Voice Capacity and Data Response Time in Cognitive Radio Networks

by

Subodha Gunawardena

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The growing interest towards wireless communication services over the recent years has increased the demand for radio spectrum. Inefficient spectrum management together with the scarcity of the radio spectrum is a limiting factor for the development of modern wireless networks. As a solution, the idea of cognitive radio networks (CRNs) is introduced to use licensed spectrum for the benefit of the unlicensed secondary users. However, the preemptive priority of the licensed users results in random resource availabilities at the secondary networks, which makes the quality-of-service (QoS) support challenging. With the increasing demand for elastic/interactive data services (internet based services) and wireless multimedia services, QoS support becomes essential for CRNs. This research investigates the voice and elastic/interactive data service support over CRNs, in terms of their delay requirements. The packet level delay requirements of the voice service and session level delay requirements of the elastic/interactive data services are studied. In particular, constant-rate and on-off voice traffic capacities are analyzed over CRNs with centralized and distributed network coordination. Some generic channel access schemes are considered as the coordination mechanism, and call admission control algorithms are developed for non-fully-connected CRNs. Advantage of supporting voice traffic flows with different delay requirements in the same network is also discussed. The mean response time of the elastic data traffic over a centralized CRN is studied, considering the shortest processor time with and without preemption and shortest remaining processor time service disciplines, in comparison with the processor sharing service discipline. Effects of the traffic load at the base station and file length (service time requirement) distribution on the mean response time are discussed. Finally, the relationship between the mean response times of interactive and elastic data traffic is studied.

Acknowledgements

First and foremost, I would like to thank my Ph.D supervisor Prof. Weihua Zhuang for the support given to me from the first day of my study at the University of Waterloo until this day. It was her excellent mentorship and guidance which took me to this place as an award winning researcher. Her kind words gave me strength to face the hardships I have encountered during the period of my Ph.D program. I am thankful for her rigorous research attitude, innovative thinking, and efficient work style, which helped to shape my work ethic.

Further, I would like to thank Prof. Patrick Mitran, Prof. Chengguo Weng, and Prof. Liang-Liang Xie for taking part in my thesis advisory committee, and Dr. Xianbin Wang form the University of Western Ontario for taking part in my thesis examination committee. I would also like to thank Dr. Guangzhe Fan for serving in my Ph.D. thesis advisory committee while he was associated with the University of Waterloo. They all devoted their precious time reading the research proposal and the final thesis to help me improve the quality and accuracy of this research. Their comments were very valuable in finalizing this thesis.

Special thanks go to Prof. Jon W. Mark and Prof. Sherman (Xuemin) Shen of the Centre for Wireless Communications (CWC) for coordinating the Broad Band Communications Research (BBCR) group and organizing weekly group meetings, which makes an excellent platform to share the knowledge among students. Group meetings helped me to broaden my knowledge and to understand some of the difficult subjects. I am also very grateful to other BBCR Lab members: Dr. Rongxing Lu, Dr. Lin Cai, Dr. Ho Ting (Anderson) Cheng, Dr. Tom (Hao) Luan, Dr. Hao Liang, Dr. Towhidul Islam, Dr. Sanaa Taha, Xiaohui Liang, Jian Qiao, Khadige Abboud, Miao Wang, and visiting scholars Dr. Tingting Yang, Hangsheng Zhao, and Hongwei Luo. Their kindness and friendship made my experience in Waterloo comfortable and pleasant.

My sincere gratitude goes to my former graduate supervisor Prof. Nandana Rajatheva who encouraged me for postgraduate studies and my friend Udaya Jayathilaka who has been there for me since my undergraduate studies. Special thanks are also due to the University of Waterloo, one of the most innovative universities in North America. I am proud of having been part of it.

Finally, I would like to thank my parents, my wife, and my friends for their love, endless support, and encouragement throughout my life.

To my dear wife, parents, and loving friends

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

BS	Base station
CAC	Call admission control
CAS	Channel access scheme
CDF	Cumulative distribution function
CR	Cognitive Radio
CRN	Cognitive radio network
DTMC	Discrete-time Markov chain
EB	Effective bandwidth
EC	Effective capacity
EDGE	Enhanced data rates for GSM evolution
FCFS	First-come first-serve
GSM	Global System for Mobile
ISM	Industrial, scientific, and medical
LHS	Left hand side
LTE	Long term evolution

MAC	Medium access control
MMPP	Markov modulated Poisson process
PDF	Probability density function
PS	Processor sharing
PU	Primary user
QoS	Quality of Service
RHS	Right hand side
SPTNP	Shortest processor time without preemption
SPTWP	Shortest processor time with preemption
SRPT	Shortest remaining processor time
STR	Service time requirement
SU	Secondary user
TCP	Transmission control protocol
VoIP	Voice over internet protocol
WLAN	Wireless local area network

Symbols

α	Shape parameter of the Weibull distribution
β	Scale parameter of the Weibull distribution
$\acute{\beta}$	Transmission probability selection parameter for slot-ALOHA
χ^2	Chi-Square value obtained for Chi-Square test
χ_{new}	Next packet with the highest waiting time (new queue head)
χ_{old}	Previous packet with the highest waiting time (old queue head)

δ^*	Common QoS exponent to make arrival and service processes equal
$\delta^*_{\min}$	Minimum δ^* required to satisfy the QoS requirement
F	State transition probability matrix of the channel
$\acute{F}$	Modified state transition probability matrix of the channel
ϵ	Maximum delay bound violation probability
Γ	Input data set to the CAC algorithm
i_r	Receiver node of call i
i_s	Source node of call i
λ	Mean arrival rate of requests
λ_i	Mean arrival rate of system unit i under consideration
λ_I	Mean arrival rate of interruptions
λ_t	Mean arrival rate of `type t` requests
$\mu_S(\cdot)$	Service process of the system
μ_i	Mean departure rate of system unit i under consideration
ω_r	Target receiver node
ω_s	Target source node
π_i	Steady state probability of state i
$\pi_{i,j}$	Steady state probability of state (i, j)
Φ	Transition rate matrix of the aggregated voice traffic flows
ϕ_i	Transition rate from state i
ρ	Probability that a voice node has a non-empty buffer
ρ_i	Probability that a voice node i has a non-empty buffer
ρ_L	The ρ value of the left hand side of an equation
ρ_R	The ρ value of the right hand side of an equation
ρ^*	The ρ value at the maximum allowable packet dropping probability

ρ_i^*	The ρ^* value corresponding to voice call i
$\rho^*(c)$	The ρ^* value corresponding to voice class c
τ	Service time requirement of a service request
σ	Utilization factor
υ	Peak-to-mean ratio of the traffic flow form a voice source
υ_a	Peak-to-mean ratio of the aggregated traffic flows
$\varphi_{busy,t}$	Residual time of `type t` busy period
$\varphi_{idle,int}$	Residual time of an idle interruption
$\varphi_{idle,int,t}$	Residual time of `type t` idle interruption
$\varphi_{t'}$	Residual time of a `type t'` request
ϱ	Probability of transmission in at an available time-slot
$\varrho(c)$	The ϱ value corresponding to voice class c
ϱ_i	The ϱ value corresponding to call i
$\varrho_{\max}$	Maximum transmission probability of the slot-ALOHA scheme to satisfy the QoS requirement
$\varrho_{\min}$	Minimum transmission probability of the slot-ALOHA scheme to satisfy the QoS requirement
ς	The variance of the traffic flow from a voice source
ς_a	Variance of the aggregated traffic flows
ϑ	Autocovariance coefficient of the traffic flow from a voice source
ϑ_a	Autocovariance coefficient of the aggregated traffic flows
$\Xi(\cdot)$	Largest real eigen value
$\zeta_b(\cdot)$	Effective bandwidth function
$\zeta_c(\cdot)$	Effective capacity function
b_t	Fraction of time having a `type t` busy period

C	Set of voice classes
$\hat{C}_i$	Set of admissible classes at the neighborhood of the receiver node i
C_ω	Set of admissible classes for voice call ω
c_i	Class of voice call i
D	Queuing delay of a voice packet
$D_{\max}$	Delay bound of a voice packet
$D_{\max}(c)$	Delay bound of a class c voice call
D_{new}	Waiting time of the new queue head χ_{new}
D_{old}	Waiting time of the previous queue head χ_{old}
$f_L(\cdot)$	Probability density function of file length L
$f_\tau(\cdot)$	Probability density function of service time requirement τ
$F_\tau(\cdot)$	Cumulative density function of service time requirement τ
$G_{\imath_r}$	Set of neighboring source nodes of receiver $\imath_r$
$G_{\imath_s}$	Set of neighboring receiver nodes of source $\imath_s$
I	Duration of an interruption
I_j	Duration of the j^{th} interruption
L	Length of a data file
L_{pk}	Size of a data chunk transmitted in a time-slot
m	Mean traffic arrival rate from a traffic source
m_a	Mean traffic arrival rate from the aggregated traffic flows
$n_a(i)$	Number of packets transmitted when the queue head waiting time is i
$n_d(i)$	Number of packets dropped when the queue head waiting time is i
n_{pk}	Maximum number of voice packets transmitted during a time-slot
n_s	Number of inter-arrival times between two adjacent packet transmissions
N	Number of voice calls or data sessions in the system

N_b	Number of type t requests in the system at the initiation of type t_b busy period
N_c	Number of type t requests in the system at the initiation of type t_c busy period
N^*	Minimum number of users in the system required for the P_e evaluation with the round-robin scheme
N_f	Number of data files in an interactive data session
$N_{I,k}$	Number of interruptions arrived during time period T_k'
N_{j_r}	Number of neighboring source nodes of receiver j_r
$N_{\max}$	System capacity in terms of voice calls or data sessions
$N_{y,k}$	Number of type t arrivals during time period T_k
p_i	Steady state probability of the channel state i
$P_{busy,t}$	Probability of a target request arriving in a type t busy period
$P_{busy,t'}$	Probability of a target request arriving in a type t' busy period
$P_{i,j}$	Transition probability of the DTMC form state i to j
$P_{(i,j),(k,l)}$	Transition probability the DTMC form state (i,j) to (k,l)
$P_{idle,av}$	Proportion of time having an idle available period
$P_{idle,av,t}$	Proportion of time having a type t idle available period
$P_{idle,int}$	Proportion of time having an idle interrupted period
$P_{idle,int,t}$	Proportion of time having a type t idle interrupted period
$P_{I,t}$	Probability of an incoming request initiating a type t busy period
P_e	Delay bound violation probability
P_e^*	Minimum P_e that can be evaluated from analytical model for the round-robin scheme
P_{fa}	Probability of false alarm

P_{md}	Probability of missed detection
P^*_{md}	Missed detection threshold to limit the interference with the primary network
P_S	Probability of a node getting successful channel access in a particular time-slot
$P_{S,md}$	Probability of the node getting successful channel access after a missed detection occurs
$P_{S,i}$	Probability of the node getting successful channel access when the channel is at state i
$P^*_{S,1}$	The $P_{S,1}$ value when ρ_i=ρ^*_i
P^*_S	Minimum $P_{S,1}$ required to satisfy the delay requirement
$P^*_S(c)$	The P^*_S of voice class c
P_T	Probability of interactive data session termination
$Q_{i,j}$	Transition rate of an on-off voice source form state i to j
R_i	Rate of a voice source at State i
$S_{i,j}$	Transition probability of the channel from state i to state j
T_{busy,t_b}	Duration of type $\mathtt{t}_b$ busy period
T_{busy,t_c}	Duration of type $\mathtt{t}_c$ busy period
$T_{busy,t}$	Duration of type $\mathtt{t}$ busy period
T_I	Voice packet inter-arrival time
T_{off}	Mean duration of channel not available for the secondary users
T_{on}	Mean duration of channel available for the secondary network
T_k	Duration of k^{th} portion target time period
$T_{i,k}$	Duration of k^{th} portion of the service time of i^{th} division of the STR having Δt duration
$T_k{}'$	Summation of the service time requirements of the service requests

	composing the k^{th} time period
T_R	Response time of a data file in an elastic/interactive traffic flow
$\bar{T}_{R,max}$	Mean response time bound of an elastic/interactive data file
T_S	Time-slot duration
T_{t_b}	Duration of $\texttt{type t}_b$ busy period
T_{t_c}	Duration of $\texttt{type t}_c$ busy period
T_T	Thinking time of an interactive data session
W_t	Waiting time of a request with service time requirement t
$x_i(\cdot)$	Arrival process of the unit i under consideration
$X_S(n)$	Channel availability index at time-slot n
X_t	Service time of a request with service time requirement t
$X_{\Delta t,i}$	Service time of the i^{th} division of the STR having Δt duration
$X_{t'}$	Service time of a `type t'` service request
X_Z	Number of time-slots required to go from state Z to $Z-1$
$y_i(\cdot)$	Departure process of the unit i under consideration
Y_t	Service time requirement of a `type t` request
$Y_{t'}$	Service time requirement of a `type t'` request
$Y_{t,j}$	Service time requirement of j^{th} `type t` request arrival
Z	Number of nodes to get the channel access right before the target node
Z_1	A node enters state Z through this state
Z_2	A node enters this state if the channel becomes unavailable when in Z_1

Chapter 1

Introduction

The extensive growth of wireless networks over the recent years has increased the demand for the radio spectrum to a great extent. Static spectrum allocation regardless of its spatiotemporal usage has led to scarcity of the spectrum. As a result, fulfilling the spectrum requirements of emerging wireless applications and technologies is a challenging task. However, it is shown by various research groups [1][2] that the spectrum scarcity is mostly due to the inefficient spectrum management. Fig.1.1 shows the average spectrum usage in Chicago and New York areas. It can be clearly seen that the utilization is well below 25% in most of the spectrum bands. As a promising solution to this problem, the concept of cognitive radio networks (CRNs) has been proposed in late 1990's [3][4].

1.1 The Concept of Cognitive Radio Networks

The concept of cognitive radio networks has evolved as a promising technique to make use of underutilized portions of the spectrum by operating unlicensed (secondary) networks over the licensed (primary) spectrum bands. The basic functionality of a CRN is to access the un-

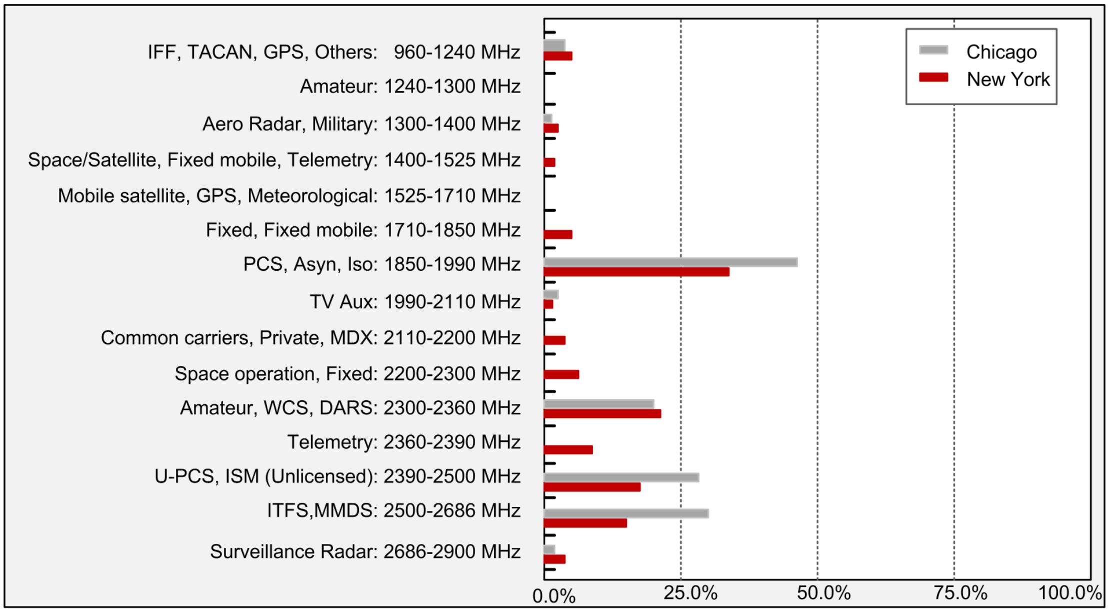


Figure 1.1: Average spectrum occupancy of Chicago and New York City regions [2].

used/underused portions of the licensed spectrum for the benefit of the unlicensed users, without causing harmful interference to the licensed users [3]-[5]. The licensed users are called primary users (PUs) and the unlicensed users are called secondary users (SUs). The concept of cognitive radio has been well accepted within the wireless communications research community, and it is predicted to be an important component of wireless communication networks.

The CRNs are categorized into two types, namely, the underlay and overlay networks. In the underlay CRNs, the SUs access a licensed channel simultaneously with the PUs, and in the overlay CRNs, the SUs access a licensed channel only when the PUs are not using the channel. Therefore, the transmit power and the channel availability are the main limitations of the underlay and overlay CRNs, respectively. As the spectrum availability and transmit power of the secondary networks exclusively depend on the behavior of the PUs, detecting/accessing the available spectrum bands and transmitting information packets without harmfully interfering the PUs are important. Further, with rapidly changing spectrum opportunities, estimating the available amount of resources and coordinating transmissions among the SUs are equally challenging. The initial research studies focused on the establishment of CRNs rather than the service satisfaction. Therefore, the early research studies on CRNs address only the best effort type of services without strict service requirements.

As the demand for Internet based interactive services and wireless multimedia services keeps on increasing, quality-of-service (QoS) support becomes essential for CRNs. Over the past few years, significant efforts have been made towards enabling the QoS support via efficient channel access [6]-[12] and performance analysis of QoS sensitive services [13]-[27] over CRNs. In the studies, channel access/power control policies, cross-layer designs, and capacity analysis for voice and video streaming are discussed. Providing service guarantees requires the knowledge of network resources, and call admission control (CAC) to maintain the service satisfaction levels.

The basic idea behind the CAC is to limit the number of traffic flows in the network in such a way that all the traffic flows can be served without violating their QoS requirements. Different from the capacity definition of the physical layer[1], the capacity is defined as the number of traffic flows that can be supported by the system, without violating their service quality requirements. As the foundation of CAC, capacity analysis is an important research area in the cognitive radio (CR) networking framework. The capacity of a CRN depends on the resource requirement of services as well as the resource availability of the network. The resource requirement of a particular service depends on its traffic model and its QoS requirements. Therefore, the capacity depends on the QoS requirements. Traffic flows of most QoS sensitive services such as voice, video, and interactive data services exhibit a time varying nature, which makes it difficult to determine their resource requirements. Furthermore, estimating the resource availability of CRNs becomes complicated with the random nature of the spectrum availabilities, which depends on the PU activities and the location of the SUs. The service disciplines of the base stations (or central controllers) in centralized networks and channel access

[1] The physical layer capacity only depends on the channel bandwidth, availability, and its condition, and it is given in terms of either bits/channel use or bits/second [28].

schemes in distributed networks can make a significant impact on the resource utilization (ratio of the exploited amount of resources from what is available) of the SUs. The randomness of the network resource availability, the time varying nature of resource requirements, and stringent QoS requirements of the wireless services make the capacity analysis and admission control challenging.

1.2 Motivation and Research Contributions

As two of the common QoS sensitive services, voice and (elastic/interactive) data services have gained an immense popularity. Both the services have service requirements in terms of delay. The voice service has a packet level service requirement in terms of end-to-end packet delay and elastic/interactive data services have a session level service requirement in terms of file transfer time (also known as the response time), respectively. These services are among the most popular services over the conventional wireless networks, and their QoS support is extensively studied over the past two decades. As CR networking is one of the next generation networking concepts, it is necessary to support these services over CRNs. In order to provide the required service levels using limited available resources in CRNs, call admission control is an essential task, and capacity analysis is the base for developing CAC algorithms.

The capacity analysis of voice services is comprehensively studied for networks such as cellular networks and IEEE 802.11 based wireless networks. There have been some studies on developing channel access schemes to support voice traffic [16][17] and to analyze the capacity of voice services [13]-[22] over CRNs. The system capacity depends on the amount of channel time that can be exploited by the channel access schemes, and it is important to have some benchmarks to compare the efficiency of the developed schemes for CRNs. The legacy channel access schemes such round-robin and slot-ALOHA schemes can be used to provide the benchmarks. Therefore, analyzing the voice capacity of CRNs with legacy channel schemes is useful to the research community. In the existing works, some important aspects such as on-off nature of the source traffic, transmitting multiple voice packets in a single transmission, and stochastic delay requirements have not been considered together for the capacity analysis. Further, there is a trade-off between the cost and the service quality. Supporting traffic flows with different delay requirements can provide SUs an option to choose the required service quality depending on the price they are willing to pay. This aspect has not been studied in the context of CRNs.

Elastic/interactive data traffic services such as web browsing are widely used over the conventional wireless networks [29][30]. These types of services fall into the non-real-time traffic category, and the mean response time is an important measure to provide a satisfactory service quality. In the conventional wireless networks, there are no significant service interruptions for the data services when compared with that of the CRNs. The service interruptions due to the presence of PUs increase the response time of SU elastic/interactive data traffic flows. The service disciplines used over the network can provide certain priorities to different users based on their requested file lengths. Only a handful of research efforts are devoted to the response time analysis of elastic/interactive data services over CRNs considering the channel availability statistics, service disciplines, and the file length distributions. Therefore, a comprehensive study about the elastic/interactive data traffic flows considering different service disciplines is expected to be a significant research contribution.

In Chapter 2, we study the voice capacity of centralized and distributed overlay CRNs. The capacity is given in terms of the number of simultaneous voice calls that can be supported by the system. Specifically speaking, we have analyzed the voice capacity of single-channel fully-connected CRNs considering the packet delay as the QoS measure. Under this, we study:

- The on-off and constant-rate voice capacity of a centralized CRN with first-come first-serve (FCFS) service discipline at the base station (BS) [31];
- The constant-rate voice capacity of a single-channel distributed fully-connected CRN

under different channel access schemes, and compared the performance of different schemes [32].

Considering the slot-ALOHA, round-robin, and random allocation schemes, we study the effect of channel rate (number of voice packets that can be transmitted in a time-slot) and the channel unavailable duration, respectively, on the system capacity. Then, we develop a CAC procedure for a distributed non-fully-connected CRN with slot-ALOHA network coordination, assuming homogeneous single-hop voice traffic flows based on the capacity analysis results of the fully-connected network. Further, we develop two CAC algorithms for the voice traffic flows with different delay requirements [33].

In Chapter 3, we study the mean response time of the elastic data service operating over a single-channel time-slotted centralized CRN with three service disciplines, namely, shortest processor time without preemption (SPTNP), shortest processor time with preemption (SPTWP), and shortest remaining processing time (SRPT), in comparison with the processor sharing (PS) service discipline [34][35]. We compare all four service disciplines under different data traffic load conditions and Weibull distributed data file lengths with different tail properties. Further, we show that the variation of mean channel unavailable duration has a significant impact on the mean response time, even when the probability of channel availability (long-term channel availability) remains unchanged. To the best of our knowledge, this is the first time that these service disciplines are considered with elastic data traffic in CRNs.

In Chapter 4, we study the mean response time of the interactive data traffic service in which each session consist of multiple data file requests (service requests). We show that the request arrivals at the BS can be approximated by a Poisson process under specific channel availability conditions. Based on the Poisson approximation, we show that the mean response time of interactive (multi-file) data sessions can be approximated by that of elastic (single-file) data traffic sessions with an equivalent session arrival rate.

1.3 Outline of the Thesis

The remainder of this thesis is organized as follows. The voice capacity analysis and the CAC algorithms are discussed in Chapter 2. Chapter 3 presents the mean response time analysis of elastic data traffic flows under different service disciplines, and Chapter 4 contains a discussion on the relationship between the mean response times of elastic (single-file) and interactive (multi-file) data traffic flows. The conclusions and future research directions are given in Chapter 5.

Chapter 2

Voice Capacity Analysis

2.1 Introduction

The voice and video streaming are the most common real-time wireless services in the modern day wireless networks [37]. The demand for video streaming is growing rapidly, and it is predicted to be the most common wireless service in the future [38]. As an upcoming networking approach, it is necessary to support the voice and video services over CRNs. As the first step, we study the voice service support over CRNs.

The cellular and satellite networks have dedicated capacity allocation to voice traffic flows, and dedicated channel time is available for each voice traffic flow. Therefore, the delay of a voice packet is a deterministic quantity, which depends on the propagation delay through the network. However, in the networks such as wireless local area networks (WLANs), voice traffic flows are sharing the network resources with other services such as video streaming and interactive data services. Therefore, fixed capacity allocation is not always feasible. Due to the random nature of the channel access, the packet delay is not deterministic. When the voice service is operating in such a network, the service quality requirement is given in terms of a stochastic delay requirement [39], such as $P(D > D_{\max}) \leq \epsilon$, where D is the end-to-end delay of a voice packet, $D_{\max}$ and ϵ are the delay bound (or the maximum delay) and maximum delay bound violation probability allowed, respectively. Further, the service requirement is relaxed in some studies by giving the QoS requirement in terms of mean of the packet delay. Voice traffic sources are categorized into constant-rate and on-off traffic sources [36] based on the output of the voice coder. The voice service support is extensively studied over the conventional wireless networks, and as an upcoming networking approach, it is an important research area in the context of CRNs.

In an overlay CRN, the channel availability for the secondary network exclusively depends on the behavior of the PUs. Therefore, the amount of spectrum resources available for the secondary network is limited. In order to provide the required service levels using limited available resources in CRNs, capacity analysis and call admission control are essential. The voice service support over CRNs has been studied in the literature with respect to the capacity analysis [16]-[19], call admission control [20][21], and developing channel access schemes [16][17][22]. The voice capacity is defined as the maximum number of voice calls that can be supported by the network satisfying their service requirements. In [18]-[22], the network under consideration is a centralized CRN, and the service quality is measured in terms of the packet dropping probability due to the buffer overflow at the base station [18][19] and the end-to-end delay of voice packets [20], respectively. The QoS requirement is given in terms of the mean end-to-end voice packet delay in [20]. However, in [16][17], a stochastic delay requirement is considered in the capacity analysis. The voice sources are treated as constant-rate traffic sources in [16][17][22], and on-off traffic sources in [18]-[21]. In [20], three different CAC strategies used for voice over IP (VoIP) traffic in wireless networks are mathematically analyzed with respect to centralized CRNs. The authors study the impact of the primary user information on the CAC algorithms in [21]. Different CAC algorithms based on the number of SUs in the system and the total number of users (both the PUs and SUs), respectively, are compared. The call blocking probability and the packet loss probability are considered as the QoS parameters. Developing channel access schemes for voice service support over distributed fully-connected CRNs and the capacities of the developed schemes are presented in [16][17]. In almost all of the works, the primary network under consideration is time-slotted, and the channel availability for the secondary users can be either independent or dependent in adjacent time-slots, respectively. Most existing studies assume perfect channel sensing by the secondary users, except in [19] which extends the capacity analysis to incorporate the effect of imperfect channel sensing. Different form the previous studies, the voice service support over TV bands is studied in [22]. The minimum number of channels required to be sensed by each SU to support a given traffic load is analyzed using an optimization technique. The channel sensing errors are also incorporated into the optimization problem. Further, deployment of long term evolution (LTE)-Advanced in the industrial, scientific, and medical (ISM) band to resolve some deployment issues is studied in [23]. The authors propose a statistical traffic control scheme to tackle critical challenges of the packet transmission coordination and the radio resources allocation in the network.

The voice capacity studies in the existing works are carried out for single voice packet transmission per time slot per frequency channel. Transmitting multiple voice packets as a composite packet over a time-slot (per channel) improves the network capacity without the requirement of a proportional increment in the channel transmission rate. However, transmitting multiple voice packets as a composite packet can increase the delay jitter, which is not desirable in the context of voice service. There is a trade-off between reducing the packet delay and delay jitter. Furthermore, the existing works are limited to the case that all the voice traffic flows require the same service quality. For example, in [16][17], all the voice traffic flows have a stochastic delay requirement with the same delay bound and maximum delay bound violation probability. In general, the larger the delay bound, the lower the probability of delay bound violation. Therefore, a larger number of users having a larger delay bound can be accommodated in the system, which gives the possibility to provide the service at a lower cost. However, the larger the delay bound, the lower the service quality. There is a trade-off between the service cost and the service quality. By incorporating different delay requirement parameters ($D_{\max}$ and ϵ), the system provides users a choice between service quality and cost, which improves the satisfaction of the users. In the literature, there are only limited works in developing channel access schemes to support voice [18][17] over CRNs, and much attention has not been payed to the generic channel access schemes in supporting voice traffic.

Motivation and Objectives

Treating the voice source as an on-off traffic source leads to higher system capacities. Further, stochastic delay guarantees are provided in shared wireless networks to provide higher service satisfaction. However, time varying nature of the on-off voice traffic flows and the random nature of the channel availability of the CRNs make the service satisfaction challenging. Therefore, on-off voice capacity has not been studied in the literature providing a stochastic delay guarantee. In this chapter, we study the on-off and constant-rate voice capacity of a centralized CRN with a stochastic delay requirement. A base station may not be available in a CRN due to the cost of installation. Therefore, the node coordination is difficult, and interruptions by the primary users makes it more challenging. In distributed CRNs, efficient channel access schemes are required for the network coordination. The system capacity depends on the efficiency of the channel access schemes, and benchmarks are important to compare the efficiency of the developed schemes. Readily available legacy channel access schemes such as round-robin and slot-ALOHA schemes can be used as the benchmarks, and studying the voice capacity of CRNs with legacy channel schemes is useful for the CR research community. In this chapter, we study the capability of legacy channel access schemes in supporting voice traffic over fully-connected CRNs. In non-fully-connected networks, the larger the number of neighbors associated with target user, the lower the channel time available per user, leading to lower service quality. Therefore, CAC plays an important role in keeping the number of neighbors per user to an acceptable level. In this chapter, we develop CAC algorithms for non-fully-connected slot-ALOHA based CRNs in supporting voice traffic.

Contributions

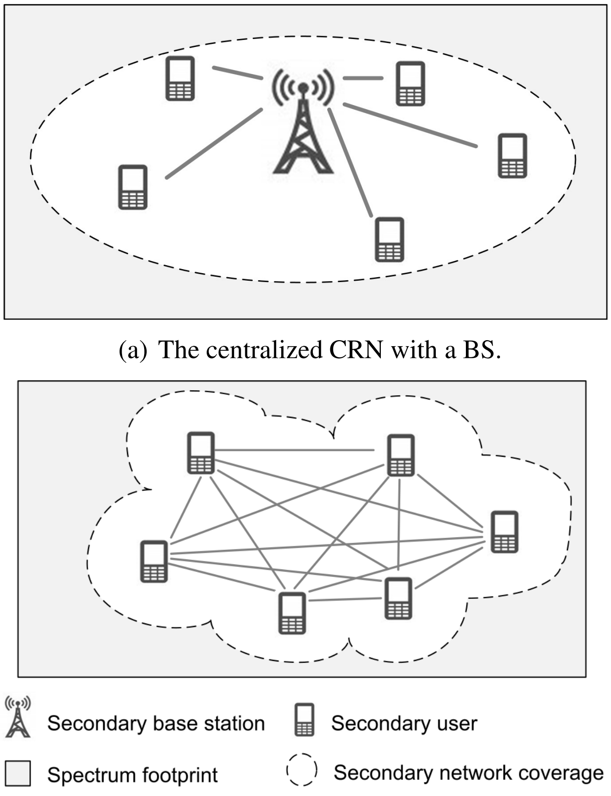
The contribution of this chapter is five fold: (i) We analyze the on-off and constant-rate voice capacity of a single-channel centralized CRN with FCFS service discipline; (ii) We analyze the constant-rate voice capacity of a single-channel distributed fully-connected CRN with slot-ALOHA channel access coordination. Different from the existing work, we consider the transmission of multiple voice packets in a single time-slot; (iii) We analyze the voice capacity of a single-channel distributed fully-connected CRN for round-robin channel access coordination. As the capacity analysis approach used for the slot-ALOHA scheme cannot be used for the round-robin scheme, a new approach is introduced. Further, possible extensions to the analytical models to incorporate sensing errors are discussed; (iv) We develop a CAC procedure for a distributed non-fully-connected CRN with slot-ALOHA network coordination, assuming homogeneous single-hop voice traffic flows. The capacity analysis results of the fully-connected network is used to limit the number of calls entering the system; (v) We develop two CAC algorithms for a distributed non-fully-connected slot-ALOHA CRN when the voice traffic flows have different delay requirements. For all the above studies, both dependent and independent channel occupancies of PUs in neighboring time-slots are considered, and the end-to-end delay of voice packets is considered as the QoS parameter. Note that (i) is presented in [31] and (ii)-(v) are presented in [32][33].

2.2 System Model

The system architecture, channel model, voice traffic model, and channel access schemes under consideration are described in this section.

2.2.1 System Architecture

In this study, we consider a centralized CRN with a base station and a distributed fully-connected CRN as illustrated in Figs.2.1(a) and Fig.2.1(b), respectively. The secondary network operates over a time-slotted single-channel primary network, and all the SUs see the same spectrum opportunities[1] (spectrum homogeneous). The secondary network is an overlay CRN in which the SUs access the channel (transmit or receive) only when the PUs are not present. Each SU is equipped with a single transceiver to sense the channel and transmit information packets.



(a) The centralized CRN with a BS.

(b) The distributed fully-connected CRN.

Figure 2.1: Centralized and distributed fully-connected CRNs.

[1] The primary user activities are consistent throughout the network. Therefore, the coverage area of the secondary network should be smaller than the one hop coverage area of the primary network.

2.2.2 Channel Availability Model

The channel time is partitioned into slots of constant duration T_S. The channel state of each time-slot is either idle (i.e., no primary activities) or busy (i.e., with primary activities). In a time-slot, the state is defined as 0 if the channel is busy, and is 1 otherwise. The state transition of the channel among adjacent time-slots can be given using a Markov chain as illustrated in Fig.2.2, where $S_{i,j}$ denotes the transition probability from state i ($\in \{0, 1\}$) to state j ($\in \{0, 1\}$). This is a widely used method to model the behavior of primary users [16]-[19][40] due to its simplicity. The channel state can be independent or dependent among adjacent time-slots.

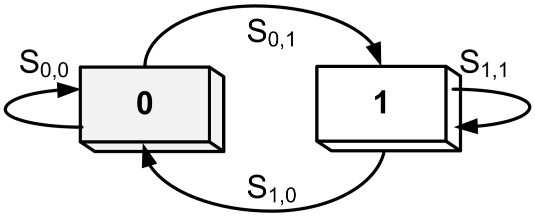


Figure 2.2: The channel state transition diagram.

In the independent case, $S_{0,1}$=$S_{1,1}$=p_1 and $S_{1,0}$=$S_{0,0}$=p_0=1-p_1. For both the independent and dependent cases $S_{0,1}$+$S_{0,0}$=1 and $S_{1,0}$+$S_{1,1}$=1. The channel state is identified by an SU by spectrum sensing[2]. For the simplicity of analysis, we assume that the final sensing decisions of the SUs and the BS are free of errors[3]. A time-slot is mainly divided into sensing and transmission phases, and in addition, distributed networks need a contention phase before the transmission phases. An SU transmits only when the channel is at State 1 (available for SUs), and the sensing and transmissions are free of errors.

[2] Different sensing techniques are listed in [41] and references there in.

[3] In reality, sensing errors are inevitable. Under the future extensions of this work, we discuss the effects of the sensing errors on the secondary network and the primary network, respectively.

2.2.3 Voice Traffic Model

All the SUs are voice nodes, and each voice call is associated with an SU and the BS or two SUs. Each of the two SUs (or an SU and the BS) initiates an independent voice traffic flow to the other. For the simplicity of our analysis, we consider only one traffic flow per voice call, and each call is limited to a single-hop voice flow. In the following, the terms voice call and voice traffic flow are used interchangeably to denote a one-way single-hop packet flow of a voice call, and the term node is used to denote an SU. Widely accepted voice traffic models, namely, the constant-rate voice model [6][42] and on-off voice model [43][44] are considered.

Constant-rate Voice Model

A voice node generates a constant-rate traffic flow with a packet inter-arrival time of T_I (normalized to the time-slot duration T_S), and the output of a voice codec is illustrated in Fig.2.3.

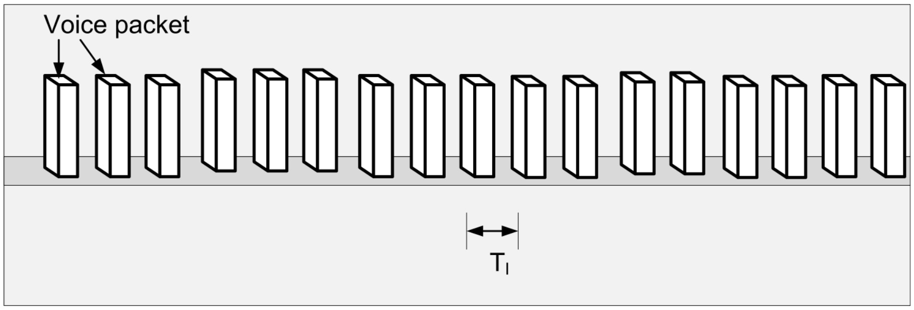


Figure 2.3: The codec output of a constant-rate voice source.

On-off Voice Model

The on-off model is a common voice traffic model used in the VoIP applications [18]-[21]. The state transition diagram of an on-off voice node is illustrated in Fig.2.4, where $Q_{i,j}$ denotes the

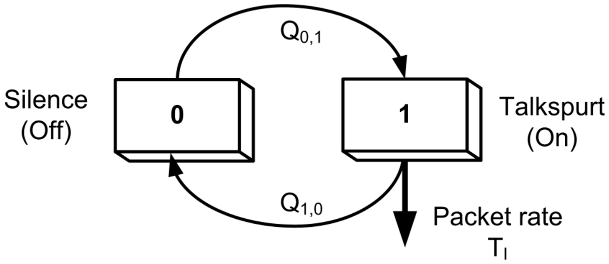


Figure 2.4: The state transition of voice traffic source [43].

transition rate from state i ($\in \{0, 1\}$) to state j ($\in \{0, 1\}$). When a voice node is at the off state, it does not generate any packet; when at the on state (talk spurt), it generates voice packets at a rate of $1/T_I$. An output of a voice codec is illustrated in Fig.2.5. When an on-off voice node is at the off state with an empty buffer, the voice node is at its inactive state, and when it is in the on state or off state with a non-empty buffer, it is in its active state.

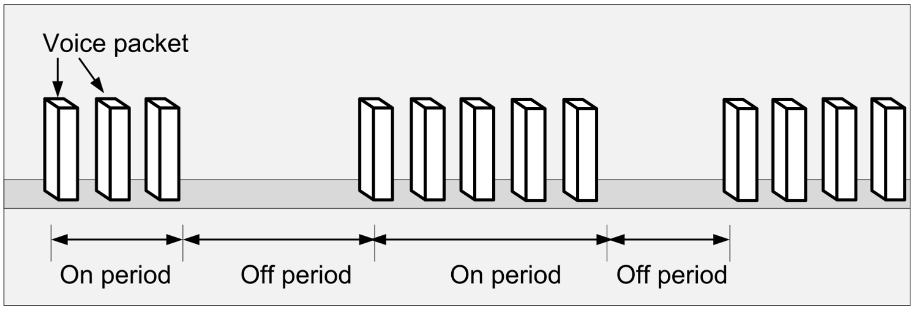


Figure 2.5: The codec output of an on-off voice source.

Voice Buffer Management

Each voice source buffers the voice packets until it gets an spectrum (transmission) opportunity. The service requirement for the voice traffic flows in the secondary network is characterized by the end-to-end delay of a voice packet (i.e., from the time that a packet is generated at the source node to the time that it is received at the receiver node). As the packet propagation delay is negligible when compared with the time that a packet spend at the source buffer, the service requirement is given in terms of the queuing delay D (normalized to T_S), from the time that a packet is generated at the source node to the time that it is transmitted from the source node. The stochastic delay requirement is given by [17][39][45]

$$P(D > D_{\max}) \leq \epsilon \tag{2.1}$$

where $D_{\max}$ (normalized to T_S) and ϵ are the delay bound and maximum delay bound violation probability allowed, respectively, in order to provide satisfactory voice quality. If the delay bound of a voice packet is violated, the packet is dropped without being transmitted. Without loss of generality, we assume integer values for T_I , D, and $D_{\max}$.

Voice Capacity

The voice capacity is defined as the maximum number of simultaneous voice calls that can be supported by the system, without violating the delay requirement given in (2.1) for all the admitted calls. In the analysis, it is important to note that the number of voice calls refers to the number of one-way voice calls.

Service Disciplines for Voice Traffic

For the centralized network, we consider an ideal scenario with the BS having the queue head waiting times of all the nodes.

- **FCFS service discipline**: The BS schedules packet transmission of the nodes in the available time slots based on the maximum queue head waiting time first principle.

For the distributed (fully-connected) CRN, two legacy channel access schemes, namely, the slot-ALOHA scheme and the round-robin scheme are considered, and the random allocation scheme is used as a benchmark for the performance comparison.

- **Round-robin scheme**: Each node will wait for its channel access right. When a particular node receives the channel access right, it transmits if it has packets in the buffer, or forwards the opportunity to the next node otherwise. Due to the cyclic nature of getting the channel access right, each node accesses the channel in a fair manner. As an approach of realizing the round-robin channel access coordination, a token based scheme [46] or a mini-slot based scheme [17] can be used. There are no packet collisions in the round-robin scheme as a node transmits only when it has the channel access right.

- **Slot-ALOHA scheme**: All the nodes with a non-empty buffer will transmit with a probability ϱ during an idle time-slot. If a collision occurs, each node will re-transmit at the next available time-slot with the same probability.

- **Random allocation**: One node act as a controller and assigns the channel access right to the other nodes randomly.

As the first step, we analyze the capacity of a centralized CRN with ideal information availability, and we extend the study to a distributed CRN with less information for the transmission decision. For simplicity of the analysis, we will only consider networks with stationary nodes.

2.3 Voice Capacity of the Centralized CRN

We analyze the voice capacity of a centralized CRN with a BS to schedule the channel access of each user, as illustrated in Fig.2.1(a). Only single-hop voice communication occurs between voice nodes, and all the voice flows have the same delay requirement. The arrival process and the service process of the centralized system are illustrated in Fig.2.6, where the outputs of all the voice codecs are either on-off or constant-rate.

