Scaling Properties of IEEE 802.11 Wireless Networks

Fragkiskos Papadopoulos, Konstantinos Psounis
University of Southern California
E-mail: fpapadop, kpsounis@usc.edu.

***Abstract*— We consider a single-hop wireless network consisting of $\alpha \cdot N$ sources, where $\alpha \geq 1$ is a scaling factor. These sources are randomly distributed around a single base-station/access-point and utilize the IEEE 802.11 standard for medium access control. The transmission speed of each node C, the minimum contention window CW_{min}, and the maximum contention window CW_{max}, are all multiplied by the scaling factor α. Further, all protocol time-intervals are multiplied by $\frac{1}{\alpha}$.**

We show that as the scaling factor α increases, the packet delays become independent of α, and therefore, of the number of sources ($\alpha \cdot N$) sharing the wireless channel. At the same time, the user's perceived throughput and drop ratio remain almost invariant.

This result is not only of theoretical interest, but also of great practical interest, as it clearly identifies the set of the system's parameters that we should aim to (simultaneously) scale in future versions of the IEEE 802.11 protocol (or in new protocols that utilize similar ideas), so that the system can support a very large number of users, while continuing to deliver to each user at least as good performance as before.

***Index Terms*—IEEE 802.11 networks, Performance-preserving scaling laws.**

I. Introduction

The IEEE 802.11 MAC protocol [1] has gained widespread popularity and has been adopted as the de-facto layer 2 protocol for wireless local area networks (WLANs). Because of its popularity, there has been a large body of work focusing on its analytical modeling, *e.g.* [2], [3], [4], [5], simulation study, *e.g.* [6], [7], and measurement-based performance evaluation, *e.g.* [8], [9], [10].

We summarize the protocol's main functionality. Time is slotted with the duration of each slot equal to a constant value, which we will be calling SlotTime. A station that wants to transmit a frame first senses the medium. If the medium is idle, the station waits for a time interval called DIFS (Distributed Inter-Frame Space) and senses again. If the medium is free the frame is transmitted. If the receiver gets the frame correctly it sends an acknowledgment (ACK) to the sender after a SIFS (Short Inter-Frame Space) time interval.

If the medium is found to be busy at any time instance that a station wishes to transmit, a back-off procedure is invoked. When invoked, the station waits until the medium is free for a DIFS time and starts a random timer uniformly distributed between $\{0...CW_{min} - 1\}$ time-slots. The parameter CW_{min} is called *minimum contention window.* The timer is decremented as long as the medium is free. If at any time instance the medium becomes busy the timer is frozen, and it is decremented when the medium becomes free again for a DIFS time. When the timer reaches zero the frame is transmitted.

A collision (which occurs when two or more stations transmit at the same time) is detected through the lack of an ACK. If a collision occurs the station backs-off again; the back-off timer is now uniformly distributed in $\{0...2CW_{min} - 1\}$. And, in general, after the $i^{th} < m$ consecutive unsuccessful attempt the back-off timer is uniformly distributed in $\{0...2^i CW_{min} - 1\}$, whereas for $i \geq m$ it is uniformly distributed in $\{0...2^m CW_{min} - 1\}$. The value $CW_{max} = 2^m CW_{min}$ is called *maximum contention window.* The maximum value of i is equal to the Long Retransmit Limit if the RTS-CTS (Ready-To-Send-Clear-To-Send) option is used, or equal to the Short Retransmit Limit if the RTS-CTS option is not used. After i reaches its maximum value the frame is dropped. [1]

In this paper we consider a wireless network consisting of αN sources, where $\alpha \geq 1$ is a scaling factor. The sources are distributed around a single base-station/access-point and utilize the IEEE 802.11 standard for medium access control. The sources generate traffic destined to the base station according to some arbitrary arrival process. We consider the single-hop case, where the base station is within the transmission range of each source (*e.g.*, such as in the popular Wi-Fi WLANs). Further, for ease of exposition, in this paper we also assume that each source is within the range of every other source, so that a transmission can be sensed by all stations. (The same results hold even if this is not the case, and will be published in a longer version of the paper.)

We scale the transmission speed of each source C, the minimum contention window CW_{min}, and the maximum contention window CW_{max}, by the factor α. In other words, C, CW_{min}, and CW_{max} become $\alpha C, \alpha CW_{min}$, and αCW_{max}. Further, we also scale all protocol time-intervals, i.e. the SIFS and DIFS durations, and the SlotTime value, by $\frac{1}{\alpha}$. [2] We show that as the scaling factor α increases the packet delays initially decrease and then quickly become independent of α, and therefore, of the factor by which the number of sources that share the wireless channel increases.

[1] Note that when the RTS-CTS option is used, a station that wishes to transmit a frame first sends an RTS (Ready-to-Send) message to the destination, in order to reserve the channel. The transmission of the actual frame starts when the station under study receives a CTS (Clear-to-Send) message from the destination. Under this scheme, the vast majority of collisions involve RTS frames and not actual data frames.

[2] Note that usually DIFS $=$ SIFS $+$ 2SlotTime, in which case we only need to scale the SIFS and the SlotTime durations.

WIOPT 2008, 1st – 3rd Apr 2008, Berlin, Germany.

