha_n\}$ from F_α and $\{\beta_n\}$ from F_β.

As a result of the cyclic structure, the departure times from each node are also the arrival times to the other node. Next, we argue that if node 1 operates under N policy in the above network, then node 2 is operating under F-policy. To understand this property, let $L(t)=\{L_1(t), L_2(t)\}$ be the queue length at nodes (1,2) at time t. Consider a departure instance, τ, from node 1 that leaves the first queue empty (i.e. $L(\tau^-)=\{1,K-1\}$ and $L(\tau^+)=\{0,K\}$). As a result of N-policy, at time τ the service process is turned off at node 1 after a delay of α_l, which in turn implies the arrival process is turned off at node 2 after a delay of α_l. Similarly, the first time after $\tau+\alpha_l$ that the service process has to be restarted is $\sigma+\beta_l$, where σ is the departure instance from node 2 $(\sigma>\tau+\alpha_l)$ such that $L(\sigma^-)=\{N-1, K-N+1\}$, and β_l is the startup delay. Therefore, at time $\sigma+\beta_l$, a restart for service is initiated which is equivalent to a restart for the arrival process for node 2. In other words, the arrival turn off is initiated at node 2 when the queue length is K with the next arrival restart initiation occuring when the queue length is $K-N$ which is F-policy (with $F=K-N$) with controlled arrivals as described earlier. That is, the sequences of inputs to node 2 can be used as the sequences of inputs to a $G_2/G_1/1/K$ $(K-N/\alpha_N/\beta_N)$ queue with $\{\omega_n\}$ as the interarrival time sequence and $\{\eta_n\}$ as the service time sequence.

Once the equivalence of the cyclic network to N-policy and F-policy queues is clear, the statement of duality in Eq. (2) immediately follows as $L_1(t)+L_2(t)=K$, for all $t\geq 0$ and the queue lengths at both queues change at the same time (either by a departure from node 1 or by a departure from node 2). As the tandem network is regenerative, by regeneration arguments the state occupation times in state, $\{j,\bullet\}$ is equal to the state occupation times in state $\{\bullet,K-j\}$ for $j=0, 1, 2,..., K$ in a regenerative cycle. In other words, starting from identical conditions and using the same sequences of random variables the duality relationship is always satisfied. Therefore we have constructed a case where duality holds with probability one. However, as the stationary queue length distributions are unique and do not depend on the initial conditions, the duality relationship holds regardless of initial conditions and the particular random variables sequence used. This establishes the equality in Eq. (2).

The above proof does not require the fact that interarrival and service time distributions be independent. Thus, the interarrival, service, startup and shutdown times can be dependent processes and can have dependencies between each other (as long as stationary queue length distributions exist).

An immediate extension to the theorem is to consider duality of the queue length probabilities at certain embedded epochs. For example, important performance measures for finite buffered queues such as blocking probabilities and waiting time distributions require the computation of the queue length distributions at arrival epochs. The following lemma establishes duality relationships for N-policy and F-policy queues for queue length probabilities at certain embedded epochs.

Lemma 1: Let $p_i^N(i=0,1,2,...,K)$ be the (steady state) queue length probability at arrival times for a $G_1/G_2/1/K$ (N/0/0) queue and let $p_j^F(j=0,1,2,...,K)$ be the queue length probability at service completion times for a $G_2/G_1/1/K$ (F/0/0) queue with controllable arrival stream. Then:

$$p_i^N = p_{K-i}^F \tag{3}$$

Proof: Once again viewing the two queues in tandem, it can be argued that the embedded points are the departure sequence from node 1, $\{\tau_i, i=0,1,2,...\}$. As $L_1(\tau_i)+L_2(\tau_i)=K$, for all τ_i, $i=1,2,...$, then the number of transitions into state $\{j,\bullet\}$ must equal the number of transitions into state $\{\bullet, K-j\}$ for all $j=0,1,2,...$, K. This implies Eq. (3).

Remark: To establish the equivalence in Eq. (3), the embedded epochs have to be selected carefully. In particular, if embedded epochs for the N-policy queue are the epochs immediately after a service completion, the embedded epochs for the F-policy queue are the epochs immediately after an arrival.

