ABSTRACT

This paper describes several capacity assignment strategies for integrated communication systems with heterogeneous traffic, and proposes the approximate methods to evaluate analytically the performance in terms of blocking probability of demand call for lost-call-cleared mode and that of reservation request for reservation mode. In the case of reservation system, this paper newly introduces the multiqueue model and the multiserver queueing model, in addition to the modified Erlang B model and the arrival process model.

1. INTRODUCTION

According to the increasing demands of business user for integrated services such as voice, data, and image information, multi-service satellite communication systems which offer integrated digital services among small earth stations located on or near customer premises have been introduced [1]-[3].

To meet user demands, we have developed an Integrated Circuit / Packet Switching System based on Demand Assignment Control (ISSDA) offering integrated services via the TDMA satellite link [3]. The main feature of the ISSDA is that the modified No.7 signalling system with the ISDN user part is adopted as a demand assignment control, which assigns satellite communication channel to a user only when needed, to utilize high-speed satellite link effectively and to cope with advanced future communication requirements. The adoption of No.7 signalling system will facilitate the integration of multi-service satellite communication system into ISDN.

According to communication purposes and its applications, this system can offer three kinds of switching services, i.e., circuit switching service, packet switching service and broadband service. In order to provide these three switching services, it becomes very important to use the satellite link efficiently.

The capacity assignment strategies for above-mentioned multi-service satellite communication systems are an important factor on traffic design to integrate a variety of services having various capacity requests and different call set up procedures such as real-time base and reservation base. The research of this problem has just been started [1]. Performance and quality of service of this system are evaluated in terms of blocking probability of demand call for lost-call-cleared mode and that of reservation request for reservation mode.

The main purpose of this paper is to evaluate analytically the blocking probability for each traffic group which has a pool of dedicated capacity units. In particular, we discuss the reservation system like the teleconference service, which requests starting time of service, service holding time, required transmission capacity. With regard to performance evaluation, in contrast to demand base, it is necessary to consider the time duration between time instant of reservation request and starting time of service, and the procedure to treat a blocked request which has been refused by the system at the time of reservation. It is very difficult, however, to analyze the general reservation system exactly, because each call has different distribution in regard to the above mentioned time duration.

In order to simplify the analysis, we assume that reservation service is performed by FCFS (first come first served) discipline in request arrival order, all the reservation calls request the same capacity and the blocked reservation request is cleared. In accordance with the above assumption, we present the modified Erlang B model and the arrival process model. In addition, in the case that the blocked reservation requests arrive with some probability again, we present the approximate methods including the multiqueue model and the multiserver queueing model.

2. CAPACITY ASSIGNMENT STRATEGY

First, users are divided in two user classes, according to traffic characteristics, that is, lost-call-cleared traffic and reservation traffic. Moreover, each class of users is divided in several user groups according to capacity demand. And the time slots have two parts, one part is dedicated to lost-call-cleared traffic and the other part is dedicated to reservation traffic. The boundary between the two user classes may be fixed.

The capacity assignment problem can be defined as follows: The capacity, N, measured in units of time slots is to be dynamically shared by a user class divided in several groups, according to traffic characteristics, for example lost-call-cleared-mode, reservation mode, and capacity demand, for example variable band-width request. Each user group in two user classes, M, has equal average arrival rate, average call duration.
and equal number of capacity units requested for a single call.

On the other hand, several capacity assignment strategies have been proposed and studied in the papers [4] - [8].

(1) Complete sharing
   Every user class and/or user group has free access to the whole capacity on a First-Come-First-Served basis.

(2) Complete partitioning
   Every user class and/or user group has a pool of dedicated capacity units.

(3) Sharing with maximum allocation
   User classes and/or user groups share the common pool of resources, but there is a limit on the number of users of a group and/or a class that are allowed in the system.

(4) Sharing with minimum allocation
   A given number of dedicated capacity units is assigned to each user class and/or user group and the rest being shared by all on a First-Come-First-Served basis.

