Saturation Rate in 802.11 Revisited

Tara Javidi, Mingyan Liu and Rajiv VijayaKumar

Abstract

There have been extensive studies on the performance of IEEE 802.11 networks, most of which focus on the *saturation* regime where user queues are always full. This paper departs from prior work in that we focus on non-saturated regimes. Specifically, we consider two scenarios: one consisting of a mixture of saturated and non-saturated users, the other consisting of users of buffer size one. In both cases we show that the system throughput achieved can be significantly greater than the saturation throughput achieved in a homogeneous environment of the same number of users with infinite queues. With these results we argue that the notion of saturation throughput is an inherently pessimistic one. When there is natural heterogeneity in the system or when queues are forced to be empty from time to time due to limited buffer size, the effect of statistical multiplexing emerges, which leads to increased system throughput.

I. INTRODUCTION

With the wide deployment of wireless LAN, its core enabling technology, the IEEE 802.11 medium access control (MAC) protocol, has been extensively studied in recent years. These studies include the performance as well as fairness properties of 802.11 basic and DCF (distributed coordination functions) options.

Many of these studies examine the behavior of a fixed number of users (or stations/clients) using 802.11 under a special operating regime known as the *saturation* regime; notable examples include [1]. This is a scenario where all users in the system are *infinite sources*, i.e., they always have a packet to send or equivalently they have infinitely many packets waiting in the queues, thus the term saturation. Saturation studies focus on deriving the *saturation throughput*, the throughput that each queue or the system as a whole can achieve under the saturation scenario. These are quantities that vary with the number of users in the system; they reflect in a sense the capacity of the system and provide significant insights in understanding the limiting behavior of 802.11. Biancci in [1] first proposed a Markov chain based modeling scheme to estimate the saturation throughput, which was then used to optimize the backoff window size. Similar models were also used in [2] and [3]. Authors in [4] further studied the saturation throughput as a fixed point problem and examined its existence and uniqueness.

By contrast, in our previous work [5], we considered a very different system operating regime, where user queues are fed with finite arrival rates. This means that any given user might not always have a packet to send. We follow our adopted notational convention and refer to this type of queues as *finite sources* to distinguish from the infinite sources/queues used in saturation studies. Whereas in [5], we investigated how the system behaves when these arrival rates approach the saturation rate (throughput) *from below*, the present paper is an attempt to achieve higher than saturation throughput. In particular, the present paper expands on the following two observations presented in [5]:

1. The queues have very good delay performance even when all queues have arrival rates approaching the saturation throughput. This is consistent with other non-saturation studies in literature [6].

2. The achievable throughput of an infinite source queue when some of the users are finite source queues (but with arrival rates approach their respective saturation throughput) exceeds that achieved when all queues are infinite sources.

The rest of the paper is organized as follows. In Section II, we motivate our study using results from previous experimental results. For that, we consider the behavior of a symmetric system where all users have similar channel and arrival characteristics and are all equipped with an infinite buffer. We, then, demonstrate the existence of a "broken ergodicity" phenomenon in the simulations, where a very good delay performance is exhibited in a system that is

T. Javid is with the ECE Department at UC San Diego (tara@ece.ucsd.edu), M. Liu is with the the EECS Department at University of Michigan, Ann Arbor (mingyan@eecs.umich.edu), and R. Vijayakumar is with the ECE Department at University of Washington, Seattle (rajivv@washington.edu).

provably unstable. This is due to a monotonically decreasing nature of the saturation throughput curve with the number of users. In other words, while the number of backlogged queues is lower than N, the instantaneous throughput of the system is strictly larger than $\lambda_{\text{sat}}(N)$, hence the arrival rate, causing a strong negative average drift away from all backlogged state. This underlines the main premise of our work in identifying the notion of saturation throughput as a pessimistic one, hence motivating our quest in search of higher than saturation throughput.

In Section III, we quantify how the unused bandwidth by the finite users is traded and being used by saturated users (who have infinite demand). For this, we propose a simple model to capture the essential characteristic of a finite source. We, then, use this model to study the performance of a network consisting of a mixture of users modeled as proposed and saturated ones. In this section, we show the increase in throughput gained via asymmetry in arrival patterns among users.

In Section IV, we answer the following important question: In the absence of saturated users, is it possible to for each and every one user to achieve a throughput higher than that of its saturation? From Section II, we know that with infinite buffers the answer to this question is no; hence, in Section IV, we focus on a finite buffer scenario. In particular, we show that reducing buffer capabilities of each user to a maximum of one significantly increases the throughput of the system. This increase in overall throughput with the reduction of buffer size might seem counter intuitive at the first glance, but can be easily explained as follows: In the presence of limited buffer space for each user, we force a certain time sharing among various backlogged states, guaranteeing a higher overall throughput. This is similar to convexifying a capacity regain. We conclude the paper in Section V.