ISBN 978-963-9799-18-9
DOI 10.4108/ICST.WIOPT2008.3114

At the same time, the user's perceived throughput and drop ratio remain almost invariant. Besides the theoretical interest, a main practical contribution of this result is that it clearly identifies the set of the system's parameters that we should aim to (simultaneously) scale in future versions of the IEEE 802.11 protocol (or in new protocols that utilize similar ideas), so that the system can support a very large number of users while continuing to deliver to each user at least as good performance as before.

The rest of the paper is organized as follows. In Section II we briefly discuss related work. In Section III we study how the system behaves under the aforementioned scaling, assuming a saturated scenario, where each source always has a packet/frame to send. We waive this assumption in Section IV, and present simulation results in Section V to verify our theoretical arguments. Conclusions along with a discussion on the applicability of the results, and future work directions, follow in Section VI.

II. Related Work

The idea of scaling a network in a manner that performance is preserved has been extensively studied for the case of *wireline* networks that resemble the Internet. For example, Psounis et al. [11], [12] have introduced a method called SHRiNK that creates a slower version of the original network and can predict significant performance measures by observing the slower replica. Further, Papadopoulos et al. [13] have introduced two methods called DSCALEd and DSCALEs that perform topological downscaling by retaining only the congested links of the original network, and extrapolate the performance of the downscaled network to that of the larger Internet.

In the context of *wireless* networks, to our best knowledge, the only relevant to this studies are the recent ones by Papadopoulos et al. [14], and by Naik et al. [15]. In [14] it has been shown that it is possible to predict the full behavior of an arbitrary mobile ad hoc network deployed in an outdoor environment at one spatial scale, by a suitably scaled replica consisting of the *same* number of nodes but deployed in an outdoor environment at another spatial scale. This is accomplished by preserving the link statistics. And, [15] does the same for static ad hoc networks operating in indoor environments. In this paper, we investigate whether performance can be preserved while *scaling* (*e.g.* increasing) the number of nodes of the wireless network.

Another somewhat relevant body of work focuses on finding the maximum achievable throughput and characterizing capacity-delay tradeoffs and connectivity in wireless ad hoc networks as the number of nodes increases, *e.g.* see [16], [17], [18], and references therein. These studies are primarily interested in the asymptotic behavior of the system and derive analytical results usually under quite simplified models for the MAC protocol. As mentioned earlier, here we study whether performance can be preserved while increasing the number of nodes in a wireless network, by appropriately scaling the system's parameters. Further, we are primarily interested on the exact network behavior as a function of the number of nodes, and not on asymptotic results. And, we study this in a realistic manner, by considering all aspects of the IEEE 802.11 MAC protocol.

III. The Saturated Case

In the analysis that follows we use the subscript α on the various parameters of interest, in order to denote that these correspond to an α-scaled system. We first assume that each station has always a frame available for transmission, that is, we first consider the saturated case scenario. We start by studying how the average contention window and the probability that there is a collision, vary with α.

A. Average contention window and collision probability

Let $m_\alpha = \log_2\left(\frac{\alpha CW_{max}}{\alpha CW_{min}}\right) = \log_2\left(\frac{CW_{max}}{CW_{min}}\right) = m_1$. Given that a station has a frame to transmit it can be easily shown, *e.g.* see [3], that its average contention window size at some arbitrary time is:

$$\overline{W_\alpha} = \frac{1 - p_\alpha - p_\alpha(2p_\alpha)^{m_\alpha}}{1 - 2p_\alpha}\frac{\alpha CW_{min}}{2}, \tag{1}$$

where p_α is the probability that the station under study experiences a collision when it attempts to transmit a frame.

The probability that a station transmits in some time-slot can be approximated by $\frac{1}{\overline{W_\alpha}}$ [3], [4]. And therefore, the collision probability, which is the probability that at least one other station transmits in the same time-slot, is given by:

$$\begin{aligned} p_\alpha &= 1 - (1 - \frac{1}{\overline{W_\alpha}})^{\alpha N - 1} \\ &= 1 - (1 - \frac{1 - 2p_\alpha}{1 - p_\alpha - p_\alpha(2p_\alpha)^{m_\alpha}}\frac{2}{\alpha CW_{min}})^{\alpha N - 1}, \end{aligned} \tag{2}$$

where αN is the total number of sources sharing the channel, as mentioned earlier. We are now ready to state our first theorem.

Theorem 1: Under the scaling we perform the collision probability is approximately independent of α, that is, $p_\alpha \approx p_1$, and $\overline{W_\alpha} \approx \alpha\overline{W_1}$.

Proof: First, recall that $m_\alpha = m_1$. Equation (2) can be approximated by:

$$\begin{aligned} p_\alpha &\approx 1 - \exp\left(-(\alpha N - 1)\frac{1 - 2p_\alpha}{1 - p_\alpha - p_\alpha(2p_\alpha)^{m_\alpha}}\frac{2}{\alpha CW_{min}}\right) \\ &= 1 - \exp\left(-\frac{(\alpha N - 1)}{\alpha}\frac{1 - 2p_\alpha}{1 - p_\alpha - p_\alpha(2p_\alpha)^{m_1}}\frac{2}{CW_{min}}\right) \\ &\approx 1 - \exp\left(-N\frac{1 - 2p_\alpha}{1 - p_\alpha - p_\alpha(2p_\alpha)^{m_1}}\frac{2}{CW_{min}}\right) \end{aligned}$$