Example: M/G/1/K under N-policy

We use the M/G/1/K (N/0/0) queue to clarify the duality relationships in Eqs (2) and (3). Note that, with Poisson arrivals the assumption of an '*interrupted arrival stream*' to the N-policy queue is unnecessary. As a result of our definition for an interrupted arrival stream, the arrival stream is turned on when space becomes available. At this instance, for the M/G/1/K queue, the remaining time until an arrival is an exponentially distributed random variable. By the memoryless property of the exponential distribution, the remaining arrival time would be identical to the regular interarrival time as if the arrivals were not interrupted.

Duality at embedded epochs

Consider the M/G/1/K (N/0/0) queue with arrival rate λ and S denoting the random variable that corresponds to service times. We denote by $F_S(x)$, the distribution function of S. Let g_i ($i=0, 1, 2,...$) be the probability of having i arrivals in a service time, i.e.:

$$g_i = \int_0^\infty \frac{e^{-\lambda x}(\lambda x)^i}{i!} dF_S(x). \tag{4}$$

Also let g_i^c be the probability that there are more than i arrivals during a service time, i.e.:

$$g_i^c = \sum_{j=i+1}^{\infty} g_j. \tag{5}$$

In the arbitrary time process, the state of the M/G/1/K (N/0/0) queue has to be represented with a couplet (i,j) where i denotes the queue length ($i=0, 1, 2,..., K$) and j denotes the status of the server with $j=0$ and $j=1$ corresponding to the server being off and on respectively. Now consider the Markov chain embedded at service completion epochs and service *restart* epochs for this system. The restart epochs are those points in time where the server is turned on after an off period. One advantage of this particular selection of embedded epochs is that it enables us to use a one dimensional state space $\{0, 1, 2,..., K-1\}$ where state 0 denotes an aggregate *off* state. In the one dimensional representation, all states $\{0,..., N-1\} \times \{0\}$ (in the two dimensional representation) are lumped into a single state 0 (disaggregation of this lumped state will be handled later). Note that, this chain resembles the embedded chain of the regular M/G/1/K queue. However, in this case, transitions starting from state 0 can only occur to state N as the service process is turned off when the queue length becomes zero.

Let p_i, $i=0, 1, 2..., K-1$, denote the steady state probability of being in state i in the embedded chain, then the probability distribution of the queue length observed at the selected times can be obtained as the solution of the following system:

$$p_i = \sum_{j=0}^{i} p_{i+1-j} g_j \; for \; i=0,1,...,K-2; i \neq N \tag{6}$$

$$p_N = P_0 + \sum_{j=0}^{N} p_{N+1-j} g_j \tag{7}$$

$$p_{K-1} = \sum_{j=0}^{K-2} p_{j+1} g_{K-2-j}^c \tag{8}$$

along with the normalizing condition:

$$\sum_{i=0}^{K-1} p_i = 1 \tag{9}$$

Similarly, consider a GI/M/1/K (F/0/0) queue with mean service time $1/\mu$ and the random variable

denoting the time between arrivals as A. We denote by h_i (i=0,1,2,...) the probability that there are i service completions between two consecutive arrivals, then:

$$h_i = \int_0^\infty \frac{e^{-\mu x}(\mu x)^i}{i!} dF_A(x). \tag{10}$$

and

$$h_i^c = \sum_{j=i+1}^{\infty} h_i \tag{11}$$

where h_i^c is the probability that there are more than i service completions between two consecutive arrivals.To match the M/G/1/K (N/0/0) system, let the embedded times be the instances immediately after arrivals and the end of off periods for the arrival stream. Note that this is a different selection than the embedded chain at pre-arrival epochs that is common for the GI/M/1/K queue. In fact our selection is valid only because the arrivals are not blocked (since $F<K$) and so it is not necessary to account for arrivals that do not change the state of the system.

Once again all *off* states can be lumped into a single state K and the state space for the embedded Markov Chain is {1, 2, 3,..., K}. The steady state queue length distribution at embedded times denoted by p_i' is obtained as the solution of:

$$p_1' = \sum_{j=1}^{K-1} p_j' h_{j-1}^c \tag{12}$$

$$p_i' = \sum_{j=0}^{K-i+1} p_{j+i-1}' h_j \text{ for } i=2,3,...,K; i \neq F \tag{13}$$