II. From Saturation to Non-Saturation

This section provides the underlying motivation for this study. We first show that saturation throughput as defined in [1] is the maximum achievable average throughput for a system of identical users. On the other hand empirical results show that for finite users, if we increase their arrival rate to approach the saturation rate from below or even slightly over, the queues exhibit very good performance in terms of packet delay and queue backlog. We provide an explanation for this apparent discrepancy, which further motivate us to consider two types of scenarios which produce higher than saturation throughput, discussed in the next two sections.

Fact 1: For any Markovian arrival process, which is independent across queues, the maximum achievable average throughput for each user is $\lambda_{\text{sat}}(N)/N$.

Proof: We have $\sum_{i=1}^{N} \lambda_i > \lambda_{\text{sat}}(N)$. This implies that for large $\min_i q_i$, where q_i is the queue backlog of user i, the total rate of service is less than arrival rate. This mean that there is a strictly positive drift with respect of the state of the system (the vector of queue backlogged) along all coordinates. In other words, $\exists \delta, M > 0$ such that for $\forall i$

$$E\left\{ |q_i(t+1) - q_i(t)| \;\middle|\; \min_i q_i(t) > M \right\} > \delta.$$

This completes the proof. ∎

On the other hand, using simulation we can obtain some rather interesting results on the behavior of such a system when we increase the load to approach the saturation rate from below. The simulations reported in this paper were run using Opnet (Release 11) and its built-in 802.11b model (the basic function). Parameters used are as follows. The physical layer rate used was 1Mbps. Data packets were generated according to independent Poisson processes at each client and were passed directly to the MAC layer (i.e. no IP or other encapsulation was used); all packets had a fixed size of 1024 bytes. The values of all the relevant MAC layer parameters (which we left at their default values) are listed in Table II. We did not use RTS/CTS for any transmissions. All nodes in the simulation were within 50 meters of each other, and the background noise level was set to 0 W to ensure that a transmission was successful if and only if there

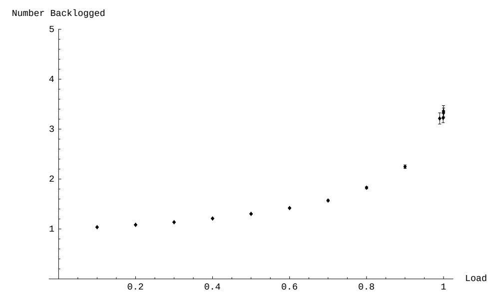

Fig. 1. Expected number of clients backlogged (out of a total of 20) as the load (fraction of saturation throughput) varies from 0.1 to 1.0.

were no other simultaneous transmissions. The size of the MAC layer buffer was set to infinity. Each data point in the results was obtained by averaging over a simulation run of 60 minutes.

slot time	50 μs
SIFS	28 μs
DIFS	128 μs
$\mathrm{CW}_{\mathrm{min}}$	15
$\mathrm{CW}_{\mathrm{max}}$	1023
Physical layer headers	128 bits
MAC layer headers	272 bits
ACK frame	240 bits
Retry Limit	7

TABLE I
802.11B PARAMETERS

Figure 1 shows the simulated expected number of backlogged clients, as the arrival rate (load) increases, in a symmetric system where all users have similar channel and arrival characteristics and are all equipped with an infinite buffer. All queues are initially empty and the simulation run time is 60 minutes. What we see here is that the observed average delay is much lower than anticipated even when the arrival rate is at or slightly above the saturation throughput. In other words, for a system of N identical users, given an initial state of all empty queues, at arrival rate $N\lambda = \epsilon + \lambda_{\text{sat}}(N)$, $\epsilon \approx 0$, where λ is the arrival rate of a single user and $\lambda_{\text{sat}}(N)$ is the total saturation throughput of a system of n saturated users (as defined and calculated in [1]), do not see queue build-up. Similar results were also reported in [6].

This observation seems to suggest that we may hope to achieve higher than saturation through and seems to contradict the theoretical result that a finite user is limited to its saturation throughput. To understand the gap between theoretical result and simulation observations in Figure 1, we consider the state of a symmetric N-user system, which includes the queue backlogs, $\underline{q} \in \mathbb{Z}^{\mathbb{N}}$. Note that in an 802.11 network, as some queues become empty, they stop competing for the channel, allowing other users to adaptively choose smaller average back-off windows. This indicates a strong negative drift on subspaces of the form $S_0 = \{\underline{q} : \exists i \text{ s.t. } q_i = 0\}$, where q_i is the backlog state of queue i. When $\lambda > \lambda_{\text{sat}}(N)/N$, there arises a *broken ergodicity* phenomenon: although the system consists of all transient states, the time averages computed over a simulation run time are significantly different from the ensemble averages. More precisely, there are effectively two classes of states, those with negative drift, denoted by S^- ($S_0 \subset S^-$), and those with positive one, denoted by S^+. If the total arrival rate is greater than λ_{sat}, all states are transient, even though, when starting at all empty queues (in S^-), the time to first hit one of the positive drift states can be significantly longer (orders of magnitude) than the simulation time. In other words, for realistic simulation times the ensemble average can be significantly different from the computed time average. To compute the ensemble average, a randomization of initial queue state is required.