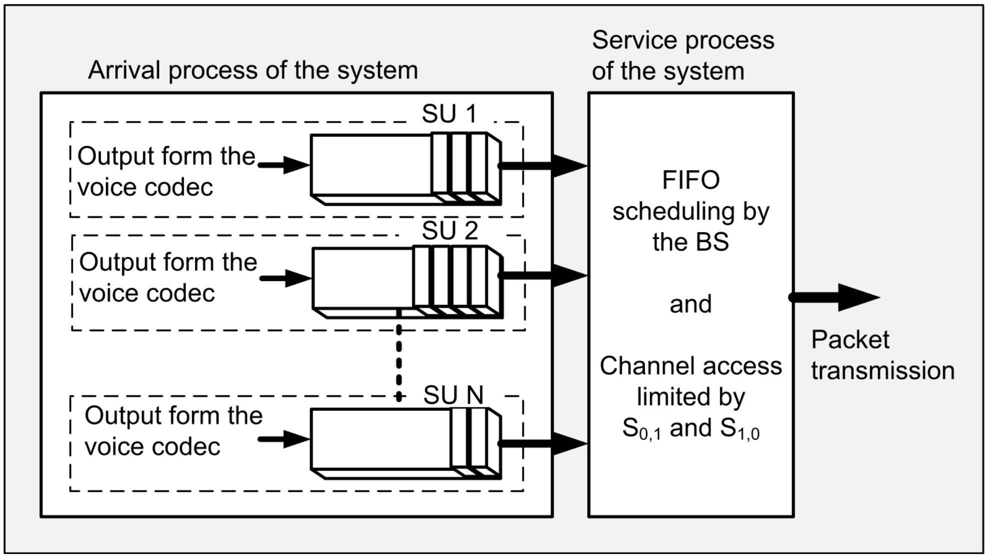

Arrival process of the system
Service process of the system
SU 1
Output form the voice codec
SU 2
Output form the voice codec
SU N
Output form the voice codec
FIFO scheduling by the BS
and
Channel access limited by $S_{0,1}$ and $S_{1,0}$
Packet transmission

Figure 2.6: The arrival process and the service process of the centralized system.

2.3.1 Service Process Analysis

Since the channel is time-slotted as discussed in Subsection 2.2.2, its service process is a discrete-time process. The service process, $\mu_S(n)$, is defined as the number of bits that can be transmitted in the time-slot n, and is given by $\mu_S(n)=X_S(n)\cdot n_{pk}$, where $X_S(n)$ ($\in \{0,1\}$) is the channel availability index of the time-slot, and n_{pk} is the maximum number of packets that can be transmitted in a time-slot. Service process analysis with respect to the QoS requirement can be carried out using the theory of effective capacity (EC), as discussed in [47]-[52]. The EC provides the constant arrival rate that can be supported by the system (service process), without violating the required service quality. The EC analysis of a block fading channel [51][54] can be adopted to analyze our channel by modeling the channel as a single block fading channel with two fading amplitudes (0 and 1). Therefore, the EC of the secondary network with dependent channel availability in adjacent time-slots is given by

$$\zeta_c(\theta) = -\frac{1}{\theta}\ln\left[\frac{S_{0,0}+S_{1,1}e^{-\theta n_{pk}}}{2} + \sqrt{\left(\frac{S_{0,0}-S_{1,1}e^{-\theta n_{pk}}}{2}\right)^2 + S_{0,1}S_{1,0}e^{-\theta n_{pk}}}\right] \qquad (2.2)$$

where, θ depends on the QoS requirement, and it is shown in [48] that $P(D \geq D_{\max}) \approx e^{-\theta D_{\max}}$. In order to satisfy the condition (2.1), the parameter θ should satisfy the condition $\theta \geq \frac{1}{D_{\max}} \ln\left(\frac{1}{\epsilon}\right)$. In order to support a constant arrival rate r with the given delay requirement (2.1), $\delta^*=r\zeta_c^{-1}(r)$ should be satisfy the condition

$$\delta^* \geq \frac{1}{D_{\max}} \ln\left(\frac{1}{\epsilon}\right). \tag{2.3}$$

The EC of the secondary network with independent channel availability scenario can be obtained by setting $S_{0,1}=S_{1,1}=p_1$ and $S_{1,0}=S_{0,0}=p_0$=1-p_1.

2.3.2 Arrival Process Analysis

The capacity requirement of the constant-rate voice traffic sources remains $1/T_I$ packets/time-slot throughout the duration of the call. However, the capacity requirement of an on-off voice source varies with time. An on-off traffic source can be characterized by the mean, m, variance, ϑ, auto covariance time coefficient, ς, and peak-to-mean ratio, υ, of the traffic flow [55]. The four parameters are given by $m=Q_{0,1}/T_I(Q_{0,1}+Q_{1,0})$, $\vartheta=m(\frac{1}{T_I}-m)$, $\varsigma=1/(Q_{0,1}+Q_{1,0})$, and $\upsilon=(Q_{0,1}+Q_{1,0})/Q_{0,1}$. The four parameters corresponding to the aggregate traffic from N independent sources are given by $m_a=N{\cdot}m$, $\vartheta_a = N{\cdot}\vartheta$, $\varsigma_a = \varsigma$, and $\upsilon_a = \upsilon$, respectively. The aggregate traffic flow can be characterized by a two-state Markov modulated Poisson process (MMPP). The MMPP can be characterized by four parameters R_1, R_2, ϕ_1, and ϕ_2, where R_i $(\in \{1,2\})$ is the mean rate of the Poisson process in state i, and ϕ_i $(\in \{1,2\})$ is the transition rate from state i. The four parameters are given by, $R_1=m_a+\sqrt{\upsilon_a\vartheta_a}$, $R_2=m_a-\sqrt{\vartheta_a/\upsilon_a}$, $\phi_1=\upsilon_a/\varsigma_a(1+\upsilon_a)$, and $\phi_2=1/\varsigma_a(1+\upsilon_a)$ [55]. The capacity requirement of a time varying arrival process considering its service requirements is carried out using the theory of effective bandwidth (EB), as explained in [56][57]. The EB of an arrival process is the required constant service rate in order to satisfy the service quality requirement of the arrival process. The effective bandwidth of the two-state MMPP is given by [55]

$$\zeta_b(\theta) = \frac{\Xi(\Phi + (e^{\theta} - 1)\acute{R})}{\theta} \tag{2.4}$$

where Φ is the transition rate matrix of the two-state Markov chain of the aggregated traffic flow, $\acute{R}$=diag(R_1, R_2), and $\Xi(\cdot)$ gives the largest real eigen value. In order to satisfy the given delay requirement (2.1) using a constant service rate u, $\delta^*=u\zeta_b^{-1}(u)$ should satisfy the condition (2.3) [47]. When both the arrival and service processes are time varying, in order to satisfy the delay requirement (2.1), $\delta^*=\theta^*\zeta_c(\theta^*)$ should be satisfy the condition (2.3), where θ^* is the solution to the equation $\zeta_b(\theta)=\zeta_c(\theta)$ [48]. Note that, in the case of N constant-rate voice traffic sources, $\zeta_b(\delta)=N/T_I$. In order to determine the maximum number of voice sessions that can be supported by the system while satisfying the stochastic delay requirement (2.1), we have to find maximum N which satisfies (2.3).

2.4 Voice Capacity of the Distributed CRN

We analyze the constant-rate voice traffic capacity of a distributed fully-connected CRN with round-robin and slot-ALOHA channel access schemes, and compare it with the random allocation scheme. In all three cases, even though the voice buffer of each node acts in the FCFS manner, the system with all the voice nodes as a whole does not behave in a FCFS manner. Therefore, the theory of EB and EC cannot be directly applied to the system, but, to each node. However, due to the complexity of analyzing the service process (of each node), we resort to packet level analysis of the voice buffer of each node.

2.4.1 Slot-ALOHA Scheme

With the initiation of a voice traffic flow, the first packet enters the source buffer becomes the queue-head, and the rest of the packets are buffered behind the queue-head. Whenever the queue-head is successfully transmitted, the next packet with the highest waiting time becomes the new queue-head, χ_{new},. While awaiting for transmission, the waiting time of the queue-head increases with time. However, when a successful transmission occurs, the waiting time of χ_{new}

is always lower than that of the queue-head, χ_{old}, which is just being transmitted. The waiting time, D_{new}, of the new queue-head (normalized to T_S), is given by

$$D_{new} = D_{old} - n_s{\cdot}T_I + 1 \tag{2.5}$$

where D_{old} (normalized to T_S) is the waiting time of χ_{old}. The term $n_s{\cdot}T_I$ is due to the n_s inter-arrival times between the arrivals of χ_{old} and χ_{new}, and the constant 1 accounts for the time-slot taken for the transmission of χ_{old}. As the voice packets whose waiting time exceeds the delay bound are dropped, the waiting time of a queue-head stays between 0 and $D_{\max}$. When a packet (queue-head) is dropped due to violation of the delay bound (i.e., $D > D_{\max}$), the waiting time of χ_{new} is given by D_{new}=($D_{\max}$ + 1) − T_I. The queue-head is dropped at the beginning of the time-slot when $D_{old} = D_{\max} + 1$. The term T_I is due to the inter-arrival time between the χ_{old} and the χ_{new}. In each idle time-slot, a target node with a non-empty buffer transmits with probability ϱ, and a successful transmission occurs if all the other non-target nodes in the network do not transmit. The probability of successful transmission (same as the probability of successful channel access), $P_{S,1}$, in an available time-slot is given by

$$P_{S,1} = \varrho\,(1 - \rho{\cdot}\varrho)^{N-1} \tag{2.6}$$

where ρ is the probability of a node having a non-empty buffer. The product $\varrho{\cdot}\rho$ is the probability of a node transmitting in an idle time-slot. Note that, the probability $P_{S,1}$ does not depend on D. The value of D at the next time-slot depends on the value of D, the state of the channel, and the success or failure of the transmission in the current time-slot. Furthermore, the state of the channel in the next time-slot either does not depend on that of the current time-slot for the independent channel availability scenario, or only depends on the state of the channel in the current time-slot for the two-state channel in Fig.2.2. Therefore, we can establish a discrete-time Markov chain (DTMC) in which the state (i, j) represents the waiting time of the queue-head and the channel state, respectively, as shown in Fig.2.7. Since there is no queue-head when the buffer is empty, the negative value of the time remaining until the next packet arrival is considered as the queue-head waiting time. Therefore, D varies from $-(T_I - 1)$ to $D_{\max}$. Theoretical aspects of this approach is discussed in [58]. Furthermore, the DTMC model is similar to the approach given in [17], in analyzing the constant-rate voice capacity of two different cognitive radio MAC protocols. Different from [17], here we consider the transmission of possible multiple (up to n_{pk}) voice packets by a node in a time-slot. The state transition probabilities of the Markov chain are given by

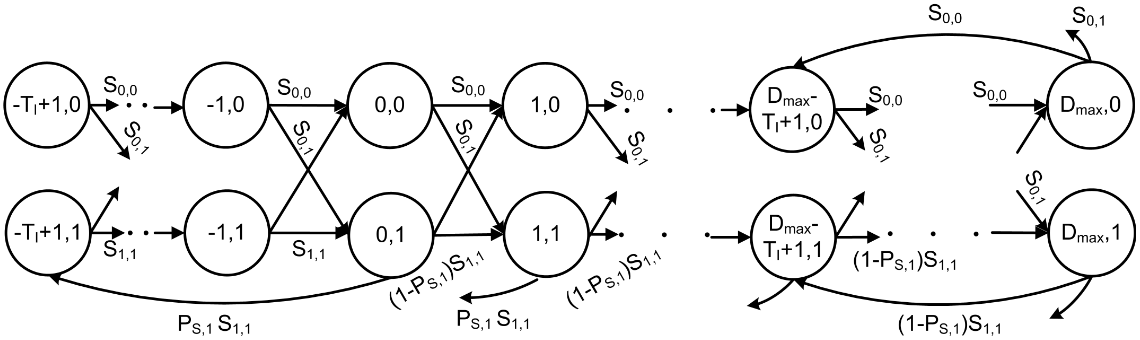


Figure 2.7: The Markov chain for the queue-head waiting time and channel state pair.

$$
\begin{array}{ll}
P_{(k,i),(k+1,j)} = S_{i,j}, & k \in \{-T_I + 1, ..., -1\} \\
P_{(k,i),(k+1,j)} = (1 - P_{S,i}) \cdot S_{i,j}, & k \in \{0, ..., D_{\max} - 1\} \\
P_{(k,i),(k-T_I+1,j)} = (1 - P_{S,i}) \cdot S_{i,j}, & k = D_{\max} \\
P_{(k,i),((k \ mod \ T_I)-T_I+1,j)} = P_{S,i} \cdot S_{i,j}, & k \in \{0, ..., (n_{pk} \cdot T_I - 1)\} \\
P_{(k,i),(k-n_{pk} \cdot T_I+1,j)} = P_{S,i} \cdot S_{i,j}, & k \in \{n_{pk} \cdot T_I, ..., D_{\max}\}
\end{array}
$$

where $P_{(k,i),(l,j)}$ denotes the transition probability from state (k, i) to state (l, j) and $i, j \in \{0, 1\}$. Since the channel is not available for the SUs when it is at state 0, $P_{S,0}$=0. As the packets whose waiting time is larger than the delay bound are dropped, the delay bound violation probability, P_e, is equal to the packet dropping probability, given by

$$
P_e = \frac{\sum_{j=0}^{1} \left(1 - P_{S,j}\right) \cdot \pi_{D_{\max},j}}{P_{S,1} \cdot \sum_{i=0}^{D_{\max}} n_a(i) \cdot \pi_{i,1} + \sum_{j=0}^{1} \left(1 - P_{S,j}\right) \cdot \pi_{D_{\max},j}} \tag{2.7}
$$

where $\pi_{i,j}$ is the steady state probability of state (i, j) and $n_a(i)$ is the number of packets that can be transmitted when the queue-head waiting time is i, given by

$$n_a(i) = \begin{cases} \lfloor \frac{i}{T_I} \rfloor + 1, & \lfloor \frac{i}{T_I} \rfloor + 1 < n_{pk} \\ n_{pk}, & \text{otherwise.} \end{cases}$$

The summation $\sum_{j=0}^{1} \left(1 - P_{S,j}\right) \cdot \pi_{D_{\max},j}$ represents the mean number of dropped packets and $P_{S,1} \cdot \sum_{i=0}^{D_{\max}} n_a(i) \cdot \pi_{i,1}$ represents the mean number of transmitted packets at the steady state, in a time slot. The capacity analysis problem can be represented as to maximize N with the constraint $P_e \leq \epsilon$. However, the relationship between the probability P_e and N is not straightforward. Therefore, we resort to numerical analysis in calculating the capacity.

We can find the probability $P_{S,1}$ for a given ρ and N by (2.6). Using $P_{S,1}$, the steady state probabilities of the Markov chain can be computed, and thereby the probability of buffer occupancy ρ is given by $\rho = \sum_{j=0}^{1} \sum_{i=0}^{D_{\max}} \pi_{i,j}$. Since probabilities $\pi_{i,j}$ ($i \in \{0, 1, ..., D_{\max}\}$ and $j \in \{0, 1\}$) can be represented in terms of ρ, the right hand side (RHS) of the equation also contains ρ. Denote the ρ in RHS as ρ_R and that in the left hand side (LHS) as ρ_L. The value of ρ_L can be computed for different values of ρ_R, and the solution for the equation is the one when $\rho_L = \rho_R$. Then, the probability of delay bound violation P_e can be obtained for a given N. Therefore, the maximum N which satisfies $P_e \leq \epsilon$ can be evaluated. The capacity analysis for the independent channel occupancy scenario can be carried out using the preceding method by substituting appropriate values for $S_{i,j}$ ($i, j \in \{0, 1\}$).

2.4.2 Random-assignment Scheme

As the assignment is random, the probability of successful transmission in an available time-slot is given by $P_{S,1} = 1/N$, and is independent in adjacent available time-slots. Therefore, the same approach used with the slot-ALOHA scheme can be used to analyze the probability of delay bound violation and the voice capacity.

2.4.3 Round-robin Scheme

The round-robin scheme guarantees that each node gets a packet transmission opportunity in an orderly manner. Whenever the node under consideration (target node) transmits, its next packet transmission does not occur before each non-target node with a non-empty buffer gets an opportunity to transmit. From (2.5), it can be seen that the queue-head waiting time of the target node drops just after a successful transmission. The probability of the target node getting the next transmission opportunity depends on the number of non-target nodes in the network having packets to transmit, the channel availability, and the time elapsed from its previous transmission. Therefore, with the round-robin scheme, the probability of a node getting a packet transmission opportunity is not the same for all D values, and the analysis for the probability of getting a transmission opportunity at the particular D value is not straightforward. Therefore, the Markov chain approach used for the slot-ALOHA scheme cannot be applied for the capacity analysis of the round-robin scheme.

Assuming that the packets of a target node are not dropped until it gets a channel access right (i.e., the packets with the waiting time larger than $D_{\max}$ will be dropped at the time the target node gets the channel access right), the range of D is $[0, \infty)$. When the target node gets a channel access right, it will drop $n_d(D)$ and transmit $n_a(D)$ voice packets, where

$$n_d(D) = \begin{cases} 0, & D \le D_{\max} \\ \lfloor \frac{D-D_{\max}-1}{T_I} \rfloor + 1, & \text{otherwise} \end{cases}$$

and

$$n_a(D) = \begin{cases} \lfloor \frac{D-n_d(D)T_I}{T_I} \rfloor + 1, & \lfloor \frac{D-n_d(D)T_I}{T_I} \rfloor + 1 \le n_{pk} \\ n_{pk}, & \text{otherwise.} \end{cases}$$

After transmitting the $n_a(D)$ packets, the D of the queue-head decreases by $(T_I \cdot n_a(D) - 1)$ time-

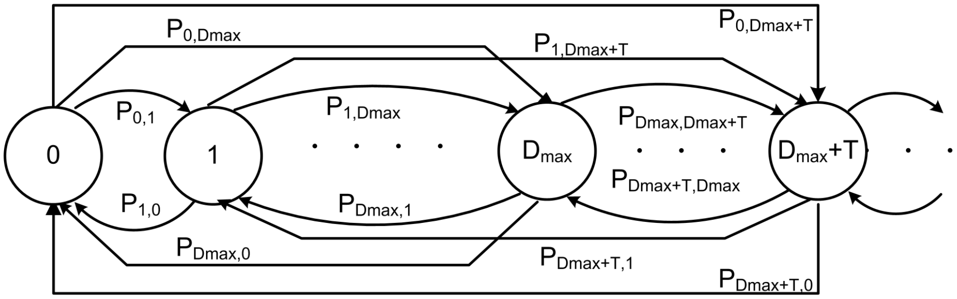


Figure 2.8: The DTMC for the queue-head delay at the time of packet transmission with round-robin channel access.

slots. Then, it increases by a random number of time-slots until the next channel access. With N voice calls in the system, for a target node, the waiting time of the queue-head at the time of packet transmission depends on the waiting time of the queue-head at the previous packet transmission and the number of time-slots required to provide a transmission opportunity to each of the $N - 1$ non-target nodes. If the number of time-slots in the shortest possible round-robin cycle is larger than or equal to the number of time-slots between two successive packet arrivals, the target source buffer will always be non-empty when it receives a transmission opportunity. As the shortest possible round-robin cycle is equal to the number of nodes in the network, N, the condition to have a non-empty buffer when a source node receives a transmission opportunity can be expressed as $N \geq T_I$. Therefore, the randomness will only be due to the channel availability, not due to the number of nodes with a non-empty buffer.

As the waiting time D at the next packet transmission depends only on that of the current packet transmission, but not on the previous packet transmissions, a DTMC can be developed with the state representing the queue-head waiting time at the time of packet transmission. With the waiting time D in $[0, \infty)$, the state space of the DTMC lies in the same range, making it an infinite-state DTMC. The Markov chain is illustrated in Fig.2.8, where $P_{i,j}$ is the transition probability from state i to state j $(i,\ j \in \{0, 1, 2, ...\})$. For a single-channel CRN with $N \geq T_I$, the state transition probabilities, $P_{i,j}$, of a target node is given by $P_{i,j}=P\left(\sum_{z=0}^{N-1} X_z = r\right)$, if $r \geq N$, and 0, otherwise, where Z is the number of nodes to access the channel before the target node gets

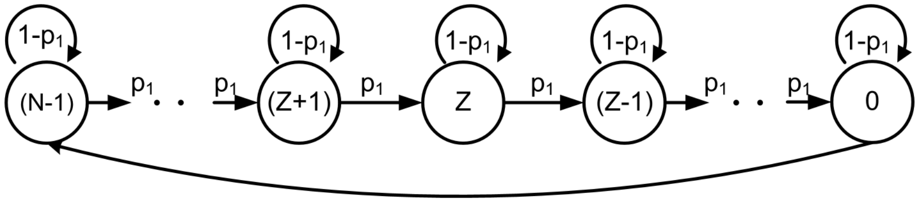


Figure 2.9: State transition of Z for independent channel occupancy of PUs.

the channel access right, X_Z is the number of time-slots required to reduce the node number from Z to $Z-1$[4], and $r{=}j-(i-(n_d(i)+n_a(i))\,T_I)$ is the elapsed number of time-slots between adjacent channel access opportunities. The number of time-slots X_Z ($Z \in \{0, 1, ..N-1\}$) are independent and identically distributed. When the channel availability for SUs in adjacent time-slots is independent, the state (the number Z) transition for a node is illustrated in Fig.2.9. When there are N source nodes in the system and they all have packets to transmit, it is impossible for a target node to have its next transmission opportunity within $N-1$ adjacent time-slots from its current transmission. Therefore, $P_{i,j}=0$ for $r<N$. In order to have $r-1$ time-slots ($r \geq N$) between two successive transmission opportunities, the target node should transmit at the r^{th} time-slot, and the rest of the $N-1$ non-target nodes should transmit during the first $r-1$ time-slots. In other words, exactly N out of the r time-slots should be idle and, out of the N idle time-slots, $N-1$ should be in the first $r-1$ time-slots. Therefore, the probability $P_{i,j}$ is given by the negative binomial distribution. The state transition probability, $P_{i,j}$, for an independent channel occupancy scenario of PUs is given by

$$P_{i,j}=\begin{cases}\binom{r-1}{r-N}p_1^N\,(1-p_1)^{r-N}\,, & \text{if } r{\geq}N\\ 0, & \text{otherwise.}\end{cases} \tag{2.8}$$

When the channel availability for SUs are dependent among adjacent time-slots, the state Z is divided into two states named Z_1 and Z_2, where a node enters state Z through state Z_1 (initial state), and enters state Z_2 if the channel is not available when it is in state Z_1. The state transition

[4] A non-target node with channel access right requires X_Z time-slots to obtain a channel opportunity and transmit its packets.

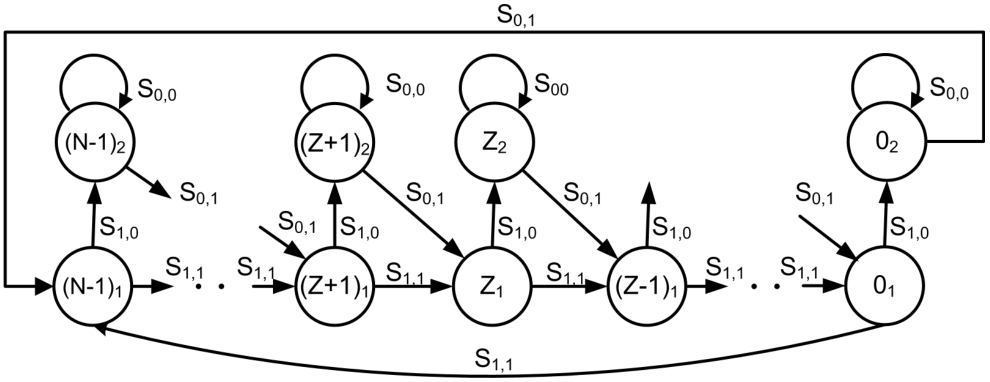


Figure 2.10: State transition of Z for dependent channel occupancy of PUs.

diagram of a node is illustrated in Fig.2.10. As explained earlier, $P_{i,j} = 0$ for $r < N$. If there are exactly N time-slots in between successive transmissions of the target node, all N time-slots should be available for the SUs[5]. Having $r > N$ time-slots between successive transmissions means that the channel has been idle for N time-slots and busy for $r - N$ time-slots. The state transition probabilities, $P_{i,j}$, for a dependent channel occupancy scenario of PUs is given by

$$P_{i,j} = \begin{cases} \sum_{l=1}^{\min(r-N,N)} \binom{N}{N-l} S_{1,1}^{N-l} S_{1,0}^{l} \cdot \binom{r-N-1}{l-1} S_{0,1}^{l} S_{0,0}^{r-N-l}, & \text{if } r > N \\ S_{1,1}^{N}, & \text{if } r = N \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

In (2.9), when $r > N$, there must be at least one transition from state 1 to state 0. The term $\binom{N}{N-l} S_{1,1}^{N-l} S_{1,0}^{l}$ represents the probability of having l state 1 to state 0 transitions out of all the transitions occur in the N idle time-slots. In order to have N idle time-slots, l state 0 to state 1 transitions are required in the remaining $r - N$ time-slots. The term $\binom{r-N-1}{l-1} S_{0,1}^{l} S_{0,0}^{r-N-l}$ represents the probability of having l state 0 to state 1 transitions in exactly $r - N$ time-slots. Since the DTMC has a countably infinite number of states, it is truncated to $D_{\max} + k{\cdot}T_I$ states for simplicity of analysis, where k (≥ 1) is a small integer. The delay bound violation probability,

[5] Being in state 1 (idle state), the channel should remain in state 1 for N successive time-slots.

P_e, is approximately given by

$$P_e \simeq \frac{\sum_{i=0}^{D_{\max}+k\cdot T_I} n_d(i)\cdot\pi_i}{\sum_{i=0}^{D_{\max}+k\cdot T_I} (n_d(i)+n_a(i))\,\pi_i} \tag{2.10}$$

where π_i is the steady state probability of state i. The terms $\sum_{i=0}^{D_{\max}+k\cdot T_I} n_d(i)\cdot\pi_i$ and $\sum_{i=0}^{D_{\max}+k\cdot T_I}$ $n_a(i)\cdot\pi_i$ represent the mean number of dropped packets and transmitted packets, respectively, at the steady state in a time-slot. The system capacity $N_{\max}$ is the maximum N which satisfies the relation $P_e \leq \epsilon$. The larger the N, the larger the P_e. The minimum P_e, P_e^*, that can be analyzed by (2.10) is for the minimum N, N^*. As $N \geq T_I$, $N^* = T_I$. Thus, the capacity can be evaluated for an ϵ value larger than P_e^*.

Capacity analysis of a fully-connected network is the first step of developing a call admission control algorithm. As we evaluate the maximum number $N_{\max}$ of simultaneous voice traffic flows that can be supported by the system without violating the delay requirement, the call admission control can be carried out by limiting the number of traffic flows in the network to $N_{\max}$.

2.5 Call Admission Control

When the slot-ALOHA scheme is used for the channel access control, collisions occur due to simultaneous transmissions of a target source node and the neighboring source nodes associated with the target receiver node. The larger the number of neighboring source nodes associated with a target receiver, the higher the chances of collisions, which leads to a lower successful transmission probability, $P_{S,1}$, of the target source node (or traffic flow). The lower the probability $P_{S,1}$, the longer the waiting time of packets in the buffer and the probability P_e of delay bound violation. Therefore, in order to keep the probability P_e within a desired limit, the number of calls admitted to the system should be controlled.

2.5.1 CAC Procedure for Homogeneous Voice Traffic

In Subsection 2.4.1, we analyze the maximum number, $N_{\max}$, of homogeneous voice traffic flows that can be carried out by a slot-ALOHA fully-connected network. Therefore, $N_{\max}$ is the maximum number of homogeneous voice source nodes that can be associated with a target receiver node. In a non-fully-connected network, each receiver node is associated with a number of source nodes. The packet transmission of a new source node increases the collisions at its associated receiver nodes, leading to a reduction in the successful transmission probability of the said receiver nodes. Therefore, to satisfy the delay requirement of the ongoing and incoming traffic flows, it is required to control the admission of new calls based on the number of source nodes associated with each receiver node (including that of the incoming call). A CAC procedure, $P1$, based on the number of neighboring nodes can be explained as follows. Denote the source and receiver nodes of the new call by target source (ω_s) and receiver (ω_r) nodes, respectively, and the set of neighboring receiver nodes of ω_s and source nodes of ω_r by $\in G_{\omega_s}$ and G_{ω_r}, respectively. Let N_{i_r} be the number of neighboring source nodes of receiver node i_r ($\in \{G_{\omega_s} \cup \omega_r\}$). It is required to limit N_{i_r} of each receiver node i_r ($\in \{G_{\omega_s} \cup \omega_r\}$) to a maximum of $N_{\max}$. Therefore, ω_s should listen to its neighbors i_r ($\in G_{\omega_s}$) and get the information N_{i_r}. At the same time, ω_r should listen to its neighbors and find N_{ω_r}. If the condition $N_{i_r} \leq N_{\max}$ can be satisfied for all i_r ($\in \{G_{\omega_s} \cup \omega_r\}$), the new call is admitted to the system, and rejected otherwise. As $N_{\max}$ is a function of ϱ, the non-fully-connected network must use the same (ϱ, $N_{\max}$) pair which used with the fully-connected network.

The capacity of a fully-connected network is under the assumption of homogeneous voice traffic. However, the capacity analysis of the fully-connected network is no longer valid for non-homogeneous voice traffic. The validity of the $N_{\max}$ used in this procedure no longer holds, and a new approach is required for the CAC of non-homogeneous voice traffic over non-fully-connected CRNs.

2.5.2 CAC Procedure for Non-homogeneous Voice Traffic

Majority of the existing CAC strategies developed for non-cognitive ad hoc networks consider only the first order statistics such as average waiting time, and are based on standard queuing analysis by using the Little's theorem. Further, there are some existing works on CAC in non-cognitive networks based on stochastic QoS guarantees using the theory of effective bandwidth and its dual effective capacity [39][45][59]. All of these works are for homogeneous/non-homogeneous traffic flows with the same delay requirement. Based on this idea, we can develop a CAC algorithm for non-fully-connected CRNs as a bench mark. However, analysis of the effective capacity of the service process of an SU is not straightforward as it depends on the channel access scheme. The approach used in Subsection 2.3.1 to analyze the effective capacity of the CRN can be adopted to analyze that of the service process of each node.

The packet buffer of each source node act in the FCFS service discipline. Therefore, in order to satisfy the delay requirement of voice packets, the effective capacity of the service process of each source node should be larger than the constant arrival rate. A successful packet transmission from a target source node occurs whenever there are no collisions at the target receiver node. Therefore, the service process of the target source node is governed by the transmissions of the neighboring source nodes of the target receiver node. The effective capacities of the discrete-time service process for independent and dependent channel availability scenarios can be obtained by (2.2). In a particular time-slot (**irrespective of its availability**), define the state of the service process of a target source node as follows: If a successful transmission occurs during the time slot, the source node is in state 1, and state 0 otherwise. The effective capacities for independent and dependent channel availability cases are given by

$$\zeta_c(\theta) = -\frac{1}{\theta}\ln\left[\frac{F_{0,0}+F_{1,1}e^{-\theta\cdot n_{pk}}}{2} + \sqrt{\left(\frac{F_{0,0}-F_{1,1}e^{-\theta\cdot n_{pk}}}{2}\right)^2 + F_{0,1}F_{1,0}e^{-\theta\cdot n_{pk}}}\right].$$

where, $F_{i,j}$ ($i, j \in \{0,1\}$) is the transition probability of a node from state i to state j. In an **available time-slot** define the state of transmission (transmission state) as follows; If a transmission is successful, the source node is in state 1, and state 0 otherwise. Consider a Markov chain in which the state is represented by the channel state and transmission state pair which consist of three states (1,1), (1,0), and (0,0). Denote the state transition probability matrix of the Markov chain by $\acute{F}$. The node is at state 1 if both the transmission state and the channel state are 1, and state 0 otherwise. The state transition probability matrix F of the service process a source node can be obtained using the state transition probability matrix $\acute{F}$. The condition to satisfy the delay requirement of a voice call i is given by $\delta^* \zeta_c(\delta^*) \geq \frac{1}{D_{\max}} \log\left(\frac{1}{\epsilon}\right)$, where δ^* is the solution to the equation $\zeta_c(\delta){=}\frac{1}{T_I}$. This condition can be given in the form $\zeta_c\left(\delta^*_{\text{min}}\right) \geq \frac{1}{T_I}$, where $\delta^*_{\text{min}}{=}\frac{T_I}{D_{\max}} \log\left(\frac{1}{\epsilon}\right)$. In the distributed non-fully-connected network scenario, the probability $P_{S,1}$ of a target source node ω_s is given by

$$P_{S,1} = \Pi_{i_s \in G_{\omega_r}} \varrho_\omega (1 - \rho_i \cdot \varrho_i) \tag{2.11}$$

where ϱ_j and ρ_j are the transmission probability given that the buffer is non-empty and the probability of having a non-empty buffer of source j_s $(\in \{G_{\omega_r} \cup \omega_s\})$, and ω_r is the target receiver node. However, the evaluation of ρ_i is not straightforward as it depends on the transmissions of the neighboring source nodes of receiver i_r. Therefore, rather than evaluating the exact value of ρ_i, we investigate the possibility of obtaining a close upper bound for the value of ρ_i. From the DTMC illustrated in Fig.2.7, it can be seen that the delay bound violation probability P_e of a constant-rate voice traffic flow and the probability ρ of a voice buffer being non-empty, monotonically decrease with the successful transmission probability $P_{S,1}$. Therefore, the delay requirement $P_e \leq \epsilon$ can be transformed to $P_{S,1} \geq P^*_S$ or $\rho \leq \rho^*$, where P^*_S is the $P_{S,1}$ value at $P_e = \epsilon$ and ρ^* is the ρ value at $P_{S,1} = P^*_S$. The variation of ρ and P_e with $P_{S,1}$, and the relationship of P^*_S, ρ^*, and ϵ are illustrated in Fig.2.11. As long as the existing source nodes satisfy the delay requirement $P_e \leq \epsilon$, the probability ρ is upper bounded by ρ^*. Therefore, instead of using $P_{S,1}$, we substitute $P^*_{S,1}{=}\Pi_{i_s \in G_{\omega_r}} \varrho_\omega (1 - \rho^*_i \cdot \varrho_i)$ $(\leq P_{S,1})$ in (2.11). When the system supports non-homogeneous voice traffic flows with different delay bounds, let C denote

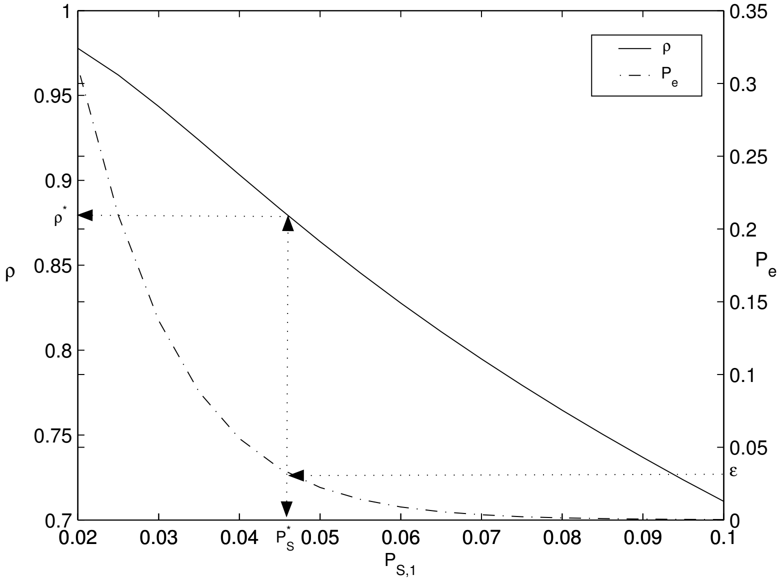


Figure 2.11: Variation of ρ and P_e with $P_{S,1}$ for $D_{\max}$=100, T_I=10, n_{pk}=5, P_{01}=0.8, P_{10}=0.2, and ϵ=0.01.

the set of all voice traffic classes in the network. Each voice traffic class c ($\in C$) has unique delay bound $D_{\max}(c)$, $P_S^*(c)$, and $\rho^*(c)$ values. Therefore, ρ_i^* and ϱ_i of $P_{S,1}^*$ should be replaced by their respective values of the traffic class c_i as $\rho^*(c_i)$ and $\varrho\,(c_i)$, where $\varrho\,(c_i)$ is the default ϱ value for the traffic class c_i. Denote the source and receiver nodes of the incoming call, ω, as the target source (ω_s) and receiver (ω_r) nodes. In order to make sure that the delay requirements of all ongoing calls and the new call are satisfied, effective capacities of each receiver node i_r ($\in \{G_{\omega_s} \cup \omega_r\}$) should be larger than the packet arrival rate $1/T_I$. The benchmark CAC algorithm based on the effective capacity is given in algorithm A1.

Each receiver node in the network should run the algorithm and identify the set $\hat{C}_j$ ($j \in \{G_{\omega_s} \cup \omega_r\}$). The set $\hat{C}_\omega$ and $\hat{C}_i$ ($i \in G_{\omega_s}$) are the set of voice classes that can be admitted by ω_r, and to the neighborhood of an existing receiver node i_r, respectively, without violating the delay requirement of the existing and incoming voice calls. The new source and receiver nodes listen to the channel and identify the set of voice classes $C_\omega = \bigcap_{i_r \in \{G_{\omega_s} \cup \omega_r\}} \hat{C}_i$ that can admit call ω.

```
Data  : $C_i$={$c_j$ : $j \in G_{i_r}$}
Result: $\hat{C}_i$
$\hat{C}_i \leftarrow \emptyset$;
repeat
    randomly select $c_k$ from $C$;
    $C \leftarrow C$-{$c_k$};
    if $i$==$\omega$ then
        $P_{S,1} \leftarrow \varrho(c_k)\Pi_{j \in G_{\omega_r}}(1-\rho^*(c_j)\cdot\varrho(c_j))$;
    else
        $P_{S,1} \leftarrow \varrho(c_i)\Pi_{j \in G_{i_r}}(1-\rho^*(c_j)\cdot\varrho(c_j))(1-\rho^*(c_k)\cdot\varrho(c_k))$;
    end
    $\delta^*_{\min} \leftarrow \frac{T_I}{D_{\max}(c_i)}\log\left(\frac{1}{\epsilon}\right)$;
    if $\zeta(\delta^*_{\min}) \geq \frac{1}{T_I}$ then
        $\hat{C}_i \leftarrow \{\hat{C}_i \cup c_k\}$ ;
    end
until $C$==$\emptyset$;
Exit;
```

Algorithm 1: CAC algorithm based on the effective capacity

If C_ω=$\emptyset$, call ω cannot be admitted to the system. The effective bandwidth/capacity approach can be applied to different types of traffic by evaluating the effective bandwidth [47][48] of the source traffic and the effective capacity of the service process via modeling the source buffer occupancy at the packet level. However, this approach is computationally complex due to the requirement of calculating the effective capacity at run-time. It is possible to introduce a less complex approach for the CAC for non-homogeneous voice traffic using the relationship of $P_{S,1}$, ρ, and P_e.

Based on Fig.2.11, guaranteeing $P_{S,1}{\geq}P^*_S(c)$ guarantees $\rho{\leq}\rho^*(c)$. Therefore, if the probability $P_{S,1}$ of source i_s ($\in G_{\omega_r}$) satisfies $P_{S,1}{\geq}P^*_S(c_i)$, the inequality $\varrho_\omega\Pi_{i_s \in G_{\omega_r}}(1-\rho_i\cdot\varrho_i) \geq \varrho_\omega\Pi_{i_s \in G_{\omega_r}}(1-\rho^*(c_i)\cdot\varrho_i)$ always stands. Provided that $P_{S,1}{\geq}P^*_S(c_i)$ for all $i_s \in G_{\omega_r}$, the delay requirement of the incoming call can be guaranteed by choosing a proper ϱ_ω value for its source ω_s, which satisfies $\varrho_\omega\Pi_{i_s \in G_{\omega_r}}(1-\rho^*(c_i)\cdot\varrho_i) \geq P^*_S(c_\omega)$. However, as discussed in Subsection 2.5.1, the admission of a new source node increases the probability P_e of delay bound violation of each source i_s, where i_s is the corresponding source node of i_r ($\in G_{\omega_s}$). Therefore, it is required to guarantee that $P_{S,1}$ values of the said source nodes and the new source node are kept above their respective $P_S^*(c_j)$ ($j_r \in \{G_{\omega_s} \cup \omega_r\}$) values by making sure that the following conditions are met respectively

$$\begin{aligned} &\varrho_j \cdot \prod_{i_s \in G_{j_r}} (1 - \rho^*(c_i) \cdot \varrho_i) \geq P_S^*(c_j), \;\; \forall\; j_r \in G_{\omega_s} \\ \text{and} \qquad\qquad & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2.12) \\ &\varrho_\omega \cdot \prod_{i_s \in G_{\omega_r}} (1 - \rho^*(c_i) \cdot \varrho_i) \geq P_S^*(c_\omega) \end{aligned}$$

where the LHSs of (2.12) are always less than or equal to $P_{S,1}$. The expressions of the LHSs of (2.12) can be evaluated using γ_j ($=\varrho_j \cdot \prod_{i \in G_{j_r}} (1 - \rho^*(c_i) \cdot \varrho_i)$) and c_j obtained from the neighboring receiver nodes of the new source node, and γ_ω ($=\prod_{i \in G_{\omega_r}} (1 - \rho^*(c_i) \cdot \varrho_i)$) obtained from the new receiver node. The CAC algorithm based on the relationship among P_e, $P_{S,1}$, and ρ is given in Algorithm 2.