From the above relation we can deduce that $p_\alpha \approx p_1$, and therefore, from Equation (1), that $\overline{W_\alpha} \approx \alpha\overline{W_1}$. [3]

[3] Notice that we have used the fact that $(1 - x)^k \approx \exp(-kx)$, which is accurate if k is not too small and x is not too large. In our case, k corresponds to $\alpha N - 1$, which is not small if the *product* αN is not small. Therefore, this requirement is satisfied even for $\alpha = 1$ as long as $N - 1$ is not too small. As we will demonstrate in Section V via simulations, an $N \geq 4$ is sufficient for these approximations to take place. Further, x corresponds to $\frac{1}{\overline{W_\alpha}} \leq \frac{1}{\alpha CW_{min}}$. Since, by the protocol's specifications, $CW_{min} = 31$ [1], $x \leq \frac{1}{\alpha} \cdot \frac{1}{31}$, which is small as required, even for $\alpha = 1$.

■

The intuition behind Theorem 1 is that while the number of competing stations increases by a factor α, the probability that each station transmits at some arbitrary slot decreases by a factor α (due to the scaling we perform to the maximum and minimum contention window sizes), thus leaving the collision probability almost unaltered. However, as we will see next, while the transmission probability decreases, the scaling we perform to the protocol's time-intervals ensures that the actual time duration until a station successfully transmits its frame, remains virtually unchanged.

B. Frame Service Time

We are now ready to study how the service time of a frame behaves as a function of the scaling factor α. The frame service time is defined as the time elapsed from the moment that the frame becomes the Head-of-Line frame in the interface transmission queue (i.e. it is ready for transmission for the first time), until it is successfully received by the destination. As we can deduce from the description of the MAC protocol in the previous section, there are four components contributing to the service time of a frame: (i) The total number of back-off slots the station has to wait before its frame is successfully transmitted, (ii) the total amount of time the back-off counter of the station under study is kept frozen because of frame transmissions and/or collisions among the other stations that share the channel, (iii) the total amount of time lost due to collisions of the frame under study, and (iv) the time needed to transmit the frame under study. For ease of exposition, we first analyze the frame service time ignoring the amount of time that the back-off counter is kept frozen. We denote this time duration by T_a. Then, we compute the time duration that the timer is kept frozen, which we denote by T_a^{tf}, and add it to T_a to get the total frame service time T_α^{tot}. Our analysis is inspired by the analysis in [3] and [4].

For simplicity, let's assume a constant frame length, and denote the (successful) frame transmission time in an α-scaled system by T_α^{frame}. Note that this time also includes the time needed to reserve the channel in the case where the RTS-CTS mechanism is used, as well as the time to receive an ACK from the receiver. In other words, $T_\alpha^{frame} = T_\alpha^{RTS} + T_\alpha^{CTS} + T_\alpha^{dframe} + \text{SIFS}_\alpha + T_\alpha^{ACK}$, where $T_\alpha^{RTS}/T_\alpha^{CTS}$ is the time required to transmit an RTS/CTS message in an α-scaled system, T_α^{dframe} is the time needed to transmit the actual data frame, T_α^{ACK} is the required time to transmit an ACK, and SIFS_α is the SIFS duration. Also, let's denote the duration of a collision by T_α^{COLL}. According to the protocol, $T_\alpha^{COLL} = \text{DIFS}_\alpha + T_\alpha^{RTS}$. In situations where RTS-CTS messages are not used, $T_\alpha^{frame} = T_\alpha^{dframe} + \text{SIFS}_\alpha + T_\alpha^{ACK}$, and the duration of a collision is simply given by $T_\alpha^{COLL} = \text{DIFS}_\alpha + T_\alpha^{dframe}$.

Before proceeding, recall that we scale all protocol time intervals by $\frac{1}{\alpha}$. This means that $\text{DIFS}_\alpha = \frac{\text{DIFS}_1}{\alpha}, \text{SIFS}_\alpha = \frac{\text{SIFS}_1}{\alpha}$, and $\text{SlotTime}_\alpha = \frac{\text{SlotTime}_1}{\alpha}$. Further, we also scale the transmission speed (C) of each node by α. Therefore, it is easy to see that $T_\alpha^{frame} = \frac{T_1^{frame}}{\alpha}$, and $T_\alpha^{COLL} = \frac{T_1^{COLL}}{\alpha}$.

Let X_α^i denote a random variable that is uniformly distributed in $\{0 ... 2^i \alpha CW_{min} - 1\} \approx \{0 ... 2^i \alpha CW_{min}\}$, and let BO_α be the random variable that represents the number of back-off slots a station needs to count down before its frame is successfully transmitted. Further, denote by T_α^k the service time of a frame given that there were k collisions (of this frame), and assume for now that there are no events that freeze the back-off counter.

