Throughput Optimal Switching in Multi-channel WLANs

Qingsi Wang and Mingyan Liu

Abstract—We observe that in a multi-channel wireless system, an opportunistic channel/spectrum access scheme that solely focuses on channel quality sensing measured by received SNR may induce users to use channels that, while providing better signals, are more congested. Ultimately the notion of channel quality should include both the signal quality and the level of congestion, and a good multi-channel access scheme should take both into account in deciding which channel to use and when. Motivated by this, we focus on the congestion aspect and examine what type of dynamic channel switching schemes may result in the best system throughput performance. Specifically we derive the stability region of a multi-user multi-channel WLAN system and determine the throughput optimal channel switching scheme within a certain class of schemes. We also empirically examine the impact of considering congestion in addition to signal quality in making channel selection decisions.

Index Terms—Wireless LAN, 802.11 DCF, multi-channel wireless system, channel switching policy, stability region, throughput optimality

1 Introduction

Advances in software defined radio in recent years have motivated numerous studies on building agile, channel-aware transceivers that are capable of sensing instantaneous channel quality [1], [2], [3]. With this opportunity comes the challenge of making effective opportunistic channel access and transmission scheduling decisions, as well as designing supporting system architectures. There have been extensive studies on dynamic channel access in a multi-user, multi-channel wireless system, see e.g., [4], [5]. By allowing users to dynamically select which channel to use for transmission, these schemes aim to improve the system performance, typically measured by the total (or per user) throughput, the average packet delay and etc, compared to a system with a single channel or more static channel allocations. The main reason behind such improvement lies in temporal, spatial and spectral diversity. That is, the quality of a channel perceived by a user is time-varying, user-dependent, and channel-dependent.

Within this context we make the additional observation that there is also a *congestion diversity* in that a channel with fewer number of competing users presents better quality for a user. This is particularly true in a random access setting, where a large number of competing users can induce large backoff timer values that in turn lead to longer waiting time and lower throughput. We note that in a multi-channel system, an opportunistic access scheme that solely focuses on channel quality sensing as a result of random fading and shadowing, e.g., by measuring received SNR [6], [5], may induce users to use channels that, while providing better signals, are more congested. This can reduce the expected performance gain, or even turn gain to loss. Ultimately the notion of "channel quality" should include both the signal quality and the level of congestion, and a good multi-channel access scheme should take both into account in deciding which channel to use and when.

Motivated by the above, in this study we examine the possibility of utilizing congestion diversity to promote certain performance measures, e.g., throughput. As mentioned above, our ultimate goal is to construct an opportunistic channel access scheme that is aware of both signal quality and user congestion. However, in the present paper we will primarily limit our attention to addressing the congestion aspect only. We do provide numerical results on the impact of considering congestion in addition to signal quality in making channel selection decisions.

Specifically, we ask the question of what type of dynamic channel switching schemes will give the best performance in a multi-channel WLAN. This will be evaluated using the notion of stability region of a scheme. This is because more effective resource allocation and sharing can achieve a lower overall congestion level, thus expanding the range of sustainable arrival rates and resulting in a larger stability region. The scheme with the largest such region is commonly known as the throughput optimal scheme. With this objective, we set out to study the stability region of a multi-channel WLAN system where users are allowed to dynamically switch between channels.

The main contributions of this paper are as follows.

- A mean-field based model is constructed to characterize the stability region of a multi-channel

- *Q. Wang and M. Liu are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109.*

This work is supported by NSF grant CIF-0910765, ARO grant W911NF-11-1-0532, and NASA grant NNX09AE91G.

WLAN system. We show that the size of the backoff window plays a decisive role in shaping the corresponding stability region: when the backoff window is sufficiently large, the stability region is convex; as the window size decreases it evolves into a concave region.
- Using this mean-field model, we provide an analytical justification of using channel-switching policies that achieve load balance in systems with symmetric channels. This is then extended to systems with asymmetric channels.
- We propose several simple heuristic implementations of the channel-switching policies presented in this paper.

802.11 DCF has been very extensively studied in the literature, ranging from throughput performance in the saturated regime [7], [8] and the non-saturated regime [9], [10], to its rate region [11], [12], to channel assignment in multi-channel WLANs [13], [14], to name a few. To the best of our knowledge, however, none has studied multi-channel WLAN in the context of stability region. Works most relevant to ours include ones on the stability region of slotted Aloha (e.g., [15]) and the rate region of 802.11 DCF [11], [12].

In the remainder of the paper, we first introduce a system of equations to characterize the stability region of a single channel WLAN consisting of multiple users within a single interference domain (Section 3) followed by numerical results (Section 4). We then extend the same method to characterize the stability region of a multi-channel system and use this result to determine the throughput optimal channel switching schemes within a class (Section 5). We also discuss how such schemes may be implemented in practice (Section 6). Due to the space limit, some technical details are omitted and may be found in our technical report [16].

2 System Model and Preliminaries

Consider a multiple access system using the IEEE 802.11 DCF. There are N nodes (or users interchangeably), indexed by the set $\mathcal{N} = \{1, 2, \ldots, N\}$, each with an infinite buffer, one transceiver (i.e., a single wireless interface) and uses the same parameterization. We assume the channel is ideal and there is no MAC-level packet discard, i.e., there is no retransmission limit of a packet after collision. Throughout the analysis we also adopt a few other simplifying assumptions to make the problem tractable; these will be stated in the context to which they apply. It should be noted that due to the complexity of the problem, successive simplification in the modeling effort is a rather common practice and has been used in most if not all previous works. We later show that these simplifications do not impact the accuracy of the model under normal operating parameter values.

The key to our method is to model the queue at each node with a service process defined by 802.11 DCF as a *slotted mean field Markov chain* [17].

Definition 1: A *virtual backoff timer* of the system (or of a virtual node) is a universal timer for all nodes in the system: it counts down indefinitely, alternating between the count-down mode (when nodes in the system are counting down) and the freezing mode (when some node in the system is transmitting). The slot time is thus a random variable.

Remark 1: The above definition provides a universal slot time for all nodes in the system, and we will assume that the real backoff timer at each node is synchronized to this virtual timer on slot boundaries. The motivation behind such a construction originates from the principal difficulty in modeling a non-saturated system: the service process at each node runs in embedded time in terms of a slot, which is a random variable, whereas the packet arrival process is more naturally described in real time [17]. This difficulty does not exist in saturated analysis, see e.g., [7], where arrival processes do not play a role.

We next introduce three key assumptions in our model, followed by a discussion on their implications and limitations.

(A1) The MAC layer arrival process at node i is Poisson with rate λ_i bits per second.

(A2) The service time of a packet, i.e. the time from the initial backoff to successful transmission, is (i) exponential with service rate μ_i at node i, and (ii) independent of all arrival processes.

(A3) Let $S(t)$ be the counting process of the number of slots accumulated up to time t. $S(t)$ is assumed to be (i) independent of $Q_i(t)$, and (ii) renewal.

Denote by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$, and by $\{Q_i(t)\}_t$ the queueing process at node i (also written as $Q_i(t)$ for simplicity), i.e., the number of packets in node i's MAC queue at time t. We now formally define the stability region of system as follows.

Definition 2: The *stability region* Λ is the set of all $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ such that $Q_i(t)$ admits a stationary distribution for all i with arrival rates $\boldsymbol{\lambda}$ under the 802.11 DCF scheme.

The above simplifying assumptions are not entirely realistic. Typically, due to congestion control by upper-layer protocols, e.g., TCP, the arrival process to the MAC layer is neither Poisson nor independent of the service process. However, as our objective is to explore the inherent properties of 802.11 DCF, the independence assumption is adopted to decouple the MAC layer from upper layers, while the Poisson and exponential assumptions are adopted to avoid technicalities that can obscure the main insight. Note that under the mean field methodology, each node is analyzed in isolation from the activities of all other nodes which are collectively regarded as an aggregate stationary process. Within such a framework the

packet service time is taken to be stationary (see e.g., Bianchi's well-known mean field Markovian model of the service process [7]).

With **A1** and **A2**, $Q_i(t)$ is then a well-defined $M/M/1$ queue, and for a given $\boldsymbol{\lambda}$, $\boldsymbol{\lambda} \in \Lambda$ if and only if $Q_i(t)$ is positive recurrent. Equivalently we may consider the utilization factor ρ_i at node i, given by $\rho_i = \min\{\frac{\lambda_i}{\mu_i}, 1\}$: the queue is stable if and only if $\rho_i < 1$. If $Q_i(t)$ is positive recurrent, then it is ergodic and we have $\lim_{t\to\infty} P(Q_i(t) > 0) = 1 - \pi_i(0) = \rho_i$, where π_i is the stationary distribution of $Q_i(t)$. If $Q_i(t)$ is transient or null recurrent, in which case $\rho_i = 1$, we have $\lim_{t\to\infty} P(Q_i(t) = 0) = 0 = 1 - \rho_i$. Therefore, ρ_i is asymptotically given by $\lim_{t\to\infty} P(Q_i(t) > 0)$ in all cases in our model.

