6

Effect of Switchover Time in Cyclically Switched Systems

Khurram Aziz
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology
Pakistan

1. Introduction

Cyclic service queueing systems have a broad range of applications in communication systems. From legacy systems like the slotted ring networks and switching systems, to more recent ones like optical burst assembly, Ethernet over SDH/SONET mapping and traffic aggregation at the edge nodes, all may employ cyclic service as a means of providing fairness to incoming traffic. This would require the server to switch to the next traffic stream after serving one. This service can be exhaustive, in which all the packets in the queue are served before the server switches to the next queue, or non-exhaustive, in which the server serves just one packet (or in case of batch service, a group of packets) before switching to the next queue.
Most of the study on systems with cyclic service has been performed on queues of unlimited size. Real systems always have finite buffers. In order to analyze real systems, we need to model queues with finite capacity. The analysis of such systems is among the most complicated as it is very difficult to obtain closed-form solutions to systems with finite capacity.
An important parameter in cyclic service queueing systems with finite capacity is the switchover time, which is the time taken for the server to switch to a different queue after a service completion. This is especially true for non-exhaustive cyclic service systems, in which the server has to switch to the next queue after serving each packet. The switchover time is usually very small as compared to the service time, and is generally ignored during analysis. In such cases, the edge node can be modeled as a server, serving the various access nodes - that can be modeled as queues - in a cyclic manner. Hence, we assume that on finding an empty queue, the server will go to the next queue with a switchover rate of, say ε, but if the queue is not empty, we ignore the switchover time and assume that the server will switch to the next queue with rate μ after serving one packet in the queue.
While this generally led to quite accurate results in the past due to a large difference in ratios between the service and switchover times, this might not be the case today as optical communication systems are getting faster and faster. Thus the switchover time cannot always be safely ignored as smaller differences between switchover times and service times may introduce significant differences in the results. In order to analyze such systems, the switchover process can be modeled as another phase in the service process.
The focus of this chapter is on the analysis of non-exhaustive cyclic service systems with finite capacity using state space modeling technique. A brief summary on the work done to date, in cyclic service systems is presented in Section 2, while some applications of such systems

are discussed in Section 3. Analytical models of systems in which the switchover times can be ignored during service are presented in Section 4, in which we start from a simple two-queue system and generalize for an n-queue system. Analytical study of edge nodes that employ non-exhaustive cyclic service to serve various incoming streams as a two stage process (serving and switchover) in which switchover times are not ignored during service is presented in Section 5, followed by a detailed comparison of systems with and without switchover times in Section 6. Scenarios in which switchover cannot be ignored are also discussed in this section. Finally, the results are summarized in Section 7.

2. Related work

The study on cyclic service queueing systems is quite extensive. It would thus be helpful if these systems can be categorized. Several types of classifications have been presented in the literature, with the most recent being the survey by Vishnevskii and Semenova (Vishnevskii & Semenova, 2006). The classification presented here, however, is based on the most widely used parameters and related work in those categories is then presented.

2.1 Categorization of cyclic service queueing models

In a cyclic service queueing system, two or more queues are associated with the same server, which scans different queues in a round-robin manner and serves the queue if a packet is present. This service can be of three types – exhaustive, non-exhaustive and gated. In an exhaustive service model, the server switches over to the next queue only after completing service for all the customers in the queue. This also includes any new customers that may arrive during this time. On the other hand, in a gated system, service is provided to only those customers that were present in the queue when the server arrives to that queue. A limited, or non-exhaustive service is one in which a fixed number of customers – typically one – are serviced by the server during one visit. Usually, the exhaustive service is considered more efficient in terms of the waiting time of the customers than the gated service, which in turn is considered more efficient than the non-exhaustive service. Hence, in case of the exhaustive and gated service policies, queues with a large number of customers get more attention than those with a small number of customers, resulting in a less fair service as compared to the non-exhaustive service policy. So depending on the definition of "fairness", the non-exhaustive service policy is the fairest. This is especially true for communication systems, as different queues usually represent different traffic streams and it is undesirable to prefer one stream to another if their priorities are equal.

Another important consideration in such systems is the switchover time which is the time taken by the server to move to the next queue, after finishing service in the current queue. The switchover time is usually quite small as compared to the service time and is ignored in most studies. However, this can cause large differences in results especially if the switchover rate is not large as compared to the service rate.

In addition to the finite switchover rate, another issue is the size of the queues. Most of the studies on systems with cyclic service have been performed on queues of unlimited size. Real systems always have finite buffers. In order to analyze real systems, queues with finite capacity need to be modelled. An important feature of such systems is blocking, which happens when the queue becomes full and any subsequent arrivals are lost.

The cyclic service queueing models can thus be mainly categorized in the following different ways:

- Service discipline – exhaustive, gated and non-exhaustive.

- Switchover times – zero and non-zero.
- Buffer capacity – infinite, finite and single buffer.

2.2 Cyclic service systems with infinite buffers

Cyclic service queueing systems have been extensively studied in the literature. The first study on the cyclic polling systems available is the patrolling machine repairman model (Mack et al., 1957) where a single repairman visits a sequence of machines in cyclic order, inspecting them and repairing them when failure has occurred. The first study on cyclic polling models relating to communication networks was in the early 1970s to model the time-sharing computer systems. Since then, there has been an extensive research in this area, especially since the range of applications in which cyclic polling models can be used is very broad.

Leibowitz (Leibowitz, 1961) was among the first to study an approximate solution for symmetrically loaded cyclic polling system with gated service and constant switchover time. Cooper and Murray (Coooper & Murray, 1969; Cooper, 1970) analyzed exhaustive and gated service systems using an imbedded Markov chain technique for zero switchover time. Eisenberg (Eisenberg, 1971) studied a two-queue system with general switchover time, while Eisenberg (Eisenberg, 1972) and Hashida (Hashida, 1972) generalized the results of Cooper and Murray for non-zero switchover times. Bux and Truong (Bux & Truong, 1983) provided a simple approximation analysis for an arbitrary number of queues, constant switchover time and exhaustive service discipline. Lee (Lee, 1996) studied a two-queue model where the server serves customers in one queue according to an exhaustive discipline and the other queue according to a limited discipline, while Boxma (Boxma, 2002) studied a combination of exhaustive and limited disciplines in the two queues along with a patient server, which waits for a certain time in case there are no customers present in one of the queues.