In the algorithm, parameter $\varrho_{\min}$ is the minimal ϱ value which satisfies the first inequality in (2.12), $\varrho_{\max}$ is the maximal ϱ value which satisfies the second inequality in (2.12), and $\acute{\beta}$ is the transmission probability selection parameter. Algorithm 2 searches for $\varrho_{\min}$ and $\varrho_{\max}$ by increasing ϱ_ω from 0 to 1 in a step size ϱ_s. The smaller the ϱ_s, the higher the accuracy of $\varrho_{\min}$ and $\varrho_{\max}$ values. However, the smaller the ϱ_s, the larger the number of iterations required to get the results, leading to a larger processing time. If the algorithm outcome is to admit the call, it needs to choose a ϱ_ω value ($\varrho_{\min} \leq \varrho_\omega \leq \varrho_{\max}$) for the transmissions of the new source node. The probability $P_{S,1}$ of the new source node and corresponding source nodes of its neighboring receiver nodes will vary depending on the chosen ϱ_ω value. Therefore, a particular $\acute{\beta}$ value should be selected for the network to obtain a ϱ_ω ($=\varrho_{\min} + \acute{\beta}(\varrho_{\max} - \varrho_{\min})$) value, such that the network capacity is maximized. This can be carried out by trial and error method off-line.

```
   Data   : $\Gamma$={$(\gamma_i, c_i)$ : $i \in \{G_{\omega_s} \cup \omega_r\}$}
   Result: $\varrho_\omega$, Admit the call or block the call
1  $\kappa \leftarrow 0$;
2  $\varrho_\omega \leftarrow \varrho_s$;
3  while $\varrho_\omega \leq 1$ do
4      if $\gamma_\omega \cdot \varrho_\omega \geq P_S^*(c_\omega)$ then
5          $\varrho_{\min} \leftarrow \varrho_\omega$;
6          Go to 10;
7      end
8      $\varrho_\omega \leftarrow \varrho_\omega + \varrho_s$;
9  end
10 while $\varrho_\omega \leq 1$ do
11     $H_g \leftarrow G_{\omega_s}$;
12     repeat
13         randomly select $j_r$ from $H_g$;
14         $H_g \leftarrow H_g$-{$j_r$};
15         if $\gamma_j\,(1 - \rho^*(c_\omega) \cdot \varrho_\omega) < P_S^*(c_j)$ then  Go to 21;
16     until $H_g$==$\emptyset$;
17     $\kappa \leftarrow 1$;
18     $\varrho_{\max} \leftarrow \varrho_\omega$;
19     $\varrho_\omega \leftarrow \varrho_\omega + \varrho_s$;
20 end
21 if $\kappa$==1 then
22     $\varrho_\omega \leftarrow \varrho_{\min} + \acute{\beta}(\varrho_{\max} - \varrho_{\min})$;
23     Admit the call;
24 else
25     Block the call;
26 end
27 Exit;
```

Algorithm 2: CAC algorithm based on the successful transmission probability

2.6 Numerical Results

Computer simulations are carried out to evaluate the accuracy of the capacity analysis of the given channel access schemes and to investigate the performance of the two CAC algorithms. In order to depict the primary user activities, the channel is made on and off according to the dependent and independent channel occupancy statistics of PUs. The voice traffic classes used in the analysis are given in Table 2.1 Note that all the time durations are normalized to T_S. The typical values of the on and off durations of an on-off voice source are around 320ms and 640 ms, respectively [43][61]. However in [61], it is shown that these durations are dependent on the factors such as conversation topics and situations of voice calls. The probability of delay bound violation, P_e, is obtained by the ratio of the number of dropped packets (at a source node due to the violation of delay bound) to the total number of packets generated by the source node. Each simulation result is obtained by averaging the results of 10 simulation runs each having a duration of 10^6 time-slots.

Table 2.1: The voice traffic classes used in the simulations

Voice traffic class	Notation	Traffic type	Mean on duration	Mean off duration	T_I	D_{max}
Class 0	c_0	on-off	320	640	10	30
Class 1	c_1	constant-rate	-	-	10	30
Class 2	c_2	constant-rate	-	-	10	100
Class 3	c_3	constant-rate	-	-	10	250

2.6.1 Capacity Analysis of Centralized and Distributed Networks

Consider homogeneous voice traffic flows of class c_i ($i \in \{0, 1, 2, 3\}$). While keeping N constant during a simulation run, the probability P_e is obtained for a particular channel access scheme and channel statistics. Starting from $N = 2$, we increase N by one for each simulation run and the resultant probability P_e is compared with ϵ to obtain $N_{\max}$, which satisfies $P_e \leq \epsilon$.

Centralized Network

Figs. 2.12-2.13 show the variation of $N_{\max}$ with n_{pk} in a centralized network with FCFS service discipline and different channel availability statistics for on-off (class c_0) and constant-rate (class c_1) voice traffic, respectively. The results demonstrate that the analytical results match closely with the simulation results. However, the analytical results stay slightly below the simulation results due to the conservative nature of the theory of effective bandwidth and its dual, effective capacity. The capacity of the system increases with the number of voice packets that can be transmitted in a single time-slot (per channel), n_{pk}. However, a proportional increment in the channel rate is not required to increase the n_{pk}. Therefore, capability of transmitting multiple packets in a single-time-slot can have a considerable impact on the system capacity. In the on-off voice sources under consideration, the duration of talkspurts are only one third

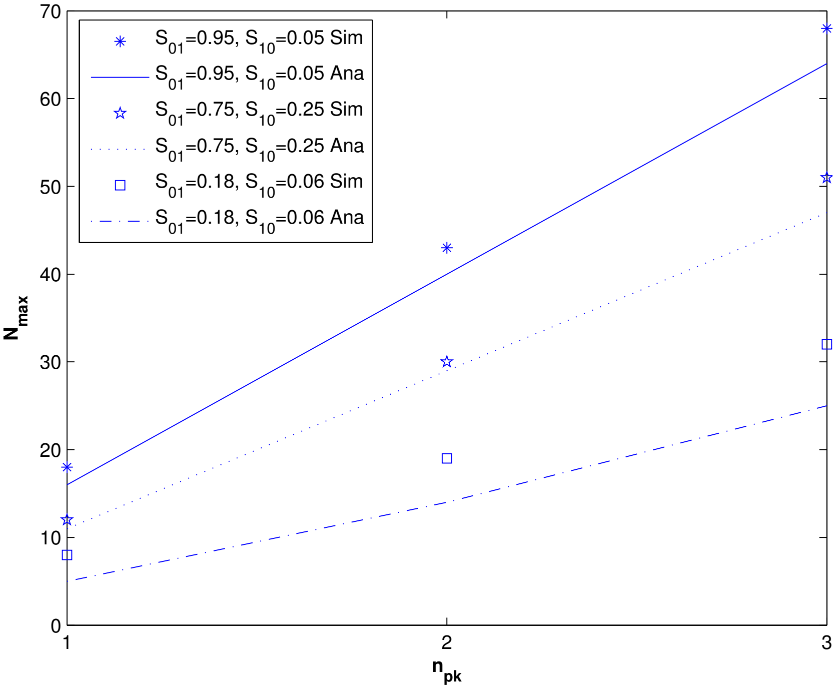


Figure 2.12: Variation of $N_{\max}$ with n_{pk} for on-off voice traffic in a cantralized network with FCFS service discipline.

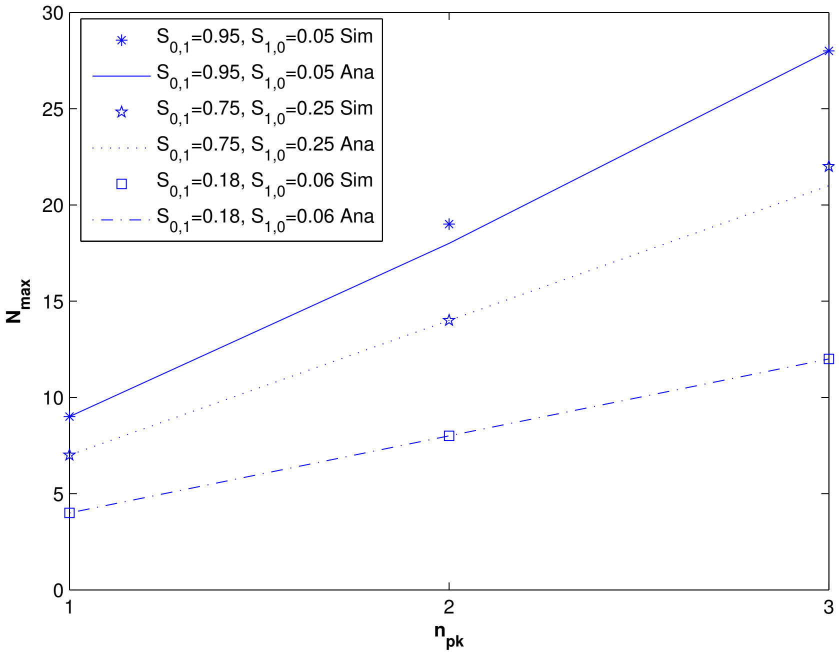


Figure 2.13: Variation of $N_{\max}$ with n_{pk} for constant-rate voice traffic in a cantralized network with FCFS service discipline.

of the call duration. The number of voice packets generated in an on-off voice traffic flow is approximately one third of that of a constant-rate voice traffic flow. Therefore, with the given FCFS service discipline (the ideal scenario), the on-off voice traffic provides more than twice the capacity of the constant-rate voice traffic.

Distributed Network

Consider homogeneous voice traffic flows of class c_2. Figs. 2.14-2.16 show the variation of P_e with N obtained from numerical analysis and simulations with slot-ALOHA, random allocation, and round-robin channel access schemes, respectively, for different channel availability statistics with n_{pk}=4. The results demonstrate that the P_e obtained form simulation match well with the analytical results in all three scenarios. The system capacity $N_{\max}$ is given by the maximum N having P_e less than ϵ (0.01 in our simulation). Further, it can be observed

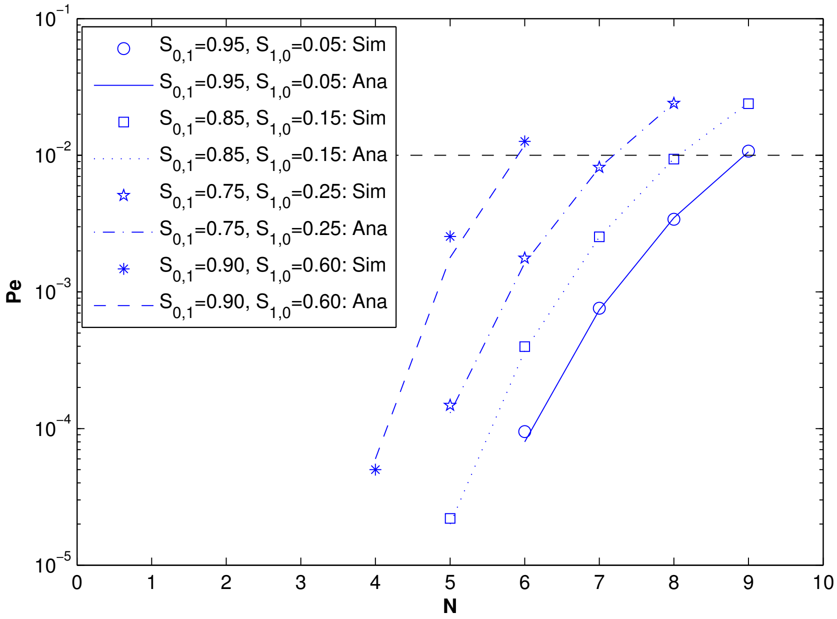


Figure 2.14: Variation of P_e with N for a fully-connected network with slot-ALOHA channel access.

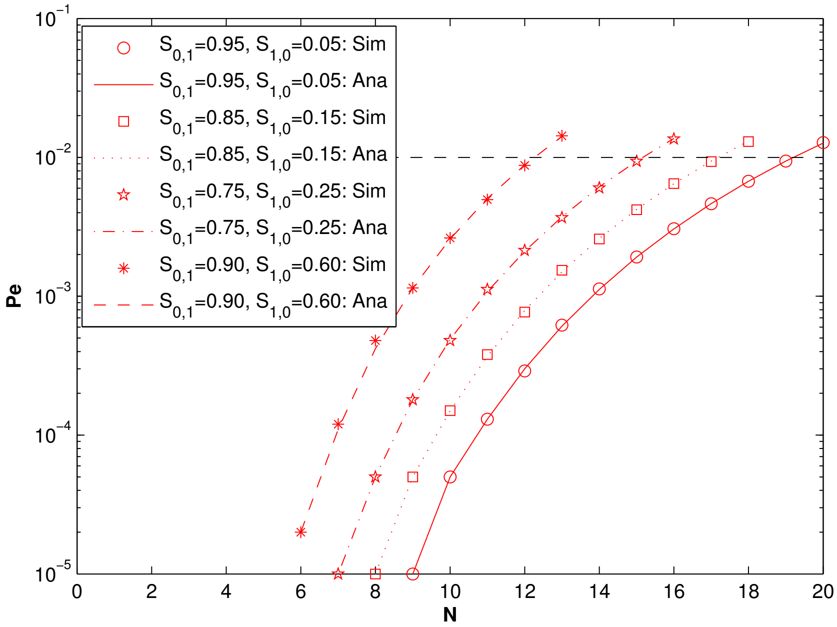


Figure 2.15: Variation of P_e with N for a fully-connected network with random allocation.

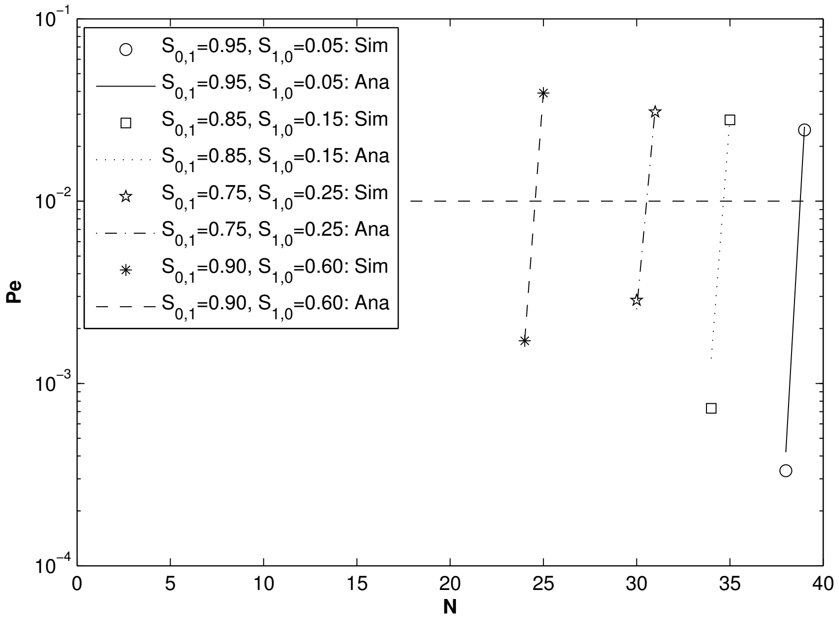


Figure 2.16: Variation of P_e with N for a fully-connected network with round-robin channel access.

that the higher the mean channel availability, p_1=$S_{0,1}/(S_{0,1}+S_{1,0})$, the higher the system capacity in all three cases. Fig. 2.17 shows the variation of $N_{\max}$ with n_{pk} for all three channel access schemes having $S_{0,1}$=0.95 and $S_{1,0}$=0.05. It is observed that the higher the mean channel availability and n_{pk}, the higher the capacity in all three channel access schemes. Further, the round-robin scheme provides the highest system capacity and slot-ALOHA provides the worst system capacity, when compared with the other two. Note that the overhead required for the establishment of the round-robin scheme is much higher than that of the slot-ALOHA scheme, as explained in Subsection 2.2.3, which is neglected in the simulation. Furthermore, it is observed that the rate of increment of system capacity with n_{pk} in the round-robin scheme is higher than the random allocation and slot-ALOHA schemes. In the slot-ALOHA and random allocation schemes, the probability of transmission is irrespective of the buffer occupancy of packets. However, in the round-robin scheme, there is a higher probability to transmit when the waiting time of the queue-head is larger (i.e., when there are more number of packets in the

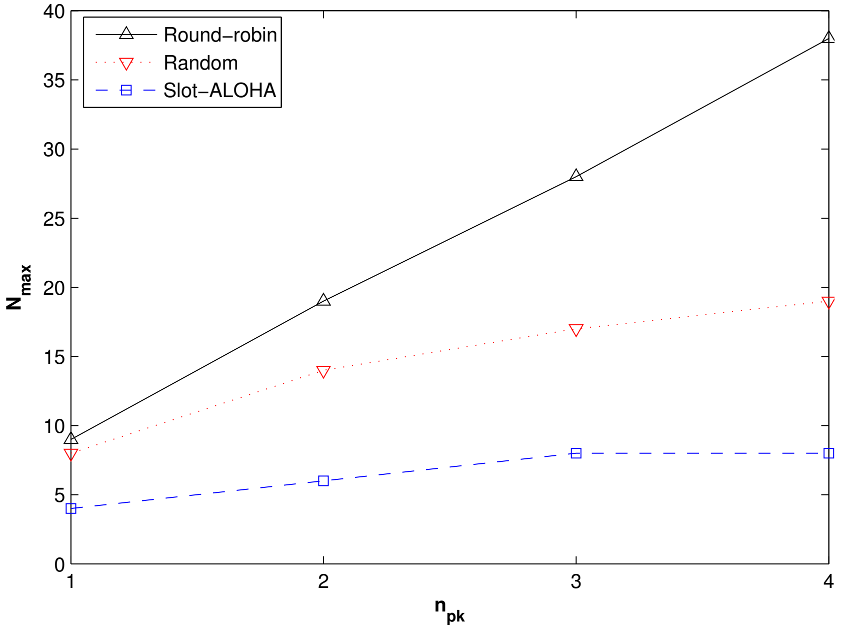


Figure 2.17: Variation of $N_{\max}$ with n_{pk} for a fully-connected network for all three channel access schemes.

buffer), which allows a node to transmit a larger number of voice packets during a transmission than in the other two schemes. Therefore, the mean number of packets transmitted during a channel access opportunity is smaller in the slot-ALOHA and random allocation schemes than that in the round-robin scheme, which explains that the latter has a higher rate of capacity improvement with n_{pk}. The system capacity with the round-robin scheme is similar to that of centralized network with FCFS service discipline. Since, the round-robin scheme does not need the packet timing information, it can be considered as a promising candidate for voice service support over CRNs.

Fig. 2.18 shows the variation of $N_{\max}$ with T_{off} for all three channel access schemes having p_1 constant in 0.8 and n_{pk}=4. The results demonstrate that the longer the T_{off}, the lower the system capacity even though the mean channel availability remains constant. The longer the T_{off}, the longer the duration of the busy periods of the channel from the viewpoint of the SUs, leading to longer the durations of packet waiting times. This increases the probability of delay

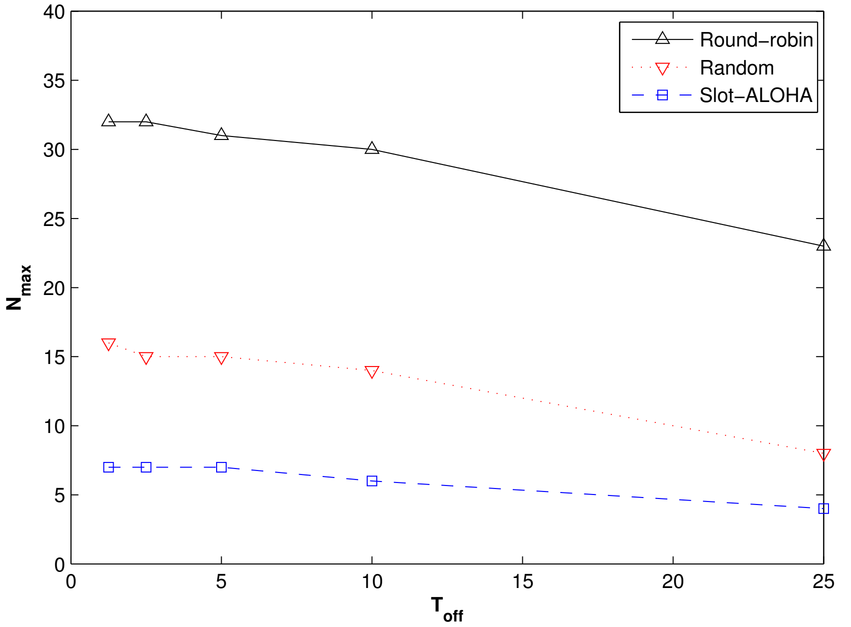


Figure 2.18: Variation of $N_{\max}$ with T_{off} for a fully-connected network for all three channel access schemes.

bound violation of the voice packets. Even if the channel available duration is longer (corresponding to longer T_{on}) it is not possible to transmit the packets which have already violated the delay bound. Therefore, the longer the T_{off}, the lower the system capacity. This shows the importance of considering the state transition probabilities of the channel in analyzing the system capacity rather than considering the mean channel availability.

Discussion

In our work, we assume ideal channel sensing (i.e., error free detection of primary activities). However, sensing errors are inevitable in practical scenarios. There are two types of sensing errors, namely, missed detection (an SU or the BS does not detect the presence of a PU) and false alarm (an SU or the BS detects presence of a PU while the PUs is not present) [60]. The missed detections lead SUs to transmit simultaneously with the PUs, causing interference.

In order to establish the CRN, the probability, P_{md}, of missed detection has to be controlled below a certain threshold to minimize the interference with the PUs. Despite the interference, the SU transmission can be successful. The false alarms reduce the channel utilization of the SUs. Therefore, the probability, P_{fa}, of false alarm has to be minimized to improve channel utilization of the SUs. From the viewpoint of an SU (or the BS), the channel availability differs from the true state of the channel due to the presence of the sensing errors. Denote the channel state from the viewpoint of an SU (or the BS) as the virtual channel state. The virtual channel state transition probabilities can be obtained using the channel state transition probability matrix and the error probabilities P_{md} and P_{fa}, respectively. We can incorporate the effect of the sensing errors into our capacity analysis after some efforts on modification of the Markov chains.

When the sensing errors are present, packet transmissions of the SUs in the centralized network depend on the virtual channel state of the BS. In order to incorporate the sensing errors into the capacity analysis, the channel state transition probabilities in (2.2) should be replaced by the virtual channel state transition probabilities. In the fully-connected slot-ALOHA network, the packet transmission of an SU depends on the channel sensing errors. Therefore, the state transition probabilities of the DTMC in Fig. 2.7 should be modified to incorporate the sensing errors. The modifications to the state transitions are illustrated in Fig. 2.19, where $P_{S,md}$ is the probability of successful transmission given the occurrence of a missed detection and $P_{S,1}$=$\varrho(1-\rho\varrho(1-P_{fa}))^{N-1}$ is the successful transmission probability given the channel is available. The transition probabilities $P_{md}P_{S,md}S_{0,j}$ and $(1-P_{fa})P_{S,1}S_{1,j}$ are due to successful transmissions given the channel is not available (transmission being successful due to missed detection) and available (when there is no false alarm), respectively, where $j \in \{0,1\}$. The transition probability $(1-(1-P_{fa})P_{S,1})S_{1,j}$ and $(1-P_{md}P_{S,md})S_{0,j}$ are due to non occurrence of packet transmissions when the channel is available and not available, respectively, where $j \in \{0,1\}$. Similar to that given in Subsection 2.4.1, the delay bound violation probability can

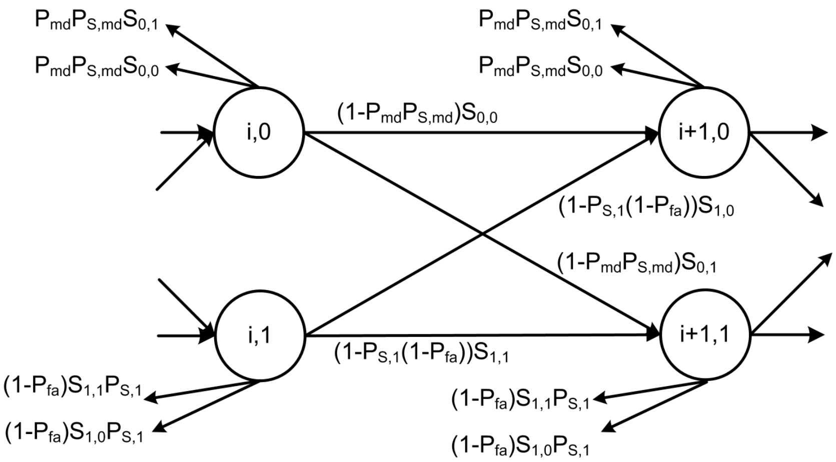

$P_{md}P_{S,md}S_{0,1}$
$P_{md}P_{S,md}S_{0,0}$
i,0
$(1-P_{md}P_{S,md})S_{0,0}$
$P_{md}P_{S,md}S_{0,1}$
$P_{md}P_{S,md}S_{0,0}$
i+1,0
$(1-P_{S,1}(1-P_{fa}))S_{1,0}$
$(1-P_{md}P_{S,md})S_{0,1}$
i,1
i+1,1
$(1-P_{S,1}(1-P_{fa}))S_{1,1}$
$(1-P_{fa})S_{1,1}P_{S,1}$
$(1-P_{fa})S_{1,0}P_{S,1}$
$(1-P_{fa})S_{1,1}P_{S,1}$
$(1-P_{fa})S_{1,0}P_{S,1}$

Figure 2.19: The modifications to the state transitions of the DTMC in Fig. 2.7 to incorporate sensing errors.

be evaluated by finding the steady state probabilities of the DTMC. The sensing errors can be incorporated into the capacity analysis of the round-robin scheme by modifying the state transition probabilities of the DTMC in Fig. 2.10, and the modification depends on the mechanism used to establish the round-robin scheme. With the above, it is clear that the sensing errors can be incorporated in our capacity analysis with slight modifications in the state transition probabilities of the DTMCs.

In a multiple channel network, an SU will either select a channel and sense for availability or sense all the channels and select an available channel. Two approaches lead to different successful transmission probabilities. We can extend our capacity analysis approach for a multiple channel CRN by evaluating the corresponding successful transmission probabilities. The service quality is given in terms of queuing delay D. The QoS requirement can be relaxed by defining the service quality in terms of the mean of queuing delay, $E[D]$.

As mobile video is predicted to generate most of the mobile traffic growth through 2005 [37], it is important to study video streaming over the CRNs. As given in [62], video frames are generated in burst according to a coding and compression algorithm, and each video burst consists of a number of video packets (with a pdf given by negative binomial distribution). The video clips are grouped into a small number of shot classes depending on the burst size, and a video traffic flow is modeled my a Markov modulated Gamma process in [63]. The author also analyze the EB of the video traffic flow for a maximum data loss rate of 10^{-2}. Therefore, the number of video traffic flows that can be supported by the centralized FCFS system can be studied using the EC evaluated in Subsection 2.3.1 and the EB approach in [63]. For the distributed networks, the possibility to carry out a packet level analysis of the source buffer can be studied, given the statistics of the video bursts and the probability of channel access. Further research is necessary to model the source buffer state using a Markov chain, and to analyze the packet dropping probability using the steady state probabilities as given in Subsection 2.4.1.

2.6.2 CAC of the Non-fully-connected Network

For the performance comparison of the CAC procedure (P1) and two CAC algorithms (A1 and A2), we consider a CRN with homogeneous voice traffic. For the performance comparison of algorithms A1 and A2, we consider a network with both traffic classes, where new call arrivals are equally likely to be of class c_2 or c_3. The network coverage area of each voice source/receiver node is a circle with a radius of unit length. The inter-arrival time of voice calls is exponentially distributed, and the location of source nodes is uniformly distributed in a square network area. Ten different data sets are generated, each containing 8000 samples of source and receiver location and call inter arrival time. In order to compare the two algorithms, 10 different simulation runs were carried out for each algorithm using the generated data sets over a constant network area. As the network is non-fully-connected, the system capacity depends on the coverage area of the network. We saturate the network with voice calls to obtain the maximum number of voice calls that can be supported by the system, and obtained the results for different network coverage areas.

Fig. 2.20 shows the comparison of the network capacity (with the 95% confidence interval) of class c_2 voice calls using procedure P1 and algorithms A1 and A2. The CAC procedure

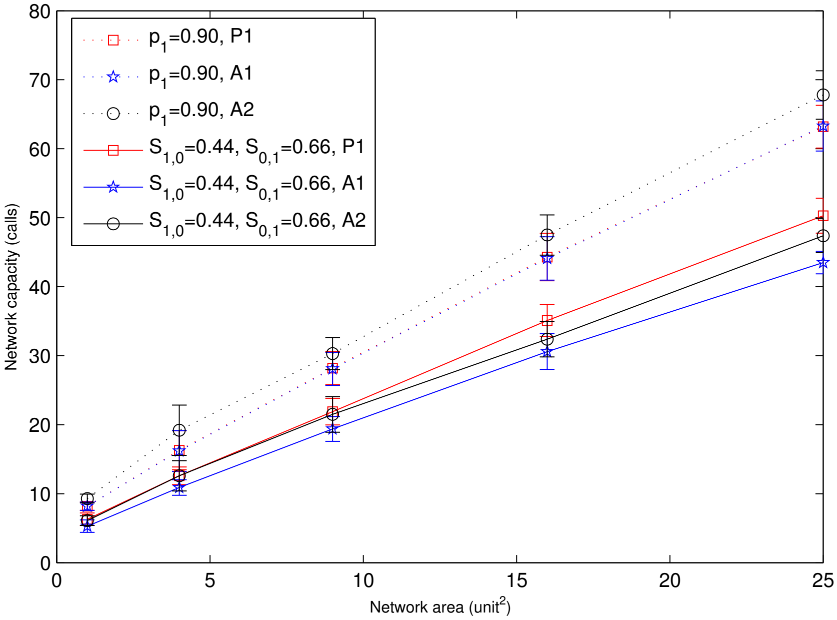


Figure 2.20: Variation of the network capacity with the network area for procedure $P1$ and algorithms A1 and A2.

P1 outperforms the algorithms A1 and A2 when the mean channel availability is lower, and the algorithm A2 outperforms the other two when the mean channel availability is higher. The algorithm A2 opportunistically chooses the probability ϱ at the instance of call admission whereas P1 has a fixed ϱ value. Therefore the opportunistic ϱ selection may choose different ϱ values for different calls leading to a probability $P_{S,1}$ which is just enough to satisfy the admission criterion (ϱ can vary from ϱ_{min} to ϱ_{max}). The lower the channel availability, the larger the P_S^*. The larger the P_S^*, the lower the tolerance for admitting a new call and vice versa. Therefore, lower channel availability can lead to lower capacities when the ϱ selection is opportunistic as in A2. The performance of algorithm A1 always stays below A2 due to the conservative nature of the theory of effective capacity. The required information from the neighboring nodes and the calculation complexity of P1 is less than those of A1 and A2.

Therefore, the procedure P1 can be a better choice over A1 and A2 at low channel availability, and A2 can be a better choice over P1 and A1 for a network with homogeneous voice traffic at high channel availability.

Fig.2.21 shows the variation of the average network capacity with the network area, using algorithms A1 and A2 for the two equally likely voice traffic classes. The results demonstrate that algorithm A2 is a better choice over A1. The average network capacity with the mixture of two voice traffic classes is higher than that for voice class c_2. The relaxed QoS requirement of voice traffic class c_3 allows more calls to be admitted. Clearly, there is a trade-off between the number of calls in the systems and the service quality, as expected. Algorithm A2 can be extended to other contention based channel access schemes (e.g. IEEE 802.11 RTS/CTS based) and traffic types, given that P_e and ρ monotonically decrease with P_s, and the channel contention is independent over adjacent time-slots.

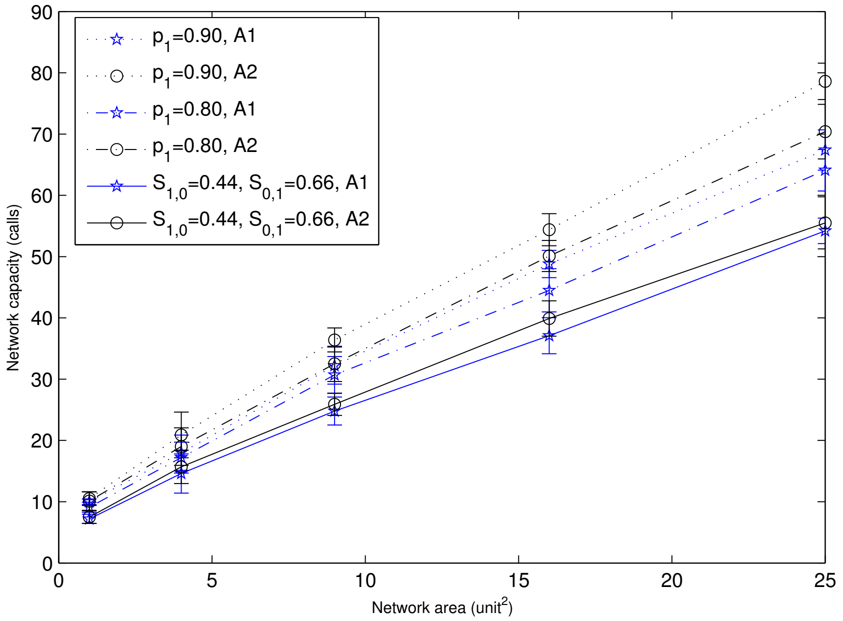


Figure 2.21: Variation of the network capacity with the network area for algorithms A1 and A2 for a network with two voice traffic classes.

2.7 Summary

In this chapter, we have studied the voice capacity and call admission control for CRNs. A stochastic delay requirement, both independent and dependent channel availability statistics, and different channel access schemes are considered. The on-off voice capacity of a centralized CRN is studied for the FCFS service discipline using the theories of effective bandwidth and its dual effective capacity. The analytical results appear to be slightly lower than the simulation results due to the conservative nature of the theory of effective bandwidth. Further, it is observed that the silent-suppressed (on-off) voice sources provides more than twice the capacity over that of constant-rate voice traffic. The existing DTMC model is modified to analyze the the capacity of slot-ALOHA scheme, and a new DTMC model is developed to analyze the capacity of round-robin scheme in supporting constant-rate voice traffic over distributed fully-connects CRNs. It is shown that the round-robin scheme performs better than the other two schemes, and the capacity is very close to that of the FCFS scheme used in the centralized network. Therefore, the round-robin scheme is a better choice in fully-connected networks to support voice traffic, and it can be established using a token based scheme or a mini-slot based scheme as explained in [46][17]. Further, the maximum number of voice packets that can be transmitted in a time-slot and the mean channel unavailable duration have a significant impact on the system capacity. The longer the mean channel unavailable durations the lower the system capacity, even when the mean channel availability remains unchanged. In order to relax the assumption of perfect channel sensing, possible extensions to the analytical models to incorporate the sensing errors are discussed.

We use the capacity analysis results of the fully-connected network to limit the number of neighboring users of each target voice user in non-fully-connected CRNs with slot-ALOHA network coordination. It is only applicable for constant-rate voice sources with the same delay requirement. However, having long delay bounds and large delay bound violation probabilities can increase the system capacity. Therefore, we develop two new CAC algorithms to support voice sources with different delay requirements (different delay bound and maximum delay bound violation probability). It is shown that the loner the delay bound, the larger the system capacity. In other words, the lower the required service quality, the higher the system capacity. A low quality service can be priced at a lower cost than a high quality service. Giving the users an option to choose the required service quality can increase the level of user satisfaction.

Chapter 3

Service Response Time of Elastic Data Traffic

3.1 Introduction

As we have discussed in Chapter 2, various studies on conversational and streaming type traffic flows over CRNs have been carried out over the recent years. So far, little attention is paid to the performance analysis of request-response type services such as web browsing over CRNs. The impact of primary user activities on traffic congestion and the economic interaction between secondary user and primary network operators are studied in [64] and [65], respectively, when the SUs are data users. This type of services does not require strict QoS as in conversational or streaming services, but has a moderate service requirement in the form of response time.

Most of the resource allocation/scheduling works in CRNs mainly focus on throughput optimization/fairness, and they do not deal with any specific data file length distributions or response time as a performance metric. The data service is a non-real-time service, where the rate of flows adjusts to fill available bandwidth [66]. Therefore, the data service is also called elastic data service. A performance analysis of elastic data traffic in non-cognitive networks is carried out in [66]. Different bandwidth sharing techniques based on maximum throughput, min-max fairness, proportional fairness, and weighted fairness are considered in the analysis. The mean response time evaluation of elastic data traffic flows is studied for cellular/WLAN integrated networks in [29] . The network supports streaming and elastic data traffic flows, and the data files are served in processor sharing service discipline. A mean response time approximation for the SRPT service discipline under a heavy traffic condition is given in [67] (and references there in). In all these works, the short-term mean channel rate available for a data user does not vary with time, and therefore the long-term mean channel rate is used for the response time analysis. However, in CRNs the channel availability for SUs varies with time due to the interruptions by the PUs (bursty PU traffic), and the short-term mean channel availability deviates from the long-term mean channel availability. Therefore, the effect of the transmission interruptions caused by the PUs should be considered in the analysis. In [27], the mean throughput and delay of transmission control protocol (TCP) and constant bit rate connections are analyzed for CRNs with on-off PU behaviors. However, there aren't many research efforts devoted on the performance analysis of elastic data traffic over CRNs. From the viewpoint of the SUs in a CRN, the available channel time can be considered as a service with break downs. The expected queue lengths and related operating characteristics of a queuing station with breakdown are studied in [68], which can also be applied in the context of CRNs. Further, the relationship between the queuing station with breakdown and a single server queuing system with preemptive priorities is also studied. However, the work in [68] is limited to the FCFS service discipline, which is not always the best service discipline. In all these works, the mean of the response time is considered as the service quality parameter, due to the complexity of analyzing its probability distribution.

Motivation and Objectives

Elastic data traffic falls into the non-real-time traffic category, and the mean response time is an important measure to provide a satisfactory service quality. The service interruptions due to the presence of PUs increase the response time of elastic data traffic flows operating over CRNs. The length of requested files vary according to the file length distribution (Weibull distribution is common for Web traffic [69]), which also has a significant impact on the mean response time. Different service disciplines can provide different priorities to the service requests based on their file lengths. For example, SPTWP service discipline provides a higher priority to the requests with short file lengths than the requests with long file lengths, whereas the PS service discipline provides equal opportunity to all the requests. Therefore, different service disciplines result in different mean response times. The longer the response time (i.e., the longer the duration a user waits to start reading/viewing the file), the lesser the user satisfaction. Choosing the correct service discipline based on the service time requirement (or the file length) distribution can reduce the mean response time of the files. Therefore, it is important to study the relationship between the mean response time and the different system parameters such as the channel availability statistics (PU behavior), data session arrival process, and the service time requirement, respectively, for a given service discipline. The mean response time results can be used in developing CAC algorithms for elastic data traffic. The objective of this chapter is to study the mean response time of elastic data traffic over CRNs with different service disciplines, and find the impact of different system parameters on the mean response time.

Contributions

The contribution of this chapter is three fold: (i) We derive mathematical expressions for the mean response time of elastic data traffic service operating over a single channel time-slotted centralized CRN with three service disciplines, namely, shortest processor time without preemption [34], shortest processor time with preemption, and shortest remaining processing time, in comparison with the processor sharing service discipline. The PU activities are considered to have an on-off behavior with on and off durations following geometric distributions; (ii) We compare the mean response times of all four service disciplines under different data traffic load conditions, and demonstrate that the SPTNP is a better choice over the PS service discipline for a heavy traffic load condition; (iii) We compare the mean response times of all four service disciplines under Weibull distributed service time requirements with different tail properties, and demonstrate that the preemption reduces the mean response time when the service time requirement (data file size) follows a heavy tailed distribution. The response time analysis can be used for call admission control to ensure service satisfaction. To the best of our knowledge, this is the first study of the service response time for the elastic data traffic under the service disciplines with service interruptions for a CRN. Note that (i)-(iii) are presented in [34][35].

3.2 System Model

The system architecture, channel model, elastic data traffic model, and channel access schemes under consideration are described in this section.

3.2.1 System Architecture

In this study, we consider a centralized CRN with a base station as illustrated in Fig.3.1. The secondary network operates over a time-slotted single-channel primary network, and the secondary network is an overlay network which is spectrum homogeneous. The BS and the SUs use a low-rate control channel to transmit control packets. Each SU is equipped with a single transceiver to sense the channel and transmit information packets.

3.2.2 Channel Availability Model

The channel model is similar to the one discussed in Subsection 2.2.2, in which the time is partitioned into slots of constant duration. The channel state in each time-slot is defined as 0

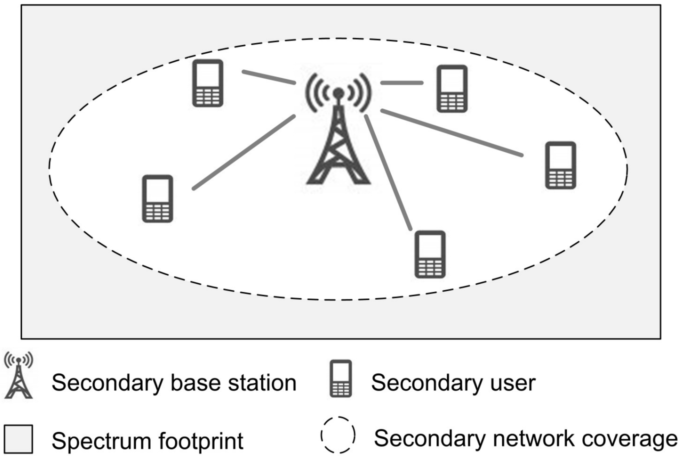


Figure 3.1: The centralized CRN with a BS.

if the channel is busy, and is 1 otherwise. The state transition of the channel among adjacent time-slots can be illustrated using a two state Markov chain as illustrated in Fig.3.2, where $S_{i,j}$ denotes the transition probability from state i ($\in \{0, 1\}$) to state j ($\in \{0, 1\}$). The channel state

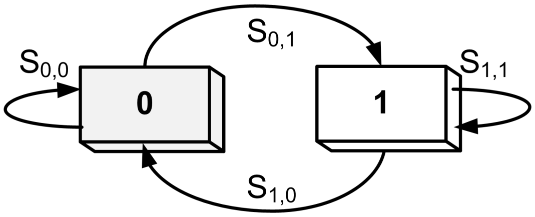


Figure 3.2: The channel state transition diagram.

can be independent or dependent among adjacent time-slots. The mean channel availability and unavailability are given by $p_1{=}S_{0,1}/(S_{0,1}{+}S_{1,0})$ and $p_0{=}S_{1,0}/(S_{0,1}{+}S_{1,0})$, respectively. The channel state is identified by an SU by spectrum sensing and the SUs transmit the sensing decision to the BS for the final channel state decision. A time-slot is mainly divided into sensing and transmission phases. The BS can transmit information packets only when the channel is at State 1 (available for SUs), and the sensing and transmissions are free of errors. The appearance of a PU and the end of an idle period (no PU is active) is denoted as service interruption to the secondary network.

