Analytical Models for Capacity Estimation of IEEE 802.11 WLANs using DCF for Internet Applications[†]

George Kuriakose, Sri Harsha, Anurag Kumar and Vinod Sharma

Abstract— We provide analytical models for capacity evaluation of an infrastructure IEEE 802.11 based network carrying TCP controlled file downloads or full-duplex packet telephone calls. In each case the analytical models utilize the attempt probabilities from a well known fixed-point based saturation analysis. For TCP controlled file downloads, following [3], we model the number of wireless stations (STAs) with ACKs as a Markov renewal process embedded at packet success instants. In our work, analysis of the evolution between the embedded instants is done by using saturation analysis to provide state dependent attempt probabilities. We show that in spite of its simplicity, our model works well, by comparing various simulated quantities, such as collision probability, with values predicted from our model.

Next we consider N constant bit rate VoIP calls terminating at N STAs. We model the number of STAs that have an up-link voice packet as a Markov renewal process embedded at so called channel slot boundaries. Analysis of the evolution over a channel slot is done using saturation analysis as before. We find that again the AP is the bottleneck, and the system can support (in the sense of a bound on the probability of delay exceeding a given value) a number of calls less than that at which the arrival rate into the AP exceeds the average service rate applied to the AP.

Finally, we extend the analytical model for VoIP calls to determine the call capacity of an 802.11b WLAN in a situation where VoIP calls originate from two different types of coders. We consider N_1 calls originating from Type 1 codecs and N_2 calls originating from Type 2 codecs. For G711 and G729 voice coders, we show that the analytical model again provides accurate results in comparison with simulations.

Index Terms— TCP throughput on WLAN, VoIP on WLAN, capacity of WLAN, performance modeling of DCF.

I. INTRODUCTION

Wireless local area networks (WLANs) based on the IEEE 802.11 standard [22] are being increasingly deployed in enterprises, academic campuses and homes, and at such places they are expected to become the access networks of choice for accessing the Internet. It therefore becomes important to study their ability to carry common Internet applications such as TCP controlled file downloading, or packet voice telephony.

In this paper we are concerned with a network in which N IEEE 802.11 stations (STAs) access a high speed local area network via an access point (AP). We consider three different traffic scenarios, and develop analytical models that yield capacity estimates for carrying such traffic over the WLAN. Thus our analysis will yield answers to the questions: "How many TCP controlled file transfers can be done in parallel so that the transfer throughput per STA is at least (say) 25 kilobytes per second?" or "How many packet telephone calls can be set up to different STAs such that the probability of packet delay over the WLAN exceeds (say) 20 ms is small?" Our goal is to provide answers to these questions using a stochastic model for the WLAN and the traffic flow through it.

In the first scenario, we consider N mobile STAs each having a TCP connection via the AP to some server. Such a TCP data transfer only situation will exist in a typical office LAN environment. Each of the connections is transmitting a long file from the server(s) to the users via the AP. We develop an analytical model for this system and obtain the system throughput.

In the second scenario, each STA is engaged in a VoIP call with some wired client via the AP. Such a situation would arise in a wireless IP PBX where the sole function is to provide telephony services in an office. In this case we will consider the quality of service (QoS) parameter to be the fraction of packets transmitted within a certain time for each connection. We form an analytical model of this system and compute the number of voice calls that can be supported.

In the third scenario, we consider the case where the VoIP calls originate from different type of codecs. The analytical model for VoIP calls (in the second scenario) is extended to analyze this case. We obtain the admissible region for the number of VoIP calls of different types, possible in the WLAN, while meeting the delay QoS constraint.

In each of the above models we identify an embedded Markov chain which we study to obtain the parameters of interest. The MAC protocol (CSMA/CA) employed in 802.11 DCF is complicated and does not really lead to a Markov system. But we replace it with a system where each station transmits its packet (if it has one) in every slot with a probability that depends only on the number of stations contending for the channel at that time. We approximate these probabilities as those obtained from the saturation results in [2] and [15]. The intervals between the instants at which

Markov chain is embedded are random, but together these constitute a Markov renewal process. We will see that the resulting stochastic model provides a good approximation to the actual system.

Remark: It is known (see for e.g., [1]) that with the default IEEE 802.11 DCF, interactive packet telephony cannot be sustained in conjunction with data downloads. Hence in this paper we analyze the two traffic classes separately. In recent work [11], [12] we have extended our approach in this paper to IEEE 802.11e WLAN where we do model voice and TCP downloads together. ∎

Related Literature: The modeling of IEEE 802.11 DCF has been a research focus since the standard has been proposed. Chhaya and Gupta in [6] analyze the effect of packet capture and hidden terminals. Cali et al., in [5], provide a theoretical throughput analysis based on a p-persistent model of the MAC. In [2], Bianchi uses a Markov model to analyze the saturation throughput of a single cell IEEE 802.11 network, and shows that the model yields accurate results. A generalization and a fixed point formalization of the Bianchi analysis is done by Kumar et al. in [15]. All the above papers assume that stations operate in saturation, i.e., they always have a packet to transmit.

There are only a few attempts to model and analyze the 802.11 MAC protocol behavior when subjected to actual traffic loads, e.g., TCP or voice traffic. Duffy et al. [8] and Sudarev et al. [23] propose models in finite load conditions by approximating the packet arrival process at the wireless stations as a Poisson process. Tickoo and Sikdar [24] derive delay and queue length characteristics for a finite load ad-hoc 802.11 WLAN by modeling each queue with an M/G/1 model. Detti et al. [7] and Pilosof et al. [20] discuss throughput unfairness between TCP controlled transfers in 802.11 WLANs. Leith and Clifford [16] discuss how TCP unfairness can be removed using the QoS extensions in 802.11e. The papers do not directly address the problem of performance evaluation of actual TCP transfers or VoIP calls in a WLAN. Bruno et al. [3] consider the scenario of STAs performing TCP controlled bulk downloads via an AP. Our modeling assumptions are drawn from this work. We discuss the relationship between [3] and ours in subsection II-B. In their recent paper [4], Bruno et al. have considered the scenario where both upload and download TCP connections are present in the WLAN. When there is a certain number of contending nodes, the authors model the state dependent attempt probabilities using an iterative analysis presented in [5]. The proposed model does not consider the delayed ACK option, an important technique that improves the TCP throughput. Miorandi et al. [18] propose a model for performance analysis of TCP download connections in the WLAN, with the delayed ACK option. The model in [18], uses a Bernoulli distribution approximation for the number of contending nodes in the WLAN.

Analytical performance modeling of packet voice telephony to estimate the call capacity over 802.11 WLANs has been done by Garg and Kappes [9], Hwang and Cho [13] and Medapalli et al. [17]. These authors do not model the

evolution of the back-off process of the 802.11 MAC layer, but consider approximate constant values for back-off parameters like average back-off time ([9] and [13]) and collision probability ([17]).

Our Contribution: We model the MAC layer queue dynamics for typical Internet applications like TCP download transfers and voice traffic, while also considering the evolution of the binary exponential back-off process of the 802.11 MAC. We provide a simple approach of using the results of saturation analysis of Bianchi [2] and Kumar et al. [15] for performance evaluation of a WLAN with finite load. The delayed ACK option is considered for TCP download transfers. In each of the scenarios, we obtain the number of contending stations through a Markov chain and obtain the performance measures through Markov regenerative analysis. In order to ascertain the accuracy of the models, we derive additional parameters like collision probability, attempt rate, etc. and show that they compare well with the simulation results.

Outline of the paper: In Section II we discuss the modeling assumptions of TCP download transfers case. We build the model that results in a Markov regenerative framework and use it to derive the performance measures, namely the aggregate download throughput and collision probability. First we consider the undelayed ACK case and then cover the delayed ACK case as well. We then provide numerical and simulation results for showing the accuracy of the model. In Section III, with some key assumptions, we model the case of duplex CBR voice calls and derive the voice capacity and other related parameters for model validation. In Section IV we justify the approach of using attempt probabilities from saturation analysis of [2] and [15], by deriving the attempt rates from the proposed voice model and comparing them with those obtained from the simulations. In Section V, we extend the voice model to capture the scenario when calls originate from different type of codecs. We obtain the admission region of voice calls in this scenario, while meeting the QoS delay constraint. Lastly (in Section VI) we conclude by listing the modeling insights obtained in this analysis.

II. MODELING TCP CONTROLLED FILE DOWNLOADS

A. Modeling Assumptions

We consider a single cell 802.11 WLAN with N STAs associated with a single AP. All *nodes* (a term we use to refer to any wireless entity and hence could be STAs or AP) contend for the channel via the DCF mechanism. Each STA has a single TCP connection to download a large file from a local file server. Hence, the AP delivers TCP data packets towards the STAs, while the STAs return TCP ACKs. We further assume that when downloading a file, RTS/CTS is used by the AP to send the data packets, while basic access is used by the STAs to send the ACKs. We begin by assuming that when an STA receives data from the AP, it immediately generates an ACK (that is queued at its MAC). Later on we also consider a model for the case in which delayed ACKs are used.
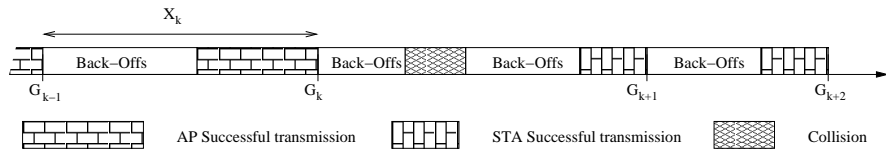
Fig. 1. An evolution of the back-offs and channel activity. G_k, $k \in 0, 1, 2, 3, \ldots$, are the instants where k^{th} successful transmission ends.

We assume that the AP and the STAs have buffers large enough so that TCP data packets or ACKs are not lost due to buffer overflows. We also assume that there are no bit errors, and packets in the channel are lost only due to collisions. Also, these collisions are recovered before TCP time-outs occur. As a result of these assumptions, for large file transfers, the TCP window will grow to its maximum value and stay there.

When there are several TCP connections (each to a different STA), since all nodes (including the AP) will contend for the channel, and no preference is given to the AP, most of the packets in the TCP window will get backlogged at the AP. The AP's buffer is served FIFO, and we can assume that the probability that a packet transmitted by the AP to a particular STA is $\frac{1}{N}$. Thus it is apparent that the larger the N, the lower is the probability that the AP sends to the same STA before receiving the ACK for the last packet sent. The number of ACKs in the STAs depends on the number of TCP data packets delivered by the AP. If there are several STAs with ACKs then the chance that AP succeeds in sending a packet is small. *Thus the system has a tendency to keep most of the packets in the AP with a few STAs having ACKs to send back.* We observe that the STA may or may not have an ACK packet. When the STA queue is non-empty, it contends for the channel. To develop the model (based on the above discussion) we assume that each STA can have a maximum of one TCP ACK packet queued up. This assumption implies two things. First, after an STA's successful transmission, the number of active STAs reduces by one. Second, each successful transmission from the AP activates a new STA. As N is increased, this assumption is close to what happens in reality.

Hence for large N, we can simply analyse the process of the number of active STAs. Before explaining the analysis we will review a similar approach from [3].

B. Discussion of Related Work [3]

The modeling assumptions mentioned above were first introduced in [3]. The authors consider the TCP transfers scenario and obtain the channel utilization achieved by the AP's transmissions. They derive the analysis for a p-persistent IEEE 802.11 protocol. The p-persistent IEEE 802.11 MAC differs from the standard protocol in the selection of the backoff interval. Instead of the binary exponential back-off used in the standard, the backoff interval is sampled from a geometric distribution with parameter p. In order to obtain the channel utilization they first obtain the mean virtual transmission time $(E[T_v])$ defined as the mean time between two AP successes. They provide a complicated

derivation of $E[T_v]_K$, the mean virtual transmission time conditioned on having K active STAs at the beginning of the virtual transmission time. Then they compute $E[T_v]$ as $\sum_k \pi(k)E[T_v]_k$, where $\pi(k)$ is the probability that there are k active STAs after an AP's successful transmission. The channel utilization is simply $\frac{T_{AP}}{E[T_v]}$ where T_{AP} is the time taken to transmit one AP packet. They obtain results only for the non-delayed ACKs case and report the delayed ACK case as a matter of further study. They provide simulation results as well to substantiate their analysis. Our approach here is similar but differs in the following ways: (i) We incorporate the IEEE 802.11 DCF backoff procedure by using the saturation analysis from [2] and [15]; in particular, as against the constant p in [3], the attempt probability in our model depends on the number of STAs having ACKs at that time. (ii) We validate this approach by calculating additional system measures (collision probability and distribution of number of non empty STAs), and compare the results against simulations. (iii) We also develop a VoIP capacity analysis (in Section III). (iv) Our analytical development is very simple.

