CHAPTER II

TIMEOUT-AWARE INTER-QUEUING FOR QOS PROVISIONING OF REAL-TIME SECONDARY USERS

2.1. Introduction

This chapter describes the proposed works of QoS achievement of cognitive radio network that has secondary users of different priorities. At the preliminary level, two types of users namely RT and NRT are considered in existing literature, in which preference is given to RT users. In the similar way, users of multiple priorities can also be considered. This chapter describes the proposals in the literature in this regard and the proposed work of the thesis. In the literature, many of the works are dealt with two levels of priorities. The need of extending it to multiple priorities and improvement of it when compared to two level priorities are explored here. This is to build solutions which support QoS of real-time SUs, while ensuring the QoS of PUs and efficient utilization of bandwidth.

Adequate care is taken in cognitive radio networks, to protect the transmissions of PUs. But there are no such measures to protect the transmissions of SUs. Opportunistic spectrum access is allowed to SUs, with which the unoccupied PU channels (called spectrum holes) are accessed by the SUs. However, SUs need to vacate the channels whenever their PUs returned back. In such instants, SU needs to search for another spectrum-hole to continue its transmissions. When it finds another spectrum hole, by using spectrum handoff, it occupies the new channel and continues its communication. But if it fails to find another spectrum hole then it should be dropped, which will lead to QoS degradation of SUs.

When delay-sensitive data needs to be transmitted by cognitive radio devices, to support some applications like real-time audio/video, satisfying their QoS in terms of delivering the data within the stipulated deadline times is very essential. In such cases, prioritizing the SUs based on their urgency and allocating the vacant channels depending on the priority basis becomes more relevant. In this backdrop of fluctuating nature of PUs' arrivals and departures, offering QoS to SUs is a challenging issue.

When SUs are dealing with delay sensitive real-time data, then, they may not complete their communication within their time deadline, if all the SUs are treated equally. It is more so when the PU activity is high, which results in lesser number of spectrum holes or demand from SU is more. In such cases, to support the real-time data, the SUs are classified into four priorities, namely SU1s, SU2s, SU3s and SU4s, based on their time deadlines. Among these, SU1s which are hard real-time users, should be treated as highest priority among all SUs. SU2s and SU3s are soft real-time users, which would take next levels of priority, out of which SU2s have more priority than SU3s. Finally, SU4s are non real-time users having least priority.

In addition to defining four levels of priorities, Timeout-aware inter-queuing is proposed here. Using this, when the users of SU2s and SU3s are about to reach their time deadline, they will be shifted to next high level priority. Until they reach that deadline, they continue with the priority of that class. If they don't get the transmission opportunity with that priority and the time deadline reaches, then they are moved to higher priority. Advantage of this method is that they don't come for competition with critical transmissions from the beginning, facilitating for better accessing of channels by time-critical transmissions of those high priority users. Some of the issues of spectrum sharing are described below.

2.2. Classification and Challenges of Spectrum Sharing

The sharing of available spectrum among cognitive users can be done in different ways. Based on the way SUs are using the spectrum opportunity, three types of spectrum sharing paradigms are available (Goldsmith, A., Jafar, S.A., Maric, I. & Srinivasa, S., 2009). They are:

(a) Underlay: Here PUs and SUs coexist together such that SUs will use less power, which doesn't disturb the PU's communications.

(b) Overlay: Here SUs will make use of spectrum holes with the help of messages and code books shared by PUs. In return, SUs need to help the PU's transmissions.

(c) Interweave: In this paradigm SUs should continuously sense the spectrum for finding and exploiting the spectrum holes.

Decision of channel-allocation can take place through central or distributed mechanisms. In the case of central decision, one central system decides about all the channel allocations in that region collecting the data from all the cognitive radios of that region about their measurements and as well as transmission requirements. In the case of distributed decision, each cognitive radio takes the decision about how to use the vacant channels on its own. This kind of decision may be required in the cases where the cognitive radios operate in an environment where proper structure of the network doesn't exist (Niyato, D.& Hossain, E., 2008).

As per another classification of spectrum-sharing, there can be cooperative vs. non-cooperative sharing schemes. In cooperative scheme, all the radios in the region cooperate with each other by exchanging measurements and channel requirements. In fact, the central decision mentioned in the previous classification needs this kind of cooperation. In (Jayaweera, S. K., Vazquez-Vilar, G. & Mosquera, C., 2010) the authors expressed the cooperation between PUs and SUs as PUs will offer bandwidth or some time slots to SUs when PUs' weak transmissions are relayed by SUs' powerful transmitter. As the PUs' transmissions get finished fast with the help of SUs, the left over time can be offered to secondary users without a need of really sensing the spectrum or without the dilemma of vacating it when the PU returns back. However, it happens only when PU needs SUs' services.

The authors of (Elkourdi, T. & Simeone, O., 2013) described cooperation as if the interference of PU is decoded by SU to mitigate the interference, then PU will help SUs by assigning its portion of bandwidth to SU for some time.

In the case of non-cooperative spectrum sharing, this kind of exchange does not take place, and hence the individual radios depend on their own measurements to learn about the radio environment. Decision of channel-usage, taken in isolation in non-cooperative schemes, is suitable for ad-hoc based networks (Niyato, D.& Hossain, E, 2008).

One more type of spectrum-sharing is based on game theory where the users will define the utility function and play maximizing the utility function. The three major types of games are enlisted below. (Yu Peng Xie, Xuezhi Tan, Yu Tao Liu, Lin Ma, Hai Yan Wu., 2011; Beibei Wang, Yongle Wu, K. J. Ray Liu, 2010; Niyato, D.& Hossain, E., 2007):

Cournot model: Here all players will play simultaneously and the utility function is in terms of quantity of spectrum.

Stackelberg model: Here the players will play one after the other and the utility function is also in terms of quantity of spectrum.

Bertrand model: Here all players will play simultaneously and the utility function is in terms of price of spectrum.

The challenges of spectrum sharing are (Akyildiz, I. F., Lee, W. Y., Vuran, M. C. & Mohanty, S., 2008):

- continuous change of operating frequency due to spectrum handoff
- assumptions of researchers that SUs know power and location of PUs so as to calculate the interference.

2.3. Related Work from Literature

In (Gelabert, X., Pérez-Romero, J. , Sallent, O. & Agustí., R., 2008), the authors have addressed the challenges of Radio Access Technology (RAT) selection strategies in multiple access, multiple service wireless networks. To analyze the RAT selection policies, 4D Markov chain is used, which has considered TDMA (Time Division Multiple Access) and WCDMA (Wide Band Code Division Multiple Access) based channel access methods. Voice and data type services correspond to the two levels of priorities in this. Five RAT selection policies were discussed in it.

a. Random (Rnd) RAT selection policy: Here voice and data are randomly given TDMA or WCDMA based channels.

b. Service-based-1 (SB-1) RAT selection policy: In this, voice users are given TDMA based access and data users are given WCDMA based channel access. If TDMA channels are overloaded, voice users can be given WCDMA based channels and if WCDMA channels are over loaded, data users can be assigned with TDMA channels.

c. Service-Based-2(SB-2) RAT selection policy: In this, voice users and data users will be given WCDMA based access and TDMA based channels respectively. If WCDMA channels are overloaded, then voice users can be given TDMA based channels and if TDMA channels are overloaded, then data users can be assigned WCDMA channels.

d. Load balancing (LB) RAT selection policy: As the aim is to maintain less load, care has to be taken while allocating the spectrum.

e. Multimode Terminal - Driven (MMTD) RAT selection policy: TDMA based channels are assigned to single mode users whereas WCDMA based channels are assigned to multi mode users. If no WCDMA based channels are free, then TDMA based channels will be assigned for multimode users.

In the above work, the authors considered two types of traffic, voice and data and assigned TDMA or WCDMA type channels, based on the type of RAT selection strategy. However, no special care is taken for real-time data, for example, in voice communications.

In (Liang, Z., Feng, S., Zhao, D. & X. S. Shen., 2011), priority is given to real-time traffic than best effort traffic, because real-time systems like health care, where abnormal condition of patient should reach the doctor immediately, is of higher priority. Two methods of channel allocation are proposed, namely reservation based channel allocation method and absolute priority channel allocation method. In reservation based channel allocation method, SUs will be given channel only if channel is existed during pre-allocated time intervals whereas in absolute priority channel allocation method, whenever free channel is available, real-time traffic is served. Best Effort Traffic is served when free channels are available and no further real-time users are waiting for service.

The authors of (Alshamrani, A., Shen, X. S. & Xie, L. L., 2011) considered heterogeneous real-time (RT) and non-real time (NRT) SUs. In view of offering better QoS to RT users, the available spectrum holes are first assigned to RT users and left over spectrum holes to NRT users. QoS in terms of blocking and dropping probabilities is considered here.

In the above two works, SUs are classified into two types although in real world, in many applications, two types of SUs cannot cover all priority levels.

The authors of (Kunert, K., Jonsson, M. & Bilstrup, U., 2012) proposed different inter frame spaces in view if giving different priorities based on applications. Less inter frame space is used for high priority applications and high inter frame space for low priority applications, for example, 10µs for voice, 12µs for video, 14µs for best effort and 16µs for background type data.

In the above work, SUs are classified into four types in view of various applications like voice, video, best effort and back ground packet types. But giving priority to different types of SUs would be better dealt with the help of queues rather than inter frame spaces.

In (Jian Wang, Aiping Huang, Wei Wang, and Tony Q.S. Quek, 2013) the authors treated all SUs as single group only. But they used two buffers, one for handoff SUs and the other for new SUs. Some channels are reserved for handoff SUs as call dropping is more painful than call blocking. They also considered the SUs leaving the queues due to impatience because of long waiting times. Under these circumstances they derived blocking and dropping probabilities using Markov chain models.

