From Saturation to Non-Saturation: A Study on 802.11 Networks

Rajiv Vijayakumar, Tara Javidi and Mingyan Liu

***Abstract*— There have been extensive studies on the performance and behavior of 802.11 networks. These studies primarily focus on the so-called *saturation* regime, where all clients/user queues always have packets to send. A prominent study is the work by Bianchi [1], as well as numerous subsequent studies. This saturation operating regime may be viewed as the limiting case when all the arrival rates in the system approach infinity. This paper departs from this class of study in that we focus on a *non-saturation* regime, where user queues are finite with arrival rates below the saturation level. We demonstrate that the widely studied saturation throughput is inherently a pessimistic notion and deserves better understanding. We provide examples of systems of users with finite arrival rates where the total throughput of the system is significantly greater than the saturation throughput. Furthermore, we study the throughput and delay performance of an 802.11 network in a non-saturation scenario, and show that for arrival rates very close to the saturation throughput, the system behavior is qualitatively very different from that in the saturation case. We attempt to explain these observations by constructing and analyzing a number of both realistic and idealized MAC schemes. We also argue that under realistic channel models and user asymmetry, the conventional measures of fairness (defined in terms of throughput) are distinct from fairness in terms of delay.**

***Index Terms*— 802.11, saturation throughput, non-saturation, delay, fairness, simulation**

I. Introduction

With the wide deployment of wireless LANs, its core enabling technology the IEEE 802.11 medium access control (MAC) protocol has been very extensively and intensively studied in recent years. These studies have focused on throughput, delay, and fairness properties of the 802.11 MAC.

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, and 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. Bianchi 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. This model has been built upon and refined by many authors; for instance, Chatzimisios et al. [3] built upon the model to compute delays in the saturation case. A different approach to saturation throughput was followed by Cali et al. [2] who used a p-persistent CSMA model to study saturation throughput. More recently, in [4], Kumar et al. further studied the saturation throughput as a fixed point problem and examined its existence and uniqueness.

By contrast, in this paper we consider a very different system operating regime, where user queues are fed with finite arrival rates and they do not always have a packet to send. We will refer to these types of queues as *finite sources* to distinguish them from the infinite sources/queues used in saturation studies. Our interest is in studying how the system behaves when the arrival rates approach the saturation rate (throughput) *from below*. Since the system may become unstable once the arrival rate exceeds the saturation throughput, our study here in a way concentrates on a critically loaded system where the system remains stable but the arrival rate is very close to the saturation point (e.g., 99.99% of the saturation throughput). For this reason, we will refer to our study in this paper as a *non-saturation* study to distinguish it from saturation studies.

In particular, we examine the following interrelated questions:

- How does the average delay increase when the arrival rate increases to the saturation level?
- When there is a mixture of infinite and finite sources (with arrival rates very close to saturation), how is the bandwidth shared and traded off among different users?
- Does the notion of fairness change in terms of delay as opposed to fairness in terms of throughput?

We show that with the finite source model, the queues exhibit very different qualitative and quantitative behavior compared to that observed under saturation. Whereas saturation studies use infinite sources to induce a saturation throughput, our study examines finite sources with arrival rates approaching the saturation throughput. In the latter there exists an interaction between traffic arrival (which can be bursty) and channel access, which is absent when all queues are infinite sources. In this sense our study complements saturation studies in revealing different aspects of the system and points to a better understanding of the notion of saturation throughput. To the best of our knowledge, this is the first comprehensive study of non-saturation behavior (particularly in terms of delay) of

R. Vijayakumar is with the Department of Electrical Engineering at the University of Washington, Seattle, rajiv@ee.washington.edu. T. Javidi is with the Department of Electrical and Computer Engineering at the University of California, San Diego, tara@ece.ucsd.edu. M. Liu is with the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, mingyan@eecs.umich.edu.

an 802.11 system.

Our main results and observations are summarized as follows.

1) The queues have very good delay performance even when all queues have arrival rates approaching the saturation throughput. This is shown to be closely connected to the fact that there are very few backlogged queues on average under the same scenario. It is generally held that random access techniques sacrifice aggregate throughput to achieve desirable delay performance. Our study strengthens this result in that it shows that the average delay in 802.11 networks remains extremely low even at arrival rates as high as 99.99% of the saturation throughput. More interestingly, there seems to be a “phase transition” in the delay performance as the arrival rate approaches the saturation throughput. This observation suggests that the notion of saturation needs to be better understood within the context of delay.
2) The achievable throughput of a single infinite source queue when the rest are finite source queues (but with arrival rates approaching their respective saturation throughput) significantly exceeds that achieved when all queues are infinite sources. This is an example where the total throughput of the system is significantly greater than the total saturation throughput achieved by the same number of users. In particular, the unused bandwidth from low rate users is efficiently used to improve the performance of high rate users. This observation suggests that saturation throughput is inherently a pessimistic notion, and does not fully reflect the capability of a realistic system with user asymmetry and statistical multiplexing.
3) Using similar scenarios with a mixture of finite sources with fixed arrival rate and aggressive sources with rates increasing beyond saturation level, we show that the increase in delay for a finite source caused by an aggressive source is bounded. Furthermore, the delay increase experienced by an aggressive user as it increases its rate is far greater than that experienced by a finite source with fixed rate. This suggests that the commonly used measures of fairness in terms of throughput have very different implications in terms of delay. In particular, 802.11 appears to provide little incentive (in terms of delay) for a user to increase its demand. Using this observation we also show how to effectively handle the now well-known 802.11 anomaly problem when users have different data rates [5].
4) We construct a number of MAC schemes, both practical and idealized, in an attempt to interpret the above observations and to identify a tractable system that shares key features with 802.11, and gives rise to such observations. We show that the low delay observed in a critically loaded 802.11 network is closely related to the fact that the throughput increases with a decrease in the number of backlogged users, causing a strong negative drift in the number backlogged. This is in drastic contrast with systems like Aloha, where the throughput decreases with a decrease in the number of backlogged users.

The rest of the paper is organized as follows. In Section II we describe the network scenarios studied as well as our simulation methodology. Section III presents detailed results on the delay behavior of the system operating below (and approaching) saturation with symmetric users. Section IV studies an asymmetric scenario where there is a mixture of finite and infinite sources. In Section V we present a number of MAC schemes in an attempt to interpret results presented in previous sections. Section VI discusses how results obtained in previous sections can be used to address the 802.11 anomaly problem due to varying channel quality. Section VII concludes the paper.

II. Network Model and Simulation Methodology

All simulations were run using Opnet (Release 11) and its built-in 802.11b model. Except for certain special cases considered in Section VI, the following holds for all simulations: The physical layer rate used was 11Mbps. 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 I; note that the physical layer headers and ACKs were always transmitted at 1Mbps. 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 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. For more detailed simulation results, including plots showing confidence intervals, see [6].

slot time	20 μs
SIFS	10 μs
DIFS	50 μs
CW$_{\text{min}}$	31
CW$_{\text{max}}$	1023
Physical layer headers	192 bits
MAC layer headers	224 bits
ACK frame	304 bits
Retry Limit	7

TABLE I

802.11b Parameters

A. Saturation Throughput

For the values of the system parameters that we use, the transmission time for a single packet (including the physical and MAC layer headers) is 0.957ms; if we add in the interframe spaces and ACK durations, the minimum time occupied by a single transmission is 1.321ms. If there is a single client in the system the mean time spent in backing off between successive transmissions is 0.31ms (15.5 slots); the bit rate obtained by a single saturated client is therefore $8192/(1.321 + 0.31) = 5.02$ Mbps.