3.2.3 Elastic Data Traffic Model

Web browsing nodes are considered as elastic data traffic sources in the network. The network supports a large number of data users, and the session requests arrive according to a Poisson process with mean λ. Each data session consists of one file request from an SU, and it places the request at the BS via the control channel. The time duration between two adjacent file requests from the same user are considered to be very long such that the arrival time of the second request is independent of that of the first one. After receiving the file request, the BS transfers the file to the particular SU according to a pre-assigned service discipline. During each available time-slot, only one data user is being served and the size, L_{pk}, of a data chunk (packet) transferred during a time-slot is same for all time-sots. The time duration from the time that the user places the file request at the BS until it receives the final data packet is denoted as the response time, T_R. We use the terms service request to denote a file request of an SU. The response time depends on the STR of the request. The STR depends on the length, L, of the requested data file, and L_{pk}. The lengths of requested files are independently and identically distributed with a Weibull distribution which is common for Internet data traffic [69][79]. The probability density function (PDF), $f_L(\cdot)$, of file length L is given by

$$f_L(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, \quad \alpha > 0, \beta > 0, x > 0 \tag{3.1}$$

where α and β are the shape parameter and the scale parameter, respectively. The shape parameter governs the tail heaviness of the Weibull distribution. The smaller the α the longer the tail of the STR distribution. Upon a service request, the whole data file is available at the BS without any delay. The mean response time, $E[T_R]$, of a data file is considered as the QoS parameter[1] of the elastic data traffic flows, and the QoS requirement is given by

$$E[T_R] \leq \bar{T}_{R,\max} \tag{3.2}$$

[1] The mean of the response time is considered as the service quality parameter due to the complexity of analyzing its probability distribution.

where $\bar{T}_{R,\max}$ is the maximum allowable mean response time to provide satisfactory service quality.

Service Disciplines for Data Traffic

When an active SU places a service request at the BS, the BS transmits packets of the requested file based on SPTNP, SPTWP, SRPT, and PS service disciplines. During each available time-slot, only one data user is being served.

- **SPT service discipline without preemption**: When a new service request (target request) arrives at the BS, it is served without any delay if there is no user currently being served (current user), or is placed in a waiting queue otherwise. Once the current user is being served, the request with the shortest STR in the queue will be served. If the channel becomes unavailable (interrupted) during the service of the current user, the service will be halted for the duration of the interruption and resumed after the interruption. If the target request arrives in an interruption period while there is no current request waiting to resume its service, it will be placed in the waiting queue until the channel becomes available, and the user with the lowest STR will be served. This type of interruption is referred to as an idle interruption.
- **SPT service discipline with preemption**: The target request preempts the current user if the **original** STR of the current user is larger than that of the target request. If the target request arrives in an interruption period, the request with the lowest original STR will be served after the interruption.
- **SRPT service discipline**: The target request preempts the current user if the **remaining** STR of the current user is larger than that of the target request. If the target request arrives in an interruption period, the request with the lowest remaining STR will be served after the interruption.

- **PS service discipline**: Users are served in a round-robin manner. If the channel is available in a particular time-slot, the BS transmits a data packet to the user who has the channel access right (current user) and the channel access right is given to the next user in a round-robin order for the next time-slot. However, if the channel is not available in the given time-slot, the current user keeps its channel access right until the next available time-slot. When a new service request arrives, it will be placed last at the round-robin order. In this way, each user gets a fair channel access opportunity, regardless of their service time requirements.

The service disciplines can be directly applied to a fully-connected network with distributed channel access control, where all the SUs are connected to each other by one-hop links. Each time-slot may consist of channel sensing, random contention, and data transmission periods [17]. Further research is necessary to develop an efficient distributed channel access scheme and to apply the service disciplines.

3.3 Response Time Analysis

Denote the queue in which the service requests are placed before they are served for the first time as the waiting queue. In other words a service request is placed in the waiting queue before the transmission of its first data packet. In order to transmit the first data packet of a request, the BS places the service request at the service queue and it remains at the service queue for the rest of its service. The packet transmission and channel available/unavailable durations are in discrete-time due to the time slotted nature of the primary network. However, for analysis tractability, the channel availability/busy durations, and the service time requirements are considered to be in continuous time. That is, the channel availability and busy durations are exponentially distributed with mean $1/\lambda_I$ $(= 1/S_{1,0})$ and $E[I]$ $(= 1/S_{0,1})$, respectively. Without loss of generality, the size of a data packet, L_{pk}, is considered as one unit. Therefore, the PDF of the service time requirement, τ, is $f_\tau(x) = f_L(x)$. The mean STR, $E[\tau]$ is given by $E[\tau]{=}\int_0^\infty v f_\tau(v) dv$. The cumulative distribution function (CDF) of STR τ is denoted by $F_\tau(x) = \int_0^x f_\tau(v) dv$.

3.3.1 SPT Service Discipline Without Preemption

The conditional mean response time, $E\left[T_R|\tau = t\right]$, of a target request given its STR, τ, equal to t is [68][70][71]

$$E\left[T_R|\tau = t\right] = E\left[W_t\right] + E\left[X_t\right] \tag{3.3}$$

where T_R is the response time of a target request, W_t is the waiting time of the target request from the time that the user places the service request until the BS starts transmitting the first data packet (i.e., duration that the request spend at the waiting queue), and X_t is the service time of the target request from the time that the BS starts transmitting the first data packet until it transmits the final data packet, which includes the interruption periods during the service (i.e., duration that the request spend at the service queue). As a new request arrives at the system with the STR exactly equal to that of the current user (i.e., $\tau{=}t$) occurs with a negligible probability, we neglect it for the clarity of presentation.

Categorization of Service Requests and Channel Time

Service requests and channel time are categorized based on the service time requirements and type of the request currently using the channel, respectively, as illustrated in Fig. 3.3. Define a `type t` request (`type t'` request) as a service request with original STR smaller (greater) than t [70]. `Type t` busy period is defined as a continuous time period during which `type t` requests are using the channel or being interrupted while using the channel. An illustration of `Type t` busy periods for a channel without interruptions is given in [70].

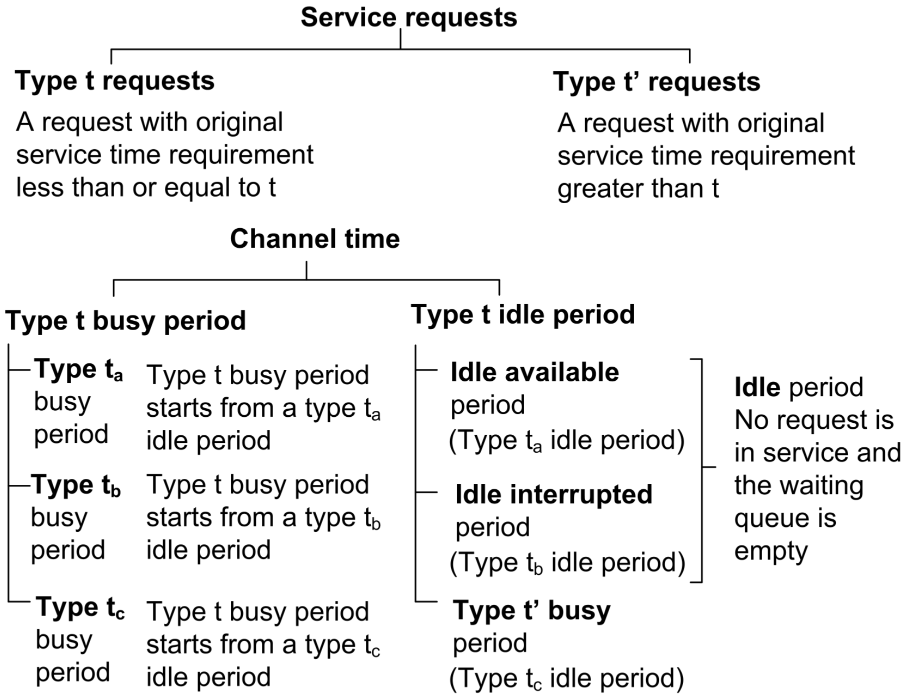

Service requests
Type t requests
A request with original service time requirement less than or equal to t
Type t' requests
A request with original service time requirement greater than t
Channel time
Type t busy period
Type t_a busy period
Type t busy period starts from a type t_a idle period
Type t_b busy period
Type t busy period starts from a type t_b idle period
Type t_c busy period
Type t busy period starts from a type t_c idle period
Type t idle period
Idle available period (Type t_a idle period)
Idle interrupted period (Type t_b idle period)
Type t' busy period (Type t_c idle period)
Idle period No request is in service and the waiting queue is empty

Figure 3.3: Categorization of service requests and channel time

If the channel is not in a `type t` busy period, it is in a `type t` idle period. A `type t` busy period starts from a request with STR less than *t* which arrives during a `type t` idle period as illustrated in Fig. 3.4, and it lasts until there is no `type t` request in the system waiting to be served. A `type t` idle period is divided into two parts, namely, `type t'` busy (`type` $\mathtt{t}_c$ idle) period and idle period. A `type t'` busy period is a continuous time period during which requests with original STR greater than *t* are using the channel or being interrupted while being served. An idle period is categorized into idle available (`type` $\mathtt{t}_a$ idle) period and idle interrupted (`type` $\mathtt{t}_b$ idle) period based on the channel availability. An idle available period is a continuous time period during which the channel is available and is not being used by any user. An idle interrupted period is an interruption period which starts from an idle available period.

A `type t` busy period is categorized into three (`type` $\mathtt{t}_a$, `type` $\mathtt{t}_b$, and `type` $\mathtt{t}_c$) busy periods based on the arrival period of the initiating type t request. The type t_a, type t_b, and type t_c busy periods initiate due to the arrival of a type t request during an idle available, idle interrupted, and type t' busy period, respectively. A type t_a busy period is initiated at the time of a type t request arrival during an idle available (type t_a idle) period. However, a type t_b busy period initiates just after the completion of an idle interruption (type t_b idle) period, and a type t_c busy period is initiated just after the completion of current type t' request (type t_c idle period). Examples for the initiation of type t_a, type t_b, and type t_c busy periods are given in Fig. 3.4. If a new service request (target request) with STR equal to t

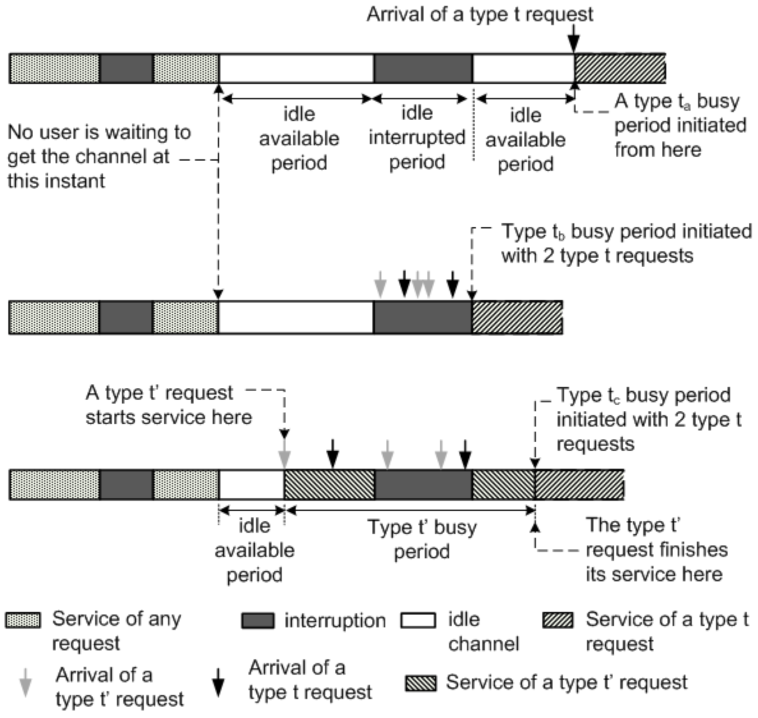

Arrival of a type t request
idle available period
idle interrupted period
idle available period
A type t$_a$ busy period initiated from here
No user is waiting to get the channel at this instant
Type t$_b$ busy period initiated with 2 type t requests
A type t' request starts service here
Type t$_c$ busy period initiated with 2 type t requests
idle available period
Type t' busy period
The type t' request finishes its service here
Service of any request
interruption
idle channel
Service of a type t request
Arrival of a type t' request
Arrival of a type t request
Service of a type t' request

Figure 3.4: Initiation of type t_a, type t_b, and type t_c busy periods.

arrives during the service time of a request (current request) with original STR smaller than t, the new service request falls into a type t busy period; otherwise, the target request falls into a type t idle period.

Target Request Arriving in a `type t` Idle Period

If a target request arrives in an idle available (`type t`$_a$ idle) period, it will get the channel access immediately. Therefore, the response time $T_R = X_t$.

If the target request arrives in an idle interrupted (`type t`$_b$ idle) period, first it waits until the interruption duration finishes. Further, any `type t` request arrivals during the idle interruption create a `type t` busy period. If so, the target request needs to wait until the end of the `type t` busy period to get the service. Therefore, the mean waiting time, $E[W_t] = E[\varphi_{idle,int}] + E[T_{busy,t_b}]P(N_b \geq 1)$, where $\varphi_{idle,int}$ is the residual time of the idle interruption, T_{busy,t_b} is the duration of a `type t`$_b$ busy period, and N_b is the number of `type t` requests arrived during an idle interrupted period.

If the target request arrives in a `type t'` busy (`type t`$_c$ idle) period, it waits first until the end of current service which is a `type t'` request. Further, any `type t` request arrivals during the service time (including the interruptions) of the current user create a `type t` busy period, and the target request needs to wait until the end of the `type t` busy period to get the channel access. Therefore, the mean waiting time, $E[W_t] = E[\varphi_{t'}] + E[T_{busy,t_c}]P(N_c \geq 1)$, where $\varphi_{t'}$ is the residual time of a `type t'` request, T_{busy,t_c} is the duration of a `type t`$_c$ busy period, and N_c is the number of `type t` requests arrived during the service time (including the interruption duration) of a `type t'` request.

Target Request Arriving in a `type t` Busy Period

If the target request arrives in a `type t` busy period, it waits until the end of the `type t` busy period to get the service. Therefore, the waiting time $E[W_t] = E[\varphi_{busy,t}]$, where $\varphi_{busy,t}$ is the residual time of the `type t` busy period. A `type t` busy period can be any of `type t`$_a$, `type t`$_b$, and `type t`$_c$ busy periods. A summary of the waiting times of a target request falling into different time periods are given in Table 3.1. The first `type t` request that arrives in an idle

Table 3.1: Waiting time of a target request arriving in different time periods

Time period	Waiting time	SPTNP	SPTWP/ SRPT
`Type t`$_a$ idle	No waiting time. Immediately receives service (accesses the channel). $E[W_t]$=0.	yes	yes
`Type t`$_b$ idle	Wait until the interruption is over. If there are any `type t` arrivals during the interruption, a `type t`$_b$ busy period is generated, wait until the end of the `type t`$_b$ busy period. $E[W_t]$=$E[\varphi_{idle,int}]$+$E[T_{busy,t_b}]P(N_b \geq 1)$.	yes	yes
`Type t`$_c$ idle	Wait until the service completion of the current (`type t'`) user. If there are any `type t` arrivals during the service of current user, a `type t`$_c$ busy period is generated, wait until the end of the `type t`$_c$ busy period. $E[W_t]$=$E[\varphi_{t'}]$+$E[T_{busy,t_c}]P(N_c \geq 1)$.	yes	no
`Type t` busy	Wait until the end of the ongoing `type t` busy period. $E[W_t]$=$E[\varphi_{busy,t}]$	yes	yes

available period initiates a `type t`$_a$ busy period. Therefore, the number of `type t` requests at the initiation of a `type t`$_a$ busy period is one. The probability of an incoming request initiating a `type t`$_a$ busy period is $P_{idle,av}F_{\tau}(t)$, where $P_{idle,av}$ is the probability of the target request arriving in an idle available period.

A `type t` request with the shortest STR that arrives in an idle interrupted period initiates a `type t`$_b$ busy period just after the interruption. However, at the initiation of the `type t`$_b$ busy period, there may be more than one `type t` request waiting to get service. As the target request has to wait until all the `type t` requests finish their service, we can treat any of the `type t` requests arriving in an idle interrupted period as the initiating request of the `type t`$_b$ busy period. The probability that an incoming service request initiates a `type t`$_b$ busy period is then $P_{idle,int}F_{\tau}(t)$, where $P_{idle,int}$ is the probability of the target request arriving in an idle interrupted period.

Similarly, the probability that a request initiates a `type t`$_c$ busy period is $P_{busy,t'}F_{\tau}(t)$, where $P_{busy,t'}$ is the probability of the target request arriving in a type t' busy period. Therefore, the probability, $P_{I,t}$, that an incoming request initiates a `type t` busy period is given by $P_{I,t}$=$F_{\tau}(t)[1 - P_{busy,t}]$, where $P_{busy,t}$ is the probability of the target request arriving in a `type t` busy period. The conditional mean response time, $E[T_R|\tau = t]$ is given by

$$\begin{aligned} E\left[T_R|\tau = t\right] &= \Big(E\left[\varphi_{idle,int}\right] + E\Big[T_{busy,t_b}\Big] P(N_b{\geq}1)\Big) P_{idle,int} + \Big(E\left[\varphi_{t'}\right] + E\Big[T_{busy,t_c}\Big] P(N_c{\geq}1)\Big) P_{busy,t'} \\ &+ E\Big[\varphi_{busy,t}\Big] P_{busy,t} + E\left[X_t\right]. \end{aligned} \tag{3.4}$$

The mean response time of a target request can be evaluated by averaging (3.4) over the PDF, $f_{\tau}(\cdot)$, of the STR.

Mean `type t` Busy Period

The duration, $T_{busy,t}$, of a `type t` busy period is given by

$$T_{busy,t} = \sum_{k=0}^{\infty} T_k \tag{3.5}$$

where $T_k = T_k' + I_k'$, $I_k' = \sum_{j=0}^{N_{I,k}} I_j$ ($k \geq 1$), $T_k' = \sum_{j=0}^{N_{y,k-1}} Y_{k-1,j}$ ($k \geq 1$), $N_{y,k-1}$ is the number of `type t` request arrivals during the period T_{k-1}, $Y_{k-1,j}$ is the STR of the j^{th} ($j \in \{0, N_{y,k-1}\}$) `type t` request arrival during T_{k-1}, $N_{I,k}$ is the number of interruptions during the period T_k', and I_j is the duration of the j^{th} ($j \in \{0, N_{I,k}\}$) interruption arrived in T_k' ($I_0, Y_{k-1,0} = 0$ by definition). The time duration T_0' in T_0 is the total STR of the $N_{y,0}$ initiating `type t` requests of the `type t` busy period. An example for a `type t` busy period is illustrated in Fig. 3.5. Similar to the analysis given in [72], it can be shown that

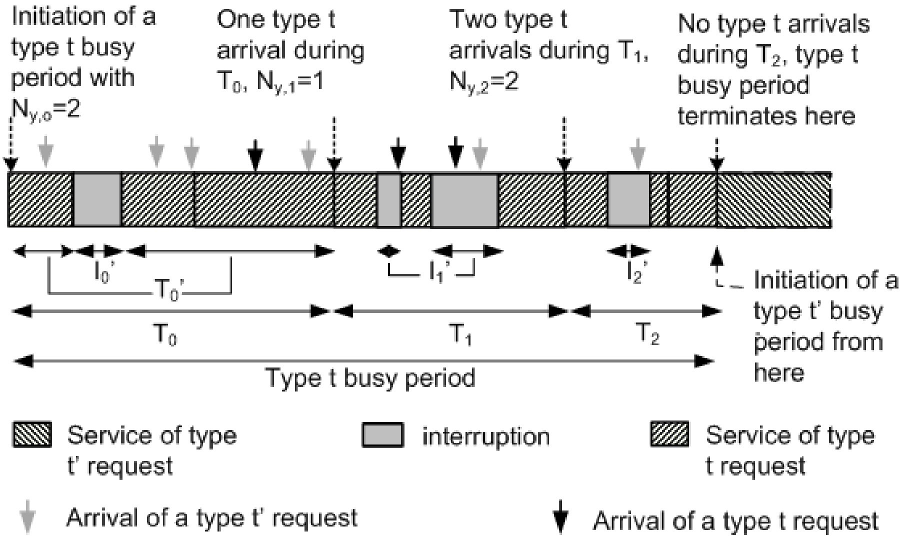

Initiation of a type t busy period with $N_{y,0}$=2
One type t arrival during T_0, $N_{y,1}$=1
Two type t arrivals during T_1, $N_{y,2}$=2
No type t arrivals during T_2, type t busy period terminates here
I_0'
T_0'
I_1'
I_2'
Initiation of a type t' busy period from here
T_0
T_1
T_2
Type t busy period
Service of type t' request
interruption
Service of type t request
Arrival of a type t' request
Arrival of a type t request

Figure 3.5: An example of a `type t` busy period initiated with $N_{y,0} = 2$.

$$E[T_{busy,t}] = E\left[\sum_{k=0}^{\infty} T_k\right] = \frac{E[T_0]}{(1 - b_t)} \tag{3.6}$$

where $E[T_0]$=$E[T_0' + \sum_{j=0}^{N_{I,0}} I_j]$, b_t=$(1 + \lambda_I E[I])\lambda_t E[Y_t]$ is the fraction of time having a `type t` busy period, λ_t is the mean arrival rate of `type t` users, and $E[Y_t]$ (= $E[\tau|\tau < t]$) is the expected value of the STR of `type t` requests. It can be shown that

$$E[T_0] = \frac{1}{p_1}E[N_{y,0}]E[Y_t], \tag{3.7}$$

where p_1=$1/(1 + \lambda_I E[I])$ is the mean channel availability. The mean waiting time of a `type t` request due to `type` t_b busy period which falls during an idle interrupted period, $E[T_{t_b}]$, can be evaluated similar to (3.6) as

$$E[T_{t_b}] = \frac{b_t}{(1 - b_t)}E[I] \tag{3.8}$$

where $\lambda_t E[I]$ is the mean number of `type t` request arrivals in an idle interrupted period. The mean duration $E[T_{t_b}]$ can be given by $E[T_{t_b}]$=$E[T_{busy,t_b}]P(N_b \geq 1)$+$0{\cdot}P(N_b = 0)$, and therefore, $E[T_{t_b}]$=$E[T_{busy,t_b}]P(N_b \geq 1)$. Further, the mean waiting time of a `type t` request due to `type` t_c busy period which falls during the service time of a `type t’` request, $E[T_{t_c}]$ (= $E[T_{busy,t_c}]P(N_c \geq 1)$), can be given similar to (3.8) as

$$E[T_{t_c}] = \frac{b_t}{(1 - b_t)}E[X_{t'}] \tag{3.9}$$

where $E[X_{t'}]$ is the mean service time of a `type t’` request including the interruption periods during the service.

Mean Residual Times

The mean residual time, $E[\varphi_{busy,t}]$, of a `type t` busy period is given by [68]

$$E[\varphi_{busy,t}] = \frac{E[T_{busy,t}{}^2|N_{y,0}\geq 1]}{2E[T_{busy,t}|N_{y,0}\geq 1]} \tag{3.10}$$

where

$$E[T_{busy,t}{}^2] = E[(\sum_{k=0}^{\infty} T_k)^2] = \frac{(1+b_t)}{(1-b_t)} E[\sum_{k=0}^{\infty} T_k{}^2]$$

The residual time of a `type t` busy period exists only if a `type t` busy period is generated. Therefore, the first and second moments of $T_{busy,t}$ are conditioned on $N_{y,0}{\geq}1$. With further manipulation, it can be shown that

$$\begin{aligned}
(1-b_t{}^2)\sum_{k=0}^{\infty} E\left[T_k{}^2\right] &= E\left[T_0{}^2\right] + \frac{E[T_0]}{(1-b_t)}\lambda_t(\frac{1}{p_1{}^2}E[Y_t{}^2] + \lambda_I E[I^2]E[Y_t]). && (3.11)\\
E[T_{busy,t}{}^2|N_{y,0}{\geq}1] &= \frac{E[T_0{}^2|N_{y,0}{\geq}1]}{(1-b_t)^2} + \frac{E[T_0|N_{y,0}{\geq}1]}{(1-b_t)^3}\lambda_t\Big\{\frac{1}{p_1{}^2}E[Y_t{}^2] + \lambda_I E[I^2]E[Y_t]\Big\} && (3.12)\\
E[T_{busy,t}|N_{y,0}{\geq}1] &= \frac{E[T_0|N_{y,0}{\geq}1]}{(1-b_t)} && (3.13)
\end{aligned}$$

where

$$\begin{aligned}
E[T_0{}^2|N_{y,0}{\geq}1] &= \frac{1}{p_1{}^2}E[T_0'^2|N_{y,0}{\geq}1] + \lambda_I E[I^2]E[T_0'|N_{y,0}{\geq}1] && (3.14)\\
E[T_0|N_{y,0}{\geq}1] &= \frac{1}{p_1}E[T_0'|N_{y,0}{\geq}1]. && (3.15)
\end{aligned}$$

As a `type t` busy period may be one of the three busy period types (`type` t_a, `type` t_b, and `type t`$_c$), the mean STR of the initiating `type t` request, $E[T_0'|N_{y,0}{\geq}1]$ is given by

$$\begin{aligned}
E[T_0'|N_{y,0}{\geq}1] &= \frac{1}{P_{I,t}}\Big\{P_{idle,av}\int_0^t \nu f_\tau(\nu)d\nu + P_{idle,int}E[N_b|N_b \geq 1]\int_0^t \nu f_\tau(\nu)d\nu/F_\tau(t) && (3.16)\\
&+ P_{busy,t'}E[N_c|N_c \geq 1]\int_t^{\infty} \nu f_\tau(\nu)d\nu/(1-F_\tau(t))\Big\}
\end{aligned}$$

where N_b and N_c are the numbers of `type t` requests at the initiation instant of `type t`$_b$ and type t$_c$ busy periods, respectively. Similar to (3.16), $E[{T_0'}^2]$ is given by

$$\begin{aligned} E[{T_0'}^2|N_{y,0}{\geq}1] &= \frac{1}{P_{I,t}}\Big\{P_{idle,av}\int_0^t v^2 f_\tau(v)dv \\ &+ P_{idle,int}\Big(E[N_b|N_b \geq 1]\int_0^t v^2 f_\tau(v)dv/F_\tau(t) \\ &+ \Big(E[{N_b}^2 - N_b|N_b \geq 1]\int_0^t v f_\tau(v)dv/F_\tau(t)\Big)^2\Big) \\ &+ P_{busy,t'}\Big(E[N_c|N_c \geq 1]\int_t^\infty v^2 f_\tau(v)dv/(1-F_\tau(t)) \\ &+ \Big(E[{N_c}^2 - N_c|N_c \geq 1]\int_t^\infty v f_\tau(v)dv/(1-F_\tau(t))\Big)^2\Big)\Big\}. \end{aligned} \tag{3.17}$$

In (3.4), the mean residual time, $E[\varphi_{idle,int}]$, of an idle interrupted period is given by $E[\varphi_{idle,int}] = E[I^2]/2E[I]$. The mean residual time, $E[\varphi_{t'}]$, of a type t' request depends on the service time of a type t' request (which includes the interruption periods during the service). The service time $X_{t'}$ is given by $X_{t'}{=}Y_{t'}{+}\sum_{j=0}^{N_{I,0}} I_j$, where $Y_{t'}$ is the original STR of a type t' request, $N_{I,0}$ is the number of interruptions occurred during T_0' $(=Y_{t'})$, and I_j is the duration of the j^{th} $(\in [0,\infty))$ interruption. The mean and the second moment of $X_{t'}$ can be given by

$$E[X_{t'}] = \frac{1}{p_1}\int_t^\infty v f_\tau(v)dv \tag{3.18}$$

$$E[{X_{t'}}^2] = \frac{1}{{p_1}^2}\int_t^\infty v^2 f_\tau(v)dv + \lambda_I E[I^2]\int_t^\infty v f_\tau(v)dv. \tag{3.19}$$

Similarly, the mean service time of a type t request is given by

$$E[X_t] = \frac{t}{p_1} \tag{3.20}$$

As there is no preemption, the mean service time does not depend on the arrivals, and it only depends on the STR and the mean channel availability. The probabilities $P_{busy,t}$, $P_{busy,t'}$, and $P_{idle,int}$ are given by $P_{busy,t} = \lambda\int_0^t v f_\tau(v)dv/p_1$, $P_{busy,t'} = \lambda\int_t^\infty v f_\tau(v)dv/p_1$, and $P_{idle,int}{=}p_0(p_1\text{-}\lambda\int_0^\infty v f_\tau(v)dv)/p_1$, where fractions (proportions) of time having a type t and type t' request occupying the channel are given by $\lambda\int_0^t v f_\tau(v)dv$ and $\lambda\int_t^\infty v f_\tau(v)dv$, respectively. The mean residual time, $E[\varphi_{t'}]$, can be evaluated from the standard equation $E[\varphi_{t'}]=E[X_{t'}{}^2]/2E[X_{t'}]$.

3.3.2 SPT Service Discipline With Preemption

The SPTWP differs from the SPTNP in that it preempts a current user to give priority to a new request with an original SRT smaller than that of the current user. Therefore, a target request with STR equal to t preempts an ongoing type t' request to get the channel access, and initiates a type t busy period. From the viewpoint of a type t user, all the interruptions of a type t idle period are idle interruptions and all the available durations of a type t idle period are idle available periods. Therefore, the channel available and interrupted periods of a type t idle period are denoted as type t idle available (type t_a idle) and type t idle interrupted (type t_b idle) periods, respectively, as illustrated in Fig. 3.6. Similar to the SPTNP, type t_a and type t_b busy periods are initiated by a type t arrival in type t_a and type t_b idle periods, respectively. However, with the SPTWP, there is no type t_c busy period. A target request with STR equal to t arriving in a type t idle available or type t idle interrupted period starts its service similar to that arrives in an idle available or idle interrupted period with the SPTNP service discipline, respectively. Examples for initiations of type t_a and type

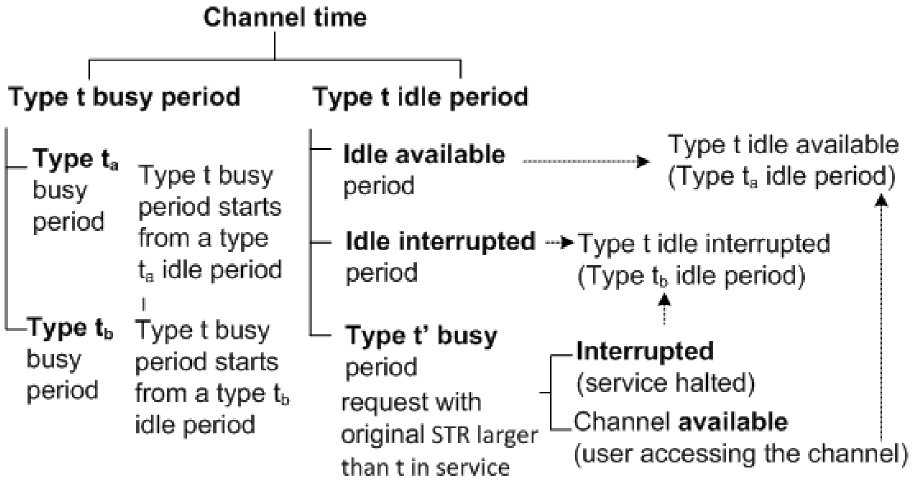

Channel time
Type t busy period
Type t idle period
Type t$_a$ busy period
Type t busy period starts from a type t$_a$ idle period
Type t$_b$ busy period
Type t busy period starts from a type t$_b$ idle period
Idle available period
Type t idle available (Type t$_a$ idle period)
Idle interrupted period
Type t idle interrupted (Type t$_b$ idle period)
Type t' busy period
request with original STR larger than t in service
Interrupted (service halted)
Channel available (user accessing the channel)

Figure 3.6: Categorization of channel time for the SPTWP service discipline

$\mathtt{t}_b$ busy periods are illustrated in Fig. 3.7. A summary of waiting times of a target request

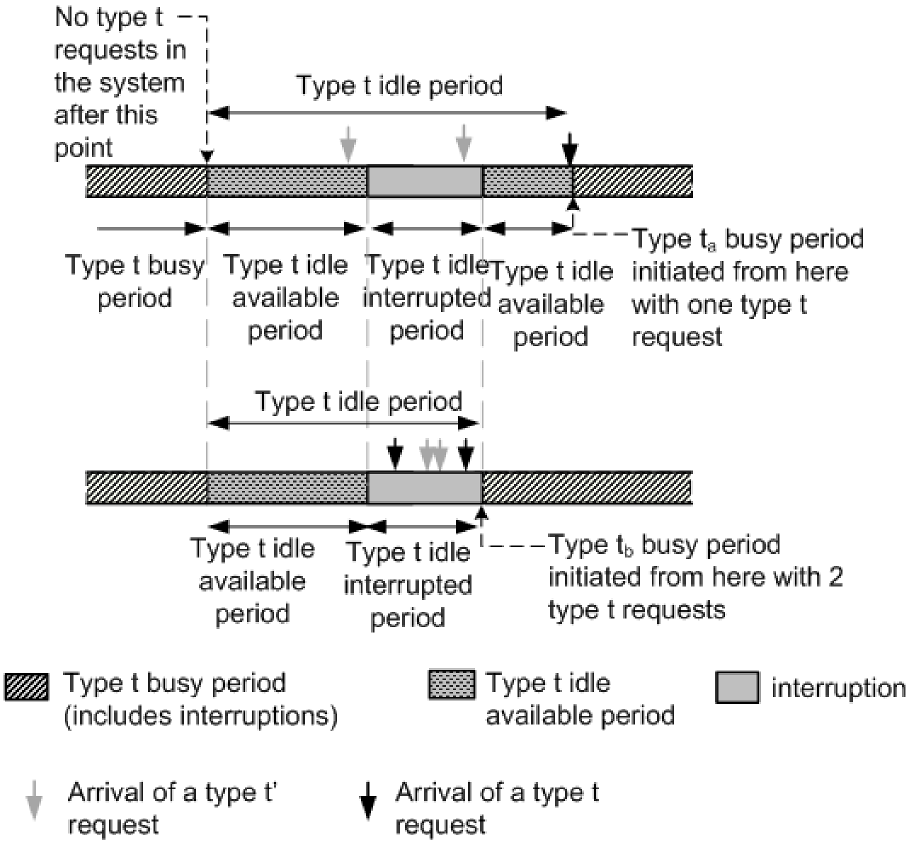

No type t requests in the system after this point
Type t idle period
Type t busy period
Type t idle available period
Type t idle interrupted period
Type t idle available period
Type t_a busy period initiated from here with one type t request
Type t idle period
Type t idle available period
Type t idle interrupted period
Type t_b busy period initiated from here with 2 type t requests
Type t busy period (includes interruptions)
Type t idle available period
interruption
Arrival of a type t' request
Arrival of a type t request

Figure 3.7: Initiation of `type t` busy periods for the SPTWP service discipline

falling into different periods is given in Table 3.1. If the target request arrives in a `type t` idle available period, it gets the channel access immediately. Therefore, the response time, T_R, is the service time, X_t. However, the service time, X_t, is different from that of the SPTNP, since any `type t` request arrival during the service time of the target request can preempt the target request. The preempted durations are the durations of `type t` busy periods initiated during the original service time of the target request. An illustration of the service time X_t is given in Fig. 3.8, where $\psi_t = t + \sum_{j=0}^{N_{I,0}} I_j$, $N_{I,0}$ is the number of interruptions occurred during $T_0' = t$ (in this example $N_{I,0}$=2), and I_j is the duration of the j^{th} interruption. Analysis of $E[X_t]$ is similar to

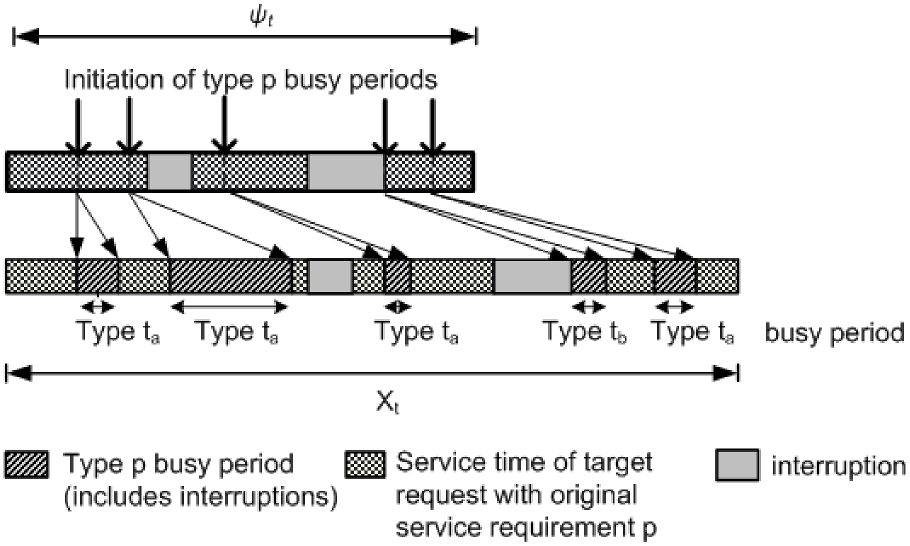

ψt
Initiation of type p busy periods
Type ta
Type ta
Type ta
Type tb
Type ta
busy period
Xt
Type p busy period (includes interruptions)
Service time of target request with original service requirement p
interruption

Figure 3.8: Service time of the target request for the SPTWP service discipline

that of $E[T_{busy,t}]$ given in (3.6) and (3.7) with $E[Y_t]$ equivalent to t and $E[N_{y,0}] = 1$. Therefore,

$$E[X_t] = \frac{t}{p_1(1 - b_t)}. \tag{3.21}$$

When the preemptions are allowed, the server (BS) stops serving a request with STR equal to t (current request) due to an arrival of a `type t` request. Therefore, the current request does not receive service during a `type t` busy period. Different from that of SPTNP the mean service time $E[X_t]$ depends on the fraction of time that a `type t` busy period is in service. Similar to a target request arrival in an idle interrupted period with the SPTNP, a target request arrival in a `type t` idle interrupted period with the SPTWP has $E[W_t] = E[\varphi_{idle,int,t}] + E[T_{busy,t_b}]P(N_b \geq 0)$, where $E[\varphi_{idle,int,t}]$ ($=E[\varphi_{idle,int}]$) is the mean residual time of the `type t` idle interruption period.

If a target request arrives in a `type t` busy period, the mean waiting time is given by $E[W_t]$=$E[\varphi_{busy,t}]$. The conditional mean response time, $E[T_R|\tau = t]$, is given by

$$E[T_R|\tau = t] = \Big(E[\varphi_{idle,int}] + E[T_{busy,t_b}]P(N_b \geq 0)\Big) P_{idle,int,t} + E[\varphi_{busy,t}]P_{busy,t} + E[X_t] \tag{3.22}$$

where $P_{idle,int,t}$is the probability of the target request arriving in a `type t` idle interrupted period, $E[\varphi_{idle,int}]= E[I^2]/2E[I]$, and $E[\varphi_{busy,t}]$ is given in (3.8).