Since we are assuming a saturated scenario, a station will sense the medium to be busy in its first transmission attempt and will set its back-off timer (after a DIFS_α interval) to $BO_\alpha = X_\alpha^1$ time-slots. When the back-off timer reaches zero the station will successfully transmit the frame with probability $1 - p_\alpha \approx 1 - p_1$ (as the collision probability is $p_\alpha \approx p_1$). Hence, with probability $1 - p_1$, $T_\alpha^0 = \text{DIFS}_\alpha + X_\alpha^1 \times (\text{SlotTime}_\alpha) + T_\alpha^{frame}$. If there is a collision and the node successfully transmits its frame on its second attempt (an event with probability $p_1(1 - p_1)$), then $BO_\alpha = X_\alpha^1 + X_\alpha^2$, and hence $T_\alpha^1 = \text{DIFS}_\alpha + X_\alpha^1 \times (\text{SlotTime}_\alpha) + T_\alpha^{COLL} + X_\alpha^2 \times (\text{SlotTime}_\alpha) + T_\alpha^{frame}$. And, in general, if there are k collisions before a successful transmission (an event that occurs with probability $p_1^k(1 - p_1)$), $T_\alpha^k = \text{DIFS}_\alpha + \sum_{i=1}^{k+1} X_\alpha^i \times (\text{SlotTime}_\alpha) + kT_\alpha^{COLL} + T_\alpha^{frame}$. Notice that the random variables X_α^i and αX_1^i have the same distribution. Hence, it is easy to see that T_α^k can be written as follows:

$$T_\alpha^k = \frac{\text{DIFS}_1}{\alpha} + \sum_{i=1}^{k+1} X_1^i \times (\text{SlotTime}_1) + \frac{kT_1^{COLL}}{\alpha} + \frac{T_1^{frame}}{\alpha}. \tag{3}$$

Let K_{max} be the maximum number of collisions allowed by the protocol before a frame is dropped (*e.g.* as defined by the Long Retransmit Limit, as explained earlier). The frame service time T_α (which ignores events that freeze the back-off timer) is therefore given by:

$$T_\alpha = \sum_{k=0}^{K_{max}} T_\alpha^k p_1^k (1 - p_1). \tag{4}$$

Notice that K_{max} is not scaled by our operations in any way. We can now state our second theorem.

Theorem 2: Under the scaling we perform, and ignoring events that freeze the back-off counter, the frame service time initially decreases as we increase α, and then its distribution converges to a limiting distribution that does not depend on α. [4]

Proof: From Equation (3) we can see that as α increases T_α^k decreases for all k. Therefore, by Equation (4), T_α also decreases. Further, as $\alpha \to \infty$, $T_\alpha^k \to \sum_{i=1}^{k+1} X_1^i \times (\text{SlotTime}_1)$, i.e. becomes independent of α for all k, and therefore, T_α also becomes independent of α. ■

Now let's account for events that freeze the back-off counter. As mentioned before, each station senses the medium in each time-slot. If other stations transmit in the same time-slot (either successfully or unsuccessfully) the back-off counter at the station under study is kept frozen. We are interested in the total time duration T_α^{tf} that the timer is kept frozen between two

[4] When we say that a (positive) random variable X is smaller compared to a (positive) random variable Y, we mean that $P(X \leq x) = P(Y \leq x + \delta)$, $\forall x$ and for some constant $\delta > 0$.

successful transmissions in an α-scaled system. This quantity can be written as follows:

$$
\begin{aligned}
T_\alpha^{tf} &= \sum_{i=1}^{BO_\alpha} 1_{[coll_i]} T_\alpha^{COLL} \\
&+ \sum_{i=1}^{BO_\alpha} 1_{[trans_i]} (T_\alpha^{frame} + \text{DIFS}_\alpha), \qquad (5)
\end{aligned}
$$

where BO_α is the back-off counter between two successful transmissions of the station under study as defined earlier, and $1_{[coll_i]}/1_{[trans_i]}$ are indicator functions, which are 1 if there was respectively a collision/transmission in slot i due to the other $\alpha N - 1$ stations that compete for the channel, and zero otherwise. The probability q_α^c that there is a collision in some time-slot among these other stations, is just the probability that two or more of these stations attempt a transmission. It is easy to see that this probability can be expressed as:

$$
q_\alpha^c = 1 - (1 - \frac{1}{\overline{W_\alpha}})^{\alpha N - 1} - \frac{\alpha N - 1}{\overline{W_\alpha}} (1 - \frac{1}{\overline{W_\alpha}})^{\alpha N - 2},
$$

where the second and third term on the right hand side of the relation are respectively the probabilities that no other station and exactly one other station transmit in some time-slot. As before (and recalling that $\overline{W_\alpha} \approx \alpha \overline{W_1}$), we can make the following approximations:

$$
\begin{aligned}
q_\alpha^c &\approx 1 - \exp\left(-\frac{\alpha N - 1}{\overline{W_\alpha}}\right) - \frac{\alpha N - 1}{\overline{W_\alpha}} \exp\left(-\frac{\alpha N - 2}{\overline{W_\alpha}}\right) \\
&\approx 1 - \exp\left(-\frac{N}{\overline{W_1}}\right) - \frac{N}{\overline{W_1}} \exp\left(-\frac{N}{\overline{W_1}}\right) \approx q_1^c.
\end{aligned}
$$

The above suggests that the collision probability among the other $\alpha N - 1$ stations is approximately independent of the scaling factor α. The intuition behind this is the same as the one for Theorem 1. And, the same holds for the probability q_α^{sc} that there is a successful transmission among the other $\alpha N - 1$ stations, which is just the probability that exactly one of these stations transmits:

$$
\begin{aligned}
q_\alpha^{sc} = \frac{\alpha N - 1}{\overline{W_\alpha}} (1 - \frac{1}{\overline{W_\alpha}})^{\alpha N - 2} &\approx \frac{\alpha N - 1}{\overline{W_\alpha}} \exp\left(-\frac{\alpha N - 2}{\overline{W_\alpha}}\right) \\
&\approx \frac{N}{\overline{W_1}} \exp\left(-\frac{N}{\overline{W_1}}\right) \approx q_1^{sc}.
\end{aligned}
$$

