Service Response Time of Elastic Data Traffic in Cognitive Radio Networks

Subodha Gunawardena, *Student Member, IEEE*, and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—Quality of service (QoS) support over cognitive radio networks (CRNs) is challenging due to the random spectrum availability. Elastic data traffic is a popular service whose service response time is an important QoS parameter. We analyze the mean response time of elastic data traffic service operating over a single channel time-slotted centralized CRN under three main service disciplines, namely, shortest processing time without preemption (SPTNP), shortest processing time with preemption, and shortest remaining processing time, in comparison with the processor sharing (PS) service discipline. It is shown that the SPTNP is a better choice over the PS service discipline when the traffic load is high, and that the preemption reduces the mean response time when the data file size (service time requirement) follows a heavy tailed distribution. The response time analysis can be used for call admission control to ensure service satisfaction.**

***Index Terms*—Cognitive radio network, elastic data traffic, quality-of-service, response time, shortest processor time, shortest remaining processor time, processor sharing.**

I. I

The spectrum underutilization [2] is becoming a major setback for the development of next generation wireless networks. Among proposed solutions, the concept of cognitive radio networks (CRNs) [3][4] has become a popular choice due to its flexibility and adaptability to use in any available frequency band. It has been well accepted within the wireless communications research community to explore the underutilized portions of the radio spectrum using new-generation smart programmable radios, without harmfully interfering with the licensed primary users (PUs). Early research studies on CRNs mainly focused on physical layer aspects such as spectrum sensing [5] and power/rate controlling [6], and link layer aspects such as channel access coordination [7]. Further, the unpredictable nature of the opportunistic channels limits the research work on CRNs to best effort services without any strict quality of service (QoS) requirements. As the demand for wireless multimedia services keeps on increasing, supporting QoS-aware services over CRNs is essential.

Research studies on voice [8]-[10] and video [11] transmission over CRNs focus on capacity/delay analysis and channel access. 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 (SU) and primary network operators are studied in [12] and [13], 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. The response time of a service request (file) is defined as the time elapsed from the instant it is placed by a user to the instant that the service (file transmission) is completed.

A. Related Work

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. A performance analysis of elastic traffic in non-cognitive networks is carried out in [14] where the rate of flows adjusts to fill available bandwidth. Different bandwidth sharing techniques based on maximum throughput, min-max fairness, proportional fairness, and weighted fairness are considered in the analysis. The response time evaluation of elastic data traffic flows is studied for cellular/WLAN (wireless local area network) integrated networks in [15]-[17]. In the studies, the network supports streaming and elastic data traffic flows. In [15] a service request supports only one data file, whereas in [16][17] a service request establishes a data session which may have a number of data files with an exponentially distributed thinking time in between adjacent file transfers. The shortest remaining processor time (SRPT) and processor sharing (PS) service disciplines are considered in [15] and [16][17], respectively. The mean response time approximation in the SRPT service discipline under a heavy traffic condition is given in [18] (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 secondary users (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 [20], the mean throughput and delay of transmission control protocol (TCP) and constant bit rate connections are analyzed for CRNs with on-off PU behaviors.

Manuscript received April 15, 2012; revised September 1, 2012; accepted October 22, 2012. The editor-in-chief coordinating the review of this paper and approving it for publication was Ying-Chang Liang.

The authors are with the Centre for Wireless Communications, Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1 (e-mail: shgunawa, wzhuang@uwaterloo.ca).

This work was supported by a research grant from the Natural Science and Engineering Research Council (NSERC) of Canada. This work is presented in part in a paper to be presented in IEEE Globecom 2012 [1].

B. Contribution

The contribution of this paper is three fold: i) We derive mathematical expressions for the 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 (SPTNP) [1], shortest processor time with preemption (SPTWP), and shortest remaining processing time, in comparison with the PS service discipline. The PU activities are considered to have an on-off behavior with on and off durations following exponential 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 service time requirement distributions 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 (CAC) 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.

This paper is organized as follows. We describe our system model, service disciplines, and the data traffic model in Section II. The response times of the three service disciplines in the centralized CRN under consideration are analyzed in Sections III, IV, and V, respectively, and Section VI presents that of the PS service discipline. Three service disciplines SPTNP, SPTWP, and SRPT are compared in Section VII. Simulation and numerical results are discussed in Section VIII and, finally, conclusions are drawn in Section IX.

II. S M

A. Channel model

The CRN under consideration consists of a base station (BS) and a number of SUs. All the SUs and the BS see the same spectrum opportunities (spectrum homogeneous). It operates over a time-slotted single-channel primary network with a constant time-slot duration. Availability of the channel in a particular time-slot is referred to as the channel state. The channel is in state 1 if it is available for the SUs, and state 0 otherwise. The channel state in the next time-slot is independent of the state in previous time-slots, given the channel state in the current time-slot. The probability of state transition from state $\boldsymbol{i}$ ($\in \{0,1\}$) to state $\boldsymbol{j}$ ($\in \{0,1\}$) is denoted by $\boldsymbol{S_{i,j}}$. This is a widely used method to model the behavior of PUs [8]-[10] due to its simplicity. Therefore, the durations of channel availability and busy ($\boldsymbol{D}$) periods are geometrically distributed with mean values $\boldsymbol{T_{on}}$ ($=\boldsymbol{1/S_{1,0}}$) and $\boldsymbol{T_{off}}$ ($=\boldsymbol{1/S_{0,1}}$), respectively. The probability, $\boldsymbol{\pi_1}$, of channel availability and the probability, $\boldsymbol{\pi_0}$, of channel non-availability are given by $\boldsymbol{\pi_1 = S_{0,1}/(S_{0,1} + S_{1,0})}$ and $\boldsymbol{\pi_0 = S_{1,0}/(S_{0,1} + S_{1,0})}$, respectively.

B. Service disciplines

The data files requested by the SUs are transmitted from the BS. The channel state is obtained by the SUs and the BS via spectrum sensing and the new requests are sent via a low-rate control channel. When an active SU requests a data file, the BS transmits packets of the file based on SPTNP, SPTWP, SRPT, and PS service disciplines. During each available time-slot, only one data user is being served and the size, $\boldsymbol{L_s}$, of a data chunk (packet) transferred during a time-slot is fixed for all available time-sots. We use the terms user and service request interchangeably to denote a service request of an SU.

1) 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 service time requirement (STR) in the queue will be served. If the channel becomes unavailable (interrupted) during the service of the current user, the service will be paused 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.

2) 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.

3) 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.

4) 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 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 [8]. Further research is necessary to develop an efficient distributed channel access scheme and to apply the service disciplines.

C. Data traffic model

The STR depends on the length, $\boldsymbol{L}$, of the requested data file. The service requests at the BS follow a Poisson arrival

process with mean arrival rate λ. The lengths of requested files are independently and identically distributed with a Weibull distribution which is common for Internet data traffic [8] in [15]. The probability density function (pdf), $f_L(\cdot)$, of file length L given by

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

where α is the shape parameter. Therefore, the pdf of the service time requirement, S, is $f_S(x) = L_s f_L(xL_s)$. The mean STR, $E[S]$ is given by $E[S]=\int_0^{\infty} u f_S(u)du$. The cumulative distribution function (CDF) of STR S is denoted by $F_S(x) = \int_0^{x} f_S(u)du$. Upon a service request, the whole data file is available at the BS without any delay.

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 λ_a $(= 1/S_{1,0})$ and $E[D]$ $(= 1/S_{0,1})$, respectively.

III. SPT

The conditional mean response time, $E[R|S = p]$, of a target request given its STR, S, equal to p is [22]-[24]

$$E[R|S = p] = E[W(p)] + E[X(p)] \tag{2}$$

where R is the response time of a target request, $W(p)$ 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, and $X(p)$ 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. As a new request arrives at the system with the STR exactly equal to that of the current user (p) occurs with a negligible probability, we neglect it for the clarity of presentation.