A target request can arrive in either of the `type t` busy periods (`type t`$_a$ or `type t`$_b$). Therefore, $E[\varphi_{busy,t}]$ can be evaluated using (3.6), (3.7), (3.10)-(3.15) with

$$E[T_0{}'|N_{y,0}{\geq}1] = \frac{1}{P_{I,t}}\Big\{P_{idle,av,t}\int_0^t \nu f_\tau(\nu)d\nu + P_{idle,int,t}E[N_b|N_b \geq 1]\int_0^t \nu f_\tau(\nu)d\nu/F_\tau(t)\Big\} \quad (3.23)$$

$$\begin{aligned} E[T_0{}'^2|N_{y,0}{\geq}1] &= \frac{1}{P_{I,t}}\Big\{P_{idle,av,t}\int_0^t \nu^2 f_\tau(\nu)d\nu + P_{idle,int,t}\Big(E[N_b|N_b \geq 1]\int_0^t \nu^2 f_\tau(\nu)d\nu/F_\tau(t) \\ &+ \Big(E[N_b{}^2 - N_b|N_b \geq 1]\int_0^t \nu f_\tau(\nu)d\nu/F_\tau(t)\Big)^2\Big)\Big\} \qquad (3.24)\end{aligned}$$

where $P_{I,t}{=}(1{-}P_{busy,t})F_\tau(t)$ is the probability of a request arrival initiating a `type t` busy period and $P_{idle,av,t}$ is the probability of the target request arriving in a `type t` idle available period. Equations (3.23) and (3.24) differ from (3.16) and (3.17) in that (3.23) and (3.24) do not contain the components for a `type t`$_c$ busy period. The probabilities $P_{busy,t}$, $P_{idle,int,t}$, and $P_{idle,av,t}$ are given by

$$P_{busy,t} = \lambda\int_0^t \nu f_\tau(\nu)d\nu/p_1 \qquad (3.25)$$

$$P_{idle,av,t} = p_1 - \lambda\int_0^t \nu f_\tau(\nu)d\nu \qquad (3.26)$$

$$P_{idle,int,t} = p_0\left(p_1 - \lambda\int_0^t \nu f_\tau(\nu)d\nu\right)/p_1. \qquad (3.27)$$

3.3.3 SRPT Service Discipline

The SRPT differs from the SPTWP in that it compares the remaining STRs of the service requests rather than their original STRs. Therefore, a `type t'` request always initiates a `type t` busy period when its remaining STR reduces to t, and an incoming request with STR equal to t can preempt a `type t'` request only when the remaining STR of the `type t'` request is larger than *t*. In order to capture the difference, we alter the definition of `type t` busy period as a continuous time period during which services with the **remaining** STR **less** than *t* are using or being interrupted while using the channel. Similarly, the definition of `type t'` busy period is altered as a continuous time period during which `type t'` requests with the **remaining** STR **greater** than *t* are using or being interrupted while using the channel. The categorization of the time periods is illustrated in Fig. 3.9. As illustrated in Fig. 3.9, a `type t` busy period may be any of `type` t_a, `type` t_b, and `type` t_d busy periods. Similar to that of the SPTWP service discipline, the `type` t_a and `type` t_b busy periods are initiated due to `type t` arrivals during `type` t_a and `type` t_b idle periods, respectively. However, a `type` t_d busy period is initiated when the remaining STR of a `type t'` request becomes *t*. The waiting times of a target request with the original STR up to *t* are given in Table 3.1. Similar to that of the SPTWP, the expression for the mean response time of a target request is given by

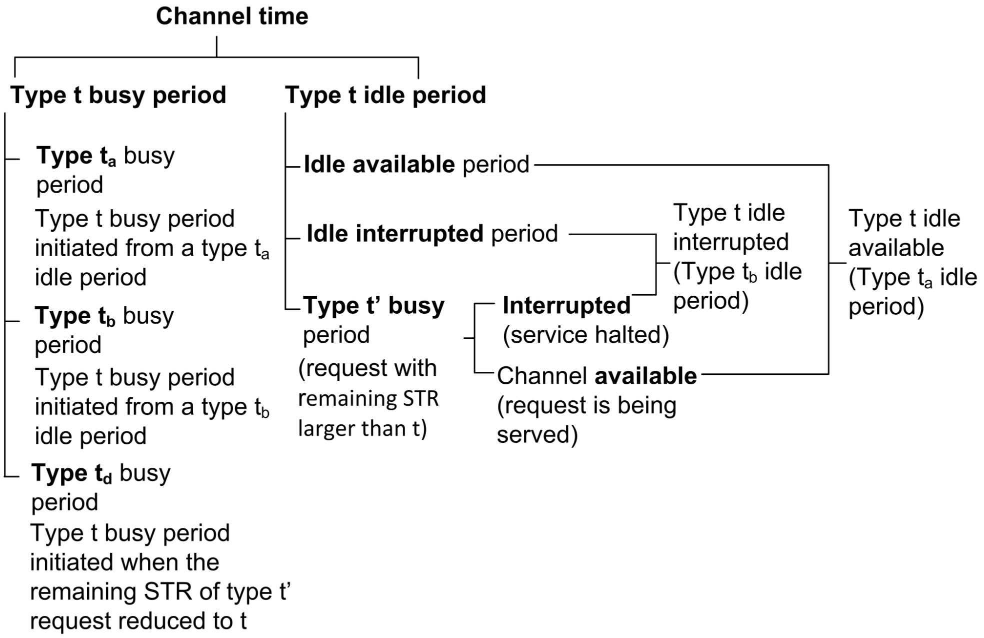

Channel time
Type t busy period
Type t idle period
Type t_a busy period
Type t busy period initiated from a type t_a idle period
Type t_b busy period
Type t busy period initiated from a type t_b idle period
Type t_d busy period
Type t busy period initiated when the remaining STR of type t' request reduced to t
Idle available period
Idle interrupted period
Type t' busy period (request with remaining STR larger than t)
Interrupted (service halted)
Channel available (request is being served)
Type t idle interrupted (Type t_b idle period)
Type t idle available (Type t_a idle period)

Figure 3.9: Categorization of channel time for the SRPT service discipline

$$E[T_R|\tau = t] = \left(E[\varphi_{idle,int,t}] + E[T_{busy,t_b}]P(N_b \geq 0)\right) P_{idle,int,t} + E[\varphi_{busy,t}]P_{busy,t} + E[X_t] \qquad (3.28)$$

where $E[\varphi_{idle,int,t}]$, $E[T_{busy,t_b}]$, and N_b are the same as those with the SPTWP. The probabilities $P_{busy,t}$ and $P_{idle,int,t}$ are given by

$$P_{busy,t} = \frac{\lambda\left[\int_0^t \nu f_\tau(\nu)d\nu + t\,(1 - F_\tau(t))\right]}{p_1},\; P_{idle,int,t} = p_0\left(1 - P_{busy,t}\right) \tag{3.29}$$

where the numerator in $P_{busy,t}$ is the fraction of time that requests with the remaining STR less than t occupies the channel (excluding the interruption durations). The evaluation of $E[\varphi_{busy,t}]$ in (3.28) is similar to that given in (3.16) with

$$\begin{aligned} E[T_0'|N_{y,0}{\geq}1] &= \frac{1}{P_{I,t}}\Big\{P_{idle,av,t}\int_0^t \nu f_\tau(\nu)d\nu + P_{idle,int,t}E[N_b|N_b \geq 1]\int_0^t \nu f_\tau(\nu)d\nu/F_\tau(t) \\ &+ t[1 - F_\tau(t)]\Big\} \qquad (3.30) \\ E[T_0'^2|N_{y,0}{\geq}1] &= \frac{1}{P_{I,t}}\Big\{P_{idle,av,t}\int_0^t \nu^2 f_\tau(\nu)d\nu + P_{idle,int,t}\Big(E[N_b|N_b \geq 1]\int_0^t \nu^2 f_\tau(\nu)d\nu/F_\tau(t)\Big) \\ &+ P_{idle,int,t}\Big(E[{N_b}^2 - N_b|N_b \geq 1]\int_0^t \nu f_\tau(\nu)d\nu/F_\tau(t)\Big)^2 + t^2\,(1 - F_\tau(t))\Big\} \qquad (3.31) \end{aligned}$$

where $P_{I,t}$ is the probability of an incoming request initiating a `type t` (`type t`$_a$, `type t`$_b$, or `type t`$_d$) busy period. Therefore, $P_{I,t}{=}P_{idle,av,t}F_\tau(t){+}P_{idle,int,t}F_\tau(t){+}1 - F_\tau(t){=}1 - P_{busy,t}F_\tau(t)$.

In the case of SRPT, a new service request can preempt the current request only if the STR of the new request is less than the remaining STR of the current request at the arrival instant. Therefore, the service time comparison has to be done exactly at the arrival instant of the new request. This comparison is not possible in continuous-time as the probability that an arrival occurs at a particular time instance is zero. It is only possible to find the probability of request arrivals with the original STR shorter than the remaining service time of the current request for a given period of time. As a result, we divide the service time requirement t (or equivalently the file length) of the target request into n units of duration Δt ($t = n{\cdot}\Delta t$) as illustrated in Fig. 3.10, where a `type` $(n-i)\Delta t$ busy period is similar to a `type t` busy period which starts from a `type` $(n-i)\Delta t$ request and ends after serving all such requests in the waiting queue, and a `type` $(n-i)\Delta t$ request being a service request with the original STR less than $(n - i)\Delta t$. The service time $X_{\Delta t,i}$

($i \in \{1, 2, ..., n\}$) is the actual duration it takes to complete the i^{th} unit of Δt, and X_t=$\sum_{i=1}^{n} X_{\Delta t,i}$ [73]. We have $E[X_t]$=$\sum_{i=1}^{n} E[X_{\Delta t,i}]$. The duration Δt in $X_{\Delta t,i}$ is equivalent to $T_0{}'$ in $T_{busy,t}$. Therefore, similar to the evaluation of $E[T_{busy,t}]$, $E[X_{\Delta t,i}]$ is given by $E[X_{\Delta t,i}] = \lim_{m\to\infty} \sum_{k=0}^{m+1} T_{i,k}$ with

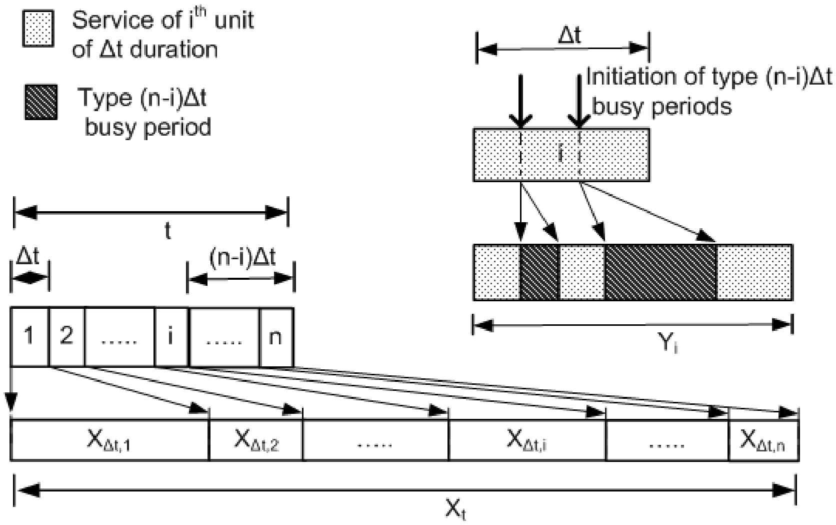


Figure 3.10: Service time for the SRPT service discipline

$$E[T_{i,k+1}|T_{i,k}] = \frac{1}{p_1}(\lambda_{(n-i)\Delta t} T_{i,k}) \frac{1}{f_\tau((n-i)\Delta t)} \int_0^{(n-i)\Delta t} v f_\tau(v) dv \tag{3.32}$$

where $\lambda_{(n-i)\Delta t} T_{i,k}$ represents the mean number of request arrivals with the STR less than $(n-i)\Delta t$ during T_k, and the integral represents the mean STR of such arrival. Similar to $E[T_{busy,t}]$, the mean duration $E[X_{\Delta t,i}]$=$E[T_0]/(1-b_{(n-i)\Delta t})$, where $E[T_0]$=$\Delta t/p_1$ and $b_{(n-i)\Delta t}$=$\lambda_{(n-i)\Delta t} \int_0^{(n-i)\Delta t} v f_\tau(v) dv/p_1$ $F_\tau((n-i)\Delta t)$. In order to make the analysis accurate (close to that for the continuous-time scenario), Δt has to be very small (i.e., n is very large). As our original system is time-slotted, we can carry out the analysis in discrete-time with $\Delta t = 1$ time unit, which is the size of a time-slot. However, since our analysis so far has been in continuous-time, we divide the time into infinitely small (i.e., a large number of) time periods. Therefore, the service time, $E[X_t]$, is obtained by making Δt very small ($\Delta t\to 0$) or equivalently n very large ($n\to\infty$) [2].

[2] In our analysis, we set $n = 10^4$.

Duration of `type t` busy periods are independent and identically distributed with SPTWP and SRPT service disciplines, and the inter-arrival time of the new requests are memory-less. In the analysis of residual time of `type t` busy period given that the target request arrives in a `type t` busy period, we can ignore the `type t` idle periods, and consider the initiation of a `type t` busy period as a renewal process. However, durations of `type t'` services and `type t` busy periods can be weakly dependent in the case of SPTNP service discipline. The dependence vanishes with the occurrence of an idle available period (when the server becomes idle). Therefore, we assume these durations to be independent and evaluate the residual times similar to the cases of SPTWP and SRPT.

3.3.4 PS Service Discipline

The conditional mean response time of a target request operating over a network following exponentially distributed channel availability durations is given by [74]

$$E[T_R|\tau = t] = \frac{t}{p_1(1-\sigma)} + p_0\frac{E[I^2]}{2E[I]} + p_0\frac{\sigma E[I^2]}{2E[I]}\frac{2-\sigma}{(1-\sigma)^2}(1 - e^{-\frac{(1-\sigma)t}{E[\tau]}}) \tag{3.33}$$

where $\sigma = \lambda E[\tau]/p_1$ is the utilization factor (ratio between the mean arrival rate and mean service rate).

3.4 Comparison of the Service Disciplines

For presentation clarity, we use subscripts NP, WP, and SRPT for the components associated with SPTNP, SPTWP, and SRPT service disciplines, respectively. The difference in the conditional mean response times between SPTNP and SPTWP is given by

$$\begin{aligned} E[T_R|\tau = t]_{NP} - E[T_R|\tau = t]_{WP} &= (\xi_{t'} - p_0\xi_{t_b})(\sigma - b_t) + b_t(E[\varphi_{busy,t}]_{NP} - E[\varphi_{busy,t}]_{WP}) \\ &- \frac{t}{p_1} \cdot \frac{b_t}{(1 - b_t)} \end{aligned} \quad (3.34)$$

where $\xi_{t'}$=$E[\varphi_{t'}] + E[T_{busy,t_c}]P(N_c \geq 1)$, ξ_{t_b}=$E[\varphi_{idle,int}]$ +$E[T_{busy,t_b}]P(N_b \geq 1)$, and $\frac{t}{p_1} \cdot \frac{b_t}{(1-b_t)}$ is the service time difference between SPTNP and SPTWP for a data file requiring a service time equal to t. The terms $(\xi_{t'} - p_0\xi_{t_b})$ and $(E[\varphi_{busy,t}]_{NP} - E[\varphi_{busy,t}]_{WP})$ contain busy periods initiated from more than one request arrival, whereas $\frac{t}{p_1} \cdot \frac{b_t}{(1-b_t)}$ only contains the service time of a target request. The probability b_t monotonically increases with t. The terms ($\xi_{t'}$-$p_0\xi_{t_b}$), $(\sigma - b_t)$, and $E[\varphi_{busy,t}]_{NP} - E[\varphi_{busy,t}]_{WP}$ decreases with t, and $\frac{t}{p_1} \cdot \frac{b_t}{(1-b_t)}$ increases with t. Therefore, the difference in the conditional mean response times $E[T_R|\tau = t]_{NP} - E[T_R|\tau = t]_{WP}$ varies from a very high positive value to a small negative value as t increases. When the file size Weibull (heavy tail) distributed as in (3.1), the probability of having a smaller file size is high and that of a larger file size is low. Therefore, the resultant mean response time difference $E[T_R]_{NP} - E[T_R]_{WP}$ obtained by averaging $E[T_R|\tau = t]_{NP} - E[T_R|\tau = t]_{WP}$ over t is a positive value. The probability of having very large and very small values for $E[T_R|\tau = t]_{NP} - E[T_R|\tau = t]_{WP}$ increases with the tail heaviness in the file size distribution. As a result, $E[T_R]_{NP} - E[T_R]_{WP}$ increases with the tail heaviness in the file size distribution. Increment of the terms $(\xi_{t'} - p_0\xi_{t_b})$ and $(E[\varphi_{busy,t}]_{NP} - E[\varphi_{busy,t}]_{WP})$ with an increasing interruption duration is larger than that of $\frac{t}{p_1} \cdot \frac{b_t}{(1-b_t)}$. Therefore, the mean response time difference $E[T_R]_{NP} - E[T_R]_{WP}$ increases with the interruption duration. The difference between the conditional mean response times for SPTWP and SRPT is given by

$$\begin{aligned} E[T_R|\tau = t]_{WP} - E[T_R|\tau = t]_{SRPT} &= E[X_t]_{WP} - E[X_t]_{SRPT} + b_t^* p_1 E[T_{t_b}] \\ &- b_t(E[\varphi_{busy,t}]_{SRPT} - E[\varphi_{busy,t}]_{WP}) - b_t^*(E[\varphi_{busy,t}]_{SRPT} \\ &- p_1 E[\varphi_{idle,int}]) \end{aligned} \quad (3.35)$$

where b_t^*=$t(1 - F_\tau(t))/p_1$. The mean service time $E[X_t]_{WP}$ is greater than $E[X_t]_{SRPT}$, due to the larger number of preemptions in SPTWP than SRPT, and the difference $E[X_t]_{WP} - E[X_t]_{SRPT}$ increases with t. The terms $E[T_{t_b}]$ and $(E[\varphi_{busy,t}]_{SRPT} - E[\varphi_{busy,t}]_{WP})$ are smaller positive values and $(E[\varphi_{busy,t}]_{SRPT} - p_1 E[\varphi_{idle,int}])$ is negative for a smaller t value. However, all three terms are larger positive values for a larger t. Therefore, difference in conditional mean response times $E[T_R|\tau = t]_{WP} - E[T_R|\tau = t]_{SRPT}$ varies from a small positive value to a small negative value with increasing t. Similar to (3.34), the mean response time difference $E[T_R]_{WP} - E[T_R]_{SRPT}$ is a positive value when the file lengths are heavy tail distributed. However, this positive value is smaller than that in (3.34). The difference between the conditional mean response times for SPTNP and SRPT is given by

$$\begin{aligned} E[T_R|\tau = t]_{NP} - E[T_R|\tau = t]_{SRPT} &= (\xi_{t'} - p_0\xi_{t_b})(\sigma - b_t) + b_t(E[\varphi_{busy,t}]_{NP} \\ &- E[\varphi_{busy,t}]_{SRPT}) + E[X_t]_{NP} - E[X_t]_{SRPT}. \end{aligned} \tag{3.36}$$

Similar to the discussion on (3.34), the terms $(\xi_{t'} - p_0\xi_{t_b})$ and $(E[\varphi_{busy,t}]_{NP} - E[\varphi_{busy,t}]_{SRPT})$ contain busy periods initiated using one or more request arrival, whereas $E[X_t]_{SRPT} - E[X_t]_{NP}$ only contains the service time of a target request. Further, the terms ($\xi_{t'}$-$p_0\xi_{t_b}$), $(\sigma - b_t)$, and $E[\varphi_{busy,t}]_{NP} - E[\varphi_{busy,t}]_{SRPT}$ decrease, and $E[X_t]_{SRPT} - E[X_t]_{NP}$ increases with t. Therefore, the difference in conditional mean response times $E[T_R|\tau = t]_{NP} - E[T_R|\tau = t]_{SRPT}$ varies from a very high positive value to a very small value with increasing t, and the unconditional mean response time difference $E[T_R]_{NP} - E[T_R]_{SRPT}$ is a positive value when the file length is heavy tail distributed. Further, the difference $E[T_R]_{NP} - E[T_R]_{SRPT}$, increases with the interruption duration and the tail heaviness of the file length distribution.

3.5 Numerical Results

Computer simulations are carried out to evaluate the accuracy of the response time analysis. As the system is time-slotted, the simulations are in discrete time and the time is measured in time-slot units. Therefore, the STR of a service request is measured in number of time-slots. The BS transmits packets to the SUs in idle time-slots (which are not being used by the PUs) based on four service disciplines, respectively. The BS transmits only one packet in each idle time-slot. Service requests are generated according to a Poisson arrival process with a Weibull distributed file length. The mean response time, $E[T_R]$, is evaluated by averaging the results of 20 simulation runs, each run having 18,000 service requests. Note that, the mean channel available and unavailable (interruption) durations are denoted as T_{on} and T_{off} throughout this section.

Fig. 3.11 shows the variation of $E[T_R]$ with T_{on} and T_{off} obtained from numerical analysis and simulations while having $T_{off} = 10$ and $T_{on} = 10$ time-slots, respectively for all four service disciplines in a light traffic load condition. We keep $\sigma = 0.6$ and $\alpha = 0.6$. The results demonstrate that the simulation results closely match with the analytical results for all four service disciplines. When preemption is allowed, the mean response time decreases considerably, and using the remaining STR instead of the original STR improves the performance. The PS outperforms the SPTNP service discipline for the lightly loaded system. The mean response time decreases exponentially with the channel availability and increases with the mean interruption duration.

Fig. 3.12 shows the variation of $E[T_R]$ with σ obtained from numerical analysis and simulations for all four service disciplines, with $T_{on} = 20$, $T_{off} = 10$ time-slots, $E[L] = 500$, and $\alpha = 0.6$. It is observed that the response times of all four service disciplines increase with σ, and the larger the σ, the larger the rate of increment of $E[T_R]$. As the mean service rate remains constant, the mean arrival rate is proportional to the offered traffic load, and the larger

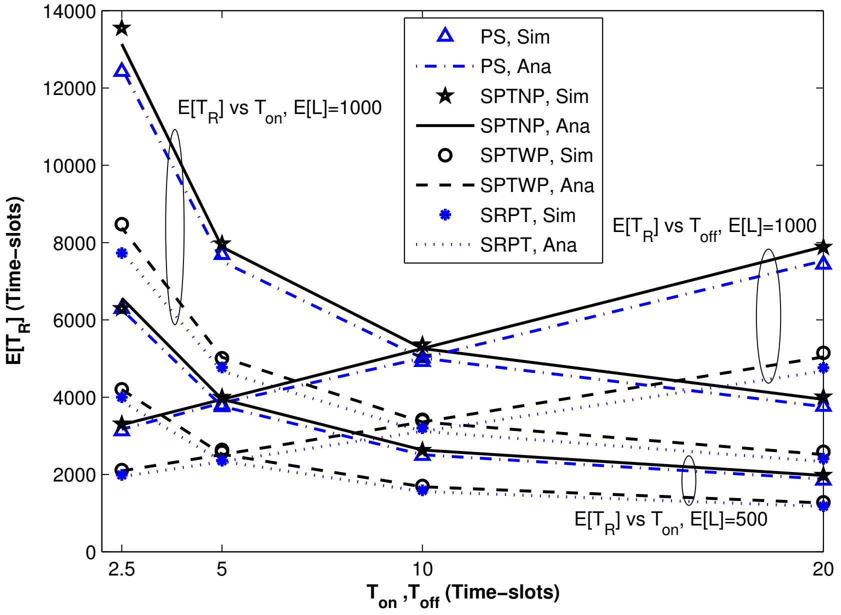


Figure 3.11: The variation of mean response time with mean channel availability and interruption durations at $\sigma = 0.6$.

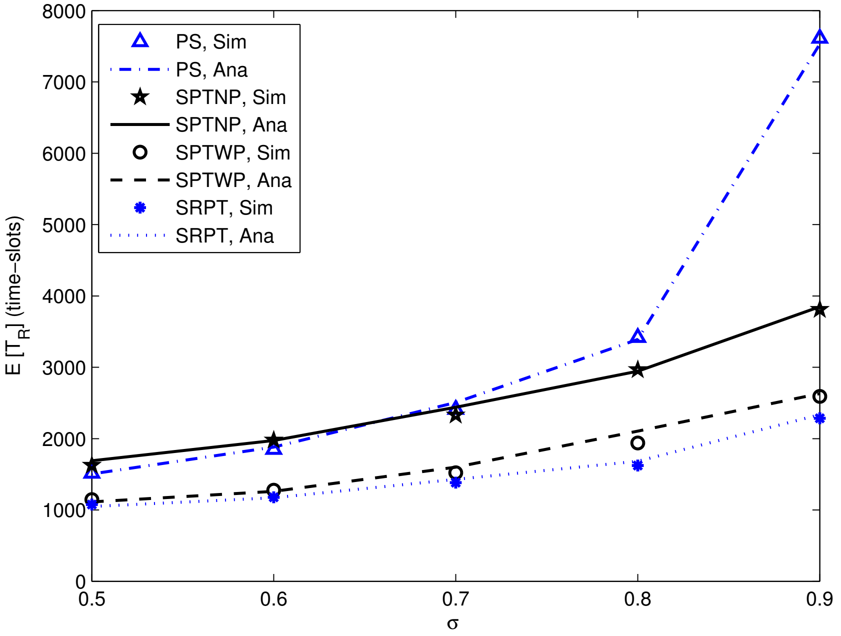


Figure 3.12: The variation of mean response time with the traffic load.

the arrival rate, the longer the waiting time of the users at the waiting queue. Therefore, the waiting time increases with the traffic load at the BS, leading to longer response times. As seen in Fig. 3.11, the service disciplines with preemption outperforms that without preemption, and the PS outperforms the SPTNP in a lightly loaded condition. For the PS service discipline, the heavier the traffic load, the larger the number of users in the round-robin order. Therefore, the mean service time increases for each request; whereas for the SPTNP service discipline, the increasing number of requests has a major impact on the waiting time (or the response time) of the requests with a long STR, and vice versa. However, the probability of request arrivals with a long STR is small. Therefore, the rate of increment of the mean response time with the traffic load is larger for the PS service discipline than that for the SPTNP service discipline. This rapid increment is indeed captured in (3.33).

Fig. 3.13 shows the $E[T_R]$ variation with T_{off} for the SRPT service discipline obtained from numerical analysis and simulations for two different traffic load conditions and α values with $T_{on} = 100$ time-slots and $E[L] = 500$. The simulation results closely match with the analytical results. That is, the discrete-time analysis in Subsection 3.3.3 is accurate for the networking scenario. Similar to what is observed in Fig. 3.12, the $E[T_R]$ increases with σ. Further, the heavier the tail of the STR distribution, the shorter the mean response time.

Fig. 3.14 shows the $E[T_R]$ variation with the tail heaviness (shape parameter) of the file length (or the STR) with $\sigma = 0.6$, $T_{on} = 20$, and $T_{off} = 10$ time slots.

The mean response time of the PS service discipline remains almost the same with the variation of α. As the PS gives an equal opportunity to all service requests, the $E[T_R]$ depends on the mean of the STR, not its distribution [74]. When the preemption is allowed, the mean response time decreases with the heaviness of the tail; otherwise, it increases with the heaviness of the tail. Preemptions result in shorter response times for requests with a short STR and longer response times for requests with a long STR. The smaller the α, the larger the number of service requests with very short STR. Therefore, when the preemption is allowed, the heavier the tail

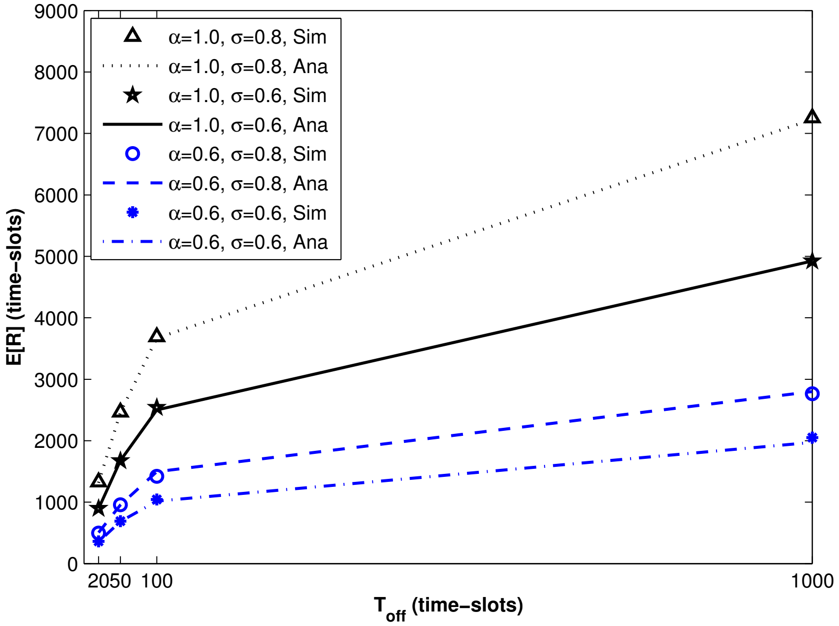


Figure 3.13: The variation of mean response time with the mean interruption duration for the SRPT service discipline

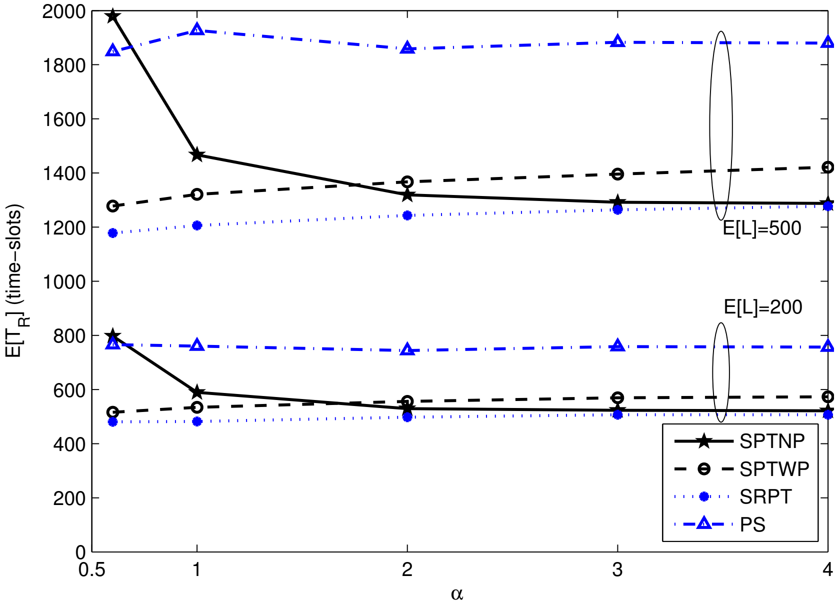


Figure 3.14: The variation of mean response time with the tail heaviness of the STR distribution.

of the STR distribution (a lower α), the shorter the $E[T_R]$. When α is large, the STR concentrate around its mean, and the probability of an incoming request having a shorter original STR than the remaining STR of the current user is very low. This reduces the probability of having preemptions by a large margin in the case of the SRPT. When the preemptions are not likely to happen, the SRPT is similar to the SPTNP. Therefore, the larger the α, the closer the $E[T_R]$ of the SPTNP to that of the SRPT. In case of the SPTWP, the preemptions are carried out with respect to the original STR of both the incoming and existing requests. When the STR concentrate around its mean, the remaining STR of most of the ongoing services are less than the original STR of the incoming service requests. Therefore, the probability of the incoming service request having a shorter original STR than that of the ongoing service is larger than the probability of incoming service request having a shorter original STR than the remaining STR of the ongoing service. As a result, the number of preemptions are larger with the SPTWP than that with the SRPT. During most of the preemptions, the ongoing services with shorter remaining STR are preempted by the service requests with long original STR. Therefore, the preemotions increase the waiting times, leading to longer mean response times in the case of the SPTWP than that in the case with the SPTNP and SRPT. Fig. 3.14 demonstrates that the heavier the tail of the STR distribution, the longer the mean response time, and the larger the rate of increment of the mean STR for the SPTNP service discipline. Therefore, the SPTNP service discipline is not a good choice when the tail of the STR distribution is very heavy.

Fig. 3.15 shows the $E[T_R]$ variation with T_{off} at $p_1 = T_{on}/(T_{on} + T_{off})$=0.66 with $E[L] = 500$ and $\sigma = 0.6$ for two different α values. The $E[T_R]$ increases with T_{off} even when the mean (long term) channel availability and the traffic load remain unchanged. When the interruption duration is exponentially distributed, the conditional mean response time for the SPTWP service discipline given in (3.22) can be simplified to $E[T_R|\tau = t]$=$T_{off}p_0$+$E[\varphi_{busy,t}]b$+$\frac{t}{p_1(1-b_t)}$. Note that $E[\varphi_{busy,t}]$ increases and b_t remains constant with T_{off} for constant p_1 and σ. As a result, the longer the T_{off} the longer the $E[T_R|\tau = t]$. Similarly, we can show that the conditional mean

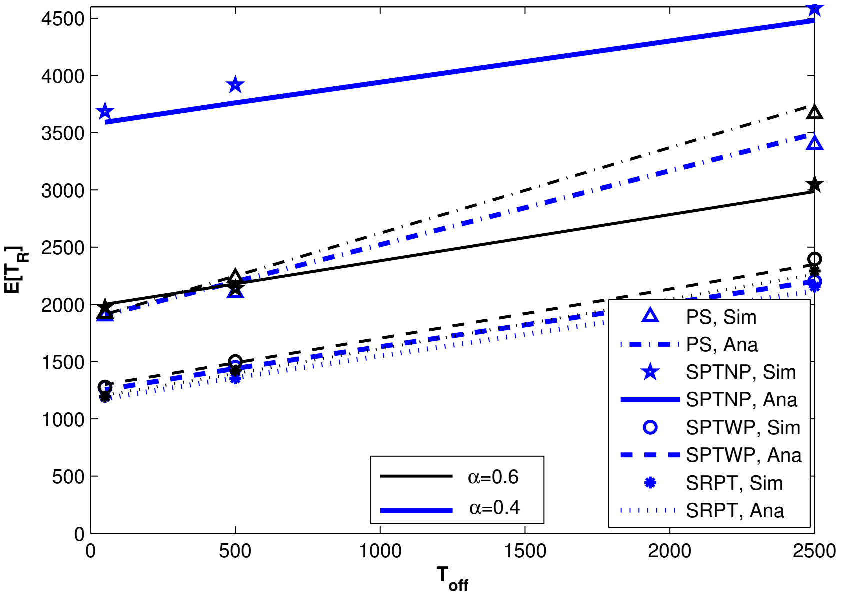


Figure 3.15: The variation of mean response time with the mean channel unavailability duration for constant p_1.

response times of the SPTNP, SRPT, and PS service disciplines increase with the mean interruption duration, when the long term channel availability and the traffic load remain unchanged. Similar to Fig. 3.14, the shorter the tail of the STR distribution (larger α), the shorter the $E[T_R]$ for the SPTNP and the longer the $E[T_R]$ for the rest of the service disciplines. As in (3.34) and (3.36), the difference between the mean response times of the SPTNP and SPTWP and that between the SPTNP and SRPT increase with T_{off}. However, there is no significant difference in the gap between the mean response times of the SPTWP and SRPT with the variation of T_{off}.

Fig. 3.16 shows the CDFs of the file length L (or STR) and the response times obtained from simulations for the SPTNP and SPTWP service disciplines, respectively, with $E[L] = 500$, $T_{on} = 10$ and $T_{off} = 20$ time-slots, and $\sigma = 0.6$. It is observed that a higher percentage of requests show a very short response times when the preemption is allowed. Further, the slope of the response time curve reduces for a high STR (file length) value with the SPTWP than

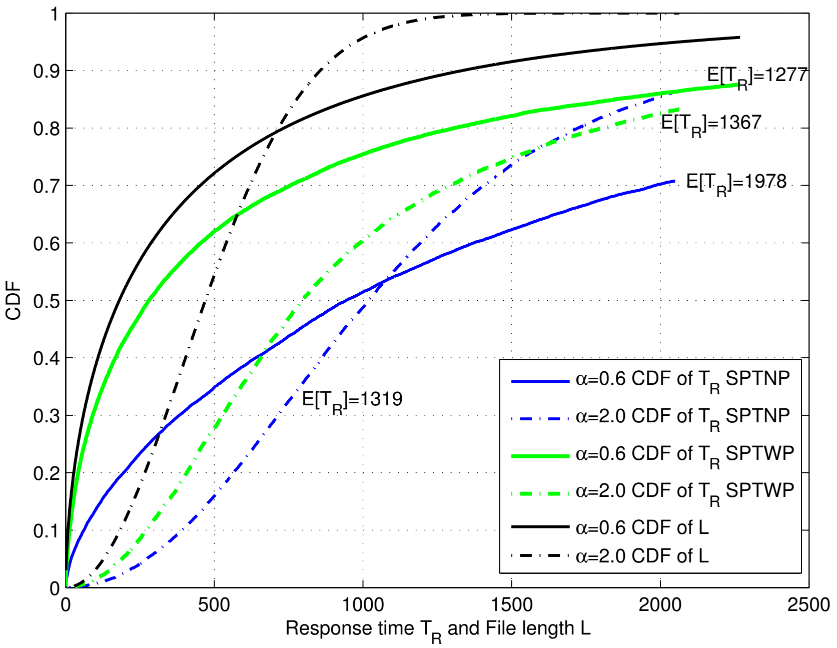


Figure 3.16: The CDF of mean response times of both SPT service disciplines and the STR.

that with the SPTNP. This means that the response time for a long STR is getting longer when the preemption is allowed. Preemption gives a higher service priority to requests with a short STR over requests with a long STR. Therefore, the requests with a short STR experience a very short response time and the requests with a long STR experience a very long response time. That is, the preemption compromises the performance of requests with a long STR to other requests with a short STR. This may not be fair in the viewpoint of the requests with a long STR. If we try to give a higher priority to the requests with a long STR over the requests with a short STR, the latter will have to wait for a very long time before getting the service, and the resulting mean response time will be longer. However, the requests with a long STR will get a shorter response time than that in the case with SPR and SRPT disciplines. Based on our analysis, the mean response time can be evaluated for a system, given the channel availability statistics, STR (file length) distribution, and request arrival rate. The best service discipline can then be selected based on the mean response time requirement of the data service service and desired trade-off with service fairness.

In this work, the mean response time is studied with respect to a time-slotted cognitive radio network. As described in Section 3.3, we conduct the analysis in continuous time, considering the fact that the interested time durations (STR, T_{on}, and T_{off}) are larger when compared with the time-slot duration. Given that an accurate channel sensing can be done, this analysis can be directly applied to the continuous-time scenario (i.e., when the channel is not time-slotted). Further, in this work, we consider only the single channel scenario. In a multiple channel network, two key approaches can be considered:

1. The BS divide incoming requests among the channels based on the arrival sequence, and SUs stay in the assigned channel for the service;
2. The BS assigns a channel to SUs instantaneously based on the channel availability in each time-slot.

The two approaches differ in signal/channel-switching overhead and in statistical multiplexing performance gain. The model here can be applied to the first scenario where the arrival rate should be normalized to the number of channels in the network. Extension to the second scenario requires further research. This analysis can be used as a benchmark for the performance in the second scenario. Further, this work is focused on the operation of a CRN with single base station. In the case of multiple base stations, data call handover between neighboring base stations should be considered. The file length (or equivalently the STR) distribution and the arrival process for each base station is a combination of those of the new request arrivals and the handover data calls. This can be treated in a way similar to that in [29][30][69] in which the authors consider handover between a cellular network and a WLAN. In extending the analysis to a system with multiple base stations, different file size distributions and different arrival processes (for new and handover calls) should be considered, which is expected to be much more complex.

3.6 Summary

In this chapter, we have evaluated the mean response time of elastic data traffic under three service disciplines (shortest processor time without preemption, shortest processor time with preemption, and shortest remaining processing time) for a single-channel single-hop synchronized CRN with a base station, in comparison with the PS service discipline. It is shown that the analytical results match well with simulation results. Numerical results demonstrate that the SRPT and the SPTWP provide very close response times and the SRPT outperforms the SPTWP. The mean response times for all four service disciplines are compared under different traffic load conditions, and it is shown that the SPTNP service discipline outperforms the PS service discipline in heavy traffic load conditions. Therefore, the SPTNP service discipline is a better choice over the PS service discipline as the traffic load increases. The mean response times of all four service disciplines are compared under the Weibull distribution with different tail parameters, and the results show that the preemption reduces the mean response time when the service time requirement follows a heavy tailed distribution. The SRPT performs better than the other service disciplines in terms of mean response time, as it achieves very short response times for service requests with short service time requirements. This mean response time analysis can be used to choose the best service discipline based on the given system parameters and the service quality requirement of the users. Further, the mean duration of the transmission interruptions (channel non-available durations) has an impact on the mean response time even when the long term channel availability and the offered traffic load at the BS remain unchanged. The mean response time analysis can be used for call admission control to ensure service satisfaction.

Chapter 4

Service Response Time of Interactive Data Traffic

4.1 Introduction

So far, we have discussed the mean response time of data sessions with only single file request (service request). After completing the reception of one file, the user leaves the system, or places the next service request after a long time compared to the inter-arrival time of the new sessions at the BS. Therefore, the arrival times of two successive file requests from the same user are independent. However, a data user (such as one browsing the web), may place the next service request just after reading/viewing the first file. From the viewpoint of the BS, the time duration from the completion of one file transfer to the reception of the next file request from the same user is not very large. A user may place multiple such requests one after another before taking a break from browsing the web. Therefore, a data session from a particular user may contain multiple file requests with a thinking (file reading/viewing) time between the completion of one service request to the placement of next service request. An interactive data session shows an on-off behavior. From the viewpoint of a secondary (data) user, an on state begins with placement of a service request at the BS, and it continues until the completion of the particular file reception, and an off state begins with the completion of a file reception, and it continues until the placement of the next service request. The service is referred to as interactive data sessions since the users place their service requests in an interactive manner. The mean response time evaluation of interactive data traffic is studied for cellular/WLAN integrated networks in [29][30]. In the studies, the network supports streaming and elastic data traffic flows. The shortest remaining processor time and processor sharing service disciplines are considered in [29] and [30], respectively. The mean response time of interactive data traffic over shared packet networks is studied in [75]. In [75] the Web traffic is considered as interactive data traffic, and an EDGE (Enhanced Data Rates for GSM Evolution) network is considered as the packet network. However, the mean response times of interactive data traffic over CRNs with different service disciplines are not studied in the previous works.

Motivation and Objectives

Interactive data traffic falls into the non-real-time traffic category, and the mean response time of a data file (service request) is an important parameter in measuring the service quality. As discussed in Section 3.1, in a cognitive radio network, the service interruptions due to the presence of PUs increase the response time. The requested files are of different lengths (service time requirements), and the response times of a service request depend on the service time requirement and the service discipline. The number of file requests in a data session can increase the offered traffic load at the BS, which result in longer response times. Longer response times can frustrate the users. Different service disciplines can be used to provide different priorities to the service requests based on their service time requirements (file lengths), which result in different response times. Therefore, it is important to keep the mean response time down, and studying the relationship between different system parameters (channel available statistics, service disciplines, service request arrival process, thinking time, and service time requirement) and the mean response time is important in choosing appropriate service discipline to provide the required service quality. This relationship can be used in developing CAC algorithms for the interactive data traffic flows. The objective of this chapter is to study the mean response time of interactive data traffic over CRNs with different service disciplines and to compare them under different system conditions. As we have already studied the mean response time of elastic data traffic, we investigate the relationship between the mean response times of elastic and interactive data traffic flows.

Contributions

The contribution of this chapter is two fold: (i) We show that the request arrival process at the BS can be approximated by the Poisson process for all four service disciplines, when the mean channel unavailable duration is comparatively long and the offered traffic load is low. Further, the larger the mean number of files (file requests) per data session, the closer the request arrival process to the Poisson process. However, the request arrival process at the BS deviates from the Poisson process for the SPTNP service discipline when the STR is heavytail distributed; (ii) Under the Poisson approximation, we show that the mean response time of a file in an interactive data session can be given by the mean response time of elastic data traffic flows in low and medium traffic load conditions, by applying an equivalent request arrival rates at the BS. However, the mean response time approximation does not work well in a heavy traffic load condition.

4.2 System Model

The system architecture, channel model, interactive data traffic model, and channel access schemes under consideration are described in this section.