Further, from our earlier discussion it is easy to see that BO_α and αBO_1 have the same distribution. Also, since $T_\alpha^{COLL} = \frac{T_1^{COLL}}{\alpha}$, $T_\alpha^{frame} = \frac{T_1^{frame}}{\alpha}$ and $\text{DIFS}_\alpha = \frac{\text{DIFS}_1}{\alpha}$, Equation (5) can be written as follows:

$$
\begin{aligned}
T_\alpha^{tf} &= \sum_{i=1}^{\alpha BO_1} \frac{1}{\alpha} 1_{[coll_i]} T_1^{COLL} \\
&+ \sum_{i=1}^{\alpha BO_1} \frac{1}{\alpha} 1_{[trans_i]} (T_1^{frame} + \text{DIFS}_1) \\
&= \sum_{i=0}^{BO_1 - 1} T_1^{COLL} \frac{1}{\alpha} \left(\sum_{j=\alpha i+1}^{\alpha i+\alpha} 1_{[coll_j]}) \right) \\
&+ \sum_{i=0}^{BO_1 - 1} (T_1^{frame} + \text{DIFS}_1) \frac{1}{\alpha} \left(\sum_{j=\alpha i+1}^{\alpha i+\alpha} 1_{[trans_j]}) \right) \\
&\approx \sum_{i=0}^{BO_1 - 1} T_1^{COLL} q_1^c + \sum_{i=0}^{BO_1 - 1} (T_1^{frame} + \text{DIFS}_1) q_1^{sc}, \qquad (6)
\end{aligned}
$$

where the last approximation holds for sufficiently large α by the Law of Large Numbers, where $\frac{1}{\alpha}\left(\sum_{j=\alpha i+1}^{\alpha i+\alpha} 1_{[coll_j]}\right) \to q_1^c$, and $\frac{1}{\alpha}\left(\sum_{j=\alpha i+1}^{\alpha i+\alpha} 1_{[trans_j]}\right) \to q_1^{sc}$. (Note that the convergence here is expected to occur in practice for small values of α, as the events of collisions or successful transmissions on different time-slots are loosely correlated due the protocol's back-off mechanism.) We can now state the following theorem, whose proof follows immediately from the above arguments.

Theorem 3: Under the scaling we perform the total time duration that the timer is kept frozen between two successful transmissions is approximately independent of α.

Since the total service time of a frame in an α-scaled system is $T_\alpha^{tot} = T_\alpha + T_\alpha^{tf}$ (where T_α as given by Equation (4) and T_α^{tf} as given by Equation (6)), we can state the following corollary for T_α^{tot}:

Corollary 1: Under the scaling we perform, as the scaling factor α increases, the frame service time T_α^{tot} first decreases, and then its distribution converges to a distribution that is independent of α.

C. *User throughput and drop ratio*

Since the collision probability is approximately independent of the scaling factor α, and considering the fact that the maximum number of allowed collisions (K_{max}) before a frame is dropped is not altered by the scaling we perform, we expect the frame drop ratio to remain almost invariant as we vary α. Further, since the frame drop ratio and service time are both independent of α, so is the user's perceived throughput.

IV. The Non Saturated Case

We now assume that frames arrive at the interface transmission queue of each source in an α-scaled system, according to some arbitrary arrival process at a rate of λ frames per unit of time. The average frame arrival rate per station per time-slot in an α-scaled system is therefore $\lambda \text{SlotTime}_\alpha = \lambda \frac{\text{SlotTime}_1}{\alpha}$, and the aggregate arrival rate (from all stations) is

$\lambda_\alpha^{tot} = \alpha N \lambda \frac{\text{SlotTime}_1}{\alpha} = N\lambda\text{SlotTime}_1 = \lambda_1^{tot}$ (i.e. independent of α). [5]

Now, assume that there are no frame collisions, and let C_α be the service rate of an α-scaled system, i.e. the total number of frames that the system can transmit per unit of time. Since we scale the transmission speed of every node in the system by α, and the SIFS and DIFS durations by $\frac{1}{\alpha}$, we are essentially speeding up the service rate of the system by the factor α. Therefore, $C_\alpha = \alpha C_1$. The total number of frames that an α-scaled system can transmit per time-slot, μ_α^{tot}, is $\mu_\alpha^{tot} = C_\alpha\text{SlotTime}_\alpha = \alpha C_1 \frac{\text{SlotTime}_1}{\alpha} = \mu_1^{tot}$ (i.e. also independent of α.)