$$P_F' = \sum_{j=0}^{K-F+1} p_{j+F-1}' h_j + p_K' \tag{14}$$

where

$$\sum_{j=1}^{K} p_j' = 1 \tag{15}$$

Lemma 2: (special case of Lemma 1) Consider the queue length distributions at above specified embedded times for the M/G/1/K (N/0/0) and the GI/M/1/K (F/0/0) queues. Let p_i (i=0, 1, 2,..., $K-1$) and p_j'(j=1,2,...,K) denote respectively the stationary queue length distribution at the specified embedded times for the M/G/1/L (N/0/0) and the GI/M/1/K (F/0/0) queues respectively. If $\lambda=\mu$, $A=S$ and $N=K-F$ then:

$$p_i = P_{K-i}' \text{ for } i=0,1,2,...,K-1 \tag{16}$$

Proof: If we relabel the states of the first (M/G/1/K (N/0/0)) embedded chain in reverse order (such that state i becomes state $K-i$ (for i=0, 1, 2,..., $K-1$), we obtain a system of equations identical to Eqs (12)–(15). Going back to the original state labeling gives the relationship in Eq. (16).

Remark: The above relationship is noted by Harris [10] and Hlynka and Wang [11] for the case of N=1 and $F=K-1$. In addition, the above relationship can be easily generalized for any arbitrary time.

DUALITY WITH EXOGENOUS ARRIVALS AND REJECTION

In this section, we study the case where the arrival stream is exogenous and therefore not controllable or interruptable. The only decision in this case is whether to accept or reject an arriving customer from the arriving stream. When arrivals are interruptable, the stream is turned on as soon as the queue length reaches the threshold, the time until the first arrival in this case is the random variable, A, with distribution F_A. In the case where arrivals are rejected when the buffer is full and accepted when there is a buffer space available, the time from the moment a buffer space becomes available until the first arrival, is not distributed as F_A. In this section we study such a model and derive its dual.

We first define the following service discipline: the server does not become idle when the queue length is zero but starts a virtual service. If an arrival occurs during a virtual service, then the service time of the first arrival is not a regular service time but the remaining time of the ongoing virtual service. If no arrival occurs, then the server starts another virtual service. In other words, the first customer served in

a busy period receives a special service while the rest of the customers receive regular service.

The above service discipline was referred to as the '*transportation variant*' of the M/G/1 queue by Keilson [17] who considered the case of Poisson arrivals. Iglehart and Whitt [18] used the identical discipline to obtain heavy traffic limits for the G/G/s queue.

Under this new service discipline, we modify the definition of N-policy as follows: the server does not become idle when the queue length becomes zero but instead starts a series of virtual services. As soon as N customers build up, the service will start for the first customer that arrived during the virtual services period and the actual service time of the first customer is equal to the remaining service time of the current virtual service time.

Theorem 2: Let $\pi_i^N,(i=0,1,2,...,K)$ be the steady state probability of having i customers in a $G_1/G_2/1/K$ (N/0/0) queue with virtual service. Also, let $\pi_i^F,(i=0,1,2,...,K)$ be the steady state queue length probability in a $G_2/G_1/1/K$ (F/0/0) queue with arrival rejection and virtual service. If $N=K-F$, then:

$$\pi_i^N=\pi_{K-i}^F \text{ for } i=0,1,2,...,K \tag{17}$$

Proof: To prove the above equality, we use the idea of *job-hole duality* of Gordon and Newell [19] to construct an analogous argument to that in the proof of Theorem 1. Let $(L_1(t),L_2(t))$ represent the state of the N-policy system where $L_1(t)$ denotes the number of jobs in the system and $L_2(t)$ denotes the number of holes (available buffer spaces). Arguing as in the proof of Theorem 1, the sequence of interarrival times $\{\omega_n\}$ for the *jobs* queue is the sequence of service times to the *holes* queue and the sequence of service times $\{\eta_n\}$ in the *jobs* queue is the sequence of interarrival times in the *holes* queue. To understand the dynamics, let τ be an instance where the server is turned off (i.e. $L_1(\tau^-)=1$ and $L_1(\tau^+)=0$) and let the first startup instance after τ be $\tau+\delta$ (i.e. $L_1(\tau+\delta^-)=N-1$ and $L_1(\tau+\delta^+)=N$). For $\tau<t<\tau+\delta$, the virtual service process continues, however the queue length does not change at the end of service completions. As $L_1(t)+L_2(t)=K$ for all $t>0$, at the instant that a virtual service period starts for the queue of jobs an arrival starts for the queue of holes where each virtual service completion corresponds to a rejected arrival and each real service completion corresponds to an accepted arrival. At time $\tau+\delta$, a virtual service is in effect. Let σ denote the remaining service time. According to our definition, σ is the actual service time for the first customer to be served after τ, i.e. the next queue length change due to a service completion takes place at $\tau+\delta+\sigma$ or in other words, $L_1(\tau+\delta+\sigma^+)=L_1(\tau+\delta+\sigma^-)-1$. Note that $\tau+\delta$ acts as a remaining arrival time for the queue of holes as $L_2(\tau+\delta+\sigma^+)=L_2(\tau+\delta+\sigma^-)+1$. Therefore, when the job queue operates as an N-policy queue, the hole queue operates as a F-policy queue and the equality in Eq. (17) follows by regeneration as $\{L_1(t)\}$ and $\{L_2(t)\}$ change states at the same time while preserving $L_1(t)+L_2(t)=K$.