For technical reasons we will also consider the *embedded* queueing process $\hat{Q}_i(n)$, $n = 1, 2, \cdots$, defined as $\hat{Q}_i(n) := Q_i(T_n)$, where T_n is the time of the nth slot boundary. $\hat{Q}_i(n)$ is thus a discrete-time process constructed by observing $Q_i(t)$ at slot boundaries.

For an arbitrary process $S(t)$, $\hat{Q}_i(n)$ is not necessarily Markovian. However, given assumption **A3**, durations between slot boundaries are i.i.d., constituting sampling periods that are independent of $Q_i(t)$. Hence $\hat{Q}_i(t)$ is a discrete-time Markov chain under our assumption. It's worth noting that **A3** does not exactly hold in reality because the slot length is a function of a node's activity, and thus the state of its queue, even with the mean field simplification of other nodes' behavior (this is more precisely shown in the appendix). However, this dependence weakens when the number of nodes or the backoff window size is sufficiently large. We empirically show that this assumption does not impact the accuracy of prediction even with a small node population and backoff window size.

Let $\hat{\rho}_i$ denote the utilization factor under the discrete-time system $\hat{Q}_i(n)$. In general $\hat{\rho}_i \neq \rho_i$. Indeed we show in Appendix A that $\hat{\rho}_i \leq \rho_i$ where equality holds if and only if $\rho_i = 1$ or $\rho_i = 0$, i.e., node i is either saturated or idle. Similar to ρ_i, $\hat{\rho}_i$ is asymptotically given by $\lim_{n\to\infty} P(\hat{Q}_i(n) > 0)$.

We will adopt Bianchi's decoupling approximation [7] as another key assumption, stated as follows. Define $C_i(j) := 1$ if the jth attempt by node i results in a collision, and $C_i(j) := 0$ if it results in a success.

(A4) [Bianchi's Decoupling Approximation] For each node $i \in \mathcal{N}$, the collision sequence $\{C_i(j)\}$ is i.i.d. with $P(C_i(j) = 1) = p_i$ for some constant p_i.

In reality successive attempts by the same node may occur if it repeatedly selects timer value 0 while other nodes' timers remain frozen. In such cases the above assumption ceases to hold. This phenomenon can be prominent when the window size is small, and has been taken into account in some recent work [18]. In this study we will ignore the possibility of successive attempts for simplicity of presentation and adopt **(A4)**. (A more precise model is possible by imposing independence not on all attempts but only the first attempt in each such sequence.) This is reasonable when the initial window size is sufficiently large. Our empirical results are fairly close between with and without consideration of successive attempts for large backoff windows. For small backoff windows, the discrepancy between the two will be illustrated in the numerical results.

We will use the term *backoff length* to mean the total number of slots that a node spends between two successive timer renewals during the service process, which is the selected timer value plus 1. Define N_i^s and N_i^{tx}, respectively, as the numbers of slots and transmission attempts that node i takes in serving one packet. $\overline{W}_i := \frac{\mathbb{E}[N_i^s]}{\mathbb{E}[N_i^{tx}]}$ is referred to as the average backoff length of node i.

Using Bianchi's approximation, we have

$$\begin{aligned}\mathbb{E}[N_i^s] &= \sum_{k=0}^{\infty}\sum_{j=0}^{k} \frac{2^{\min\{j,m\}}W + 1}{2}(p_i)^k(1-p_i) \\ &= \sum_{j=0}^{\infty} \frac{2^{\min\{j,m\}}W + 1}{2}\left(\sum_{k=j}^{\infty}(p_i)^k(1-p_i)\right) \\ &= \sum_{j=0}^{\infty} \frac{2^{\min\{j,m\}}W + 1}{2}(p_i)^j\end{aligned}$$

where W is the size of the initial backoff window and m is the value of the maximum backoff stage. Also note $\mathbb{E}[N_i^{tx}] = \frac{1}{1-p_i}$. Therefore, $\overline{W}_i$ is given by

$$\overline{W}_i = \frac{1}{2}\left[W\left((1-p_i)\sum_{j=0}^{m-1}(2p_i)^j + (2p_i)^m\right) + 1\right].$$

We next derive a relationship between the transmission attempt probability and $\hat{\rho}_i$. Let $\tau_i(n)$ be the probability that node i initiates a transmission attempt in the nth slot.

Lemma 1: $\tau_i := \lim_{n\to\infty} \tau_i(n)$ exists and is given by $\tau_i = \hat{\rho}_i/\overline{W}_i$.

Proof: Let $T_X(n)$ denote the event that node i initiates an attempt in the nth slot. Then

$$\begin{aligned}\tau_i(n) = P(T_X(n)|\hat{Q}_i(n) > 0) \cdot P(\hat{Q}_i(n) > 0) + \\ + P(T_X(n)|\hat{Q}_i(n) = 0) \cdot P(\hat{Q}_i(n) = 0).\end{aligned}$$

Consider now the sequence of slots in which node i has a packet in service. Given the decoupling among nodes, the occurrences of slots in which node i starts the service for a packet thus form renewal events. Regarding each transmission attempt as one-unit reward and using the renewal reward theory, we then obtain

$$\lim_{n\to\infty} P(T_X(n)|\hat{Q}_i(n) > 0) = \frac{\mathbb{E}[N_i^{tx}]}{\mathbb{E}[N_i^s]} = \frac{1}{\overline{W}_i}.^1$$

1. We note that [19] used a similar technique in computing the conditional transmission probability defined therein.

Since $\lim_{n\to\infty} P(\hat{Q}_i(n) > 0) = \hat{\rho}_i$, and $P(T_X(n)|\hat{Q}_i(n) = 0) = 0$, the result follows. □

To put the above result in context, one easily verifies that in the extreme case where all nodes are saturated and identical, we have $\hat{\rho}_i = \rho_i = \rho = 1$ and $p_i = p$ for all i. Consequently,

$$\tau_i = \tau = \frac{2}{W\left((1-p)\sum_{j=0}^{m-1}(2p)^j + (2p)^m\right) + 1} = \frac{2(1-2p)}{(1-2p)(W+1) + pW(1-(2p)^m)},$$

which is exactly the same as obtained in [7] Eqn (7).

3 Single Channel Stability Region

3.1 The stability region equation Σ

Our first main result is the following theorem on the quantitative description of Λ. Let $\mathbb{E}[S_{i,Q,\overline{T}_x}]$ denote the conditional average length of a slot given that the queue at node i is non-empty but i does not transmit in this slot. T_s and T_c denote the lengths of a successful transmission and a collision, respectively.

Theorem 1: $\boldsymbol{\lambda} \in \Lambda$ if and only if there exists at least one solution $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_N)$ to the following constrained system of equations $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$:

$$\Sigma: \begin{cases} \tau_i = \dfrac{\hat{\rho}_i}{\overline{W}_i}, \forall i & \text{(a)} \\ p_i = 1 - \prod_{j\neq i}(1-\tau_j), \forall i & \text{(b)} \\ \rho_i = \min\left\{\dfrac{\lambda_i}{P}\left(\dfrac{\overline{W}_i - 1}{1-p_i}\mathbb{E}[S_{i,Q,\overline{T}_x}] + \right.\right. & \\ \left.\left. \quad + T_c\dfrac{p_i}{1-p_i} + T_s\right), 1\right\}, \forall i & \text{(c)} \end{cases}$$

subject to

$$\mathrm{C}: \begin{cases} 0 \le \tau_i \le 1, \forall i & \text{(i)} \\ 0 \le \rho_i < 1, \forall i & \text{(ii)} \end{cases}$$

where P is the packet payload size.

Proof: Σ(a) is the result of Lemma 1, and Σ(b) is an immediate consequence of the definition of p_i. Let the average service time at node i be $\overline{X}_i$ seconds per bit; the average service time per packet is thus $P\overline{X}_i$. Define $\overline{Y}_i(j)$ as

$$\overline{Y}_i(j) = T_c + \left(\frac{2^{\min\{j,m\}}W + 1}{2} - 1\right)\mathbb{E}[S_{i,Q,\overline{T}_x}].$$

Physically, $\overline{Y}_i(j)$ is the average time between the beginning of the jth transmission attempt, which results in a collision, and the beginning of the $(j+1)$th attempt, given that node i encounters at least j collisions before completing the service of some packet. Since the collision sequence is geometric, we have

$$\begin{aligned} P\overline{X}_i &= \sum_{k=0}^{\infty}\left[\left(\frac{W+1}{2} - 1\right)\mathbb{E}[S_{i,Q,\overline{T}_x}] + \sum_{j=1}^{k}\overline{Y}_i(j) + \right. \\ &\quad \left. + T_s\right] \times (p_i)^k(1-p_i) \\ &= \sum_{j=1}^{\infty}\sum_{k=j}^{\infty}\overline{Y}_i(j)(p_i)^k(1-p_i) + \left(\frac{W+1}{2} - 1\right)\times \\ &\quad \times \mathbb{E}[S_{i,Q,\overline{T}_x}] + T_s \\ &= \sum_{j=1}^{\infty}(p_i)^j\overline{Y}_i(j) + \left(\frac{W+1}{2} - 1\right)\mathbb{E}[S_{i,Q,\overline{T}_x}] + T_s. \end{aligned}$$