(5) Sharing with priorities
   Priorities are assigned to user classes and/or user groups. Users may preempt other users of lower priority.

We define a system state as the number of calls in progress for each user group. That is, a system state is given by the vector n = (n₁, n₂, ..., nₘ), where nᵢ is the number of the i-th group calls in progress. The ith group, i = 1, ..., M, is characterized by calling rate from an infinite population, the traffic statistics are assumed Poisson distribution with mean λᵢ, average call duration 1/μᵢ, number of capacity units requested for a single ith group user call Kᵢ, maximum number of ith group user calls allowed Nᵢ, number of user calls with dedicated time slots to the ith group Lᵢ. Access disciplines are characterized by assigning to every group a pair of numbers (Nᵢ, Lᵢ), that is, the maximum number of user calls and their dedicated time slots. This is an important subset of possible access disciplines. Resource size imposes a physical constraint for each user class:

\[ \sum_{i=1}^{M} K_i n_i \leq N. \]  

where nᵢ is the number of the i-th group user calls.

In the case of complete sharing, every user group has free access to the whole capacity on a First-Come-First-Served basis.

\[ L_i = 0, \quad N_i = [N/K_i]^+, \quad i=1, \ldots, M. \]  

where \([A]^+\) is the integer part of A. As for complete partitioning, every user group has a pool of dedicated capacity units.

\[ N_i = L_i, \quad \sum_{i=1}^{M} L_i K_i = N, \quad i=1, \ldots, M. \]  

For sharing with maximum allocation,

\[ L_i = 0, \quad N_i \leq P_i, \]  

where Pᵢ is the maximum number of time slots from the common pool allowed for each user call.

For sharing with minimum allocation, \(\sum_{i=1}^{M} L_i = N\) and \(\sum_{i=1}^{M} K_i n_i \leq N\) for each user class. Access disciplines are characterized by assigning to every group a pair of numbers (Nᵢ, Lᵢ), that is, the maximum number of user calls and their dedicated time slots. This is a generalization of the above two disciplines (both complete sharing and complete partitioning).

3. LOST-CALL-CLEARED MODE

In this section, we discuss only lost-call-cleared traffic when the boundary between the lost-call-cleared mode and reservation mode is fixed.

3.1 CALL HAVING THE SAME CAPACITY

Assume that lost-call traffic originates according to a Poisson distribution with mean rate \(A\), the holding time is negative exponential with mean \(1/\mu\), frame time duration is \(b\) seconds and the number of lost-call traffic slots per frame is \(N\). In the following, we show that, under the condition that \(b+1\) and \(N\mu b^N\leq 1\), the blocking probability for the lost-call traffic can be closely approximated by the well-known Erlang B equation:

\[ B = a^N/N! \sum_{i=0}^{N} a^i/i! \]  

where \(a = \lambda/\mu\) is the offered lost-call traffic load in Erlang.

3.2 CALL HAVING THE VARIOUS CAPACITY

In the case of lost-call having various transmission capacity requests, we assume that lost-call traffic originates according to a Poisson distribution with mean rate \(\lambda_i (i=1, \ldots, M)\), the holding time is negative exponential with mean \(1/\mu_i (i=1, \ldots, M)\), number of capacity units requested for a single ith user group call Kᵢ, calls whose
requirements cannot be satisfied are blocked and depart without affecting the system further. If the constraint is given by eq. (2), we call such a strategy as complete sharing. In this section, we consider in detail complete sharing strategy.