Remark: It is possible to incorporate setup times to the above argument as in Theorem 1. The result remains unchanged.

EXAMPLES

In this section we consider two examples that emphasize the utility of the results obtained in this article. The first example is the case where the queue length distribution of one of the systems is already known. This immediately yields the queue length distribution of the second system. In the second example, we solve a new problem which implies that we have solved its dual as well.

M/G/1/K queue under N-policy

Consider the M/G/1/K (N/0/0) queue which was studied in detail in the second section. We continue to use the notation introduced in that section with the following additions:

B_K^N: random variable denoting the length of a busy (on) period for a K capacity queue operating under N-policy

B_K: random variable denoting the busy period for a regularly operating K capacity queue ($=B_K^1$).

Teghem Jr [20] gives the steady state queue length probability distribution for the M/G/1/K (N/0/0) queue in closed form in terms of expected busy periods of M/G/1/K queues. Takagi [21] covers the case with nonzero start up times. By the results of the second section, the queue length distributions of the dual systems are readily available. We discuss the M/G/1/K (N/0/0) case here (see [20] for details).

The key relationship for any busy periods of the M/G/1/K (N/0/0) queue is:

$$E[B_K^N] = \sum_{i=K-N+1}^{K} E[B_i] \tag{18}$$

The queue length distribution can be written in terms of the busy periods as (see Teghem Jr [20]):

$$\pi_0^N = \frac{1}{N+\lambda E[B_k^N]} \tag{19}$$

$$\pi_i^N = \frac{E[B_{i+1}]}{E[S](N+\lambda E[B_K^N])} \quad for\ 1 \le i \le N-1 \tag{20}$$

$$\pi_i^N = \frac{E[B_{i+1}^N] - E[B_i^N]}{E[S](N+\lambda E[B_K^N])} \quad for\ N \le i \le K-1 \tag{21}$$

$$\pi_K^N = \frac{NE[S] - (1-\rho)E[B_K^N]}{E[S](N+\lambda E[B_K^N])} \tag{22}$$

where $\rho = \lambda/\mu$ Using Theorem 1, we can obtain the dual queueing system by inspection. The dual of the above system is a GI/M/1/K (F/0/0) queue with the arrival and service processes interchanged in the first system and $F = K - N$. This yields the following steady state distribution for the queue length in the dual queue:

$$\pi_0^F = \frac{(K-F)E[A] - (1-(\mu/\lambda))E[B_K^{K-F}]}{E[A]((K-F)+\mu E[B_K^{K-F}])} \tag{23}$$

$$\pi_i^F = \frac{E[B_{(K-i)+1}^{K-F}] - E[B_{K-i}^{K-F}]}{E[A]((K-F)+\mu E[B_K^{K-F}])} \quad for\ 1 \le i \le F \tag{24}$$

$$\pi_i^F = \frac{E[B_{(K-i)+1}]}{E[A]((K-F)+\mu E[B_K^{K-F}])} \quad for\ F+1 \le i \le K-1 \tag{25}$$

$$\pi_K^F = \frac{1}{(K-F)+\mu E[B_K^{K-F}]} \tag{26}$$

GI/M/1/K queue under N-policy with controlled arrivals

Here, we study the GI/M/1/K (N/0/0) queue. As the arrival process can be non-Poisson, it is necessary to emphasize that the arrival mechanism is of the interrupted type. Hence, the arrival stream is stopped when the system is full and restarted when a space becomes available.

