Delay Performance Modeling and Analysis in Clustered Cognitive Radio Networks

Nadia Adem and Bechir Hamdaoui
School of Electrical Engineering and Computer Science
Oregon State University, Corvallis, Oregon 97331
Email:ademn,hamdaoui@eecs.oregonstate.edu

***Abstract*—Cognitive radio networks (CRNs) emerge as a promising solution for overcoming the shortage and inefficient use of bandwidth resources by allowing secondary users (SUs) to access the primary users' (PUs) channels so long as they do not interfere with them. The random availability of the PU channels makes the delay analysis of the SU, which accesses the channels opportunistically, plays a crucial role as a quality of service measure. In this paper, we model and characterize the total average delay the SUs experience in a CRN. The cognitive radio system is modeled as a discrete-time queueing system. The availability of the N independent and identical PU channels is modeled as a two states Markov chain. Our contributions in this paper is that we provide a solid performance evaluation that gives a closed-formula for the two delay components experienced by the SUs, namely the waiting delay and the service delay. We derive the waiting delay using the residual time concept. We characterize the service time distribution by considering the buffered-slotted-ALOHA systems. We also provide numerical results to show the effects of the analysis on the CRN design.**

I. Introduction

The increasing demand and usage of wireless technologies and applications are causing a shortage in the spectrum resource supply. However, the allocated spectrum is considered to be not fully utilized and far from reaching its capacity [1]. The dynamic spectrum access provided through cognitive radios (CRs) is considered as a promising solution to overcome this shortage of the spectrum and exploit its inefficient usage. In addition to the primary users (PUs), which have the priority to access a number of communication channels, the secondary users (SUs), which implement the CRs, can access to these channels but opportunistically, in that they can do so as long as they do not interfere with the PUs. This opportunistic access has great potentials for improving the spectrum utilization by giving an opportunity to the SUs to access the licensed bands in an economical way. However, the lack of access priority could cause drastic performance degradation.

A. Motivation

Depending on the PUs' usages of their channels, a channel could be available for SUs or not randomly, leading to intermittent SU transmissions, which affects their delay performance. This makes it of interest to analyze the SU average packet delay and study the implication of the PUs' activities on such a delay.

The delay analysis has its consequences on the CRN design. For a given PUs' activity level, the SUs might not be able to meet certain performance criteria. The required transparency of the SU activities could result in excessive packet drops or queue instability in the buffered CRN. Hence, modifications need to be introduced to the CRN settings to make sure that it meets the required performance criteria. This is explained in details in section IV. The CRN parameters that could be affected include SUs' data rates, their numbers, their number of interfaces, their packet lengths, the number of channels they can access to, etc.

In this paper we consider a clustered CRN, where a number of nodes along with a cluster head form a cluster. The cluster head is able to access to a number of PU channels opportunistically. This model can very well apply to a cognitive radio sensor network, where the sensor nodes send their data to a sink, which can be the cluster head in our model, that accesses to PU channels opportunistically. Sensor applications usually generate data in small rates and hence there is no need for acquiring a licensed band, and having opportunistic spectrum access can be enough to achieve a desired quality of service (QoS).

B. Summary of Contributions

Although focus has already been given to studying and analyzing the performance of various wireless systems (e.g., WiMAX [2], [3], ALOHA [4]), little has been given to studying delay performance of CRNs [5], [6]. The complexity and difficulty of analyzing delay performance in cognitive radio networks and the broad aspects of such performance analysis seem to be the reasons for making researchers shy away from such studies. The SUs need to adapt their operating conditions to the PUs which have channel access priority and possess different transmission characteristics. Different delay components come to the picture as a result of that. A SU experiences a delay while identifying and exploiting spectrum access opportunities. To the best of our knowledge, there are no comprehensive delay models that account for most of these components in the literature. Hence, more thorough investigations need to be done in this area. Most of the related work does not provide closed formulas for the average delay, nor does it explicitly characterize SUs' service time distribution, which is important when it comes to assessing delay performance. In this paper, we provide a solid delay performance analysis that gives a closed-formula for two delay components experienced by the SUs. Our work derives analytically both the mean waiting delay and mean service delay, which capture the delay SUs experience as a result of exploiting spectrum access opportunities in the presence of PUs' activities. We also characterize and derive the service time distribution. Our work provides insights into the understanding of the cognitive radio networks, and serves as the basis for analyzing more delay components analytically.