Define the overall occupancy distribution

\[ Q(n) = \Pr (\sum_{i=1}^{M} K_i n_i = n) \]  

so that, let the time congestion of ith group calls be \( E_i \),

\[ E_i = \sum_{n=0}^{N-K_i} Q(n) \]

In this section, we consider in detail complete sharing strategy.

\[ M = E_i = \sum_{n=0}^{N-K_i} Q(n) \]

(8)

Now, for Poisson arrivals for all i, we have the simple recurrence relation which provides the distribution of the number of the time slots occupied for the complete sharing strategy

\[ n Q(n) = \sum_{i=1}^{M} a_i K_i Q(n-K_i) / \mu_i \]

(9)

where \( a_i = \lambda_i / \mu_i \), \( Q(n) = 0 \) for \( n < 0 \) and

\[ \sum_{n=0}^{N} Q(n) = 1. \]

(10)

The blocking probability \( PB_i \) for each user group can be written

\[ PB_i = \sum_{i=0}^{N-K_i} Q(n-i) \]

(11)

In Fig. 2, we present a sample of the results obtained for this model. Suppose that the number of total time slots are 6, \((K_1,K_2)=(1,2)\) and \( a_1 = 0.25 \). This figure gives blocking probability as a function of offered load \((a_2)\). Table 1 shows the state distribution \((Q(n)\) \( n=0,\ldots,6 )\) of the number of time slots occupied for the complete sharing system of 6 time slots in the case of offering two kinds of traffic.

In heterogeneous traffic with various bit rates, the blocking probability \( PB \) for high bit rate calls is greater than that for low bit rate calls, when time slots are shared between them. As for the evaluation of the blocking probability for each user group, several criteria may be proposed for the optimization problem.

(1) Average blocking probability is minimized

\[ \min PB = \frac{\sum_{i=1}^{M} \lambda_i PB_i}{\sum_{i=1}^{M} \lambda_i} \]

(12)

(2) Minimization of the maximum \( PB_i \)

\[ \min \max (PB_1, \ldots, PB_M) \]

(13)

On the other hand, there is the capacity allocation problem in reverse way. In this problem, the blocking requirements for each user group is given, and the best strategy which achieves minimum capacity size is obtained.

One method has been envisaged in order to put the blocking probability of high bit rate calls close to that of low bit rate calls, the so-called trunk reservation, restricting the number of time slots for the low bit rate call unless sufficient capacity is available for the high bit rate call.

4. RESERVATION MODE

In this chapter, we discuss only reservation traffic on condition that the boundary between the lost-call-cleared mode and reservation mode is fixed. We consider the reservation system which requests starting time of communication and predetermined service time at the time of reservation such as the teleconference service and the reserved circuit switched service with X.21. In general, this system is characterized by three elements:

<table>
<thead>
<tr>
<th>( KB )</th>
<th>((K_1, K_2) = (1, 2))</th>
<th>((K_1, K_2) = (1, 3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Q(0) )</td>
<td>0.476</td>
<td>0.476</td>
</tr>
<tr>
<td>( Q(1) )</td>
<td>0.119</td>
<td>0.119</td>
</tr>
<tr>
<td>( Q(2) )</td>
<td>0.252</td>
<td>0.03</td>
</tr>
<tr>
<td>( Q(3) )</td>
<td>0.061</td>
<td>0.241</td>
</tr>
<tr>
<td>( Q(4) )</td>
<td>0.067</td>
<td>0.06</td>
</tr>
<tr>
<td>( Q(5) )</td>
<td>0.015</td>
<td>0.012</td>
</tr>
<tr>
<td>( Q(6) )</td>
<td>0.012</td>
<td>0.061</td>
</tr>
</tbody>
</table>

Table 1 State distribution of the number of time slots

2.1A-2-3
starting time of service, service holding time, required transmission capacity. With regard to performance evaluation, in contrast to demand base, it is necessary to consider the notice time (the time duration between time instant of reservation request and starting time of service) and the procedure to treat a blocked request which would be refused by the system at the time of reservation.

4.1 RESERVATION SYSTEM

The reservation system which we discuss is characterized by the following control information elements such as required capacity C, service time H and communication starting time T.

Actual reservation procedures are as follows. User makes reservation by means of speech conversation or data communication between user's reservation terminal and network reservation center, and requests a circuit by indicating communication starting starting time, the service time, the required capacity and so on. When the reservation is succeeded, that is, capacity is available at T, the system informs the user of the results of the reservation. When a user's reservation request is refused by the system, that is, capacity is not available at T, user may abandon the reservation request or inquire about available time of the system.