4.2.1 System Architecture

In this study, we consider a centralized CRN with a base station as illustrated in Fig. 4.1. The secondary network operates over a time-slotted single-channel primary network, and the secondary network is an overlay network which is spectrum homogeneous. The BS and the SUs use a low-rate control channel to transmit control packets. Each SU is equipped with a single transceiver to sense the channel and transmit information packets.

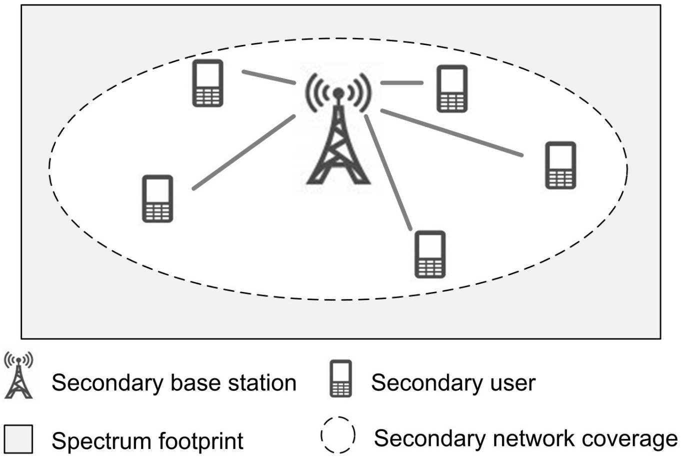


Figure 4.1: The centralized CRN with a BS.

4.2.2 Channel Availability Model

The channel model is similar to the one discussed in Subsection 2.2.2, in which the time is partitioned into slots of constant duration. The channel state of each time-slot is defined as 0 if the channel is busy (PUs are active), and is 1 otherwise. The state transition of the channel among adjacent time-slots can be illustrated using a two state Markov chain, where $S_{i,j}$ denotes the transition probability from state i ($\in \{0, 1\}$) to state j ($\in \{0, 1\}$). The appearance of a PU at the end of an idle period (no PU is active) is denoted as service interruption to the secondary network.

4.2.3 Interactive Data Traffic

The secondary network supports a large number of interactive traffic users, and the interactive data sessions arrive according to a Poisson process with mean arrival rate λ. The interactive class comprises non-real time services with a request-response pattern. Each data session is a collection of a finite number of file requests and thinking processes. The number, N_f, of file requests in a data session is geometrically distributed with mean $1/P_T$. An SU places a file request at the BS via a low rate control channel. Each data session can be modeled as a collection of alternating on-off phases [29], where an on phase corresponds to the response time, T_R, and an off phase correspond to an exponentially distributed thinking time, T_T, before the next file request. The response time is the time duration from the instant that a user places a file request at the BS until it receives the final data packet of the file. We use the terms service request to denote a file request of an SU, user arrival to denote a session arrival at the BS, and completion of a service request to denote the completion of a file transfer. The structure of an interactive session is illustrated in Fig.4.2. After completion of each service request, an SU decides whether to end the session or to enter the thinking state. If the number of data files per session is one ($N_f = 1$), the data session is similar to an elastic data traffic session. Similar to that with elastic data files, during each available time-slot, only one data user is being served and the size, L_{pk}, of a data chunk (packet) transferred during a time-slot is the

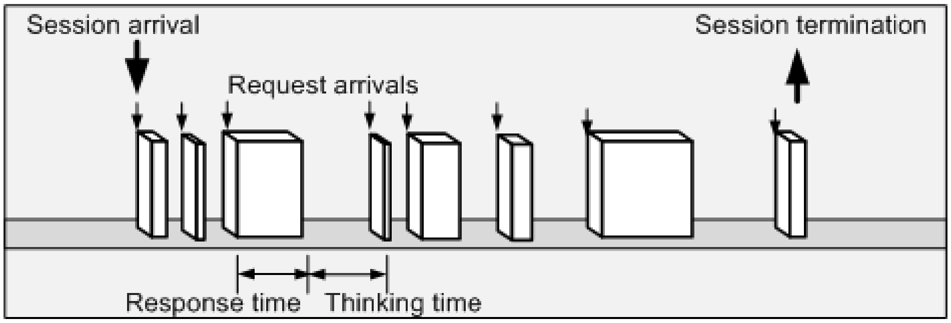


Figure 4.2: Structure of an interactive session [29]

same for all time-sots. The duration of an on phase depends on the STR of the request. The STR depends on the length, L, of the requested data file, and L_{pk}. The lengths of requested files are independently and identically distributed with a Weibull distribution which characterizes for Internet data traffic [69][79]. The PDF, $f_L(\cdot)$, of file length L is given by

$$f_L(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, \quad \alpha > 0, \beta > 0, x > 0 \tag{4.1}$$

where α and β are the shape parameter and the scale parameter, respectively. The shape parameter governs the tail heaviness of the Weibull distribution. Upon a service request, the whole data file is available at the BS without any delay. The mean response time, $E[T_R]$, of a data file is considered as the QoS parameter of the interactive data traffic flows, and the QoS requirement is given by

$$E[T_R]\leq\bar{T}_{R,\max} \tag{4.2}$$

where $\bar{T}_{R,\max}$ is the maximum allowable mean response time to provide satisfactory service quality. The interactive data capacity, $N_{\max}$, is defined as the maximum number of interactive data sessions that can be supported by the system without violating the service requirement in (4.2).

4.2.4 Service Disciplines

The secondary base station serves the secondary data users in shortest processor time without preemption, shortest processor time with preemption, shortest remaining processor time, and processor sharing service disciplines, respectively, as described in Subsection 3.2.3.

4.2.5 Assumptions

The simulation results for time durations such as the response time, and inter-arrival/departure time of the service requests are measured in discrete-time (in terms of number of time-slots).

Further, the channel availability and busy durations are geometrically distributed with mean $1/S_{1,0}$ and $1/S_{0,1}$, respectively. For analysis tractability, the channel availability/busy durations, the service time requirements, and BS service process, and the thinking durations are considered to be in continuous time. Without loss of generality the data packet size, L_{pk}, is considered as unit length. Therefore, the PDF of the service time requirement, τ, is $f_\tau(x) = f_L(x)$. Further, the inter-arrival/departure time distributions in the analysis are treated as continuous-time distributions. However, when comparing an inter-arrival/departure time distribution with a continuous-time distribution using the simulation results, the discrete-time equivalent of the continuous-time distribution is used for the comparison. For example, in order to find the match between in inter-departure time distribution using the simulation results with an exponential distribution, the inter-departure time samples are compared with the geometric distribution (which is the discrete-time analogue of the exponential distribution).

The mean response time of the service requests depends on the request arrivals at the BS. Unlike in the elastic data traffic, in addition to the requests from the new session arrivals, at the BS, there are request arrivals from the users returning from the thinking state (after reading/viewing the last file). In order to analyze the mean response time, it is required to study the request arrival/departure processes. We model the system as a queuing network to study the arrival/departure processes at the BS.

4.3 Secondary Network as a Queuing Model

A service request arrives at the BS stays at the BS queue[1] until its service completion, and it enters the thinking state afterwards. The BS service disciplines (PS, SPTNP, SPTWP, and SRPT) are general processes and the session arrivals follow a Poisson process. Therefore, for elastic data traffic (N_f=1), the BS queue can be considered as an M/G/1 queue. However, when

[1] As discussed in Section 3.3 a request goes through the waiting queue and the service queue to complete the service at the BS, and the BS queue is a combination of both the queues.

serving interactive data traffic (N_f >1), a portion of completed service requests return to the BS as new service requests after completion of the thinking time. The request arrival process at the BS is a combination of the service requests placed by the new users and the existing users returning from the thinking state. Therefore, the request arrival process at the BS may not be a Poisson process, and the service at the BS should be modeled as a G/G/1 queue. All the users entering the thinking state can be modeled by a thinking queue with an exponentially distributed service time and an infinite number of servers to serve any number of incoming users. As the arrivals at the thinking queue depend on the departures of the BS queue, which is a general process, the arrival process of the thinking queue is also a general process. The system can be modeled as two cascaded queues with feedback, as illustrated in Fig. 4.3, where i=0, i=1, and i=2 denote the whole system (which is a combination of the BS and the SUs), the BS queue, and the thinking queue, respectively. Let $x_i(t)$ and $y_i(t)$ denote the arrival and departure processes, respectively, of the system component i, and λ_i and μ_i denote the mean values of $x_i(t)$ and $y_i(t)$, respectively, where $i \in \{0, 1, 2\}$.

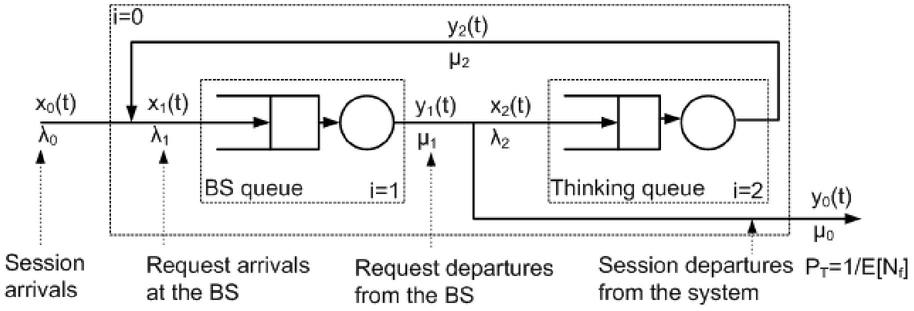


Figure 4.3: The cascaded queues with feedback [29].

4.4 Mean Response Time of Interactiv Data Sessions

The mean response time of a service request depends on the file length or the service time requirement, the channel availability statistics, the service discipline, and the request arrival process $x_1(t)$ of the BS queue. The request arrival process to the BS queue $x_1(t)$ is a combination of the Poisson session arrival process $x_0(t)$ and a general departure process $y_2(t)$. Therefore, $x_1(t)$ is a general process. Given the STR distribution, the channel availability statistics, and the service disciplines we want to characterize $x_1(t)$, which is the first step to analyze the mean response time. As described in Section 4.3, the thinking queue is a G/M/∞ queue. If the arrival process $x_2(t)$ is Poisson, the thinking queue becomes an M/M/∞ queue, and Poisson arrivals at the thinking queue results in Poisson departures with mean identical to that of the arrival process. This would lead to Poisson arrivals at the BS queue. Since the request arrival process $x_1(t)$ at the BS queue depends on the departure process $y_1(t)$, we need to model $y_1(t)$. The departure processors of the BS queue is complex to analyze with the given service disciplines and the service interruptions (due to the presence of the PUs). As a result, we resort to simulations to study the behavior of the departure process $y_1(t)$ of the BS queue. From this point onward, we use the term distribution to denote the CDF of inter-arrival/departure times.

4.4.1 Inter-departure Time Distribution of the BS Queue

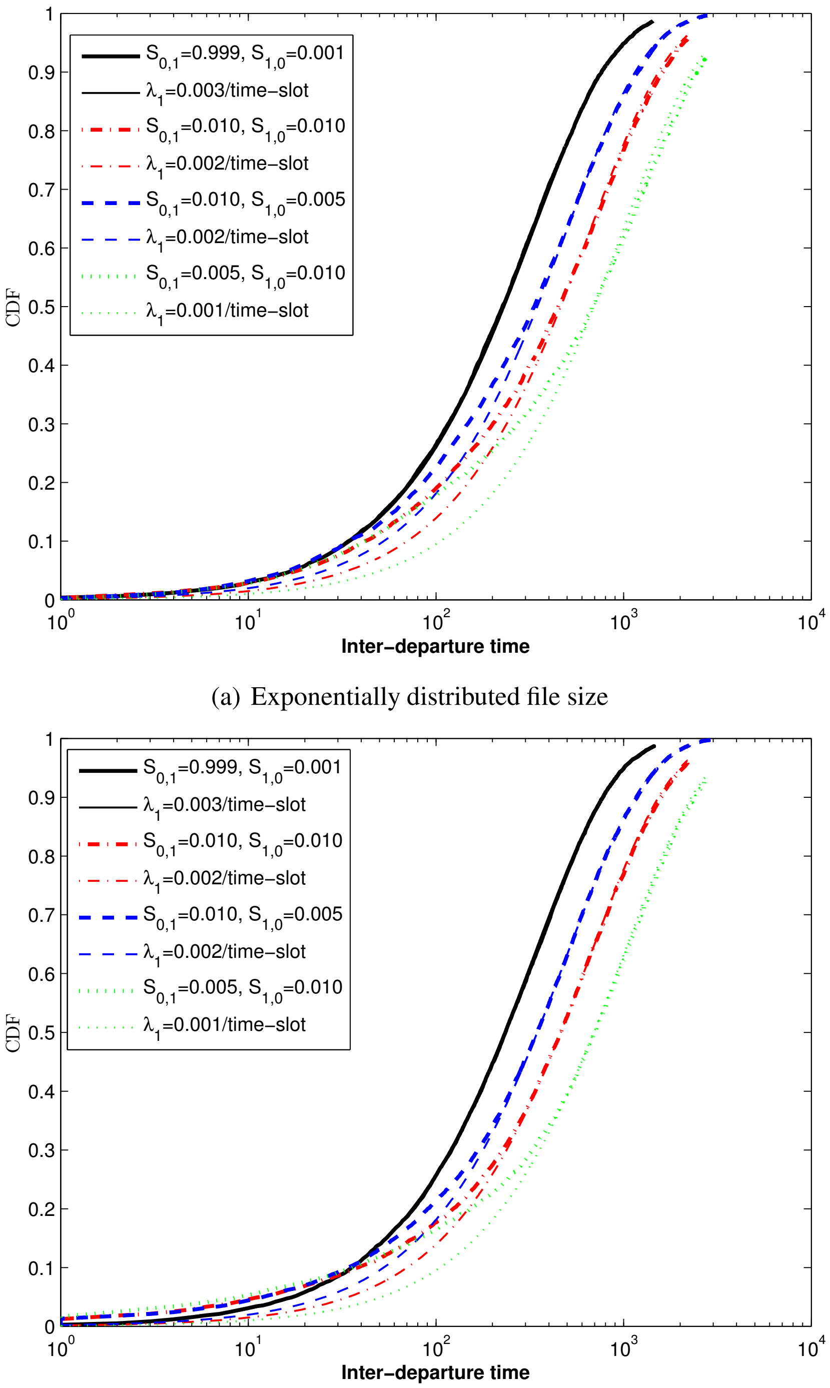
Computer simulations are carried out to compare the inter-departure time distribution of the BS queue with an exponential distribution. As the system is time-slotted, the simulations are in discrete time and the time durations are measured in time-slot units. Therefore, the inter-departure time distribution is compared with a geometric distribution with the same mean value as that of the exponential distribution. The simulation setup is similar to that in Section 3.5. As the first step, we obtain the inter-departure times of the BS queue by keeping the number of files per sessions N_f=1 (i.e., elastic data traffic), having Poisson session arrival process, and mean STR equal to 200 time-slots. We carry out the simulations for different channel availability statistics and different tail properties of the STR distributions. For each system parameter set, we complete 10 simulation runs, and in each simulation run the BS completes 9,000 service requests. Figs. 4.4-4.7 illustrates the comparison of the inter-departure time distribution and the corresponding geometric distribution (CDF) for different channel availability statistics and different tail properties of the STR distribution for all four service disciplines, respectively.

The simulation results demonstrate that the longer the inter-departure time the closer its distribution to the geometric distribution. However, the inter-departure time distribution shows a large deviation from the geometric distribution for the SPTNP service discipline, when the STR is heavytail distributed. When the STR is exponentially distributed and the service disciplines do not allow preemptions, the longer the T_{off} and the shorter the T_{on}, the larger the deviation of inter-departure time distribution from the corresponding geometric distribution. On the other hand, (when the STRs are exponentially distributed and the service disciplines allow preemptions) the longer the T_{off} and the shorter the T_{on}, the smaller the deviation between the two distributions. Even though the inter-departure time distribution deviates form the geometric distribution at short inter-departure times, for the tractability of the mean response time analysis, we assume that the inter-departure time distribution of the BS queue is exponentially distributed in the following scenarios:

1. For all four service disciplines, when the STR is exponentially distributed;
2. For the PS, SRPT, and SPTWP service disciplines, when the STR is heavytail distributed, given that the T_{off} is not very long and T_{on} is not very short in comparison with mean STR.

4.4.2 Exponentially Distributed Inter-departure Times of the BS Queue

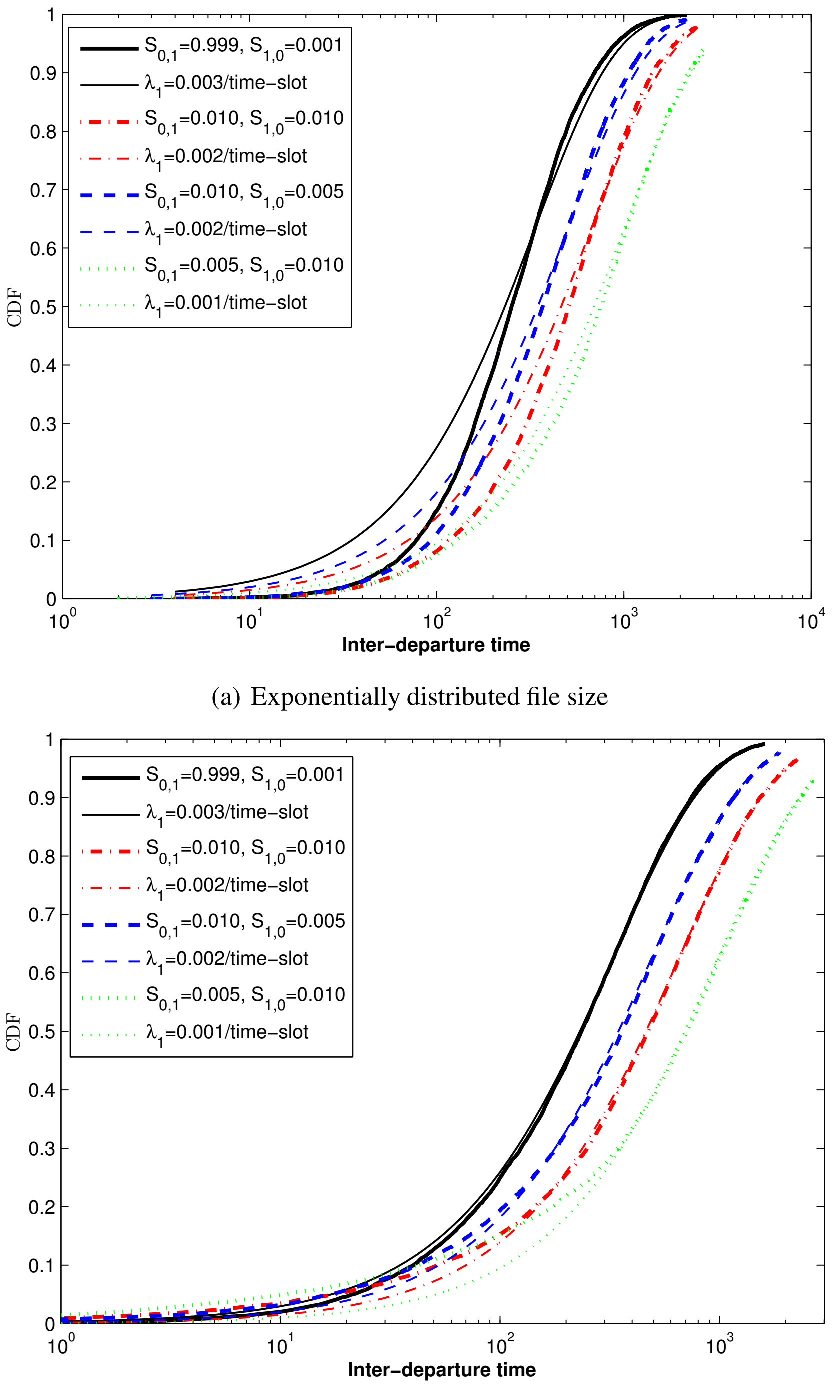
From Fig. 4.3, if the departure process $y_1(t)$ is Poisson, the arrival process $x_2(t)$ is also Poisson with mean λ_2=(1-P_T)μ_1. The term P_T accounts for the session departures form the system. Therefore, the mean arrival rate at the thinking queue is given by, $\lambda_2=(1-P_T)\mu_1=(1-P_T)\lambda_1$. Further, the request arrivals at the BS queue is Poisson distributed with mean arrival rate



(a) Exponentially distributed file size

(b) Heavytail distributed file size

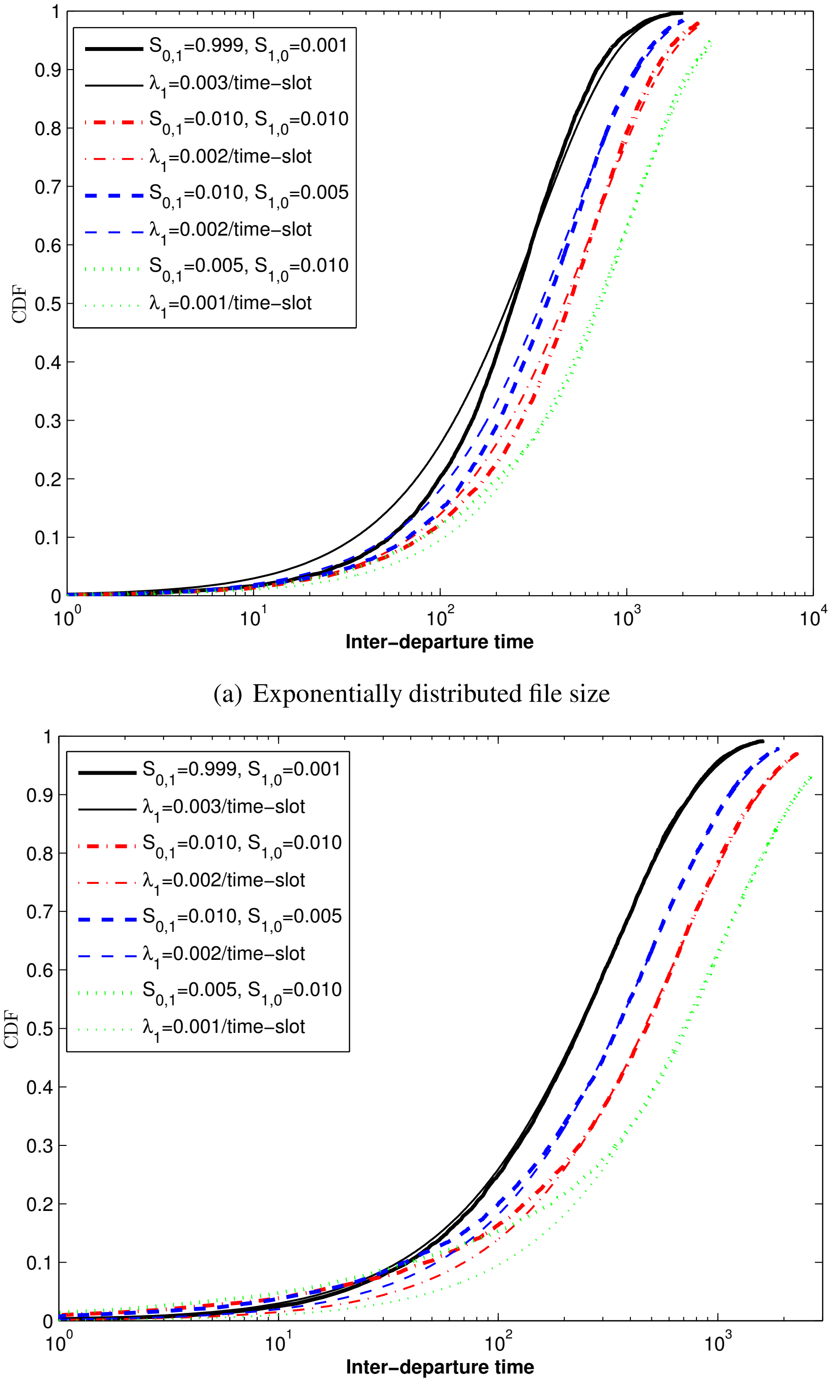
Figure 4.4: Comparison of inter-departure time distribution with the geometric distribution for the PS service discipline



(a) Exponentially distributed file size

(b) Heavytail distributed file size

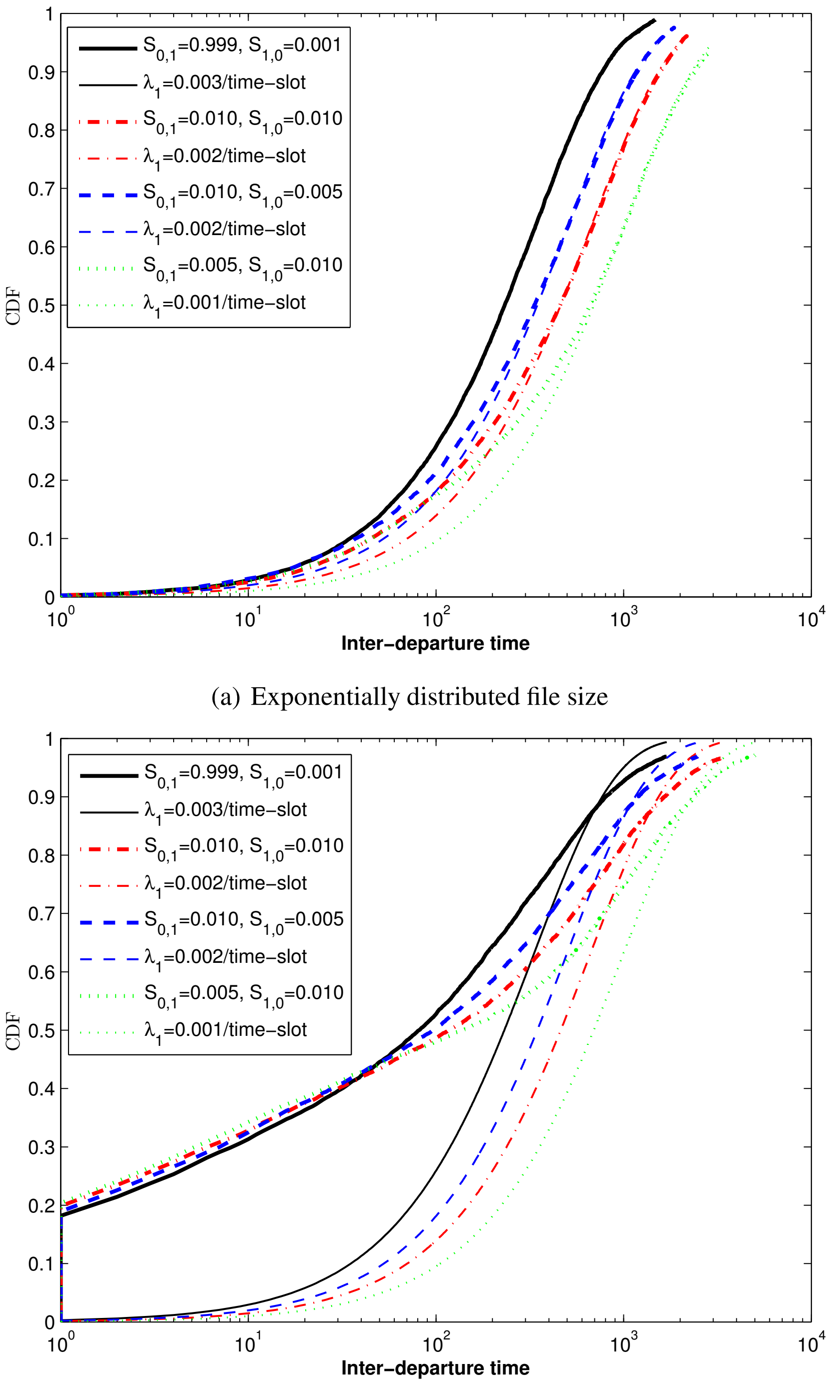
Figure 4.5: Comparison of inter-departure time distribution with the geometric distribution for the SRPT service discipline



(a) Exponentially distributed file size

(b) Heavytail distributed file size

Figure 4.6: Comparison of inter-departure time distribution with the geometric distribution for the SPTWP service discipline



(a) Exponentially distributed file size

(b) Heavytail distributed file size

Figure 4.7: Comparison of inter-departure time distribution with the geometric distribution for the SPTNP service discipline.

λ_1=λ_0+μ_2. With further manipulations, it can be shown that

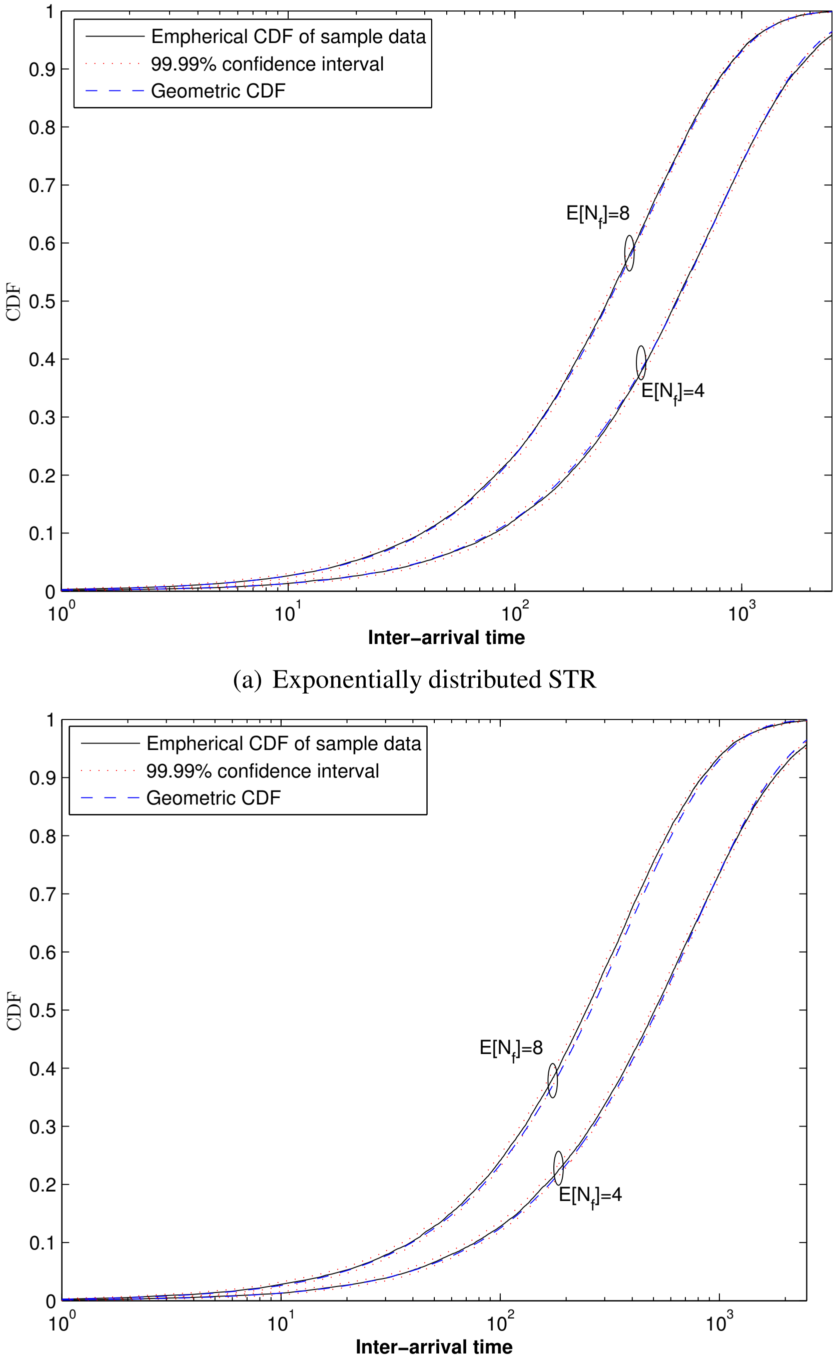
$$\lambda_1 = \lambda_0[1 + (1 - P_T)/P_T] = E[N_f]\lambda_0. \tag{4.3}$$

Therefore, the arrival process of the BS queue is Poisson distributed with mean λ_1=$E[N_f]\lambda_0$, and the offered traffic load at the BS from the interactive (multi-file) data sessions is equal to $E[N_f]$ times the offered traffic load by single-file data sessions with same STR distribution. In order to find the closeness of $x_1(t)$ to a Poisson process with mean arrival rate $E[N_f]\lambda_0$, we compare the service request inter-arrival time distributions of the BS queue with an exponential distribution with mean $1/E[N_f]\lambda_0$.

Computer simulations are carried out to compare the service request inter-arrival time distributions of the BS queue with geometric distribution (the discrete-time equivalent of the exponential distribution). The BS transmits only one packet in each idle time-slot, and it follows the four service disciplines, respectively. We obtain the service request inter-arrival times of the BS queue for different $E[N_f]$ values, having Poisson session arrival process and Weibull distributed STR with mean equal to 200 time-slots. The STRs are rounded off to the next higher integer value. We carry out the simulations for different channel availability statistics and different tail properties of the STR distributions. For each system parameter set, the BS completes 30,000 service requests.

Figs. 4.8-4.11 and 4.12-4.15 compare the service request inter-arrival time distributions of the BS queue for multi-file data sessions with the geometric distribution (CDF) with mean $1/E[N_f]\lambda_0$ for two different channel availability scenarios, respectively. We keep the mean STR $E[\tau]$=200 time-slots and mean the thinking duration $E[T_T]$=5,000 time-slots for all the four service disciplines.

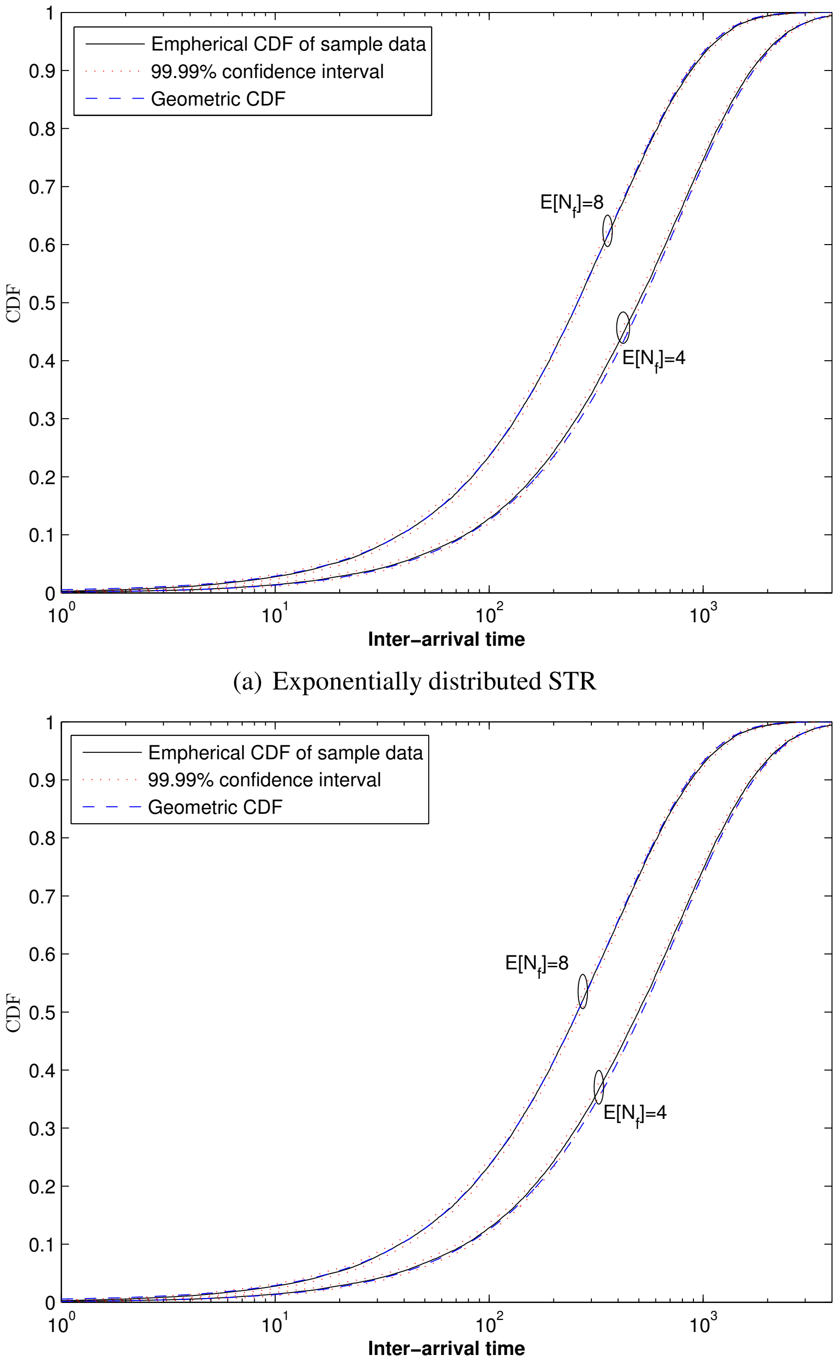
It can be observed that the service request inter-arrival time distribution match closely with the geometric distribution (CDF with mean $1/E[N_f]\lambda_0$), and the match is closer than that between



(a) Exponentially distributed STR

(b) Heavytail distributed STR

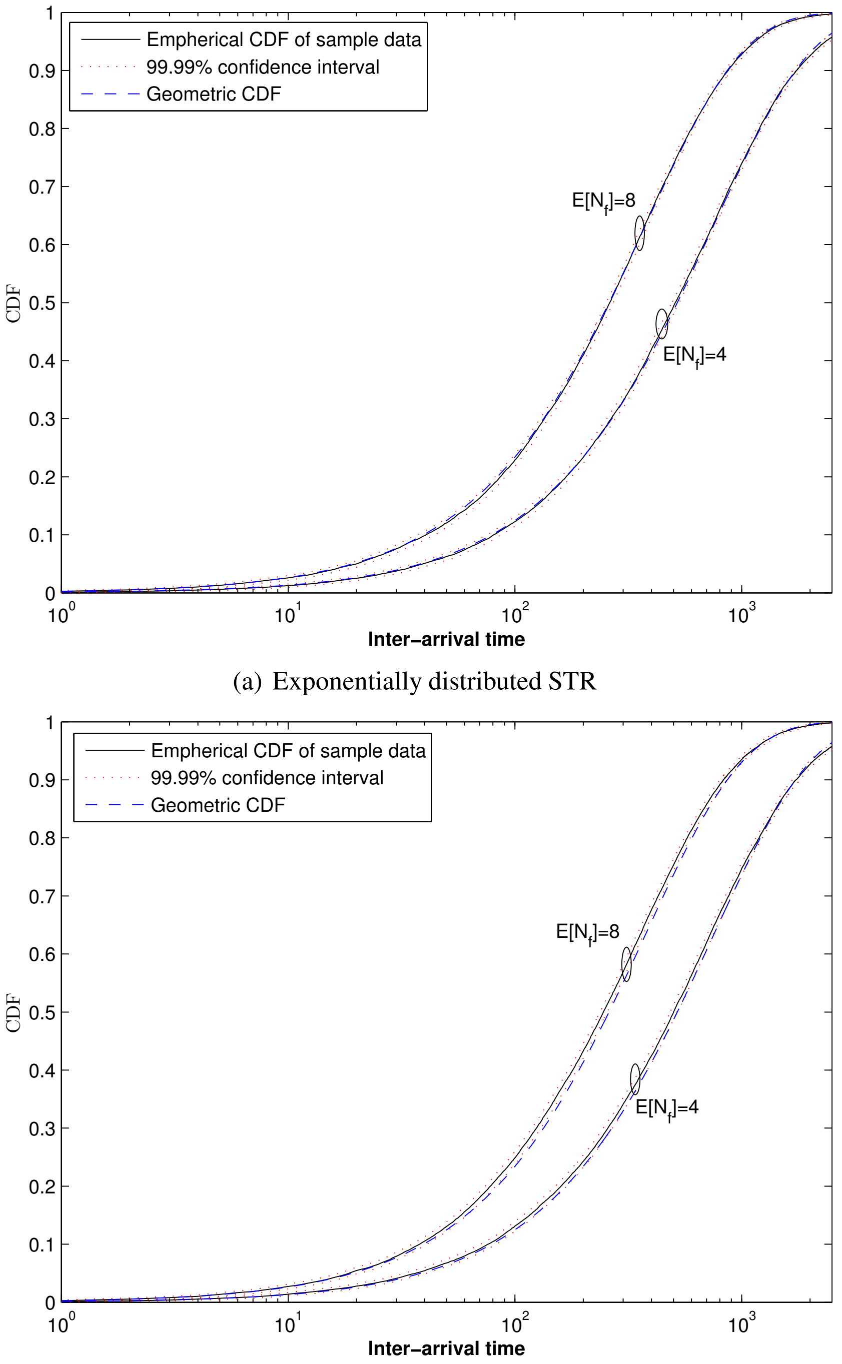
Figure 4.8: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the PS service discipline having $T_{on} = 200$ and $T_{off} = 100$ time-slots.