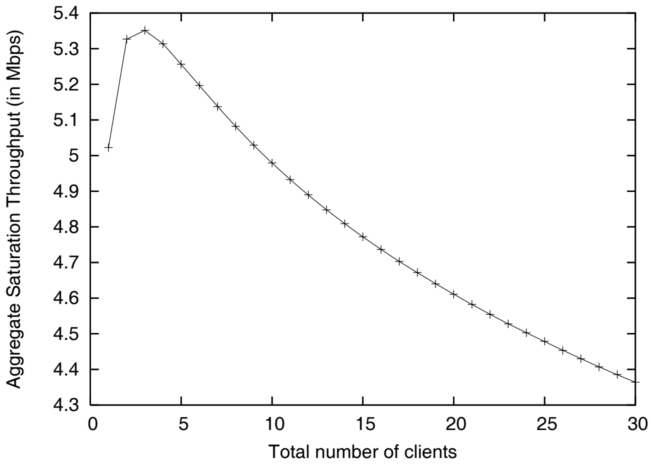

Fig. 1. Saturation throughput as a function of the total number of clients in the system

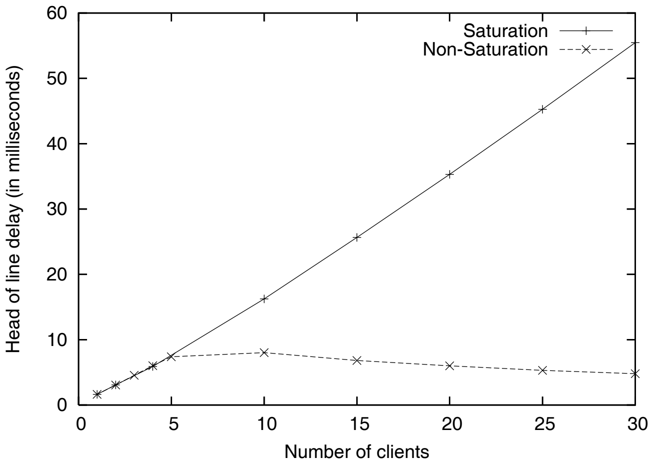

Fig. 2. HOL delay in the saturation and non-saturation cases as a function of the total number of clients in the system. In the non-saturation case the throughput of the clients was 99.99% of saturation throughput.

Figure 1 plots the aggregate saturation throughput (i.e. the total number of bits received successfully per unit time across all clients) as a function of the total number of clients in the system. The aggregate throughput initially increases from 5.02 Mbps to 5.35 Mbps as the number of clients increases from 1 to 3, but then decreases as the number of clients increases beyond 3. The initial increase is attributable to the corresponding decrease in time spent in the backing off between successive transmissions. Similar saturation throughput curves for 802.11b at 11Mbps have been presented by various authors (e.g. [7]). We use the notation $\lambda_{\text{max}}(n)$ for the saturation throughput of a single client when there are n clients in all in the system. In this notation, Figure 1 is a plot of $n\lambda_{\text{max}}(n)$ against n.

Note that the departure rate from the queue of a saturated client exceeds its saturation throughput by a small amount because the 802.11 MAC drops packets which exceed their retry limit. For example, with 20 clients, the measured saturation throughput was 4.611 Mbps (with a 95% confidence interval of $\pm 0.0066\%$), whereas the departure rate from all the queues was 4.619 Mbps. In cases where we are interested in driving a client's queue near saturation, we use the departure rate at saturation instead of saturation throughput as our reference point. However for simplicity of exposition, we use the term "saturation throughput" for both quantities.

III. Symmetric Users

In this section we study a symmetric scenario where all clients in the system have the same arrival rates. Specifically, we consider a system with n users each with arrival rate $\lambda < \lambda_{max}(n)$, where $\lambda_{max}(n)$ is the saturation throughput as defined in the previous section. As mentioned in the introduction, the case with arrival rates strictly below $\lambda_{max}(n)$ will be referred to as the *non-saturation* scenario to be distinguished from the *saturation* scenario where queues are always full and each client gets a throughput of $\lambda_{max}(n)$. The goal of this section is to study the delay behavior of these queues in the non-saturation scenario when the arrival rates approach the saturation rate.

We consider two types of delays. The first is the head-of-line (HOL) packet delay, defined as the time from when a packet first reaches the head of its queue until it is either successfully transmitted or dropped. The second is the end-to-end (E2E) delay, defined as the time from when a packet enters a queue until it is successfully transmitted (E2E delay is not defined for dropped packets). Therefore the end-to-end delay is the sum of the head-of-line delay and the queueing delay. Note that in a saturated system the only meaningful notion of delay is HOL delay, whereas both HOL and E2E delays can be defined and measured for a non-saturated system.

We first contrast HOL delay for the saturated case with HOL delay for the symmetric non-saturated case where the arrival rates λ equal 99.99% of $\lambda_{max}(n)$, i.e. the arrival rates are very close to the saturation throughput. The HOL delays for both these systems are plotted in Figure 2 for different values of the total number of clients in the system n. To place the delay values in context, note that since the transmission time for each packet is itself about 1 ms (1KB @ 11Mbps), the minimum possible delay is about 1ms.

The difference between the two delay curves in Figure 2 is striking. The delay in the saturation case increases linearly with the number of clients n, but the delay in the non-saturation case varies relatively little, and in fact actually drops as n grows large; consequently the non-saturation delay is significantly smaller than the saturation delay for larger n.

Figure 2 shows that the delays in the non-saturation case stay low even when the arrival rate is close to the saturation throughput. For lower arrival rates, the delays are correspondingly lower, as illustrated in Figure 3 where we plot the HOL and E2E delays in the symmetric non-saturation case for a system with 20 clients as the arrival rates increase from low values up to the saturation throughput. The aggregate saturation throughput is marked at $20 \cdot \lambda_{\text{max}}(20) = 4.61$ Mbps.

We see here that even at very high loads (e.g., 4.2 Mbps) both delays are only about 4 or 5 packet transmission times. The closeness of the two curves also shows that there is very little queueing until we reach close to saturation.

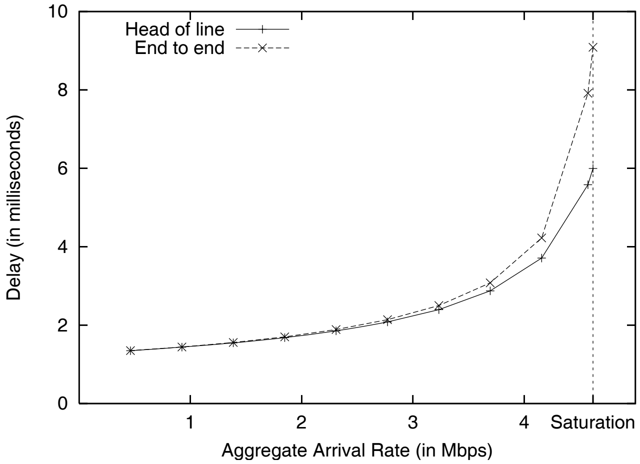

Fig. 3. Head of line and end to end delay as functions of the aggregate arrival rate to a system of 20 clients. The aggregate saturation throughput is 4.61Mbps.