While the above phenomenon and explanation seem nothing more than a lessen in how we should run simulation and read simulation results, they do lead us to think differently about such a system. One can essentially view the system as a globally unstable one with a desirable (higher throughput) equilibrium which is *practically* "robust". This "robustness" depends on the statistics of a stopping time associated with the first transition from S^- to S^+, which by itself heavily depends on the statistics of the arrival processes (beyond the average mean). This brings us to the question as to whether in a real engineered system, a higher rate of arrival can be sustained if all queues are periodically emptied. Or more generally, what are the mechanisms, if any, to stabilize more desirable (higher throughput) equilibria.

In the next two sections we consider two scenarios which produce higher than saturation throughput.

III. An Asymmetric Scenario

The first case we consider is an asymmetric scenario with a mixture of saturated/infinite sources and non-saturated/finite sources. The purpose is to observe the throughput or bandwidth distribution among these queues as well as the total system throughput achieved.

In what follows we will first present a model that attempts to describe a finite source model (as an extension to the Markov chain model for an infinite source given by Bianchi in [1]). We then use this model along with Bianchi's model to derive a fixed point that represents the steady state transmission attempt probabilities and transmission success probabilities of both the finite and infinite queues, respectively. These quantities are then used to calculate the throughput of both types of queues. We begin by illustrating the main idea of this computation, which is based on the fixed point approach as was used in [4].

A. A Fixed Point Approach

Specifically, consider a system consisting of N_s identical saturated sources and N_f identical non-saturated sources. Denote by p_1 and p_2 the probability of experiencing a packet collision by the saturated and the finite sources, respectively, and denote by τ_1 and τ_2 the probability of attempting a transmission by the saturated and the finite sources, respectively.

We adopt the model by Bianchi in [1] to describe an infinite source, and present (in the next subsection) an extension of this model to describe a finite source. These two models can then be used together to compute the above quantities in steady state. These quantities can be obtained via iterating the following set of fixed point equations:

$$\tau_1 = f_1(p_1) = \frac{2(1-2p_1)}{(1-2p_1)(W+1)+p_1 W(1-(2p_1)^m)} \quad (1)$$

$$p_1 = 1-(1-\tau_1)^{N_s-1}(1-\tau_2)^{N_f} \quad (2)$$

$$\tau_2 = f_2(p_2,\lambda,\alpha,\tau_1,\tau_2) \quad (3)$$

$$p_2 = 1-(1-\tau_1)^{N_s}(1-\tau_2)^{N_f-1}\,. \quad (4)$$

Together these four equations are used to solve four unknowns: p_1, p_2, τ_1, τ_2.

Here $f_1(\cdot)$ in (1) is taken from [1]; it expresses the attempt probability of an infinite source as a function of its experienced collision probability, where W is the maximum window size of the initial backoff stage, and m is the maximum number of allowed backoff stages. $f_2(\cdot,\cdot,\cdot,\cdot,\cdot)$ in (3) expresses the attempt probability of a finite source as a function of other parameters of the system. Here λ and α are tunable parameters associated with the finite source model. This model as well as the derivation of f_2 are described below.