For non-exhaustive cyclic service and general switchover times, Kuehn (Kuehn, 1979) developed an approximation technique based on the concept of conditional cycle times and derived a stability criteria for the general case of $GI/G/1$ systems with a cyclic priority service. Boxma (Boxma, 1989) related the amount of work in a polling system with switchover times to the amount of work in the same polling system without switchover times, leading to several studies on this relationship, notably by Cooper et al. (Cooper et al., 1996), Fuhrmann (Fuhrmann, 1992), Srinivasan et al. (Srinivasan et al., 1995), and Borst and Boxma (Borst & Boxma, 1997). An important question is that how large should the switchover rate be as compared to the service rate, so that it can be safely ignored. The answer is not simple and this study will attempt to answer this question in relation to the cyclic service queueing models with finite buffers and non-exhaustive service in later sections.

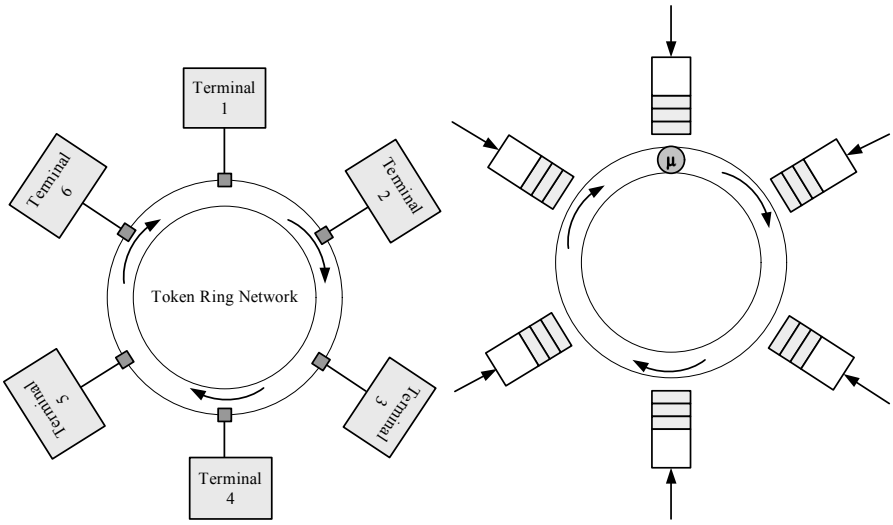
2.3 Cyclic service systems with finite buffers

While the study of cyclic service systems with infinite buffer capacity has been very extensive and closed form solutions for several such systems with exhaustive service discipline exist, the study of cyclic service systems with finite capacity queues and non-exhaustive service discipline is rather limited in the literature. Single buffer systems have been studied by Chung and Jung (Chung & Jung, 1994), and Takine et al., (Takine et al., 1986; 1987; 1990). Magalhaes et al., (Magalhaes et al., 1998) present a distribution function for the interval between the instant when the customers leave each queue, in a two-queue $M/M/1/1$ polling system. Titenko (Titenko, 1984) established formulae for the calculation of the moments of any order of the waiting times for single-buffer queues. Takagi (Takagi, 1992) presented the Laplace-Stieltjes transform (LST) of the cycle time for an exhaustive service, $M/G/1/n$ polling

system. A virtual buffer scheme for customers entering the system when the queue is full is suggested by Jung (Jung & Un, 1994). Tran-Gia and Raith have several important studies in this area. In (Tran-Gia & Raith, 1985a;b), a non-exhaustive cyclic queueing systems with finite buffers is analyzed based on the imbedded Markov chain approach in conjunction with a two-moment approximation for the cycle time. In (Tran-Gia, 1992), the stationary probability distributions of the number of waiting customers at polling instants as well as arbitrary instants for a $GI/G/1/n$ polling system with a 1-limited service discipline is obtained using discrete time analysis. Onvural and Perros (Onvural & Perros, 1989) present an approximation method for obtaining the throughput of cyclic queueing networks with blocking as a function of the number of customers. A polling system with Munit capacity queues and one infinite capacity queue with exhaustive service is described in (Takine et al., 1990).
The work on polling systems has been well summarized by Takagi in various papers. In (Takagi, 1986), all the results available till 1986 were organized, while an up-to-date summary on polling systems was presented in (Takagi, 1988). This survey was updated twice in 1990 (Takagi, 1990) for all work until 1989 and 1997 (Takagi, 1997) for the advances made after his previous update between the years 1990 to 1994. A more recent survey by Vishnevskii and Semenova (Vishnevskii & Semenova, 2006) covers various polling models, including the cyclic service models with finite service in great detail. It is clear, however, that accurate and generalized results for cyclic service finite queueing models are still not available. Some authors have provided a few closed form solutions for some specific models, but most of the time, these are approximate solutions, mostly for single buffer systems.