C. The Mathematical Model and its Analysis

Let us consider Figure 1 which shows the back-offs and the channel activity. The instants G_k, $k \in 0, 1, 2, 3, \ldots$, are the instants where the kth successful transmission ends.

First consider N large, and let S_k be the number of active STAs at the instants G_k. Since the AP has TCP data packets to transmit all the time, it is sufficient to keep track of S_k, in order to model the channel contention. We also assume that whenever there are n active STAs then these STAs and the AP each attempt in a slot with probability β_{n+1}, where β_{n+1} is the attempt rate obtained via saturation analysis ([2] and [15]) when there are $n + 1$ saturated nodes.

Since the back-off parameters for both the AP and the STAs are the same, it is assumed that when there are n STAs active, the probability of the AP to win the contention is $1/(n+1)$ while the probability of one of the STAs to win the contention is $n/(n+1)$ [15]. As explained earlier in Section II-A, since the AP is carrying the traffic of all the N STAs, the number of contending STAs cannot become large. Hence the number of STAs that are active with a high probability is insensitive to N for large N. See also [4], [18]. Hence with the above observations and assumptions, S_k is modeled as a Markov chain, over all nonnegative integers. The transition probabilities of the Markov chain are shown in Figure 2. This approximation also helps us to obtain a simple closed form expression of the stationary probability distribution, π, which we will derive below. We will show via simulations that this simplification yields accurate results for large N (in fact, N
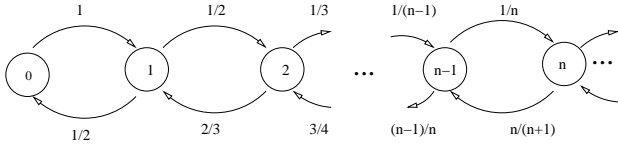
Fig. 2.   Transition probability diagram of the Markov chain S_k.

just needs to be greater than 4 for the infinite N model to suffice).

It is easy to see that *for $N = 1$*, the situation is different from that described for N large. Since nodes contend for access independent of their packet lengths, in steady state (for large file downloads) the TCP window will be equally split between the AP and the single STA. Both nodes are thus saturated and the AP throughput is the connection throughput. This observation was also made in [15].

The following subsections provide the analysis of the model for N large, followed by the analysis for $N = 1$. We will see from simulations how large N needs to be for the "large N" analysis to apply.

1) Aggregate Download Throughput: The throughput of the AP is the main performance metric for this system. Consider Figure 1. Let $X_k = G_k - G_{k-1}$. Under our assumptions $\{(S_k; G_k), k \geq 0\}$ forms a Markov renewal process. Let the number of successful attempts made by the AP in the kth cycle be denoted by H_k ($= 0$ or 1). We view H_k as a reward associated with the kth cycle. Let $H(t)$ denote the total number of AP successes in $(0, t)$. Then by Markov regenerative analysis (or a renewal reward theorem) [14] we obtain, with probability one,

$$\lim_{t \to \infty} \frac{H(t)}{t} = \frac{\sum_{n=0}^{\infty} \pi_n \left(\frac{1}{n+1}\right)}{\sum_{n=0}^{\infty} \pi_n \, \mathsf{E}_n X} =: \Theta_{AP-ftp}$$

where π_n is the stationary probability of having n contending STAs in a cycle, and $\mathsf{E}_n X$ is the average time until the end of the next success when the number of contending STAs at the end of a success is n. In the following we compute π_n and $\mathsf{E}_n X$. Θ_{AP-ftp} is the total throughput (in packets per second) obtained by all the TCP connections together. The ith TCP connection will get the throughput, θ_i (in packets per second) proportional to its maximum window size. The throughput of each connection, in bits per second, will be proportional to the product of the maximum window size and the packet length. We are assuming here that each connection has the same maximum window size and equal packet length and so each of the connection will obtain an equal share of the aggregate download throughput Θ_{AP-ftp}.

2) The Stationary Distribution, π_n: The balance equations for the Markov chain are (see Figure 2)

$$\pi_n = \frac{1/n}{n/(n+1)}\pi_{n-1} = \frac{n+1}{n^2}\pi_{n-1}, \; n \in \{0, 1, 2, \ldots\}.$$

Using the above equations and the fact that $\sum_n \pi_n = 1$, one can obtain the stationary probability π_n as

$$\pi_n = \frac{n+1}{(n!)(2e)} \; n \in \{0, 1, 2, \ldots\}.$$

Parameter	Symbol	Value
PHY data rate	C_d	11 Mbps
Control rate	C_c	2 Mbps
PLCP preamble time	T_P	$144\mu s$
PHY Header time	T_{PHY}	$48\mu s$
MAC Header Size	L_{MAC}	34 bytes
RTS Packet Size	L_{RTS}	20 bytes
CTS Packet Size	L_{CTS}	14 bytes
MAC ACK Header Size	L_{ACK}	14 bytes
IP Header	L_{IPH}	20 bytes
TCP Header	L_{TCPH}	20 bytes
TCP ACK packet size	$L_{TCP-ACK}$	20 bytes
TCP data payload size	L_{TCP}	1500 bytes
VoIP packet size: G 711	L_{voice}, L_{voice1}	200 bytes
VoIP packet size: G 729	L_{voice2}	60 bytes
System slot time	δ	$20\mu s$
DIFS Time	T_{DIFS}	$50\mu s$
SIFS Time	T_{SIFS}	$10\mu s$
EIFS Time	T_{EIFS}	$364\mu s$
Min. Contention Window	CW_{min}	31
Max. Contention Window	CW_{max}	1023

TABLE I

VARIOUS PARAMETERS USED IN ANALYSIS AND SIMULATION.

Since we have a positive invariant probability vector, the Markov chain is positive recurrent. We notice that $\sum_{n=0}^{\infty} \pi_n \left(\frac{1}{n+1}\right) = \frac{1}{2}$, as expected, i.e., in the undelayed ACK case, the AP must transmit half the successful transmissions.

3) Mean Cycle Length, $\mathsf{E}_n X$: Let the attempt probability of a node obtained from fixed point analysis be β_{n+1} [15] when there are $n+1$ contenders. Then the following equation holds (this takes into account the fact that the following events take different times: the time wasted in collision, when a slot goes idle, when TCP packet is successfully transmitted by AP and when a TCP ACK packet is successfully transmitted by an STA)

$$\mathsf{E}_n X = P_{idle}(\delta + \mathsf{E}_n X) + P_{sAP}T_{sAP} + \\ P_{sSTA}T_{sSTA} + P_c(T_c + \mathsf{E}_n X)$$

which yields:

$$\mathsf{E}_n X = \frac{P_{idle}\delta + P_{sAP}T_{sAP} + P_{sSTA}T_{sSTA} + P_c T_c}{1 - P_{idle} - P_c}.$$

The above equation uses the following notations. These use the IEEE 802.11b parameters provided in Table I.

δ is the system slot time. A system slot is the time unit employed for discrete-time backoff countdown in IEEE 802.11 MAC standard.

P_{idle} is the probability of a slot being idle $= (1 - \beta_{n+1})^{n+1}$.

P_{sAP} is the probability that the AP wins the contention $= \beta_{n+1}(1 - \beta_{n+1})^n$.

P_{sSTA} is the probability that an STA wins the contention $= n\beta_{n+1}(1 - \beta_{n+1})^n$.

P_c is the probability that there is a collision $= 1 - P_{idle} - P_{sAP} - P_{sSTA}$.

T_{sAP} is the time required for transmitting one TCP packet (from AP) including MAC and PHY overhead $= T_P + T_{PHY} + \frac{L_{RTS}}{C_c} + T_{SIFS} + T_P + T_{PHY} + \frac{L_{CTS}}{C_c} + T_{SIFS} + T_P + T_{PHY} +$

$\frac{L_{MAC}+L_{IPH}+L_{TCPH}+L_{TCP}}{C_d} + T_{SIFS} + T_P + T_{PHY} + \frac{L_{ACK}}{C_c} + T_{DIFS}$.

T_{sSTA} is the time required for transmitting one TCP ACK packet including MAC and PHY overhead $= T_P + T_{PHY} + \frac{L_{MAC}+L_{IPH}+L_{TCP-ACK}}{C_d} + T_{SIFS} + T_P + T_{PHY} + \frac{L_{ACK}}{C_c} + T_{DIFS}$.

T_c is the time spent in collision $= T_P + T_{PHY} + \frac{L_{MAC}+L_{IPH}+L_{TCP-ACK}}{C_d} + T_{EIFS}$.

In the above calculations, we have assumed that TCP data packets are larger than the RTS threshold and hence the AP uses the RTS/CTS access mechanism, and since TCP ACKs are small, the STAs use the basic access mechanism. Also, we note that whenever there is a collision, either between an RTS packet from the AP and one or more TCP ACK packets from the STAs, or between two or more TCP ACK packets from STAs, the channel time wasted is that due to the TCP ACK packet, since, the RTS packet is smaller than a TCP ACK packet. This gives us only one collision time, given by T_c.

4) Collision Probability: To further check the accuracy of the model, we give an expression for the conditional collision probability defined as the probability that an attempt of the AP fails due to a collision. Again let us consider the Markov renewal process $\{(S_k, G_k), k \geq 0\}$ mentioned earlier. Let us define, for the kth cycle, $\{A_k, k \geq 0\}$ as the number of attempts made by the AP and $\{C_k, k \geq 0\}$ as the number of collisions of these attempts by the AP. Let $C(t)$ and $A(t)$ denote the total number of collisions and attempts, respectively, in $(0, t)$. Then,

$$\lim_{t\to\infty} \frac{C(t)}{A(t)} \overset{a.s.}{=} \frac{\sum_{n=0}^{\infty} \pi_n \, \mathsf{E}_n C}{\sum_{n=0}^{\infty} \pi_n \, \mathsf{E}_n A} =: \gamma_{AP-ftp}$$

$\mathsf{E}_n A$ and $\mathsf{E}_n C$ can be calculated as follows. We use the assumption that after every collision, success or idle slot, the nodes attempt with a probability which depends only upon the total number of nodes in active contention and is independent of the previous state of the system. Then,

$\mathsf{E}_n A$ = Prob { None of the nodes attempt}$(\mathsf{E}_n A)$ +
Prob { AP attempts and succeeds}(1) +
Prob { AP attempts and collides}$(1 + \mathsf{E}_n A)$ +
Prob { Some STA attempts and succeeds}(0) +
Prob { AP does not attempt, STAs collide}$(\mathsf{E}_n A)$

= $(1 - \beta_{n+1})^{n+1}(\mathsf{E}_n A) +$
$\beta_{n+1}(1 - \beta_{n+1})^n (1) +$
$\beta_{n+1}(1 - (1 - \beta_{n+1})^n)(1 + \mathsf{E}_n A) +$
$(1 - \beta_{n+1})n\beta_{n+1}(1 - \beta_{n+1})^{n-1}(0) +$
$(1 - \beta_{n+1})\left(1 - (1 - \beta_{n+1})^n - n\beta_{n+1}(1 - \beta_{n+1})^{n-1}\right)(\mathsf{E}_n A)$

and

$\mathsf{E}_n C$ = Prob { None of the nodes attempt}$(\mathsf{E}_n C)$ +
Prob { AP attempts and succeeds}(0) +
Prob { AP attempts and collides}$(1 + \mathsf{E}_n C)$ +
Prob { Some STA attempts and succeeds}(0) +
Prob { AP does not attempt, STAs collide}$(\mathsf{E}_n C)$

= $\beta_{n+1}(1 - \beta_{n+1})^n (0) +$
$\beta_{n+1}(1 - (1 - \beta_{n+1})^n)(1 + \mathsf{E}_n C) +$
$(1 - \beta_{n+1})^{n+1}(\mathsf{E}_n C) +$
$(1 - \beta_{n+1})n\beta_{n+1}(1 - \beta_{n+1})^{n-1}(0) +$
$(1 - \beta_{n+1})\left(1 - (1 - \beta_{n+1})^n - n\beta_{n+1}(1 - \beta_{n+1})^{n-1}\right)(\mathsf{E}_n C)$

5) Single TCP Session ($N = 1$): As explained earlier in this section, when only one STA is engaged in a download file transfer, we have just 2 nodes and the assumption of asymmetry in the queues of AP and STA does not hold. The two nodes eventually reach a steady state wherein both are saturated [15]. Then the throughput is simply obtained as

$$\Theta_{AP-ftp} = \lim_{t\to\infty} \frac{H(t)}{t} = \frac{1/2}{\mathsf{E}_1 X} \qquad (1)$$

since each success is a data packet or a TCP ACK packet, with equal probability.