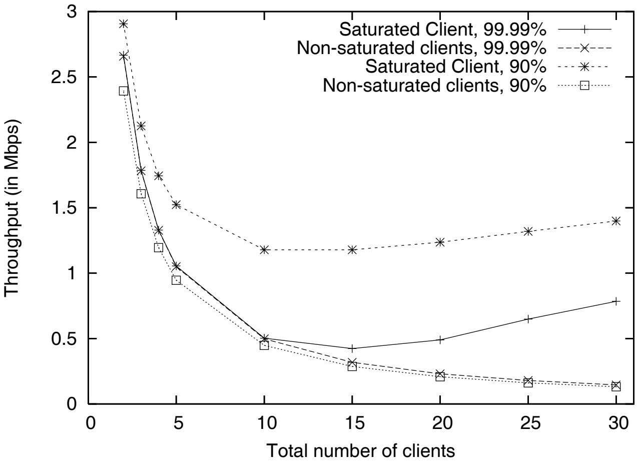

Fig. 4. Individual throughputs of the saturated client and the $n-1$ finite source clients as a function of the total number of clients n. Throughputs are shown for two values of the arrival rates of the finite source clients, corresponding to 90% and 99.99% of $\lambda_{\text{max}}(n)$.

Figures 2 and 3 show that an 802.11 system with symmetric users and arrival rates very close to the saturation rate has very interesting delay behavior. By definition, the saturation rate $\lambda_{max}(n)$ is the highest throughput that a queue can obtain in a system of n symmetric, saturated queues. One might therefore expect that delays will grow large as the arrival rates approach $\lambda_{\text{max}}(n)$. However what we see here is that if we stay below this rate, then the queues have very good delay performance even when the arrival rates approach the saturation throughput. In particular, the HOL packet delay remains nearly constant with respect to the number of clients in the system even when operating at arrival rates very close to saturation.

It is generally held that random access techniques sacrifice aggregate throughput in order to achieve desirable delay performance. Our observation here strengthens this belief in that it shows that the average delay in 802.11 networks remains extremely low even at arrival rates as high as 99.99% of the saturation throughput. More interestingly, our observation also appears to suggest a certain *discontinuity* as we go from non-saturation toward saturation. In other words, there seems to be a “phase transition” in the delay performance as the arrival rate approaches the saturation rate. This also suggests that the notion of saturation needs to be better understood within the context of delay. We defer further discussion of this issue to Section V where we tie the observed delay behavior to the number of backlogged users in the system.

In the next section we examine *asymmetric* scenarios where some of the clients have higher arrival rates than the other clients.

IV. Asymmetric Users

In this section we examine throughputs and delays in a system with finite arrival rates and asymmetric users.

A. Increased Aggregate Throughput

Consider a system with n users in all where $n-1$ users have a traffic arrival rate of $(1-\epsilon)\lambda_{\text{max}}(n)$, $0<\epsilon<1$, and the last user is saturated (always backlogged). We are interested in investigating the rate of service obtained by this saturated user. Here we can think of the throughput obtained by the saturated client as the *freed up bandwidth* due to an ϵ drop in arrivals of others. One might expect this freed up bandwidth to be no more than $(n-1)\epsilon\lambda_{\text{max}}(n)$. More generally, one might not expect the saturated user to obtain a throughput significantly more than $\lambda_{\text{max}}(n)$ when the arrival rate of the first $n-1$ user is close to the saturation throughput (i.e. for ϵ close to zero). However, we shall see that the freed up bandwidth can in fact be significantly higher.

Figure 4 shows the throughput obtained by the saturated client and the (per-client) throughput of the finite rate clients as a function of the total number of clients for $\epsilon \in \{0.0001, 0.1\}$. Since the finite rate clients obtain a throughput equal to their arrival rate, their throughput decays with the total number n as the saturation throughput does. But the throughput obtained by the saturated client remains much higher than $\lambda_{\text{max}}(n)$ and actually increases after a point as the total number of clients increases. From the figure we see that, in general, the freed up bandwidth consumed by the saturated user can be much larger than 19 times the throughput each unsaturated user has given up. More importantly, the gap between the consumed bandwidth and the bandwidth given up depends on the total number of users, as well as ϵ.

In Figure 5 we consider the case of $\epsilon = .9$, i.e. when the non-saturated clients have a throughput which is 10% of their saturation throughput. In this figure, notice that the throughput for the saturated user is only slightly less than the total saturation throughput for a one user system. To appreciate this better, imagine a system with one user transmitting or receiving a large volume of data at 4.6 Mbps (which is about 92% of $\lambda_{\text{max}}(1)$). Then what Figure 5 shows is that this exact system can serve an additional 19 low rate users at 23 kbps each, with no significant loss to the first user.

Figure 6 shows the aggregate throughput (i.e. the combined throughput of the saturated client and all finite rate clients)

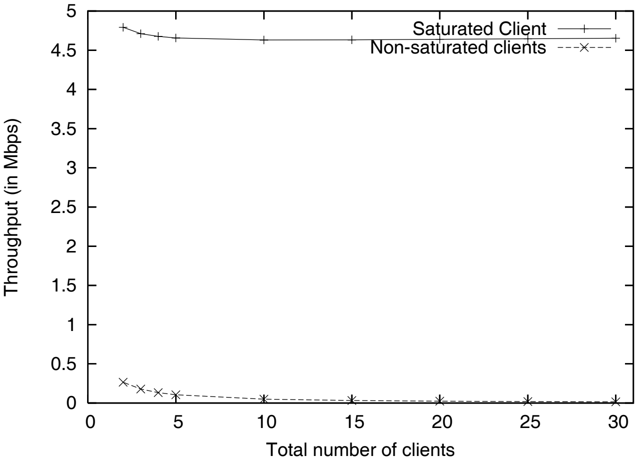

Fig. 5. Individual throughputs of the saturated client and the $n-1$ finite source clients as a function of the total number of clients n. The arrival rate of the finite source clients is $10\%\lambda_{\max}(n)$.

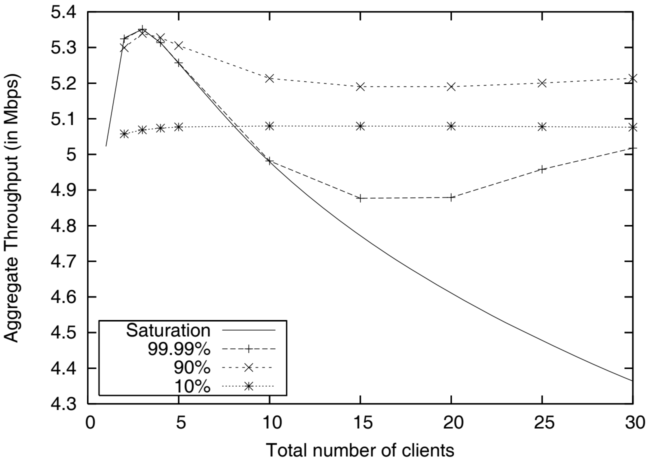

Fig. 6. Aggregate throughput as a function of the total number of clients.Throughputs are shown for three values of the arrival rates of the finite source clients, corresponding to 10%, 90% and 99.99% of $\lambda_{\max}(n)$. The throughput when *all* users are saturated is also shown for comparison.

as a function of the number of clients. Note that there is a significant gain in the total throughput for $n > 10$. This gain in aggregate throughput may have been expected considering the reduction in the rate of arrivals for some clients and the statistical multiplexing nature of the system. Indeed, in any contention based MAC scheme the collisions (or mechanisms devised to avoid it) are the root causes of the inefficiency in bandwidth use. The result here means that the uniform distribution of the load to many users is much less efficient than asymmetric loading of only one user (where the packets do not contend/collide). What is perhaps more surprising is the significance of this increase even when there is no significant reduction in the arrival rates of the finite source queues e.g. when the finite rate clients have throughputs of $99.99\%\lambda_{\max}(n)$.