3. Applications of cyclic service queueing systems

Polling models with cyclic service can be used in a wide range of applications, from computer communications to robotics, production, manufacturing, and transportation. In computer communications, the queueing model with cyclic service was first used in the analysis of time-sharing computer systems in the early 1970's. In the 1980's, the token passing systems such as the token ring and token bus, as well as other demand-based channel access schemes in local area networks, such as the one shown in Figure 1, were analysed using such queueing systems with cyclic service.
From legacy systems like the slotted ring networks and switching systems, to more recent systems like wireless networks, optical burst assembly, Ethernet over SDH/SONET mapping and traffic aggregation at the edge nodes, and all may employ cyclic service as a means of providing fairness to incoming traffic. Queueing systems with cyclic service are extensively used especially at the edge nodes to provide fairness to the different flows that arrive at the node. One such example is the fair queueing system proposed by Nagel (Nagle, 1987). Another example is the mapping of Ethernet over SDH/SONET, as shown in Figure 2. Cyclic service can also be employed by the burst assembler in an optical burst switching node as shown in Figure 3.
Ibe and Trivedi (Ibe & Trivedi, 1990) propose the use of stochastic Petri Net models for obtaining the performance measures of a finite buffer polling system using the exhaustive, gated and limited service disciplines. Choi (Choi, 2004) proposes a cyclic polling based algorithm for differentiated class of services in Ethernet passive optical networks. Takagi highlights three classical but instructive applications of polling models to the performance evaluation of communication networks in (Takagi, 2000). The three applications discussed are the half-duplex transmission for an inquiry system, the polling data link control and the token ring network. Bruneel and Kim (Bruneel & Kim, 1993), Grillo (Grillo, 1990), and Levy and Sidi (Levy & Sidi,



(a) Token ring network (b) Equivalent queueing model

Fig. 1. Token ring network and its equivalent queueing model

1990) analyze several applications examples of communication networks, including the ATM systems that employ the cyclic polling systems. In the late nineties, rapid development of the broadband wireless transmission networks prompted several studies of the polling system models in this area, especially by Ziouva and Antonakopoulos (Ziouva & Antonakopoulos, 2002/2007; 2003), and Vishnevskii (Vishnevsky et al., 1999; 2004). Miorandi et al., (Miorandi et al., 2004) performed an interesting study on the performance evaluation of the Bluetooth polling schemes.

The focus in this chapter is to study the basic polling models that employ non-exhaustive cyclic service and finite queues, independent of the communication system involved, and study the effect of switchover time on these systems.

4. Systems with zero switchover times

In this section, cyclic service queueing systems that ignore the switchover times during service are studied. Typically, a server spends some time serving a customer and then switches over to the next queue. The time taken for the server from the completion of service in one queue to the commencement of service in the next queue is known as the switchover time. This switchover takes a small amount of time as compared to the service time and is usually ignored. The assumption here is that the switchover times in such systems will be very small as compared to the service times and when ignored, they will not have a considerable effect on the overall system performance.

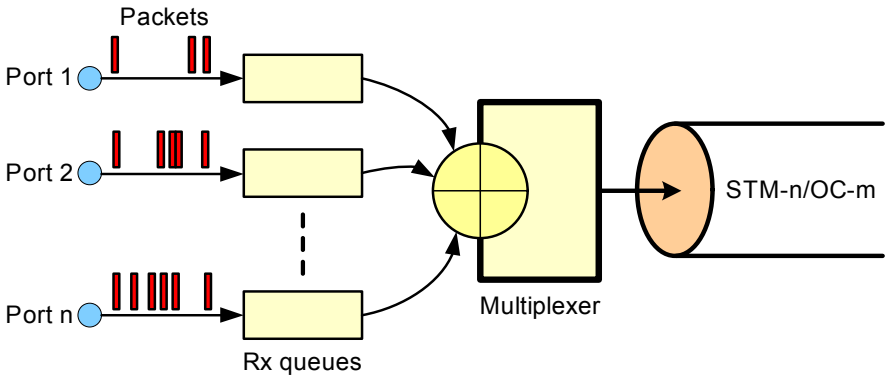


Fig. 2. Mapping Ethernet over SDH/SONET in an edge node

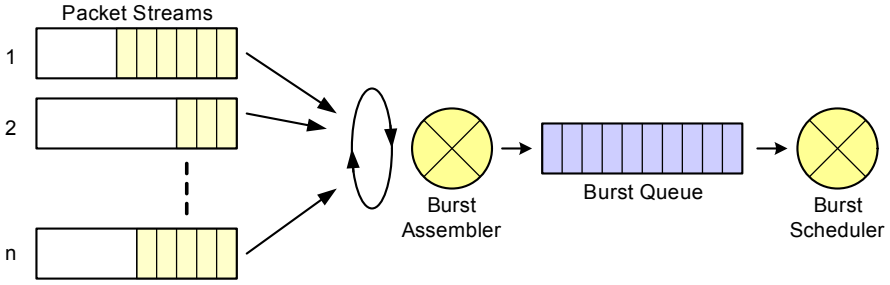


Fig. 3. Burst assembly in an OBS edge node using cyclic service

4.1 Model description

Cyclic service systems can be modelled as shown in Figure 4, which shows N queues, each of size s_i $(i = 1,\ldots,N)$, being served in a round-robin manner by a server with an exponentially distributed service rate of mean μ. The arrival rate to each queue is also exponentially distributed with mean λ_i $(i = 1,\ldots,N)$. The average time taken by the server to switch over from one queue to the next is given by $1/\varepsilon$ where ε is the mean switchover rate.

At each scanning epoch, the server processes one packet in the queue if there is at least one packet waiting. In case there is no waiting packet in the queue, the server switches over to the next queue with a switchover rate of ε.

The following parameters are used:

N = number of queues in the system
λ_i = arrival rate of packets offered to queue i; $i = 1,\ldots,N$
S_i = capacity of queue i; $i = 1,\ldots,N$
μ = mean service rate of the server
ε = mean switchover rate of the server

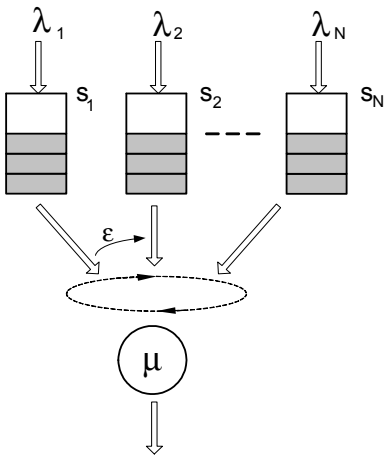


Fig. 4. System model for a cyclic service queueing system

4.2 Basic two-queue system

The analysis of cyclic service queueing systems is presented with a model that has only two queues as shown in Figure 5. Such a system can be considered as an $M/M/1$-s system.
The two-queue cyclic service system consists of one server and two queues with a capacity of s_1 and s_2 respectively, as shown in Figure 5. The mean arrival rates to the two queues are given by λ_1 and λ_2 respectively, while server completes each service with a mean rate of μ.