Therefore,

$$\begin{aligned} P\overline{X}_i &= \sum_{j=1}^{\infty}\left[(p_i)^j\left(T_c + \left(\frac{2^{\min\{j,m\}}W+1}{2} - 1\right)\times\right.\right. \\ &\quad \left.\left. \times \mathbb{E}[S_{i,Q,\overline{T}_x}]\right)\right] + \left(\frac{W+1}{2} - 1\right)\mathbb{E}[S_{i,Q,\overline{T}_x}] + T_s \\ &= \sum_{j=0}^{\infty}\left[\frac{2^{\min\{j,m\}}W - 1}{2}(p_i)^j\right]\mathbb{E}[S_{i,Q,\overline{T}_x}] + \\ &\quad + T_c\sum_{j=1}^{\infty}(p_i)^j + T_s \\ &= \frac{\overline{W}_i - 1}{1-p_i}\mathbb{E}[S_{i,Q,\overline{T}_x}] + T_c\frac{p_i}{1-p_i} + T_s. \end{aligned}$$

Note that $\tau_i < 1$ for all i, and we have $p_i < 1$ for all i as a result. In addition, $\mathbb{E}[S_{i,Q,\overline{T}_x}]$ is finite (computed in the appendix). Hence we conclude that the packet service time is finite. Thus, the utilization factor of node i is given by $\rho_i = \min\{\lambda_i\overline{X}_i, 1\}$ and Σ(c) follows. C(i) is for the validity of τ as a probability measure. $(\Sigma$, C(i), $\boldsymbol{\lambda})$ then constitutes a full description on the system utilization. C(ii) is the necessary and sufficient condition for stability as commented in the previous section. □

For a given set of system parameter values, two sets of quantities are needed to compute Σ: $\mathbb{E}[S_{i,Q,\overline{T}_x}]$ and $\hat{\rho}_i$, $\forall i \in \mathcal{N}$. These are computed in Appendix B and C, respectively. In particular, in Appendix C we show that though it is analytically intractable, $\hat{\rho}_i$ is well approximated by

$$\hat{\rho}_i \approx \frac{\rho_i\mathbb{E}[S_{i,\overline{Q}}]}{\rho_i\mathbb{E}[S_{i,\overline{Q}}] + (1-\rho_i)\mathbb{E}[S_{i,Q}]},$$

where $\mathbb{E}[S_{i,Q}]$ (resp. $\mathbb{E}[S_{i,\overline{Q}}]$) is the conditional average length of a slot given that the queue at node i is non-empty (resp. empty) at the beginning of this slot.

3.2 Characterizing the solutions to Σ

Without the stability constraint C(ii), $(\Sigma$, C(i), $\boldsymbol{\lambda})$ can be rewritten as a vector equation in $[0,1]^N$, $\boldsymbol{\tau} = \boldsymbol{\Gamma}(\boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_N) \in [0,1]^N$, and the existence of solutions can be shown by Brouwer's fixed point theorem. However, the uniqueness of its solution is in general difficult to prove; nevertheless, under the condition of a sufficiently large initial backoff window W, we have the following result on the uniqueness of its solution.

With a large initial backoff window W, the probability of collision is small, so we have $\overline{W}_i \approx \frac{W+1}{2}$. We also observe that $\mathbb{E}[S_{i,Q}] \approx \mathbb{E}[S_{i,\overline{Q}}]$ when W is large (cf. Appendix B). Consequently, we can approximate $\hat{\rho}_i$ by ρ_i. Also, using the first-order Taylor approximation, we have $\prod_{j\neq i}\frac{1}{1-\tau_j} \approx 1+\sum_{j\neq i}\tau_j$ for small τ. Note that the minimization operator in Σ is redundant when combined with C(ii). Hence, let $T_s = T_c = T$ for simplicity of presentation, and $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$ can be then approximated by the following constrained system of equations,

$$\widetilde{\Sigma}: \begin{cases} \tau_i = \frac{\rho_i}{\frac{W+1}{2}}, \forall i & \text{(a)} \\ \rho_i = \frac{\lambda_i}{P}\left[\frac{W-1}{2}\left(\sigma + T\sum_{j\neq i}\tau_i\right) + \right. \\ \qquad \left. + T\left(1+\sum_{j\neq i}\tau_i\right)\right], \forall i & \text{(b)} \end{cases}$$

subject to the same set of constraints.

Proposition 1: $(\widetilde{\Sigma}, \boldsymbol{\lambda})$ admits a unique solution.

Proof: See Appendix D. □

Remark 2: 1) The above result suggests that Σ has a unique solution when W, the initial window size, is sufficient large. As an approximation we will take this condition to be equivalent to a large average backoff window. This is because the probability of a (first-attempt) collision decays inverse-linearly in W, and thus $\overline{W}_i$ is dominated by W when W is sufficiently large.

2) As we will see numerically in the next section, multiple fixed point solutions may arise when W is small; this will be referred to as multi-equilibrium (as opposed to "multistable" or "metastable" [17] to avoid confusion).

In the proof of Proposition 1, we in fact obtained the approximated unique solution to $(\Sigma, \boldsymbol{\lambda})$. Therefore, by imposing feasibility constraints C, we can induce a simplified version of $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$ which is an approximation to Λ and is easier to compute.

Corollary 1: When W is sufficiently large, Λ is approximated by

$$\tilde{\Lambda} = \left\{ \boldsymbol{\lambda} \in \mathbb{R}_+^N \;\middle|\; 0 < \frac{\gamma_i^1(\lambda_i)\sum_j \gamma_j^2(\lambda_i)}{1-\sum_i \gamma_j^1(\lambda_i)} + \gamma_i^2(\lambda_i) < \frac{2}{W+1}, \forall i \right\}$$

where $\gamma_i^1(\lambda_i) = \frac{\lambda_i T}{P} \Big/ \left(1+\frac{\lambda_i T}{P}\right)$, and $\gamma_i^2(\lambda_i) = \frac{\lambda_i((W-1)\sigma+2T)}{P(W+1)} \Big/ \left(1+\frac{\lambda_i T}{P}\right)$.

Within the context of a unique solution to $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$, consider $\boldsymbol{\lambda}$ as input parameters and rewrite Σ as $\mathbf{F}(\boldsymbol{\tau}, \boldsymbol{\lambda}) = 0$, with $(n+n)$ unknowns (τ_i's and λ_i's). We can then inspect the existence of an implicit function of $\boldsymbol{\tau}$ in terms of $\boldsymbol{\lambda}$, and for this we need to examine the invertibility of the corresponding Jacobian matrix. Note also that the correspondence between ρ_i and $(\boldsymbol{\lambda}, \boldsymbol{\tau})$ given by Σ(c) is a continuous function. If the Jacobian is invertible on the boundary of the stability region Λ in the space $\mathbb{R}_+^N$, then the continuity of $\rho_i = \rho_i(\boldsymbol{\lambda})$ is established. Hence, on the boundary of Λ, denoted by $\partial\Lambda$, there exists at least one node i such that $\rho_i = 1$. Unfortunately, the invertibility of the Jacobian on $\partial\Lambda$ is highly non-trivial to determine and in general analytically intractable when the number of nodes is large. In the next section we numerically evaluate $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$ and more is discussed.

4 Numerical Results: Single Channel

Using $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$, we can quantitatively describe the stability region of a single channel system, and some numerical results for the two-user case are illustrated in this section. The parameters used in both the numerical computation and the simulation are reported in Table 1 in Appendix F. Under the basic access mechanism of DCF we have

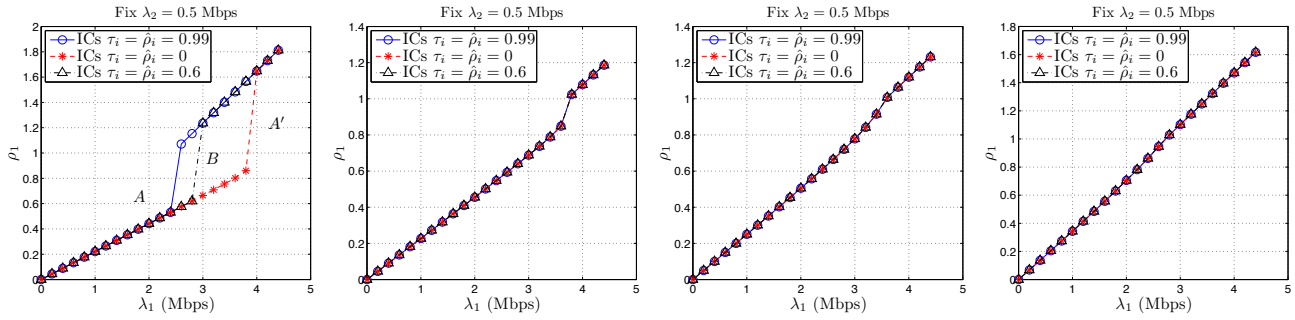
$$\begin{cases} T_s = \frac{P}{\text{Tx. Rate}} + \text{Header} + \text{ACK} + \text{DIFS} + \text{SIFS} + 2\delta \\ T_c = \frac{P}{\text{Tx. Rate}} + \text{Header} + \text{DIFS} + \delta \end{cases}$$

where δ is the propagation delay.