In this paper, we analyze the delay performance of a clustered CRN modeled as a discrete-time queueing system. Our model as well our methodology for analyzing the CRN performance are different from those considered in the related papers, which usually model their system as a continuous-time system. We consider the average residual service time, which we find using a graphical argument, to come up with the formula of the average waiting delay. The concept of the mean residual service time has been considered for evaluating the performance of some continuous-time systems. However, to the best of our knowledge, it has not been considered for discrete-time systems performance analysis. Inspired by the buffered slotted-ALOHA (S-ALOHA) system, we determine the service time distribution and hence find the mean service delay. We provide some numerical results through which we study the effects of considering different network parameters, including some parameters that reflect the dynamic availability of the resources, on the delay performance.

C. Related Work

Below we discuss the previous works that are most related to our work. Using fluid queue approximation approach, the authors in [5] studied the steady-state delay of the SUs that are contending to acquire the PUs' available channels. The authors characterized the moments of the SUs queues by representing their queue dynamics as Poisson driven stochastic differential equations. They considered the average lengths of the queues as a way to study the delay performance. In [10], the authors studied a cluster-based cognitive radio sensor network that supports both real-time traffic and best-effort traffic. They considered different resource allocation policies. The authors considered the service time to be random and not following a standard distribution, and hence they ended up analyzing the delay through approximating the average length of the SU queue size. In [6], the authors analyzed the stationary queue tail distribution of a SU using a large deviation approach. They assumed the arrival process at the SU to be constant, and modeled the PUs' activity as a Markov chain. In case there are two PU channels, they provided a closed-form expressions for the tail distribution of the queue length, and in more general cases they provided an upper and lower bounds for the distribution. [8] presents a queuing analytic framework to study queuing delay and buffer statistics of SUs' packets by modeling PUs' activity as a two state Markov chain and SUs' channel quality variations as a finite state Markov chain. The authors in [9] proposed dynamic spectrum access schemes for SUs with two priority classes. The schemes perform spectrum handoff for the SUs' traffic to protect PUs' transmission. The authors in this paper analyzed the mean handoff delay along with some other parameters to measure the schemes' performance.

Besides these works that are related to studying CRN delays, [11] introduced and presented the concept of residual service time, a concept that has been used to derive the Pollaczek-Khinchin formula, which expresses the average waiting time in M/G/1 systems. [11] also used the residual service time concept in a number of subsequent developments; it has, for e.g., been used to analyze M/G/1 systems with priorities.

The authors in [4] presented an approximation approach to analyze S-ALOHA systems with finite user population having either finite or infinite user buffer capacity. Using a state flow graph, they found the service time distribution of a user in the system. They proved that when considering deferring first transmission as the transmission protocol, a protocol in which all packets are transmitted with a given probability in each slot, the service time follows a geometric distribution. We, in this work, consider some of the results provided in [4] to characterize the service time in our studied network.

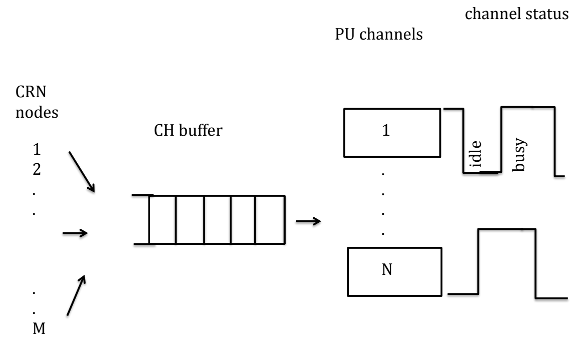

Fig. 1. Clustered cognitive radio network

II. Network Model

The cognitive radio network is considered to have access to N PU channels which can be available (idle) or unavailable (busy). The idle and busy intervals are defined to be the times during which a given channel is continuously available or unavailable respectively. These intervals are assumed to be independent of one another. Each channel's idle and busy intervals, denoted by T_{idle} and T_{busy} respectively, are assumed to be exponentially distributed with parameters u and v, respectively.

In our model we are considering a clustered CRN, where a number of nodes along with a cluster head (CH), which equips the CR, form a cluster. Fig. 1 illustrates the architecture of the network under consideration.