D. Analysis for TCP with Delayed ACKs

The analysis can be applied to a system with TCP connections with delayed ACKs as well with a small modification in the model. Let us assume that instead of every TCP packet, every alternate packet is acknowledged. (This analysis can be easily extended to the case in which every m^{th} packet is acknowledged.)

In our model without delayed ACKs, when the AP succeeds it generates an ACK at an STA due to which the state of the system increases by one. In the delayed ACK case an AP success generates an immediate ACK at an STA only half of the time. Thus if the number of STAs with ACK packets is n and the AP succeeds then S_k goes to the state $n + 1$ with probability $1/2(n + 1)$ and S_k will stay at the same state with probability $1/2(n + 1)$. The rest of the transitions remain unchanged. The new transition diagram is shown in Figure 3.
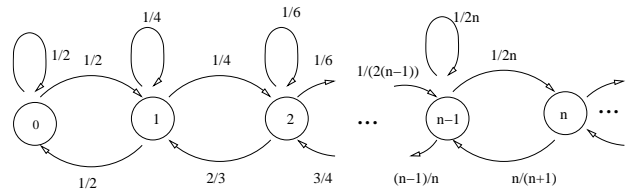

Fig. 3. Transition probability diagram for the infinite Markov process S_k with delayed ACKs.

The balance equations for this Markov chain are

$$\pi_n = \frac{1/2n}{n/(n+1)}\pi_{n-1} = \frac{n+1}{2\,n^2}\pi_{n-1}, \; n \in \{0, 1, 2, \ldots\}$$

from which we obtain

$$\pi_n = \frac{n+1}{2^n\,n!}\,\pi_0, \; n \in \{0, 1, 2, \ldots\}.$$

Using the above equations and the fact that $\sum_n \pi_n = 1$, one can obtain the stationary probability π_n. All other calculations for throughput and collision probabilities remain unchanged.

Since we are reducing the number of packets generated at the STAs, the AP's share of transmitted packets increases. Thus the throughput of this system will be more than that of the system with non-delayed ACKs.

Remark: The analysis above assumes strictly that every other packet is acknowledged. If N is large, due to the increase in queue length at the AP, the time between successful packet transmissions for the same STA might exceed the delayed ACK timeout, and as a result a delayed ACK will be generated at the STA. Thus, for large N the throughput is expected to decrease, which our analysis will not capture. Thus, this analysis gives an upper bound on the throughput (see Figure 7). ∎

E. Simulation Results and Comparisons

In this section we compare the results obtained by our analysis with those obtained by simulations (done in Network Simulator ns-2 [19]). The various parameters used were taken from the 802.11b standard (given in Table I). The TCP packet size is 1500B and the RTS threshold is 300B. The error bars in simulation curves denote 95% confidence intervals. The analysis yields two throughput numbers, one for $N = \infty$ (for each PHY rate), and one for $N = 1$ for each PHY rate. The values are shown in Table II. Figures 4, 5, 6 and 7 show the distribution $\{\pi_n\}$, the aggregate throughputs (without delayed ACKs), the collision probabilities and the aggregate throughputs with delayed ACKs, respectively. The throughput is in Mbps and is obtained as $8 \times L_{TCP} \times \Theta_{AP-ftp}$. The following are some of our observations:

PHY Data Rate, C_d (Mbps)	Θ_{AP-ftp} (Mbps)		
	Undelayed ACK		Delayed ACK
	$N = 1$	$N = \infty$	$N = \infty$
2	1.41	1.41	1.51
5.5	2.80	2.78	3.04
11	3.88	3.86	4.30

TABLE II

Θ_{AP-ftp} FOR VARIOUS PHY DATA RATES OBTAINED VIA ANALYSIS.

1) In Figure 4 we compare π_n obtained via simulations for $N = 5, 10$ and 30, and via analysis (using $N = \infty$). As predicted by the analysis, π_n is independent of N for such values of N. Note that the shape of the distribution and its support is captured quite well by the analysis. We see that for $N \geq 5$, the distribution of the number of active STAs is insensitive to N and hence an analysis for $N = \infty$ can be expected to work well. Interestingly, it works well for $N < 5$ as well (Figures 5, 6).

2) The plot of aggregate download throughput with different values of N for PHY bit rates of 11, 5.5 and 2 Mbps are shown in Figure 5. The values obtained via the analysis are shown in Table II. In Figure 5 we show the single throughput number obtained from the $N = \infty$ analysis, plotted for $N \geq 5$ (in view of the observation in Point 1 just above). The value obtained for $N = 1$ is shown with an ×. The analysis is remarkably accurate, and we find that the throughput for $N = 2, 3, 4$ is the same as that for the other values of N.

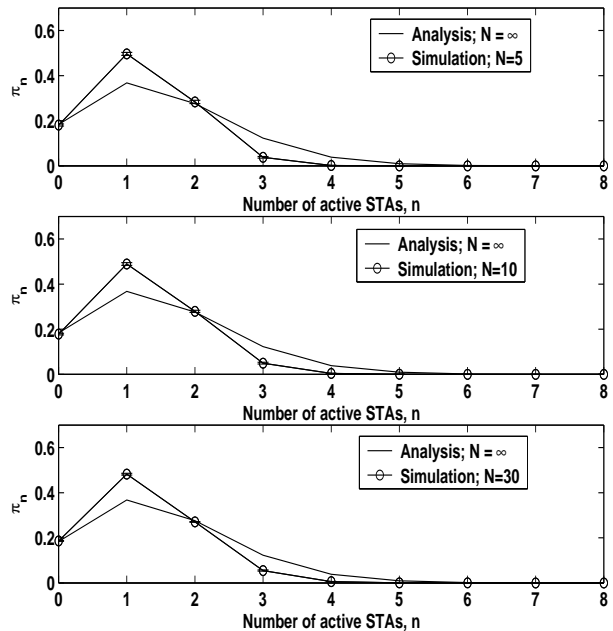3) We compare the collision probabilities in Figure 6 which gives a further check on the accuracy of our

Fig. 4. Simulation results for stationary distribution, π_n of number of active STAs n, for $N = 5, 10$ and 30. Also shown alongside is the analytical result using $N = \infty$. The TCP sessions use undelayed ACKs; the PHY data rate is 11 Mbps.
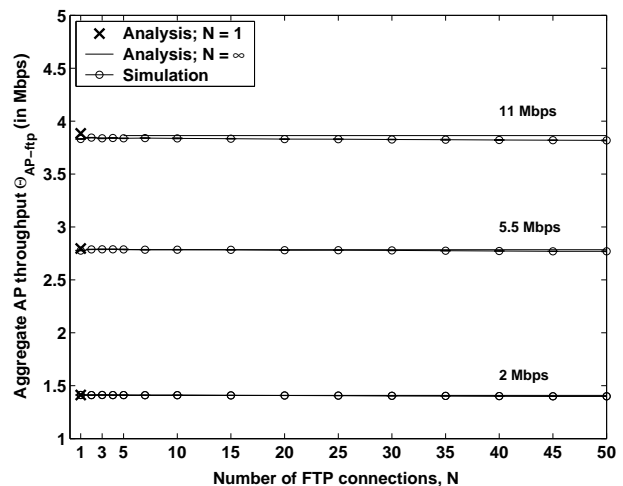

Fig. 5. Analysis and simulation results for the downlink FTP aggregate download throughput vs number of FTP connections (one per STA) for various PHY rates. The TCP sessions use undelayed ACKs.

model. The equation for γ_{AP-ftp} shows that it is *independent of the PHY rate*. This is verified by the simulation plots. This insensitivity with the PHY rate is as expected, since the evolution of the contention process does not depend on PHY rates.

4) In Figure 7 we compare the aggregate downstream throughput with different values of N for PHY bit rate of 11, 5.5 and 2 Mbps for the *delayed* ACK case. As commented on before, as N increases there is a drop in the throughput which our model does not capture.

As a general rule of thumb, we can conclude that the FTP download capacities (using TCP with delayed ACKs) for an
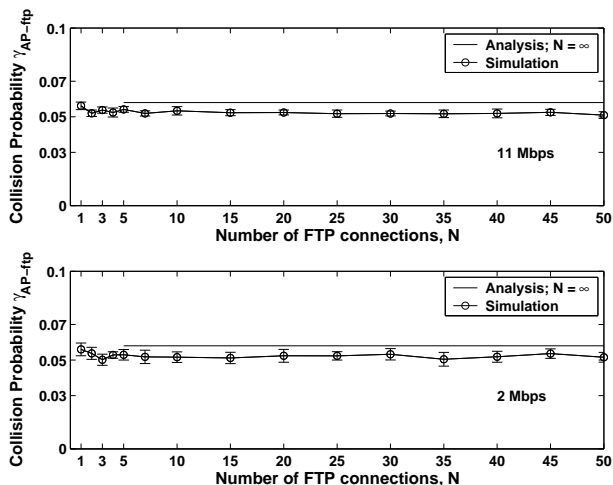
Fig. 6. Analysis and simulation results for the collision probability vs number of FTP connections (one per STA), for 11Mbps and 2 Mbps PHY rates. The TCP sessions use undelayed ACKs.
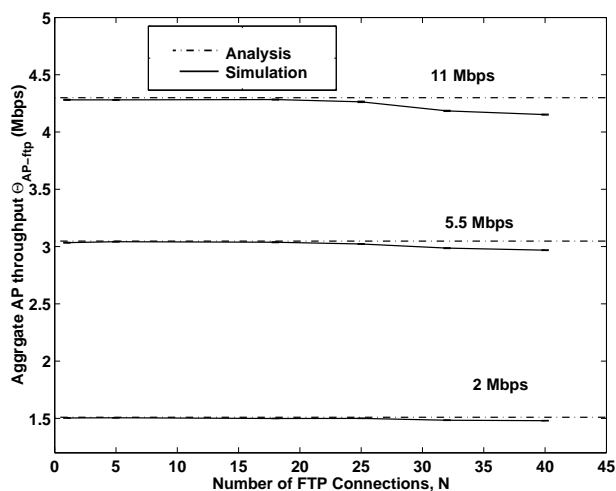

Fig. 7. Analysis and simulation results for the downlink FTP aggregate download throughput vs number of FTP connections (one per STA) for various PHY rates. Delayed ACK option is enabled.

infrastructure IEEE 802.11 WLAN with all STAs associated at 11 Mbps, 5.5 Mbps or 2 Mbps are roughly 4.3 Mbps, 3 Mbps or 1.5 Mbps. These aggregate rates are shared equally (for equal maximum window sizes and packet lengths for each connection) among the STAs performing the downloads, if there is one FTP session per STA.