4.2.1 Analysis

For an exact analysis, the system states can be described by a vector $\{Q_1(t), Q_2(t), \ldots, Q_n(t), I(t), X(t)\}$, where $Q_i(t)$ is the number of packets in the ith queue, $I(t)$ is the current location of the server within the cycle and $X(t)$ is the age of the current service (Kuehn, 1979). In this study, the single-stage service process is taken to be a Markov process having a mean rate of μ. $X(t)$ can then be ignored due to the PASTA (Poisson Arrivals See Time Averages) property of the service process, which leaves us the vector $\{Q_1(t), Q_2(t), \ldots, Q_n(t), I(t)\}$ that accurately describes the system states. Hence for this two-queue system, three variables for each system state are required – one each for the number of occupied queue places – while another to show which queue's customer is currently undergoing service. Each state is then defined by the vector $\{Q_1(t), Q_2(t), I(t)\}$, where $Q_1(t)$ is the number of customers in the system coming through the first queue, $Q_2(t)$ is the number of customers in the system coming through the second queue and $I(t)$ is the current location of the server within the cycle. Clearly, $I(t)$ can have only two values where a value of 1 means that the server is serving a customer from queue 1 while 2 means that the server is serving a customer from queue 2. $Q_1(t)$ and $Q_2(t)$ can vary from zero to s_1 and s_2, respectively. The state diagram will hence be three-dimensional as shown in Figure 6, where transitions along the x-axis show arrivals of customers from queue 1 while transitions along the y-axis show arrivals of customers from queue 2. The z-axis shows the current location of the server within the cycle, with the front xy-plane showing the service of packets from queue 1 and the back xy-plane showing the service of packets from queue 2. State diagram of

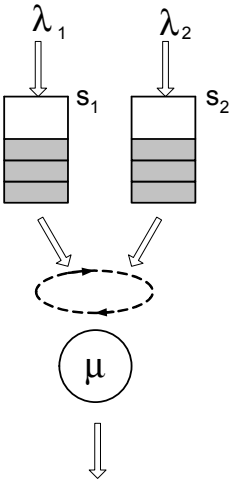


Fig. 5. System diagram for a two-queue cyclic service system

such systems usually consists of two parts – a boundary portion and a repeating portion. The boundary portion usually shows the states and transitions when the queues of the system are either empty or full, while the repeating portion usually shows the states and transitions when there is something in the queues but the queues are still not full. For very large state diagrams, such a depiction is very useful in studying the behavior of the system. Figure 7 shows a simplified view of the repeating portion of the state diagram in which transitions to and from just one state are shown. The server will switch from one queue to the other with a mean rate of ε.
Using the state diagram, the state probabilities p_i of all the states can be calculated by solving the system of linear equations. Using these state probabilities, the mean number in system and mean number in queue can then be found using the following equations.
Mean number of customers in system:

$$E[N] = \sum_{x=0}^{s+1} x p_x \tag{1}$$

Mean number of customers in queue:

$$E[Q] = \sum_{x=1}^{s+1} (x-1) p_x \tag{2}$$

From these equations, using the Little's theorem (Little, 1961), we get
Mean time in system:

$$T_S = \frac{E[N]}{\lambda} \tag{3}$$

Mean waiting time:

$$T_W = \frac{E[Q]}{\lambda} \tag{4}$$

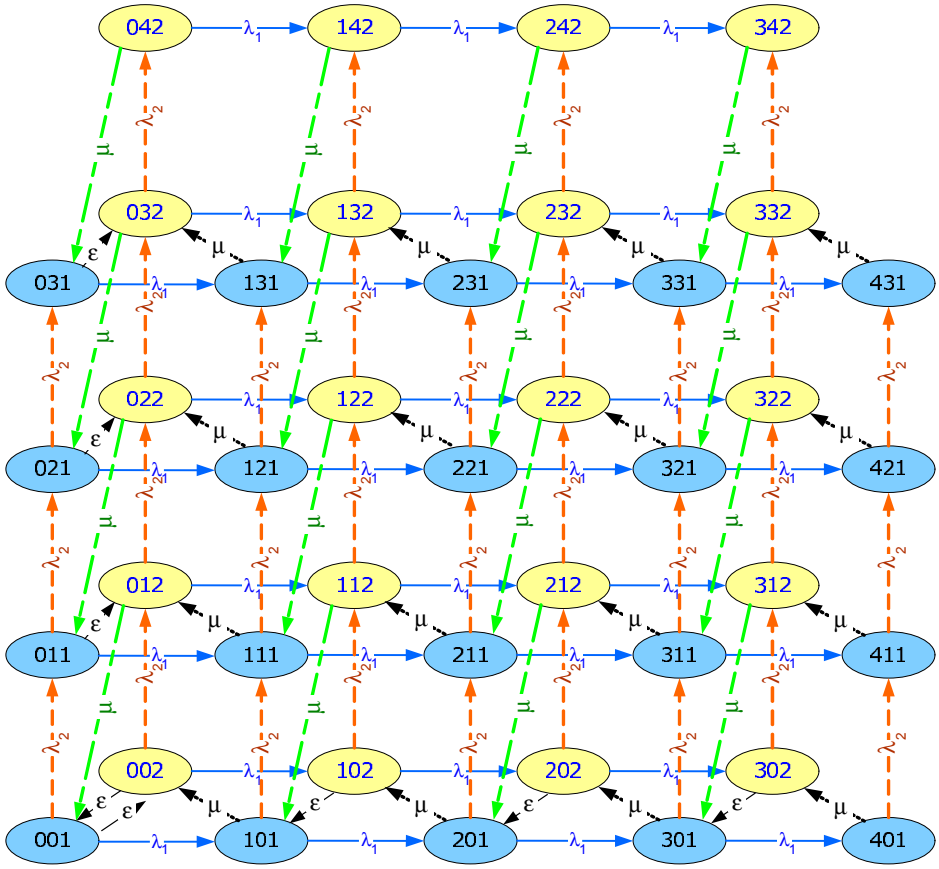


Fig. 6. State diagram for a two-queue cyclic service system

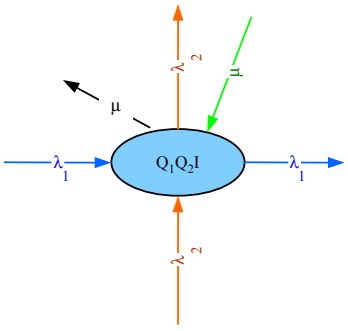


Fig. 7. Simplified view of the transitions to and from a state for a two-queue cyclic service system

An important point to note here is that the number in system are being considered, i.e., number in queue plus any customer that may be in service, and not just the number in queue. Hence in the state diagram of Figure 6 as well as the equations, Q_1 goes from 0 to $s_1 + 1$ and not s_1, while Q_2 goes from 0 to $s_2 + 1$ and not s_2.
Using (1) to (4), the various characteristic measures can be calculated for each queue as given in (5) to (8).

$$E[N_1] = \sum_{i_2=0}^{s_2} \sum_{i_1=0}^{s_1+1} i_1 P(i_1, i_2, 1) + \sum_{i_2=0}^{s_2+1} \sum_{i_1=0}^{s_1} i_1 P(i_1, i_2, 2) \tag{5}$$

$$E[Q_1] = \sum_{i_2=0}^{s_2} \sum_{i_1=2}^{s_1+1} (i_1 - 1) P(i_1, i_2, 1) + \sum_{i_2=0}^{s_2+1} \sum_{i_1=1}^{s_1} i_1 P(i_1, i_2, 2) \tag{6}$$

$$T_{S_1} = \frac{E[N_1]}{\lambda_1} \tag{7}$$

$$T_{W_1} = \frac{E[Q_1]}{\lambda_1} \tag{8}$$

When a customer arrives in a system and finds the server busy, it has to wait. If all the probabilities for the states in which the customer has to wait are summed up, the probability of waiting is obtained. Similarly, when a customer arrives to a system and finds the queue full, it will be blocked. If all the probabilities of such states are summed, the probability of blocking is obtained. The probabilities of waiting and blocking for this system are as follows:

$$W_1 = \sum_{i_2=0}^{s_2} \sum_{i_1=1}^{s_1} P(i_1, i_2, 1) + \sum_{i_2=0}^{s_2+1} \sum_{i_1=0}^{s_1-1} P(i_1, i_2, 2) \tag{9}$$

$$B_1 = \sum_{i_2=0}^{s_2} P(s_1 + 1, i_2, 1) + \sum_{i_2=0}^{s_2+1} P(s_1, i_2, 2) \tag{10}$$

4.2.2 Results

The various characteristic measures for customers in queue 1 will be affected not only by the queue length and arrival rate in queue 1, but also the arrival rate and maximum queue size of queue 2. Similarly, the switchover rate, although ignored during service, may still have an effect on the characteristic measures, especially at lower arrival rates and needs to be studied further.
In order to study these effects, various characteristic measures for customers in queue 1 given by (5) to (10) are plotted against arrival rate in queue 1 for different queue sizes and different arrival rates in queue 2. Symmetric as well as asymmetric traffic loads and queue sizes for both queues are studied.
Figure 8 shows the mean number of customers in queue 1 against varying arrival rate in queue 1, for various queue capacities. The graph shows that the mean number of customers in queue 1 increases slowly for low arrival rates up to 0.4, but increases rapidly from 0.4 to 0.7. It then stabilizes and levels out after the saturation point (arrival rate of 1.0). The graph also shows that increasing the capacity in queue 2 from 3 to 10 has a very small effect on the mean number of customers in queue 1. On the other hand, Figure 9 shows the mean number of customers

in queue 1 against varying arrival rate in queue 1, for various arrival rates in queue 2. It can be clearly seen that the arrival rate of queue 2 has a significant effect on the queue length distribution in queue 1. At low arrival rates in queue 2, the rate of increase in the queue length of queue 1 is much slower than the rate of increase observed for a high arrival rate in queue 2, as on average, the server spends more time serving customers of queue 2, especially at lower arrival rates of queue 1.

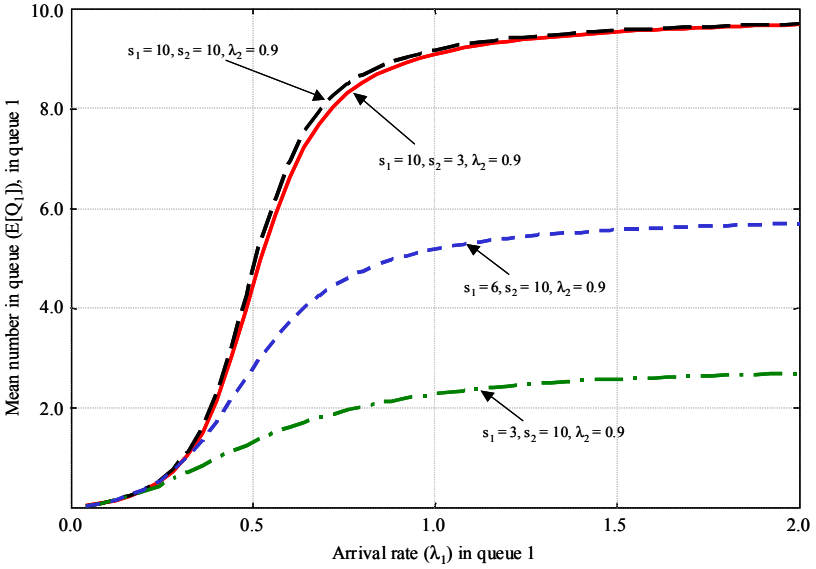


Fig. 8. Effect of varying queue sizes of queues 1 and 2 on number of customers in queue 1 for a two-queue system