It is also worth noting that the aggregate throughput when 19 users have throughput demands at 10% of $\lambda_{\max}(20)$ is lower than that in the case of 90% of $\lambda_{\max}(20)$, but higher than that in the case of 99.99% of $\lambda_{\max}(20)$. This shows that the aggregate throughput is not a monotonic increasing function of ϵ, suggesting that as ϵ grows beyond a threshold, the freed up bandwidth remain unused. In other words, no one user can consume too much bandwidth. In reality, however, the number of high-rate users (those willing to consume as much bandwidth as possible, hence acting like a saturated user) would likely be more than one. So we need to investigate the issue of freed-up bandwidth in the presence of multiple saturated users. Figure 7 addresses this question. Here we consider a system with 20 clients in all, and examine the aggregate throughput as the number of these clients who are saturated increases. As seen from this figure, the overall freed up bandwidth that is consumed by the saturated clients is an increasing function of ϵ if at least 3 or 4 users are saturated but decreases as the number of saturated users increases.

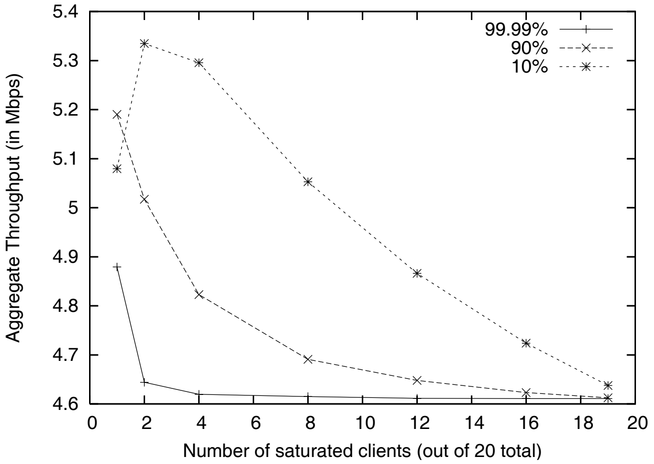

Fig. 7. Aggregate throughput as a function of the number of saturated clients (out of 20). Throughputs are shown for three values of the arrival rates of the finite source clients, corresponding to 10%, 90% and 99.99% of $\lambda_{\max}(n)$.

B. Bounded Delay Degradation

A natural question that follows from the results above is whether the increased throughput for some clients is obtained at the cost of increased delays for other clients. We next examine how the increase in the throughput of one user can impact the delay characteristics of low rate users. To do this, we consider the following scenarios: $n-1$ of the n nodes have a fixed arrival rate of $(1-\epsilon)\lambda_{\max}(n)$, and we increase the arrival rate of the nth user (starting from $(1-\epsilon)\lambda_{\max}(n)$) and observe the delays experienced by the first $n-1$ as well as the last user.

From Figure 4, we see that, for $n = 20$, the saturated node gets a throughput of about 6 times that of the non-saturated clients when the 19 non-saturated clients have arrival rates equal to 90% of their saturation rate (which works out to $\approx$ 200Kbps). To study the impact on delay, we set all 20 nodes to initially have an arrival rate of 200 kbps, and then increase the rate of one of the nodes (which we'll call the "high-rate" node) up to and beyond the 1.23 Mbps limit that we expect to see based on Figure 4.

Figures 8 and 9 plot the delays experienced by high rate and low rate users, respectively, as a function of arrival rate of the high rate user. The two delays are plotted in different figures due to the differing scales on the y-axis. The main observation here is that the high-rate node can obtain a throughput of 2-3 times that of the other nodes without a significant impact on their delays. As the high-rate node approaches saturation (which is at 1.23 Mbps in this case), its own delay as well as other nodes' delay grow significantly; when the high-rate node achieves 1 Mbps (close to saturation), the delay at the other nodes increases by a factor of 3-4 times over the delay in the symmetric case. Note that the delay for the high-rate node is much larger than the delay for the low rate nodes, and that the delay of the low rate clients remains bounded even when the high rate client reaches saturation.

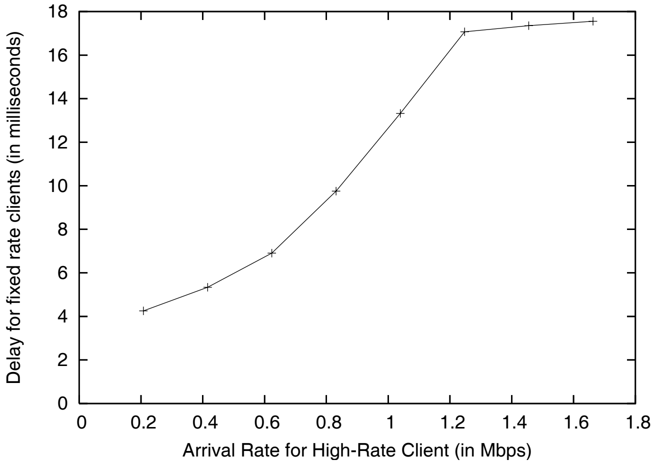

Fig. 8. Delay for the 19 fixed rate clients as the rate of the high-rate client increases. The fixed rate clients throughput is 90% of saturation, which is 0.2 Mbps. The high rate client's throughput saturates at 1.23 Mbps.

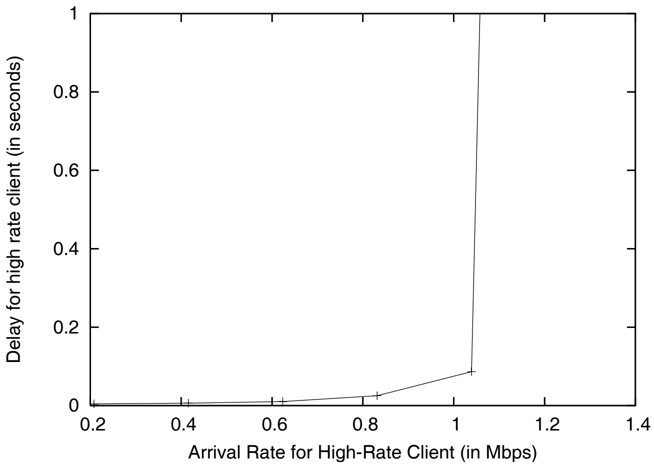

Fig. 9. Delay (in seconds) for the high-rate client as its rate increases. The fixed rate clients throughput is 90% of saturation, which is 0.2 Mbps.

Figure 10 shows the delays for the low-rate clients when the experiment is repeated with the low rate clients' throughput set to 10% of saturation (about 23 kbps). In this case, the delay of the low rate clients only increases by a factor of about two even when the high rate client is driven into saturation and achieves a throughput of 4.85 Mbps.