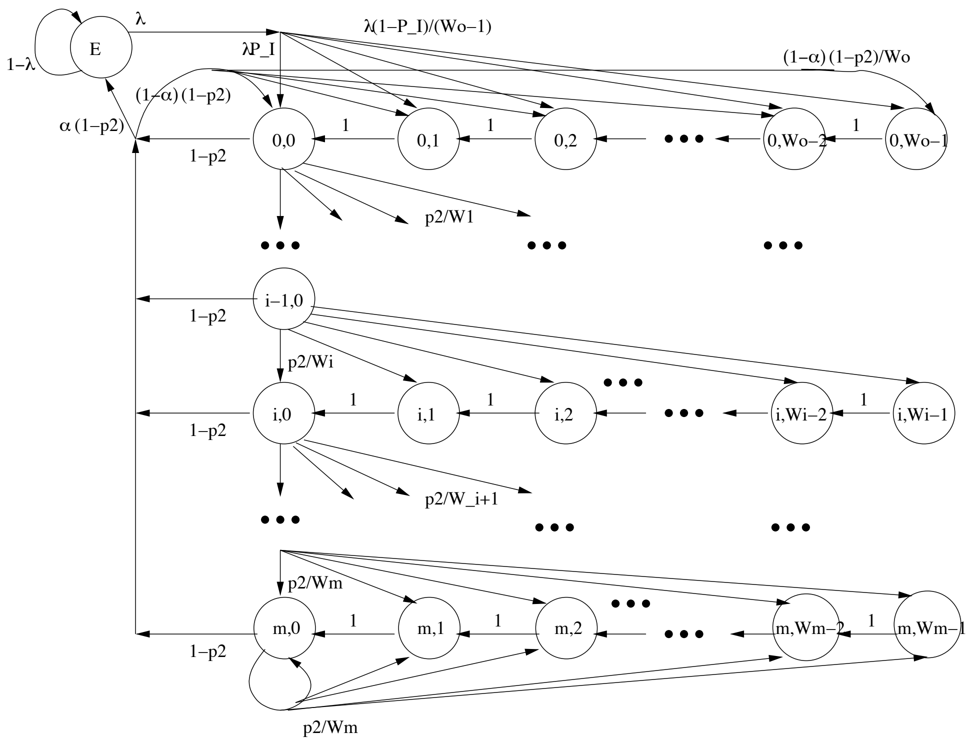

Fig. 2. A finite source model

B. A Finite Source Model

Consider the Markov chain model introduced by Bianchi in [1]. To model a finite source, we will simply add a state E to represent the state when the queue becomes empty. Upon a successful transmission, with probability α the chain enters this state; while in this state, with probability λ an arrival occurs [1]. This model is illustrated in Figure 2. Note that P_I is the probability that the channel is sensed idle by the node in the empty state.

There are a few things worth noting regarding this model. The main motivation for adopting this model is simplicity in representation and analysis as we shall see in the next subsection. There are, however, two meaningful ways of interpreting this model. Firstly, when the extra state denotes the empty state, this model represents a finite source that empties out from time to time. We note that while the couple (λ, α) seems a rather unconvential description of the arrival process, it is possible in some cases to translate an otherwise defined arrival process (e.g., by its average arrival rate) into these two quantities, possibly correlated. In this sense this model represents a finite source that empties out infinitely often. In this paper, we will not further explore the process of such a mapping, and will simply take these two quantities as given. In our numerical experiment, these are chosen to roughly reflect the intensity of the arrival.

Alternatively, this model also describes a queue that is infinite but *transmission-controlled*, in that whenever there is a success, with a fixed probability the queue is forced out of the contention/competition for the channel. We note that this mechanism may have similar effect on the system throughput, for properly selected λ and α, as adjusting the maximum window size in the original backoff scheme. That is, by properly tune these two parameters, one can maximize the system throughput for a given window backoff scheme (with fixed initial window size, maximum window size, etc.). This however is out of the scope of this paper and will be examined in a future study. In what follows we will use this model to represent a finite source with given λ and α.

[1]Note that a similar idea was used in [7], where two extra states were introduced.

C. Deriving the Steady State

We summarize all the transition probabilities in the finite source model described above as follows. We have states denoted by the couple (i,k), where $i=0,1,\cdots,m$ denotes backoff stage, $k=0,1,\cdots,W_i-1$ denotes the backoff timer count-down, and W_i is the maximum window size of the i-th stage. Lastly, E denotes the empty state.

1. $p(i,k|i,k+1)=1$, for $k=0,\cdots,W_i-2, i=0,\cdots,m$.
This is simply counting down the backoff timer.
2. $p(E|i,o)=\alpha(1-p_2)$, for $i=0,\cdots,m$.
This is the probability of transitioning to the empty state upon a successful transmission.
3. $p(0,k|i,0)=\frac{(1-p_2)(1-\alpha)}{W_o}$, for $k=0,\cdots,W_o-1, i=0,\cdots,m$.
This says that upon a transmission, there is probability $(1-p_2)(1-\alpha)$ that the transmission is successful *and* the queue is not empty. Therefore with equal probability the queue enters any of the backoff state in the first stage.
4. $p(i,k|i-1,0)=p_2/W_i$, for $k=0,\cdots,W_i-1, i=1,\cdots,m$.
This is the probability of having a collision and entering another backoff state.
5. $p(m,k|m,0)=p_1/W_m$, for $k=0,\cdots,W_m-1$.
This is the probability of staying in the maximum backoff stage upon collision.
6. $p(E|E)=1-\lambda$. This is the probability of remaining in the empty state.
7. $p(0,0|E)=\lambda P_I$.
Here P_I is the probability that there is an arrival and the channel is sensed idle given that this queue is the empty state; this will be calculated as a function of τ_1 and τ_2 below.
8. $p(0,k|E)=\frac{\lambda(1-P_I)}{W_o-1}$, for $k=1,\cdots,W_o-1$.
When the channel is sensed busy upon an arrival, the queue enters a backoff state with equal probability.