The queue length distribution. Let $p_{i,j}$ denote embedded probabilities where (i,j) is the queue length and the status of the server, respectively, at epochs immediately after an arrival. Let S_e denote the state space, then:

$$S_e = \{1,2,...N-1\} \times \{0\} \cup \{2,3,...K\} \times \{1\}$$

It can be seen that the embedded probabilities satisfy the following system of equations:

$$p_{1,0} = \sum_{i=2}^{K-1} h_i^c p_{i,1} + p_{K,1} h_{K-1}^c \tag{27}$$

$$p_{n,0} = p_{1,0}\ for\ 2 \le n \le N-1 \tag{28}$$

$$p_{n,1} = \sum_{i=0}^{K-n} p_{i+n-1,1} h_i + p_{K,1} h_{K-n}\ for\ 2 \le n \le N-1 \tag{29}$$

$$p_{N,1} = \sum_{i=0}^{K-N-1} p_{i+N,1} h_i + p_{K,1} h_{K-N-1} + p_{N-1,0} \tag{30}$$

$$p_{n,1} = \sum_{i=0}^{K-n} p_{i+n-1,1} h_i + p_{K,1} h_{K-n}\ for\ N+1 \le n \le K \tag{31}$$

The normalizing condition is:

$$\sum_{i=1}^{N-1} p_{i,0} + \sum_{i=2}^{K} p_{i,1} = 1 \tag{32}$$

To pass to the semi-Markov process, we need to consider the expected sojourn times, $T_{i,j}$ in each state. We have $T_{i,0}=E[A]=1/\lambda$ for $i=1, 2,\ldots N-1$ and $T_{i,1}=E[A]=1/\lambda$ for $i=2, 3,\ldots, K-1$. But the sojourn time in state $(K,1)$ includes the time that the arrival process is shut off and is given by $T_{K,1}=E[A]+E[S]$. For the steady state distribution of the semi-Markov process, we have:

$$q_{i,j}=\frac{p_{i,j}}{((1-p_{K,1})/\lambda)+p_{K,1}T_{K,1}} \quad for\ (i,j)\in S_e \tag{33}$$

where in writing the denominator, we have used the identity:

$$\sum_{(i,j)\in S_e-\{(K,1)\}} p_{i,j}=1-p_{K,1} \tag{34}$$

Finally, the above semi Markov process is related to the general time queue length process as follows:

$$\pi_{0,0}=\sum_{n=1}^{K-1} q_{n,1}(h_n^+)^c+(1-\kappa)q_{K,1}(h_K^+-1)^c \tag{35}$$

$$\pi_{n,0}=q_{n,0}\ for\ n=1,2,\ldots,N-1 \tag{36}$$

$$\pi_{n,1}=\sum_{i=0}^{K-n-1} q_{i+n,1}h_i^+ +(1-\kappa)q_{K,1}h_{K-n-1}^+\ for\ n=1,2,3,\ldots,K-1 \tag{37}$$

$$\pi_{K,1}=\kappa q_{K,1} \tag{38}$$

where

$$h_K^+=\int_0^x \frac{e^{-\mu x}(\mu x)^K}{K!}\,dF_{\hat{A}}, \tag{39}$$

$$(h_K^+)^c=\sum_{i=K+1}^{\infty} h_K^+ \tag{40}$$

and $\hat{A}$ is the backward recurrence time of A with distribution function

$$F_{\hat{A}}(x)=\int_0^x \frac{1-F_A(y)}{E[A]}\,dy \tag{41}$$

and $\kappa=E[S]/(E[S]+E[A])$. To gain some insight into the probabilistic meaning of κ, consider a random point that falls into state $(K, 1)$ of the semi Markov process. As mentioned before, a sojourn time in this state consists of a service interval and an arrival interval. In other words the n'th sojourn in state $(K, 1)$, $T_{(K,1),n}$ consists of two periods where the first period has distribution F_S and the second period has distribution F_A. Hence, κ is the probability that the random point falls in a service interval given that it falls in state $(K, 1)$.

Numerical results. As a numerical example for the GI/M/1/K (N/0/0) queue, consider the case where the arrival distribution is Erlang-2, i.e. $f_A(x)=(2\lambda)^2 e^{-2\lambda x}$. Let $\lambda=0.9$, $\mu=1$ and $K=15$.