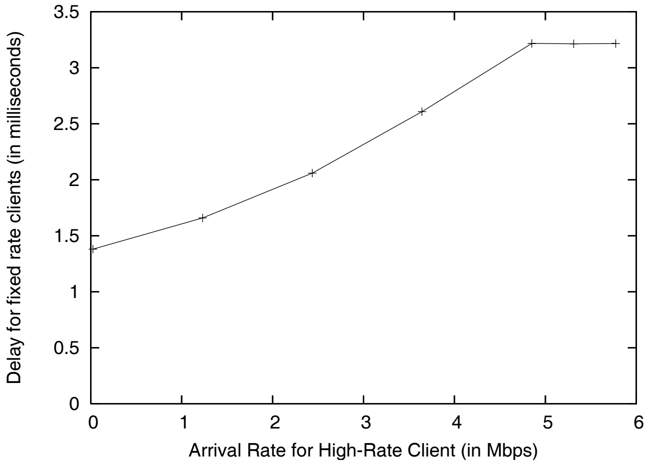

Fig. 10. Delay for the 19 fixed rate clients as the rate of the high-rate client increases. The fixed rate clients throughput is 10% of saturation, which is 0.023 Mbps. The high rate client's throughput saturates at 4.85 Mbps.

Some important conclusions to be drawn from our observations are as follows. When the demand of some users is below the achievable rate, the unused bandwidth is efficiently traded off to improve performance for other users with higher demand. In addition, for the most part, such resource multiplexing causes minimal impact on the service provided to the low demand users. In fact, as the demand of the high rate user increases beyond the admissible traffic pattern, the delay degradation experienced by low rate users remains bounded. More importantly, such performance degradation (increased delay) experienced by low rate users is orders of magnitude less than the degradation experienced by the high rate user. This in turn implies an extremely desirable incentive structure in the following sense: a user has little incentive in increasing its demand, as unacceptably high demand severely decreases his own received quality of service long before it degrades the environment for others.

V. Discussion and Analysis

In this section, we attempt to explain the qualitative differences between the finite source and saturation scenarios. In order to do so, we refer back to Figure 2 showing the HOL delay.

The head of line delay at a given queue is essentially the time between successive departures from that queue and is therefore inversely proportional to the service rate at the queue. In the saturation case, the service rate is the saturation throughput and hence the mean HOL delay can be estimated from the saturation throughput by $E\{T_{HOL}(n)\} =$ packet size$/\lambda_{\max}(n)$. For example, when $n = 20$, $\lambda_{\max}(n)$ is 0.23 Mbps and therefore $E\{T_{HOL}(n)\} = 8192/0.23 \approxeq 35ms$, which is indeed the empirical saturation HOL delay in Figure 2. Since $\lambda_{\max}(n)$ is inversely proportional to n, this explains the linear increase in mean HOL delay versus

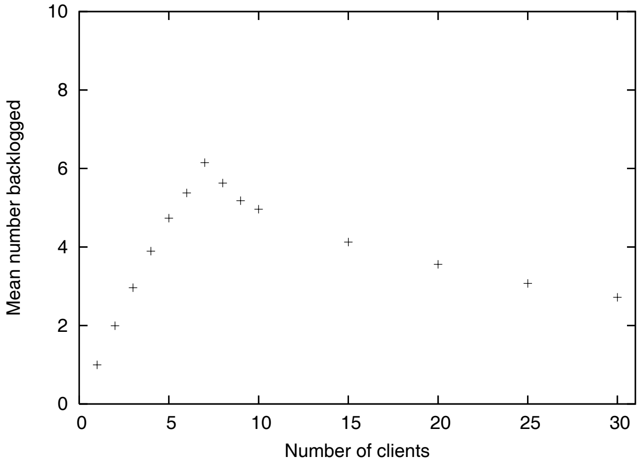

Fig. 11. Mean number of clients backlogged at the start of a cycle ($\lambda = 99.99\%\lambda_{\max}$)

n for the saturation case. The low values of HOL delay in the non-saturation case suggest that in this case, the queues are receiving a higher service rate than in the saturation case. On the other hand, the level of service in 802.11 systems is determined by the level of contention, which in turn depends on the number of backlogged users in the system.

In general, if there are k users who are contending simultaneously for the channel, then there will be on average $k-1$ transmissions interspersed between two successive transmissions of a given queue. In other words, the HOL delay is always proportional to k. Now this combined with the sharp variation in HOL from non-saturation to saturation suggests that the number of backlogged users, k, has a similar sharp transition, i.e. k is kept low even for arrival rates very close to that of saturation. This rather counter-intuitive fact is borne out by Figure 11, which plots the mean number of backlogged clients in the non-saturation case with arrival rates equal to 99.99% of saturation. It shows that the expected number of backlogged users first grows linearly in n, but after $n > 7$ it drops to much lower values than n, and remains steady around 2-4 as n increases. We believe that this unexpected characteristic of 802.11 is essential in understanding the desirable delay properties of the system at moderate loads (remember that all contention-based access schemes sacrifice some throughput).

We believe that the tendency to keep number of contenders at low numbers (such as 4 or 5) is the key mechanism behind the observed behavioral characteristics of 802.11 which we provided in Sections III and IV. In order to fully understand the system behavior, we need to study *why and how* the mean number of backlogged clients in an 802.11 system is kept low even at rates close to the saturation rate.

Our hypothesis is that the number of backlogged users in a multi-queue system would be "stabilized" at lower values if a system has the property that the total rate of service increases as the number of backlogged users decrease (creating a negative drift). The remainder of this section is our attempt to verify this hypothesis by introducing simple queueing models which isolate and capture the above phenomenon. In Section V-A we describe a few contention-based MAC schemes and observe the service rate as a function of k[1]. We observe that the service rate in an 802.11 network does, in fact, drop with the number of backlogged users. Then in Section V-B we provide a simple queuing model which abstracts out the details of each MAC scheme to isolate impact of varying service rate on the mean number of backlogged users. We demonstrate, via the simplified queuing model, the validity of our hypothesis.

One could argue that another (or maybe even more) interesting question is how the service rates in 802.11 are regulated to have the desirable properties discussed. This is especially significant since the access scheme in 802.11 does not explicitly estimate the number of backlogged users at any point. This question remains a topic of future study.

A. Number of Backlogged Users and the Aggregate Service Rates

In this section we attempt to understand the mechanism which enables 802.11 to "stabilize" the number of contenders at and around 2-4. To do so, we refer to Figure 12, plotting the expected number of successful transmissions per unit time as a function of the backlogged users (out of 20 clients) for four multi-access schemes. We refer to this quantity as $p_s(k)$, and it is nothing but the departure rate of the system when there are k backlogged users. The four MAC schemes are as follows:

MAC1 (p-Persistent Slotted Aloha): At each time slot, each user transmits a packet with probability $1/n = 1/20$.

MAC2 (p-Persistent CSMA Slotted Aloha): Each user uses carrier sensing; if idle state is identified by a user, the user attempts to transmit with $p^*_{CSMA}(20)$, a quantity chosen to maximize the saturation throughput of the CSMA system with 20 users [2].

MAC3 (Optimal Slotted Aloha): At each time slot, each user knows exactly the current backlog k and transmits a packet with probability $1/k$. This is an idealized case [8].

MAC4 (802.11): Each user acts like an 802.11 client with parameters given in Section II

Our choices of MACs represent adaptive (3 and 4) as well as non-adaptive (1 and 2) schemes. In addition, schemes 2 and 4 use carrier sensing while 1 and 3 do not. From Figure 12, we make the following observations: The optimal slotted Aloha as well as 802.11 have a similar trait in that, in both systems, the expected number of successfully transmitted packets per unit of time drops as the number of backlogged users increases (for $k > 2$). We hypothesize that *the decreasing property of function $p_s(k)$, in the adaptive MAC schemes MC3 and MC4, is key to keeping the mean number of backlogged users in the system low even at arrival rates close to saturation*. We argue that the negative slope of the graph implies a negative drift in the number of backlogged clients when the arrival rate is kept below the maximum departure rate.