Blocking probability that a reservation request would be refused is a function of required capacity, service time and the notice time.

4.2 CALL HAVING THE SAME CAPACITY

We consider reservation system having relatively long interval between time instant of reservation request and starting time of actual communication. Mathematical model is as follows. The time interval \( 0 \leq t \leq \infty \) is divided into contiguous fixed intervals (e.g. one day): \( (j-1)D \leq t \leq jD \), \( 1 \leq j \leq \infty \). Moreover, each fixed interval is divided into fixed time segments (e.g. one hour): \( (i-1)\Delta \leq t \leq i\Delta \), \( 1 \leq i \leq \infty \), each of which accommodates S reservation calls. S is the number of time slots per one time segment and d is the length of the time segment. Reservation is made for a time slot in the time segment. It is very difficult, however, to analyze the reservation system exactly. In order to simplify reservation system, we assume reservation service is performed by FCFS (first come first served) discipline in request arrival order and all the reservation calls request the same capacity. And we examine a few approximate methods as described below.

We deal with the case that all blocked reservation requests are cleared.

First, we consider the M/G/S model as shown in Fig.3. Supposing that the reservation request arrives at random by a Poisson process with the rate \( \lambda \) per D and reservation request per one time segment arrives according to the Poisson distribution with mean rate \( \lambda / R \). The holding time is general distribution with mean g and the number of servers is S which is equivalent to that of time slots. In this model, mean holding time is approximated by the sum of constant service time d and mean notice time f.

\[
g = d + f \quad (14)
\]

Therefore, the blocking probability (Br) for the reservation request can be approximated by the Erlang B formula:

\[
Br = a/S! \sum_{i=0}^{S} (a/i)! (15)
\]

\[
where \ a = \lambda (d + f)/R \text{ is the offered traffic load in Erlang.}
\]

This approximate method is a simplified model. Because of the mutual dependence between arrival process and holding time, this model may not be valid in the strict sense.

Second, we discuss the arrival process model of reservation requests for one time segment during the interval I (e.g. one month) in which user can make reservation. Let the Poisson process \( \lambda / R \) denote the mean reservation request arrival rate for one time segment during I. Then, the probability of n reservation request arrivals for a time segment during I is

\[
P_{n}(I) = e^{-\lambda I/R} \left( \frac{\lambda I}{R} \right)^{n}/n! 
\]

\[
\quad n = 0, 1, 2, \ldots \quad (16)
\]

Since S is the number of the time slots per time segment, blocking state occurs in case that more than S reservation requests arrive. Therefore, the blocking probability of reservation request (Br) is

\[
Br = \sum_{i=1}^{S} P_{S+i}(I) \frac{i}{(S+i)}
\]

\[
= a_{i=1} e^{-\lambda I/R} \left( \frac{\lambda I}{R} \right)^{i}/(S+i)! (17)
\]

where \( i/(S+i) \) is conditional blocking probability in the case that S+i reservation requests arrive during I.

Third, we consider the model of multiserver queueing system with finite waiting room. This model is shown in Fig.4. This model consists of a first stage queue of unlimited capacity, a second stage queue having S finite capacity and S servers. The server attaches to the second stage queue. The arrivals at first stage queue occur in accordance with Poisson process with mean \( \lambda / R \) and the service time has constant distribution with mean d. The server's operation is assumed to start at the beginning of the time segment. Operation is as follows. The instant a switch is closed, reservation calls in the first stage queue
enter the second stage queue and the servers begin the reservation service. However, as the maximum number of capacity of second stage queue are $S$, reservation requests more than $S$ are blocked and the blocked calls are cleared. Switching period is $1/d$. This model is similar to the above mentioned model which considers the arrival process of reservation request.

Next, we deal with the case that blocked reservation requests are scheduled for other time segments.