A. 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. 1. Define a `type p` request (`type p'` request) as a service request with original STR smaller (greater) than p [23]. `Type p` busy period is defined as a continuous time period during which `type p` requests are using the channel or being interrupted while using the channel. An illustration of `Type p` busy periods for a channel without interruptions is given in [23].

If the channel is not in a `type p` busy period, it is in a `type p` idle period. A `type p` busy period starts from a request with STR less than p which arrives during a `type p` idle period as illustrated in Fig. 2, and it lasts until there is no `type p` request in the system waiting to be served. A `type p` idle period is divided into two parts, namely `type p'` busy (`type` p_c idle) period and idle period. A `type p'` busy period is a continuous time period during which requests with original STR greater than p are using the channel or being interrupted while being served. An idle period is categorized into idle available (`type` p_a idle) period and idle interrupted (`type` p_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.

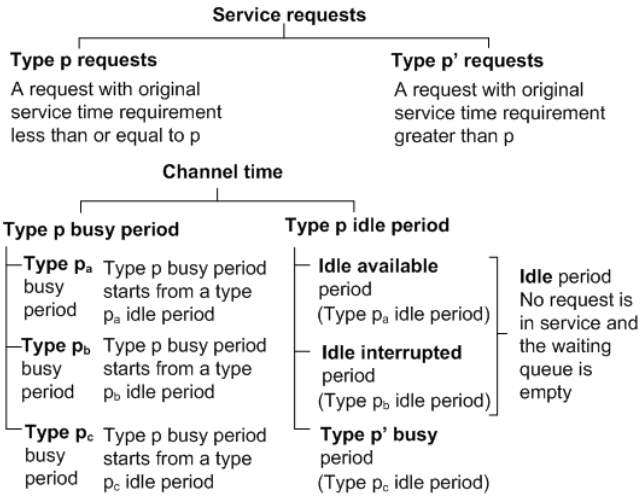

Fig. 1: Categorization of service requests and channel time

A `type p` busy period is categorized into three (`type` p_a, `type` p_b, and `type` p_c) busy periods based on the arrival period of the initiating `type p` request. The `type` p_a, `type` p_b, and `type` p_c busy periods initiate due to the arrival of a `type p` request during an idle available, idle interrupted, and `type p'` busy period, respectively. A `type` p_a busy period is initiated at the time of a `type p` request arrival during an idle available (`type` p_a idle) period. However, a `type` p_b busy period initiates just after the completion of an idle interruption (`type` p_b idle) period, and a `type` p_c busy period is initiated just after the completion of current `type p'` request (`type` p_c idle period). Examples for the initiation of `type` p_a, `type` p_b, and `type` p_c busy periods are given in Fig. 2.

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

B. Target request arriving in a `type p` *idle period*

If a target request arrives in an idle available (`type` p_a idle) period, it will get the channel access immediately. Therefore, the response time $R = X(p)$.

If the target request arrives in an idle interrupted (`type` p_b idle) period, first it waits until the interruption duration finishes. Further, any `type p` request arrivals during the idle interruption create a `type p` busy period. If so, the target request needs to wait until the end of the `type p` busy

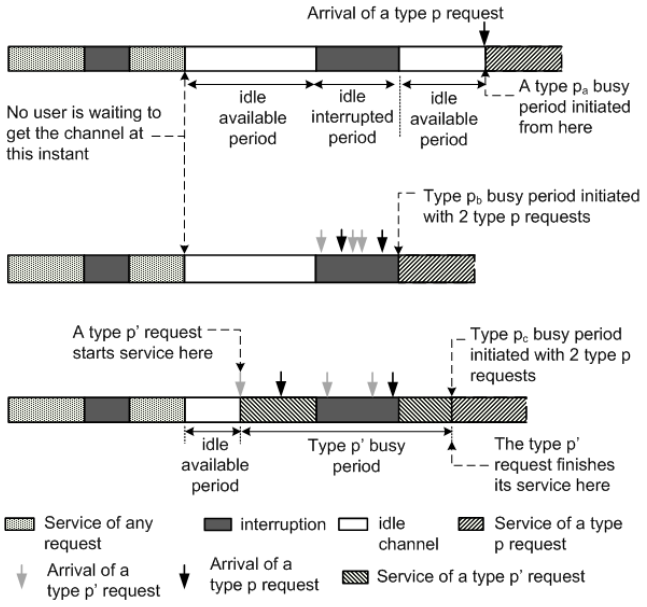

Fig. 2: Initiation of type p_a, type p_b, and type p_c busy periods.

period to get the service. Therefore, the mean waiting time, $E[W(p)] = E[\varphi_{idle,int}] + E[T_{busy,p_b}]P(N_b \geq 0)$, where $\varphi_{idle,int}$ is the residual time of the idle interruption, T_{busy,p_b} is the duration of a type p_b busy period, and N_b is the number of type p requests arrived during an idle interrupted period.

If the target request arrives in a type p' busy (type p_c idle) period, it waits first until the end of current service which is a type p' request. Further, any type p request arrivals during the service time (including the interruptions) of the current user create a type p busy period, and the target request needs to wait until the end of the type p busy period to get the channel access. Therefore, the mean waiting time, $E[W(p)] = E[\varphi_{p'}] + E[T_{busy,p_c}]P(N_c \geq 0)$, where $\varphi_{p'}$ is the residual time of a type p' request, T_{busy,p_c} is the duration of a type p_c busy period, and N_c is the number of type p requests arrived during the service time (including the interruption duration) of a type p' request.

C. Target request arriving in a type p busy period

If the target request arrives in a type p busy period, it waits until the end of the type p busy period to get the service. Therefore, the waiting time $E[W(p)] = E[\varphi_{busy,p}]$, where $\varphi_{busy,p}$ is the residual time of the type p busy period. A type p busy period can be any of type p_a, type p_b, and type p_c busy periods. A summary of the waiting times of a target request falling into different time periods are given in Table I. The first type p request that arrives in an idle available period initiates a type p_a busy period. Therefore, the number of type p requests at the initiation of a type p_a busy period is one. The probability of an incoming request initiating a type p_a busy period is $P_{idle,av}F_S(p)$, where $P_{idle,av}$ is the probability of the target request arriving in an idle available period.

A type p request with the shortest STR that arrives in an idle interrupted period initiates a type p_b busy period just after the interruption. However, at the initiation of the type p_b busy period, there may be more than one type p request waiting to get service. As the target request has to wait until all the type p requests finish their service, we can treat any of the type p requests arriving in an idle interrupted period as the initiating request of the type p_b busy period. The probability that an incoming service request initiates a type p_b busy period is then $P_{idle,int}F_S(p)$, 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 p_c busy period is $P_{busy,p'}F_S(p)$, where $P_{busy,p'}$ is the probability of the target request arriving in a type p' busy period. Therefore, the probability, I_p, that an incoming request initiates a type p busy period is given by $I_p = F_S(p)[1 - P_{busy,p}]$, where $P_{busy,p}$ is the probability of the target request arriving in a type p busy period. The conditional mean response time, $E[R|S = p]$ is given by

$$\begin{aligned} E[R|S=p] &= \left(E[\varphi_{idle,int}] + E[T_{busy,p_b}]P(N_b \geq 1)\right)P_{idle,int} \\ &+ \left(E[\varphi_{p'}] + E[T_{busy,p_c}]P(N_c \geq 1)\right)P_{busy,p'} \\ &+ E[\varphi_{busy,p}]P_{busy,p} + E[X(p)]. \end{aligned} \quad (3)$$

The mean response time of a target request can be evaluated by averaging (3) over the pdf, $f_S(\cdot)$, of the STR.