In the next section we try to provide a model to substantiate and verify our hypothesis via study of a queuing system. We

[1] If a MAC is not contention-based, and packet sizes are all equal, then the service rate becomes independent of number of backlogged users.

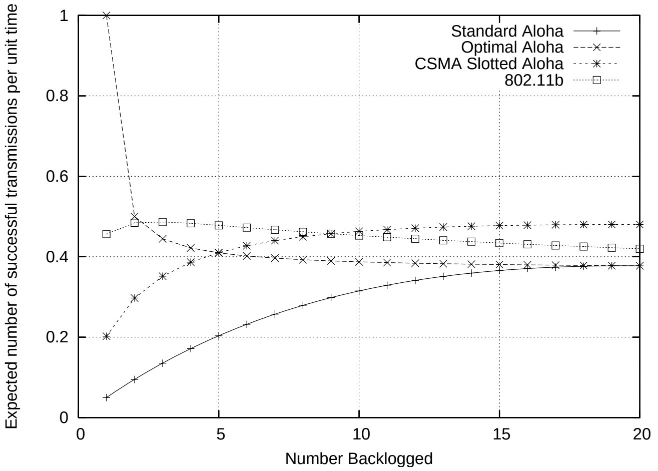

Fig. 12. Expected number of successful transmissions per unit time (throughput) as a function of the number of backlogged clients for various schemes. The total number of clients is 20 clients in each case.

emphasize that the introduced queuing model is not meant for practical MAC design, but rather to serve the purpose of illustrating the above claim.

B. A Multi-Queue Model with Random Service

Consider a slotted Aloha system with n clients in which a backlogged client attempts to transmit a packet in a slot with probability p. In such a system, the probability that there is a successful transmission in a slot in which there are a total of k backlogged clients is $p_s(k) = kp(1-p)^{k-1}$. Further, if there is a successful transmission then it is equally likely to have come from any of the k backlogged clients. In what follows, we use these two characteristics to provide a general multi-queue model with a single server whose availability to serve is random and state dependent. We use such a model to investigate the impact of function $p_s(k)$ on the delay performance of MAC schemes we studied in Section V.

Consider a single server system consisting of n queues buffering jobs of equal length. We assume a finite arrival rate λ and we assume that given the server is allocated to a queue, one job is served (departs) from that queue. The main difference of our model with a regular single server, multi-queue system is that we allow for a random and state-dependent availability for server, i.e. a server is available with probability $0 \leq p_s(k) \leq 1$, where k is the number of backlogged clients. Now we use each curve given in Figure 12 to obtain what we call queue models 1 to 4 (QM1-QM4), corresponding to MAC schemes 1-4 introduced in Section V-A:

QM1 This is the queueing model corresponding to p-Persistent Slotted Aloha, i.e. $p_s(k)$ is chosen from the very bottom curve in Figure 12

QM2 This is the queueing model corresponding to p-Persistent CSMA Slotted Aloha, i.e. $p_s(k)$ is chosen from the CSMA curve (with the highest throughput) in Figure 12

QM3 This is the queue model corresponding to Optimal Slotted Aloha and the construction similarly uses $p_s(k)$ from Figure 12

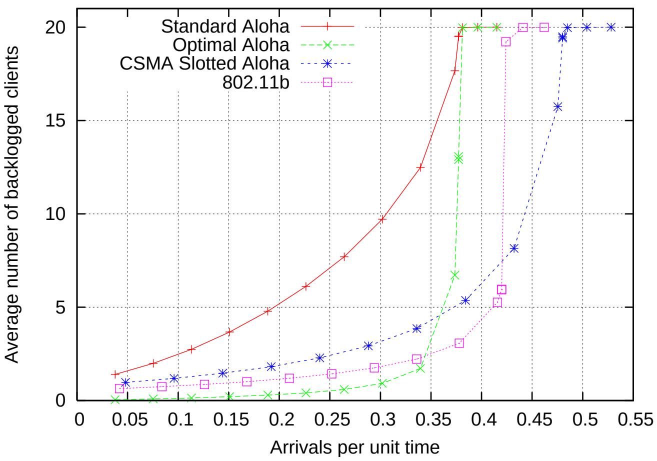

Fig. 13. Mean number of backlogged clients (out of 20) versus load for systems QM1-QM4

QM4 This is the queue model corresponding to 802.11 curve

Figure 13 plots the expected number of clients backlogged as a function of the load for the case of 20 clients for QM1-QM4. We see that QM3 and QM4 show a much sharper knee close to their saturation throughput. This is a result of the strong negative pull away from 20 toward 0 in both systems.

Furthermore, we notice that function $p_s(k)$ associated with MAC1 is always less than all other schemes, and at the same time total number of backlogged servers under slotted Aloha is much greater than the number of backlogged users for other MAC schemes. We formalize this observation via a sample path argument in the appendix.

VI. Fairness, Anomalous Behavior, and Practical Design

Results shown in the previous sections have significant implications in the context of practical design and improvements to existing 802.11 systems and the existing transport layer mechanisms operating and interacting on top of an 802.11 system. For example, the above study can shed light on the benefits of scheduling at higher layers (proxy-level) as studied recently in the literature (see for example [9]). Similarly, as we will see next, these results can also significantly alter our understanding of the impact of wireless channel conditions on the performance of 802.11.

In this section we restrict our attention to what has become known as the “anomalous behavior of 802.11” [5] when considering the impact of asymmetric wireless channel conditions. The behavior of 802.11 in the presence of channel asymmetry (sometimes called rate diversity) has been studied in [5], [10], [11], and [12]. It has been shown that when the packet transmission duration varies with the quality of channel, the low quality of one user’s channel reduces the aggregate throughput, and hence individual users’ throughput, by a factor of 10%-16%. Intuitively this is because 802.11 achieves long-term fair channel access for all participating users (in terms of their probability of successfully reserving the channel for transmission). As low-quality users occupy the channel for longer periods of time (due to fixed packet size and low bit rate), whenever they get hold of the channel, equal channel access in effect reduces the actual amount of time a high-quality user gets to occupy the channel, thus causing the reduction in aggregate throughput. This is an undesirable property since it in effect can force some users to an unstable operating point. In addition, this creates an incentive for a rational user to drive the system to an inefficient equilibrium.

It can be argued that it is beneficial to provide fairness in terms of access-time. The authors in [10] guarantee this access-time fairness via a time-based regulator. Another alternative would be if packet durations are kept equal to guarantee equal channel access time among users. Here we propose a re-packeting of arriving packets according to the requirements of the channel to ensure a fixed time duration for each packet, which will be referred to as the *reference transmission time*. In particular, we suggest that the anomalous behavior of 802.11 can be resolved if the notion of maximum packet size in the standards is refined to be dependent on the bit rate or PHY rate determined by the channel quality. This, in fact, is consistent with the second condition for an "ideal MAC protocol" proposed by Tan et. al. in [12]. It would guarantee service to each user at a rate greater than or equal to the saturation throughput $\lambda_{\text{max}}(n)$ (in terms of packets/sec). This achieves similar fairness properties as the schemes proposed in [10] in that it avoids any reduction in the aggregate throughput while guaranteeing that users transmitting at a lower PHY rate receive a throughput higher than what they would achieve in a single-rate system where all users are at the lower PHY rate.