Here, instead of treating the real-time SUs separately, they have given importance to handoff SUs than new requesting SUs.

In (Motamedi, N. , Kumar, S., Hu, F. & Rowe, N., 2013), best channel with high channel holding time is assigned to user with highest priority to overcome the disturbances to be experienced by high priority SUs. For this, it needs prediction to anticipate the holding times of the channels.

In (Doost-Mohammady, R., Naderi, M. Y. & Chowdhury, K. R., 2014), the authors considered two types of priorities for two types of traffic of SUs: Streaming and Non-Streaming. Their idea is to find successive channels such that it will satisfy the bandwidth requirement of secondary users. They have derived the optimum number of channels to be required for reservation of channels for streaming type SUs. Here they support the real-time SUs by reserving some channels for them. But reservation of channels may increase unnecessary blocking. When no channels are free except the reserved channels and if a non-streaming SU requests for service and no streaming SU needs the service, it will be blocked even when reserved channels are free. Another challenge here is that assigning contiguous channels may not be possible all the times.

In (El Helou, M., Ibrahim, M., Lahoud, S., Khawam, K. , Mezher, D.& Cousin, B., 2015), the authors addressed the problem of selfish users, with whom system performance degrades. Their idea is if secondary users are provided with network parameters like QoS parameters and cost, SUs combine them with needs and preferences and will take better decisions. As supplying the network parameters is not an easy task, they introduced reinforcement learning algorithm to obtain the network parameters. Semi-Markov Decision Process (SMDP) is used to derive the network information, which meets the objectives of both operator and SU. Here the reinforcement algorithm may not assure in providing network information.

To maintain QoS of SUs, authors of (Yang, W. & Zu, Y., 2016) used M-LDWF (Maximum Largest Weighted Delay First) algorithm in which the users who have waited for more time will be selected each time from the queue. In addition to this, they have classified SUs into two types as latency sensitive and latency insensitive. For each type of SUs they have used two queues, one for newly arriving SUs and the other for handoff SUs. The priority order of these four queues is as follows:

Handoff Latency Sensitive> New Latency Sensitive> Handoff Latency insensitive> New Latency insensitive.

Here, only two priorities of SUs are considered. Furthermore, there are chances of some RT SUs might be lost due to their early deadlines. The M-LDWF algorithm can't assure the successful service of all RT SUs.

The authors of (Jong- Hong Park & Jong Moon Chung, 2016) have considered 'M' channels and each channel contains 'N' sub channels. They have considered that one channel is required for PUs and one sub channel is sufficient for SUs. They have divided the SU traffic into two groups: SU1s and SU2s, out of which SU1s are given high priority compared to SU2s. They have implemented prioritized channel allocation in a different manner as shown in Figure 2.1 (Jong- Hong Park & Jong Moon Chung, 2016). Channel assignment to PUs is done from left most and towards right. Channel allocation of SU1s is done from right most and towards left. SU2s' channel allocation is done from middle and towards left. From this it can be observed that SU1s will experience less blocking and dropping compared to SU2s.

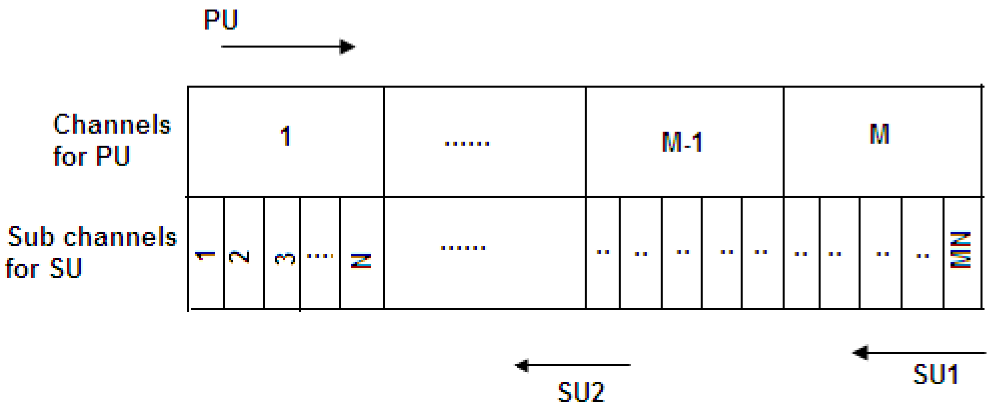

Figure 2.1. System Model of Channel Assignment for Prioritized SUs

But this type of channel allocation is suitable to systems which don't have fixed frequency bands for their PUs like the cellular, but not suitable for other cases like television white spaces.

The cognitive radio networks involve full of arbitrariness due to random data generation and dynamic availability of spectrum holes. Random data generation is due to random transmission requirements of different users, who want to download chapters, movies, songs etc., and dynamic changing of data flow during conversations based on the situations and videos with variable rate.

To deal with this randomness in the generation of data and dynamic availability of spectrum holes, queues are introduced.

2.4. Multiple Priority Users Serviced Through Queuing

2.4.1.Queue

Queuing theory deals with the study of queues. Formation of queues is common in real life situations like ticket reservation counters, banks, super market billing counters, for Bhagavan darshan, at ration shops and so on. Queues form when the arrival rate is more than service rate of users. The basic model of queuing is shown in Figure 2.2.

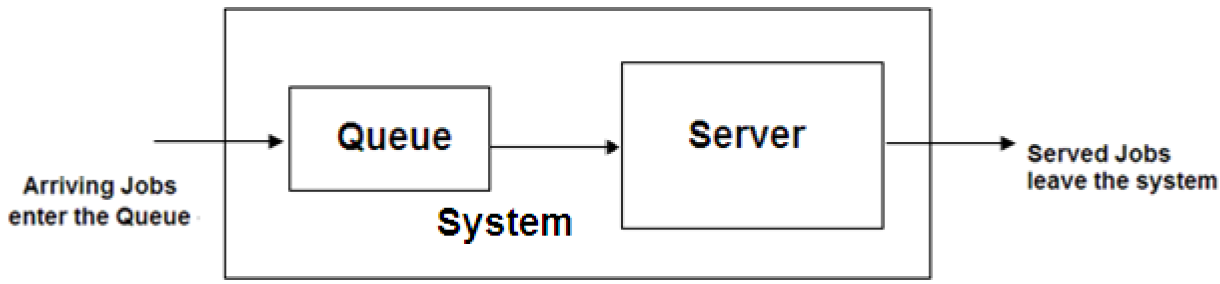

Figure 2.2. Basic Queuing Model

Though arriving jobs may follow any probability distribution, most of the times Poisson distribution is considered, as number of arrivals at one time instant is independent of number of arrivals at the remaining time instants. The arrival rate of users or jobs is represented as 'λ'. Once the users enter the queue, they will be given service based on queue-discipline. The length of the queue can be finite or infinite. In finite length queues users cannot enter the queue when queue is full. Usually the users enter the rear end of the queue and clear from the front end.

Preemptive priority queues are used sometimes to carry out the prioritized job. When a high priority SU is requesting for channel and no channel is free, then the low priority SU which is using the channel will vacate it for the requested high priority SU and wait in the queue. The service completion of interrupted transmissions due to preemption will resume its transmission instead of retransmission, whenever it gets the next chance of utilizing the spectrum hole.

2.4.2. Terminology of Queues

a. Balking: Users don't enter the queue due to presence of large number of users in the queue.

b. Reneging: Users will drop in the middle from the queue due to impatience of waiting for much time.

c. Jockeying: Users change from one queue to another queue when the queue length of destination queue is reducing fast.

2.4.3. Queuing Disciplines

There are mainly three types of queuing disciplines. They are:

i) **FIFO (First In First Out):** The user who enters first will be given service first. It is also called FCFS (First Come First Serve). Figure 2.3 illustrates the FIFO queue discipline. This queuing discipline is used in most of the applications. Single queue is used here, for all the users.

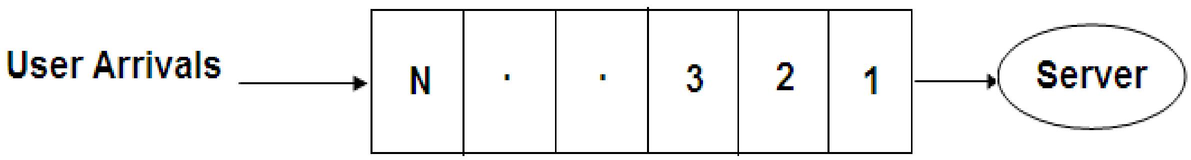

Figure 2.3. FIFO Queue Discipline

ii) **Priority queuing:** The high priority and low priority users are maintained in separate queues. The service will be always be given to high priority users queue unless it is empty. Figure 2.4 shows this priority queuing discipline.

iii) **Round Robin or Fair Queuing:** Multiple queues are used here. In this discipline, all users will be given equal priority and service will be given to users of all queues one after the other continuously. This is described in Figure 2.5.