D. Mean type p busy period

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

$$T_{busy,p} = \sum_{k=0}^{\infty} T_k \quad (4)$$

TABLE I: Waiting time of a target request arriving in different time periods

Time period	Waiting time	SPTNP	SPTWP/ SRPT
Type p_a idle period	No waiting time. Immediately receives service (accesses the channel). $E[W(p)]=0$.	yes	yes
Type p_b idle period	Wait until the interruption is over. If there are any type p arrivals during the interruption, a type p_b busy period is generated, wait until the end of the type p_b busy period. $E[W(p)]=E[\varphi_{idle,int}]+E[T_{busy,p_b}]$.	yes	yes
Type p_c idle period	Wait until the service completion of the current (type p') user. If there are any type p arrivals during the service of current user, a type p_c busy period is generated, wait until the end of the type p_c busy period. $E[W(p)]=E[\varphi_{p'}]+E[T_{busy,p_c}]$.	yes	no
Type p busy period	Wait until the end of the ongoing type p busy period. $E[W(p)]=E[\varphi_{busy,p}]$	yes	yes

where $T_k = T_k' + \sum_{i=0}^{N_{d,T_k'}} D_i$ ($k \geq 1$), $T_k' = \sum_{i=0}^{N_{y,T_{k-1}}} Y_{k-1,i}$ ($k \geq 1$), $N_{y,T_{k-1}}$ is the number of type p request arrivals during the period T_{k-1}, $Y_{k-1,i}$ is the STR of the i^{th} ($i \in \{0, N_{y,T_{k-1}}\}$) type p request arrival during T_{k-1}, $N_{d,T_k'}$ is the number of interruptions during the period T_k', and D_i is the duration of the i^{th} ($i \in \{0, N_{d,T_k'}\}$) interruption arrived in T_k' ($D_0, Y_{k-1,0} = 0$ by definition). The time duration T_0' in T_0 is the total STR of the $N_{y,0}$ initiating type p requests of the type p busy period. An example for a type p busy period is illustrated in Fig. 3. Similar to the analysis given in [22], it can be shown that

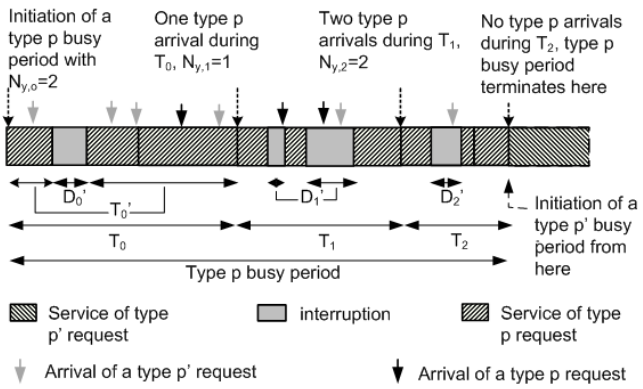

Fig. 3: An example of a type p busy period initiated with $N_{y,0} = 2$.

$$E[T_{busy,p}] = E\left[\sum_{k=0}^{\infty} T_k\right] = \frac{E[T_0]}{(1-b)} \tag{5}$$

where $E[T_0] = E[T_0' + \sum_{i=0}^{N_{d,T_0'}} D_i]$, $b = (1 + \lambda_a E[D])\lambda_p E[Y_p]$ is the fraction of time having a type p busy period, λ_p is the arrival rate of type p users, and $E[Y_p]$ ($= E[S|S < p]$) is the expected value of the STR of type p requests. It can be shown that

$$E[T_0] = \frac{1}{\pi_1} E[N_{y,0}] E[Y_p]. \tag{6}$$

The mean waiting time of a type p request due to type p_b busy period which falls during an idle interrupted period, $E[T_{p_b}]$, can be evaluated similar to (5) as

$$E[T_{p_b}] = \frac{b}{(1-b)} E[D] \tag{7}$$

where $\lambda_p E[D]$ is the mean number of type p request arrivals in an idle interrupted period. The mean duration $E[T_{p_b}]$ can be given by $E[T_{p_b}] = E[T_{busy,p_b}] P(N_b \geq 1) + 0 \cdot P(N_b = 0)$, and therefore, $E[T_{p_b}] = E[T_{busy,p_b}] P(N_b \geq 1)$. Further, the mean waiting time of a type p request due to type p_c busy period which falls during the service time of a type p' request, $E[T_{p_c}]$ ($= E[T_{busy,p_c}] P(N_c \geq 1)$), can be given similar to (7) as

$$E[T_{p_c}] = \frac{b}{(1-b)} E[X_{p'}] \tag{8}$$

where $E[X_{p'}]$ is the mean service time of a type p' request including the interruption periods during the service.

E. Mean residual times

The mean residual time, $E[\varphi_{busy,p}]$, of a type p busy period is given by [22]

$$E[\varphi_{busy,p}] = \frac{E[(T_{busy,p})^2 | N_{y,0} \geq 1]}{2E[T_{busy,p} | N_{y,0} \geq 1]} \tag{9}$$

where $E[(T_{busy,p})^2] = E[(\sum_{k=0}^{\infty} T_k)^2] = \frac{(1+b)}{(1-b)} E[\sum_{k=0}^{\infty} T_k^2]$. The residual time of a type p busy period exists only if a type p busy period is generated. Therefore, the first and second moments of $T_{busy,p}$ are conditioned on $N_{y,0} \geq 1$. With further manipulation, it can be shown that

$$(1-b^2) \sum_{k=0}^{\infty} E\left[T_k^2\right] = E\left[T_0^2\right] + \frac{E[T_0]}{(1-b)} \lambda_p \left(\frac{1}{{\pi_1}^2} E[{Y_p}^2] + \lambda_a E[D^2] E[Y_p]\right).$$

$$E[(T_{busy,p})^2 | N_{y,0} \geq 1] = \frac{E[T_0^2 | N_{y,0} \geq 1]}{(1-b)^2} + \frac{E[T_0 | N_{y,0} \geq 1]}{(1-b)^3} \lambda_p \left\{\frac{1}{{\pi_1}^2} E[{Y_p}^2] + \lambda_a E[D^2] E[Y_p]\right\} \tag{10}$$

$$E[T_{busy,p} | N_{y,0} \geq 1] = \frac{E[T_0 | N_{y,0} \geq 1]}{(1-b)} \tag{11}$$

where

$$E[T_0^2 | N_{y,0} \geq 1] = \frac{1}{{\pi_1}^2} E[T_0'^2 | N_{y,0} \geq 1] + \lambda_a E[D^2] E[T_0' | N_{y,0} \geq 1] \tag{12}$$

$$E[T_0 | N_{y,0} \geq 1] = \frac{1}{\pi_1} E[T_0' | N_{y,0} \geq 1]. \tag{13}$$

As a type p busy period may be one of the three busy period types (type p_a, type p_b, and type p_c), the mean STR of the initiating type p request, $E[T_0' | N_{y,0} \geq 1]$ is given by

$$\begin{aligned} E[T_0' | N_{y,0} \geq 1] = \frac{1}{I_p} \Big\{ & P_{idle,av} \int_0^p t f_S(t) dt \\ & + P_{idle,int} E[N_b | N_b \geq 1] \int_0^p t f_S(t) dt / F_S(p) \\ & + P_{busy,p'} E[N_c | N_c \geq 1] \int_p^{\infty} t f_S(t) dt / (1 - F_S(p)) \Big\} \end{aligned} \tag{14}$$

where N_b and N_c are the numbers of type p requests at the initiation instant of type p_b and type p_c busy periods, respectively. Similar to (14), $E[T_0'^2]$ is given by

$$\begin{aligned} E[T_0'^2 | N_{y,0} \geq 1] = \frac{1}{I_p} \Big\{ & P_{idle,av} \int_0^p t^2 f_S(t) dt \\ & + P_{idle,int} \Big(E[N_b | N_b \geq 1] \int_0^p t^2 f_S(t) dt / F_S(p) \\ & + \Big(E[{N_b}^2 - N_b | N_b \geq 1] \int_0^p t f_S(t) dt / F_S(p) \Big)^2 \Big) \\ & + P_{busy,p'} \Big(E[N_c | N_c \geq 1] \int_p^{\infty} t^2 f_S(t) dt / (1 - F_S(p)) \\ & + \Big(E[{N_c}^2 - N_c | N_c \geq 1] \int_p^{\infty} t f_S(t) dt / (1 - F_S(p)) \Big)^2 \Big) \Big\}. \end{aligned} \tag{15}$$