4.1 Multi-equilibrium and discontinuity in ρ

We first illustrate the existence of multi-equilibrium solutions and discontinuity of $\rho_i(\boldsymbol{\lambda})$ in $\boldsymbol{\lambda}$; this is shown in Figure 1. We fix the value of λ_2 and increase λ_1 from 0 to 4.5 Mbps. For each pair $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, we solve for the fixed point(s) of $\boldsymbol{\Sigma}$ with the same set of initial values of τ_i and $\hat{\rho}_i$ for $i \in \mathcal{N}$ to which we refer as a set of initial conditions (ICs). We then convert the results to $\boldsymbol{\rho} = (\rho_i, i \in \mathcal{N})$ using Eqn. Σ(c). The collection of the pairs $(\boldsymbol{\lambda}, \boldsymbol{\rho}(\boldsymbol{\lambda}))$ then constitutes a *solution component* for this set of ICs. Note that this is obtained by solving (Σ, C(i), $\boldsymbol{\lambda}$) without considering the stability constraint C(ii). We repeat the above computation for different sets of ICs under the same system parameters including W and m. The entire process is then repeated for different pairs (W, m). For each pair (W, m), the resulting solution components constitute an overall correspondence between the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}(\boldsymbol{\lambda})$, and this is plotted for ρ_1 vs. λ_1 in Figure 1.

In the first scenario as shown in Figure 1(a), where the initial window is of the smallest possible size for two users and window expansion is disallowed ($m = 0$), three different zones of the correspondence $\rho_1(\lambda_1)$ are present, labeled as A, A' and B in the figure. In zones A and A', a single fixed point is admitted and $\rho_1(\lambda_1)$ reduces to a function, while in zone B we see two solutions. Along each solution component, there is a jump in ρ_1 in zone B as λ_1 increases; this is essentially a phase transition from stable to unstable regions. What this result illustrates is that depending on the initial condition, certain input rates may or may not lead to a feasible solution (a point in the stability region). Thus when such multi-equilibrium exists, we



(a) $W = 2, m = 0$ (b) $W = 8, m = 5$ (c) $W = 32, m = 5$ (d) $W = 128, m = 5$

Fig. 1. Solution components for various scenarios: an illustration.

may have a collection of stability region Λ's given different initial conditions, and this phenomenon is illustrated in Figure 3 and discussed in the next subsection in detail. Recall that under our definition of stability region and Theorem 1, an arrival rate vector is considered within the stability region as long as there exists such an initial condition that induces so; the stability region thus defined is therefore the supremum of this collection when multiple equilibria exist. The advantage of the "stability region" A is that the points within are stable independent of the initial condition. With a slight abuse of terminology, we would later refer to this region as the stability region with multi-equilibrium.

Intuitively, initial conditions with large values suggest a pessimistic prediction on the system stability under $\boldsymbol{\lambda}$, and it may thus result in a small Λ; by contrast, ICs with small values render an optimistic one and a larger Λ. Empirically, we find that the set of ICs with $\tau_i = \rho_i \approx 1$ for $i \in \mathcal{N}$ results in the earliest jump in ρ_1 and the one with $\tau_i = \rho_i = 0$ for $i \in \mathcal{N}$ gives the latest. Consequently, solution components resulting from these two sets of ICs define the boundary of zone B and the corresponding stability regions, forming the empirical supremum and infimum of the collection of Λ's.

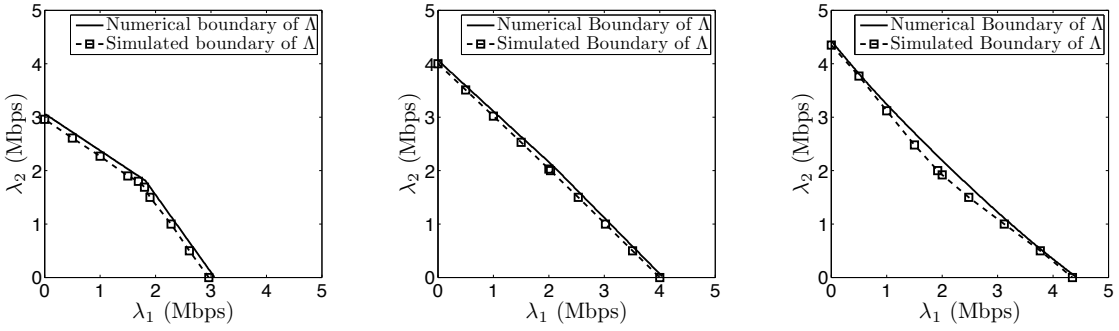
Inspecting the set of figures Fig. 1(a)-1(d), we see that as the initial window increases, the multi-equilibrium gradually vanishes and the gap in ρ_1 caused by the jump discontinuity closes.

4.2 Numerical and empirical stability regions

We numerically solve $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$ with two nodes to obtain the corresponding Λ, and then compare it with the simulated boundary. In simulation, for each fixed λ_2, we increase λ_1 with a step size $\Delta\lambda$, and compute the empirical throughput of node i obtained under $\boldsymbol{\lambda}$, denoted as $S_i^{\boldsymbol{\lambda}}$, and the number of backlogged packets at node i by the end of simulation, denoted as $B_i^{\boldsymbol{\lambda}}$. The simulator declares a point $\boldsymbol{\lambda}$ unstable if there exists at least one i such that $S_i^{\boldsymbol{\lambda}} < \lambda_i$ and $B_i^{\boldsymbol{\lambda}} P/(\lambda_i T_f) > \beta_i$, by the simulation time T_f, where β_i is an instability threshold, $0 < \beta_i < 1$. In the experiment we set $\Delta\lambda = 0.1$ Mbps (100 Kbps), $T_f = 10$ sec and $\beta_i = \beta = 1\%$. The stable point (λ_1, λ_2) such that $(\lambda_1 + \Delta\lambda, \lambda_2)$ is unstable is declared a point on the simulated boundary; the experiment is repeated for each λ_2 and the empirical mean value of λ_1 is recorded. Due to symmetry, only half of the boundary points are evaluated. The results are shown in Figure 2.

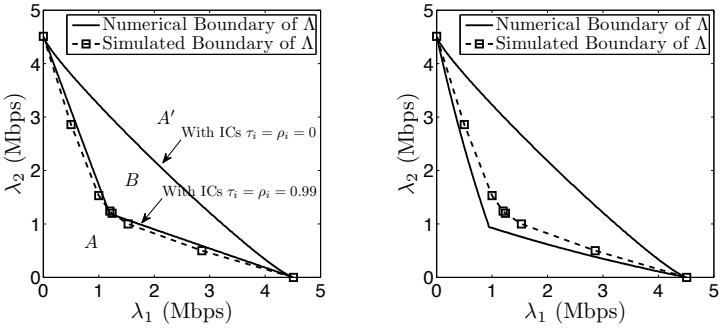
Our main observation is that when the initial (or average) backoff window is large, the stability region is convex (Figure 2(a)). The convexity gradually disappears as the window size decreases and the region is given by a near-linear boundary in Figure 2(b). It becomes clearly concave when the window size is small (Figure 2(c)). Interestingly, the case of $W = 32$ is the most frequently studied in the literature, and a linear boundary of the capacity region has been observed in [11]. As shown here, this linear boundary is only a special case in a spectrum of convex-concave boundaries. It is worth noting that in [12], Leith *et al.* established the general log-convexity of the rate region of 802.11 WLANs. This implies that the rate region could be either convex or concave, though [12] did not associate this with the window size as we have explicitly done here. It also suggests that the rate region and the stability region may be quite similar in nature; this however is not a formally proven statement, nor are we aware of such in the case of 802.11.

The change in the shape of the stability region as W changes may be explained as follows. Small W represents a highly aggressive configuration. This is much more beneficial when there is a high degree of asymmetry between the users' arrival rates. This is reflected in the concave shape of the region. When W is large, users are non-aggressive, which is more beneficial when arrival rates are similar, resulting in the convex shape. Numerically, the $W = 8$ case gives the largest stability region. This seems to suggest that the largest stability region is given by the smallest choice of W such that a unique feasible solution to $(\Sigma, \mathrm{C}, \boldsymbol{\lambda})$ exists. It would be very interesting to see if this could be established rigorously.



(a) $W = 128, m = 5$ (b) $W = 32, m = 5$ (c) $W = 8, m = 5$

Fig. 2. The stability regions in various scenarios - part I.



(a) Using Bianchi's decoupling approximation (b) Using the first-attempt decoupling approximation

Fig. 3. The stability regions in various scenarios - part II: $W = 2$ and $m = 0$.

In Figure 3, we compute the stability regions of the case where $W = 2$ and $m = 0$ for two different sets of ICs. As discussed earlier, when multi-equilibrium exists we may have a collection of stability regions. This is clearly seen in Figure 3: three different zones A, A' and B in the correspondence $\rho_1(\lambda_1)$ are mapped accordingly onto Λ. From these results, we may interpret that in zones A (A'), the system is uniformly stable (resp. unstable) regardless of the IC, while in zone B the stability of system depends on the IC. As noted in [17], the simulated boundary reflects time-averages of multiple equilibria.