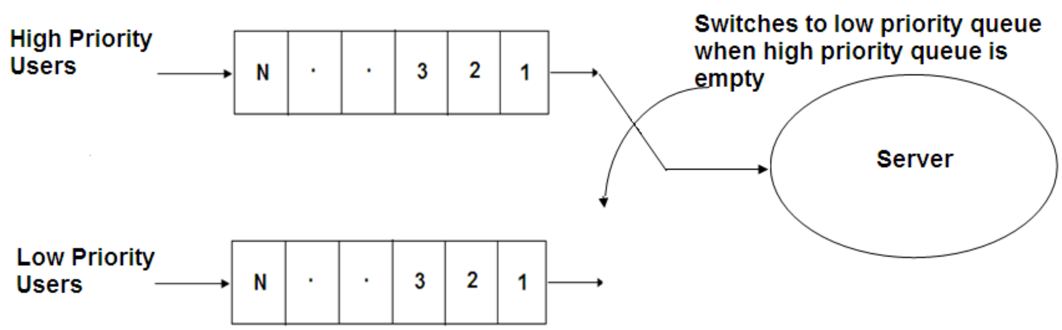

Figure 2.4. Priority Queue Discipline

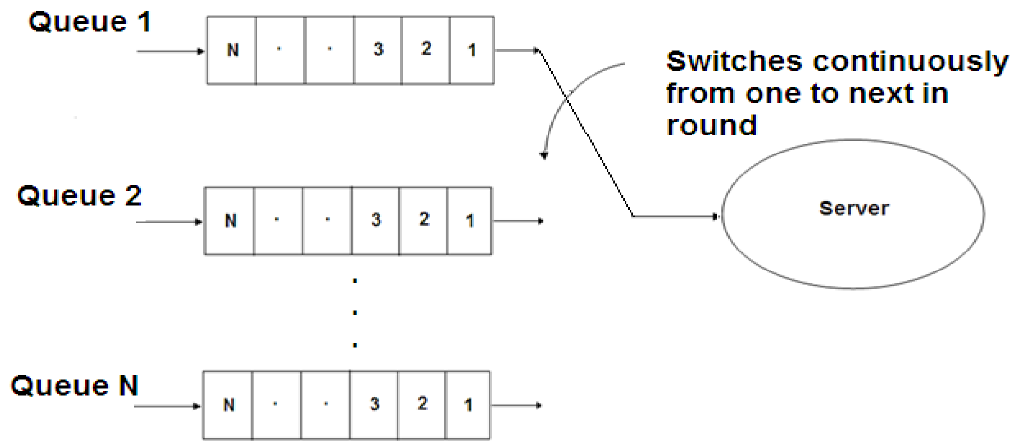

Figure 2.5. Round Robin Queue Discipline

The service rate of the server can follow any random distribution. In most of the applications, exponential distribution is cosidered for service rate, which is represented by the symbol ‘μ’.

2.4.4. Nomenclature of Queues

The expression used to describe the queuing system is called Kendall’s notation. It is in the form of:

A/B/C/D/E, where

‘A’ represents the inter arrival time or arrival rate probability distribution of users. It is ‘M’ (Markovian) if it follows exponential distribution for inter arrival times and Poisson for arrival rates. Markovian is memoryless and independent. That is, the number of users arriving in one time instant is independent of number of users coming in other time instants.

‘B’ represents the service time probability distribution of server. It is represented as ‘M’ (Markovian) if it follows exponential distribution and ‘D’ if they are constant time intervals.

‘C’ denotes number of servers,

‘D’ is the System capacity or Queue capacity and

‘E’ is Queue discipline.

Traffic intensity is defined as ratio of arrival rate of users to service rate of server.

$$\rho = \frac{\lambda}{\mu}$$

2.5. Proposed Queuing Technique

To support real-time SUs, SUs are classified into four types SU1s, SU2s, SU3s and SU4s, out of which first three belong to real-time traffic. The most delay sensitive SUs come under SU1s, treated as highest priority SUs. From SU1s to SU3s, delay sensitivity is considered to be decreasing in that order, and so priorities will be decreased. The last type is non-real time, which doesn't have any time deadline.

By categorizing the real-time SUs into three groups, the probability of giving service to real-time SUs within their time deadline can be improved compared to the case where there are no priority categories or only two types of SU priority types exist. In addition to the distribution of real-time SUs into three groups, timeout-aware inter-queuing is proposed here, to improve further the chances of offering channels to real-time SUs within their time deadlines. The results of proposed technique are compared with the two-level priority SU counterpart.

2.5.1 System Model

Centralized decision making, cooperative type cognitive radio network is considered. That is, all sensing information and channel requests from individual SUs are sent to a central coordinator. Based on the available spectrum holes and number of type-wise SU requests, the central coordinator assigns the channels. Initially it assigns the available vacant channels to SU1s. After fulfilling all SU1 requests, if vacant channels are still available, then it proceeds to allot them to SU2s, then to SU3s and finally to SU4s in that order.

The arrival times and conference times of PUs and the four types of SUs are taken as uniform distribution type. 'N' non-overlapping channels are considered. Each PU is considered to have its own channel. The PUs have freedom to use the channels at any time of their interest as they are licensed users. Each SU of the network has its own transceiver. SUs should use the channels opportunistically after knowing the status of spectrum holes through sensing. Whenever SU knows that a free channel is available, then SU can use it and thereafter should sense the spectrum periodically to notice the reappearance of PU. Whenever it notices the reappearance of the PU, the SU should vacate and occupy another free channel if it is available or drop the transmission otherwise. Four queues are considered for the four types of SUs.

Further refinement of this scheme can be obtained by incorporating the mechanism of queue shifting. Deadline times are considered here to implement this mechanism. For example, within a priority class, if there are some users whose transmissions are having lesser deadline times compared to some other users whose transmission deadlines are little higher, then the later users can be put into next priority class that has lesser priority than the priority given to former users. Then they continue with that lesser priority and try to get the transmission opportunities. As the time elapses, their deadline times would be approaching the values of former users. At those instants, at which the deadline values fall below the threshold value that decides the criticality, these later users are also brought to high-priority users' category.

Advantage of this approach is that, these later devices which have higher time deadlines don't compete immediately with those later devices which have smaller time deadlines. They enter into the competition for high-priority, only when their deadlines also reach that criticality. Due to this mechanism, critical transmissions are handled effectively, with little impact on the performance of non-critical transmissions. It is proposed as timeout-aware inter-queuing, in which the user of a low priority queue will be shifted to high priority queue when it is about to reach its time deadline.

In the simulated environment, all the devices are able to read all the available data (variables) from MATLAB environment. Hence the central entity knows about the availability of free channels and priority requirements of all the devices. However, in the real environment this information will be transmitted on the forward and reverse control channels, in the specified frame formats.

The PU occupancy information can be modelled as Markov chain as shown in Figure 2.6 (Alshamrani, A., Shen, X. S. & Xie, L. L., 2011). The primary channel occupancy can be written as (Alshamrani, A., Shen, X. S. & Xie, L. L., 2011).

$$\delta_i = \frac{\beta_i}{\alpha_i + \beta_i} \quad \text{for i=1,2,3,…N} \qquad (2.1)$$

where

α_i represents the channel transition probability from occupied to idle state, and

β_i represents the probability of channel transition from idle to occupied state

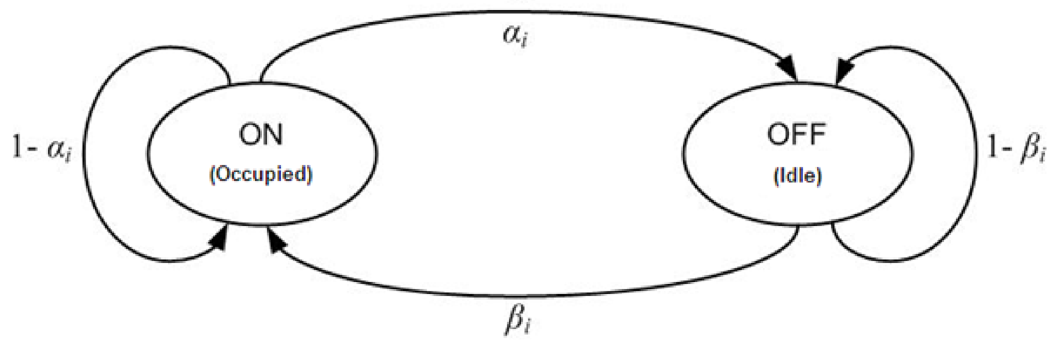

Figure 2.6. Model of channel state

The priority based channel allocation is illustrated in Figure 2.7. When SU1, SU2 and SU3 are competing for one spectrum hole, SU1 becomes the winner as it is the highest priority SU.