With these transition probabilities, we can proceed to calculate τ_2. We begin by expressing all state probabilities in terms of state $(0,0)$. Denoting the steady state probability of state s by b_s, we have

$$\begin{aligned} b_{i,0} &= p_2 b_{i-1,0}, \quad i=1,\cdots,m-1 && (5)\\ \rightarrow \quad b_{i,0} &= p_2^i b_{0,0}, \quad i=1,\cdots,m-1 && (6)\\ b_{m-1,0}\cdot p_2 &= b_{m,0}\cdot(1-p_2) && (7)\\ \rightarrow \quad b_{m,0} &= \frac{p_2^m}{1-p_2} b_{0,0} && (8)\\ b_E &= \sum_{i=0}^{m} b_{i,0}\frac{\alpha(1-p_2)}{\lambda} = \frac{\alpha}{\lambda} b_{0,0} && (9)\end{aligned}$$

Next, we have

$$\begin{aligned} b_{i,k} &= \frac{(W_i-k)p_2}{W_i} b_{i-1,0}, \quad i=1,\cdots,m && (10)\\ b_{0,k} &= (W_0-k)\cdot\left(\frac{\lambda(1-P_I)b_E}{W_0-1}+(1-p_2)(1-\alpha)\sum_{j=1}^{m} b_{j,0}\right) \\ &= \frac{W_0-k}{W_0}(\beta\cdot\alpha+1-\alpha)b_{0,0}, \ \ k=0,\cdots,W_0-1 && (11)\end{aligned}$$

where $\beta=\frac{(1-P_I)W_o}{W_o-1}$.

Using total probability, we have

$$\begin{aligned}
1 &= \sum_{i=0}^{m}\sum_{k=0}^{W_i-1} b_{i,k} + b_E \\
&= \sum_{i=1}^{m} b_{i,0}\sum_{k=0}^{W_i-1}\frac{W_i-k}{W_i} + \sum_{k=0}^{W_0-1} b_{0,k} + b_E \\
&= \sum_{i=1}^{m} b_{i,0}\frac{W_i+1}{2} + \sum_{k=0}^{W_0-1}\frac{W_0-k}{W_0}(\beta\alpha+1-\alpha)b_{0,0} + b_E \\
&= \sum_{i=1}^{m} b_{i,0}\frac{W_i+1}{2} + \frac{W_0+1}{2}(\beta\alpha+1-\alpha)b_{0,0} + \frac{\alpha}{\lambda}b_{0,0} \\
&= \sum_{i=1}^{m-1} p_2^i b_{0,0}\frac{2^iW+1}{2} + \frac{p_2^m}{1-p_2}b_{0,0}\frac{2^mW+1}{2} + \frac{W+1}{2}(\beta\alpha+1-\alpha)b_{0,0} + \frac{\alpha}{\lambda}b_{0,0}
\end{aligned}$$

where $W = W_0$ is the initial (first stage) window size.

From the above, we derive the following:

$$b_{0,0} = \frac{2(1-2p_2)(1-p_2)}{(1-2p_2)(W+1)(1-\alpha(1-p_2)(1-\beta)) + p_2W(1-(2p_2)^m) + 2\alpha(1-2p_2)(1-p_2)/\lambda}$$

and

$$\begin{aligned}
\tau_2 &= \sum_{i=0}^{m} b_{i,0} = \frac{b_{0,0}}{1-p_2} \\
&= \frac{2(1-2p_2)}{(1-2p_2)(W+1)(1-\alpha(1-p_2)(1-\beta)) + p_2W(1-(2p_2)^m) + 2\alpha(1-2p_2)(1-p_2)/\lambda}
\end{aligned}$$

The above gives us f_2 as a function of λ, α, p_2 and β, which in turns is a function of P_I. P_I can be obtained as a function of τ_1 and τ_2, as shown below.

Following methods using in [1], let P_t denote the probability that there is at least 1 transmission in a slot, P_s^1 the probability that a successful transmission is from a saturated user, P_s^2 the probability that a successful transmission is from a finite user, δ the propagation delay, σ the duration of an empty slot, T_s the average time the channel is sensed busy due to a successful transmission, and T_c the average time the channel is sensed busy due to a collision. Then

$$P_t = 1-(1-\tau_1)^{N_s}(1-\tau_2)^{N_f} \quad (12)$$

$$P_s^1 = \frac{N_s\tau_1(1-\tau_1)^{N_s-1}(1-\tau_2)^{N_f}}{P_t} \quad (13)$$

$$P_s^2 = \frac{N_f\tau_2(1-\tau_2)^{N_f-1}(1-\tau_1)^{N_s}}{P_t} \quad (14)$$

$$T_s = H + EP + SIFS + \delta + ACK + DIFS + \delta \quad (15)$$

$$T_c = H + EP + DIFS + \delta \quad (16)$$

$$P_I' = \frac{(1-P_t)\sigma}{(1-P_t)\sigma + P_t(P_s^1+P_s^2)T_s + P_t(1-P_s^1-P_s^2)Tc} \quad (17)$$

$$P_I = \frac{P_I'}{b_E} = \frac{\lambda P_I'}{\alpha b_{0,0}} = \frac{\lambda P_I'}{\alpha(1-p_2)\tau_2} \quad (18)$$

where P_I' is the unconditioned probability that the channel is idle. H, EP, ACK, $SIFS$, $DIFS$ denotes the packet header, average packet payload size, size of the ACK packet, the durations of the intervals SIFS and DIFS, respectively,