Below we show how the proposed re-packeting scheme impacts the system under realistic arrival traffic (where most users have finite rate of arrivals). Our delay results presented earlier can be used to provide unintuitive predictions regarding the delay characteristics and fairness (in terms of delay) of the proposed system. Assuming the *information rate* (this is the arrival rate measured in information bits per second) at each user is independent of the channel conditions (we ignore the overhead associated with re-packeting and ignore the impact of the transport layer's rate control mechanisms, at least on short time scales), the question is how the channel degradation for one user affects the delay characteristics of others. Under the proposed re-packeting scheme, newly arrived packets are split and reassembled into smaller equal-sized packets (of lower information content) such that the resulting transmission time of such a smaller packet, under the lower channel quality and lower PHY rate, would remain at the reference transmission time (thus higher channel quality leads to bigger packets). It is thus not hard to see that when the arrival rate (in bits per second) is fixed, a user with channel degradation is equivalent to one with a higher arrival rate in packets per second (but each packet is smaller in data bits) and burstier packet arrivals. Let us, for a moment, neglect the increased burstiness associated with re-packeting and consider only the impact of increasing the rate. From Figures 8 and 9, we know that an increase in one user's arrival rate (in packets per second) has minimal impact on the delay performance of the other fixed arrival rate users within a certain region, and this impact in the worst case is bounded. It is, then, not hard to convince oneself that (modulo our neglecting the increased burstiness) *the proposed re-packeting scheme will demonstrate the same desirable delay characteristics*.

In the remainder of this section, we provide simulations to confirm the above statement in a realistic setting. Notice that our proposed re-packeting scheme guarantees 100% throughput for users with high PHY rate and stable arrival rates (less than λ_{max}), hence bounding the impact on their delay. We go even further and predict that the delay degradation of high PHY rate users due to the drop in one user's PHY rate, is in many cases fairly insignificant.

The following simulation setting, approximately, illustrates the above prediction. We consider a simple example of $n = 20$ users. We assume that the first $n - 1 = 19$ users have good channel quality, allowing them to operate at 11Mbps PHY rate, while the last user's channel degradation is such that physical layer transmission is only possible at 5.5 Mbps PHY rate. We assume that the first 19 users have an arrival rate equal to 228 Kbps (99% of $\lambda_{\text{max}}(20)$). Figure 14 shows that, under the re-packeting scheme, the user with 5.5 Mbps PHY rate is able to sustain an arrival rate of 228 Kbps (same as the others) without significant delay degradations caused to other users (around 20ms). Moreover, if the client with a bad channel adapts to its channel conditions and reduces its arrival rate to around 120Kbps (almost half of its previous rate when it had a good channel), not only is the delay degradation to other users minimal (14ms), but also the delays of this client will stay reasonable in the 20ms range. Notice that in the absence of re-packeting, the drop in the overall throughput would be sufficient to drive all users into the saturation regime (including the user with the bad channel) where E2E delay would grow without bound. This situation is an instance where fairness improves the performance of individual users, even though it might at first appear that re-packeting a user's packets into many packets would mean increasing its delay compared to the case when it did not split up its packets.

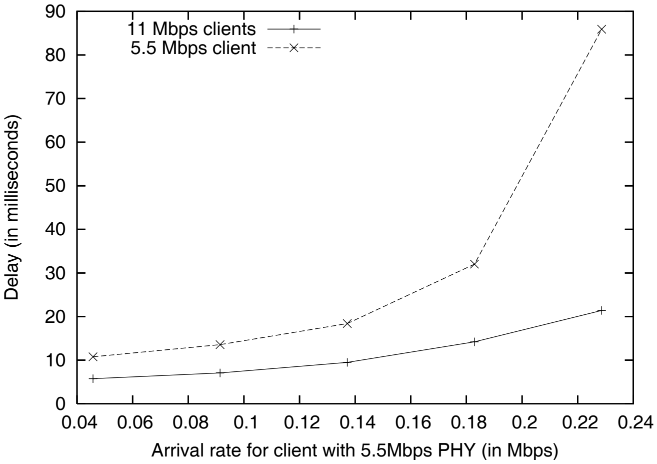

Fig. 14. End to end delays (in milliseconds) as a function of the arrival rate of a user with lower channel quality. Here 19 clients transmit at 11Mbps, while the channel conditions of one client necessitates that it drops its physical layer data rate down to 5.5 Mbps.

Unfortunately such an insignificant delay degradation to high PHY rate users is not always possible. In fact, if the low

PHY rate user has to drop its rate to 1Mbps rather than 5Mbps, it can cause much more significant performance degradation to others. As mentioned before, such a drop, under the re-packeting scenario, does not impact the throughput of the high PHY rate users (still at 228 Kbps), but it does significantly increase their delay (see Figures 15 and 16). Now the low PHY rate node's delay and throughput under the re-packeting scheme can potentially be worse than those under the current 802.11 implementation. This is because such a drop in the current 802.11 MAC (with no re-packeting) might or might not cause instability for the original users with low quality channel, despite its impact on all other users who will certainly experience an unbounded delay degradation due to instability. For instance, we know that at low PHY rates (say less than 100 Kbps), this user will not be saturated (as all users will get around 160 Kbps), even though the drop in its channel quality will saturate all others. What we do know is that re-packeting would not decrease the original users' throughput beyond the throughput it would achieve in a single rate system with 1 Mbps PHY rates (32 Kbps). In other words, the re-packeting scheme and standard 802.11 provide very different notions of fairness in this system. In such a scenario, investigating the trade-off between fairness and overall system performance is rather subjective and beyond the scope of this paper (we refer the reader to the very thorough discussions provided in [10] and [12]).

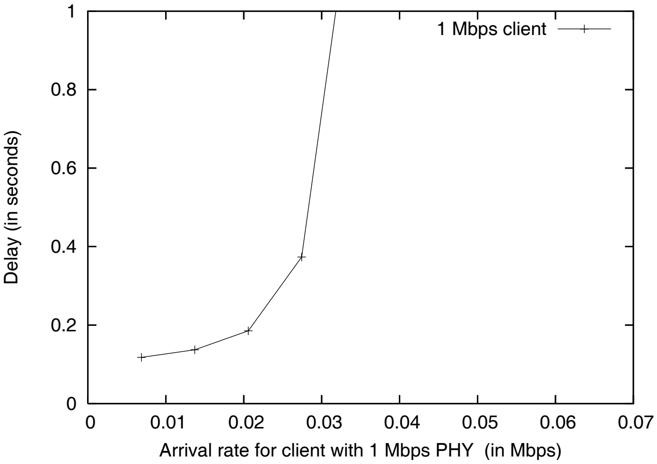

Fig. 15. End to end delay (in seconds) for the client with lower channel quality as a function of its arrival rate. Here 19 clients transmit at 11Mbps, and one client transmits at 1 Mbps.

VII. Conclusion

In this paper we presented a comprehensive study of 802.11 networks in the non-saturation case, especially when the users queues are critically loaded (i.e., with arrival rates approaching the saturation throughput). We examined the system performance under symmetric and asymmetric cases. In the former, all users have the same arrival rate and approach the saturation throughput from below. We showed that the system has a rather unintuitive delay performance that is closely connected to a very low level of backlogged users. In the asymmetric case we investigated how unused bandwidth from finite source queues is shared among saturated sources. We also showed that the negative impact an aggressive user has on finite source queues is bounded and that the system provides little incentive for a user to increase its demand. We also presented a sequence of MAC schemes in an attempt to isolate the key mechanism responsible for the above observations.