Accounting for collisions, the *effective* service rate (per time-slot) of an α-scaled system is $\mu_\alpha^{toteff} = (1-p_\alpha)\mu_\alpha^{tot} = (1-p_\alpha)\mu_1^{tot}$, where p_α denotes the collision probability as before. An arriving frame is backlogged if at the instant of arrival the system is non-empty. The probability that the system is empty when an arbitrary arrival occurs is:

$$\pi_\alpha^0 = 1 - \frac{\lambda_\alpha^{tot}}{\mu_\alpha^{toteff}} = 1 - \frac{\lambda_1^{tot}}{(1-p_\alpha)\mu_1^{tot}}.$$

Therefore, an arbitrary arriving frame is transmitted immediately (after a DIFS$_\alpha$ time interval) with probability π_α^0, in which case the contention window size is 0, and with probability $1-\pi_\alpha^0$ it is backlogged, in which case the corresponding average contention window is given by Equation (1). Thus, the average contention window size is now given by:

$$\overline{W_\alpha} = (1-\pi_\alpha^0)\frac{1-p_\alpha-p_\alpha(2p_\alpha)^{m_\alpha}}{1-2p_\alpha}\frac{\alpha CW_{min}}{2}$$
$$= \frac{\lambda_1^{tot}}{(1-p_\alpha)\mu_1^{tot}}\frac{1-p_\alpha-p_\alpha(2p_\alpha)^{m_1}}{1-2p_\alpha}\frac{\alpha CW_{min}}{2}.$$

And, the collision probability is:

$$p_\alpha = 1-(1-\frac{1}{\overline{W_\alpha}})^{\alpha N-1}$$
$$= 1-(1-\frac{2(1-2p_\alpha)(1-p_\alpha)\mu_1^{tot}}{\alpha CW_{min}(1-p_\alpha-p_\alpha(2p_\alpha)^{m_1})\lambda_1^{tot}})^{\alpha N-1}.$$

By performing the same approximations as in the saturated case scenario, it is easy to show again that $p_1 \approx p_\alpha$ and $W_\alpha \approx \alpha W_1$. Notice that this also means that $\mu_\alpha^{toteff} = \mu_1^{toteff}$ and $\pi_\alpha^0 = \pi_1^0$. Further, it is also easy to see (given the description of the MAC protocol in Section I) that Equation (4) now becomes:

$$T_\alpha = \pi_1^0(\frac{T_1^{frame}}{\alpha}+\frac{\text{DIFS}_1}{\alpha})+(1-\pi_1^0)\left(\sum_{k=0}^{K_{max}} T_\alpha^k p_1^k(1-p_1)\right), \tag{7}$$

where T_α^k is given by Equation (3), as before. In addition, as with p_α, we can show again that $q_\alpha^c \approx q_1^c$ and $q_\alpha^{sc} \approx q_1^{sc}$, and therefore, that Equation (6) holds here as well. Thus, we can conclude that all theorems and corollaries that we have stated for the saturated case, hold for the non-saturated case as well. In addition, for the non-saturated case, we can also state the following lemma:

Lemma 1: Under the scaling we perform, as the scaling factor α increases, the queueing delay of a frame (i.e. its waiting time in the interface transmission queue) initially decreases and then its distribution converges to a distribution that is independent of α.

Proof: Since the frame arrival process at each station remains the same as we scale the system, and the frame service time initially decreases, the queueing delay also decreases. Since the frame service time distribution becomes independent of α, the queueing delay distribution also becomes independent of α. ■

V. Simulations

In this section we perform experiments with the ns-2 simulator [19] in order to verify our theoretical arguments. The ns-2 simulator provides one of the most accurate IEEE 802.11 MAC layer implementations [19], and it is perhaps the most popular simulator for wireless network performance evaluation.

We consider two scenarios that yield a qualitatively different behavior. In both scenarios, packets are generated at each source according to a Poisson process and are destined to the base station. (Similar results hold for any other packet arrival process.) In the first scenario, the packet arrival/generation rate at each source is 100packets/sec. This corresponds to a scenario where there is high contention. In the second scenario, this rate is 67packets/sec. This corresponds to a scenario where there is low contention. In both cases the packet size is 256bytes. The initial number of sources/stations is $N = 4$, and we show results for $4, 8, 16, 32, 64$ and 128 sources, i.e. when $\alpha = 1, 2, 4, 8, 16$ and 32 respectively. In both scenarios we scale the system's parameters as described before. (The initial values for the system's parameters, i.e. before performing any scaling, are the ones used by default in the ns-2 simulator.)

Figure 1 refers to the first scenario and shows how the packet drop ratio, the source throughput, and the average packet delay (including *both* queueing and service time), behave as we vary the number of sources. Figure 2 does the same for the second scenario.

From Figure 1 we observe that the system's performance remains almost invariant as the number of sources increases. And, this is the case for all performance metrics we consider, even for small values of α. This is expected for the drop ratio, and hence for the source's throughput, according to our earlier theoretical arguments. Notice that an N as small as 4 is sufficient to invoke the approximations of Theorem 1. (Also, notice that since the drop ratio is around 34%, the source's throughput is around its expected value, which is 100packets/sec × 256bytes/packet × $(1-0.34) \approx$ 17000bytes/sec.) The reason that the delay remains almost invariant for even small α's as well, is because there is high contention. In this case, since the collision probability is pretty

[5] Notice that we assume homogeneous sources, i.e. sources with the same arrival rate λ. In the case of non-homogeneous sources, if the arrival rate of each source when $\alpha = 1$ is λ_i, $i \in (1...N)$, the aggregate arrival rate is $\lambda_1^{tot} = \sum_{i=1}^N \lambda_i\text{SlotTime}_1$. To have $\lambda_\alpha^{tot} = \lambda_1^{tot}$, we can assume, for example, that the number of sources of each rate λ_i is scaled by α, $\forall i \in (1...N)$.