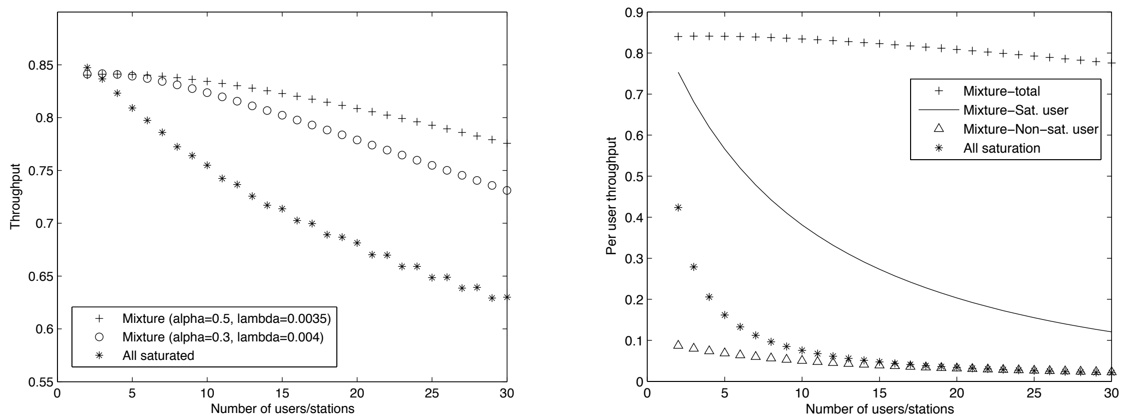

Fig. 3. Throughput comparison of a all-saturated system and a mixed system: Total system throughput (left) and per-user throughput (right).

all measured in time (assuming a certain channel bit rate). The above calculates P_I as a function of τ_1 and τ_2. This completes the derivation of $\tau_2 = f_2(p_2, \lambda, \alpha, \tau_1, \tau_2)$.

D. *Numerical Results*

Using results derived above, we obtain the throughput of both types of queues as

$$S_1 = \frac{P_s^1 P_t EP}{(1-P_t)\sigma + P_t(P_s^1 + P_s^2)T_s + P_t(1 - P_s^1 - P_s^2)T_c} \tag{19}$$

$$S_2 = \frac{P_s^2 P_t EP}{(1-P_t)\sigma + P_t(P_s^1 + P_s^2)T_s + P_t(1 - P_s^1 - P_s^2)T_c} \tag{20}$$

Below we show some numerical results using the above calculation. All parameter values are the same as used in our simulations and can be found in [8]. The computation of our finite source model is partially validated in that when we set $\alpha = 0$ we get exactly the same result obtained in [1] for an infinite source. Numerical values for the slot time σ, DIFS, SIFS, packet header H and ACK size are given earlier in Section II. In addition, the channel bit rate is assumed to be 1Mbps, the propagation delay δ is 1μs, and packets of equal size (payload) $EP = 8184$ bits.

We fix the total number of users in the system to be 30 while varying the number of finite/non-saturated users from 1 to 20. We compare both the total system throughput and the per user throughput achieved in these scenarios with that achieved in a homogeneous scenario where all users are saturated.

Figure 3(left) shows the total system throughput under this asymmetric scenario compared to the symmetric scenario where all users are saturated. In the asymmetric case we fix the number of saturated users at 1, and increase the non-saturated users from 1 to 29. In the symmetric case we increase the number of saturated users from 2 to 30. As we can see the introduction of asymmetry results in increased total system throughput. The parameter λ used for a finite source roughly translates into 0.6 Mbps arrival rate in the empty state, which is about the same (slightly below) as the saturation throughput of a system of 30 saturated users. The net result of such a mixture of heterogeneous sources is the much increased total throughput. Indeed our finite source model forces a source to empty out, as discussed in the previous section.

We further break this down into the per-user throughput for the saturated and the non-saturated users, respectively, for the first set of parameters shown in Figure 3(left). This is compared with the per-user throughput of the all-saturated system. This comparison is shown in Figure 3(right). Not surprisingly, we see that in the non-homogeneous system the increase in total throughput is almost entirely due to the fact that the saturated queue is able to "grab" the unused

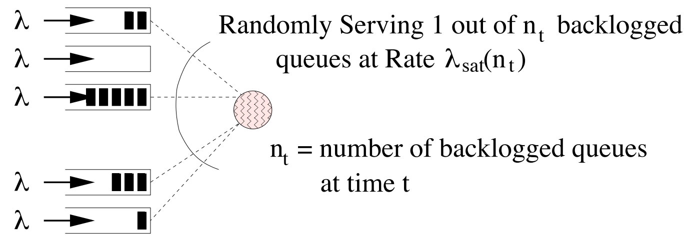

Fig. 4. Rate of departure μ (both from analysis as well as OPNET simulations) as a function of arrival rate

bandwidth from the non-saturated users. However, what is interesting is that the throughput of a non-saturated user in this system is not below that of a saturated user in the homogeneous system when the total number of users exceeds 17.