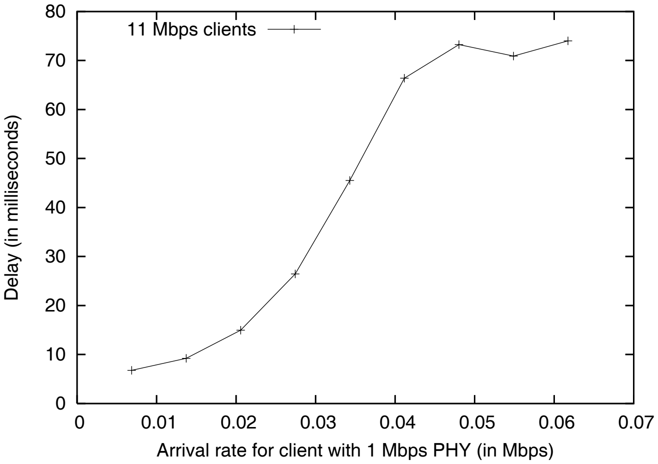

Fig. 16. End to end delays (in milliseconds) for the clients with higher channel quality as a function of the arrival rate of the client with lower channel quality. Here 19 clients transmit at 11Mbps, and one client transmits at 1 Mbps.

As mentioned earlier, our study in the non-saturation regime complements the numerous studies in the saturation regime of 802.11 networks. We believe that there remain many open issues in this direction and that this paper serves as a first step toward fully understanding the intricate dynamics of 802.11 networks. In summary, we believe that our study points out the necessity of further studies in the following directions:

- The notion of saturation is pessimistic as it arrives at the maximum possible contention; instead notions of capacity regions, similar to the same notions in Aloha ([13], [14], and [15]), must be developed and considered.
- The delay performance for 802.11 networks in the saturation regime is distinctly (orders of magnitude) worse than in realistic settings; any useful delay studies of the system should address the performance under finite arrival rates.

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] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *Proceedings of IEEE INFOCOM 2003*, vol. 22, 2003, pp. 836–843.

[6] R. Vijayakumar, T. Javidi, and M. Liu, “From saturation to non-saturation: A study on 802.11 networks,” Dept. of Electrical Engg, University of Michigan, Tech. Rep. CSPL-363, 2005.

[7] S. Choi, K. Park, and C. kwon Kim, “On the performance characteristics of wlans: revisited,” in *Proceedings of the 2005 ACM SIGMETRICS*, 2005, pp. Pages: 97 – 108.

[8] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Prentice-Hall, 1992.

[9] T. Simunic, W. Quadeer, and G. D. Micheli, “Managing heterogeneous wireless environments via hotspot servers,” in *Proceedings of MMCN 2005*.

[10] G. Tan and J. Guttag, “Time-based fairness improves performance in multi-rate wlans,” in *Proc. of USENIX’04*, Boston, MA, 2004.

[11] ——, “Long-term time-share guarantees are necessary for wireless lans,” in *Proc. of SIGOPS European Workshop*, Leuven, Belgium, Sept. 2004.

[12] ——, “The 802.11 mac protocol leads to inefficient equilibria,” in *Proc. of IEEE INFOCOM’05*, Miami, FL, 2005.

[13] R. R. Rao and A. Ephremides, “On the stability of interacting queues in a multiple-access system,” *IEEE Trans. Inform. Theory*, vol. 34, no. 5, pp. 918–930, Sept. 1988.

[14] V. Anantharam, “The stability region of the finite-user slotted ALOHA protocol,” *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 535–540, May 1991.

[15] W. Szpankowski, “Stability conditions for some distributed systems: Buffered random access systems,” *Advances in Applied Probability*, vol. 26, pp. 498–515, June 1994.

Appendix

In this appendix, we attempt to formalize our observation regarding the relationship between $p_s(k)$ curves and the expected number of users backlogged. This is, in particular, relevant to explain the poor performance of regular slotted Aloha.

We propose an alternate construction of the stochastic processes governing the multi-queue systems discussed in Section V-B. This alternate construction provides a way to make sample path comparisons of two systems which have different values for the $p_s(k)$.

The probability space is defined by the following independent processes, each of which consists of an iid sequence of random variables:

1) Aggregate arrival process: $A(t)$ is the aggregate number of arrivals to the system in slot t; it is Poisson distributed with parameter λ.
2) Splitting of arrivals: $X(k)$ is uniformly distributed on the integers $1 \dots n$; the k-th aggregate arrival is assigned to client $X(k)$.
3) Successful transmission process: $Y(t)$ is a continuous random variable uniformly distributed on $[0, 1)$. There is a successful transmission in slot t if $Y(t) < p_s(k)$, where k is the number of clients backlogged in slot t
4) Selection of successful transmitter: Roughly speaking, we want a process that will select which of the k backlogged clients was successful if there is a successful transmission, i.e. if $Y(t) < p_s(k)$. It will be convenient to construct this as follows: For each time slot t, let $Z_{t1}, Z_{t2}, \dots$ be an iid integer valued sequence uniformly distributed on $1 \dots n$. Let $B_t \subset \{1 \dots n\}$ denote the set of backlogged clients in slot t. We pick the successful transmitter to be the first of the Z_{ti} which actually lies in the set B_t, i.e. the successful transmitter is Z_{tj}, where $j = \arg\min_i Z_{ti} \in B_t$. It is easy to see that this construction picks the successful transmitter uniformly among the set of backlogged clients.

The above construction can be used for any slotted multiple access system with the properties that

1) The probability of a successful transmission in a slot is a deterministic function $p_s(k)$ of the number of backlogged clients k independent of the history of the system and
2) Given that a successful transmission occurred, it is equally likely to have come from any of the backlogged clients.

Suppose that we have two different systems S_1 and S_2 with success probability functions $p_s(k)$ and $q_s(k)$ respectively, and with the property that $p_s(k) \leq q_s(k)$ and $q_s(k) \geq q_s(k+1)$ for $k \geq 1$; i.e. we require the success probabilities to be higher under S_2 than under S_1, and that the success probabilities under S_2 are non-increasing. As an example, S_1 could be standard Aloha (where the $p_s(k)$ are increasing in k) and S_2 could be optimal Aloha.

Theorem 1: When the system starts from the empty state, then along every sample path, the length of each of the n queues under S_2 is less than or equal to the length of the corresponding queue under S_1. Note that since this implies that every arriving packet leaves S_2 no later than it leaves S_1, it follows that the average delay under S_2 is no more than the average delay under S_1.

Proof:

Proceed by induction. Suppose that the claim holds up to slot $t - 1$. The arrivals to both systems are identical (and we assume that new arrivals cannot be transmitted until the slot following their arrival), and therefore we only need to show that if there is a successful transmission from client i under S_1 *and* client i is also backlogged under S_2, then client i must also transmit successfully under S_2.

Let $k1$ (resp. $k2$) denote the number backlogged under S_1 (resp. S_2) in slot t. By the induction hypothesis, $k2 \leq k1$. If $k2 = 0$, we are done; so suppose that $k2 \neq 0$. The conditions we impose on $p_s(k)$ and $q_s(k)$ ensure that $k2 \leq k1 \Rightarrow q_s(k2) \geq p_s(k1)$. Therefore if there is a successful transmission under S_1, i.e. $Y(t) < p_s(k1)$, then we must have $Y(t) < q_s(k2)$, i.e. we also have a successful transmission under S_2. Further, the construction of the choice of the successful transmitter via the Z_{ti} ensures that if the successful client under S_1 is also backlogged under S_2, then it will also be the successful client under S_2.

■