The CR system works as follows:

- Through a local channel the users in the cluster send their data to the CH which has an infinite queue capacity.
- The CH traffic is modeled as a Bernoulli arrival process. The arrivals are independent from each other. One packet, at the most, can arrive at any given slot with a probability λ, which can be thought of as the arrival rate per slot ($0 < \lambda < 1$).
- The CH sends the data in first-in first-serve basis over an available channel. If more than one channel is idle, it picks one randomly with an equally likely probability. It switches, in a neglected amount of time, in the same manner from one to another whenever the last assigned channel becomes unavailable. The delay overhead of discovering the transmission opportunities can be avoided through dedicating an interface for channel sensing.
- The CR system is considered to be time-slotted.
- The service (transmission) time of any packet starts and ends at the slot boundaries. It equals an integral (random) multiple of the slot duration. The service times are assumed to be independent and identically distributed with an unspecified general distribution. The service process is assumed to be independent from the arrival process.

The CR system can be modeled as a Geo/G/1 system. Geo refers to the Bernoulli arrival process, G indicates that the service times are generally distributed, and the number one here refers to the fact that the data that queues up at the CH is served over only one channel at any given time.

It is worth mentioning that we are not assuming any specific arrival models. The analysis to be made in this paper can be applied to the late as well as early arrival models. The late arrival model assumes that the packets arrive late during the slot, while the early model assumes the packets arrive early during the slot.

III. Delay Modeling and Characterization

In this section we are interested in finding the total average time a packet spends in the clustered CRN. We are considering the two delay components, waiting and service.

The waiting delay is the time a packet spends in the queue until it gets served. If a packet arrives to the system while there is a packet under service, the remaining of this service time is counted in its waiting delay. In addition, if a packet arrives while the queue is not empty, then the waiting time will also include the service time of all the packets ahead of it in the queue. In this section we derive the expression for the mean waiting delay and show how it is related to the mean service delay.

The service delay is the time a packet spends while it is being served. It is in fact the time spent in transmitting the packet. If the CRN has access priority, it takes only one slot to serve a packet. However, since the CH access the channel opportunistically, it takes integral (random) multiple of the slot duration to transmit the packet. In this section, we characterize the service time distribution and find the corresponding average delay.

A. Waiting Delay

We now derive the average waiting delay for the Geo/G/1 system, which is what we ended up modeling our system as, using the service residual time concept.

The concept of the mean residual service time has been considered for evaluating the performance of some continuous-time systems, M/G/1 for example. However, to the best of our knowledge, it has not been considered for evaluating the performance of the discrete-time systems. The analysis made in continuous-time systems can not be applied to discrete-time systems. In this section, we determine the mean residual service time in the discrete systems and use it to analyze the delay performance.

1) The Residual Service Time Concept: An arrival to the system may experience some delay resulting from the residual service time of one of the packets arrived ahead of it. Let R_i denotes the residual service time seen by the i^{th} arrival. If the j^{th} packet is being served when the i^{th} packet arrives, then R_i corresponds to the remaining time until packet j completes its service. When packet i arrives while the system is empty, then R_i equals zero.

Fig. 2 illustrates the concept of the residual time. In this figure we draw the number of arrivals and departures over time and show the residual service time corresponding to each arrival. X_i denotes the service time of the i^{th} arrival. t_i represents the time at which the i^{th} arrival arrives, and $t_i^{'}$ represents the time at which the i^{th} arrival leaves the system.

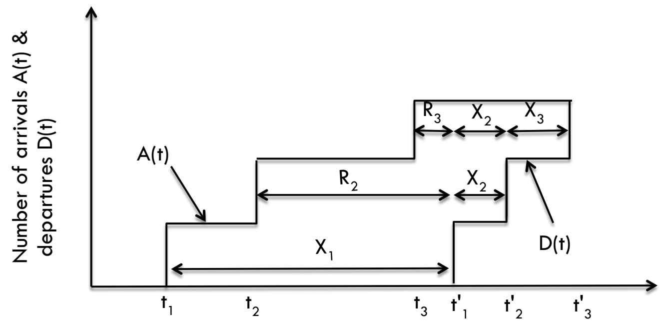

Fig. 2. The concept of the residual service time

2) The Residual Service Time in the Discrete Systems: In the discrete systems, the residual time can take a non zero value only at the instants at which an arrival can occur. The arrival instants depend on the arrival models. In our analysis we are considering those time instants to be the slot boundary. The analysis applies to the early arrival model, where the arrivals arrive just post to the start of the slot, as well as to the late arrival model, where the arrivals arrive just prior to the end of the slot.