Thus one may imagine having a system of N heavily loaded users (or saturated) where at any time only one user is allowed to be aggressive and "grab" the unused bandwidth while others are forced to the state E. With the above result, as we rotate the role of this aggressive user, over time all users will achieve an average throughput higher than $\lambda_{\text{sat}}(N)/N$. This of course is nothing but the TDMA type of channel access based on time-share, which requires some type of central control.

This case study demonstrates that if natural asymmetry or heterogeneity exists in a practical setting, then saturation throughput alone does not convey the full picture of the capability of the system. In order to see how we might "force" the system into higher throughput without resorting to time-share, in the next section we consider the option of limiting the buffer size.

IV. A Symmetric Scenario

The second case we consider is a symmetric one where each queue has a perfect channel and a finite arrival rate λ, and a finite buffer space. In this study, we take advantage of finite buffer sizes to stabilize the system in order to force a time-share among operating points, most of which provide with a higher throughput. In particular, we consider a special case, where each user's buffer capabilities is limited to one packet at a time, and show an increase in the achievable throughput. We will show that even though we incur a loss when dropping packets due to buffer overflow, we are capable of compensating this loss with an increase in the departure rate!

To analyze the above system, we consider a simplified model of the network. This is a continuous time extension of the models proposed both in [5] and Section VI of [4]. We assume the following model. Consider the system model shown in Figure 4. There are N identical nodes with buffers of size B. Arrival rate to each node is Poisson with rate λ (if the buffer at a node is full at the arrival time of a new packet, the packet is dropped). Furthermore, the arrivals to nodes are assumed to be independent among nodes. We, here, discard the back-off times and focus only on the times when successful attempts are made, i.e. we assume that when there are k backlogged users a successful attempt leads to the channel being allocated to one of the k contending nodes with equal probability. We also assume that the time between two successful attempts, given k contending nodes, is an exponential random variable with rate $\lambda_{\text{sat}}(k)$. In this paper, we assume that $\lambda_{\text{sat}}(\cdot)$ is a monotonic decreasing function.

Let $\mu(\lambda)$ denote the average departure rate (average throughput). Note that from previous sections we know that if $B = \infty$, then the maximum achievable average throughput is $\lambda_{\text{sat}}(N)$. The following theorem shows that $\mu(\lambda)$ can be strictly greater than $\lambda_{\text{sat}}(N)$, when $B = 1$.

Theorem 1: Assume $B = 1$. There exist a rate of arrival, λ^* such that for $\forall \lambda \geq \lambda^*$ the achievable average throughput $\mu(\lambda)$ is strictly greater than $\lambda_{\text{sat}}(N)$.

Proof: The number of backlogged users in the system forms a birth-death continuous time Markov chain. The state transition diagram for such a system is shown in Figure 5.

Note that the total departure rate in this model is

$$\mu(\lambda) = \sum_{m=1}^{N} \lambda_{\text{sat}}(m)\pi_m.$$

where π_m is the stationary distribution for state m.

On the other hand, from balance equation we arrive at the following:

$$\pi_m = \pi_N \frac{\lambda_{\text{sat}}(m+1)\lambda_{\text{sat}}(m+2)\cdots\lambda_{\text{sat}}(N)}{(n-m)!\lambda^{(n-m)}}. \tag{21}$$

Solving the system of equations (21) together with the fact that $\sum_m \pi_m = 1$, we calculate π_N:

$$\pi_N = \left(1 + \sum_{m=0}^{N-1} \frac{\lambda_{\text{sat}}(m+1)\lambda_{\text{sat}}(m+2)\cdots\lambda_{\text{sat}}(N)}{(n-m)!\lambda^{(n-m)}}\right)^{-1}.$$

In other words, we have

$$\mu(\lambda) = \sum_{m=1}^{N} \lambda_{\text{sat}}(m)\pi_m = \lambda_{\text{sat}}(N)\frac{N!\lambda^N + \sum_{m=1}^{N-1}(N)_m\lambda_{\text{sat}}(m)\lambda^m\Pi_{k=m+1}^{N-1}\lambda_{\text{sat}}(k)}{N!\lambda^N + \sum_{m=1}^{N-1}(N)_m\lambda_{\text{sat}}(N)\lambda^m\Pi_{k=m+1}^{N-1}\lambda_{\text{sat}}(k) + \Pi_{k=1}^{N}\lambda_{\text{sat}}(k)}$$

Here we have skipped the details in the derivation, which can be found in [8].