Remark: An extension to the case where different STAs are associated at different rates can be done as follows. The Markov chains $\{S_k\}$ (see Figures 2 and 3) remain unchanged. The success and collision probabilities will not depend on the rates. Suppose a fraction a_i of the STAs are associated with rate $r_i(\in 11, 5.5, 2\ Mbps)$. Then an STA success can be ascribed to an STA associated with rate r_i w.p. a_i. An AP success can similarly be ascribed to an STA associated with rate r_i w.p. a_i. ∎

F. Remarks on our Modeling Assumptions

Under certain modeling assumptions, we have provided an accurate analytical model for TCP controlled downlink file transfers in an IEEE 802.11 WLAN. In this section we discuss some of these assumptions.

1) Finite AP Buffer: One of our modeling simplifications is that the buffer at the AP is infinite and hence there are no packet drops due to buffer overflow. A consequence of the infinite buffer assumption is that the TCP window grows to its maximum value, the AP buffer never empties out and hence the AP always contends. It may be recalled that we have assumed this in our analytical model. In practice, however, the buffer at the AP is finite. Recall that we are modeling the situation in which the file transfers are taking place from a server on the high speed LAN to which the AP is connected. Hence the round trip propagation delay is very small. Then, it can be easily seen that, if the number of transfers is not very small (5 or more), a TCP window of 1 suffices to keep the AP from emptying out. In fact, our analytical model continues to hold in all aspects. The concern remains that if the maximum TCP window is large (denoted usually by W_{max}, a typical value being 20 packets) then buffer losses and the consequent timeouts can result in starvation of the AP buffer. We therefore conducted ns2 simulations with an AP buffer of 300 KBytes, or 200 packets. With 5 TCP connections there were no packet losses, as expected. With 50 connections, we observed packet losses, some of which resulted in timeouts and others in triple-duplicate ACK based recovery. The packet loss probability observed was 10%. However, the simulations showed that the stationary probability distribution of the number of contending STAs, the aggregate download throughput, and the collision probability were still the same as in Figures 4, 5 and 6, respectively. This is explained as follows. One packet from each transfer suffices to keep the AP from starving, as observed earlier. The TCP window never drops below 1. Also, even when timeouts occur in some connections, there are enough active connections to keep the AP from starving. In fact, we have observed that even with a very small AP buffer, e.g., just 10 packets, the aggregate performance measures are the same as with an infinite AP buffer, but there is a large short term throughput variability across connections. With a 300 KBytes AP buffer this variability becomes insignificant.

2) Bidirectional Transfers: In our model, we have only considered TCP controlled donwlink file transfers. If we retain the infinite AP buffer model, then it can be seen that the same model works for uplink file transfers. This is easily observed when delayed ACKs are not used, i.e., for each received data packet the TCP receiver sends back an ACK. First consider only uplink file transfers. Now, in our model, we only need to replace downlink data packets with ACKs, and uplink ACKs with data packets. Exactly the same analysis works. This is basically a consequence of the fact that in the IEEE 802.11 DCF the attempt behaviour of the nodes does not depend on the length or type of the packet being attempted. Now, suppose that some STAs are performing downlink transfers, whereas others are performing uplink transfers (with each

STA being involved in only one transfer). Again the same model holds, and we have the same Markov model for the number of STAs with a packet to send (ACK or data). We just need to observe that, if all the TCP windows are equal, then the head-of-the-line packet at the AP is a data packet with probability equal to the fraction of STAs that are performing downloads. Even different window sizes can be handled by this approach.

Although, numerical results from our model match the finite buffer simulations, the detailed analytical modeling of TCP transfers over a WLAN with a finite AP buffer remains a challenging problem. With simultaneous transfers in both directions, and finite AP buffers, unfairness between downlink and uplink transfers has been reported in empirical and simulation studies [10]. It is also of interest to obtain a performance model when transfers take place from a remote server across a wide area Internet. Modeling of such situations is a topic of our ongoing research.

III. A Model for Packet Voice Telephony

There are N STAs, all associated with a single AP. Each STA has a single full duplex VoIP call to a wired client on the wired LAN via the AP. The calls are not synchronized with each other. Each call results in two RTP/UDP streams, one from a remote client to a wireless STA, and another in the reverse direction. We begin by considering the case where each call uses the ITU G711 codec. Packets are generated every 20 ms. Including the IP, UDP and RTP headers, the size of the packet emitted in each call in each direction is 200 bytes every 20 ms. We also present results for the G729 codec which compresses 20 ms speech to 20 bytes; this results in a packet of size 60 bytes including the IP, UDP and RTP headers. We do not model voice activity detection (and consequent packet suppression) since not all instances of packet voice can be expected to utilize this optimization.
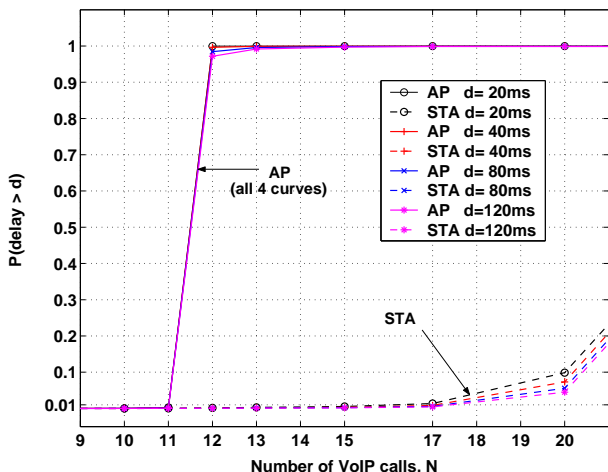

Fig. 8. Simulation results showing the probability of delay of packets at AP and STA being greater than d, $d \in \{20ms, 40ms, 80ms, 120ms\}$ vs. the number of calls (N). Packet size is 200 bytes (G711 codec); MAC protocol is 802.11b; PHY data rate is 11Mbps; control rate is 2Mbps

We set an objective that each arriving packet of a call should get served with a high probability before the next packet of the same call arrives, i.e., "with a high probability the packet delay should be less than 20 ms". To justify this delay objective, we present some useful simulation results, in Figure 8. The figure shows the probability that the voice packet delay, at the AP and at an STA, exceeds d, $d \in \{20ms, 40ms, 80ms, 120ms\}$ vs. the number of voice calls in the WLAN. The solid lines are for the AP while the dashed lines are for an STA. We make the following observations:

1) The AP packet delays shoot up earlier than that of the STA. This implies, as is to be expected, that AP is the capacity bottleneck.
2) All the AP delay curves (for different values of d), shoot up after 11 voice calls. These simulation results show that the IEEE 802.11 service is such that there is a sharp change from an uncongested regime to a congested one. Such an observation can also be made from the results reported in [21] and [24], where for an open-loop arrival model of a WLAN it is found that the delay is very small but sharply increases as the arrival rate approaches saturation.

Thus, though a more relaxed delay QoS may be acceptable, we make an important conclusion that even "an objective of $Prob(delay \geq 120ms)$ is small", yields no increase in the call capacity. For our model, the choice of delay bound of 20 ms is convenient as it permits us to assume that a device (AP or STA) will rarely have more than one packet of the same call if QoS has to be met.

A. A Stochastic Model

In this subsection we develop a Markov renewal model for the number of active senders when there are N calls in the system, each call terminating on a different STA.

We make some assumptions that permit us to formulate as a discrete time Markov chain the number of STAs that have packets to transmit, i.e., that contend for the channel. Packets arrive at the STAs every 20 ms. As discussed earlier (just before Subsection III-A), as a QoS requirement we demand that the probability that a packet is transmitted successfully within 20 ms is close to 1. Since the packets will experience delays in the rest of the network also, this is a reasonable target to achieve. Then, if the target is met, whenever a new packet arrives at an STA, it will find the queue empty. Thus the following two assumptions will be acceptable in the region where we want to operate: (1) the buffer of every STA has a queue occupancy of at most one packet, and (2) new packets arriving to the STAs arrive only at empty queues. The latter assumption implies that if there are k STAs with voice packets then a new voice packet arrival comes to a $(k+1)^{th}$ STA. Since the AP handles packets from N streams we expect that it is the bottleneck (as also demonstrated by the simulation results in Figure 8) and we assume that it will contend at all times. This is a realistic assumption near the system capacity. Note however that the AP can have up to N packets of *different* calls.

$$\text{Prob}\left(B_{j+1} = b | (Y_j = y; L_{j+1} = l)\right) = \binom{N-y}{b} (p_l)^b (1-p_l)^{N-y-b} \tag{2}$$
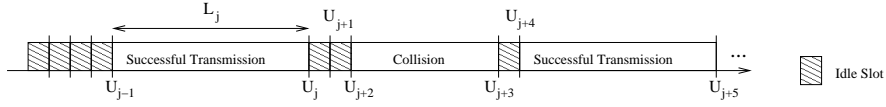

Fig. 9. An evolution of the back-offs and channel activity. U_j, $j \in 0, 1, 2, 3, \ldots$ are the instants where j^{th} channel slots ends.

As mentioned earlier, packets arrive every 20 ms in every stream. We use this model in our simulations. However, since our analytical approach is via Markov chains, we assume that the probability that a voice call generates a packet in an interval of length l slots is $p_l = 1 - (1-\lambda)^l$, where λ is obtained as follows. Each system slot is of $20\mu s$ duration. Thus in 1000 system slots there is one arrival. Therefore, for the 802.11b PHY we take $\lambda = 0.001$. This simplification turns out to yield a good approximation.

Figure 9 shows the evolution of the back-offs and channel activity in the network. U_j, $j \in 0, 1, 2, 3, \ldots$, are the random instants when either an idle slot, or a successful transmission, or a collision ends. Let us define the time between two such successive instants as a *channel slot*. The interval $[U_{j-1}, U_j)$ is called the jth channel slot. Let Y_j be the number of non-empty STAs at the instant U_j. Let B_j be the number of new VoIP packet arrivals at all the STAs, $V_j^{(AP)}$ the number of departures from AP and $V_j^{(STA)}$ the number of departures from STAs in the jth channel slot. We note that new arrivals in $[U_j, U_{j+1})$ cannot contend until U_{j+1}. Hence,

$$Y_{j+1} = Y_j - V_{j+1}^{(STA)} + B_{j+1}.$$

with the condition $V_j^{(STA)} + V_j^{(AP)} \in \{0, 1\}, \forall j$. By our earlier assumptions in this subsection, it is sufficient to keep track of Y_j in order to model the channel contention.

The distribution of the number of arrivals in one channel slot, B_j, can be obtained as follows. The probability with which a packet arrives at a node in a slot is λ. Then the probability that at least one packet arrives in l slots will be $1 - (1-\lambda)^l = p_l$. Since we assume that packets arrive at only empty STAs, B_j will have the distribution as given by Equation (2) (displayed above).

We also assume that whenever there are k nonempty STAs then these STAs and the AP each attempt in a slot with probability β_{k+1}, where β_{k+1} is the attempt rate obtained via fixed point analysis [15] when there are $k+1$ saturated nodes. We can then express the conditional distributions $V_{j+1}^{(STA)}$ and $V_{j+1}^{(AP)}$, as follows. $V_{j+1}^{(STA)}$ is 1 if an STA wins the contention for the channel and 0 otherwise. Thus

$$V_{j+1}^{(STA)} = \begin{cases} 1 & \text{w.p. } (Y_j)\beta_{Y_j+1}(1-\beta_{Y_j+1})^{Y_j} \\ 0 & \text{otherwise} \end{cases}$$

and $V_{j+1}^{(AP)}$ is 1 if an AP wins the contention for the channel

and 0 otherwise. *i.e.*

$$V_{j+1}^{(AP)} = \begin{cases} 1 & \text{w.p. } \beta_{Y_j+1}(1-\beta_{Y_j+1})^{Y_j} \\ 0 & \text{otherwise} \end{cases}$$

With the assumed binomial distribution for voice packet arrivals and the state dependent probabilities of attempt, it is easily seen that for $\lambda > 0$, $\{Y_j; j \geq 0\}$ is an irreducible, finite state DTMC and hence positive recurrent. The stationary probabilities, π_n, $0 \leq n \leq N$ of this DTMC can be numerically obtained. Note that π_n is the fraction of *channel slot boundaries* at which the number of STAs is n.