Also, since service times are integral multiples of the slot duration, the remaining of a service time as seen by an arrival can only equal integral multiples of the slot duration.

Let the service time of the i^{th} arrival start at the beginning of the k^{th} slot (just post the boundary of slot $k-1$ and k), and let it be X_i slots. Let's also refer to the residual time at the end of a slot k (the boundary of slot k and $k+1$) by r_k, where r_k is measured in slots. The residual times corresponding to the arrivals arrive during the service of the i^{th} arrival are denoted by $r_k, r_{k+1}, \ldots, r_{k+X_i-1}$. Their values are $X_i - 1, X_i - 2, \ldots, 1, 0$ slots, respectively. At the end of the first slot of the service time, the residual time is $X_i - 1$ slots, and its value decreases by one slot at the end of the next slot, and keeps doing so until the service time completes. It equals zero at the end of the last slot of the service time. See Fig. 3 for illustration.

It is worth mentioning here that for an outside observer, any service corresponds to an arrival arrives right prior to the start of the service. The service that starts at the k^{th} slot corresponds to an arrival arrives at the boundary between the slot $k-1$ and k. Since one packet at most can arrive at any given time slot, no other arrival can arrive at this particular arrival instant. That is said, the residual times at the beginning of any slot at which a service starts is zero. See Fig. 3 for illustration.

3) The Mean Residual Service Time: According to [11], the mean residual time as seen by an arrival, $\overline{R}$, is equal to the mean residual time seen by an outside observer at a random time. This is valid for any arrivals satisfying the PASTA (Poisson Arrivals See Time Averages) property, which is the case for the M/G/1 systems. The question arises here is what about the Geo/G/1 systems? Since the BASTA (Bernoulli Arrivals See Time Averages) property in the Geo/G/1 systems is analogous to the PASTA property in continuous-time systems, we can also define $\overline{R}$ in the discrete-time system to be the mean residual time seen by an outside observer at a random time.

Considering this definition, we use a graphical argument to find $\overline{R}$. In Fig. 3 we plot a sample path of a server status over

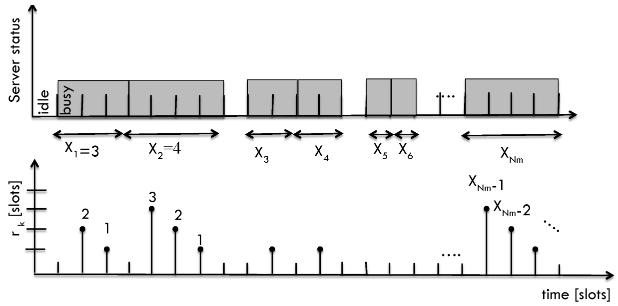

Fig. 3. A sample path of a server status and the corresponding residual service time.

time. During any time slot, the server could be either busy or idle. When the server starts serving a packet i, it stays busy for X_i slots. We also plot the corresponding sample path of the residual service time from which its time average can be derived.

Consider the time interval [0,τ], where τ is the time instant corresponding to the end of the m^{th} slot. We are assuming that up to the m^{th} slot, N_m packets have already been served. The average residual time (measured in slots) in this interval is given by $E_m = \frac{1}{m}\sum_{k=1}^{m} r_k$.

Since we know the values of r_k during the service time of each packet, the sum of the r_k over the m slots can be determined by summing the r_k corresponding to the service times. The average residual time can then be rewritten as

$$E_m = \frac{1}{2}\frac{N_m}{m}(\frac{\sum_{i=1}^{N_m} {X_i}^2}{N_m} - \frac{\sum_{i=1}^{N_m} X_i}{N_m})$$

Taking the limit as $m \to \infty$, assuming it exists, we obtain

$$\lim_{m\to\infty} E_m = \frac{1}{2}\lim_{m\to\infty}\frac{N_m}{m}\lim_{m\to\infty}(\frac{\sum_{i=1}^{N_m} {X_i}^2}{N_m} - \frac{\sum_{i=1}^{N_m} X_i}{N_m})$$

The left-hand side limit is the time average of the residual time. The limits on the right-hand side are the departure rate (which equals the arrival rate), the service time second and first moments respectively. Assuming that the time averages can be replaced by the ensemble averages, the average residual time can then be expressed as

$$\overline{R} = \frac{1}{2}\lambda(\overline{X^2} - \overline{X}) \tag{1}$$

where $\overline{X}$ and $\overline{X^2}$ denote the service time first and second moment respectively. $\overline{X}$ by definition equals $1/\mu$, where μ is the service rate (in packets per slot).