For comparison, we consider the dual system to the one above, i.e. the M/G/1/K (F/0/0) queue with service distribution $f_S(x)=(2\mu)^2 e^{-2\mu x}$ where $\mu=0.9$, $\lambda=1$ and $K=15$. Fig. 1 displays the average queue

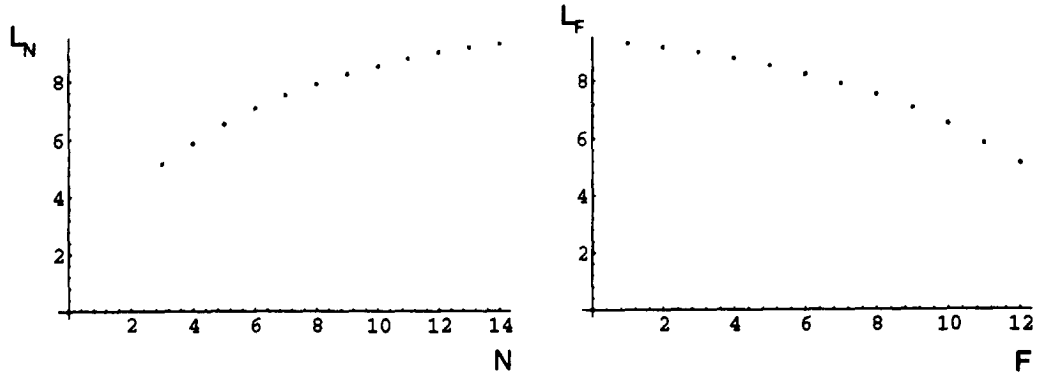

Fig. 1. The average queue lengths as a function of N and F.

length L_N as N varies between 3 and 14 for the first queue and the average queue length, L_F as F varies between 1 and 12 for the dual queue.

It is known that the variability of the interarrival and service time distributions has an important effect on the performance of a queueing system. A good measure of variability of a random variable is its coefficient of variation (standard deviation divided by the mean). As a second example, we compare the expected queue lengths for a number of GI/M/1/K (N/0/0) queues that have arrival processes with different coefficients of variation. By duality, same comparison can easily be performed on the dual system to analyze the effects of changing the coefficient of variation of the service times.

Let $\lambda=0.95$, $\mu=1$ and $K=20$. We use five different interarrival time distributions: the Erlang-2 distribution (denoted by E_2) which has a coefficient of variation (CV) of 0.77, the Poisson (denoted by Po) distribution which has a CV of 1 and three distributions from the two stage hyperexponential distribution family with CV's of 1.18, 1.46 and 2.13 (denoted by H1, H2 and H3 respectively). Therefore, H1, H2 and H3 have a probability density function of the type:

$$f_H(x)=q\lambda_1 e^{-\lambda_1 x}+(1-q)\lambda_2 e^{-\lambda_2 x} \tag{42}$$

where q is a mixing probability between 0 and 1 and λ_1 and λ_2 are the arrival rates in stages 1 and 2 of the distribution respectively. As the effective arrival rate λ is 0.95 for this example, it is required that:

$$q\lambda_1^{-1}+(1-q)\lambda_2^{-1}=0.95 \tag{43}$$

Further, the distributions H1, H2 and H3 have the property that they have balanced means in each of the two stages, i.e.:

$$q\lambda_1^{-1}=(1-q)\lambda_2^{-1} \tag{44}$$

The above restrictions ensure that H1, H2 and H3 are similar except for the differences in their respective variances.

Table 1 reports the expected queue lengths for each arrival distribution for the threshold values of $N=5$, 10 and 15. Also Table 1 confirms the deteriorating effect of variability on the performance of the queueing system. As the coefficient of variation increases, the average queue lengths also increase for the identical threshold value, N. In contrast Table 1 also implies an interesting property for the dual M/G/1/K (F/0/0) queue. Using duality properties, we know that queue lengths in this queue change in the reverse direction to the change in the original system (GI/M/1/K (N/0/0). Consequently, the average queue lengths in the M/G/1/K (F/0/0) queue decrease with increasing variability in the service process.

Remark: Until recently, the computations similar to those presented in the above examples were considered a major challenge. The advances in symbolic computational tools have facilitated this computation. All numerical values reported here are exact and have been computed through symbolic integration using Mathematica Version 2.2 for SPARC stations.

CONCLUSION

In this article, the queue length duality relationship for controlled arrival and service processes has been extended to non-Markovian queues. Computation of the performance measures for controlled queues is important for optimization purposes. If N-policy or F-policy is implemented for controlling the service or the arrival process, the optimal values of N or F must be obtained through a search procedure. This search procedure requires computation of the average queue length and the average length of the busy period for the server or the on period for the arrival process. Through duality, these performance measures can be obtained in pairs.