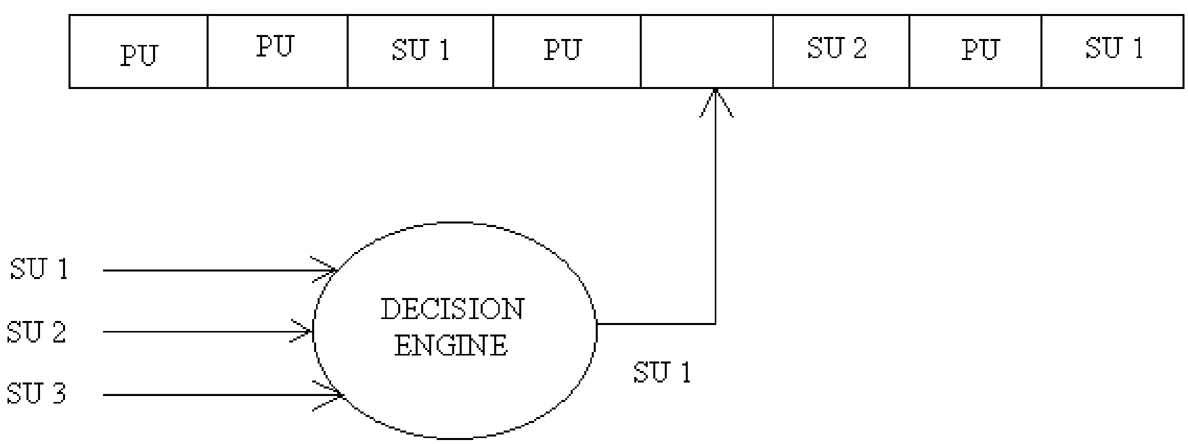

Figure 2.7. Priority based Channel Allocation

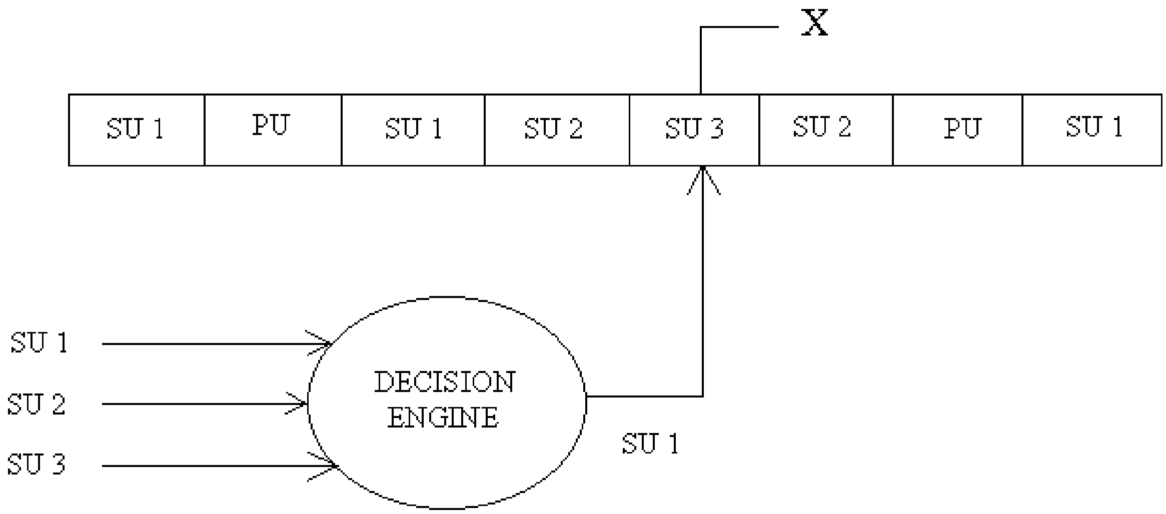

Figure 2.8. Pre-emption

The system supports pre-emption, in which a low priority SU needs to vacate the channel when a high priority SU needs it. This is illustrated in Figure 2.8. In this case, the winner SU1 is unable to detect a free channel to its use as all channels are completely occupied by PUs and SUs. So the least priority among the SUs which are using the channels, here SU3, needs to vacate the channel for the sake of honouring the request from SU1.

2.5.2. State Space Model to Derive Blocking Probability

In (Peter J. Smith, Abdulla Firag, Pawel A. Dmochowski, and MansoorShafi, 2012), blocking probabilities are derived for PUs and SUs for three cases namely, Secondary User Cleared (SUC), Secondary User Equality (SUE) and Secondary User Partially Cleared (SUPC). Here, SUC means - when a PU arrives and no free channels are available then the SU that is in service should be dropped, which is nothing but preemption. In this scenario, no sensing errors are considered. In SUE both SUs and PUs are treated equally - that means no preemption occurs and it is equal to the case of imperfect sensing and SU will continue its transmission due to misdetection. In SUPC, the SUs will drop partially when PU arrives. That is, sensing is neither perfect nor imperfect.

In the proposed work, SUs are divided into four priority classes SU1s, SU2s, SU3s and SU4s. The highest priority is given to SU1s and SU4s are given least priority.

Let there are N channels in the system where PUs and SUs follow Poisson distribution for arrivals and exponential distributions for service times. Let the arrival rates of PUs is λ_p, SU1s is λ_{s1}, SU2s is λ_{s2}, SU3s is λ_{s3} and SU4s is λ_{s4}. Let the service rates of PUs is μ_p, SU1s is μ_{s1}, SU2s is μ_{s2}, SU3s is μ_{s3} and SU4s is μ_{s4}. Let the state of the system is represented by (n_p, n_{s1}, n_{s2}, n_{s3}, n_{s4}),

where,

n_p= Number of in-service PUs

n_{s1}= Number of in-service SU1s

n_{s2}= Number of in-service SU2s

n_{s3}= Number of in-service SU3s

n_{s4}= Number of in- service SU4s

At every minute, based on the arrivals or departures of PUs and SUs, state may be changed. Here three cases are considered for deriving blocking probability by considering N=1.

Case-1: System with four prioritized SUs, imperfect sensing, without preemption and timeout-aware inter-queuing.

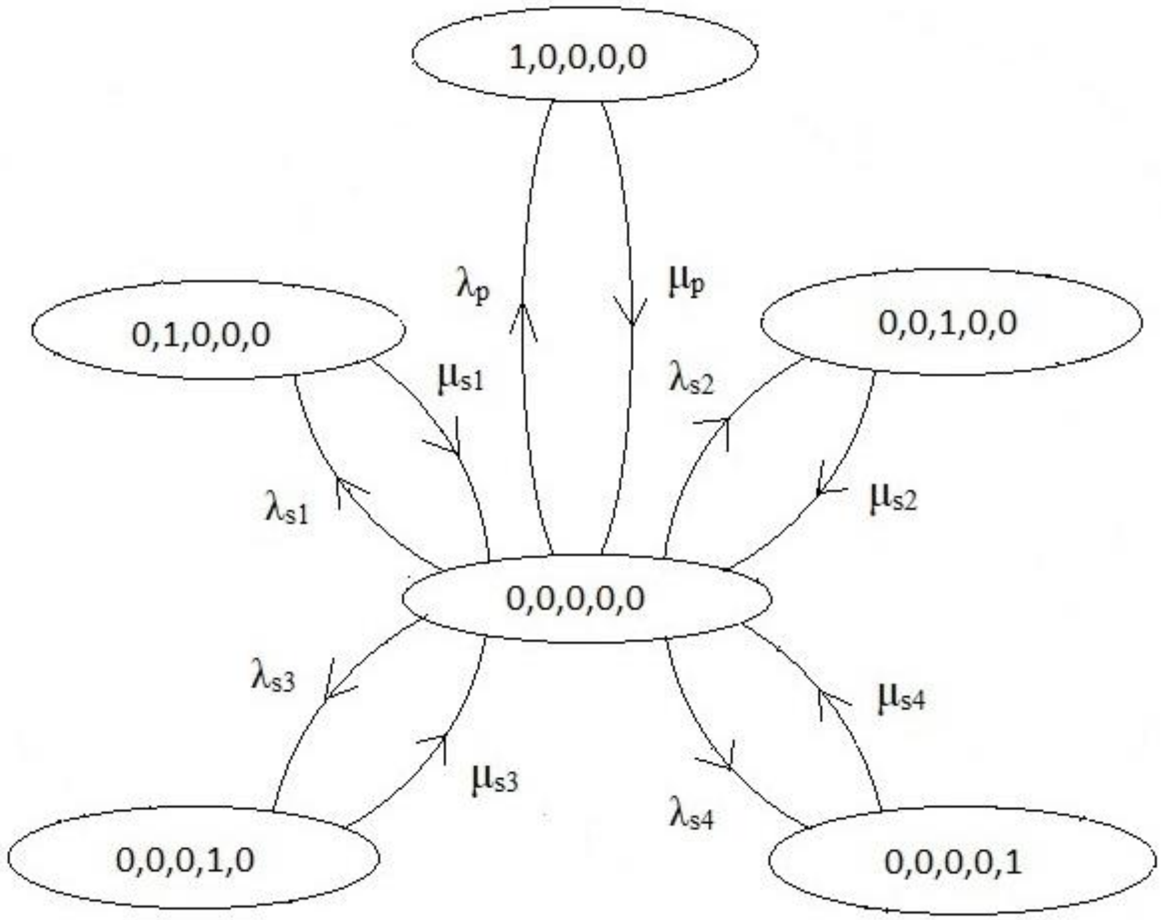

Figure 2.9. Markov Model for case-1

Markov model for case-1 is shown in Figure 2.9. Writing the flow balance equations at each state results in

$P_{0,0,0,0,0}(\lambda_p+\lambda_{s1}+\lambda_{s2}+\lambda_{s3}+\lambda_{s4})=$
$\mu_p p_{1,0,0,0,0} + \mu_{s1} p_{0,1,0,0,0} + \mu_{s2} p_{0,0,1,0,0} + \mu_{s3} p_{0,0,0,1,0} + \mu_{s4} p_{0,0,0,0,1}$ ……….....(2.2)

$p_{0,0,0,0,0}\lambda_p = p_{1,0,0,0,0}\mu_p$ …………………………………………………..(2.3)

$p_{0,0,0,0,0}\lambda_{s1} = p_{0,1,0,0,0}\mu_{s1}$ ………………………………………………….(2.4)

$p_{0,0,0,0,0}\lambda_{s2} = p_{0,0,1,0,0}\mu_{s2}$ …………………………………………………………….…(2.5)

$p_{0,0,0,0,0}\lambda_{s3} = p_{0,0,0,1,0}\mu_{s3}$ …………………………………………………………….…(2.6)

$$p_{0,0,0,0,0}\lambda_{s4} = p_{0,0,0,0,1}\mu_{s4} \qquad (2.7)$$

Sum of all steady state probabilities is one. So

$$p_{0,0,0,0,0} + p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1} = 1 \qquad (2.8)$$

By solving 2.2 to 2.8 one can get all state probabilities.