In (3), the mean residual time, $E[\varphi_{idle,int}]$, of an idle interrupted period is given by $E[\varphi_{idle,int}] = E[D^2]/2E[D]$. The mean residual time, $E[\varphi_{p'}]$, of a type p' request depends on the service time of a type p' request (which includes the interruption

periods during the service). The service time $X_{p'}$ is given by $X_{p'}=X_{p'}'+\sum_{i=0}^{N_{d,X_{p'}}'} D_i$, where $X_{p'}'$ is the original STR of a `type p'` request, $N_{d,X_{p'}'}$ is the number of interruptions occurred during $X_{p'}'$ and D_i is the duration of the i^{th} $(\in [0,\infty))$ interruption. The mean and the second moment of $X_{p'}'$ can be given by

$$E[X_{p'}] = \frac{1}{\pi_1}\int_p^\infty x f_S(x)dx \quad (16)$$

$$E[X_{p'}^2] = \frac{1}{\pi_1^2}\int_p^\infty x^2 f_S(x)dx + \lambda_a E[D^2]\int_p^\infty x f_S(x)dx. \quad (17)$$

The probabilities $P_{busy,p}$, $P_{busy,p'}$, and $P_{idle,int}$ are given by $P_{busy,p}=\lambda\int_0^p x f_S(x)dx/\pi_1$, $P_{busy,p'}=\lambda\int_p^\infty x f_S(x)dx/\pi_1$, and $P_{idle,int}=\pi_0(\pi_1-\lambda\int_0^\infty x f_S(x)dx)/\pi_1$, where fractions (proportions) of time having a `type p` and `type p'` request occupying the channel are given by $\lambda\int_0^p x f_S(x)dx$ and $\lambda\int_p^\infty x f_S(x)dx$, respectively. The mean residual time, $E[\varphi_{p'}]$, can be evaluated from the standard equation $E[\varphi_{p'}]=E[X_{p'}^2]/2E[X_{p'}]$.

IV. SPT

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 p preempts an ongoing `type p'` request to get the channel access, and initiates a `type p` busy period. From the view point of a `type p` user, all the interruptions of a `type p` idle period are idle interruptions and all the available durations of a `type p` idle period are idle available periods. Therefore, the channel available and interrupted periods of a `type p` idle period are denoted as `type p` idle available (`type` p_a idle) and `type p` idle interrupted (`type` p_b idle) periods, respectively, as illustrated in Fig. 4. Similar to the SPTNP, `type` p_a and `type` p_b busy

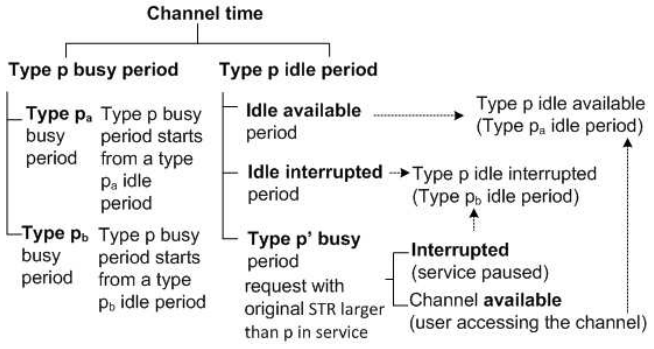

Fig. 4: Categorization of channel time for the SPTWP service discipline

periods are initiated by a `type p` arrival in `type` p_a and `type` p_b idle periods, respectively. However, with the SPTWP, there is no `type` p_c busy period. A target request with STR equal to p arriving in a `type p` idle available or `type p` 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` p_a and `type` p_b busy periods are illustrated in Fig. 5. A summary of waiting times of a target request falling into

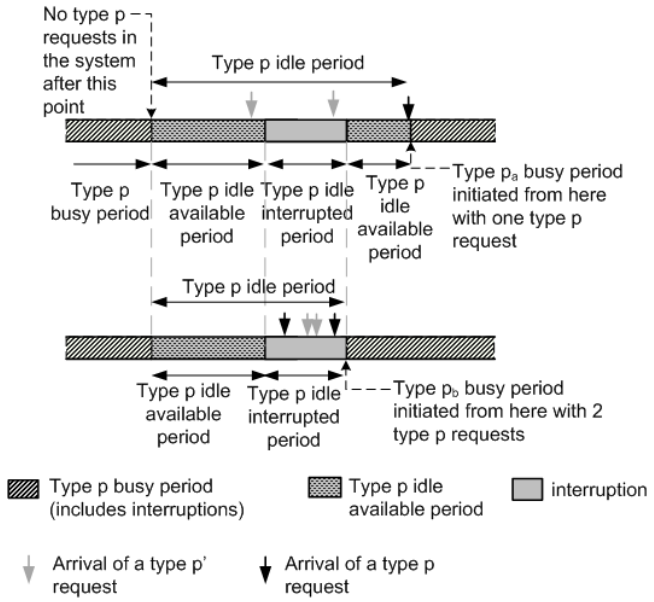

Fig. 5: Initiation of `type p` busy periods for the SPTWP service discipline

different periods is given in Table I. If the target request arrives in a `type p` idle available period, it gets the channel access immediately. Therefore, the response time, R, is the service time, $X(p)$. However, the service time, $X(p)$, is different from that of the SPTNP, since any `type p` request arrival during the service time of the target request can preempt the target request. The preempted durations are the durations of `type p` busy periods initiated during the original service time of the target request. An illustration of the service time $X(p)$ is given in Fig. 6, where $Z_p=p+\sum_{i=0}^{N_{d,p}} D_i$, $N_{d,p}$ is the number of interruptions occurred during p (in this example $N_{d,p}$=2), and D_i is the duration of the i^{th} interruption. Analysis of $E[X(p)]$

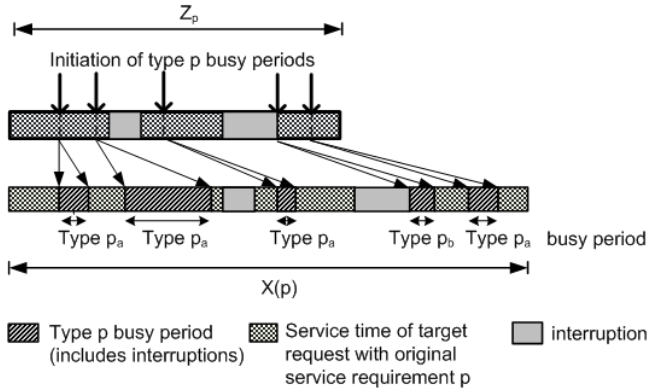

Fig. 6: Service time of the target request for the SPTWP service discipline

is similar to that of $E[T_{busy,p}]$ given in (5) and (6) with $E[Y_p]=p$ and $E[N_{y,0}]=1$. Therefore,

$$E[X(p)] = \frac{p}{\pi_1(1-b)}. \quad (18)$$

Similar to a target request arrival in an idle inter-

rupted period with the SPTNP, a target request arrival in a type p idle interrupted period with the SPTWP has $E[W(p)]=E[\varphi_{idle,int,p}]+E[T_{busy,p_b}]P(N_b \geq 0)$, where $E[\varphi_{idle,int,p}]$ ($=E[\varphi_{idle,int}]$) is the mean residual time of the type p idle interruption period.

If a target request arrives in a type p busy period, the mean waiting time is given by $E[W(p)]=E[\varphi_{busy,p}]$. The conditional mean response time, $E[R|S=p]$, is given by

$$E[R|S=p] = \left(E[\varphi_{idle,int}]+E[T_{busy,p_b}]P(N_b \geq 0)\right)P_{idle,int,p} + E[\varphi_{busy,p}]P_{busy,p}+E[X(p)] \quad (19)$$

where $P_{idle,int,p}$is the probability of the target request arriving in a type p idle interrupted period, $E[\varphi_{idle,int}]= E[D^2]/2E[D]$, and $E[\varphi_{busy,p}]$ is given in (7).