We now find the distribution of channel slot length as follows: Let L_j be the length of the jth channel slot which can take three possible values in units of system slots (δ): (1) one slot, when nobody attempts, or (2) T_s slots, when a successful transmission takes place, or (3) T_{col} slots when a collision takes place.

Remark: The values of T_s and T_{col} depend on the access mechanism employed. Since voice packets are of small size, we use the *basic access* mechanism. Let, L_{voice} be the length of a voice packet (including upper layer headers). Then $T_s = T_P + T_{PHY} + \frac{L_{MAC}+L_{voice}}{C_d} + T_{SIFS} + T_P + T_{PHY} + \frac{L_{ACK}}{C_c} + T_{DIFS}$ and $T_{col} = T_P + T_{PHY} + \frac{L_{MAC}+L_{voice}}{C_d} + T_{EIFS}$ where the notation is as in previous section (see Table I). Table IV gives the values of T_s and T_{col}, for different values of C_c and C_d. ∎

Then the distribution of L_j, given $Y_{j-1} = n$, is

$$L_j = \begin{cases} 1 & \text{w.p. } (1-\beta_{n+1})^{(n+1)}, \\ T_s & \text{w.p. } (n+1)\beta_{n+1}(1-\beta_{n+1})^n, \\ T_{col} & \text{otherwise.} \end{cases}$$

The process $\{(Y_j; U_j), j \geq 0\}$ can be seen to be a Markov renewal process, with cycle time L_j.

1) Obtaining the Voice Call Capacity : Let $A_j, j \geq 0$, be the number of successes of the AP in successive channel slots. A_j is 1 if the AP wins the channel contention and 0 otherwise. If $Y_{j-1} = n$, then,

$$A_j = \begin{cases} 1 & \text{w.p. } \beta_{n+1}(1-\beta_{n+1})^n, \\ 0 & \text{otherwise.} \end{cases}$$

Let $A(t)$ denote the number of successes of the AP until time t. We view the number of successes for the AP in a channel slot as the "reward" associated with that channel slot. Applying Markov regenerative analysis [14], we obtain, with probability one

$$\lim_{t \to \infty} \frac{A(t)}{t} = \frac{\sum_{n=0}^N \pi_n \, \mathsf{E}_n A}{\sum_{n=0}^N \pi_n \, \mathsf{E}_n L} =: \Theta_{AP-VoIP}(N)$$

where, $\mathsf{E}_n A = \mathsf{E}(A_j|Y_{j-1} = n)$ and $\mathsf{E}_n L = \mathsf{E}(L_j|Y_{j-1} = n)$. $\Theta_{AP-VoIP}(N)$ is the service rate of the AP in packets per slot.

The rate at which a single call sends data to the AP is λ. Since the AP serves N such calls the total input rate to the AP is $N\lambda$. Obviously, this rate should be less than $\Theta_{AP-VoIP}(N)$. Thus, we define

$$N_{max} = \max_N \ (\ \Theta_{AP-VoIP}(N) > N\lambda \ )$$

Note that we are asserting that using $N \leq N_{max}$ also ensures the delay QoS. As discussed earlier in relation to Figure 8, this is based on the observation in earlier research ([21] and [24]) that when the arrival rate is less than the saturation throughput then the delay is very small. We validate this approach in Section III-B.

Since each STA serves only one call, the number of calls at which its service rate becomes less than the input load will be more than the N_{max} obtained by the above equation. We already saw this in Figure 8, and it will be reconfirmed by additional simulations, that the AP is the capacity bottleneck in this problem.

Remark: To appreciate the importance of our refined analysis of calculating N_{max} developed above, we examine a simpler approach to find N_{max}. Instead of calculating π_n, we assume that the STAs are always non-empty, i.e., there are $N + 1$ non-empty nodes in the system always. Then the service rate applied to the AP will be

$$\Theta'_{AP-VoIP}(N) = \frac{\mathsf{E}_N A}{\mathsf{E}_N L}$$

and the maximum number of calls is then given by

$$N'_{max} = \max_N \ (\ \Theta'_{AP-VoIP}(N) > N\lambda \ )$$

We give the results in Section III-B and show that this simpler approach does not work and yields half the number of calls as compared to N_{max}. ∎

2) Mean Number of non-empty STAs : In this section we determine the time average mean number of STAs non-empty and later compare it with simulation results. This gives a further check on the accuracy of our model. Let $Y(t)$ denote the number of STAs with packets at time t. Then we need $\lim_{t\to\infty} \frac{Y(t)}{t}$. To determine the time average mean number of STAs active, we need the time average distribution of the number of non-empty STAs, ν_n, i.e., $\nu_n = \lim_{t\to\infty} \frac{\int_0^t I_{\{Y(u)=w\}} du}{t}$. We determine ν_n as follows. The process $\{(Y_j; U_j), j \geq 0\}$, is a Markov renewal process. We consider the channel slots in time units of the system slot δ. Let the reward associated with the jth cycle be the total number of slots in which the number of STAs having packets is equal to a particular value, say w, and be denoted by $R_j(w)$ in the jth channel slot. Let $\mathsf{E}_n R(w)$ be the expected reward if we have n busy STAs at the start of the channel slot. This is calculated as follows. Let $g_n^{(w)}(l)$ be the mean time spent in state w in l system slots starting with state n. $g_n^{(w)}(l)$ can

be obtained by the following recursive equations,

$$g_n^{(w)}(l) = \sum_{k=0}^{w-n} a_{(N-n),k} \ g_{n+k}^{(w)}(l-1), \text{for } n < w.$$

$$g_w^{(w)}(l) = 1 + a_{(N-w),0} \ g_w^{(w)}(l-1)$$

$$g_n^{(w)}(l) = 0, \text{for } n > w$$

where, $a_{x,k} = \text{Prob}(k \text{ arrivals from } x \text{ STAs in a system slot})$, $0 \leq k \leq x$

$\mathsf{E}_n R(w)$ will be given by,

$$\mathsf{E}_n R(w) = \sum_{l \in \{1, T_s, T_{col}\}} \text{Prob}(L_j = l|Y_{j-1} = n) \ g_n^{(w)}(l)$$

Now, we obtain

$$\lim_{t\to\infty} \frac{\int_0^t I_{\{Y(u)=w\}} du}{t} \overset{a.s.}{=} \frac{\sum_{n=0}^N \pi_n \mathsf{E}_n R(w)}{\sum_{n=0}^N \pi_n \mathsf{E}_n L} = \nu_w$$

where, $\mathsf{E}_n L = \mathsf{E}(L_j|Y_{j-1} = n)$.

Then the mean number of STAs active is given by

$$\lim_{t\to\infty} \frac{Y(t)}{t} \overset{a.s}{=} \sum_{n=0}^N n \ \nu_n$$

B. Analytical and Simulation Results

In this subsection we present the simulation results and compare them with results obtained from the analysis. The simulations were done using $ns2$ [19]. The PHY parameters were taken from the 802.11b standard which are shown in Table I. In simulations, the start time of a VoIP call is uniformly distributed in $[0, 20ms]$. This ensures that the voice packets do not arrive in bursts and remain unsynchronized.

1) Maximum Number of Calls : In Figure 10 we show the plot of AP service rate, $\Theta_{AP-VoIP}$, versus the number of calls, N, for two PHY rates $C_d = 11$Mbps and 2Mbps, for each codec. Also shown is the line $N\lambda$. We proposed the design objective $\Theta_{AP-VoIP} > N\lambda$. From the graph we can find the largest N that satisfies this requirement. For example, from Figure 10 upper graph (G711 codec), for 11 Mbps data rate, we note that the AP service rate crosses the load rate, after 12 calls. This implies that a maximum of 12 calls are possible while meeting the delay QoS, on a 802.11 WLAN. The values of N_{max} obtained for various data rates and codecs are shown in Table III. Also shown in the table are the values of N'_{max} (see the Remark at the end of Subsection III-A.1) that are almost half of the values of N_{max} obtained via our refined Markov analysis of the system.

In Figure 11, we show the simulation results for the QoS objective of $P(delay > 20ms)$, for both AP and STA packets, for different data rates and codecs. Note that the $P(delay : AP > 20ms)$ is greater than $P(delay : STA > 20ms)$ and that the AP delay shoots up before the STA delay, for any given packet size $\in \{200B, 60B\}$ and PHY data rate $\in \{11Mbps, 2Mbps\}$. This *confirms our assumption that the AP is the capacity bottleneck*. We observe that there is a
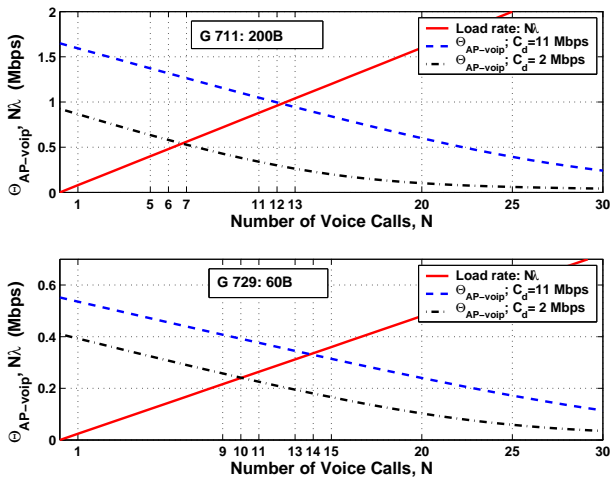
Fig. 10. The service rate $\Theta_{AP-VoIP}$ (in Mbps) applied to the AP as a function of number of voice calls, N. Also shown is the line $N\lambda$. The point where the line $N\lambda$ crosses the curves gives the maximum number of calls supported. The upper graph is for G711 codec and the lower graph is for G729 codec.
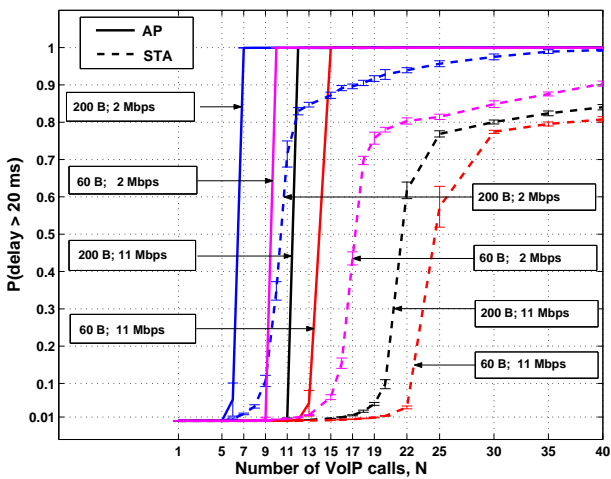

Fig. 11. Simulation results showing the probability of delay of AP and STA packets being greater than 20ms vs. the number of calls (N), for various data rates and codecs. The error bars denote 95% CI.

C_d in Mbps	G 711			G 729		
	Analysis		Sim	Analysis		Sim
	N_{max}	N'_{max}	N_{max}	N_{max}	N'_{max}	N_{max}
11	12	5	11	13	5	12
2	6	3	5	10	4	9

TABLE III

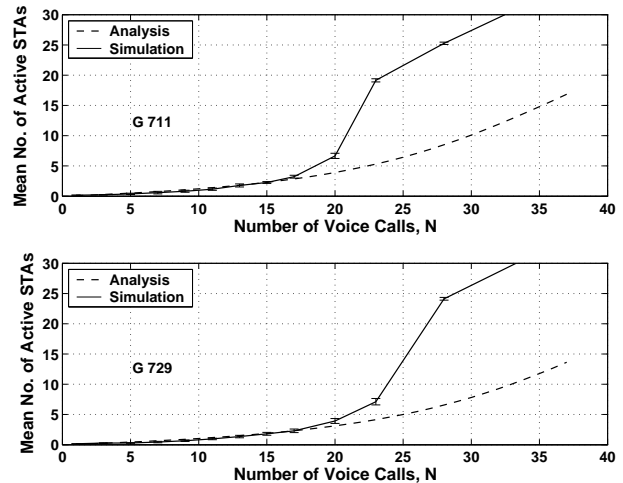ANALYTICAL AND SIMULATION RESULTS OF N_{max} FOR VARIOUS DATA RATES AND CODECS

Fig. 12. Analysis and simulation results showing mean number of active STAs vs N, for different codecs. In the upper graph, packet Size is 200B (G711 codec) and in the lower graph, packet Size is 60B (G729 codec); PHY data rate is 11Mbps.

for different codecs. We see an exact match of the plots in the region where QoS requirement is met. For both codecs (see Figure 12), beyond $N = 17$, the analysis underestimates the attempt rate, but this is well beyond the normal operating point (See Table III), and for these larger N, our model itself does not apply. The match is poor for large N (beyond the capacity) because the theoretical assumption that the STAs have only 0 or 1 packet, which is typical of the regime in which the QoS is met, is no more valid.