Figures 10 and 11 show the mean waiting time for customers of queue 1 against the arrival rate of customers in queue 1, for varying queue capacities of both queues and varying arrival rate of customers in queue 2. Here again, a similar behavior is seen, whereby the queue capacity of queue 2 has a very small effect on the waiting time of customers in queue 1, as shown in Figure 10, but the increase of the arrival rate in queue 2 significantly increases the mean waiting time of customers in queue 1.
Finally, in Figures 12 and 13, the effect of queue 2 on the probability of blocking and the probability of waiting for customers in queue 1 is observed. Only the effect of increasing the arrival rate in queue 2 are shown as it has been observed that queue capacity of queue 2 has little effect on measures of queue 1. Here again, it is observed that a lower arrival rate in queue 2 results in a gradual increase in the blocking and waiting for customers of queue 1 as compared to a higher arrival rate, in which case this increase is quite abrupt.

4.3 Generalization to n-queue systems

An n-queue cyclic service system requires $n+1$ state variables to describe a state and hence, an $n+1$ dimensional state diagram. An important feature that is observed in these systems is the symmetry of the model. Extending the two-queue model to a more general n-queue model

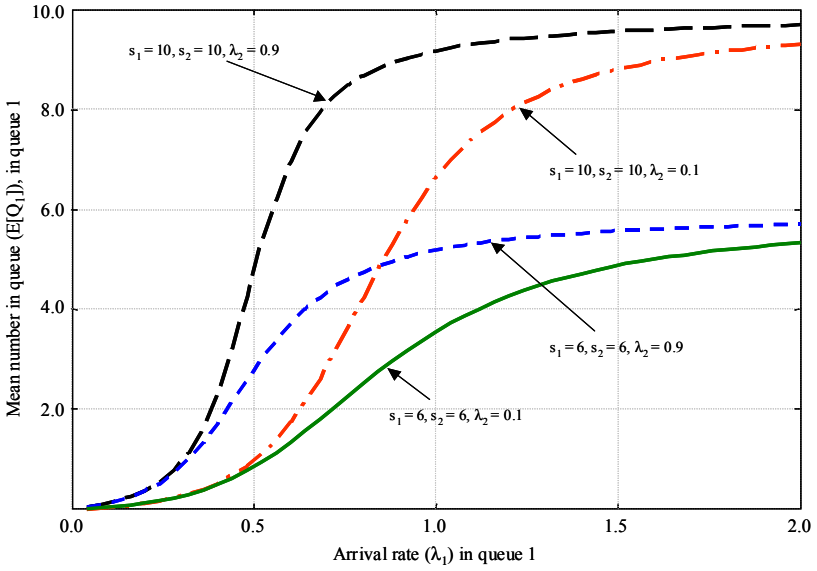


Fig. 9. Effect of varying arrival rate to queue 2, on number of customers in queue 1 for a two-queue system

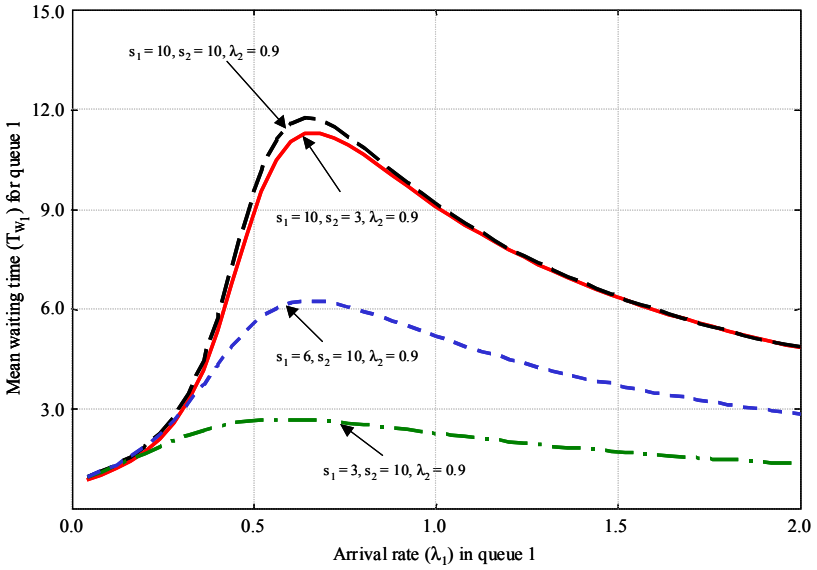


Fig. 10. Effect of varying queue sizes of queues 1 and 2, on waiting time of customers in queue 1 for a two-queue system

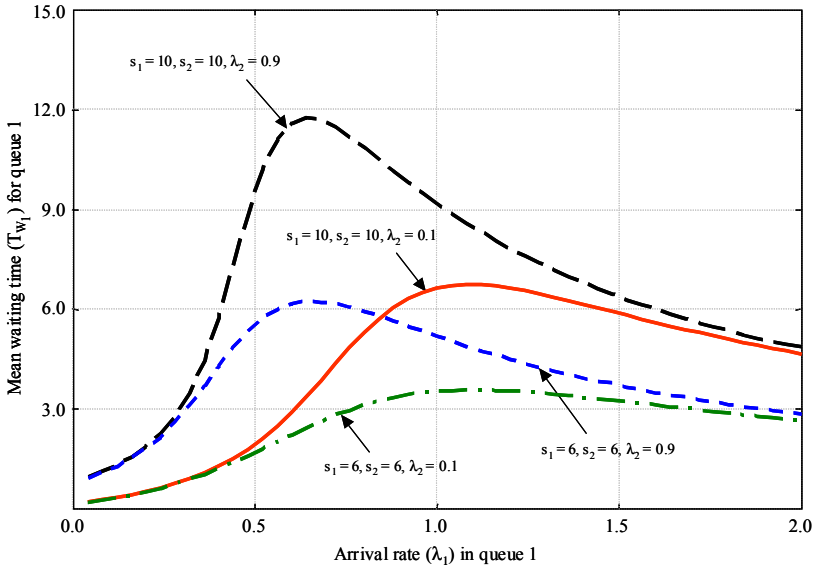


Fig. 11. Effect of varying arrival rate to queue 2, on waiting time of customers in queue 1 for a two-queue system