We made use of physically dual systems to establish stochastic duality relationships. Most queue length duality relations established so far were algebraic results based on the analysis of the probability transition matrices. In comparison, our approach yields easier and more general proofs for the

Table 1. Expected queue lengths as a function of N for different arrival distributions

N	E_2	Po	H1	H2	H3
5	9.08605	9.43632	9.64230	9.93319	10.4693
10	10.6237	10.8659	11.0223	11.2417	11.6247
15	11.8522	12.0146	12.1330	12.3004	12.5817

relationships we sought avoiding the algebra and the Markovian structure used in the previous results.

As an example for the utility of the results obtained here, we provided an analysis of the GI/M/1/K queue operating under N-policy. By obtaining the queue length distribution for this server control problem, the queue length distribution for the dual arrival control problem was automatically obtained.

REFERENCES

1. Yadin, M. and Naor, P., Queueing systems with a removable service station. *Opl Res. Quart.* , 1963, **14**, 393–405.
2. Heyman, D., Optimal operating systems in M/G/1 queueing systems. *Oper. Res.* , 1968, **16**, 362–382.
3. Gupta, S. M., Interrelationships between controlling arrival and service in queueing systems. *Comp. Oper. Res.* , 1995, **22**, 1005–1014.
4. Gupta, S. M., N-policy queueing system with finite population. *Trans. Opl Res.* , 1995, **7**, 45–62.
5. Gupta, S. M., Duality in truncated steady state erlang distibution based queueing processes. *J. Opl Res. Soc.* , 1993, **44**, 253–257.
6. Gupta, S. M., Finite source erlang based queuing systems: complementarity, equivalence and their implications. *Comp. Math. Appl.* , 1994, **28**, 57–74.
7. Gupta, S. M., Interrelationship between queueing models with balking and reneging and machine repair problems with warm spares. *Microelectronics and Reliability*, 1994, **34**, 201–209.
8. Gupta, S. M., Queueing model with state dependent balking and reneging: its complementary and equivalence. *Performance Evaluation Review*, 1995, **22**, 63–72.
9. Gupta, S.M. and Melachrinoudis, E., Complementarity and equivalence in finite source queueing models with spares. *Comp. Oper. Res.* , 1994, **21**, 289–296.
10. Harris, T., Duality of finite Markovian queues. *Oper. Res.* , 1967, **15**, 575–576.
11. Hlynka, M. and Wang, T., Comments on duality of queues with finite buffer size. *Oper. Res. Lett.* , 1993, **14**, 29–33.
12. Yang, P., A unified algorithm for computing the stationary queue length distributions in M(k)/G/1/N and GI/M(k)/1/N Queues. *Queueing Systems*, 1994, **17**, 383–401.
13. Sparragis, P., Cassandras, C. and Towsley, D., On the duality between routing and scheduling systems with finite buffer space. *IEEE Transactions on Automatic Control*, 1993, **38**, 1440–1446.
14. Xu, S. H. and Shantikumar, J. G., Optimal expulsion control—a dual approach to admission control of an ordered entry system. *Oper. Res.* , 1993, **41**, 1137–1152.
15. Xu, S. H., A duality approach to admission and scheduling control of queues. *Queueing Systems*, 1994, **18**, 273–300.
16. Lavenberg, S., Steady-state queueing time of M/G/1 finite capacity queue. *Mgmt Sci.* , 1975, **21**, 501–506.
17. Keilson, J., The Ergodic queue length distribution for queueing systems with finite capacity. *J. Roy. Stat. Soc. Series B*, 1966, **28**, 190–201.
18. Iglehart, D. and Whitt, W., Multiple channel queues in heavy traffic I. *Advances Appl. Probability*, 1970, **2**, 150–177.
19. Gordon, W. and Newell, G., Cyclic queueing networks with restricted queue lengths. *Oper. Res.* , 1967, **15**, 266–278.
20. Teghem, J.Jr, Optimal control of a removable server in an M/G/1 queue with finite capacity. *Eur. J. Opl Res.* , 1987, **31**, 358–367.
21. Takagi, H., M/G/1/K queues with N-policy and setup times. *Queueing Systems*, 1993, **14**, 79–98.