The per-packet average waiting time, $\overline{W}$, can be expressed in terms of the average residual time as $\overline{W} = \overline{R}/(1-\rho)$, where $\rho = \lambda/\mu$ is the utilization factor [11]. Replacing $\overline{R}$ with its expression presented in Equation (1) yields

$$\overline{W} = \frac{1}{2}\frac{\lambda(\overline{X^2} - \overline{X})}{(1-\rho)} \tag{2}$$

We derived the expression of the average waiting delay for the discrete-time system, the Geo/G/1 system. The formula is given in terms of the service time statistics and the rate of arrivals. More elaboration on the behavior of this delay component is given in section IV.

B. Service Delay

As we have mentioned earlier, the service time is the time spent in transmitting a packet. The service time distribution is a prerequisite for analyzing the CRN delay performance. The average waiting delay formula we derived earlier for example involves the first as well as the second moments of the service time. Delay analysis can still be made if the service time distribution is not realized. However, the analysis can be very complicated, and usually closed-form formulas cannot be derived. Most of the related work consider the service rate to be random and does not follow a standard distribution.

Depending on the model of the system under consideration, the service time can turn out to be not following any standard distribution. Let's assume that a channel needs to be available for an S amount of time continuously so that a packet can be served. Let's also assume that the CH starts to serve packets whenever there is a channel available, i.e., the service does not necessary start at a slot boundary. It is possible that a CH starts to serve a packet and then before it completes its packet transmission, the channel gets occupied by a PU. This could happen many times in a random manner. This causes the service time to be random and not follow any standard distribution.

Our system is time-slotted with Bernoulli arrivals. Inspired by the S-ALOHA system presented in [4] and [11], we make the following arguments. Let S denotes the slot duration. As it has been mentioned earlier, the service of our CR system packets starts and ends at the slot boundaries. At any given slot, if there is a packet ready for the service, the CH will try to send it. Given that the channels' availability is time-variant, with some probability P_s, we call it probability of success, the trial succeeds. If the trial succeeds, the time spent in serving that particular packet is one slot. Otherwise, another trial needs to be made. That said, the probability that the CH takes l slots to serve a packet is $(1-P_s)^{l-1}P_s$. That is to say that the service time distribution is geometric with parameter P_s.

Assuming noiseless channels, or alternatively assuming the CRN operates at a signal-to-noise ratio that does not make the noise a concern, a packet service is successful if there is at least one channel available during at least a slot duration, i.e., during at least S amount of time. Remember that T_{idle} is the time during which a channel is continuously available. The probability of success can be written as $P_s = Pr\{no\ outage\}Pr\{T_{idle} > S\}$.

Considering that T_{idle} and T_{busy} are assumed to be exponentially distributed with parameters u and v, the probability that a CH succeeds in sending a packet in one slot can be written as

$$P_s = (1 - \frac{1}{(1+v/u)^N})e^{-uS} \tag{3}$$

Given that the per-packet total average delay $\overline{T}$ equals $\overline{X}+\overline{W}$, it follows from Equation (2) that

$$\overline{T} = \overline{X} + \frac{1}{2}\frac{\lambda(\overline{X^2} - \overline{X})}{(1-\lambda\overline{X})} = \frac{1-\lambda}{P_s - \lambda} \tag{4}$$

This formula gives the total average delay per packet in the CR system. The behavior of this delay is to be studied in the following section.