Blocking probability of all SU types = $p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1}$

……………...…….(2.9)

That is, if the channel is occupied by any user, the new SU request will be blocked.

Case-2: System with four prioritized SUs, perfect sensing without preemption and timeout-aware inter-queuing.

Markov chain for case-2 is shown in Figure 2.10. Writing the flow balance equations at each state results in

$$P_{0,0,0,0,0}(\lambda_p+\lambda_{s1}+\lambda_{s2}+\lambda_{s3}+\lambda_{s4}) =$$
$$\mu_p p_{1,0,0,0,0} + \mu_{s1} p_{0,1,0,0,0} + \mu_{s2} p_{0,0,1,0,0} + \mu_{s3} p_{0,0,0,1,0} + \mu_{s4} p_{0,0,0,0,1} \qquad (2.10)$$

$$(p_{0,0,0,0,1} + p_{0,0,0,1,0} + p_{0,0,1,0,0} + p_{0,1,0,0,0} + p_{0,0,0,0,0})\lambda_p = p_{1,0,0,0,0}\mu_p \qquad (2.11)$$

$$p_{0,0,0,0,0}\lambda_{s1} = p_{0,1,0,0,0}(\mu_{s1} + \lambda_p) \qquad (2.12)$$

$$p_{0,0,0,0,0}\lambda_{s2} = p_{0,0,1,0,0}(\mu_{s2} + \lambda_p) \qquad (2.13)$$

$$p_{0,0,0,0,0}\lambda_{s3} = p_{0,0,0,1,0}(\mu_{s3} + \lambda_p) \qquad (2.14)$$

$$p_{0,0,0,0,0}\lambda_{s4} = p_{0,0,0,0,1}(\mu_{s4} + \lambda_p) \qquad (2.15)$$

Sum of all steady state probabilities is one. So

$$p_{0,0,0,0,0} + p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1} = 1 \qquad (2.16)$$

By solving 2.10 to 2.16 one can get all state probabilities.

Blocking probability of all SU types = $p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1}$

……………...(2.17)

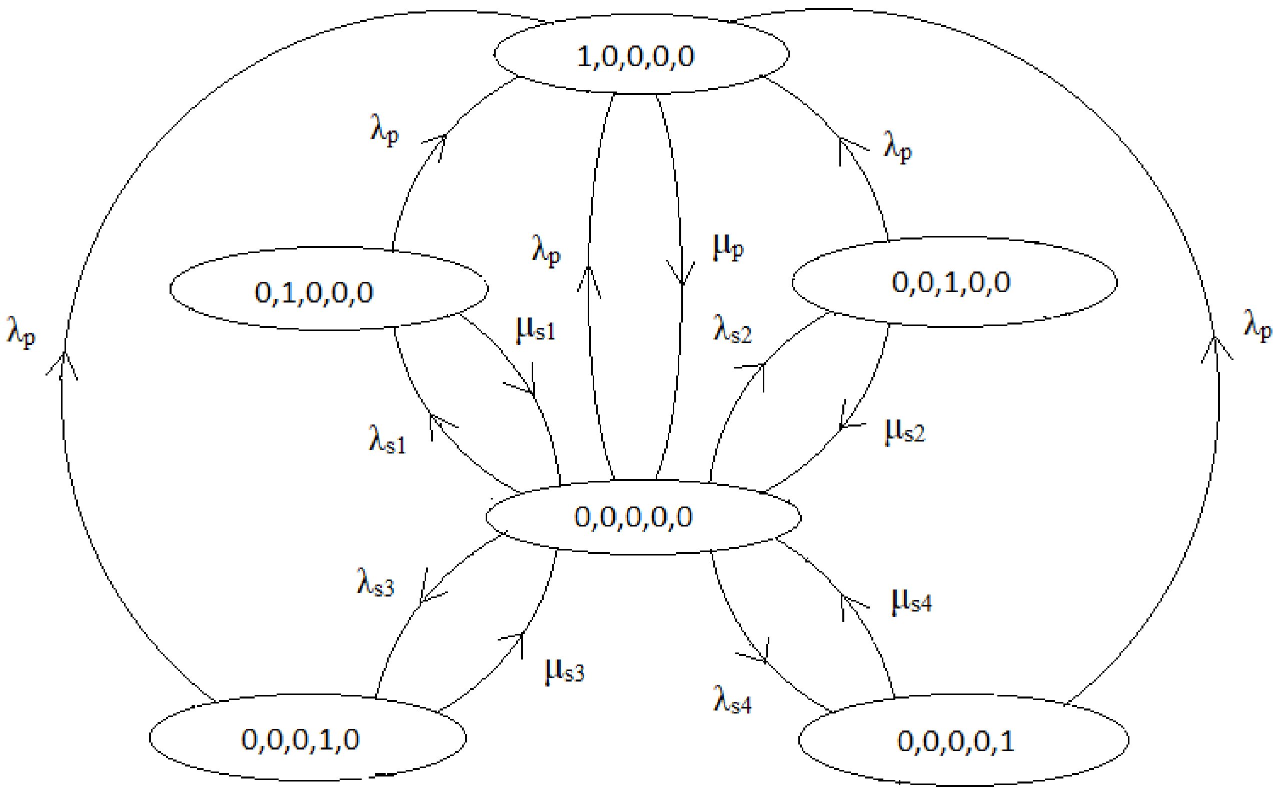

Figure 2.10. Markov Model for case-2

Case-3: System with four prioritized SUs, perfect sensing with preemption and without timeout-aware inter-queuing.

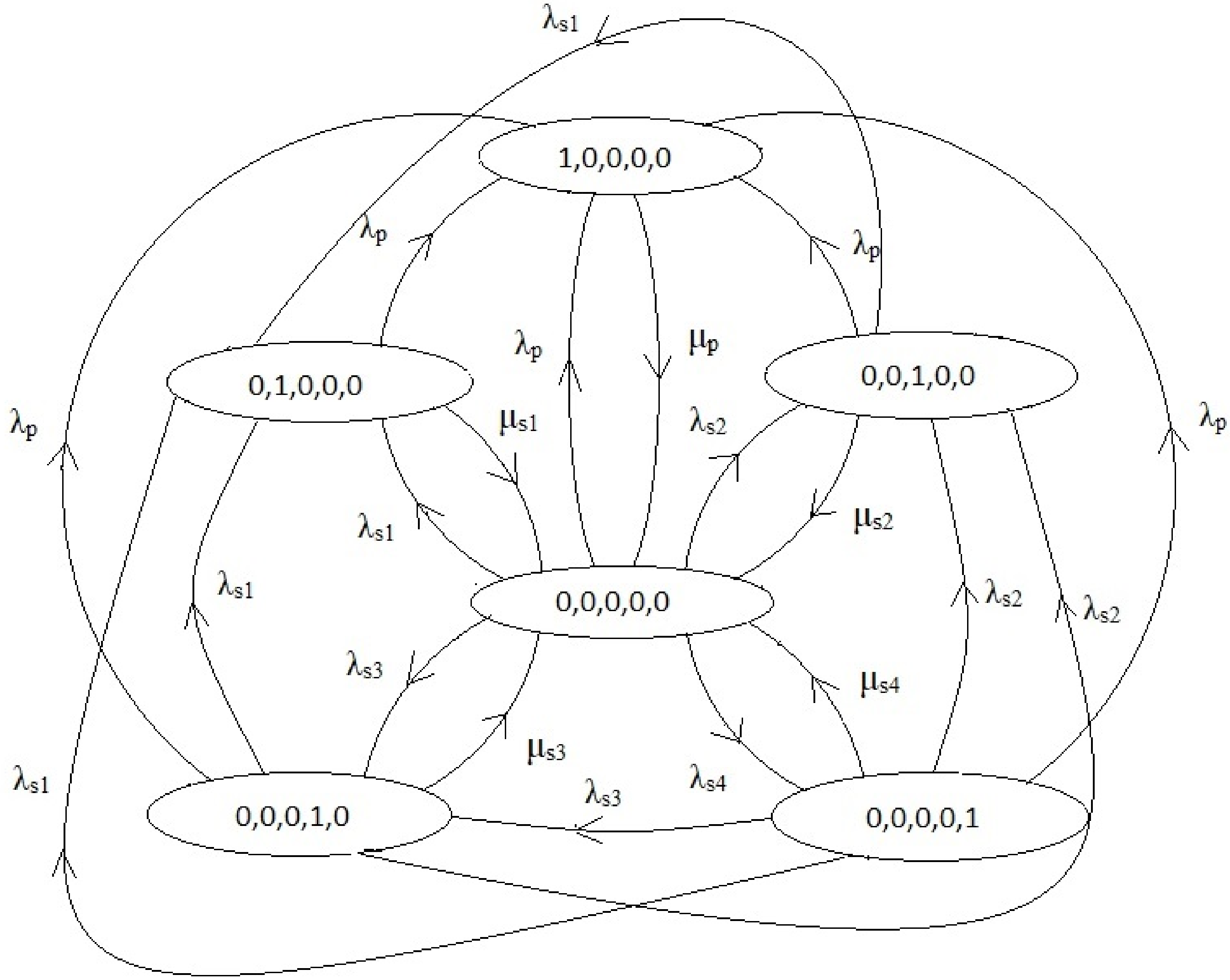

Figure 2.11. Markov Model for case-3

Markov chain for case-3 is shown in Figure 2.11. Writing the flow balance equations at each state results in