A target request can arrive in either of the type p busy periods (type p_a or type p_b). Therefore, $E[\varphi_{busy,p}]$ can be evaluated using (5), (6), (9)-(13) with

$$E[T_0'|N_{y,0}\geq 1] = \frac{1}{I_p}\Big\{P_{idle,av,p}\int_0^p t f_S(t)dt + P_{idle,int,p}E[N_b|N_b\geq 1]\int_0^p t f_S(t)dt/F_S(p)\Big\} \quad (20)$$

$$E[T_0'^2|N_{y,0}\geq 1] = \frac{1}{I_p}\Big\{P_{idle,av,p}\int_0^p t^2 f_S(t)dt + P_{idle,int,p}\Big(E[N_b|N_b\geq 1]\int_0^p t^2 f_S(t)dt/F_S(p) + \left(E[N_b^2-N_b|N_b\geq 1]\int_0^p t f_S(t)dt/F_S(p)\right)^2\Big)\Big\} \quad (21)$$

where $I_p=(1-P_{busy,p})F_S(p)$ is the probability of a request arrival initiating a type p busy period and $P_{idle,av,p}$is the probability of the target request arriving in a type p idle available period. Equations (20) and (21) differ from (14) and (15) in that (20) and (21) do not contain the components for a type p_c busy period. The probabilities $P_{busy,p}$, $P_{idle,int,p}$, and $P_{idle,av,p}$ are given by

$$P_{busy,p} = \lambda\int_0^p x f_S(x)dx/\pi_1 \quad (22)$$

$$P_{idle,av,p} = \pi_1 - \lambda\int_0^p x f_S(x)dx \quad (23)$$

$$P_{idle,int,p} = \pi_0\left(\pi_1-\lambda\int_0^p x f_S(x)dx\right)/\pi_1. \quad (24)$$

V. SRPT

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 p' request always initiates a type p busy period when its remaining STR reduces to p, and an incoming request with STR equal to p can preempt a type p' request only when the remaining STR of the type p' request is larger than p. In order to capture the difference, we alter the definition of type p busy period as a continuous time period during which services with the **remaining** STR **less** than p are using or being interrupted while using the channel. Similarly, the definition of type p' busy period is altered as a continuous time period during which type p' requests with the **remaining** STR **greater** than p are using or being interrupted while using the channel. The categorization of the time periods is illustrated in Fig. 7. As illustrated in

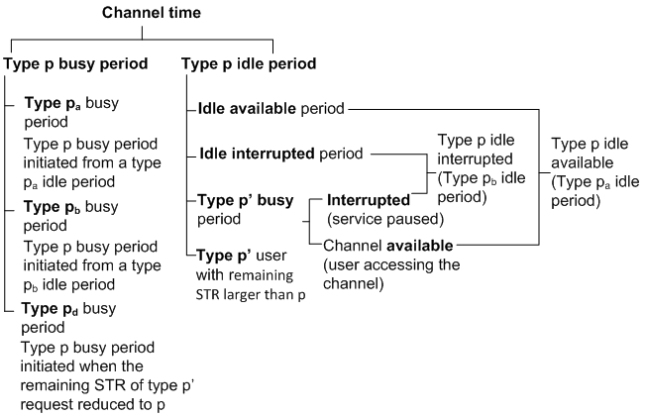

Fig. 7: Categorization of channel time for the SRPT service discipline

Fig. 7, a type p busy period may be any of type p_a, type p_b, and type p_d busy periods. Similar to that of the SPTWP service discipline, the type p_a and type p_b busy periods are initiated due to type p arrivals during p_a and type p_b idle periods, respectively. However, a type p_d busy period is initiated when the remaining STR of a type p' request becomes p. The waiting times of a target request with the original STR up to p are given in Table I. Similar to that of the SPTWP, the expression for the mean response time of a target request is given by

$$E[R|S=p] = \left(E[\varphi_{idle,int,p}]+E[T_{busy,p_b}]P(N_b\geq 0)\right)P_{idle,int,p} + E[\varphi_{busy,p}]P_{busy,p}+E[X(p)] \quad (25)$$

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

$$P_{busy,p} = \frac{\lambda\left[\int_0^p x f_s(x)dx + p(1-F_s(p))\right]}{\pi_1},\ P_{idle,int,p} = \pi_0\left(1-P_{busy,p}\right) \quad (26)$$

where the numerator in $P_{busy,p}$ is the fraction of time that requests with the remaining STR less than p occupies the channel (excluding the interruption durations). The evaluation of $E[\varphi_{busy,p}]$ in (25) is similar to that given in (14) with

$$E[T_0'|N_{y,0}\geq 1] = \frac{1}{I_p}\Big\{P_{idle,av,p}\int_0^p t f_S(t)dt + P_{idle,int,p}E[N_b|N_b\geq 1]\int_0^p t f_S(t)dt/F_S(p) + p[1-F_s(p)]\Big\} \quad (27)$$

$$E[T_0'^2|N_{y,0}\geq 1] = \frac{1}{I_p}\Big\{P_{idle,av,p}\int_0^p t^2 f_S(t)dt + P_{idle,int,p}\Big(E[N_b|N_b\geq 1]\int_0^p t^2 f_S(t)dt/F_S(p)\Big) + P_{idle,int,p}\Big(E[N_b^2-N_b|N_b\geq 1]\int_0^p t f_S(t)dt/F_S(p)\Big)^2 + p^2(1-F_s(p))\Big\} \quad (28)$$

where I_p is the probability of an incoming request initiating a `type p` (`type` p_a, `type` p_b, or `type` p_d) busy period. Therefore, $I_p = P_{idle,av,p}F_s(p) + P_{idle,int,p}F_s(p) + 1 - F_s(p) = 1 - P_{busy,p}F_s(p)$.

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 p (or equivalently the file length) of the target request into n units of duration Δt ($p = n \cdot \Delta t$) as illustrated in Fig. 8, where a `type` $(n-i)\Delta t$ busy period is similar to a `type p` 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 Y_i ($i \in \{1, 2, ..., n\}$) is the actual duration it takes to complete the i^{th} unit of Δt, and $X(p) = \sum_{i=1}^{n} Y_i$ [23]. We have $E[X(p)] = \sum_{i=1}^{n} E[Y_i]$. The duration Δt in Y_i is equivalent to T_0'

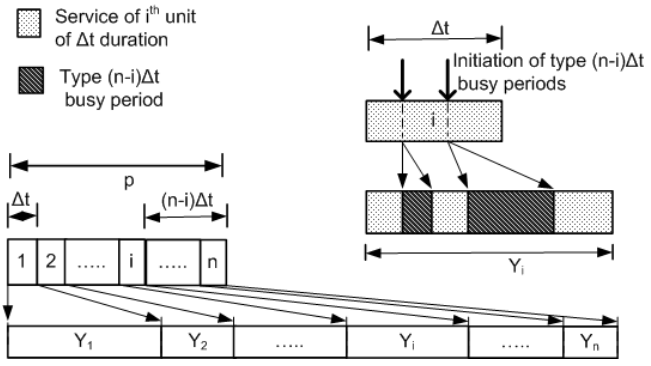

Fig. 8: Service time for the SRPT service discipline

in $T_{busy,p}$. Therefore, similar to the evaluation of $E[T_{busy,p}]$, $E[Y_i]$ is given by $E[Y_i] = \lim_{m\to\infty} \sum_{k=0}^{m+1} T_{i,k}$ with

$$E[T_{i,k+1}|T_{i,k}] = \frac{1}{\pi_1}(\lambda_{(n-i)\Delta t}T_{i,k})\frac{1}{F_s((n-i)\Delta t)}\int_0^{(n-i)\Delta t} x f_s(x)dx \quad (29)$$

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 $E[T_{busy,p}]$, the mean duration $E[Y_i] = E[T_0]/(1-b_i)$, where $E[T_0] = \Delta t/\pi_1$ and $b_i = \lambda_{(n-i)\Delta t}\int_0^{(n-i)\Delta t} x f_s(x)dx/\pi_1 F_s((n-i)\Delta t)$. 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, $E[X(p)]$, is obtained by making Δt very small ($\Delta t \to 0$) or equivalently n very large ($n \to \infty$) [1].