IV. VALIDATION OF USING SATURATION ANALYSIS ATTEMPT PROBABILITIES

The key approximation of the above analysis is,"if n nodes have non-empty queues at any channel slot boundary, then the attempt probability of a node is taken to be β_n". The β_n values are obtained from [15] where if there are n saturated nodes, the attempt probability of each node is β_n. We would like to cross-check the average attempt probability used (in our analysis) with the average attempt probability obtained from simulations in a non-saturated WLAN. This further validates the use of state dependent attempt probabilities in the analysis.

It is difficult to obtain β through simulations. This is because by definition (see [15]), β is the probability that a node attempts, conditioned on an idle channel slot having just elapsed. These events are not easily readable from the simulation trace file unless modifications in the *ns* source

value of N at which the $P(delay : AP > 20ms)$ sharply increases from a value below 0.01. This can be taken to be the voice capacity. For example, consider packet size of 200B (of G711 codec) and PHY data rate of 11 Mbps. We find that the $P(delay : AP > 20ms)$ curve sharply increases after $N = 11$, implying that $N_{max} = 11$, and is one call less than that obtained from the analysis.

Table III lists the values of N_{max} obtained from simulations, for different data rates and codecs. In all cases, our analytical N_{max} is one more than that from the simulation. Thus we may infer a rule of thumb that the system can support 1 call less than the analytical N_{max}, while providing the desired QoS.

2) Mean Number of non-empty STAs : As a further check on our model we compare the mean number of active STAs. In Figure 12, we show the plots of mean number of STAs as obtained by our analysis and as obtained via simulations,

$$\text{Prob}\ \left(B_{j+1}^{(AP)} = b | (X_j = x; L_{j+1} = l)\right) = \binom{N - x}{b} (p_l)^b (1 - p_l)^{N-x-b} \tag{3}$$

code [19] are carried out. To circumvent this problem, we derive an expression for attempt rate, i.e., if $D(t)$ denotes the cumulative number of attempts until time t, $\lim_{t\to\infty} \frac{D(t)}{t}$ shall be the average attempt rate in the WLAN. This is the average rate at which the nodes attempt or contend for the channel. The attempt rate is easily obtained from the simulation trace file since we just have to count the total number of attempts in the network and average out on the total simulation time. The analytic attempt rate can easily be obtained using the regenerative analysis (as we will show). Thus, we derive the attempt rate from the above analysis and compare it with simulation for varying number of nodes.

A. Analytical Calculation of the Attempt Rate

In order to derive the attempt rate, we need to drop the assumption that the AP is always saturated from the voice model of Section III discussed above. This is because, in the real scenario, the AP gets saturated only when the number of voice calls nears the *maximum number of calls sustainable*, while meeting the delay QoS. With the saturation assumption dropped, depending on whether the AP queue contains a packet, the total number of active nodes will be Y_j (in case no packet is there in the AP queue) or Y_j+1 (if the AP queue has at least one packet). The Markov Chain $\{Y_j; j \geq 0\}$ only provides the number of active STAs in the WLAN at the channel slot boundaries. Additionally, we need to know the state of AP queue so as to know the number of active nodes at the channel slot boundaries. Therefore we now model the buffer occupancy of the AP.

Let X_j be the number of packets in the AP queue and $B_j^{(AP)}$ be the number of new packets arriving at AP queue at the end of jth channel slot. Then

$$Y_{j+1} = Y_j - V_{j+1}^{(STA)} + B_{j+1}$$
$$X_{j+1} = X_j - V_{j+1}^{(AP)} + B_{j+1}^{(AP)}$$

with the condition $V_{j+1}^{(STA)} + V_{j+1}^{(AP)} \in \{0,1\}$. $V_j^{(STA)}$, $V_j^{(AP)}$ and B_j are as defined in Section III. On similar lines as B_j, $B_j^{(AP)}$ can be modeled as having a binomial distribution. Observe that if x packets are already there in AP queue, at most only $N - x$ packets can arrive before the QoS delay bound of the earliest arrived packet gets exceeded. Then the probability $prob(B_{j+1}^{(AP)}|X_j, L_{j+1})$, is given by Equation (3) (displayed above). It can be seen that $\{(Y_j, X_j); j \geq 0\}$ forms a positive recurrent DTMC and the stationary probabilities, $\pi_{y,x}$, $0 \leq y, x \leq N$, can be numerically found.

We make use of Markov regenerative framework to find the attempt rate. In order to apply the renewal reward theorem

for point processes, we need the mean renewal cycle time and hence we identify the distribution of L_j as follows: Define $Z_j := Y_j + 1$ if $X_j \neq 0$ and $Z_j := Y_j$ if $X_j = 0$, at the instant U_j. Let $\eta(Z_j)$ be the probability of the $(j+1)^{th}$ channel slot being idle, $\alpha(Z_j)$ be the probability that a STA succeeds, $\sigma(Z_j)$ be the probability that the AP succeeds and $\zeta(Z_j)$ be the probability that there is a collision. Then L_{j+1} takes the three values with the following probabilities.

$$L_{j+1} = \begin{cases} 1 & \text{w.p. } \eta(Z_j) \\ T_s & \text{w.p. } \sigma(Z_j) + \alpha(Z_j) \\ T_{col} & \text{w.p. } \zeta(Z_j) \end{cases}$$

where T_s and T_{col} are as defined before and

$$\begin{aligned} \eta(Z_j) &= (1 - \beta_{Z_j})^{Z_j}, \\ \alpha(Z_j) &= Y_j \beta_{Z_j} (1 - \beta_{Z_j})^{Z_j - 1}, \\ \sigma(Z_j) &= I_{\{X_j \neq 0\}} \beta_{Z_j} (1 - \beta_{Z_j})^{Z_j - 1}, \\ \zeta(Z_j) &= 1 - (\alpha(Z_j) + \sigma(Z_j) + \eta(Z_j)) \end{aligned}$$

$I_{\{X_j \neq 0\}}$ is as usual, an indicator function denoting that AP has packets to send.

The process $\{(Y_j, X_j; U_j), j \geq 0\}$ can be seen to be a Markov renewal process with L_j being the renewal cycle time. Let D_j be the number of attempts in the network when any node contends for the channel in the channel slot j. Since we are interested in the system attempt rate, D_j is the reward in cycle j. If there are n nodes active at the $(j-1)^{th}$ channel slot boundary, (i.e., $Z_{j-1} = n$), then we have,

$$D_j = i \ \text{w.p.} \ \binom{n}{i} \beta_n^i (1 - \beta_n)^{n-i}$$

Let $\mathsf{E}D$ be the mean number of attempts. Then

$$\mathsf{E}D = \sum_{i=1}^{n} i \binom{n}{i} \beta_n^i (1 - \beta_n)^{n-i} = n\beta_n$$

Let $D(t)$ denote the cumulative number of attempts until time t. Applying Markov regenerative analysis [14], we obtain the net attempt rate of nodes in the WLAN, $\Phi(N)$ as

$$\Phi(N) = \lim_{t\to\infty} \frac{D(t)}{t} \overset{a.s.}{=} \frac{\sum_{y=0}^{N} \sum_{x=0}^{N} \pi_{y,x} \ \mathsf{E}_{y,x}D}{\sum_{y=0}^{N} \sum_{x=0}^{N} \pi_{y,x} \ \mathsf{E}_{y,x}L}$$

where, $\mathsf{E}_{y,x}D = \mathsf{E}(D_j|(Y_{j-1}, X_{j-1}) = (y, x))$ and $\mathsf{E}_{y,x}L = \mathsf{E}(L_j|(Y_{j-1}, X_{j-1}) = (y, x))$ and $\Phi(N)$ is in attempts/slot.

B. Numerical Results and Validation

Figure 13 shows the attempt rate of a node vs number of VoIP calls obtained from analysis and simulation in the region of operation. The simulation is done using the parameters as in Table I. As before, in simulations, the start time of a VoIP call is uniformly distributed in $[0, 20ms]$. The error bars in simulation curve denote the 95% confidence intervals. The error between analysis and simulation is less than 5%. Thus we have further validated the approach of Section III.

$$\text{Prob}\left(B_{j+1}^{(1)} = b | ((Y_j^{(1)}, Y_j^{(2)}) = (y_1, y_2); L_{j+1} = l)\right) = \binom{N_1 - y_1}{b} (p_l)^b (1 - p_l)^{N_1 - y_1 - b} \qquad (4)$$

$$\text{Prob}\left(B_{j+1}^{(2)} = b | ((Y_j^{(1)}, Y_j^{(2)}) = (y_1, y_2); L_{j+1} = l)\right) = \binom{N_2 - y_2}{b} (p_l)^b (1 - p_l)^{N_2 - y_2 - b} \qquad (5)$$
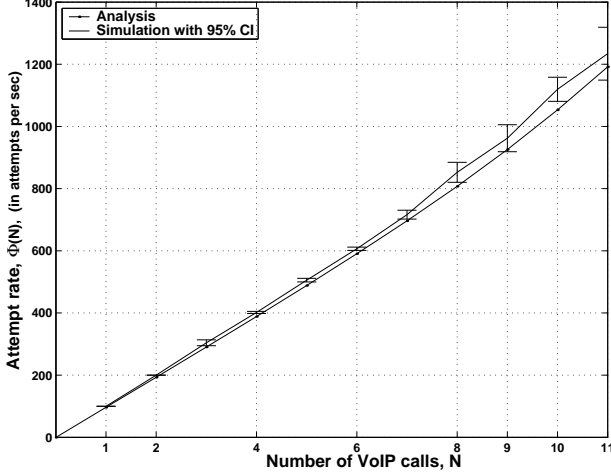

Fig. 13. Results from analysis and simulation: The total attempt rate $\Phi(N)$ obtained vs number of voice calls, N. Packet size is 200B (G711 Codec); data rate is 11 Mbps and control rate is 2Mbps.

V. MODEL FOR TWO TYPES OF VoIP CALLS

We now consider a case where the VoIP calls originate from two types of codec. We answer the question: "When two different types of codecs are used, how many VoIP calls can be set up to different STAs such that VoIP call QoS is met?" We assume that Type 1 voice calls have a larger packet size than Type 2 calls. Let Type 1 calls use the G.711 codec and the Type 2 calls use G.729 codec. Then, as assumed before, Type 1 calls generate 1 packet of 200 bytes every 20 ms and Type 2 calls generate 1 packet of 60 bytes every 20 ms. We obtain an analytical approximation for the number of calls of each type that can be admitted so that QoS is met. We extend the analysis of Section III for this scenario.