As mentioned earlier, for small backoff windows the occurrence of successive attempts is non-trivial, which our model has ignored. The first-attempt decoupling approximation mentioned after **A4** captures the nodal behavior more accurately, and the adaptation of Σ using this alternative assumption is detailed in our technical report [16]. In Figure 3(b), we plot the counterpart of Figure 3(a) using the first-attempt decoupling approximation, and the discrepancy between results obtained using these two assumptions does exists. This is most notably shown in the numerical boundary A. The fact that the simulated boundary is now in between the two numerical boundaries verifies that this alternative assumption is more accurate. We do note however that for large windows this gap diminishes judging from numerical observation, which is to be expected.

4.3 Discussion: from 802.11 DCF back to Aloha

We next recall results on the stability region of slotted Aloha, the natural prototype of modern 802.11 DCF, and provide an intuitive argument on why the qualitative properties of the stability region shown in the previous section are to be expected.

In [20], Massey and Mathys studied an information theoretical model of multiaccess channel which shares several fundamental features with slotted Aloha. They investigated the Shannon capacity region of this channel with n users, which is shown to be the following subset of $\mathbb{R}_+^n$:

$$C = \left\{ \mathrm{vect}\left(\tau_i \prod_{j \neq i} (1 - \tau_j) \right) \;\middle|\; 0 \leq \tau_i \leq 1, 1 \leq i \leq n \right\},$$

where $\mathrm{vect}(v_i) = (v_1, v_2, \ldots, v_n)$, and τ_i is the transmission attempt rate of user i. In [15], Anantharam showed that the closure of the stability region of slotted Aloha is also given by C, under a geometrically distributed aggregate arrival process with parameter $1/(\sum_i \lambda_i)$ and probability $\lambda_i / \sum_j \lambda_j$ that such an arrival is at node i.

The above result on slotted Aloha can be used to explain the stability region of 802.11 DCF. Note that the main difference between the two lies in the

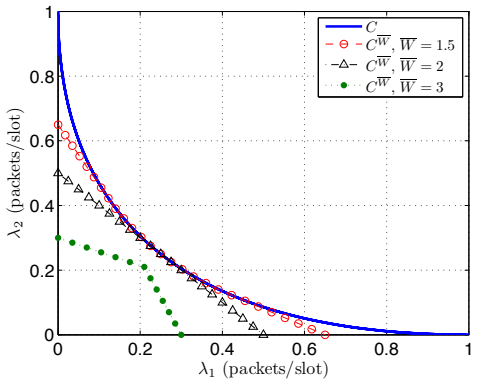


Fig. 4. The stability region of slotted ALOHA and induced subsets.

collision avoidance mechanism. Instead of attempting transmission with probability $0 \leq p \leq 1$ in a slot under slotted Aloha, under DCF each user randomly chooses a backoff timer value within a window. The effect the average backoff length $\overline{W}$ has on transmission under DCF is akin to that of restricting the attempt rate p within an upper bound $\frac{1}{\overline{W}}$ under slotted Aloha. Hence, the stability region of 802.11 DCF may be viewed as a subset of C provided that we properly scale a slot to real time.

To verify this intuition, let $C^{\overline{W}}$ be the subset of C when $0 \leq p_i \leq \frac{1}{\overline{W}}$ for all i. In Figure 4, we plot C and $C^{\overline{W}}$ with different values of $\overline{W}$. We see that as $\overline{W}$ grows, $C^{\overline{W}}$ evolves from a concave set to a convex set, consistent with what we observed of 802.11 DCF in the previous subsection. It must be pointed out that this connection, while intuitive, is not a precise one technically. For instance, this connection might suggest that the stability region of 802.11 DCF will reduce to C when the average backoff length is 1. This is however not true. In this trivial case, the stability region of 802.11 DCF is reduced to one dimension, i.e., the system is unstable for $n \geq 2$. This is because the retransmission probability of DCF is also lower bounded by the reciprocal of the window size at its backoff stage, and in the case when the backoff length is one another collision occurs with certainty.

5 Multi-channel Analysis

Using a similar, mean-field Markovian model as we did in the single channel case, we can show that the stability region of a multi-channel system under a certain switching policy $\mathbf{g}$ is given by another system of equations denoted as $(\Sigma^{\mathbf{g}}, \mathrm{C}, \boldsymbol{\lambda})$, under the arrival rates $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$, and subject to the feasibility constraints C; this is given later in the section. In addition to the same set of assumptions made in the single channel model, we assume that the system has K channels, indexed by the set $\mathcal{C} = \{1, 2, \ldots, K\}$.

The fundamental conceptual issue accompanying channelization is the notion of a channel switching policy, either centralized or distributed, that introduces channel occupancy and packet assignment distributions for each node. An additional technical issue induced by channelization is the heterogeneity of embedded time units among different channels. Since the slot length in a channel is by nature a random variable that depends on random packet arrivals, channels are in general strongly asynchronous in the embedded time units. Thus, as nodes switch among channels, we may need to switch the corresponding reference of embedded time in the slot based analysis. We therefore define the notion of a slot in different contexts as follows.

Definition 3: Consider the virtual backoff timer defined earlier separately for a *single channel*. A *channel-slot (c-slot)* is defined as the time interval between two consecutive decrements on this virtual timer for a given channel.

Definition 4: Consider a virtual backoff timer *at each node* that counts down indefinitely according to the node's backoff state, and is synchronized to the virtual timer of the channel in which the node resides and is done upon switching. A *node-slot (n-slot)* is defined as the time interval between two consecutive decrements on a given node's virtual backoff timer.

Remark 3: There is no inherent difference between the two types of slots. However, this differentiation of time references becomes crucial when we define quantities based on the random embedded time. This observation will be made more concrete in the analysis. We will also omit the explicit association of a channel (node) index with a slot whenever it does not cause ambiguity.

A channel switching or scheduling policy $\mathbf{g}$ induces a number of distributions related to $\Sigma^{\mathbf{g}}$. Denote by $\mathcal{Q}_i^n(j) = \{q_i^{(k)}(j), k \in \mathcal{C}\}$, where $q_i^{(k)}(j)$ is the probability that node i is in channel k at the beginning of its jth n-slot, t_j^-. $\mathcal{Q}_i^n(j)$ is referred to as the the channel occupancy distribution *in n-slots* of node i in the jth n-slot.

Denote by $\mathcal{Q}_i^c(j) = \{\hat{q}_i^{(k)}(j), k \in \mathcal{C}\}$, where $\hat{q}_i^{(k)}(j)$ is the probability that node i is in channel k at the beginning of its jth c-slot, $\hat{t}_j^-$. $\mathcal{Q}_i^c(j)$ is referred to as the channel occupancy *profile* of node i at the jth c-slot. Note that $\mathcal{Q}_i^c(j)$ is not necessarily a distribution and $\sum_{k\in\mathcal{C}} \hat{q}_i^{(k)}(j)$ need not be 1 for a given j.

Denote by $\mathcal{Q}_i^p(l) = \{\tilde{q}_i^{(k)}(l), k \in \mathcal{C}\}$, where $\tilde{q}_i^{(k)}(l)$ is the probability that the lth packet of node i is served in channel k, and $\mathcal{Q}_i^p(l)$ is referred to as the packet assignment distribution of node i.

We have the following assumptions on policy $\mathbf{g}$.

- **(A5)** Under $\mathbf{g}$, Bianchi's approximation is still satisfied.
- **(A6)** $\mathbf{g}$ is independent of the binary state of the queue at any node (empty vs. non-empty).
- **(A7)** $\mathbf{g}$ is *nonpreemptive* in a channel for the entire service process of a packet; that is, a channel-switching decision is only made before or after

the service process of a packet.

(**A8**) The limits of $\mathcal{Q}_i^n(j)$, $\mathcal{Q}_i^c(j)$ and $\mathcal{Q}_i^p(l)$ exist under $\mathbf{g}$ as their respective arguments tend to infinity, and are denoted by $\mathcal{Q}_i^n$, $\mathcal{Q}_i^c$ and $\mathcal{Q}_i^p$, respectively.[2]

Similar as in single channel analysis, we impose the Markovian assumption on the discrete-time queueing process $\hat{Q}_i^{(k)}(n)$, which is the embedded process of $Q_i(t)$ (queue state of node i) sampled at the boundaries of c-slots of channel k, and define $\hat{\rho}_i^{(k)} = \lim_{n\to\infty} P(\hat{Q}_i^{(k)}(n) > 0)$. Also, let $\tau_i^{(k)}(n)$ be the probability that node i initiates a transmission attempt in the nth c-slot of channel k. Then we have the following lemma; its proof is similar to that of Lemma 1 (based on **A6** and **A8**) and omitted.