$P_{0,0,0,0,0}(\lambda_p+\lambda_{s1}+\lambda_{s2}+\lambda_{s3}+\lambda_{s4})=$
$\mu_p p_{1,0,0,0,0} + \mu_{s1} p_{0,1,0,0,0} + \mu_{s2} p_{0,0,1,0,0} + \mu_{s3} p_{0,0,0,1,0} + \mu_{s4} p_{0,0,0,0,1}$ ……………………...…(2.18)

$(p_{0,0,0,0,1} + p_{0,0,0,1,0} + p_{0,0,1,0,0} + p_{0,1,0,0,0} + p_{0,0,0,0,0})\lambda_p = p_{1,0,0,0,0}\mu_p$ ……………...……(2.19)

$(p_{0,0,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1})\lambda_{s1} = p_{0,1,0,0,0}(\mu_{s1} + \lambda_p)$ ……………………..(2.20)

$(p_{0,0,0,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1})\lambda_{s2} = p_{0,0,1,0,0}(\mu_{s2} + \lambda_p + \lambda_{s1})$ …………………........... (2.21)

$$(p_{0,0,0,0,0} + p_{0,0,0,0,1})\lambda_{s3} = p_{0,0,0,1,0}(\mu_{s3} + \lambda_p + \lambda_{s1} + \lambda_{s2}) \quad (2.22)$$

$$p_{0,0,0,0,0}\lambda_{s4} = p_{0,0,0,0,1}(\mu_{s4} + \lambda_p + \lambda_{s1} + \lambda_{s2} + \lambda_{s3}) \quad (2.23)$$

Sum of all steady state probabilities is one. So

$$p_{0,0,0,0,0} + p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1} = 1 \quad (2.24)$$

By solving 2.18 to 2.24 one can get all state probabilities.

Blocking probability of SU1s = $p_{1,0,0,0,0} + p_{0,1,0,0,0}$ ……(2.25)

Blocking probability of SU2s = $p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0}$ ……(2.26)

Blocking probability of SU3s = $p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0}$ ……(2.27)

Blocking probability of SU4s = $p_{1,0,0,0,0} + p_{0,1,0,0,0} + p_{0,0,1,0,0} + p_{0,0,0,1,0} + p_{0,0,0,0,1}$ ……(2.28)

It can be understood from equations 2.25 to2.28 that the blocking probability is increasing when priority is getting reduced.

Case-4: System with four prioritized SUs, perfect sensing with preemption and timeout-aware inter-queuing.

Here when a RT user's Time-out is about to reach, then that real-time SU will be shifted from its priority level to its next higher priority level. So blocking of real-time SUs is further reduced with the introduction of timeout-aware inter-queuing.

2.6. Results and Discussion

The simulation parameters of PUs, SUs and system are given in Tables 2.1, 2.2 and 2.3 respectively. These are applicable for the simulations carried out in chapters 2 to 5.

In the simulation environment considered here, the priority needs of the devices are made available to the central entity through the data available in the MATLAB environment, because all the devices can read all the available data (variables) from MATLAB environment. Through this, the central entity knows about the availability of free channels and priority requirements of the devices. However, in real scenarios, this information is communicated through control channels.

Blocking probability of SUs where the secondary users are categorized only into two types is shown in Figure 2.12 for high PU occupancy and high SU demand. The two types of SUs are RT and NRT. The spectrum hole usage opportunities are first offered to RT SUs as they belong to high priority, then to NRT SUs if still spectrum holes are available. So the blocking probability of RT SUs is less compared to NRT SUs.

Table 2.1 Simulation parameters of PUs

Parameter	High PU occupancy	Low PU Occupancy
Maximum number of channels	10	10
Maximum holding time of Pus	20 minutes	10 minutes
Maximum number of times each PU is reappearing	20	10

Table 2.2 Simulation parameters of SUs

Parameter	High SU demand	Low SU Demand
Maximum holding time of SUs	20 minutes	10 minutes
Maximum number of times each SU is reappearing	20	10
Number of SUs in each type	20	10

Table 2.3 Simulation parameters of system

Parameter	**Value**
Modulation	8PSK
Samples/frame	300
Sample time	0.33µs
Output samples/symbol	8
Channel type	AWGN
Signal to noise ratio	10dB
Input signal level referenced to one ohm	1/8
Symbol period	1µs
Filter	Raised cosine
Rolloff factor of filter	0.22
Filter span in symbols	12
Output samples per symbol	8
Linear amplitude filter gain	1
Decision type	Hard decision

Generally RT SUs will have a time deadline and if service to RT SUs is offered only within their time deadline conditions is considered then their corresponding blocking is increased with respect to the case where channel allocations are done even after their time deadlines but considering only time deadline based priority for channel allocation. The drawback of this technique is that some RT SUs are missing the opportunity of getting the channel as all RT SUs are in single queue, which won't allow the SUs whose time deadline is about to reach. They are offering services to the SUs based on first come first serve principle. The blocking of NRT SUs is very large as its priority is least.

The blocking probability of secondary users with Timeout-Aware Inter-Queuing is shown in Figure 2.13. It can be noticed that the blocking probability of highest priority SUs, that is SU1s, is less compared to the remaining SUs. It is also observed that its value is 0.71 when only one channel is available and is gradually decreasing and reaches '0' when number of channels become five. The blocking probability of next highest priority SUs, that is SU2s, is '1' when only one channel is available and gradually decreasing and reaches zero when number of channels become eight. Blocking of next level RT SUs is '1' upto five channels and then starts decreasing and becomes 0.12 when number of channels reaches ten. The last priority secondary users, called NRT SUs or SU4s are experiencing a highest blocking probability of '1' up to six channels case, and then starts decreasing and becomes 0.75 when number of channels reach ten. It also describes the blocking scenario of SU1s, SU2s and SU3s when channel access is given to them within their time deadline.

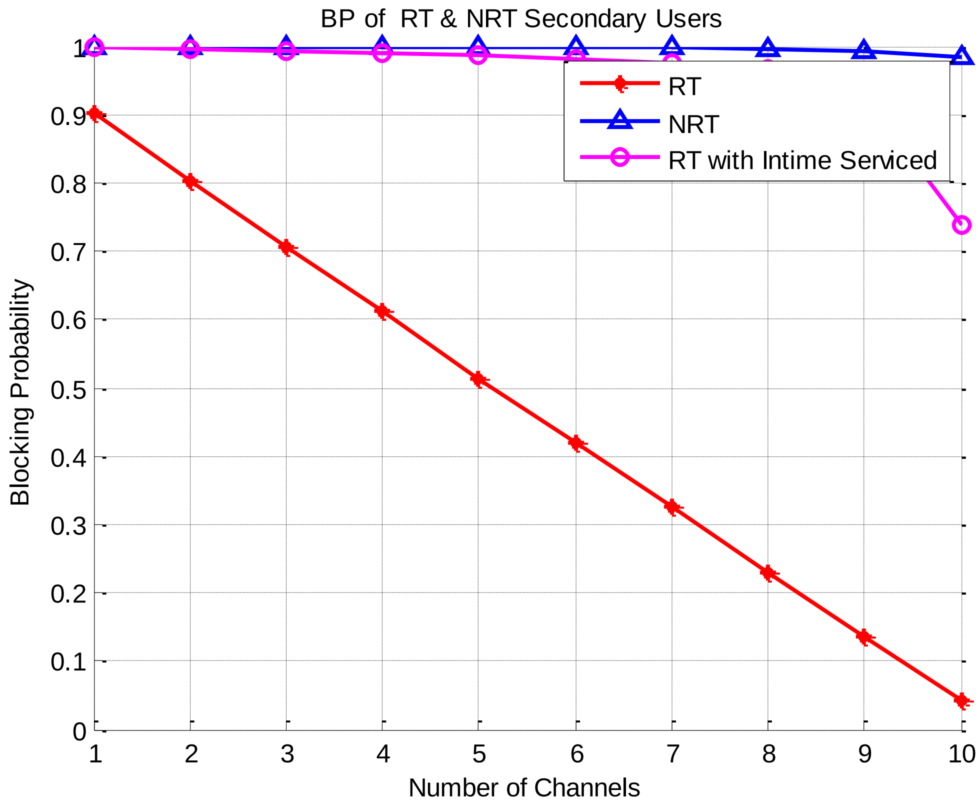

Figure 2.12. Blocking Probability of RT and NRT Secondary Users for High PU Occupancy and High SU Demand

The comparative blocking of real-time SUs with and without Timeout-Aware Inter-Queuing is shown in Figure 2.14. It can be observed that the blocking probability of RT SUs is very less in the Timeout- Aware Inter-Queuing concept.