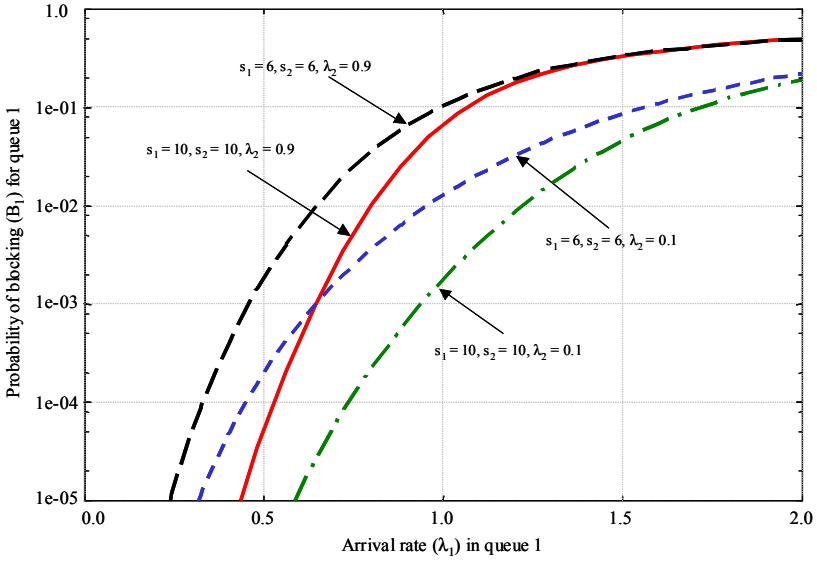


Fig. 12. Effect of varying arrival rate and maximum queue size of queue 2 on probability of blocking for customers in queue 1, for a two-queue system

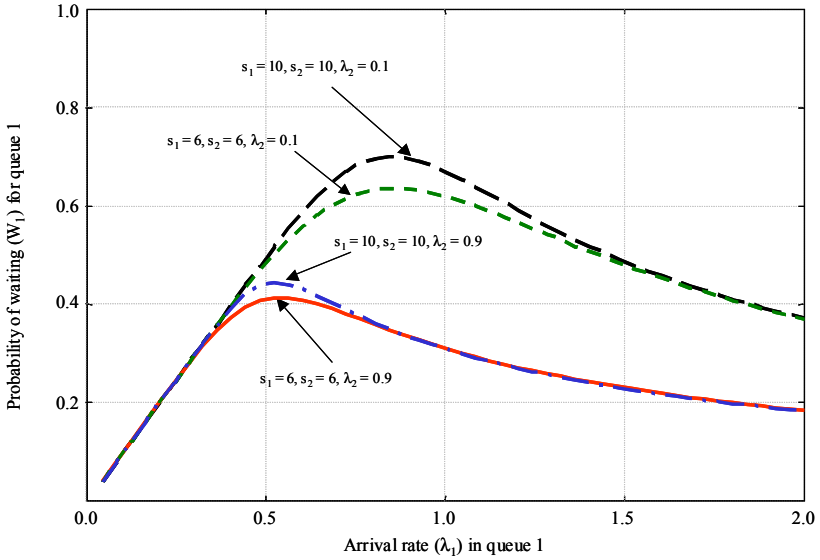


Fig. 13. Effect of varying arrival rate and maximum queue size of queue 2, on probability of waiting for customers in queue 1, for a two-queue system

is quite straight-forward. The complex part is the difficulty in drawing a state diagram with more than three dimensions. Due to the symmetry of the model, however, it is quite sufficient to draw a subset of the diagram for the boundary portion and the repeating portion of the system. The derivation of the system equations is also straightforward and (11) to (16) give the various measures for an n-queue system with switchover time ignored during service. The mean number in system and mean number in queue are given by:

$$\begin{aligned} E[N_1] = & \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=0}^{s_1+1} i_1 P(i_1, i_2, \cdots, i_n, 1) \\ & + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} \sum_{i_1=0}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, 2) \\ & + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=0}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, n) \end{aligned} \tag{11}$$

$$
\begin{aligned}
E[Q_1] = & \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=2}^{s_1+1} (i_1 - 1) P(i_1, i_2, \cdots, i_n, 1) \\
& + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} \sum_{i_1=1}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, 2) \\
& + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=1}^{s_1} i_1 P(i_1, i_2, \cdots, i_n, n)
\end{aligned} \tag{12}
$$

Using Little's theorem, the mean time in system and the mean waiting time can be obtained as follows:

$$T_{S_1} = \frac{E[N_1]}{\lambda_1} \tag{13}$$

$$T_{W_1} = \frac{E[Q_1]}{\lambda_1} \tag{14}$$

The probability of waiting and probability of blocking can be calculated from the following equations.

$$
\begin{aligned}
W_1 = & \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=1}^{s_1} P(i_1, i_2, \cdots, i_n, 1) \\
& + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} \sum_{i_1=0}^{s_1-1} P(i_1, i_2, \cdots, i_n, 2) \\
& + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} \sum_{i_1=0}^{s_1-1} P(i_1, i_2, \cdots, i_n, n)
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
B_1 = & \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2} P(s_1 + 1, i_2, \cdots, i_n, 1) \\
& + \sum_{i_n=0}^{s_n} \cdots \sum_{i_2=0}^{s_2+1} P(s_1, i_2, \cdots, i_n, 2) \\
& + \cdots + \sum_{i_n=0}^{s_n+1} \cdots \sum_{i_2=0}^{s_2} P(s_1, i_2, \cdots, i_n, n)
\end{aligned} \tag{16}
$$