Duration of `type p` busy periods are independent and identically distributed with SPTWP and SRPT service disciplines, and the inter-arrival time of the new requests are memoryless. In the analysis of residual time of `type p` busy period given that the target request arrives in a `type p` busy period, we can ignore the `type p` idle periods, and consider the initiation of a `type p` busy period as a renewal process. However, durations of `type p'` services and `type p` 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.

VI. PS

The conditional mean response time of a target request operating over a network following exponentially distributed channel availability durations is given by [21]

$$E[R|S=p] = \frac{p}{\pi_1(1-\rho)} + \pi_0\frac{E[D^2]}{2E[D]} + \pi_0\frac{\rho E[D^2]}{2E[D]}\frac{2-\rho}{(1-\rho)^2}\left(1 - e^{-\frac{(1-\rho)p}{E[S]}}\right) \quad (30)$$

where $\rho = \lambda E[S]/\pi_1$ is the utilization factor (ratio between the mean arrival rate and mean service rate).

VII. C

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

$$E[R|S=p]_{NP} - E[R|S=p]_{WP} = (\Omega_{p'} - \pi_0\Omega_{p_b})(\rho - b) + b(E[\varphi_{busy,p}]_{NP} - E[\varphi_{busy,p}]_{WP}) - \frac{p}{\pi_1}\cdot\frac{b}{(1-b)} \quad (31)$$

where $\Omega_{p'} = E[\varphi_{p'}] + E[T_{busy,p_c}]P(N_c \geq 1)$, $\Omega_{p_b} = E[\varphi_{idle,int}] + E[T_{busy,p_b}]P(N_b \geq 1)$, and $\frac{p}{\pi_1}\cdot\frac{b}{(1-b)}$ is the service time difference between SPTNP and SPTWP for a data file requiring a service time equal to p. The terms $(\Omega_{p'} - \pi_0\Omega_{p_b})$ and $(E[\varphi_{busy,p}]_{NP} - E[\varphi_{busy,p}]_{WP})$ contain busy periods initiated from more than one request arrival, whereas $\frac{p}{\pi_1}\cdot\frac{b}{(1-b)}$ only contains the service time of a target request. The probability b monotonically increases with p. The terms $(\Omega_{p'} - \pi_0\Omega_{p_b})$, $(\rho - b)$, and $E[\varphi_{busy,p}]_{NP} - E[\varphi_{busy,p}]_{WP}$ decreases with p, and $\frac{p}{\pi_1}\cdot\frac{b}{(1-b)}$ increases with p. Therefore, the difference in the conditional mean response times $E[R|S=p]_{NP} - E[R|S=p]_{WP}$ varies from a very high positive value to a small negative value as p increases. When the file size Weibull (heavy tail) distributed as in (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 $E[R]_{NP} - E[R]_{WP}$ obtained by averaging $E[R|S=p]_{NP} - E[R|S=p]_{WP}$ over p is a positive value. The probability of having very large and very small values for $E[R|S=p]_{NP} - E[R|S=p]_{WP}$

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

increases with the tail heaviness in the file size distribution. As a result, $E[R]_{NP} - E[R]_{WP}$ increases with the tail heaviness in the file size distribution. Increment of the terms $(\Omega_{p'} - \pi_0\Omega_{p_b})$ and $(E[\varphi_{busy,p}]_{NP} - E[\varphi_{busy,p}]_{WP})$ with an increasing interruption duration is larger than that of $\frac{p}{\pi_1} \cdot \frac{b}{(1-b)}$. Therefore, the mean response time difference $E[R]_{NP} - E[R]_{WP}$ increases with the interruption duration. The difference between the conditional mean response times for SPTWP and SRPT is given by

$$\begin{aligned} E[R|S=p]_{WP} - E[R|S=p]_{SRPT} &= E[X(p)]_{WP} - E[X(p)]_{SRPT} \\ &+ b^*\pi_1 E[T_{p_b}] - b(E[\varphi_{busy,p}]_{SRPT} - E[\varphi_{busy,p}]_{WP}) \\ &- b^*(E[\varphi_{busy,p}]_{SRPT} - \pi_1 E[\varphi_{idle,int}]) \end{aligned} \quad (32)$$

where $b^*=p(1 - F_s(p))/\pi_1$. The mean service time $E[X(p)]_{WP}$ is greater than $E[X(p)]_{SRPT}$, due to the higher number of preemptions in SPTWP than SRPT, and the difference $E[X(p)]_{WP} - E[X(p)]_{SRPT}$ increases with p. The terms $E[T_{p_b}]$ and $(E[\varphi_{busy,p}]_{SRPT} - E[\varphi_{busy,p}]_{WP})$ are smaller positive values and $(E[\varphi_{busy,p}]_{SRPT} - \pi_1 E[\varphi_{idle,int}])$ is negative for a smaller p value. However, all three terms are larger positive values for a larger p. Therefore, difference in conditional mean response times $E[R|S=p]_{WP} - E[R|S=p]_{SRPT}$ varies from a small positive value to a small negative value with increasing p. Similar to (31), the mean response time difference $E[R]_{WP} - E[R]_{SRPT}$ is a positive value when the file lengths are heavy tail distributed. However, this positive value is smaller than that in (31). The difference between the conditional mean response times for SPTNP and SRPT is given by

$$\begin{aligned} E[R|S=p]_{NP} - E[R|S=p]_{SRPT} &= (\Omega_{p'} - \pi_0\Omega_{p_b})(\rho - b) \\ &+ b(E[\varphi_{busy,p}]_{NP} - E[\varphi_{busy,p}]_{SRPT}) \\ &+ E[X(p)]_{NP} - E[X(p)]_{SRPT}. \end{aligned} \quad (33)$$

Similar to the discussion on (31), the terms $(\Omega_{p'} - \pi_0\Omega_{p_b})$ and $(E[\varphi_{busy,p}]_{NP} - E[\varphi_{busy,p}]_{SRPT})$ contain busy periods initiated using one or more request arrival, whereas $E[X(p)]_{SRPT} - E[X(p)]_{NP}$ only contains the service time of a target request. Further, the terms $(\Omega_{p'}\text{-}\Omega_{p_b})$, $(\rho - b)$, and $E[\varphi_{busy,p}]_{NP} - E[\varphi_{busy,p}]_{SRPT}$ decrease, and $E[X(p)]_{SRPT} - E[X(p)]_{NP}$ increases with p. Therefore, the difference in conditional mean response times $E[R|S=p]_{NP} - E[R|S=p]_{SRPT}$ varies from a very high positive value to a very small value with increasing p, and the unconditional mean response time difference $E[R]_{NP} - E[R]_{SRPT}$ is a positive value when the file length is heavy tail distributed. Further, the difference $E[R]_{NP} - E[R]_{SRPT}$, increases with the interruption duration and the tail heaviness of the file length distribution.

VIII. N R

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. Without loss of generality, we consider $L_s = 1$. 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, $E[R]$, is evaluated by averaging the results of 20 simulation runs, each run having 18,000 service requests.