(a) Exponentially distributed STR

(b) Heavytail distributed STR

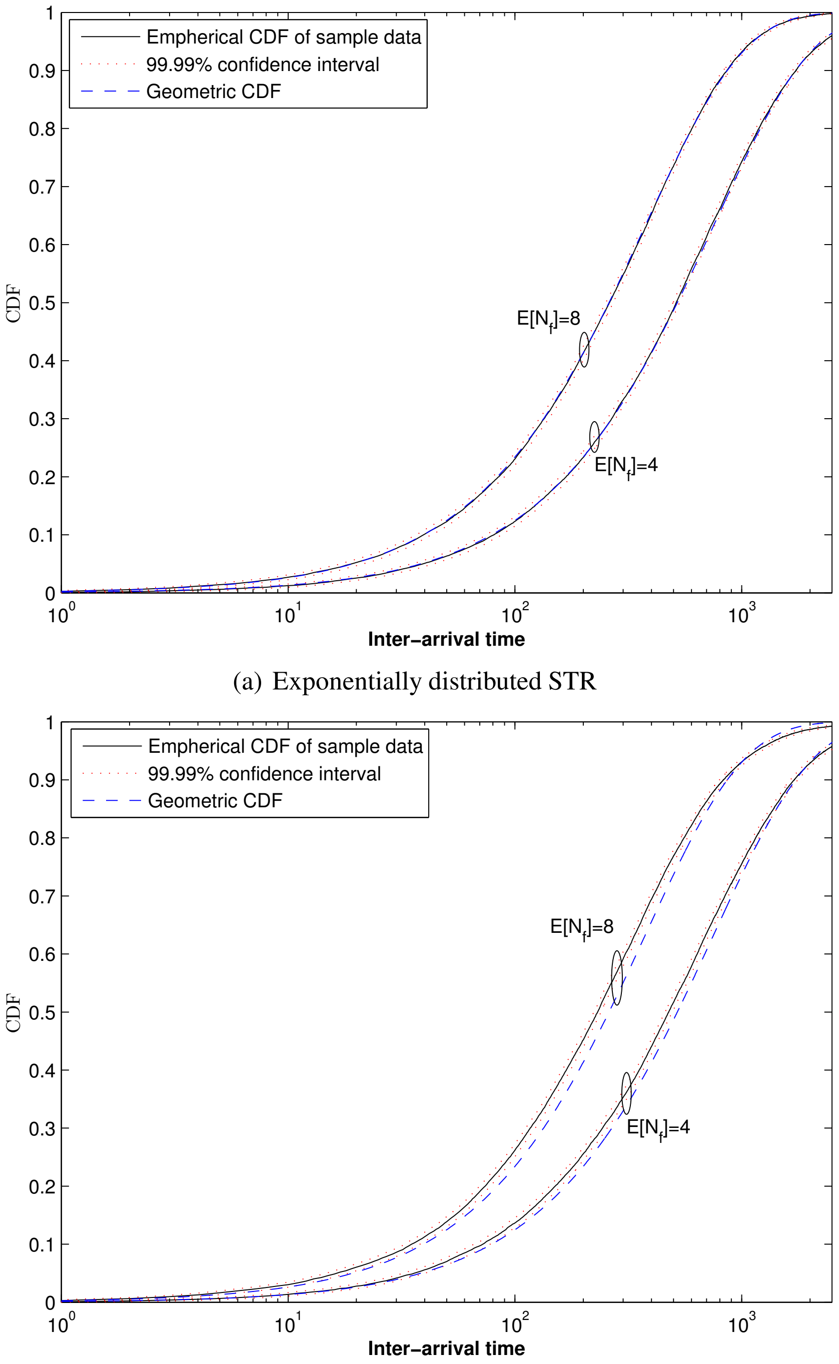
Figure 4.9: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the SRPT service discipline having $T_{on} = 200$ and $T_{off} = 100$ time-slots.



(a) Exponentially distributed STR

(b) Heavytail distributed STR

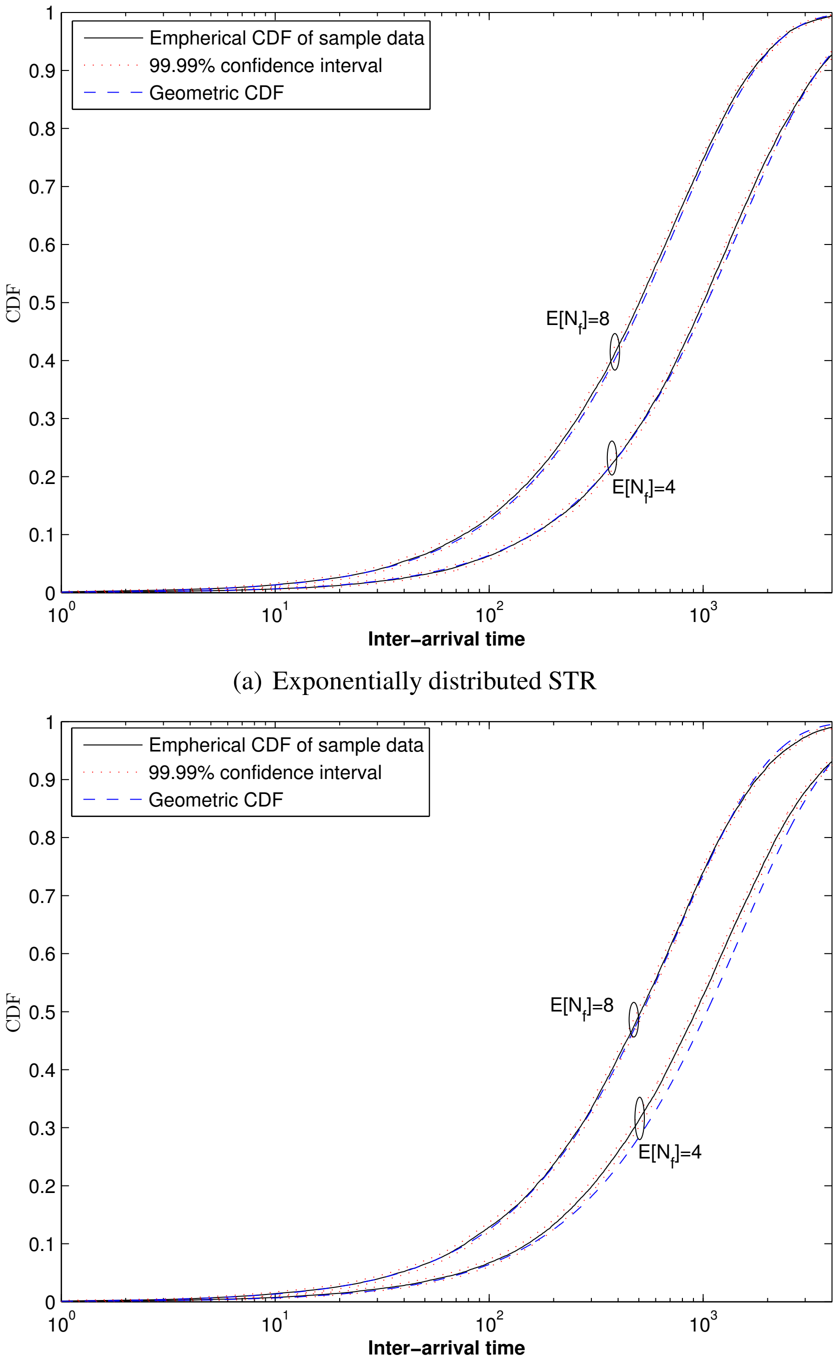
Figure 4.10: Comparison between the service request inter-arrival time distributions and the geometric distribution having mean equal to $1/E[N_f]\lambda_0$ for the SPTWP service discipline having $T_{on} = 200$ and $T_{off} = 100$ time-slots.



(a) Exponentially distributed STR

(b) Heavytail distributed file size

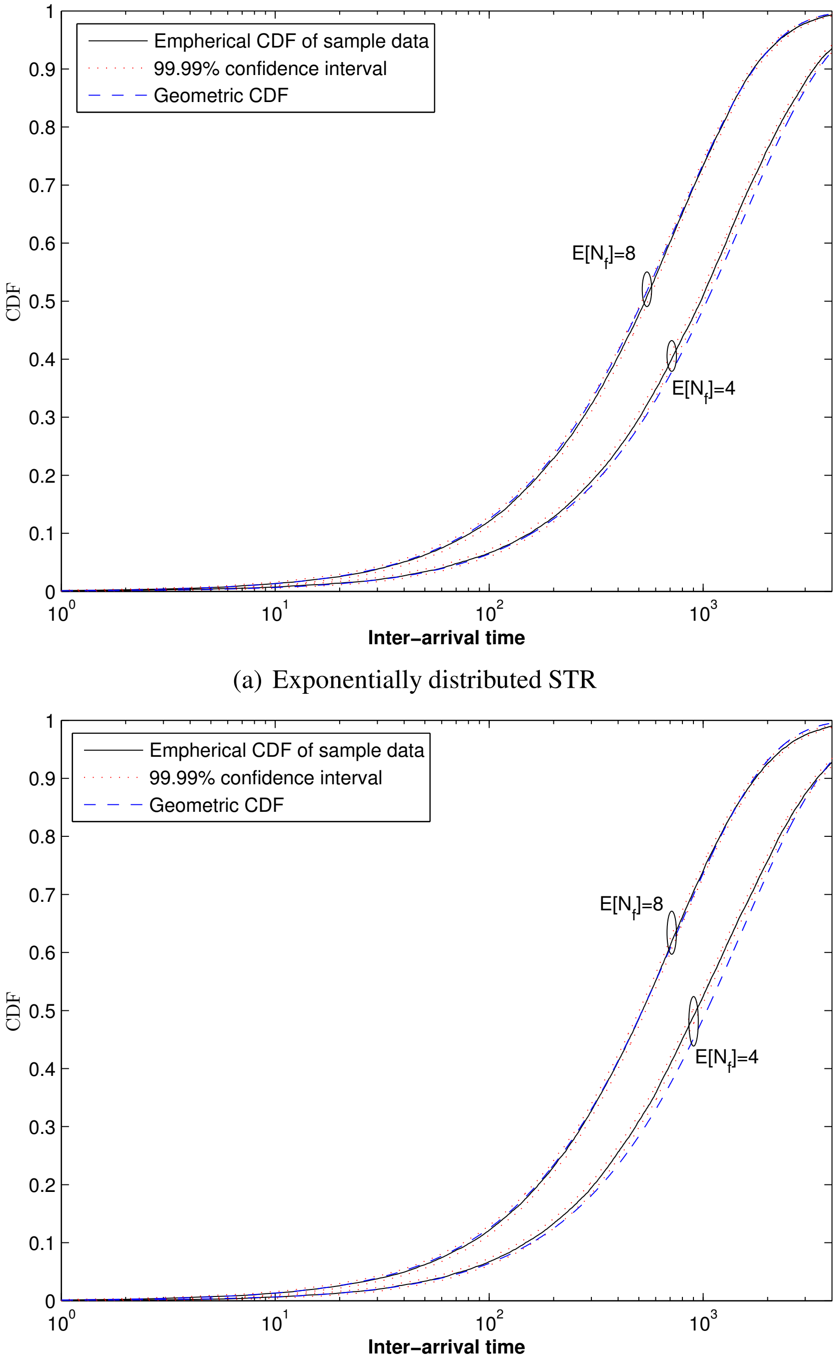
Figure 4.11: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the SPTNP service discipline having $T_{on} = 200$ and $T_{off} = 100$ time-slots.



(a) Exponentially distributed STR

(b) Heavytail distributed STR

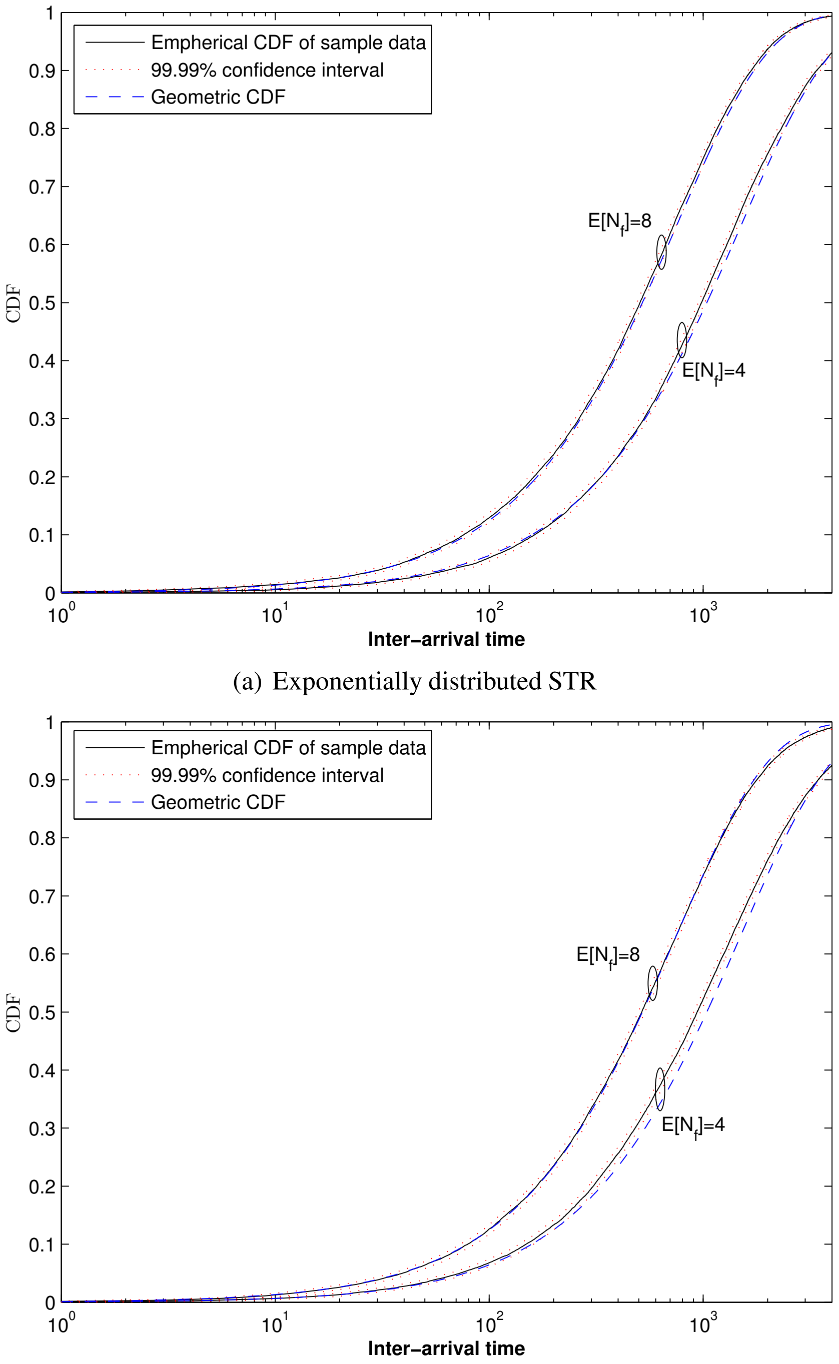
Figure 4.12: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the PS service discipline having T_{on} = 100 and T_{off} = 200 time-slots.



(a) Exponentially distributed STR

(b) Heavytail distributed STR

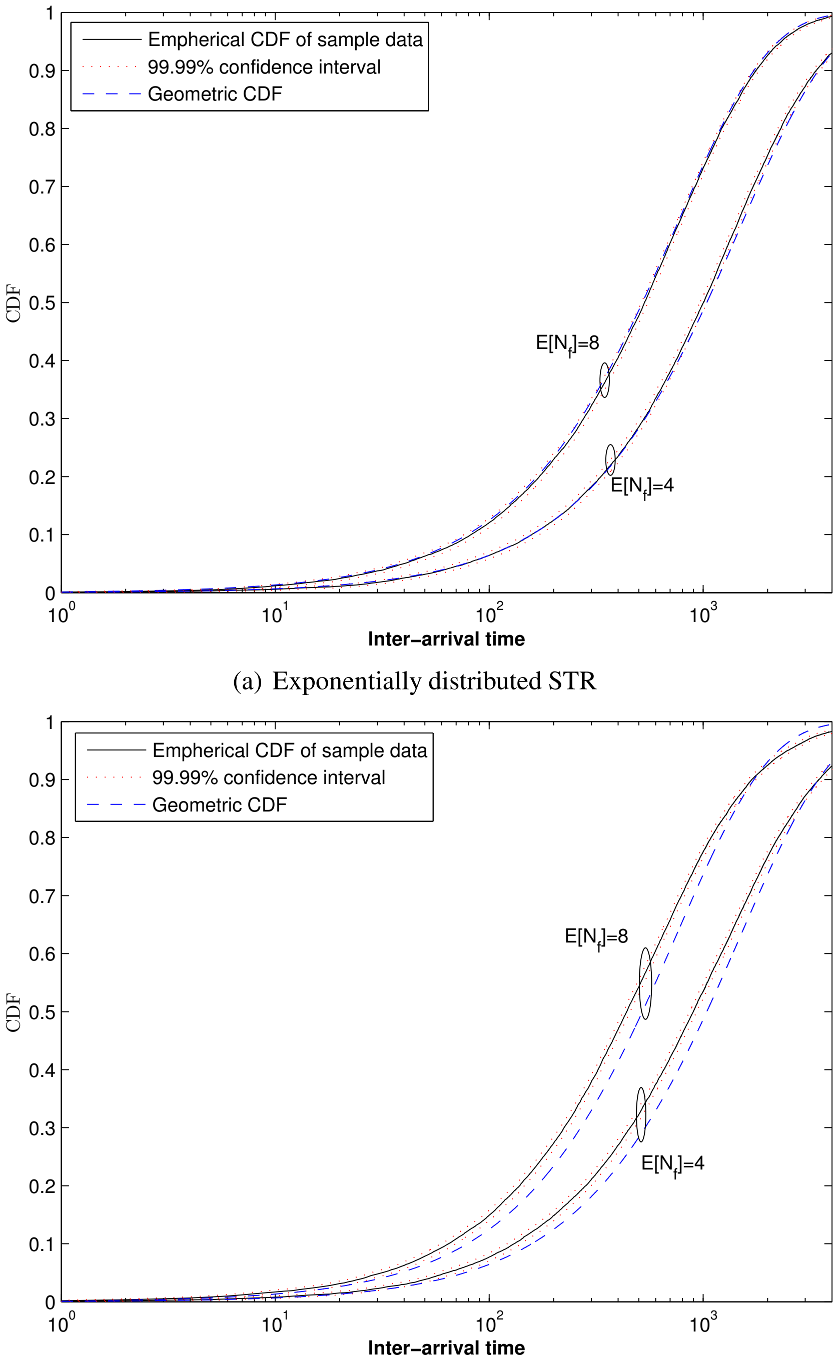
Figure 4.13: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the SRPT service discipline having $T_{on} = 100$ and $T_{off} = 200$ time-slots.



(a) Exponentially distributed STR

(b) Heavytail distributed STR

Figure 4.14: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the SPTWP service discipline having $T_{on} = 100$ and $T_{off} = 200$ time-slots.



(a) Exponentially distributed STR

(b) Heavytail distributed STR

Figure 4.15: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the SPTNP service discipline having $T_{on} = 100$ and $T_{off} = 200$ time-slots.

the inter-departure time distribution and the geometric distribution for all the four service disciplines. From Figs. 4.8-4.11, when the T_{on} is longer and the T_{off} is shorter than the mean STR, the geometric distribution stays within the 99.99% confidence bounds of the observed service request inter-arrival time samples. However, with SPTNP service discipline, the service request inter-arrival time distribution deviates from the geometric distribution, when the mean STR is heavytail distribution. It is due to the large deviation of the inter-departure time distribution of the BS queue, which is illustrated in Fig. 4.7. From Figs. 4.12-4.15, when the T_{on} is shorter and the T_{off} is longer than the mean STR, the service request inter-arrival time distribution slightly deviate from the geometric distribution. Figs. A.1-A.2 also demonstrate the deviation of the two distributions at shorter T_{on} and longer T_{off} when compared with the mean STR. The larger the mean number of files in a data session, the closer the match between service request inter-arrival time distribution and the geometric distribution (i.e., in continuous-time, the larger the mean number of files in a data session, the closer the service request inter-arrival time distribution of the BS queue to an exponential distribution with mean $1/E[N_f]\lambda_0$). The larger the $E[N_f]$, the larger the offered load to the BS queue, leading to large departure rates from the BS queue. The larger the departure rate of the BS queue, the lower the chances of having an empty thinking queue. As the thinking durations are exponentially distributed, the lower the probability of having an empty thinking queue, the closer the inter-departure time distribution of the thinking queue to the exponential distribution. As the simulation result for the service request inter-arrival time distribution appear to match well with the geometric distribution, in our continuous-time analysis, the request arrival process at the BS can be approximated by a Poisson process with mean arrival rate λ_1=$E[N_f]\lambda_0$ for longer T_{on} and shorter T_{off} than the mean mean STR.

4.4.3 Mean Response Time of Multi-file Data Sessions

Denote the mean arrival rate and the offered traffic load from the first file of multi-file data sessions (or considering multi-file data sessions as single-file data sessions) as the base arrival rate, λ_0, and base traffic load, σ_0, respectively. From (4.3), under the Poisson approximation, it is clear that the request arrivals at the BS from a system with multi-file data sessions can be represented by an equivalent system with single-file (elastic) data sessions. Therefore, the mean response time for the SPTNP, SPTWP, SRPT, and PS service disciplines can be approximated using (3.4), (3.22), (3.28), and (3.33), respectively, with mean session arrival rate $\lambda{=}E[N_f]\lambda_0$. Note that, at the steady state, when the arrival process of the thinking queue is Poisson, the mean thinking time has no impact on the mean departure rate. Therefore, the mean thinking time has no impact on the mean arrival rate at the BS queue and the mean response time, when the system reaches steady state.

4.5 Numerical Results for the Mean Response Time

Computer simulations are carried out to evaluate the accuracy of the mean response time approximation for the interactive data traffic, which is the mean response time obtained using the mean response time equations derived for the elastic data traffic by setting the $\lambda{=}E[N_f]\lambda_0$. From this point onward, we denote the mean response time approximation as the analytical result. The simulation setup is similar to the one used in Subsection 4.4.2 to obtain the inter-arrival/departure times of the BS queue for interactive data traffic ($E[N_f] > 1$). The mean response time, $E[T_R]$, is evaluated by averaging the results from 20 simulation runs, each run having 18,000 service requests.

Figs. 4.16-4.19 show the variation of the mean response time with $E[N_f]$ obtained from the simulations and the analytical results for all the service disciplines, respectively. The mean response times for the four service disciplines are obtained from (3.4), (3.22), (3.28), and (3.33), respectively. We keep the mean STR $E[\tau]$=200 time-slots and base traffic load σ_0=0.1 throughout the simulations.

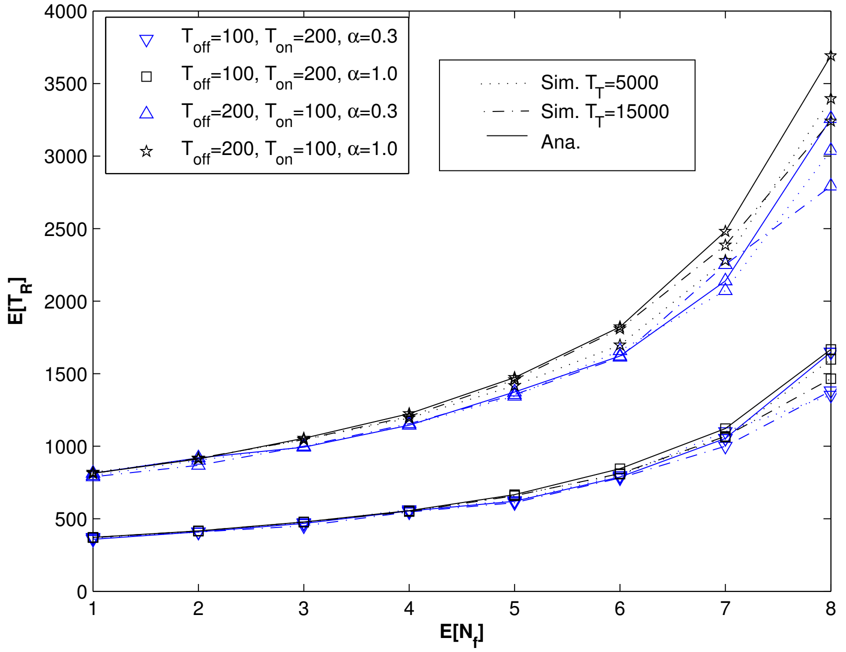


Figure 4.16: Variation of the mean response time with $E[N_f]$ for the PS service discipline.

It can be observed that the smaller the $E[N_f]$, the closer the match between the simulation results and the analytical results. Since the offered traffic load at the BS queue is $E[N_f]{\cdot}\sigma_0$, the results can be explained with respect to the offered traffic load at the BS. The simulation results match closely with the numerical results at low and medium traffic load conditions, and they start to deviate when the traffic load is increasing. Figs. 4.8-4.15 illustrate that the larger the $E[N_f]$ (or the offered traffic load) the closer the service request inter-arrival time distribution to the exponential distribution with mean $1/E[N_f]{\cdot}\lambda_0$. However, the results from Figs. 4.16-4.19 illustrates that the larger the $E[N_f]$ (or the offered traffic load), the larger the deviation of the simulation results from the analytical results. This phenomena can be explained using Fig. 4.20 (which is the same as Fig. 3.12) which illustrates the variation of the mean response

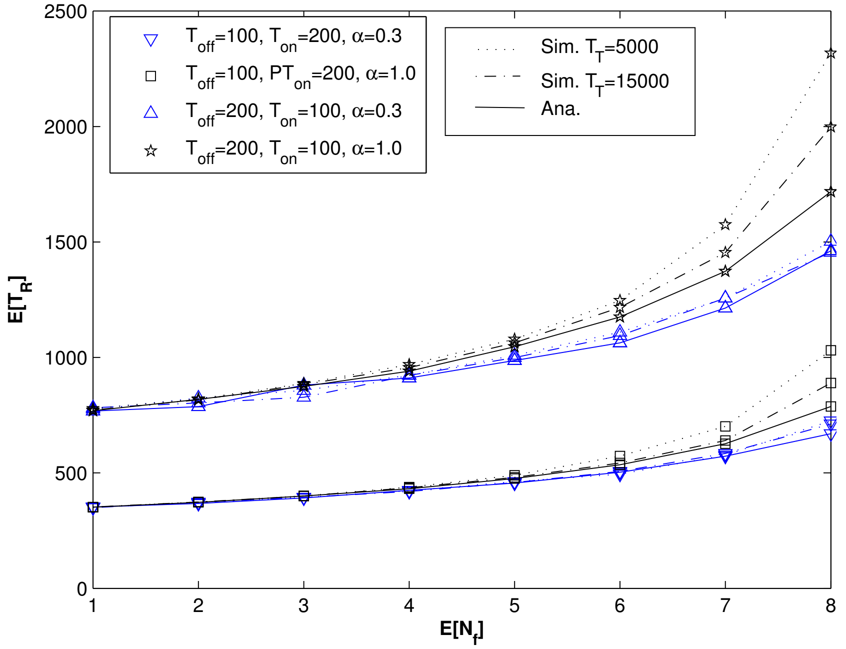


Figure 4.17: Variation of the mean response time with $E[N_f]$ for the SRPT service discipline.

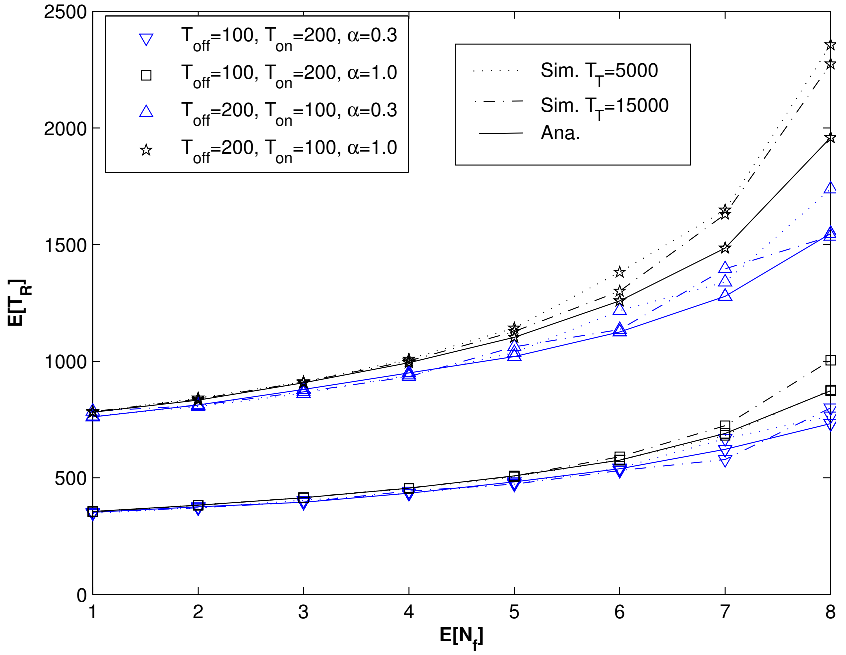


Figure 4.18: Variation of the mean response time with $E[N_f]$ for the SPTWP service discipline.

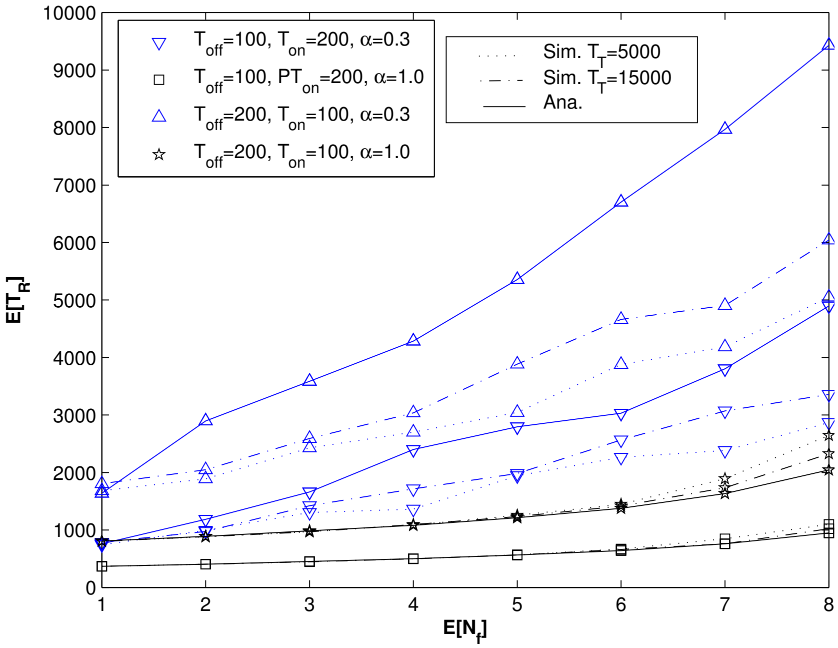


Figure 4.19: Variation of the mean response time with $E[N_f]$ for the SPTNP service discipline.

time of elastic data traffic with the offered traffic load at the BS. The larger the offered traffic load, the larger the rate of increment of the mean response time for all the service disciplines, respectively. This phenomena leads to large variations in the mean response time for small variations in the offered traffic load, at high offered traffic load conditions. Therefore, even though the deviation of the service request inter-arrival time distribution from the approximated exponential distribution is small, the variation of the mean response time is larger at high traffic load conditions than that at the low traffic load conditions. Figs. 4.12-4.15 demonstrate that the longer the T_{off} and the shorter the T_{on} in comparison with the mean STR, the larger the deviation of the service request inter-arrival time distribution form the expected geometric (or exponential in continuous-time) distribution. It is confirmed by the results in Appendix 5.2. However, Figs. 4.16-4.19 demonstrate that the mean response time approximation works well even when the T_{off} is comparable and the T_{on} is shorter in comparison with the mean STR. Therefore, we can conclude that the sensitivity of the mean response time to the variations of

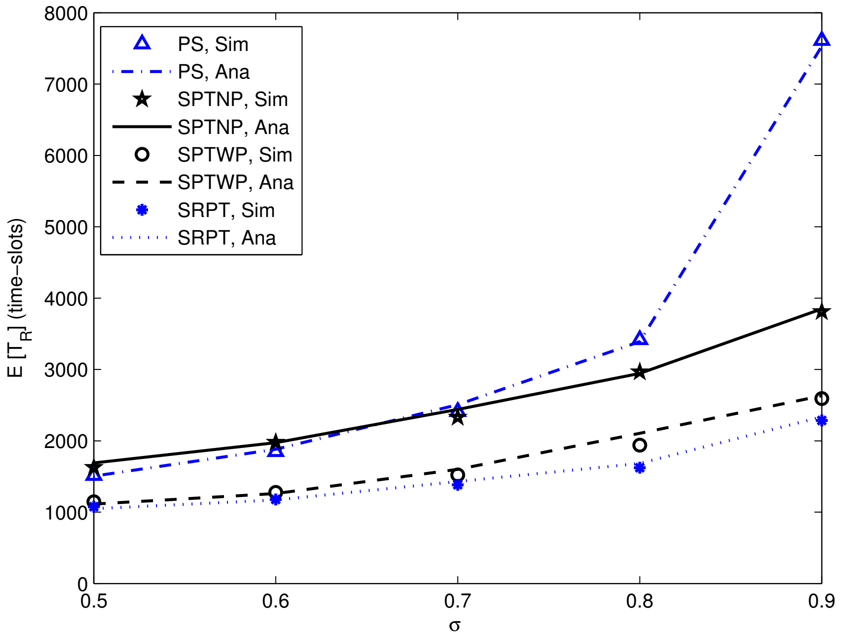


Figure 4.20: The variation of mean response time with the traffic load.

the service request inter-arrival time distribution are very low for the SPTWP, SRPT, and PS service disciplines. Further, the simulation results for the mean response time of the SPTNP service discipline have large deviations form the approximated (analytical) results, when the STR is heavytail distributed. It is due to the deviation of the service request inter-arrival time distribution from the exponential distribution. Therefore, we can conclude that the mean response time for the SPTNP service discipline is sensitive to the variations of the service request inter-arrival time distribution when the STR is heavy tail distributed than when it is exponentially distributed (i.e., the sensitivity of the mean response time to the variations of the service request inter-arrival time distribution varies with the heaviness of the tail of the STR distribution). From Fig. 3.12, it is shown that the SPTNP is not a good choice when the tail of the STR distribution is very heavy. Further, for the other three service disciplines, the sensitivity of the mean response time to the variations of the service request inter-arrival time distribution does not vary with the heaviness of the tail of the STR distribution. The results demonstrate that the shorter the mean thinking time, the larger the deviation of the mean response times. The shorter the mean thinking time, the faster the requests leave the thinking queue and the higher chances of having an empty thinking queue. The shorter the mean thinking time, the higher the chances of the inter-departure time distribution of the thinking queue deviate from the exponential distribution, and the higher the chances of service request inter-arrival time distribution of the BS queue to deviate from the approximated exponential distribution. When the service request inter-arrival time distribution of the BS queue deviates from the approximated exponential distribution, the analytical results deviate form the simulation results. It is clear from the numerical results that mean response time of the interactive data traffic can be closely approximated by that of the elastic data traffic, at low and medium traffic load conditions in the following scenarios: (1) for all four service disciplines with exponentially distributed STRs; (2) for the PS, SRPT, SPTWP service disciplines with heavytail distributed STRs.

4.6 Summary

In this chapter, we have investigated the relationship between the interactive and elastic data traffic flows, in terms of the mean response time. It is shown that the departure process of the BS queue deviates from the Poisson process when the mean channel available duration is shorter and the unavailable duration is longer than the mean STR. However, the request arrival process at the BS can be approximated by the Poisson process under specific channel availability conditions. The request arrival/departure processes of the BS queue show a larger deviation from the Poisson process for the SPTNP service discipline when the mean STR is heavytail distributed. It is shown that the mean response time of multi-file data sessions can be approximated by an equivalent single-file (elastic data) setup for low and medium traffic load conditions. Significant deviation between the analytical and simulation results are observed at heavy traffic load conditions. Further, it is shown that there is no significant impact from the mean thinking time on the mean response time at low and medium traffic load conditions when the system is at the steady state.

Chapter 5

Conclusions and Future works

In this chapter, we summarize the main research results and identify some further works.

5.1 Research Contributions

The objective of this research is to study the voice capacity of CRNs and the mean response time of elastic/interactive data traffic over the CRNs with different service processes. As there is an uncertainty in spectrum resource availability, rather than a strict delay requirement, a stochastic delay requirement is considered as the main QoS parameter for voice capacity analysis. The voice capacity is given in terms of the number of single-hop (up-link) voice traffic flows that can be supported by the system. A packet level analysis of the source buffer is carried out to find the packet dropping probability due to violation of the delay bound. The results demonstrate that the silent suppression in the voice traffic (on-off voice traffic) provides approximately twice the capacity than that of the constant-rate voice traffic. An existing DTMC model is modified to analyze the the capacity of the slot-ALOHA scheme, and a new DTMC model is developed to analyze the capacity of the round-robin scheme in supporting constant-rate voice traffic over distributed fully-connects CRNs. It is shown that the round-robin scheme performs better than the random allocation and the slot-ALOHA schemes. Further, the impact of the number of packets that can be transmitted in a time-slot (per channel) in the system capacity is larger in the round-robin scheme than in the other two schemes. As the overhead required in implementing the FCFS service disciple larger than that of implementing round-robin scheme, the round-robin scheme can be considered as an alternative in a centralized network. Given the mean channel availability, the system capacity decreases with the mean channel unavailable duration. Therefore, it is clear that the mean channel unavailable duration has a significant impact on the system capacity. For the simplicity in the capacity analysis, effects of the sensing errors has not been considered in our analysis. However, we have discussed the possibility of extending the DTMC models to incorporate the sensing errors into the capacity analysis. The analytical models used for the constant-rate voice capacity based on the stochastic delay requirement can be used to analyze the capacity of any type of constant-rate traffic flow. Further, our capacity analysis can be used as benchmarks in the performance analysis phase of the development of new channel access schemes to support voice traffic over CRNs. As the capacity analysis is limited to homogeneous voice traffic flows, two CAC algorithms are developed to support constant-rat voice traffic with different delay requirements (different delay bound and maximum delay bound violation probability) over non-fully-connected slot-ALOHA CRNs based on an empirical relationship between the successful transmission probability and the delay bound violation probability. It is shown that the longer the delay bound, the larger the system capacity. In other words, the lower the required service quality, the higher the system capacity. A low quality service service can be priced at a lower rate than a high quality service to increase the user satisfaction

We have studied the mean response time of elastic data traffic via a session level analysis. Three service disciplines are studied, namely, shortest processor time without preemption, shortest processor time with preemption, and shortest remaining processing time in comparison with the processor sharing service discipline. The works on non-cognitive networks only consider the mean channel availability (the long term channel availability) in the mean response time analysis. However, we have shown that the variation of the mean channel available available/non-available durations have a significant impact on the mean response time even when the mean channel availability is constant. Therefore, the mean response time evaluations based on the mean channel availability is not suitable in the context of the CRNs. The SRPT service discipline outperforms the other three service disciplines, and the preemption reduces the mean response time for heavytail distributed STRs. However, if the original STR is compared for preemptions rather than remaining STR, the advantage of the preemptions disappears with decreasing tail heaviness of the STR distribution. Therefore, when the STR distribution is concentrated around its mean, preemption gives a negative impact on the mean response time, and the SPTWP gives longer mean response times than the SPTNP. Even though the equal bandwidth sharing scheme in the PS outperforms the SPTNP when the tail of the STR is heavy, the lighter the tail of the STR distribution, the shorter the mean response times of the SPTNP in comparison with that of the PS. Further, the heavier the traffic load, the larger the rate of increment of the mean response time for the PS than that for the SPTNP. However, when the tail of the STR is heavier, it is shown that the mean response times corresponding to the SPTNP are extremely longer than that corresponding to the other three service disciplines. Therefore, the SPTNP service discipline is not a good choice when the tail of the STR distribution is very heavy, and it is a better choice over the PS when the tail of the STR distribution is light. The behavior of the BS of a CRN is analogues to a service station with random breakdowns, in a machine repair problem in the operations and research studies. Therefore, this analysis can be used in machine repair problems with service station subjected to breakdown, considering the same service disciplines at the service station.

We have studied the interactive data service over CRNs by modeling the system as two cascaded queues with feedback. It is shown that the request arrival process at the BS for interactive data traffic is close to the Poisson process, given Poisson session arrivals and longer (shorter) mean channel available (unavailable) durations in comparison with the mean STR. A relationship between the mean response time of elastic data traffic and the interactive data traffic is obtained under the Poisson approximation of the request arrivals at the BS. As an interactive data traffic session represents the behavior of a Web traffic user, the obtained relationship between the mean response times of interactive and elastic data traffic can be used for QoS provisioning in Web traffic over CRNs.

5.2 Future Works

In this thesis, we consider a time-slotted primary network in which the PU activities are perfectly synchronized with the time slots. Further, we assume the capability of the SUs to synchronize with the time-slots and to carry-out perfect channel sensing. However, the sensing errors are inevitable in the practical networking scenarios. If the presence of a PU is not detected at the spectrum sensing stage (by SUs), it is denoted as a missed detection error, and it can be caused by channel fading. The missed detections lead the SUs to transmit simultaneously with the PUs, which may cause harmful interference to the PUs. As a fundamental requirement in CRNs, it is required to limit the interference to the PUs below a predefined threshold by minimizing the missed detection errors. The interference to the PUs can be quantified by the probability, P_{md}, of missed detection and the duration of interference. In the network under consideration, the duration of interference is the time-slot duration[1], which is a fixed quantity. Therefore, the interference to the PUs is characterized by P_{md}, which should be controlled below a certain threshold, P^{*}_{md}, to satisfy the condition on interference with the PUs. The longer the sensing duration, the more accurate the sensing decision and the shorter the time available for information packet transmission [76]. There is a trade-off between the sensing accuracy and

[1] When the primary network is time-slotted, as the worst case scenario, whole information frame is assumed to be lost due an interference caused by the secondary network. However, the actual interference duration depends on how the information packets are organized and transmitted within the time-slot (by a PU).

the spectrum time available for the secondary network. In the voice and data networks under consideration, the shorter the time available for information packet transmission, the smaller the n_{pk} and L_{pk}. It is shown in Section 2.6 that the voice capacity monotonically increases with n_{pk}. Further, in Section 3.5 it is shown that the longer the STR, the longer the mean response time. The longer the L_{pk}, the shorter the STR. Therefore, the longer the L_{pk}, the shorter the mean response time and the larger the n_{pk}, the larger the voice capacity. Further, the smaller the probability, P_{fa}, of false alarm (the accurate the sensing decision), the larger the amount of spectrum opportunities available for the secondary network. The larger the amount of spectrum opportunities available for the secondary network, the larger the voice capacity and the shorter the mean response time. For the optimum performance of the systems under consideration, the voice capacity should be maximized and the mean response time of elastic/interactive data traffic has to be minimized, while satisfying the constraint $P_{md} \leq P^{*}_{md}$.

In a CRN network with a common channel (known as common control channel) to share the sensing decisions, the sensing decisions of multiple SUs can be combined to obtained the final sensing decision. This procedure is known as cooperative spectrum sensing [77], which results in a better sensing accuracy. It is an interesting research area in the context of CRNs. Even with the application of cooperative spectrum sensing, missed detections (interference with the PUs) are inevitable. In the centralized CRNs under consideration, cooperative spectrum sensing can be accomplished via a control channel, by incorporating sensing decisions of multiple SUs at the BS to make the final decision. The BS can transmit the final sensing decision back to the SUs. In this way, the consistency in spectrum availability information can be achieved. In the presence of cooperative spectrum sensing, the missed detections and the false alarms probabilities are reduced. However, cooperative spectrum sensing can only be accomplished at the expense of available spectrum time for the secondary network, which has a negative impact on n_{pk} and L_{pk}. There is a trade-off between the gain in the spectrum availability by reducing the probability of false alarms and the spectrum time lost due to the cooperation.