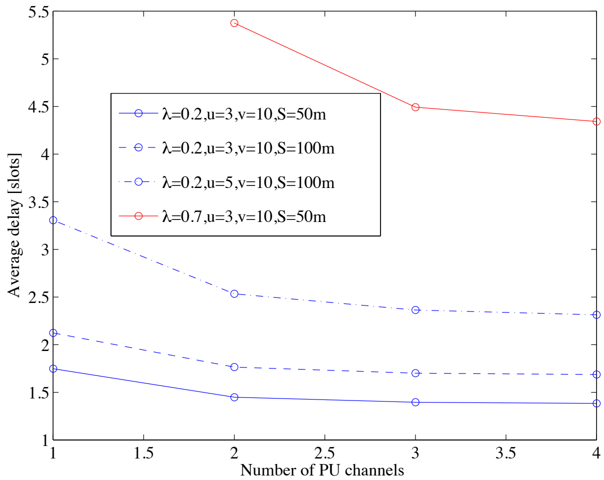

Fig. 4. Average delay vs. number of PU channels

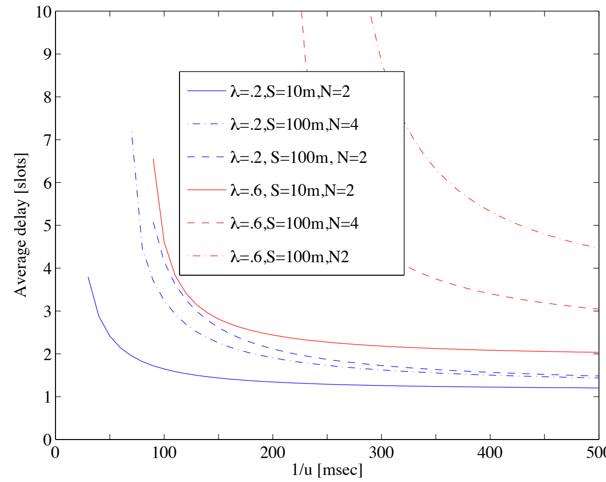

Fig. 5. Average delay vs average of channel availability time ($\overline{T}_{idle}$)

IV. Performance Evaluation and Analysis

In this section, we numerically evaluate CRN delay performance and analyze the impact of several network parameters on this performance. We also explain the effect of this performance analysis on the CRN design.

For a given arrival rate, slot duration, average time on which a channel stays available ($\overline{T}_{idle} = 1/u$) and unavailable ($\overline{T}_{busy} = 1/v$), we illustrate in Fig. 4 how the delay performance behaves when varying the number of PU channels. Note that the average total delay decreases as the number of channels increases, which is intuitive. However, how fast the delay decreases and what value the delay converges to depend on the slot duration, the SUs' packet arrival rate, and the PUs' usage of the channels reflected via the rate at which a channel leaves the idle and the busy states, u and v, respectively.

As the rate at which a channel leaves the idle state, u, increases, the outage probability increases. Also, the probability that a channel stays idle for a period of length (at least) S, which is the least amount of time required to serve a packet, decreases. Hence, the average service rate decreases. In addition, since no more than one channel can be accessed at any given time, having more channels might not improve the performance considerably.

It is worth mentioning here that for a given network setting whenever the results, in any of the figures presented in this section, are not shown, that should be read as violating the stability condition. The network parameters such as the CR system packets arrival rate, the slot duration, number of primary user channels, as well as the PUs statistics can all affect the service rate. If the service rate turns out to be less than the arrival rate, the delay increases drastically and the CR system becomes unstable.

Fig. 5 illustrates how the average delay behaves when the average time on which a channel stays available changes. As expected, as $\overline{T}_{idle}$ increases, the CR system performance improves. However, how significant the improvement is depends on other network settings.

It can be noticed here that for a small arrival rate (here 0.2), going from $N = 2$ to $N = 4$ while fixing the other parameters does not have the same impact on the performance as it does in the case the arrival rate is relatively large. Similarly, the impact of increasing the slot duration becomes more severe as the value of λ increases. Increasing the slot duration decreases the probability that a channel can stay idle over the slot duration. Hence, the average service time increases, and so consequently does the average waiting time, especially for large values of λ.

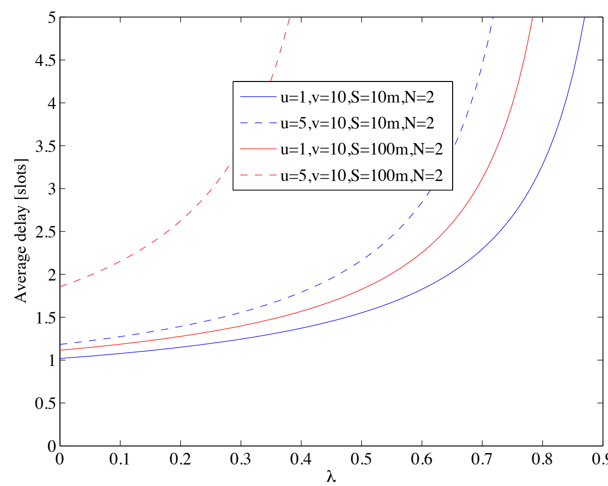

Fig. 6. Average delay vs the arrival rate

Note here the range of $\overline{T}_{idle}$ through which the system is unstable varies depending on the other system settings. To be able to maintain the system stability, the value(s) of λ, $\overline{T}_{idle}$, S, or N (or combination of all) need to be modified. This in fact demonstrates the importance of considering the CRN delay performance when designing CRN networks.