A. Stochastic Modeling

The modeling assumptions remain the same as in Section III. The STAs shall have at most one packet in their queue and the AP is assumed to be saturated. Let N_1 and N_2 be the total number of calls of Type 1 and Type 2 respectively. Let $Y_j^{(1)}$ be the number of non-empty STAs of Type 1 and $Y_j^{(2)}$ be the number of non-empty STAs of Type 2 call stations at the instant U_j. Thus $0 \le Y_j^{(1)} \le N_1$ and $0 \le Y_j^{(2)} \le N_2$. Let L_j be the length of the channel slot, j, as defined earlier. Let $B_j^{(1)}$ and $B_j^{(2)}$ be the number of new packet arrivals of Type 1 and Type 2 calls respectively. Let $V_j^{(AP)}$ be the number of departures from AP, and $V_j^{(STA1)}$ and $V_j^{(STA2)}$ be the number of departures from STAs of Type 1 calls and Type 2 calls respectively in the jth channel slot. At most one

departure can happen in any channel slot. Thus,

$$Y_{j+1}^{(1)} = Y_j^{(1)} - V_{j+1}^{(STA1)} + B_{j+1}^{(1)}$$
$$Y_{j+1}^{(2)} = Y_j^{(2)} - V_{j+1}^{(STA2)} + B_{j+1}^{(2)}$$

with the condition $V_{j+1}^{(STA1)} + V_{j+1}^{(STA2)} + V_{j+1}^{(AP)} \in \{0, 1\}$.

Since we assume that packets arrive at only empty STAs, $B_j^{(1)}$ and $B_j^{(2)}$ can be modeled as having a binomial distribution, as done in Section III, and the probabilities $prob(B_{j+1}^{(1)} | Y_j, L_{j+1})$ and $prob(B_{j+1}^{(2)} | Y_j, L_{j+1})$ are given by Equations (4) and (5) respectively (displayed above).

We again employ the approximation that if n nodes are contending (i.e., have non empty queues), then the attempt probability is taken to be β_n and is obtained from [15] with n saturated nodes. Thus when there are $Y_j^{(1)}$ Type 1 STAs and $Y_j^{(2)}$ Type 2 STAs contending, the total number of contending STAs is $Y_j := Y_j^{(1)} + Y_j^{(2)}$. Hence, including the AP we take the attempt probability to be β_{Y_j+1}.

For convenience, let us define the following probability functions of the activities in the $(j+1)^{th}$ channel slot: Let $\eta(Y_j^{(1)}, Y_j^{(2)})$ be the probability of channel slot being idle, $\alpha_1(Y_j^{(1)}, Y_j^{(2)})$ be the probability that a STA with Type 1 packet succeeds, $\alpha_2(Y_j^{(1)}, Y_j^{(2)})$ be the probability that a STA with Type 2 packet succeeds, $\sigma_1(Y_j^{(1)}, Y_j^{(2)})$ be the probability that the AP succeeds and sends Type 1 packet, $\sigma_2(Y_j^{(1)}, Y_j^{(2)})$ be the probability that the AP succeeds and sends Type 2 packet, $\zeta_1(Y_j^{(1)}, Y_j^{(2)})$ be the probability that there is a long collision (involving at least one Type 1 packet) and $\zeta_2(Y_j^{(1)}, Y_j^{(2)})$ be the probability that there is a short collision (not involving a Type 1 packet). These are expressed, using the state dependent attempt probabilities, as below:

$$\begin{aligned}
\eta(Y_j^{(1)}, Y_j^{(2)}) &= (1 - \beta_{Y_j+1})^{(Y_j+1)}, \\
\alpha_1(Y_j^{(1)}, Y_j^{(2)}) &= Y_j^{(1)} \beta_{Y_j+1} (1 - \beta_{Y_j+1})^{Y_j}, \\
\alpha_2(Y_j^{(1)}, Y_j^{(2)}) &= Y_j^{(2)} \beta_{Y_j+1} (1 - \beta_{Y_j+1})^{Y_j}, \\
\sigma_1(Y_j^{(1)}, Y_j^{(2)}) &= p_1 \beta_{Y_j+1} (1 - \beta_{Y_j+1})^{Y_j}, \\
\sigma_2(Y_j^{(1)}, Y_j^{(2)}) &= p_2 \beta_{Y_j+1} (1 - \beta_{Y_j+1})^{Y_j},
\end{aligned}$$

$\zeta_1(Y_j^{(1)}, Y_j^{(2)})$ and $\zeta_2(Y_j^{(1)}, Y_j^{(2)})$ are given by Equations (6) and (7), and $p_1 = \frac{N_1}{N_1 + N_2}; p_2 = \frac{N_2}{N_1 + N_2}$

Then $V_j^{(STA1)}$ is 1 if an STA with Type 1 call wins the contention for the channel and 0 otherwise and is given as

$$V_{j+1}^{(STA1)} = \begin{cases} 1 & \text{w.p. } \alpha_1(Y_j^{(1)}, Y_j^{(2)}) \\ 0 & \text{otherwise} \end{cases}$$

$$\zeta_1\left(Y_j^{(1)}, Y_j^{(2)}\right) = p_1\beta_{Y_j+1}\sum_{l_2=1}^{Y_j^{(2)}}\binom{Y_j^{(2)}}{l_2}\beta_{Y_j+1}^{l_2}(1-\beta_{Y_j+1})^{Y_j-l_2} + \sum_{l_1=2}^{Y_j^{(1)}}\binom{Y_j^{(1)}}{l_1}\beta_{Y_j+1}^{l_1}(1-\beta_{Y_j+1})^{Y_j+1-l_1}$$
$$+ \sum_{l_1=1}^{Y_j^{(1)}}\sum_{l_2=1}^{Y_j^{(2)}+1}\binom{Y_j^{(1)}}{l_1}\beta_{Y_j+1}^{l_1}\binom{Y_j^{(2)}}{l_2}\beta_{Y_j+1}^{l_2}(1-\beta_{Y_j+1})^{Y_j+1-l_1-l_2} \tag{6}$$

$$\zeta_2\left(Y_j^{(1)}, Y_j^{(2)}\right) = p_2\beta_{Y_j+1}\sum_{l_2=1}^{Y_j^{(2)}}\binom{Y_j^{(2)}}{l_2}\beta_{Y_j+1}^{l_2}(1-\beta_{Y_j+1})^{Y_j-l_2} + \sum_{l_2=2}^{Y_j^{(2)}}\binom{Y_j^{(2)}}{l_2}\beta_{Y_j+1}^{l_2}(1-\beta_{Y_j+1})^{Y_j+1-l_2} \tag{7}$$

C_c	C_d	L_j in system slots			
		$T_s = T_{s1}$	T_{s2}	$T_{col} = T_{c-long}$	$T_{c-short}$
2	2	72	44	75	47
2	5.5	43	32	45	35
2	11	34	29	37	32
1	2	75	47	75	47
1	5.5	45	35	45	35
1	11	37	32	37	32

TABLE IV

VALUES OF L_j FOR VARIOUS DATA RATES AND CONTROL RATES, USING BASIC ACCESS MECHANISM.

Similarly $V_j^{(STA2)}$ and $V_{j+1}^{(AP)}$ can be expressed as below

$$V_{j+1}^{(STA2)} = \begin{cases} 1 & \text{w.p. } \alpha_2(Y_j^{(1)}, Y_j^{(2)}) \\ 0 & \text{otherwise} \end{cases}$$

$$V_{j+1}^{(AP)} = \begin{cases} 1 & \text{w.p. } \sigma_1(Y_j^{(1)}, Y_j^{(2)}) + \sigma_2(Y_j^{(1)}, Y_j^{(2)}) \\ 0 & \text{otherwise} \end{cases}$$

Then it is easily seen that $\{Y_j^{(1)}, Y_j^{(2)}; j \geq 0\}$ forms a finite irreducible two dimensional discrete time Markov chain on the channel slot boundaries and hence is positive recurrent. The stationary probabilities π_{n_1,n_2} of the Markov Chain $\{Y_j^{(1)}, Y_j^{(2)}; j \geq 0\}$ can then be numerically determined using distributions of $B_j^{(1)}$, $B_j^{(2)}$, $V_j^{(STA1)}$, $V_j^{(STA2)}$ and $V_j^{(AP)}$, and the probability functions defined before. L_j, the channel slot duration, can take five values (in number of system slots): 1 if it is an idle slot, T_{s1} if it corresponds to a successful transmission of a node with a Type 1 call, T_{s2} if it corresponds to a successful transmission of a node with a Type 2 call, T_{c-long} if it corresponds to a collision between one Type 1 node and any node, and $T_{c-short}$ if it corresponds to a collision involving only Type 2 packets. Let L_{voice1} and L_{voice2} be the lengths of G711 voice packet and G729 voice packet respectively. The expressions for various channel slot values are: $T_{s1} = T_P + T_{PHY} + \frac{L_{MAC}+L_{voice1}}{C_d} + T_{SIFS} + T_P + T_{PHY} + \frac{L_{ACK}}{C_c} + T_{DIFS}$, $T_{s2} = T_P + T_{PHY} + \frac{L_{MAC}+L_{voice2}}{C_d} + T_{SIFS} + T_P + T_{PHY} + \frac{L_{ACK}}{C_c} + T_{DIFS}$, $T_{c-long} = T_P + T_{PHY} + \frac{L_{MAC}+L_{voice1}}{C_d} + T_{EIFS}$, and $T_{c-short} = T_P + T_{PHY} + \frac{L_{MAC}+L_{voice2}}{C_d} + T_{EIFS}$. See Table I for values of parameters. Table IV gives the different values of L_j for various rates, using 802.11b. The distribution of L_j is then given as

$$L_{j+1} = \begin{cases} 1 & \text{w.p. } \eta(Y_j^{(1)}, Y_j^{(2)}) \\ T_{s1} & \text{w.p. } \alpha_1(Y_j^{(1)}, Y_j^{(2)}) + \sigma_1(Y_j^{(1)}, Y_j^{(2)}) \\ T_{s2} & \text{w.p. } \alpha_2(Y_j^{(1)}, Y_j^{(2)}) + \sigma_2(Y_j^{(1)}, Y_j^{(2)}) \\ T_{c-long} & \text{w.p. } \zeta_1(Y_j^{(1)}, Y_j^{(2)}) \\ T_{c-short} & \text{w.p. } \zeta_2(Y_j^{(1)}, Y_j^{(2)}) \end{cases}$$

The process $\{(Y_j^{(1)}, Y_j^{(2)}; U_j), j \geq 0\}$ can be seen to be a Markov renewal process with L_j being the renewal cycle time. As before, we use the Markov regenerative framework to find the WLAN VoIP call capacity, as follows.

B. VoIP Call Capacity

Let A_j be the reward when the AP wins the channel contention. If there are n_1 STAs of Type 1 calls active and n_2 STAs of Type 2 calls active at the $(j-1)^{th}$ channel slot boundary, by taking $n = n_1 + n_2$, we have,

$$A_j = \begin{cases} 1 \text{ w.p. } \beta_{n+1}(1-\beta_{n+1})^n \\ 0 \text{ otherwise} \end{cases}$$

Let $A(t)$ denote the cumulative reward of the AP until time t. Applying Markov regenerative analysis (or the renewal reward theorem) we obtain the service rate of the AP, in packets per slot, as

$$\Theta_{AP-VoIP}(N_1, N_2) = \lim_{t\to\infty}\frac{A(t)}{t} \overset{a.s.}{=} \frac{\sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2}\pi_{n_1,n_2}\mathsf{E}_{n_1,n_2}A}{\sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2}\pi_{n_1,n_2}\mathsf{E}_{n_1,n_2}L}$$

where, $\mathsf{E}_{n_1,n_2}A = \mathsf{E}(A_j|(Y_{j-1}^{(1)}, Y_{j-1}^{(2)}) = (n_1, n_2))$ and $\mathsf{E}_{n_1,n_2}L = \mathsf{E}(L_j|(Y_{j-1}^{(1)}, Y_{j-1}^{(2)}) = (n_1, n_2))$. Since the rate at which a single call sends data to the AP is λ, and the AP serves $N(= N_1 + N_2)$ such calls the total load rate at the AP is $(N_1 + N_2)\lambda$ (= $\gamma(N_1, N_2)$ say). Obviously, this rate should be less than $\Theta_{AP-VoIP}(N_1, N_2)$ for stability. Thus, for permissible combination of N_1 and N_2 calls we need $\Theta_{AP-VoIP}(N_1, N_2) > (N_1 + N_2)\lambda$. This inequality defines the admission region.