So far, we have discussed the scenarios with a time-slotted primary network (synchronous primary network). However, it is important to study the CRNs operating over asynchronous primary networks. Even though the primary network is not synchronized, the secondary network can operate in a time-slotted manner with the accomplishment of periodic channel sensing. In this type of CRNs, in addition to the missed detections (occur at the sensing periods), the presence of a PU in between the finishing and stating time instances of two adjacent sensing periods, respectively, can cause interference to the primary network. Therefore, the interference to the primary network is larger than that when the primary network is synchronized (time-slotted). Due to the asynchronous nature of the primary network, the interference duration is also different from that when the primary network is synchronized . The interference duration due to a missed detection depends on the probability distribution of the duration that a PU occupies the channel, whereas the interference duration due to the presence of a PU in between the finishing and the starting time instances, receptively, in adjacent time-slots depends on the probability distribution of the duration that a PU is idle. Given the corresponding probability distributions, the probability of interference to the primary network can be analyzed. In order to optimize the network performance, the sensing period and the sensing duration need to be fine tuned such that the voice capacity is maximized (or the mean response time of elastic/interactive data is minimized), while the interference with the primary network is kept under control. For the simplicity of our analysis, we did not incorporate the sensing errors and the interference to the primary users in our analysis. However, as the spectrum sensing errors are inevitable, and it is worthwhile to study their effects on the voice capacity and the mean response time of elastic/interactive data traffic.

In our research, the capacity analysis of on-off voice traffic flows is carried out only for the centralized network with FCFS service discipline, which is the ideal situation. As the capacity of on-off voice traffic is higher that of the constant-rate voice traffic, it is important to extend the on-off voice capacity analysis to the other channel access schemes as well. In [78], we have provided some insights on the packet level analysis of the on-off voice traffic in distributed CRNs. However, further research is required to do a comprehensive study on the on-off voice capacity of the distributed CRNs.

In the capacity and mean response time analysis, the channel available durations are geometrically (or exponentially) distributed. However, these durations can follow different probability distributions based on the behavior of the PUs. Analysis of the system capacity and mean response time for different probability distributions will be complex, and it would be a significant research contribution to extend the analysis considering different channel availability distributions. Further, in the mean response time analysis, we have only considered the Weibull distributed file lengths in our comparisons. However, the analysis can be used with any file length distribution. It is interesting to compare the four service disciplines under different file length distributions such as uniform distribution and generalized Pareto distribution [79].

In our analysis, we consider the mean of the response time as the QoS parameter. However, some knowledge about the probability distribution of the response time can provide more information on the service quality. Therefore, it is interesting to study the probability distribution of the response time of elastic/interactive data traffic.

Bibliography

[1] Federal Communication Commission, “Spectrum policy task force,” Rep. ET docket no. 02135, Nov. 2002.

[2] M. McHenry, P. Tenhula, D. McCloskey, D. Roberson, and C. Hood, “Chicago spectrum occupancy measurements & analysis and a long-term studies proposal,” in *Proc. of ACM TAPAS ’06*, Aug. 2006.

[3] J. Mitola III, “Cognitive radio: an integrated agent architecture for software defined radio,” *Dissertation, Doctor of Technology, Royal Institute of Technology*, Sweden, May 2000.

[4] J. Mitola III and G. Q. Maguire Jr., “Cognitive radio: making software radios more personal,” *IEEE Pers. Commun. Mag.*, vol. 6, no. 4, pp. 13-18, Aug. 1999.

[5] J. Mitola, “Cognitive radio for flexible mobile multimedia communications,” in *Proc. of IEEE MOMUC ’99*, pp. 3-10, 1999.

[6] L. Wang, A. Chen, and D. Wei, “A cognitive MAC protocol for QoS provisioning in overlaying ad hoc networks,” in *Proc. of IEEE CCNC ’07*, pp. 1139-1143, Jan. 2007.

[7] H. Su and X. Zhang, “Cross-layer based opportunistic MAC protocols for QoS provisioning over cognitive radio wireless networks,” *IEEE J. Select. Areas Commun.*, vol. 26, no. 1, pp. 118-129 , Jan. 2008.

[8] B. Ishibashi, N. Bouabdallah, and R. Boutaba, “QoS performance analysis of cognitive radio-based virtual wireless networks,” in *Proc. of IEEE INFOCOM ’08*, April 2008.

[9] X. Hong, C.-X. Wang, H.-H. Chen, and J. Thompson, “Performance analysis of cognitive radio networks with average interference power constraints,” in *Proc. of IEEE ICC ’08*, pp. 3578-3582, May 2008.

[10] M. Zhang, S. Jiang, G. Wei, and H. Wang, “Performance analysis of the cognitive radio network with a call level queue for secondary users,” in *Proc. of IEEE WiCOM ’09*, Sep. 2009.

[11] M. Zhang, S. Jiang, G. Wei, and H. Wang, “Performance analysis of cognitive radio networks against secondary user’s policies,” in *Proc. of IEEE ACIS ’09*, June 2009.

[12] L. Ruan and V. K. N. Lau, “Power control and performance analysis of cognitive radio systems under dynamic spectrum activity and imperfect knowledge of system state,” *IEEE Trans. Wireless Commun.* vol. 8, no. 9, pp. 4616-4622, 2009.

[13] W. Wong and H. Foh, “Analysis of cognitive radio spectrum access with finite user population,” *IEEE Commun. Letters*, vol. 13, no. 5, May 2009.

[14] S. Akin and M. C. Gursoy, “Performance analysis of cognitive radio systems under QoS constraints and channel uncertainty,” *IEEE Trans. Wireless Commun.* vol. 10, no. 9, pp. 2883-2895, 2011.

[15] Y.-U. Jang, “Performance analysis of cognitive radio networks based on sensing and secondary-to-primary interference,” *IEEE Trans. Signal Proc. of* vol. 59, no. 11, pp. 5663-5668, 2011.

[16] P. Wang, D. Niyato, and H. Jiang, “Voice service support over cognitive radio networks,” in *Proc. of IEEE ICC ’09*, June 2009.

[17] P. Wang, D. Niyato, and H. Jiang, "Voice service capacity analysis for cognitive radio networks," *IEEE Trans. Veh. Tech.*, vol. 59, no. 4, pp. 1779-1790, May 2010.

[18] H. Lee and D.-H Cho, "VoIP capacity analysis in cognitive radio system," *IEEE Commun. Letters*, vol. 13, no. 6, pp. 393-395, June 2009.

[19] H. Lee and D.-H. Cho, "Capacity improvement and analysis of VoIP service in a cognitive radio system," *IEEE Trans. Veh. Technol.*, vol. 59, pp. 1646-1651, May 2010.

[20] F. A. Cruz-Perez, M. E. Rivero-Angeles, G. Hernandez-Valdez, and S. L. Castellanos-Lopez, "Joint call and packet level performance analysis of CAC Strategies for VoIP traffic in wireless," in *Proc. of IEEE GLOBECOM '11*, Dec. 2011.

[21] S. L. Castellanos-Lopez, F. A. Cruz-Perez, M. E. Rivero-Angeles, and G. Hernandez-Valdez, "Impact of the primary resource occupancy information on the performance of cognitive radio networks with VoIP traffic," in *Proc. of IEEE CROWNCOM '12,* pp. 338-343, June 2012.

[22] K. Koufos, K. Ruttik, and R. Jantti, "Voice service in cognitive networks over the TV spectrum," *IET Commun. Magazine*, vol. 6, no. 8, pp. 991-1003, 2012.

[23] S. Lien and K.-C. Chen, "Statistical Traffic Control for Cognitive Radio Empowered LTE-Advanced with Network MIMO" in *Proc. of INFOCOM '11*, pp. 80-84, April 2011.

[24] H. Qin and Y. Cui, "Cross-layer design of cognitive radio network for real time video streaming transmission," in *Proc. of IEEE CCCM '09*, Aug. 2009.

[25] D .Hu and S. Mao, "Streaming scalable videos over multi-hop cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3501-3511, 2010.

[26] B. Guan and Y. He, "Optimal resource allocation for video streaming over cognitive radio networks," in *Proc. of IEEE MMSP '11*, Oct. 2011.

[27] M. Kartheek, R. Misra, and V. Sharma, “Performance analysis of data and voice connections in a cognitive radio network,” in *Proc. of IEEE NCC ’11*, pp. 1-5, Jan. 2011.

[28] A. M. Wyglinski, M. Nekovee, Y. T. Hou, “Cognitive Radio Communications and Networks: Principles and Practice,” *Elsevier*, Dec. 2009.

[29] W. Song and W. Zhuang, “Multi-class resource management in a cellular/WLAN Integrated network,” in *Proc. of IEEE WCNC ’07*, pp. 3070-3075, Mar. 2007.

[30] W. Song and W. Zhuang, “Resource allocation for conversational, streaming, and interactive services in cellular/WLAN interworking,” in *Proc. of IEEE GLOBECOM ’07*, pp. 4785-4789, Nov. 2007.

[31] S. Gunawardena and W. Zhuang, “Voice capacity of cognitive radio networks,” *Proc. of IEEE ICC ’10*, May 2010.

[32] S. Gunawardena and W. Zhuang, “Voice Capacity of Cognitive Radio Networks for Both Centralized and Distributed Channel Access Control,” *Proc. of IEEE GLOBECOM ’10*, Dec. 2010.

[33] S. Gunawardena and W. Zhuang, “Capacity analysis and call admission control in distributed cognitive radio networks,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 3110-3120, 2012.

[34] S. Gunawardena and W. Zhuang, “Service response time of elastic data traffic in cognitive radio networks with SPT service discipline,” in *Proc. of IEEE GLOBECOM ’12*, Dec. 2012.

[35] S. Gunawardena and W. Zhuang, “Service response time of elastic data traffic in cognitive radio networks,” *IEEE J. Select. Areas of Commun.*, vol. 31, no. 3, pp. 559-570, March 2013.

[36] L. Cai, Y. Xiao, X. Shen, L. Cai, and J. Mark, "VoIP over WLAN: voice capacity, admission control, QoS, and MAC," *Wiley InterScience, Int. J. Commun. Syst.*, pp. 491-508, 2006.

[37] Cisco, "Cisco visual networking index: global mobile data traffic forecast update, 2010-2015," Feb. 2011.

[38] Allot, "Allot mobile trends: global mobile broadband traffic report," Sep. 2010.

[39] A. Abdrabou and W. Zhuang, "Stochastic delay guarantees and statistical call admission control for IEEE 802.11 single-hop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3972-3981, Oct. 2008.

[40] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Select. Areas Commun.: Special Issue on Adaptive, Spectrum Agile and Cognitive Wireless Networks*, vol. 25, no. 3, pp. 589-600, April 2007.

[41] A. Bagwari and B. Singh, "Comparative performance evaluation of spectrum sensing techniques for cognitive radio networks," in *Proc. of IEEE CICN '12,* pp. 98-105, Nov. 2012.

[42] M. Timmers, A. Dejonghe, L. Van der Perre, and F. Catthoor, "A distributed multi-channel MAC protocol for cognitive radio networks with primary user recognition," in *Proc. of IEEE CROWNCOM '07*, pp. 216-223, Aug. 2007.

[43] M. Schwartz, *Broadband Integrated Networks*, Prentice Hall, 1998.

[44] C. Chang, *Performance guarantees in communication networks*, Springer-Verlag, 2000.

[45] A. Abdrabou and W. Zhuang, "Statistical QoS routing for IEEE 802.11 multi-hop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1542-1552, March 2009.

[46] P. Wang and W. Zhuang, "A token based scheduling scheme for WLANs supporting voice/data traffic and its performance analysis," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, part 1, pp. 1708-1718, May 2008.

[47] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless commun.*, vol. 2, no. 4, pp. 630-643, July 2003.

[48] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," in *Proc. of IEEE BROADNETS '04*, pp. 527-536, Oct. 2004.

[49] S. Shakkottai, "Effective capacity and QoS for wireless scheduling," *IEEE Trans. Automatic Control*, vol. 53, no. 3, April 2008.

[50] D. Wu and R. Negi, "Effective capacity channel model for frequency-selective fading channels," *SpringerLink J. Wireless Networks*, vol. 13, no. 3, pp. 299-310, July 2006.

[51] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058-3068, Aug. 2007.

[52] L. Liu and J. Chamberland, "On the effective capacities of multiple antenna Gaussian channels," in *Proc. of IEEE ISIT '08*, pp. 2583-2587, July 2008.

[53] D. Niyato and E. Hossain, "A game-theoretic approach to competitive spectrum sharing in cognitive radio networks," in *Proc. of IEEE WCNC '07*, pp. 16-20, March 2007.

[54] C.-S . Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automatic Control*, vol. 39, no. 5, pp. 913-931, May 1994.

[55] Z. Fan and P. Mars, "Effective bandwidth approach to connection admission control for multimedia traffic in ATM networks," *IEE Electronics Letters,* vol. 32, no. 16, pp. 1438-1439, 1996.

[56] E. Hossain and V. Bhargava, “Cognitive wireless communication networks,” *Springer US*, 2007.

[57] D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630-643, July 2003.

[58] L. D. Servi, “D/G/1 queues with vacations,” *Operations Research*, vol. 34, no. 4, pp. 619-629, Aug. 1986.

[59] L. Lin, H. Fu, and W. Jia, “An efficient admission control for IEEE 802.11 networks based on throughput analysis of unsaturated traffic,” in *Proc. of IEEE GLOBECOM ’05*, pp. 3017-3021, Dec. 2005.

[60] H. Jiang, L. Lai, R. Fan, and H. V. Poor, “Optimal Selection of channel sensing order in cognitive radio,” *IEEE Trans. Wireless Comm.*, vol. 8, no. 1, Jan. 2009.

[61] S. Deng, “Traffic characteristics of packet voice,” in *Proc. of IEEE ICC ’95*, vol. 3, pp. 1369-1374, June 1995.

[62] W. Song and W. Zhuang “Performance analysis of probabilistic multipath transmission of video streaming traffic over multi-Radio wireless devices,” *IEEE Trans. Wireless Comm.*, vol. 11 , no. 4, pp. 1554-1564, April 2012.

[63] W. Song “Resource reservation for mobile hotspots in vehicular environments with cellular/WLAN interworking,” *Springer EURASIP J. Wireless Comm. and Networks*, Jan. 2012.

[64] H. Li, “Impact of primary user interruptions on data traffic in cognitive radio networks Phantom jam on highway,” in *Proc. of IEEE GLOBECOM ’11,* pp. 1-5, Dec. 2011.

[65] J. Elias and F. Martignon, “Joint spectrum access and pricing in cognitive radio networks with elastic traffic,” in *Proc. of IEEE ICC ’10,* pp. 1-5, May. 2010.

[66] L. Massoulíe and J. W. Roberts, “Bandwidth sharing and admission control for elastic traffic,” *Telecommunication systems,* vol. 15, no. 1-2, pp. 185-201, June 2006.

[67] M. Lin, A. Wierman, and B. Zwart, “The average response time in a heavy-traffic SRPT queue,” *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 2, pp. 12-14, Sep. 2010.

[68] B. Avi-Itzhak and P. Naor, “Some queuing problems with the service station subject to breakdown,” *Operations and Research*, vol. 11, no. 3, pp. 303-320, May-June 1963.

[69] W. Song and W. Zhuang, “Multi-service load sharing for resource management in the cellular/WLAN integrated network,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, part 1, pp. 725-735, Feb. 2009.

[70] L. E. Schrage and L. W. Miller, “The queue M/G/1 with the shortest remaining processor time discipline,” *Operations and Research*, vol. 14, no. 4, pp. 670-684, July-Aug. 1966.

[71] A. Cobham, “Priority assignment in waiting line problems,” *Operations and Research*, vol. 2, no. 1, pp. 70-76, Feb. 1954.

[72] H. Nan, T. Hyon, and S. Yoo, “Distributed coordinated spectrum sharing MAC protocol for cognitive radio,” in *Proc. of IEEE DySPAN ’07*, pp. 240-249, April 2007.

[73] A. Hsu, D. Wei, and C. Kuo, “A cognitive MAC protocol using statistical channel allocation for wireless ad-hoc networks,” in *Proc. of IEEE WCNC ’07*, pp. 105-110, March 2007.

[74] J. Jia, Q. Zhang, and X. Shen, “HC-MAC: A hardware-constrained cognitive MAC for efficient spectrum management,” *IEEE J. Select. Areas Commun.*, vol. 26, no. 1, pp. 106-117, Jan. 2008.

[75] N. K. Shankaranarayanan, Z. Jiang, and P. Mishra, “Performance of a shared packet wireless network with interactive data users,” *ACM J. Mobile Networks and Applications*, vol. 8, no. 3, pp. 279-293, June 2003.

[76] Y. Pei, A. T. Hoang, Y.-C. Liang, “Sensing throughput trade-off in cognitive radio networks: how frequently should spectrum sensing be carried out?,” in *Proc. of IEEE PIMRC 2007*, pp. 5330-5335, Sep. 2007.

[77] R. K. Mallik and K. Letaief, “Optimization of cooperative spectrum sensing with energy detection in cognitive radio networks,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5761-5766, Dec. 2009.

[78] S. Gunawardena and W. Zhuang, “On-off voice capacity of single-hop cognitive radio networks with distributed channel access control,” *Proc. of IEEE ICC ’12*, June 2012.

[79] K. M. Rezaul and A. Pakštas, “Web traffic analysis based on EDF statistics,” in *Proc. of 7th Annual Post Graduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNet),* June 2006.

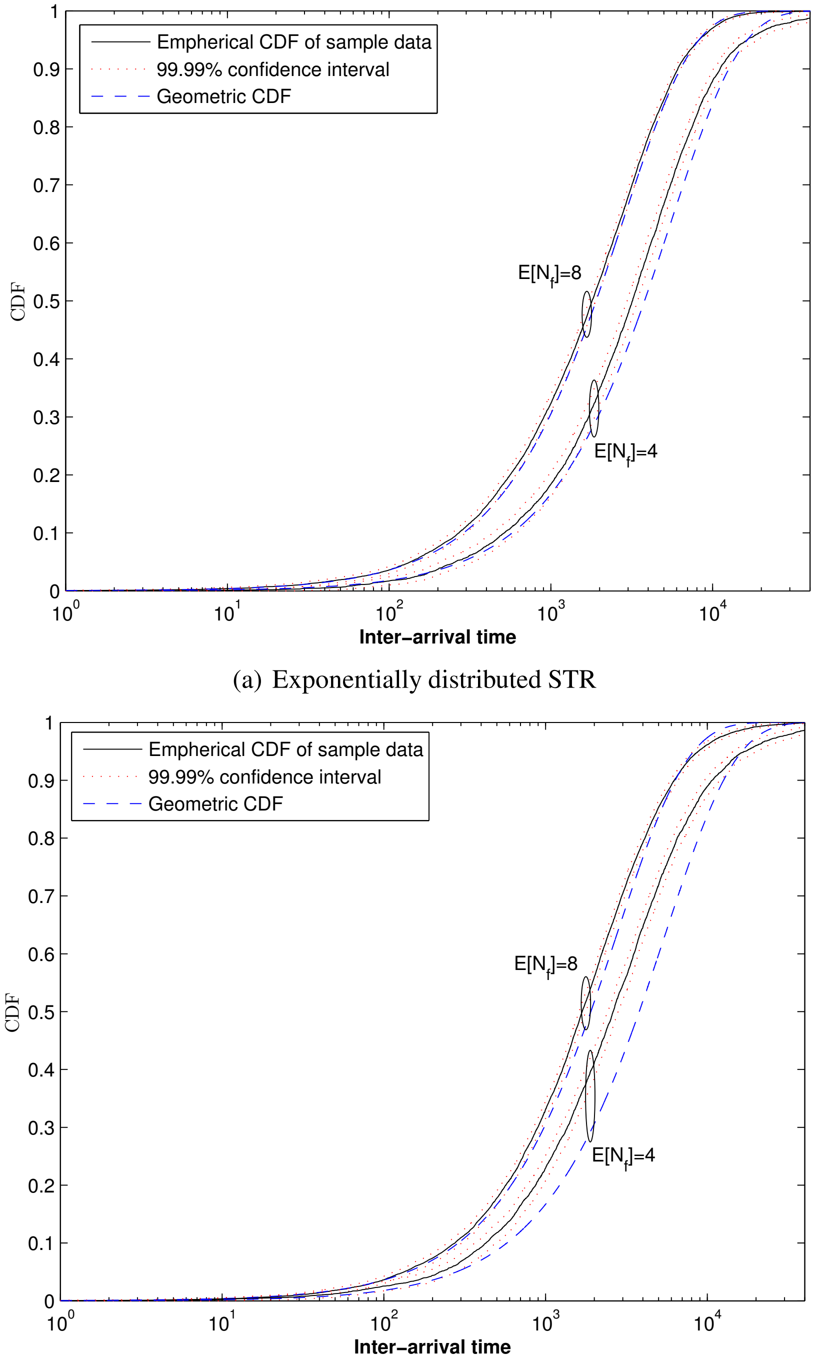
Appendix A

Mean Response Time of Interactive Data for Long Interruption Durations

This appendix presents numerical results for the comparison between the service request inter-arrival time distribution at the BS queue and the geometric distribution, and between the simulation and numerical results for $E[T_R]$, having longer T_{off} and shorter T_{on} in comparison with the mean STR.

The comparison between the service request inter-arrival time distribution at the BS queue and the geometric distribution is an extension to that presented in Subsection 4.4.2, and the same simulation setup is used in this comparison. Figs. A.1-A.4 present a comparison between the service request inter-arrival time distributions of the BS queue for multi-file data sessions and the geometric distribution (in terms of the CDF) having mean $1/E[N_f]\lambda_0$ for T_{on}=50 and T_{off}=500 time-slots. We keep the mean STR $E[\tau]$=200 time-slots and mean thinking duration $E[T_T]$=5,000 time-slots for all the four service disciplines, respectively.

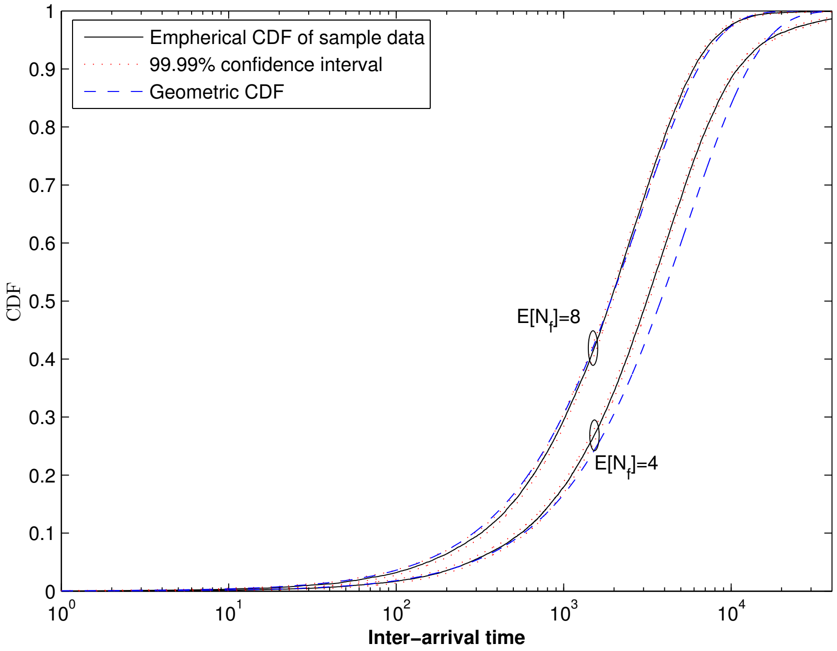
Similar to that observed in Subsection 4.4.2, the larger the $E[N_f]$, the closer the match between the service request inter-arrival time distribution of the BS queue and the geometric distribution.



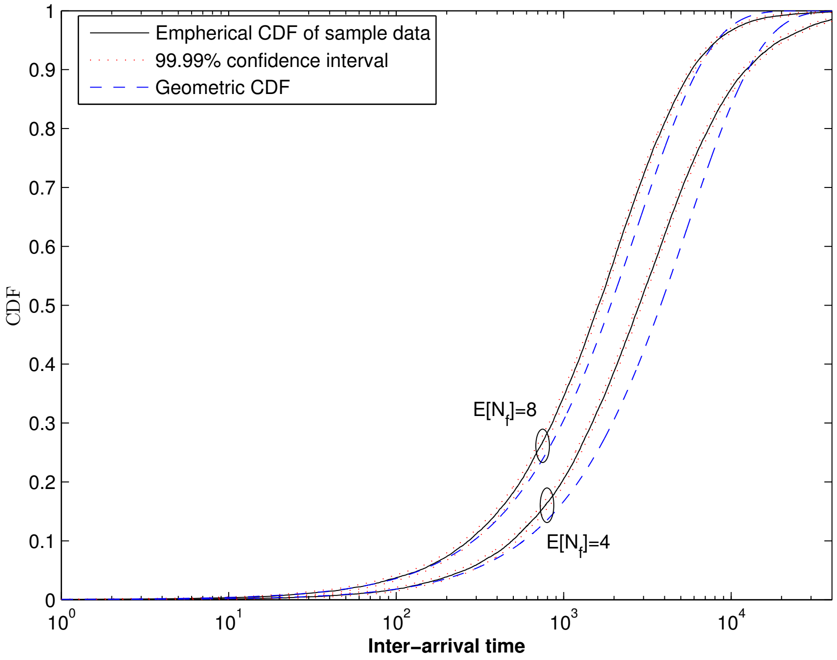
(a) Exponentially distributed STR

(b) Heavytail distributed STR

Figure A.1: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the PS service discipline having T_{on} = 50 and T_{off} = 500 time-slots.

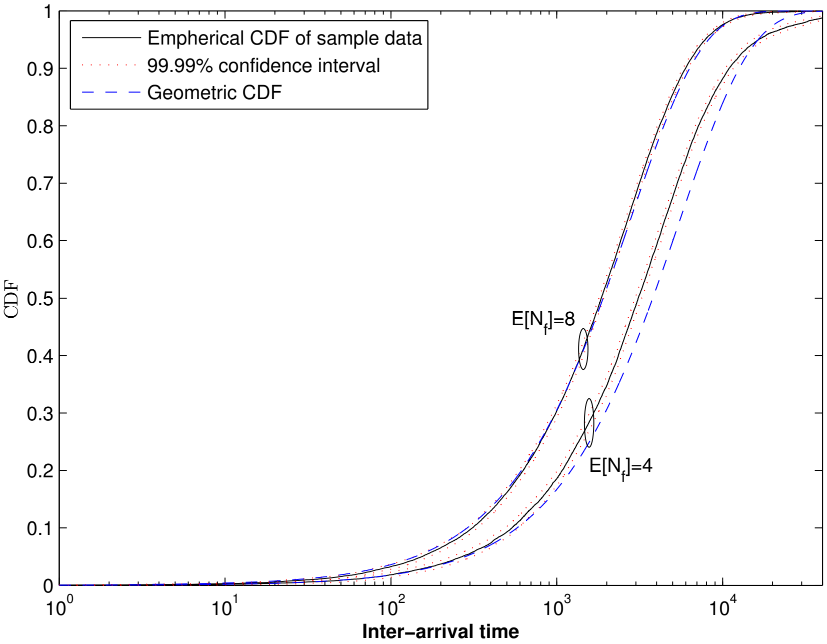


(a) Exponentially distributed STR

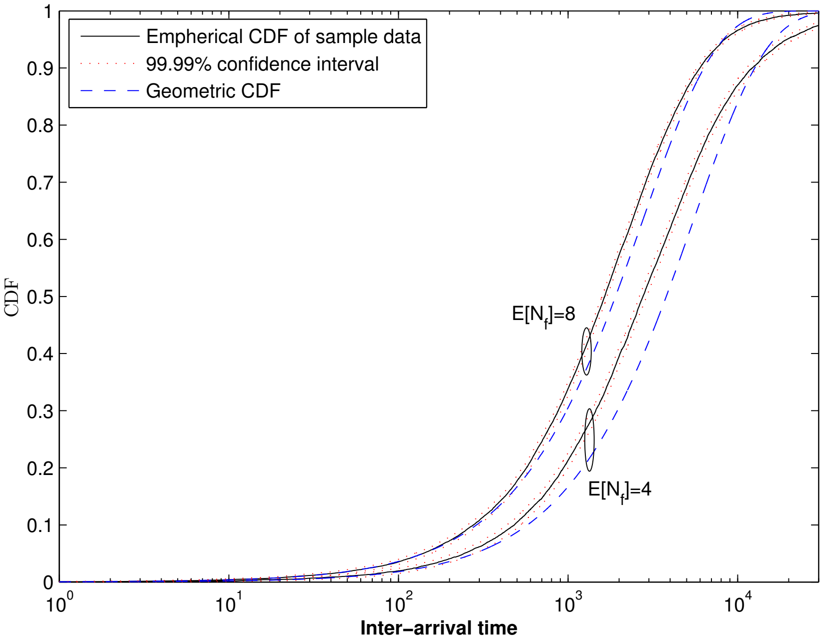


(b) Heavytail distributed STR

Figure A.2: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the SRPT service discipline having $T_{on} = 50$ and $T_{off} = 500$ time-slots.

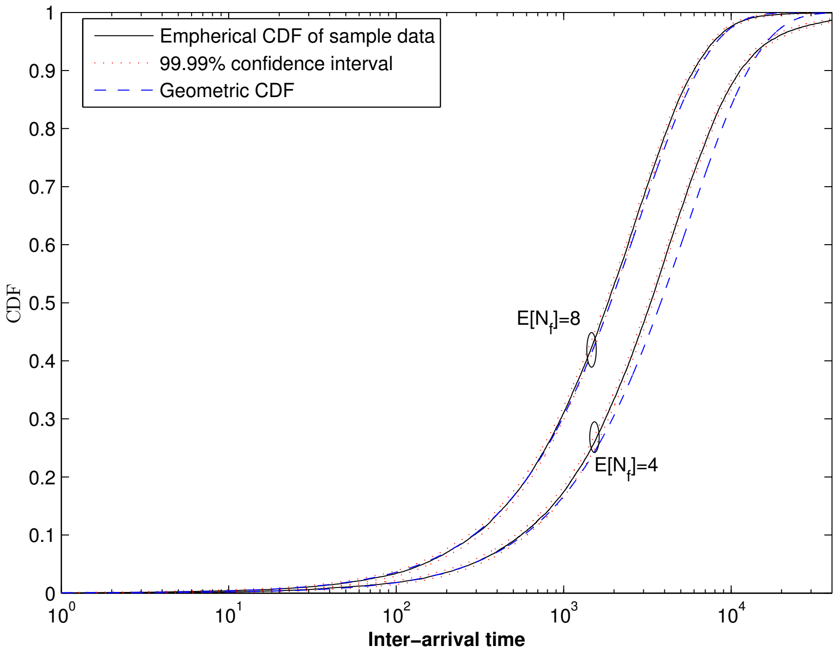


(a) Exponentially distributed STR

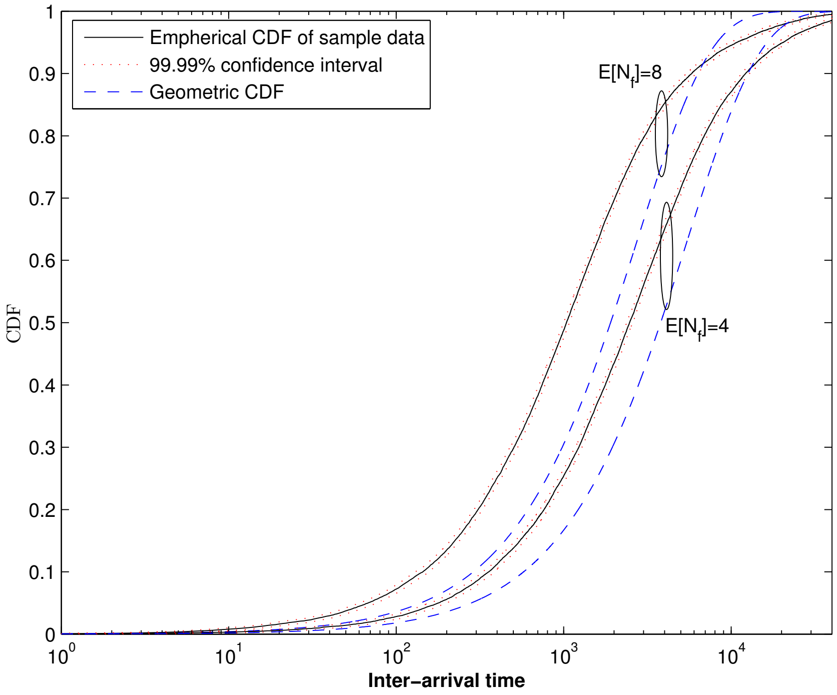


(b) Heavytail distributed STR

Figure A.3: Comparison between the service request inter-arrival time distributions and the geometric distribution having mean equal to $1/E[N_f]\lambda_0$ for the SPTWP service discipline having $T_{on} = 50$ and $T_{off} = 500$ time-slots.



(a) Exponentially distributed STR



(b) Heavytail distributed file size

Figure A.4: Comparison between the service request inter-arrival time distributions and the geometric distribution with mean equal to $1/E[N_f]\lambda_0$ for the SPTNP service discipline having $T_{on} = 50$ and $T_{off} = 500$ time-slots.

The deviation between the two distributions is larger when $T_{on} = 50$ and $T_{off} = 500$ than that when $T_{on} = 100$ and $T_{off} = 200$ and $T_{on} = 200$ and $T_{off} = 100$. Therefore, it is clear that the longer the T_{off} and the shorter the T_{on} when compared with the mean STR, the larger the deviation of the service request inter-arrival time distribution from the geometric distribution. Figs. A.5-A.8 show the variation of the mean response time with $E[N_f]$ obtained from the simulations and the analytical results (i.e., the mean response time obtained using the equations derived for the elastic data traffic by setting the $\lambda=E[N_f]\lambda_0$) for all the service disciplines, respectively.

It can be observed that the smaller the $E[N_f]$, the closer the match between the simulation results and the analytical results. However, due to the longer T_{off} and shorter T_{on} when compared with that used in Subsection 4.4.2, the deviation between the simulation results and the analytical results are larger than that observed in Figs. 4.8-4.15. The larger deviations between the mean response time (simulation and analytical) results are due to the larger devi-

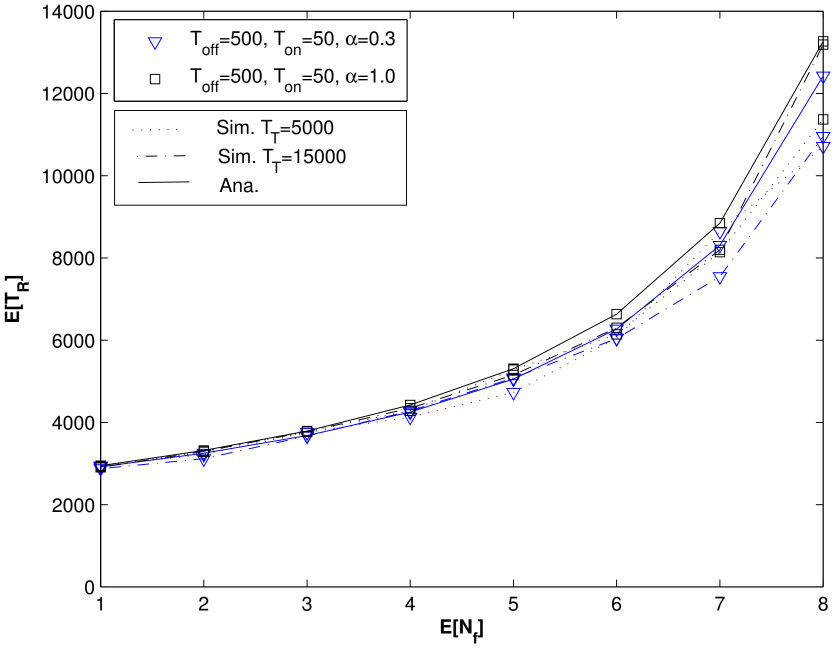


Figure A.5: Variation of the mean response time with $E[N_f]$ for the PS service discipline.

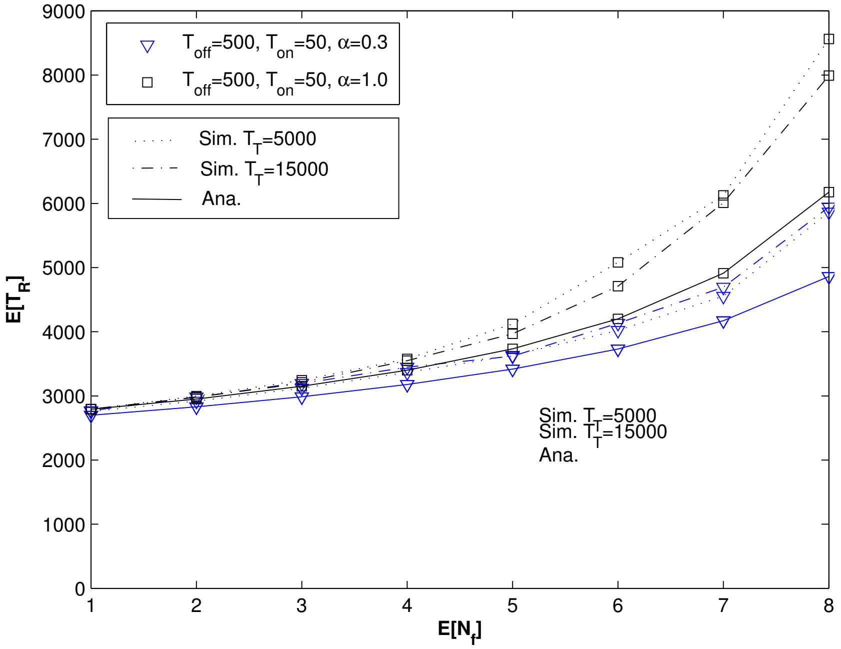


Figure A.6: Variation of the mean response time with $E[N_f]$ for the SRPT service discipline.

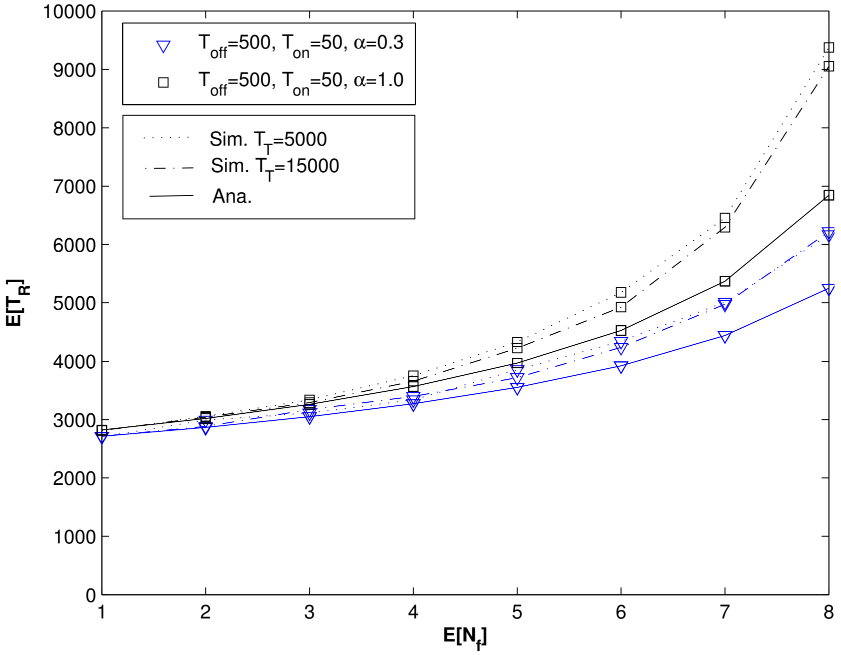


Figure A.7: Variation of the mean response time with $E[N_f]$ for the SPTWP service discipline.

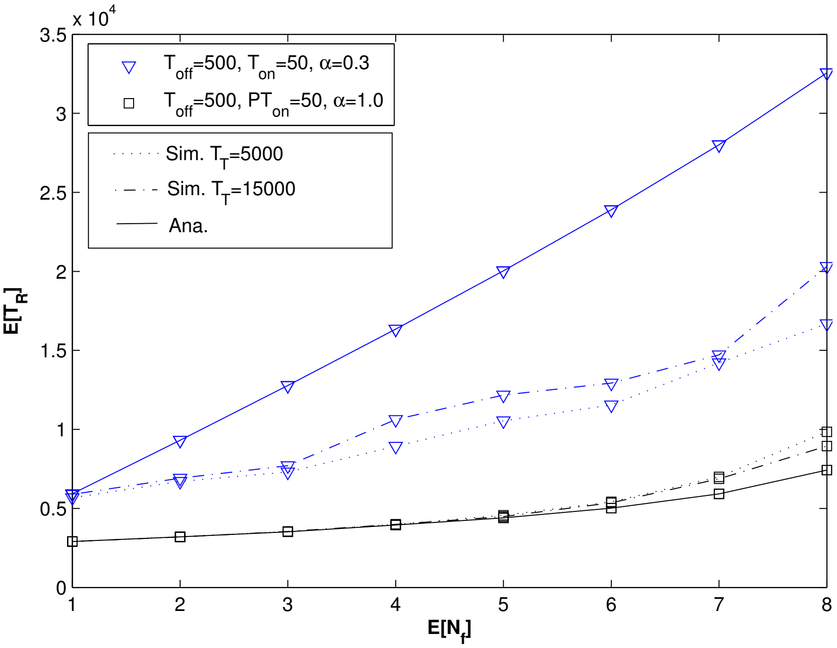


Figure A.8: Variation of the mean response time with $E[N_f]$ for the SPTNP service discipline.

ations between the service request inter-arrival time distributions and the expected geometric distribution, which is illustrated in Figs. A.1-A.4.