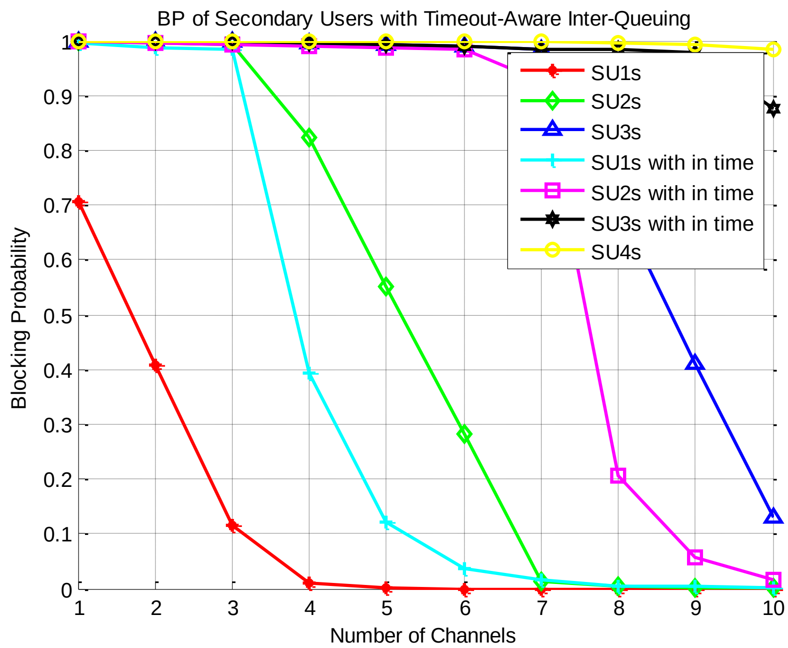

Figure 2.13. Blocking Probability of Secondary users with Timeout-Aware Inter-Queuing for High PU Occupancy and High SU Demand

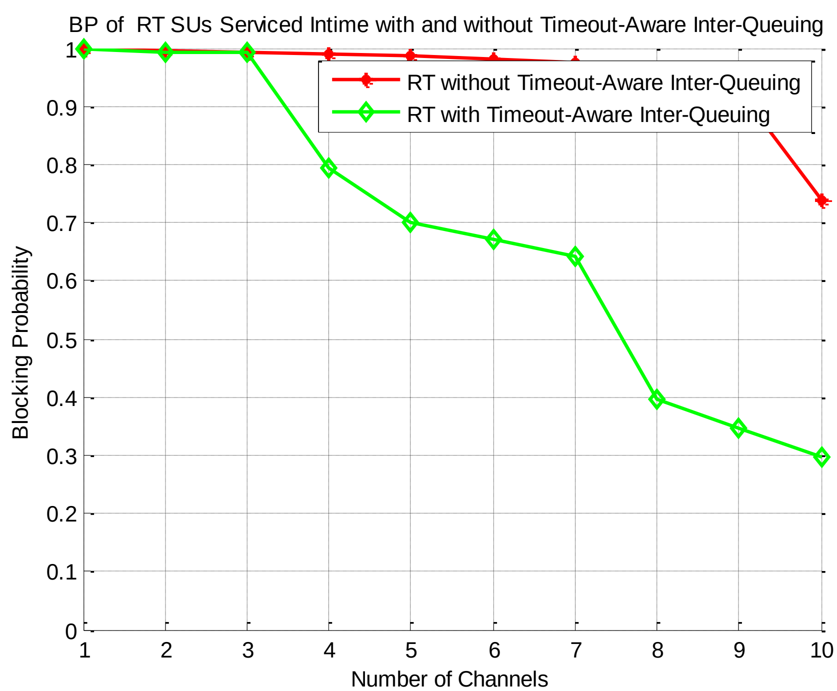

Figure 2.14. Blocking Probability of RT SUs with and without Timeout-Aware Inter-Queuing for High PU Occupancy and High SU Demand

When SU demand is changing from high to low and PU occupancy remains the same, the corresponding results are shown in Figures 2.15 to 2.17.

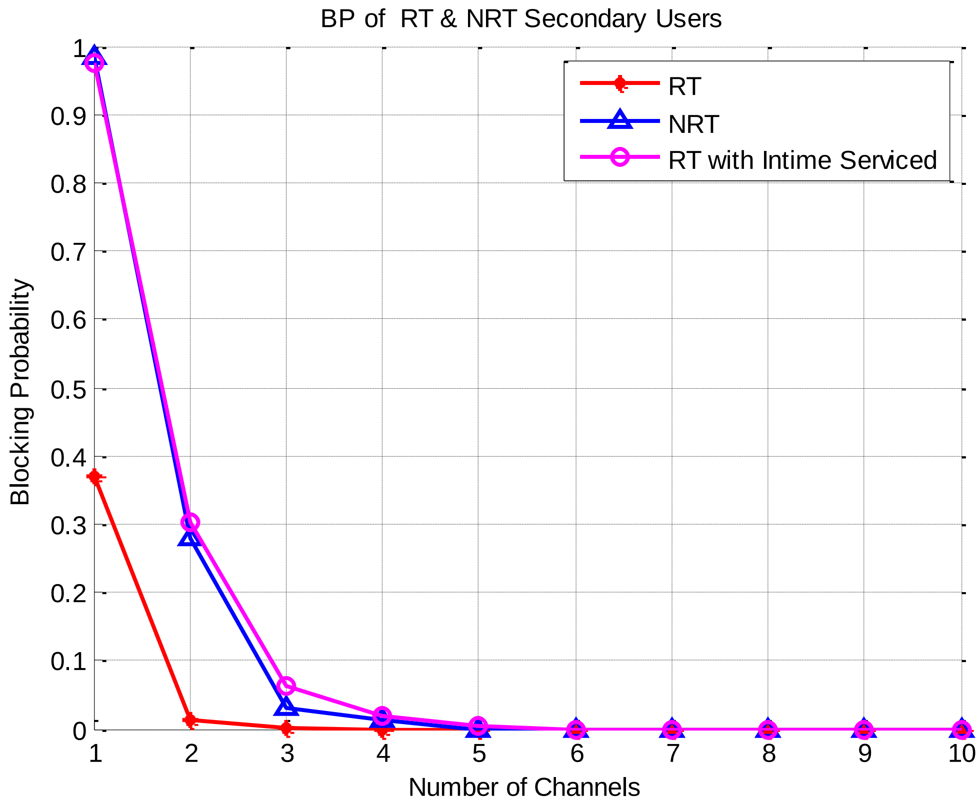

Figure 2.15. Blocking Probability of RT and NRT Secondary Users for High PU Occupancy and Low SU Demand

From Figures 2.14 and 2.17 it can be noticed that the proposed method offers higher improvement in high PU occupancy and high SU demand case than high PU occupancy and low SU demand case.

When SU demand is kept at high and by changing the PU occupancy from high to low, the corresponding results are shown in Figures 2.18 to 2.20.

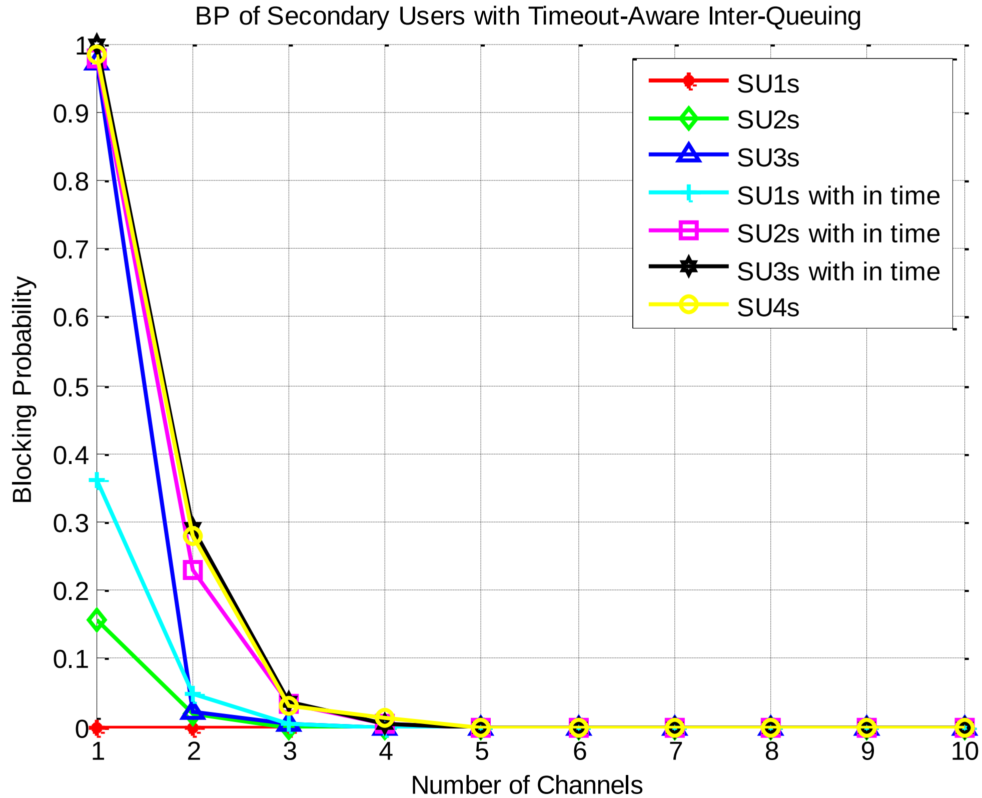

Figure 2.16. Blocking Probability of Secondary users with Timeout- Aware Inter-Queuing for High PU Occupancy and Low SU Demand