5. Systems with non-zero switchover times

Cyclic service queueing systems have a broad range of applications in communication systems mainly as a means of providing fairness to incoming traffic. In such systems, the server is required to switch to the next traffic stream after serving one. Usually in optical networks, this switching is done very fast as compared to the service and hence, the switchover time – which is the time taken by the server to switch from one stream to the next – is usually ignored. This however, can cause large differences in the results, especially if the switchover time is not very small as compared to the service time.
This section presents an analytical study of an edge node that employs non-exhaustive cyclic service to serve various incoming streams as a two-stage process (serving and switchover), and finite size queues to model real systems as closely as possible. The effect of switchover time on the performance of such systems is then studied. Comparison of systems with various ratios of switchover times to the service times is also done and the scenarios under which switchover times cannot be ignored are discussed.

5.1 Cyclic service system with two-stage service

An n-queue cyclic service system with two-stage non-exhaustive service can be modelled as shown in Figure 14. The n queues, each of size $s_i (i = 1, \ldots, n)$, are served in a round-robin manner by a server with a negative exponentially distributed service rate of mean μ. The arrival rate to each queue is Poisson with mean $\lambda_i (i = 1, \ldots, n)$. The average time taken by the server to switch over from one queue to the next is given by $1/\varepsilon$ where ε is the mean switchover rate.
At each scanning epoch, the server processes one packet in the queue if there is at least one packet waiting, with a rate of μ and then switches over to the next queue with a rate of ε. In case there is no waiting packet in the queue, the server simply switches over to the next queue with a switchover rate of ε.

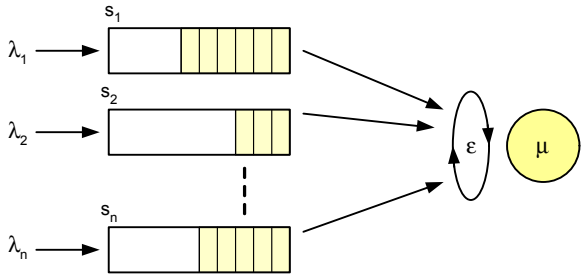


Fig. 14. System model for a queueing system with two-stage cyclic service

5.1.1 Analysis

For an exact analysis, the system states can be described by a vector $\{Q_1(t), Q_2(t), \ldots, Q_n(t), I(t), X(t)\}$, where $Q_i(t)$ is the number of packets in the ith queue, $I(t)$ is the current location of the server within the cycle and $X(t)$ is the age of the current service (Kuehn, 1979). In this study, a two-stage service process is assumed, with each of its two stages as a Markov process having mean rates of μ and ε for service and switchover,

respectively. $X(t)$ can then be ignored due to the PASTA (Poisson Arrivals See Time Averages) property of the service process, which leaves us the vector $\{Q_1(t), Q_2(t), \ldots, Q_n(t), I(t)\}$ that accurately describes the system states.
In case of the two-stage process presented here however, an additional state is needed to differentiate between two switchover cases. The first occurs after the processing of the packet in a non-empty queue, in which case the service will be two-stage; and the switchover that occurs for an empty queue, in which case the service will consist of only the switchover. The state space is hence modelled using the vector $\{Q_1(t), Q_2(t), \ldots, Q_n(t), I(t), K(t)\}$, where $K(t)$ is defined as the status of the server.
This study is restricted to two queues ($n = 2$). Each state is described by four parameters (i_1, i_2, j, k), where i_1 is the number of packets in queue 1, i_2 is the number of packets in queue 2, j is the current location of the server within the cycle ($j = 1$ means the server is scanning queue 1 while $j = 2$ means the server is scanning queue 2) and k is the status of the server. This server status can have three possible values: $k = 0$ means that the server encountered an empty queue and will simply switchover to the next queue without processing a packet, $k = 1$ means that the server is processing a packet, and $k = 2$ means that the server is switching over to the next queue after processing a packet in a non-empty queue. The resulting three-dimensional state diagram with a queue capacity of two for each queue is shown in Figure 15. It can be easily seen that a state diagram for more than two queues needs a four-dimension representation, which is not possible to draw on paper.
Figure 16 shows a simplified view of the repeating portion of the state diagram in which transitions to and from just one state are shown. Note that these transitions can be divided into four main parts:

- Arrivals to queue 1, which result in an increment of i_1. All other parameters describing the state remain unchanged.
- Arrivals to queue 2, which result in an increment of i_2. All other parameters remain unchanged.
- Service followed by switchover from queue 1. In this case, processing of a packet results in a decrement of i_1, while the server status changes to 2 from 1. Switchover then changes j to 2 from 1, indicating that server is now pointing at queue 2, while the server status k changes to 1 from 2 indicating that the server is ready to serve queue 2.
- Service followed by switchover from queue 2. In this case, processing of a packet results in a decrement of i_2, while the server status k changes to 2 from 1. Switchover then changes j to 1 from 2, indicating that the server is now pointing at queue 1, while the server status changes to 1 from 2 indicating that the server is ready for service.

Figure 17 shows the states of the system and the transitions that occur when the queues are empty. Note that in this case, no packets will be processed and the service process will consist of just the switchover. The transitions would be similar to those described above with the exception that after service and switchover, if the server sees an empty queue, the server status k will go to 0 instead of 1.
The steady state solution of the state-space model shown in Figure 15 can be obtained by solving the system of equations that is obtained from the state diagram. The steady state probabilities thus obtained can be used to solve for the various system measures. The mean number in the system, $E[N_1]$ and the mean number in queue $E[Q_1]$ for packets of type 1, i.e., packets that arrive to queue 1 are given by (17) and (18), respectively.