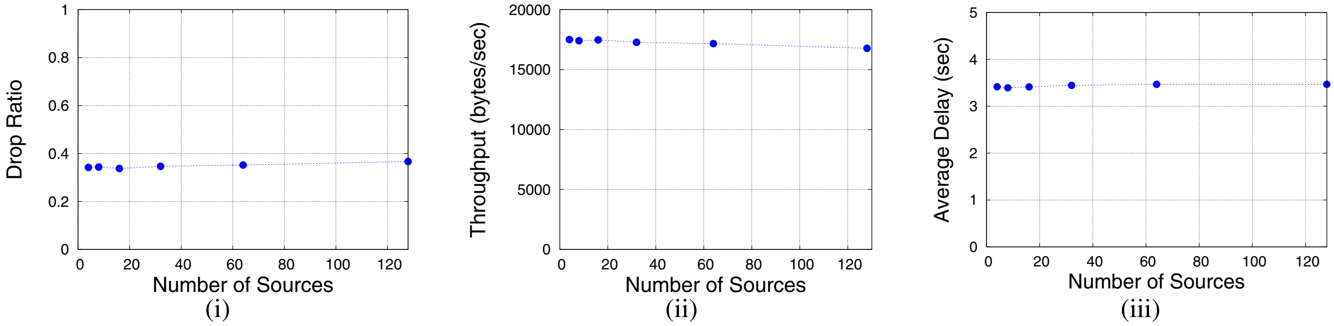

Fig. 1. (i) Drop ratio, (ii) source throughput, and (iii) average packet delay, as a function of the number of sources. (Scenario 1.)

high and the sum in Equation (3) dominates (when there is a large number of collisions) the rest of the relation's terms that depend on α, the one portion of the frame service time that is given by Equation (7) is virtually independent of α. Further, as mentioned earlier, the other portion of the frame service time, which is given by Equation (6) is also approximately independent of α (even for small α's).

Figure 2 shows again that the packet drop ratio and source throughput remain almost invariant as we increase the number of sources, as expected. (Notice that in this case the drop ratio is pretty low, around 5%, and the source's throughput is again around its expected value, which is 67packets/sec $\times$ 256bytes/packet $\times$ $(1 - 0.05)$ $\approx$ 16300bytes/sec.) However, in contrast to the previous scenario, here we first observe a notable decrease in the average packet delay. Further, we also observe that this notable decrease stops after the number of sources becomes 16, which corresponds to $\alpha = 4$. This is because, in this case, there is lower contention and the terms that depend on α have a greater influence on the frame service time, especially when α is not large, as we can deduce from our theoretical analysis. However, as we observe, convergence is taking place quite fast.

Therefore, both of the above scenarios are in agreement with our theoretical arguments and the approximations that took place.

VI. Discussion and Future Work

In this paper we have studied some important scaling properties of single-hop IEEE 802.11 wireless networks. In particular, we have identified a set of system's parameters that we should aim to scale as the number of users sharing the channel increases, in order *not* to degrade each individual user's perceived performance. We have supported our results using both rigorous theoretical analysis and ns-2 simulations.

A natural question to ask is how easy it is to scale the system parameters that we have identified in this paper, in practice. Clearly, one can easily scale the minimum and maximum contention window sizes (CW_{min} and CW_{max}) of the IEEE 802.11 protocol, however the transmission speed C as well as all the IEEE 802.11 protocol's time-intervals (i.e. the SIFS, DIFS, and SlotTime durations) that we also wish to scale, depend on the hardware technology that is being used. Therefore, to be able to support a large number of users in currently deployed and future IEEE 802.11 networks, we believe that we should aim in developing technology that will allow the scaling of these parameters by the desired factors. Notice that developing technology that allows the scaling of some of the parameters that we have identified in this paper, has been the trend for increasing the capacity of these networks. For example, C = 11Mbps for IEEE 802.11b [20], whereas C = 54Mbps for IEEE 802.11g [21]. Further, SlotTime $= 20\mu$sec and SIFS $= 10\mu$sec in IEEE 802.11b, whereas SlotTime $= 9\mu$sec and SIFS $= 5\mu$sec in IEEE 802.11g. In this paper, we have rigorously established the exact amount of scaling that it is required for the system parameters, in order to preserve individual user performance as the total number of users increases. Interestingly enough, we have found that a scaling factor, which is equal to the factor by which the number of users increases, is sufficient to preserve performance.

However, note that it may not be possible to have arbitrarily large scaling factors. For example, in the IEEE 802.11 specifications [20], [21], the SlotTime duration should be larger than the sum of the MAC-layer processing time and the air propagation time. Therefore, since we can only improve the processing time, the SlotTime duration cannot get smaller than the air propagation time ($<< 1\mu$sec), which immediately gives an upper bound on the scaling factor that we could ever have. How close to this upper bound can we get, is an interesting open question.

Further, it is interesting to point out that in this paper we were starting from smaller networks and moving to larger networks (i.e. we have studied the network behavior as the number of users increases). One can also move the opposite direction, i.e. downscale larger networks. It is easy to see that as long as the downscaling factor $0 < \alpha < 1$ is not too small, one can accurately predict the performance of larger networks from scaled-down replicas that consist of fewer nodes. This is important for simulations and experiments with testbeds where one could experiment with network miniatures, which are much easier to manage, and have much lower computational requirements and costs.