Lemma 2: $\tau_i^{(k)} := \lim_{n\to\infty} \tau_i^{(k)}(n)$ exists and is given by $\tau_i^{(k)} = \hat{q}_i^{(k)}\hat{\rho}_i^{(k)}/\overline{W}_i^{(k)}$, where $\overline{W}_i^{(k)} := \frac{\mathbb{E}[N_i^{s,(k)}]}{\mathbb{E}[N_i^{tx,(k)}]}$ is the average backoff length of node i in channel k, with $N_i^{s,(k)}$ and $N_i^{tx,(k)}$ defined in parallel as in the single channel case.

Remark 4: Under **A7**, $\overline{W}_i^{(k)}$ is given by

$$\overline{W}_i^{(k)} = \frac{1}{2}\left[W\left((1-p_i^{(k)})\sum_{j=0}^{m-1}(2p_i^{(k)})^j + (2p_i^{(k)})^m\right) + 1\right],$$

where $p_i^{(k)}$ is the probability of collision in channel k given a transmission attempt and W is the initial backoff window size.

Given any scheduling policy $\mathbf{g}$, let $\Lambda^{\mathbf{g}}$ be the corresponding stability region, and we have the following theorem characterizing $\Lambda^{\mathbf{g}}$.

Theorem 2: $\boldsymbol{\lambda} \in \Lambda^{\mathbf{g}}$ if and only if there exists at least one solution $\boldsymbol{\tau} = (\boldsymbol{\tau}^{(k)}, k \in \mathcal{C})$ where $\boldsymbol{\tau}^{(k)} = (\tau_i^{(k)}, i \in \mathcal{N})$ to the following constrained system of equations $(\Sigma^{\mathbf{g}}, \mathrm{C}, \boldsymbol{\lambda})$,

$$\Sigma^{\mathbf{g}}: \begin{cases} \tau_i^{(k)} = \dfrac{\hat{q}_i^{(k)}\hat{\rho}_i^{(k)}}{\overline{W}_i^{(k)}}, \ \forall i,k & \text{(a)} \\ p_i^{(k)} = 1 - \displaystyle\prod_{j\neq i}(1-\tau_j^{(k)}), \ \forall i,k & \text{(b)} \\ \rho_i = \min\left\{\dfrac{\lambda_i}{P}\displaystyle\sum_{k\in\mathcal{C}}\left[\tilde{q}_i^{(k)}\left(\dfrac{\overline{W}_i^{(k)}-1}{1-p_i^{(k)}}\mathbb{E}[S_{i,Q,\overline{T}_x}^{(k)}] + \right.\right.\right. \\ \qquad \left.\left.\left. + T_c^{(k)}\dfrac{p_i^{(k)}}{1-p_i^{(k)}} + T_s^{(k)}\right)\right], 1\right\}, \ \forall i,k & \text{(c)} \end{cases}$$

subject to

$$\mathrm{C}: \begin{cases} 0 \le \tau_i^{(k)} \le 1, \ \forall i,k & \text{(i)} \\ 0 \le \rho_i < 1, \ \forall i & \text{(ii)} \end{cases}$$

where $i \in \mathcal{N}$ and $k \in \mathcal{C}$; P is the packet payload size; $\mathbb{E}[S_{i,Q,\overline{T}_x}^{(k)}]$ is the conditional average length of a c-slot in channel k given that the queue at node i is non-empty but i does not transmit in this slot.

Proof: The proof is an immediate extension of the proof of Theorem 1, given assumptions on $\mathbf{g}$. □

The existence of a solution to $\Sigma^{\mathbf{g}}$ can be similarly established using Brouwer's fixed point theorem. We next study its uniqueness and the throughput optimality of a switching policy by resorting to an approximation given below, due to the complexity of $\Sigma^{\mathbf{g}}$. For the rest of this section, we will limit our discussion to the symmetric case where the channels have the same bandwidth and the system uses the same parameterization in all channels. We extend our discussion to more generic settings in the next section.

Definition 5: A scheduling policy is *unbiased* if the stationary channel occupancy distribution induced by such a policy is identical for every node, i.e., $q_i^{(k)} = q^{(k)}$ for all $i \in \mathcal{N}$ and $k \in \mathcal{C}$. It is denoted by $\mathbf{g}^U$, and the space of unbiased policies is $\mathcal{G}^U$.

We can obtain an approximation to $(\Sigma^{\mathbf{g}^U}, \mathrm{C}, \boldsymbol{\lambda})$ similarly as we did for Σ, using $\hat{q}^{(k)} \approx \tilde{q}^{(k)} \approx q^{(k)}$:

$$\widetilde{\Sigma}^{\mathbf{g}^U}: \begin{cases} \tau_i^{(k)} = \dfrac{q^{(k)}\rho_i}{\frac{W+1}{2}} & \text{(a)} \\ \rho_i = \dfrac{\lambda_i}{P}\displaystyle\sum_{k\in\mathcal{C}}\left\{q^{(k)}\left[\frac{W-1}{2}\left(\sigma + T\sum_{j\neq i}\tau_j^{(k)}\right)\right.\right. \\ \qquad \left.\left. + T\left(1 + \displaystyle\sum_{j\neq i}\tau_j^{(k)}\right)\right]\right\} & \text{(b)} \end{cases}$$

and we have the following result.

Theorem 3: Consider a system modeled by $\widetilde{\Sigma}^{\mathbf{g}^U}$ and the associated stability region $\Lambda^{\mathbf{g}^U}$. For all sufficiently large initial window sizes W, (i) the system of equations $(\Sigma^{\mathbf{g}^U}, \boldsymbol{\lambda})$ admits a unique solution, and (ii) $\mathbf{g}^U$ is throughput optimal within the class $\mathcal{G}^U$ if $q^{(k)} = \frac{1}{K}$ for all k. These are referred to as *equi-occupancy* policies.

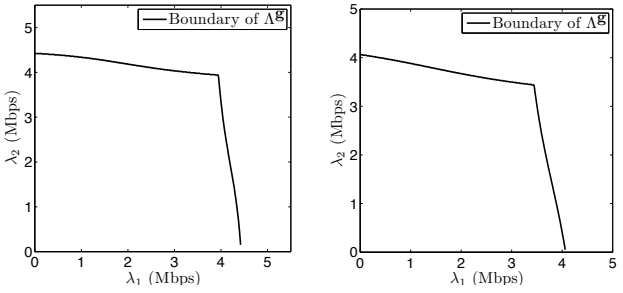
Proof: We omit the proof on uniqueness, which is similar to the single-channel case; see Appendix E for the proof on throughput optimality. □

The above results provide the following insights in addition to what we have observed in the single-channel case. Firstly, it's worth noting that $\Sigma^{\mathbf{g}}$ reduces to Σ in the single-channel case by properly configuring related parameters, and $\Sigma^{\mathbf{g}}$ thus constitutes a unified framework in describing the stability region of 802.11 DCF.

Secondly, the uniqueness of the solution to $(\Sigma^{\mathbf{g}^U}, \boldsymbol{\lambda})$ is in fact true for even small windows. As an example, in Figure 5, we plot the numerical boundaries of stability regions for various window settings with equal channel occupancy. Compared to results in the single-channel case, convexity of the stability region is observed even with small backoff windows in the bi-channel case. Also, the numerical multi-equilibrium phenomenon disappears in this case. One way to explain this is by considering the discounting effect of channelization on the attempt rate. The attempt rate

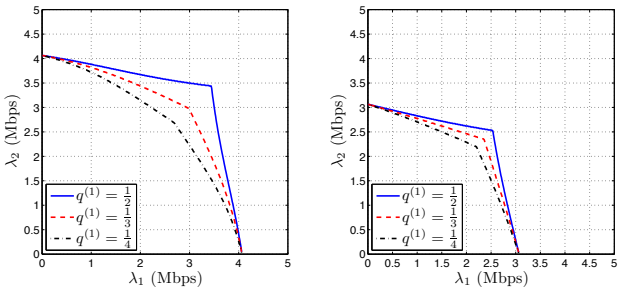
2. These limiting quantities are related by well-define correspondences, which are detailed in our technical report [16], and those relations are used to numerically evaluate the stability region equation for a multi-channel system presented in this section.

of each node in a channel is discounted by the occupancy probability in that channel. As discussed in the single-channel case, the attempt rate is roughly upper bounded by the reciprocal of the average backoff window size. Hence channelization has the effect of window expansion. The same explanation also applies to the observation that the stability region in a multi-channel system is nearly always convex.



(a) $W = 8, m = 5$ (b) $W = 32, m = 5$

Fig. 5. The stability region of bi-channel 802.11 DCF under the equi-occupancy policy.



(a) $W = 32, m = 5$ (b) $W = 128, m = 5$

Fig. 6. Throughput optimality of equi-occupancy distribution.

Thirdly, given symmetric channelization, equal occupancy time is equivalent to equal packet assignment in each channel. The optimality of equi-occupancy policies therefore confirms the intuitive notion that load balancing (either in the number of active nodes or in the amount of date flow) optimizes the system performance in terms of expanding the stability region. In Figure 6, we plot the analytical boundaries of stability regions corresponding to different unbiased policies in two scenarios. As can been seen, the equi-occupancy policy results in a stability region that is the superset of those of the other unbiased policies. It is also worth noting that as the backoff window increases, the gap between the superset region and other inferior regions decreases, as the reciprocal of the window size becomes the dominant factor in upper bounding the attempt rate.