We first notice that as $\lambda \to \infty$, we have $\mu(\lambda) \to \lambda_{\text{sat}}(N)$. We also notice that due to monotonic decreasing property of $\lambda_{\text{sat}}(\cdot)$, for sufficiently large λ the numerator becomes greater than the denominator, i.e. for large λ we have $\mu(\lambda) > \lambda_{\text{sat}}(N)$, completing the proof. ■

Below we provide numerical results on a network of 50 users. Parameters are the same as in previous sections, and the specific values of $\lambda_{\text{sat}}(\cdot)$ are obtained from the same calculation as in [1]. Figure 6 shows the departure rate $\mu(\cdot)$, given by Equation 22 as well as that obtained from an OPNET simulation when MAC layer buffers are set to 1. Even though, as expected, the simple MAC model proposed earlier fails to capture the precise departure rate, our simulations confirm the general trend specified by Theorem 1. In other words, limiting the buffer size to 1 unit of buffering improves the maximum achievable throughput of the system. Note that such a buffer constraint, on the other hand, causes packet drops in the admissible regime (when $N\lambda < \lambda_{\text{sat}}(N)$), which can be compensated by retransmission. A precise study of this phenomenon is the subject of future studies.

V. Conclusion

In this paper we studied the system throughput achieved in an 802.11 network with fixed number of users. Different from prior work that mostly focuse on the saturation study where all users are assumed to always have packets to send, we consider a finite regime, where some or all users do not always have packets to send. We present two case studies. In the first case we consider a non-homogeneous system where there is a mixture of saturated and non-saturated/finite users, and present a fixed point approach to deriving the throughput of such a system. The second case consists of a system of fixed number of users, each with a buffer of size 1. We present a queueing model to analyze this system. In both cases we show that the total throughput achieved in such a system can be significantly greater than the saturation

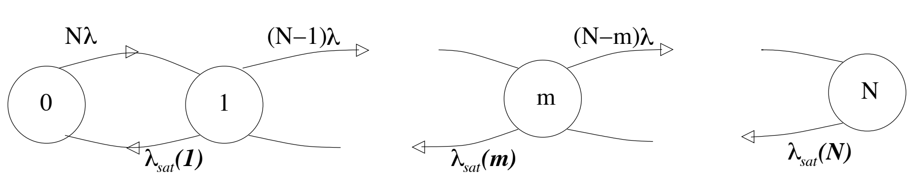

Fig. 5. Continuous time Markov Chain describing the number of backlogged users

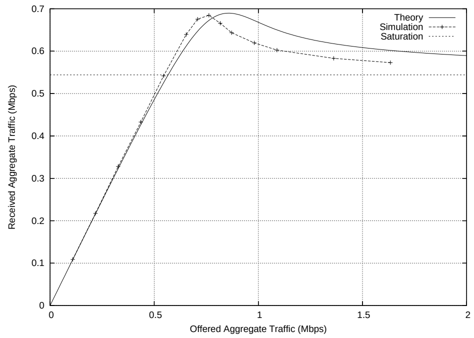

Fig. 6. Rate of departure μ (both from analysis as well as OPNET simulations) as a function of arrival rate

throughput achieved in a homogeneous environment with the same number of users of infinite buffer sizes. With these two case we argue that the notion of saturation throughput is inherently a pessimistic one in that it does not fully convey the capability of the system. In particular, when there is natural heterogeneity in the system or when queues are forced to be empty from time to time due to limited buffer sizes, the effect of statistical multiplexing emerges, which leads to increased system throughput. The particular mechanisms to optimally take advantage of these facts remains and is the subject of future studies.

REFERENCES

[1] G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[2] F. Cali, M. Conti, and E. Gregori, “Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit,” *IEEE/ACM Trans. Networking*, vol. 8, no. 6, pp. 785–799, Dec. 2000.

[3] P. Chatzimisios, A. C. Boucouvalas, and V. Vitsas, “IEEE 802.11 packet delay-a finite retry limit analysis,” in *Global Telecommunications Conference, 2003. GLOBECOM ’03. IEEE*, vol. 2, 2003, pp. 950–954 Vol.2.

[4] A. Kumar, M. Goyal, E. Altman, and D. Miorandi, “New insights from a fixed point analysis of single cell ieee 802.11 wlans,” in *Proc. of IEEE INFOCOM’05*, Miami, FL, 2005.

[5] R. Vijayakumar, T. Javidi, and M. Liu, “From saturation to non-saturation: A study on 802.11 networks,” CSPL Technical Report Series: TR-363. [Online]. Available: http://www.eecs.umich.edu/systems/TechReportList.html

[6] H. Zhai, Y. Kwon, and Y. Fang, “Performance analysis of IEEE 802.11 MAC protocols in wireless LANs,” *WIRELESS COMMUNICATIONS AND MOBILE COMPUTING*, 2004.

[7] M. Ergen and P. Varaiya, “Throughput analysis and admission control of IEEE 802.11,” *Mobile Networks and Applications*, 2005.

[8] T. Javidi, M. Liu, and R. Vijayakumar, “Saturation throughput of 802.1 revisited,” CSPL Technical Report Series: TR-371. [Online]. Available: http://www.eecs.umich.edu/systems/TechReportList.html