First, we consider the blocking probability of reservation request in (17). Now, we assume that if a new reservation request is blocked, the blocked reservation request arrives repeatedly up to $R$ times according to the Poisson process. From the above assumption, total arrival rate ($A$) for one time segment during 1 is given by

$$A = \lambda (1 - B^R + 1) / (1 - B) \quad (18)$$

Applying the total arrival rate in (18) to (17), the blocking probability of reservation request including blocked requests is obtained. In this model, as blocked call arrival process is treated by Poisson process, total arrival rate is smaller than actual arrival rate.

Second, we consider the model of multiserver model with finite waiting room. This system is similar to the polling system consisting of a number of buffered input terminals connected to a computer by transmission line. Multiserver model of reservation system is shown in Fig. 5. This system consists of $R$ first stage queues having $S$ finite capacity, $R$ second stage queues having $S$ finite capacity and $S$ servers. The arrivals at each first stage queue occur in accordance with independent Poisson process (mean arrival rate to first stage queue $i$ is denoted by $\lambda_i$, $i = 1, ..., R$) and the service time of each queue has constant distribution (mean service time is $d$). The service operation is assumed to start at the beginning of the time segment $d_i$ and the service time of each queue has constant distribution (mean service time is $d$). The service operation is assumed to start at the beginning of the time segment $d_i$ and the service time of each queue has constant distribution (mean service time is $d$).

In this model, in case a reservation request is refused in the first stage queue, the modified request arrives at another first stage queue immediately with some probability again. In accordance with the above assumption and operation, it is shown that reservation system can be modeled as a multi queue model with the overflow call phenomenon. Although arrival process of the overflow call from the first stage queue $i$ to queue $j$ loses the nature of random process, the variance of the overflow call in the queueing system is smaller than that of loss system. Therefore, in this model, the arrival process of the overflow call is approximated by the Poisson process. Most actual reservation system in which blocked reservation requests are scheduled for other time segments is well reflected in this approximate method. As for the analysis of the blocking probability, this model is independent of servers. Therefore, it is sufficient to consider only the arrival process, the mean number of capacity and the finite queue.

Third, from the view point of the multiserver queueing model with finite capacity as shown in Fig. 4, we examine an approximate method. We assume that the modified request
arrives with some probability (a) again in case a request is refused. In accordance with the above assumption, reservation system can be modeled as a multiserver queueing model with repeated call phenomenon which relates to the loss system, as shown in Fig.6. This model may be analyzed by the approximate method in reference [13].

5. SHARING BETWEEN LOST-CALL AND RESERVATION CALL

In this section, we discuss the capacity assignment strategy sharing between lost-call and reservation call.

If a reservation service shares a transmission facility with lost-call services (complete sharing policy), it seems that the quality of service of the latter may be inferior to the former and vary widely as a function of random fluctuations of carried reservation traffic. However, it is expected to improve resource utilization by sharing resources between lost-call and reservation call. Therefore, it is necessary to consider the traffic model that priority is assigned to lost-call (the sharing with priorities), with some schemes, for example, trunk reservation. On the other hand, all the lost calls are blocked as long as all shared time slots are reserved by reservation traffic. As for this point, it is necessary to consider the strategy which allocates a given number of dedicated time slots exclusively to lost call and shares the other time slots. This strategy is the sharing with minimum allocation.

6. CONCLUSION

In this paper, we have described several capacity assignment strategies for integrated communication system with heterogeneous traffic and discussed the approximate method giving the blocking probability of demand call for lost-call-cleared mode and that of reservation request for reservation mode.

In particular, we have investigated the reservation system in which all reservation calls request the same capacity. And we have presented the simple approximate methods such as the modified Erlang B model, the arrival process model, the multiserver queueing model with overflow phenomenon and the multiserver queueing model with repeated call phenomenon.

The results of this study will be applied to a multi-service digital satellite network based on DA/TDMA and a terrestrial-based ISDN.

ACKNOWLEDGEMENTS

The authors express their gratitude to Dr. H. Kaji, director of KDD R and D Laboratories, Dr. K. Nosaka, vice director and Dr. K. Ono, deputy director, for their continuous guidance.

REFERENCES