6 Applicability and Implementation of Unbiased Policies in Both Symmetric and Asymmetric Systems

In this section we discuss the applicability of the class of unbiased policies. We then present a number of practical implementations and their use in both symmetric and asymmetric systems.

6.1 Unbiased policies

We have so far restricted our policy space to unbiased policies that induce a node-independent channel occupancy or packet assignment distribution. Note that while nodes in the same system are typically programmed with the same protocol stack, the same protocol may not necessarily yield the same statistical behavior among different nodes. Nevertheless, there are a number of circumstances in which node-independent behaviors are induced, which justifies our focus on unbiased policies. Firstly, if the protocol explicitly prescribes packet allocation to each channel, the resulting packet assignment distributions are identical for all nodes. Secondly, if nodes have identical arrival processes, they then have unbiased behavior as well. Unbiasedness can also be observed in a saturated network (however, such a system is unstable).

More generally, we note that when a node is active (i.e., its queue is non-empty and it is in the service process), from a mean-field point of view the channel conditions observed by this node is fully characterized by $p_i^{(k)}$ for each k (as a result of the decoupling assumption), which is a function of $\tau_j^{(k)}$ for all $j \neq i$. Therefore, the set of attempt rates $\{\tau_i^{(k)}; \forall i, \forall k\}$ characterizes the contention condition in the system. If nodes are asymptotically symmetric, that is, $\lim_{N\to\infty} \tau_i^{(k)}/\tau_j^{(k)} = 1$, for all $i \neq j$ and k, then we have

$$\lim_{N\to\infty} \frac{p_i^{(k)}}{p_j^{(k)}} = \lim_{N\to\infty} \frac{1 - \prod_{l\neq i}(1-\tau_l^{(k)})}{1 - \prod_{l\neq j}(1-\tau_l^{(k)})} = 1 + \lim_{N\to\infty} \frac{A(\tau_j^{(k)} - \tau_i^{(k)})}{A\tau_i^{(k)} + (1-A)} = 1,$$

where $A = \prod_{l\neq i,j}(1-\tau_l^{(k)})$. In this case we may consider the behavior induced by the underlying protocol on each node identical, and the corresponding policy unbiased. Note that the decoupling assumption is regarded as asymptotically true for a large number of nodes, so we may consider the asymptotic symmetry as an adjoint condition if we impose the decoupling approximation in modeling.

6.2 Practical implementation of throughput optimal unbiased policies: symmetric channels

We have shown that when channels are symmetric the optimal switching policy within the class of unbiased

policies is the equi-occupancy policy that balances load precisely. When channels are asymmetric, i.e., have different bandwidths, it is natural to expect that a load balancing policy yields throughput optimal performance, and to interpret a balanced load as having a packet assignment distribution proportional to the channel bandwidths. We will see that this interpretation is reasonable though not precise.

We begin by commenting on how such policies may be realized in a symmetric system.

We describe two very simple heuristics that implement an unbiased policy, and in particular, the equi-occupancy policy when channels are symmetric. The description is given in the bi-channel case for simplicity. The first is called SAS (switching after success), and the second SAC (switching after collision). In both schemes, a switching probability is assigned to each backoff stage. Under SAS (resp. SAC), a node switches to the other channel with probability $\alpha_l^{(k)}$ upon a successful transmission (resp. collision) if it is at the lth backoff stage in channel k when this success (resp. collision) occurs. In addition, in SAC, after switching to the other channel, a node does not reset its backoff stage; instead, it continues the exponential backoff due to the last collision. Note that SAS can be used to implement any arbitrary packet assignment distribution (and thus load distribution), which is a useful feature when we proceed to the implementation under asymmetric channels. This is because with the assumption of nonpreemptiveness of the policy, i.e., **A7**, switching after each successful transmission is equivalent to assigning packets.

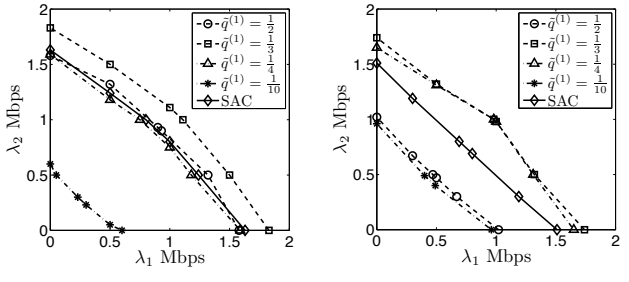
These two schemes heuristically implement the equi-occupancy policy in the following sense, when the switching probability profiles are identical in all channels and the channels are symmetric. Consider the two-dimensional Markov chains for a bi-channel system in the form of Bianchi's model [7], where each state in one channel has a mirror state in the other. Since for both SAS and SAC, the corresponding Markov chain is irreducible with a finite number of states, using the argument of symmetry, the symmetric solution is the unique stationary distribution that reflects equi-occupancy. It should be noted however that neither of the above is a perfect solution and the key may be a proper combination of the two. The problem with SAS is that it can result in empty channels (the node that succeeded in the transmission happens to be the only node in that channel). When this happens nodes can tend to cluster in the non-empty channel for significant periods of time due to collision and backoff, while our mean field Markov analysis implicitly assumes no channels are empty for long. On the other hand, the problem with SAC (SAC rarely results in empty channels and avoids clustering in one channel) is that it interrupts the service process of a packet in a given channel, thus violating the nonpreemptive assumption about the policy.

It is also worth noting that when SAS or SAC implements the equi-occupancy policy, or more generally known occupancy (or packet assignment) distributions, our model and assumptions admit an $M/M/1$ type of delay analysis. For instance, the average packet delay of a stable node i is given by $\frac{\rho_i}{\lambda_i(1-\rho_i)}$ and can be numerically evaluated through the stability equations.

6.3 Practical implementation of throughput optimal unbiased policies: asymmetric channels

We next proceed to asymmetric channels and examine how these two heuristics perform in this setting, and in doing so also empirically examine when the stability region is maximized. In particular, we focus on the performance of a policy when the majority of the nodes have similar arrival rates, and we examine the advantage of load balancing in improving stability. In our experiment, we fix 10 nodes with an arrival rate 0.5Mbps that creates a mean-field background in a bi-channel network while inspecting the stability region of another two nodes, which is the projection of the aggregate stability region onto a plane of these two nodes' arrival rates. All nodes use the same policy in a single experiment.

In Figure 7, we plot the empirical boundary of stability regions under different packet assignment distributions (implemented using SAS). As shown, policies with packet assignment ratio close to the bandwidth ratio indeed result in larger stability regions. However, while it seems safe to claim that properly balancing active time among channels according to their bandwidths improves the system performance, it remains unknown whether an exact match in load assignment is the optimal policy due to the nonlinearity of slot length in each channel w.r.t. active nodes. In addition, in practice we may not even know the effective bandwidth of each channel when channel conditions are imperfect.



(a) The bandwidth ratio = 1:2 (b) The bandwidth ratio = 1:3

Fig. 7. The projection of simulated stability region onto a plane of arrival rates of the two nodes under inspection.

It is therefore highly desirable to have an adaptive mechanism that dynamically adjusts the load distribution in practical implementation. Below we show that

SAC to a large extent can achieve this goal, with the reason being that collision rate reflects the contention level and bandwidth information. Figure 7 also shows the empirical stability region obtained using SAC with a switching probability at the lth backoff stage $\alpha_l^{(k)} = \frac{l}{m}$ for all k, where m is the maximum backoff stage. SAC is clearly not optimal, but it maintains good performance under different bandwidth ratios.

We further highlight the adaptiveness of SAC in comparison to SAS. Assume that the active node population in each channel is the same and static, given then the same period of time, faster channels experience more transmission successes than slower ones. Therefore, if a SAS-like switching policy is adopted for a relatively congested network, nodes would cluster in the slower channels and the throughput performance degrades significantly. However, if the congestion is due to bandwidth asymmetry, then this is reflected in the collision rate of transmission, which in turns triggers channel reallocation under SAC. We illustrate this point using the following experiment. Consider a bi-channel system with strongly asymmetric channels, where the bandwidth of channel 1 (2) is 1Mbps (10Mbps). The system consists of 60 nodes each with an arrival rate 0.1Mbps, and this aggregate arrival rate (6Mbps) is slightly below the empirical saturation throughput under this setting. In the first test, we compare the resulting distribution of number of nodes in channel 1 between SAC and SAS with the switching probability $\alpha_l^{(k)} = 0.5$ for all stages in both channels, and we repeat the inspection with the switching probability $\alpha_l^{(k)} = \frac{l}{m}$ at stage l in the second test; the duration of simulation is 180 seconds. The switching probability profile in the first test can be regarded as a blind configuration, while the second profile can be taken as an adaptive configuration that partially incorporates collision history into switching decisions. In Figure 8, we plot the histograms of the number of nodes in channel 1, as well as the empirical throughput obtained. As can be seen, the blindly configured SAS drives nodes to cluster in the slower channel, while SAC avoids this problem. Interestingly, SAS has comparable performance as SAC if we adjust the switching probabilities as we did in the second test, which reflects the congestion level in the residing channel, and both distributions "match" the bandwidth ratio. It suggests that while SAS is not as adaptive as SAC, it remains a valid alternative implementation and could achieve comparable performance when configured appropriately, as did above.