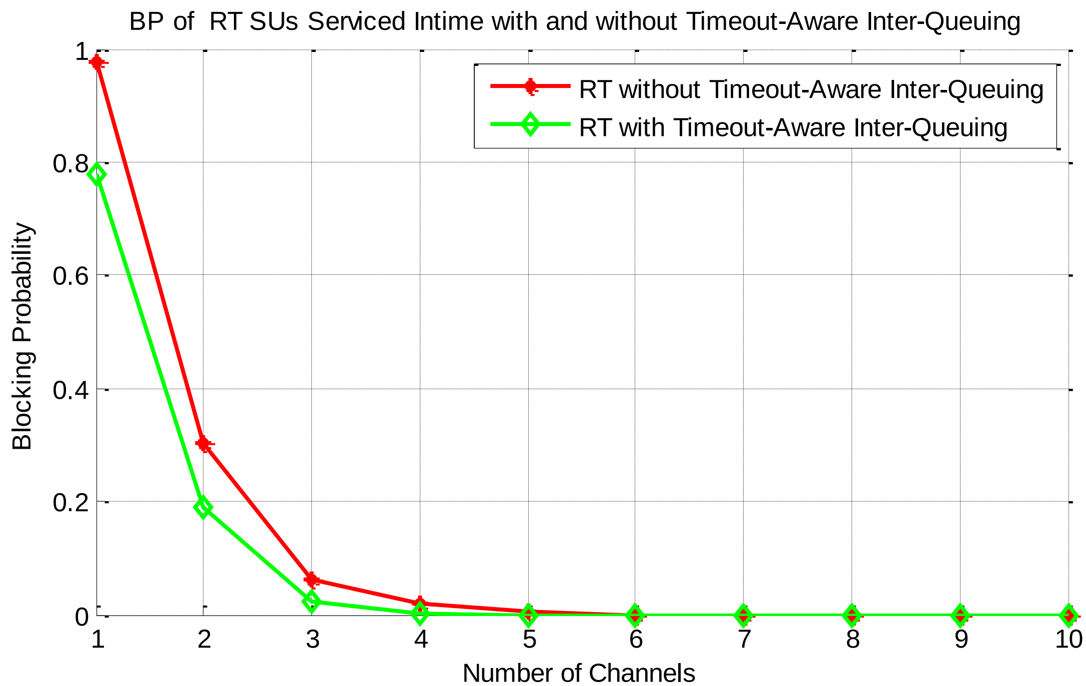

Figure 2.17. Blocking Probability of RT SUs with and without Timeout-Aware Inter-Queuing for High PU Occupancy and Low SU Demand

From Figures 2.18 to 2.20, it can be observed that the blocking probabilities are greatly reduced for all type SUs compared to high PU occupancy and high SU demand case. Low PU occupancy means the possibility of finding spectrum holes is increased and hence blocking probability is increased.

Here also the improvement of the proposed method is more significant when compared to conventional two priority case.

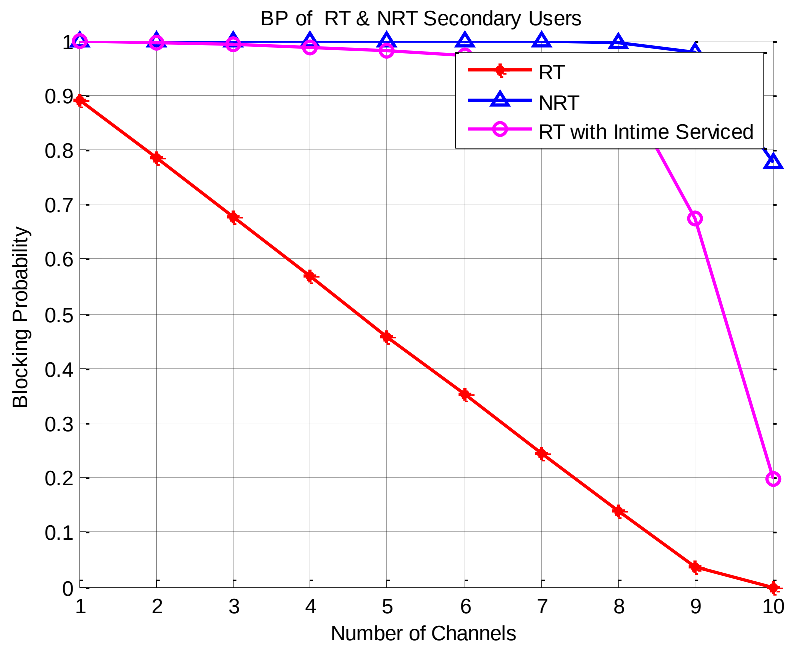

Figure 2.18. Blocking Probability of RT and NRT Secondary Users for Low PU Occupancy and High SU Demand

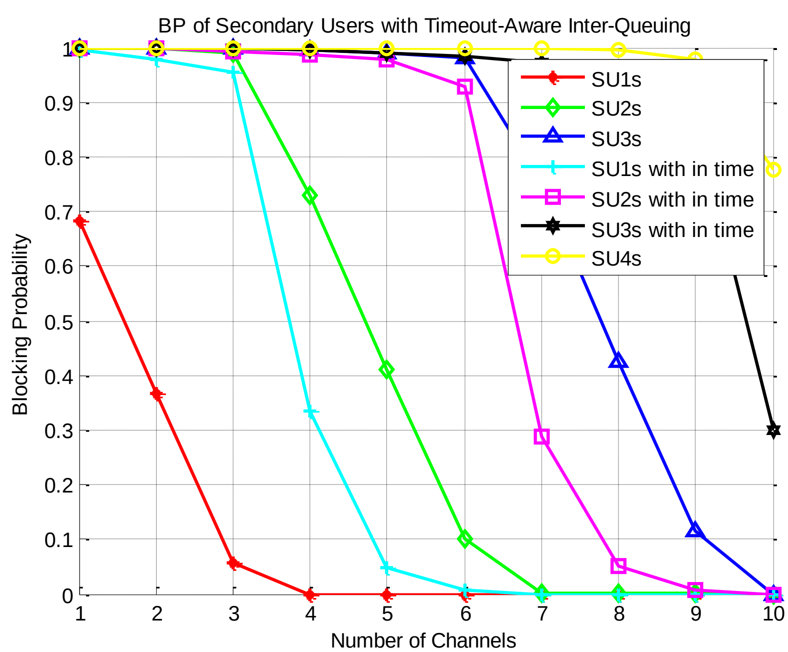

Figure 2.19. Blocking Probability of Secondary users with Timeout-Aware Inter-Queuing for Low PU Occupancy and High SU Demand

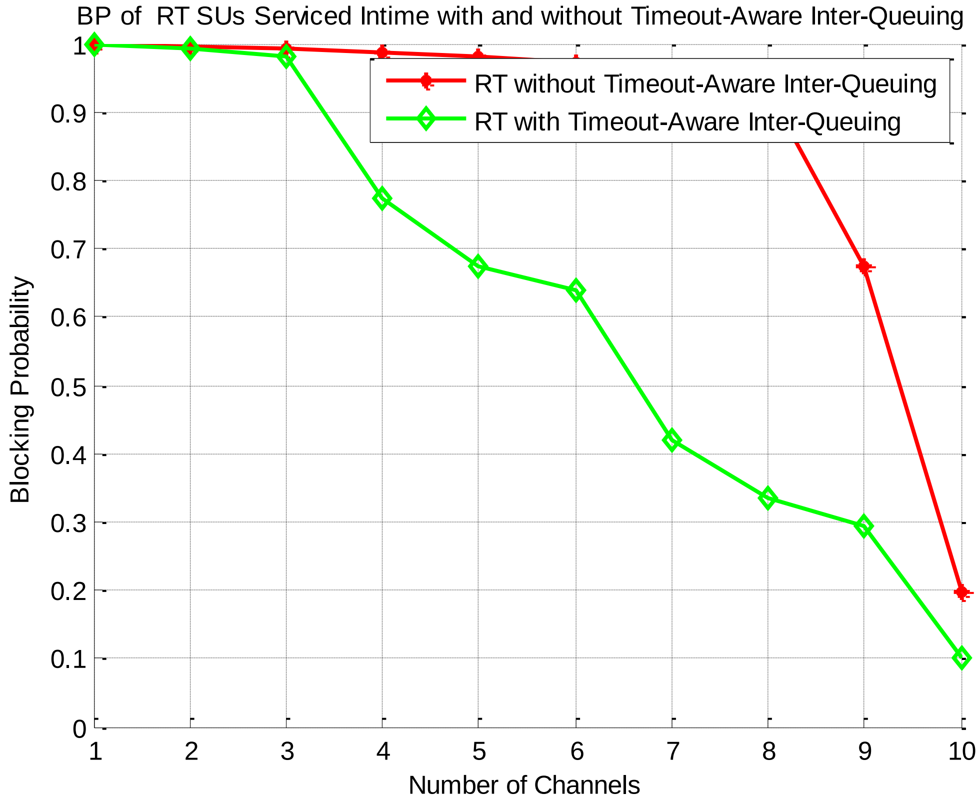

Figure 2.20. Blocking Probability of RT SUs with and without Timeout-Aware Inter-Queuing for Low PU Occupancy and High SU Demand

The comparative results of proposed technique and conventional techniques are shown in Table 2.4.

Table 2.4. Performance improvement of proposed technique with respect to conventional existing techniques

	High PU occupancy and High SU demand	High PU occupancy and Low SU Demand	Low PU Occupancy and High SU Demand	Low PU Occupancy and Low SU Demand
Percentage improvement	29.6	18.3	31.56	18.4

From Table 2.3 it is clear that the proposed technique yields better improvement under High SU demand, that is for high competition than low competition and there is a small improvement from high PU occupancy to low PU occupancy scenario.

2.7. Summary

Quality of service to secondary users based on multiple priorities is explored in this work. Users of four priority levels, SU1 to SU4 are considered, with SU1 being the highest priority and SU4, the least priority. In addition to priority levels, inter queue shifting, i.e shifting of the users from one priority level when their time deadlines are approached is also introduced. The proposed 'Timeout-aware inter queuing' is compared with existing method of RT and NRT user classification. Performance improvement in the form of reduced blocking probabilities in the range of 18% to 31% is observed for the high priority users. As High priority SUs are the first ones that get channel allocations as per their priority order, the low-priority SUs will suffer higher delays. If the delays of these low-priority SUs are inordinate, then appropriate mechanisms like 'credit based shaper' (Johennes, S. & Soheil, S., 2017 ; Luxi, z., Paul,P., Zhong, Z. & Qiao, L., 2018) can be adopted, to support the 'best effort' traffic of these devices. However, focus of this thesis is limited to address the priority needs of high priority users.