Fig. 9 shows the variation of $E[R]$ 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 $\rho = 0.6$ and $\alpha = 0.6$. It can be clearly seen that the simulation

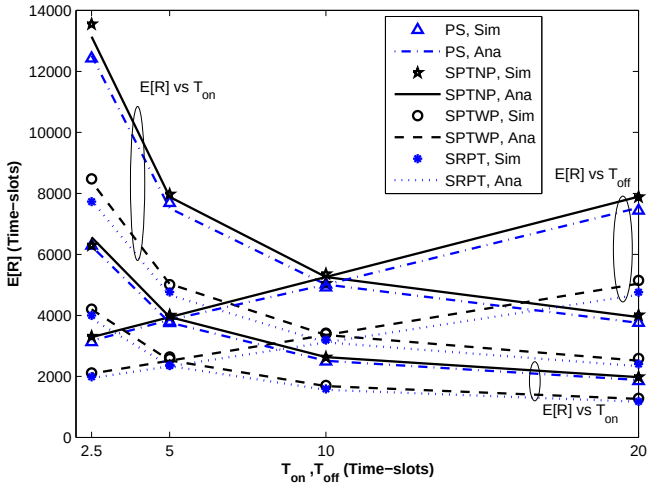

Fig. 9: The variation of mean response time with mean channel availability and interruption durations at $\rho = 0.6$.

results closely match with the numerical 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. 10 shows the variation of $E[R]$ with ρ 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 ρ, and the larger the ρ, the larger the rate of increment of $E[R]$. As the mean service rate remains constant, the mean arrival rate is proportional to the system load, and the higher the arrival rate, the higher the waiting time of the users at the waiting queue. Therefore, the waiting time increases with the system load, leading to longer response times. As seen in Fig. 9, the service disciplines with preemption outperforms that without preemption, and the PS outperforms the SPTNP at lightly loaded condition. For the PS service discipline, the heavier the 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 have 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.

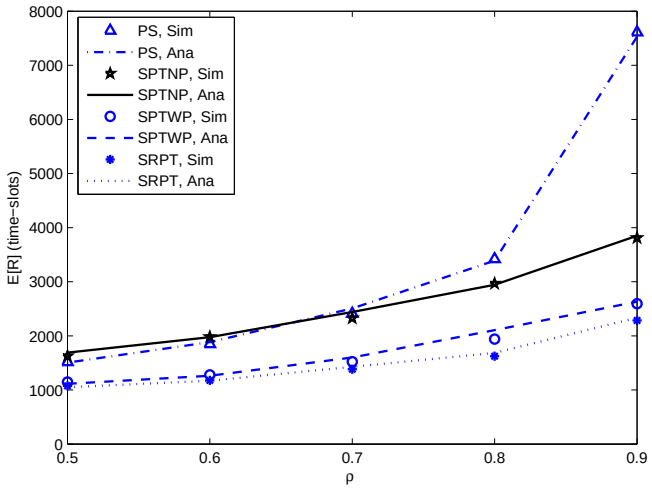

Fig. 10: The variation of mean response time with the traffic load.

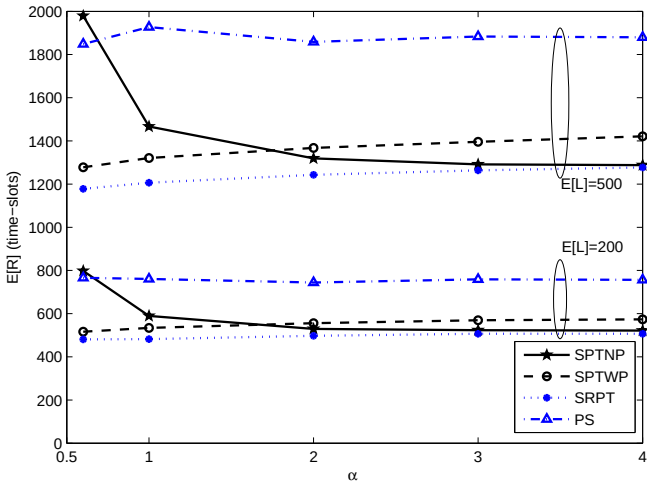

Fig. 12: The variation of mean response time with the tail heaviness of the STR distribution.

Therefore, the rate of increment of the mean response time with the system load is larger for the PS service discipline than that for the SPTNP service discipline. This rapid increment is indeed captured in (30).

Fig. 11 shows the $\boldsymbol{E[R]}$ variation with $\boldsymbol{T_{off}}$ for the SRPT service discipline obtained from numerical analysis and simulations for two different traffic load conditions and two different $\boldsymbol{\alpha}$ values with $\boldsymbol{T_{on}}$ **= 100** time-slots and $\boldsymbol{E[L]}$ **= 500**. The

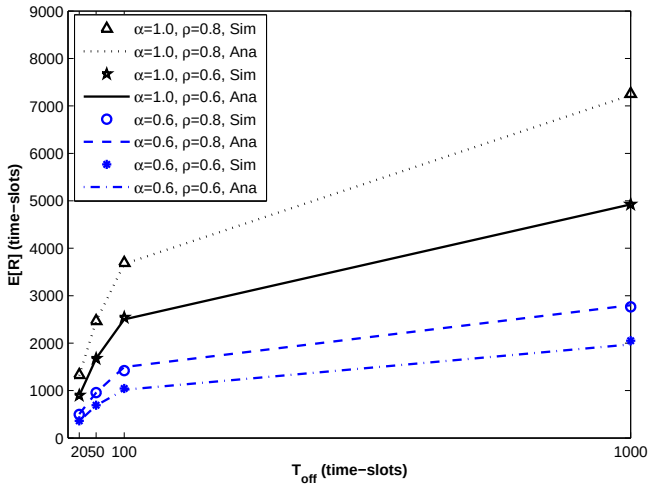

Fig. 11: The variation of mean response time with the mean interruption duration for SRPT service discipline

simulation results closely match with the numerical results. That is, the discrete-time analysis in section V is accurate for the networking scenario. Similar to what is observed in Fig. 10, the $\boldsymbol{E[R]}$ increases with $\boldsymbol{\rho}$. Further, the heavier the tail of the STR distribution, the longer the mean response time.

Fig. 12 shows the $\boldsymbol{E[R]}$ variation with the shape parameter $\boldsymbol{\alpha}$ (heaviness of the tail) with $\boldsymbol{\rho}$ **= 0.6**, $\boldsymbol{T_{on}}$ **= 20**, and $\boldsymbol{T_{off}}$ **= 10** time slots. The mean response time of the PS service discipline remains almost the same with the variation of $\boldsymbol{\alpha}$. As the PS gives an equal opportunity to all service requests, the $\boldsymbol{E[R]}$ depends on the mean of the STR, not its distribution [21]. 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 $\boldsymbol{\alpha}$, the larger the number of service requests with very short STR. Therefore, when the preemption is allowed, the heavier the tail of the STR distribution (a lower $\boldsymbol{\alpha}$), the lower the $\boldsymbol{E[R]}$. When $\boldsymbol{\alpha}$ is large, the STR concentrate around the mean STR, 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. Therefore, the larger the $\boldsymbol{\alpha}$, the closer the $\boldsymbol{E[R]}$ of the SPTNP to that of the SRPT. When the probability of having a longer STR is relatively larger and the preemption is based on the original STR, the requests with a longer STR gets preempted more often. Therefore, these unnecessary preemptions increase the response time in the case of SPTWP.