C. Numerical Results and Validation

We present our simulation results and compare them with results obtained from the simulation. The simulations were done using ns 2 [19]. Again, as before, in simulations, the start time of all VoIP calls is uniformly distributed in
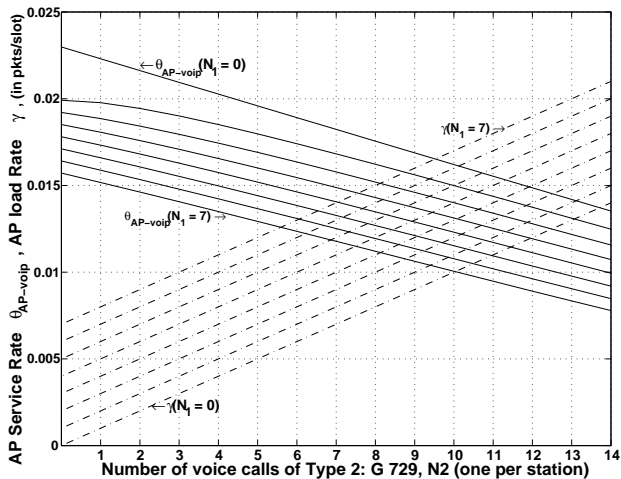
Fig. 14. Results from analysis: The service rate $\Theta(N_1, N_2)$ applied to the AP vs number of voice calls, N_2 for different values of N_1. Also shown are lines $\gamma(N_1, N_2) = (N_1 + N_2)\lambda$ for different values of N_1. The point where the γ line crosses the curve for a fixed value of N_1 gives the maximum number of calls supported; N_1 use G711 Codec and N_2 use G729 Codec. The PHY data rate is 11 Mbps and control rate is 2Mbps.
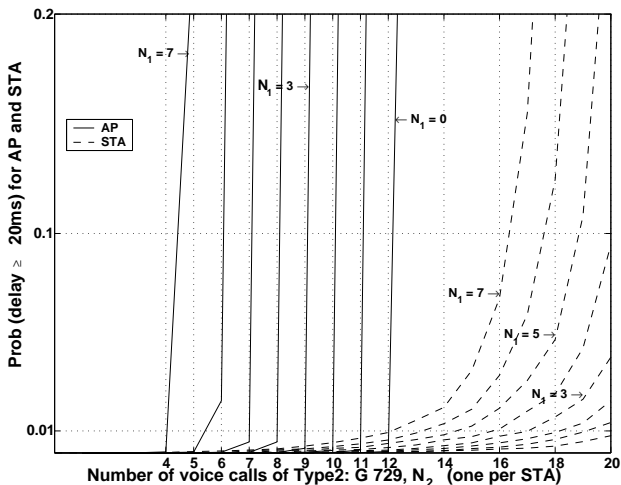

Fig. 15. Results from simulation: The $Prob(delay \geq 20ms)$ at AP and STA vs number of voice calls, N_2. N_1 use G711 Codec and N_2 use G729 Codec. The PHY data rate is 11 Mbps and control rate is 2Mbps.

$[0, 20ms]$. In Figure 14 we plot the numerical results for the AP service rate (solid lines) and load arrival rate (dot-dashed lines) at the AP vs. values of N_2. The different curves correspond to different values of N_1 starting from 0. The simulation results for the QoS objective of $Prob(delay \geq 20ms)$ for the AP and the STAs are shown in Figure 15.

From Figure 14 we observe that for each value N_2, as we increase the value of N_1 the service rate available to the AP decreases. This is, of course, because more service needs to be given to the STAs as the number of calls increases. Observe that for $N_1 = 0$, the rate of packets arriving into the AP is $N_2\lambda$ packets per slot. This exceeds the curve $\theta_{AP-VoIP}(0, N_2)$ after $N_2 = 13$ but before $N_2 = 14$. Hence, from the analysis, we can conclude that the pair $(N_1 = 0, N_2 = 13)$ can be admitted. Looking at Figure

15, we find that for $N_1 = 0$, the $Prob(delay : AP \geq 20ms)$ shoots up after $N_2 = 12$. As in Section III we find that our analysis overestimates the capacity by 1 call. Similarly, for $N_1 = 7$, the analysis says that we can permit $N_2 = 5$, whereas the simulations show that we can permit $N_2 = 4$.

These observations are also summarized in Figure 16, where the $\circ$ symbols show the (N_1, N_2) pair, admissible by the simulations and the $*$ symbols show the call admission points obtained by analysis. Thus the analysis captures the admissible region very well, and in practice we can use the rule of thumb of accepting one call less than that given by the analysis.
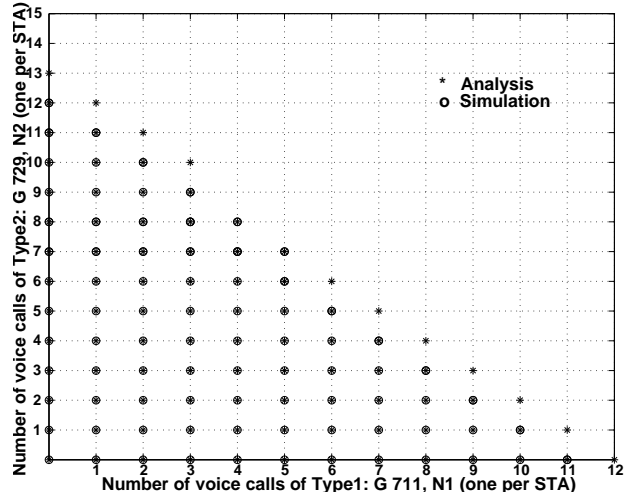

Fig. 16. Analysis and simulation results: The admissible combinations of Type 1 and Type 2 calls. N_1 use G711 Codec and N_2 use G729 Codec.

VI. CONCLUSION

In this paper we analysed two traffic scenarios that represent two of the most common applications that are carried over WLANs.

Firstly, we considered a system with N TCP connections downloading files in a single cell of an IEEE 802.11 WLAN. The system throughput was accurately determined. To further check the model's accuracy other quantities such as the distribution of the number of STAs with ACKs and the collision probability of the AP were provided. They matched well with the simulations.

We also formed an accurate analytical model for VoIP calls over a single cell of an 802.11 WLAN. Our model was able to determine the maximum number of calls that can be supported by a single cell infrastructure 802.11 WLAN. Results were provided for different PHY data rates and codecs. The results obtained were verified with simulations. We further validated the modeling approach of using the saturated attempt probabilities of [2] and [15] as state dependent attempt probabilities. Then, we extended the VoIP model for a special case where the VoIP calls are from different codecs. Again the analytical results match well with the simulation results.

Our work provides the following modeling insights:

1) The idea of using saturation attempt probabilities as state dependent attempt rates yields an accurate model in the unsaturated case.

2) Using this approximation, an IEEE 802.11 infrastructure WLAN can be well modeled by a Markov renewal process embedded at channel slot boundaries.

In related work, we have used the approach of this paper to model the performance of voice calls, video streaming sessions and data transfers, in an IEEE 802.11e WLAN. Our preliminary results, with combined TCP transfers and packet voice, are reported in [12] and [11]. The model including streaming video has recently been submitted for publication.

REFERENCES

[1] F. Anjum, M. Elaoud, D. Famolari, A. Ghosh, R. Vaidyanathan, A. Dutta, P. Agrawal, T. Kodama, and Y. Katsube. Voice performance in WLAN networks - an experimental study. In *IEEE GLOBECOM '03*, volume 6, pages 3504 – 3508, 1-5 Dec 2003.

[2] G. Bianchi. Performance Analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE JSAC*, pages 535–547, March 2000.

[3] R. Bruno, M. Conti, and E. Gregori. Throughput Analysis of TCP clients in Wi-Fi Hot Spot Networks. In *Networking '04*, LNCS 2042, pages 626–637, 2004.

[4] R. Bruno, M. Conti, and E. Gregori. Performance Modelling and Measurements of TCP Transfer Throughput in 802.11based WLANs. In *MSWiMS '06*, 2-6 Oct 2006.

[5] F. Cali, M. Conti, and E. Gregori. Dynamic Tuning of the IEEE 802.11 protocol to achieve a Theoretical Throughput Limit. *IEEE/ACM Trans. On Networking*, 8(6):785–799, dec 2000.

[6] H. S. Chhaya and S. Gupta. Performance Modelling of Asychronous Data Transfer Methods of IEEE 802.11 MAC Protocol. *Wireless Networks*, 3(3):217–234, 1997.

[7] A. Detti, E. Graziosi, V. Minichiello, S. Salsano, and V. Sangregorio. TCP fairness issues in IEEE 802.11 based access networks. preprint, 2005.

[8] K. Duffy, D. Malone, and D. J. Leith. Modelling 802.11 Wireless Links. In *CDC-ECC '05*, volume 3, pages 6952 – 6957, 12-15 Dec 2005.

[9] S. Garg and M. Kappes. Can I add a VoIP call? In *IEEE ICC 2003*, volume 2, pages 779–783, 11 - 15 May 2003.

[10] M. Gong, Q. Wu, and C. Williamson. Queue Management Strategies to improve TCP Fairness in IEEE 802.11 Wireless LANs. In *RAWNET '06 workshop*, 3 April 2006.

[11] S. Harsha, SVR Anand, A. Kumar, and V. Sharma. An analytical model for capacity evaluation of VoIP on HCCA and TCP file transfers over EDCA in an IEEE 802.11e WLAN. In *ICDCN '06*, pages 245–256, 27-30 Dec 2006.

[12] S. Harsha, A. Kumar, and V. Sharma. An Analytical Model for the Capacity Estimation of Combined VoIP and TCP File Transfers over EDCA in an IEEE 802.11e WLAN. In *IEEE IWQoS '06*, pages 178 – 187, 19-21 June 2006.

[13] G. H. Hwang and D. H Cho. Voice capacity in IEEE 802.11 wireless LANs. In *Electronic Letters*, volume 40, issue 18, pages 1137 – 1138, 2 Sept 2004.

[14] V. G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman and Hall/ CRC Press, 1996.

[15] A. Kumar, E. Altman, D. Miorandi, and M. Goyal. New Insights from a Fixed Point Analysis of Single Cell IEEE 802.11 WLANs. In *IEEE INFOCOM '05*, pages 1550 – 1561, 13 - 17 March 2005.

[16] D. J. Leith and P. Clifford. Using the 802.11e EDCF to achieve TCP upload fairness over WLAN Links. In *WIOPT*, pages 109–118, 2005.

[17] K. Medepalli, P. Gopalakrishnan, D. Famolari, and T. Kodama. Voice Capacity of IEEE 802.11b, 802.11a and 802.11g Wireless LANs. In *IEEE GLOBECOM 2004*, volume 3, pages 1549–1553, 29 Nov - 3 Dec 2004.

[18] D. Miorandi, A.A. Kherani, and E. Altman. A Queueing Model for HTTP traffic over IEEE 802.11 WLANs. *Computer Networks*, 50:63–79, 2006.

[19] The Network Simulator ns2 http://www.isi.edu/nsnam/ns/.

[20] S. Pilosof, R. Ramjee, Y. Shavitt, and P. Sinha. Understanding TCP fairness over wireless LAN. In *IEEE INFOCOM '05*, pages 863–872, 30 March - 3 April 2003.

[21] M. Prabhakaran. Design and Analysis of Link Scheduling Algorithms for Wireless Networks. Master's thesis, Indian Institue of Science, Bangalore, June 2004.

[22] IEEE 802.11 standard for Wireless Local Area Networks. *h*ttp://standards.ieee.org/getieee802/802.11.html.

[23] J.V. Sudarev, L.B. White, and S. Perreau. Performance Analysis of 802.11 CSMA/CA for Infrastructure Networks under Finite Load Conditions. In *LANMAN '05*, volume 3, pages 1 – 6, 18-21 Sept 2005.

[24] O. Tickoo and B. Sikdar. A Queueing Model for Finite Load IEEE 802.11 Random Access MAC. In *IEEE ICC, 2004*, volume 1, pages 175 – 179, 20-24 June 2004.