One of the most interesting future work directions is to investigate whether similar scaling properties hold for multi-hop wireless networks, which can be either static or mobile.

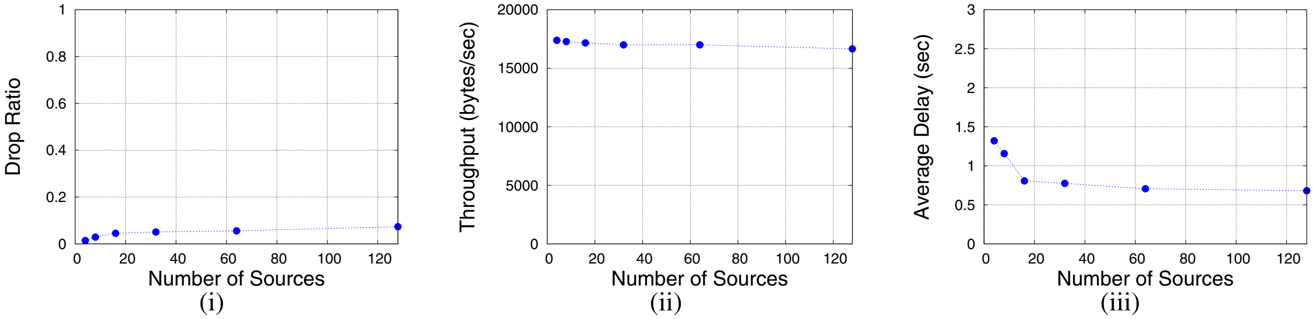

Fig. 2. (i) Drop ratio, (ii) source throughput, and (iii) average packet delay, as a function of the number of sources. (Scenario 2.)

References

[1] "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE standards 802.11, January 1997 (http://standards.ieee.org/getieee802/802.11.html)," .

[2] G. Bianchi, "Performance analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on selected areas in communications*, vol. 18, no. 3, 2000.

[3] O. Tickoo and B. Sikdar, "Queueing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks," in *Proceedings of IEEE INFOCOM*, 2004.

[4] H. L. Vu and T. Sakurai, "Accurate delay distribution for IEEE 802.11 DCF," *IEEE Communications Letters*, vol. 10, no. 4, 2006.

[5] G. Sharma, A. Ganesh, and P. Key, "Performance analysis of contention based medium access control protocols," in *Proceedings of IEEE INFOCOM*, 2006.

[6] A. Lindgren and A. Almquist, "Quality of Service schemes for IEEE 802.11 - a simulation study," M.S. thesis, Lule University of Technology, May 2001.

[7] L. Scalia and I. Tinnirello, "A Low-level Simulation Study of Prioritization in IEEE 802.11e Contention-based Networks," in *Proceedings of IEEE COSMWARE*, 2006.

[8] G. Anastasi, E. Borgia, M. Conti, and E. Gregory, "IEEE 802.11b Ad Hoc Networks: Performance Measurements," *Journal of Cluster Computing, Springer*, vol. 8, no. 2–3, 2005.

[9] G. Anastasi, E. Borgia, M. Conti, and E. Gregory, "Wi-fi in ad hoc mode: a measurement study," in *IEEE PerCom*, 2004.

[10] E. Pelletta and H. Velayos, "Performance Measurements of the Saturation Throughput in IEEE 802.11 Access Points," in *WiOpt*, 2005.

[11] K. Psounis, R. Pan, B. Prabhakar, and D. Wischik, "The scaling hypothesis: Simplifying the prediction of network performance using scaled-down simulations," in *Proceedings of ACM HOTNETS*, 2002.

[12] R. Pan, B. Prabhakar, K. Psounis, and D. Wischik, "Shrink: Enabling scaleable performance prediction and efficient simulation of networks," *IEEE/ACM Transactions on Networking*, October 2005.

[13] F. Papadopoulos, K. Psounis, and R. Govindan, "Performance preserving topological downscaling of internet-like networks," *IEEE Journal on Selected Areas in Communications (JSAC), Special Issue on Sampling the Internet: Techniques and Applications*, December 2006.

[14] F. Papadopoulos and K. Psounis, "Predicting the performance of mobile ad hoc networks using scaled-down replicas," in *Proc. of IEEE ICC*, June 2007.

[15] V. Naik, E. Ertin, H. Zhang, and A. Arora, "Wireless testbed Bonsai," in *Proc. of WiNMee*, 2006.

[16] P. Gupta and P.R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, 2000.

[17] A. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-off in wireless networks," in *In Proc. 2004 INFOCOM*, 2004.

[18] G. Sharma, R. R. Mazumdar, and N. B. Shroff, "Delay and Capacity Trade-off in Mobile Ad hoc Networks: A Global Perspective," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, 2007.

[19] "Network simulator," http://www.isi.edu/nsnam/ns.

[20] "Supplement to 802.11-1999,Wireless LAN MAC and PHY specifications: Higher speed Physical Layer (PHY) extension in the 2.4 GHz band (http://standards.ieee.org/getieee802/802.11.html)," .

[21] "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications–Amendment 4: Further Higher-Speed Physical Layer Extension in the 2.4 GHz Band (http://standards.ieee.org/getieee802/802.11.html)," .