Fig. 13 shows the $\boldsymbol{E[R]}$ variation with $\boldsymbol{T_{off}}$ at $\boldsymbol{\pi_1 = T_{on}/(T_{on} + T_{off})}$=0.66 with $\boldsymbol{E[L]}$ **= 500** and $\boldsymbol{\rho}$ **= 0.6** for two different $\boldsymbol{\alpha}$ values. The $\boldsymbol{E[R]}$ increases with $\boldsymbol{T_{off}}$ even when the long term channel availability and the system load remain unchanged. When the interruption duration is exponentially distributed, the conditional mean response time for the SPTWP service discipline given in (19) can be simplified to $\boldsymbol{E[R|S = p]=E[D]\pi_0+E[\varphi_{busy,p}]b+\frac{p}{\pi_1(1-b)}}$. The $\boldsymbol{E[\varphi_{busy,p}]}$ increases and $\boldsymbol{b}$ remains constant with $\boldsymbol{E[D]}$ for constant $\boldsymbol{\pi_1}$ and $\boldsymbol{\rho}$. As a result, the higher the $\boldsymbol{E[D]}$ the higher the $\boldsymbol{E[R|S = p]}$. Similarly, we can show that the conditional mean response times of the SPTNP, SRPT, and PS service disciplines increase with the mean interruption duration when the long term channel availability and the system load remain unchanged. Similar to Fig. 12, the shorter the tail of the STR distribution (larger $\boldsymbol{\alpha}$), the lower the $\boldsymbol{E[R]}$ for the SPTNP and the larger the $\boldsymbol{E[R]}$ for the rest of the service disciplines. As in (31) and (33), the difference between the mean response times of the SPTNP and SPTWP and that between the SPTNP and SRPT increase

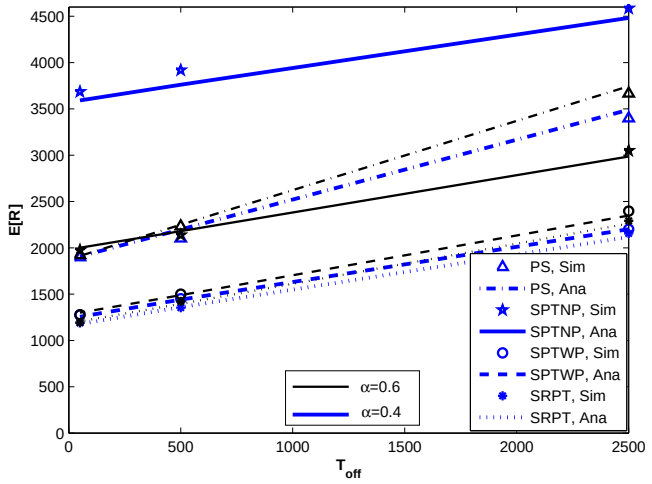

Fig. 13: The variation of mean response time with the mean channel unavailability duration at constant $\boldsymbol{\pi_1}$

with $\boldsymbol{T_{off}}$. However, there is no significant difference in the gap between the mean response times of the SPTWP and SRPT with the variation of $\boldsymbol{T_{off}}$.

Fig. 14 shows the CDFs of the file length $\boldsymbol{L}$ (or STR) and the response times obtained from simulations for the SPTNP and SPTWP service disciplines, respectively, with $\boldsymbol{E[L] = 500}$, $\boldsymbol{T_{on} = 10}$ and $\boldsymbol{T_{off} = 20}$ time-slots, and $\boldsymbol{\rho = 0.6}$. It is observed

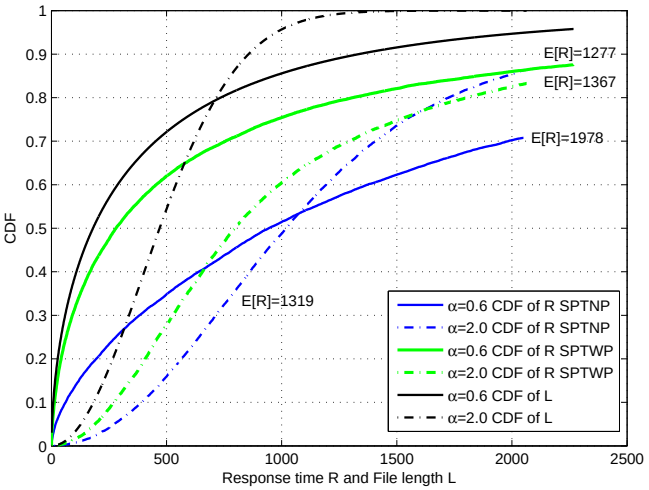

Fig. 14: The CDF of mean response times of both SPT service disciplines and the STR.

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 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 view point 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 larger. 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, 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 [15]-[17] 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.

IX. C

In this paper, we evaluate the mean response time of elastic data traffic under three service disciplines (namely, 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 load conditions, and it is shown that the SPTNP service discipline outperforms the PS service discipline in heavily loaded systems. Therefore, the SPT service discipline is a better choice over the PS service discipline as the system load increases. The mean response times of all four service disciplines are compared under the Weibull distribution with different 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. Further, the mean duration of the transmission interruptions (channel non-availability) has an impact on the mean response time even when the long term channel availability and the system load remain unchanged. This response time analysis can be used for call admission control to ensure service satisfaction.

R

[1] S. Gunawardena and W. Zhuang, "Service response time of elastic data traffic in cognitive radio networks with SPT service discipline," in *Proc. IEEE Globecom'12*, 2012 (to appear).

[2] FCC, "Spectrum Policy Task Force Report," ET docket no. 02-155, Nov. 2002.

[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 Pres. Communications.*, vol. 6, no. 4, pp. 13-18, Aug. 1999.

[5] A. Ghasemi and E. S. Sousa, "Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 32-39, April 2008.

[6] A. Guha and V. Ganapathy, "Power allocation schemes for cognitive radios," in *proc. of IEEE COMSWARE'08*, pp. 51-56, Jan. 2008.

[7] P. Pawelczak, S. Pollin, H. W. So, A. Motamedi, A. Bahai, R. V. Prasad, and R. Hekmat, "State of the art in opportunistic spectrum access medium access control design," in *Proc. of IEEE CrownCom'08*, pp. 1-6, May 2008.

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

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

[10] 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.

[11] D. Hu and S. Mao, "Streaming scalable videos over multi-hop cognitive radio networks," in *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 3501 - 3511, Nov 2010.

[12] H. Li, "Impact of primary user interruptions on data traffic in cognitive radio networks Phantom jam on highway," in *Proc. of IEEE Globecom 2011*, pp. 1-5, Dec. 2011.

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

[14] 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.

[15] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/WLAN integrated network," *IEEE Trans. Wireless Commun.*, vol. 8, issue. 2, part 1, pp. 725-735, Feb 2009.

[16] 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.

[17] W. Song and W. Zhuang, "Resource allocation for conversational, streaming, and interactive services in cellular/WLAN interworking," in *Proc. IEEE Globecom07*, pp. 4785-4789, Nov. 2007.

[18] 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.

[19] K. M. Rezaul and A. Pakstas, "Web traffic analysis based on EDF statistics, in *Proc. 7th Annual PostGraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PGNet)*, June 2006.

[20] 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.

[21] F. Delcoigne, A. Proutière, and G.Régnié, "Modeling and integration of streaming and data traffic," *Perform. Eval.*, vol. 55, no. 3-4, pp.185-209, Feb. 2004.

[22] 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.

[23] 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.

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

Subodha Gunawardena (S'08) received the B.Sc. degree in Electronics and Telecommunications Engineering from the University of Moratuwa Sri Lanka in 2004 and M.Eng in Telecommunications Engineering from the Asian Institute of Technology Thailand in 2008. He has worked at the Canon research labs in Japan form 2005 to 2006 in collaboration with Metatechno Japan Inc.. He received the Best Conference Paper Award in 2011 awarded by the IEEE Multimedia Communications Technical Committee. Subodha is currently working toward his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include capacity and resource allocation in Cognitive Radio Networks, cooperative communication, and robotics.

Weihua Zhuang (M'93-SM'01-F'08) received the B.Sc. and M.Sc. degrees from Dalian Maritime University, China, and the Ph.D. degree from the University of New Brunswick, Canada, all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where she is a Professor and a Tier I Canada Research Chair in wireless communication networks. Her current research focuses on resource allocation and QoS provisioning in wireless networks. Dr. Zhuang is a co-recipient of the Best Paper Awards from the IEEE VTC Fall-2010, IEEE WCNC 2007 and 2010, IEEE ICC 2007, and the International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine) 2007 and 2008. She received the Outstanding Performance Award 4 times since 2005 from the University of Waterloo for outstanding achievements in teaching, research, and service, and the Premiers Research Excellence Award in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. Dr. Zhuang is the Editor-in-Chief of the IEEE Transactions on Vehicular Technology, and the TPC Symposia Chair of the IEEE Globecom 2011. She is a Fellow of the Canadian Academy of Engineering (CAE) and the IEEE, and an elected BoG member of the IEEE Vehicular Technology Society.