Fig. 6 shows the impact of varying the arrival rate on the total delay. When $\lambda = 0$, the waiting delay is zero and the average total delay equals the service delay. The service delay does not change as λ changes since service process is independent from the arrival process.

The figure shows that the average service delay increases as the rate at which a channel leaves the idle state, u, and/or the slot duration, S, increase. As u and/or S increase, the probability that a channel stays idle over a slot duration gets smaller, and hence the average time required to serve a SU packet increases. In addition, as the slot duration increases along with increasing u and the arrival rate, the probability that an arrival occurs while serving another packet increases, thus increasing the average time a packet ends up waiting in the queue. This explains why the performance degrades significantly as λ increases specially as S and u increases.

V. Conclusions and Future Work

This paper studies the delay performance of a clustered cognitive radio network which is modeled as a discrete-time system. The primary users' activity is modeled as two states Markov chain. We derived both the mean waiting and service delay. The results provided emphasize the importance of the delay analysis and show its consequences on the network design. We showed that depending on the PUs' statistics, a modification in the CR system settings might be required to maintain its stability. Extending the current work to a multi-cluster network where the SUs, that are equipped with multiple interfaces, contend to acquire the channels is underway. It is also of interest to perform the same analysis considering more empirical channel model. Our work provides insights not only on the understanding of the cognitive radio networks, but also on the queueing analysis of the discrete-time systems.

Acknowledgment

This work was supported in part by the US National Science Foundation under NSF award CNS-1162296.

References

[1] F. S. P. T. Force, " Report of the spectrum efficiency working group, " Nov. 2002.

[2] J. Ben Othman, L. Mokdad, and S. Ghazal, "Performance analysis of WiMAX networks AC," *Wireless Personal Commun.* , vol. 74, no. 1, pp. 133-146, Jan. 2014.

[3] L. Mokdad and J. Ben Othman, "Admission control mechanism and performance analysis based on stochastic automata networks formalism," *Journal of P. and D. Computing*, vol. 71, no. 4, pp. 594-602, April 2011.

[4] T. Wan and A. U. Sheikh, "Performance and stability analysis of buffered slotted aloha protocols using tagged user approach," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 582-593, Mar. 2000.

[5] S. Wang, J. Zhang, and L. Tong,"Delay analysis for cognitive radio networks with random access: a fluid queue view," *in Proc. IEEE INFOCOM*, San Diego, CA, 2010.

[6] A. Laourine, S. Chen, and L. Tong, "Queueing analysis in multichannel cognitive spectrum access: a large deviation approach," *in Proc. IEEE INFOCOM*, San Diego, CA, 2010.

[7] O. Akan, O. Karli, and O. Ergul, " Cognitive radio sensor networks," *IEEE Network*, vol. 23, no. 4, pp. 34-40, July 2009.

[8] M. M. Rashid, J. Hossain, E. Hossain, and V. K. Bhargava, "Opportunistic spectrum access in cognitive radio networks: A queueing analytic model and admission controller design," *in Proc. IEEE Globecom*, Washington, DC, 2007, pp. 4647-4652.

[9] Vamsi Krishna Tumuluru,Ping Wang, Dusit Niyato, and Wei Song, , "Performance Analysis of Cognitive Radio Spectrum Access With Prioritized Traffic," *IEEE Trans. Veh. Tech.*, vol. 61, no. 4, pp. 1895-1906, May 2012.

[10] S. Feng and D. Zhao, "Supporting real-time CBR traffic in a cognitive radio sensor network," *in Proc. IEEE WCNC*, Sydney, 2010.

[11] D. Bertsekas and R. Gallager, *Data Networks,* 2nd E. Prentice Hall, 1992.

[12] Hideaki Takagi, *Queuing Analysis: Discrete-time systems,* Vol. 3. North-Holland, 1993 .

[13] U. Narayan Bhat, "Lecture notes: Introduction to Queuing Theory," Southern Methodist University, 2005.