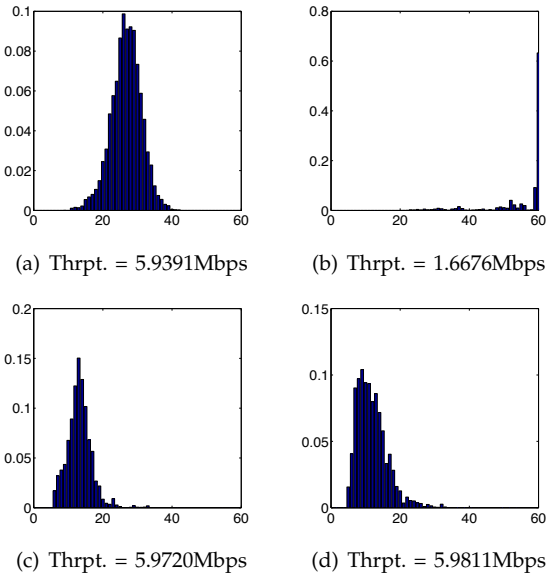


Fig. 8. Histogram of node population in the slower channel: (a)((b)) SAC (SAS) with $\alpha_l = 0.5$; (c)((d)) SAC (SAS) with $\alpha_l = \frac{l}{m}$.

6.4 Fairness under throughput-optimal policies

The general philosophy of SAS is that a node immediately vacates a channel in which it just had a success so other nodes can have a chance, while that of SAC is to keep using that channel until it gets inferior. While at opposite ends of the spectrum, this altruism and egoism respectively achieves the same system level fairness when universally adopted by all nodes in the network due to symmetry[3].

To illustrate further, consider a possibly asymmetric bi-channel system with a mixture of saturated and unsaturated nodes, and consider two notions of fairness. Under the first notion, fairness is measured by the individual throughput achieved by a node, compared to other similarly loaded nodes. For stable nodes, their throughput is simply their arrival rates. For saturated nodes, their attempt rates become essentially the same after queues have built up. This together with the fact that the implementation of SAS and SAC are not user-specific suggests the individual throughput is identical among saturated nodes.

Under the second notion, we measure fairness by the portion of a user's packets served in the better channel. Recall that SAS can be used to implement any arbitrary packet assignment distribution by tuning the conditional switching probabilities at each backoff stage after a successful transmission. For instance, if the switching probabilities are set to 1 in the worse channel at all stages while $1/2$ in the better one, each node should then have on average $2/3$ of its packets served in the better channel in the long term. This is independent of the arrival process or attempt rate of any node, and hence this type of fairness is also achieved.

3. *Strategic* behavior could lead to unfair advantage if users deviate from the preset rule. Consider for instance a bi-channel example where all but one node adopt SAS thus clustering in an inferior channel, while one node persists in the good channel using SAC.

7 Signal Quality plus Congestion Level in Channel Selection

Our primary intention is to study how congestion should be factored into switching decisions in a multi-channel system, and have so far assumed a perfect channel condition in terms of signal quality. In this section we consider the impact of considering congestion *in addition* to signal quality in making channel switching decisions. Below we first consider extending the current model to include packet loss due to poor channel/signal quality, and then empirically study how SAS and SAC perform under imperfect channel conditions compared to a switching policy that solely relies on signal quality estimates.

Different signal quality can be captured by a probability of packet failure loss for each transmission attempt, independent from losses due to collision, denoted by $\pi^{(k)}$ for channel k. We consider two cases depending on whether we will assume that a node can distinguish a collision loss from a packet failure loss due to poor signal quality. In the first case when a node is able to distinguish the two, then Automatic Repeat reQuest (ARQ) can be applied upon a failed transmission within the same channel reservation (i.e., a node does not release the channel upon a packet failure but will continue to retransmit). For simplicity we will assume there is no re-try limit, and thus the introduction of packet failure losses only affects the duration of a data session after a successful channel reservation, which was denoted by $T_s^{(k)}$ in the origin model for a successful transmission. This effectively leads to asymmetric channels even if they have the same amount of bandwidth. Since the duration of a single data session is generally much greater than the channel coherence time, we will assume that packet failures occur independently in each re-transmission attempt with probability $\pi^{(k)}$. The number of retransmissions then follows a geometric distribution, and the expected duration of a data session after a successful reservation of channel k is given by $\frac{T_q^{(k)}}{1-\pi^{(k)}} + T_s^{(k)}$, where $T_q^{(k)}$ is the duration of a transmission that resulted in packet failure.

In the second case, when a node is not able to distinguish a packet failure loss from collision, it will simply regard each unsuccessful transmission attempt as being involved in a collision. As a result the conditional collision probability given a transmission attempt in $\Sigma^{\mathfrak{g}}$(b) is updated as

$$p_i^{(k)} = 1 - (1 - \pi^{(k)}) \prod_{j \neq i} (1 - \tau_j^{(k)}).$$

In both cases, the original model can be extended to compute the corresponding stability regions.

We now numerically compare the proposed congestion-aware switching algorithm to a method that uses only signal quality. Consider three channels with equal bandwidth (a third of 11Mbps) but different signal qualities modeled as packet loss probabilities for a given transmission attempt (0.1, 0.2 and 0.3 for the three channels, respectively). Assume nodes can tell collision loss from failure loss. We fix 20 nodes each with an arrival rate 0.1Mbps that creates a mean-field background as in the previous section, while tuning the arrival rates of two additional nodes. We then inspect the stability region projected onto the plane where these two nodes' arrival rates reside.

In one scenario, all nodes use SAC together with ARQ within each data session until success. In the other scenario, all nodes use a signal-based (SB) switching method that essentially performs an online estimate of the packet failure loss rate in each channel, by tracking the total number of successful transmissions and the total number of transmission attempts within each data session (after successful channel reservation), and switches to (or remains in) the channel with the lowest current estimate upon each successful packet transmission. In the long run one expects nodes to cluster in the best channel even while it gets more congested. This is indeed observed in our simulation; the resulting stability regions are depicted in Figure 9[4]. We also report the average number of nodes in each channel at near-saturated points during a simulation of 30 seconds in Table 1, which confirms our intuition.

In this study we have used a rather simple signal-based algorithm. Nevertheless it validates our observation that considering *only* signal quality can be a very detrimental thing to do when there is significant congestion in the system.

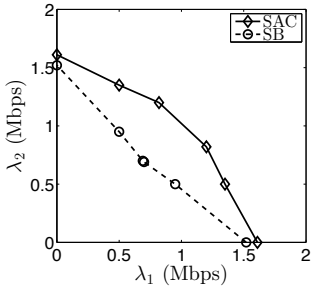


Fig. 9. Congestion-based vs. signal-based: stability region.

TABLE 1
Congestion-based vs. signal-based: node distribution.

Channel		1	2	3	
Node Distr'n	SAC	6.69	7.21	8.10	$(\lambda_1, \lambda_2) = (1, 1)$
	SB	14.22	5.30	2.48	$(\lambda_1, \lambda_2) = (.6, .6)$

4. Note that only a limited number of boundary points are identified to sketch the stability regions; the connecting lines are hence not necessarily the exact boundary.

8 Conclusion

Using the characterization of the stability region of a multi-channel multi-user WLAN system, we investigated the throughput optimal channel switching schemes in such systems. In particular, we showed that a balanced load distribution (channel occupancy time, packet assignment) in general improves the system performance in terms of the stability region, and we proposed simple and adaptive online switching algorithms to achieve load balance in a general system with asymmetric channels. While the modeling effort primarily focused on the congestion aspect assuming perfect signal quality, we also presented extensions of the basic model to incorporate noisy channels, which in essence can be considered as asymmetric channels in bandwidth. We also performed an empirical comparison between our channel switching method and one that is solely based on signal quality; the latter induces nodes to cluster in one channel. Our ultimate intention is to promote a definition of channel quality that reflects both signal quality and congestion levels in a multi-channel wireless network. This work can be extended in the following directions: 1) the effect of asymmetric channels on the characterization of stability region; 2) throughput optimal switching when considering the larger space of biased policies.

References

[1] K. Tan, J. Zhang, J. Fang, H. Liu, Y. Ye, S. Wang, Y. Zhang, H. Wu, W. Wang, and G. Voelker, "Sora: High Performance Software Radio using General Purpose Multi-core Processors," *USENIX NSDI 2009*, 2009.

[2] Y. Li, J. Fang, K. Tan, J. Zhang, Q. Cui, and X. Tao, "Soft-LTE: A Software Radio Implementation of 3GPP Long Term Evolution Based on Sora Platform," *Demo in ACM MobiCom 2009*, 2009.

[3] F. K. Jondral, "Software-Defined Radio: Basics and Evolution to Cognitive Radio," *EURASIP Journal Wireless Communications and Networking*, vol. 2005, pp. 275–283, August 2005.

[4] N. B. Chang and M. Liu, "Optimal Channel Probing and Transmission Scheduling for Opportunistic Spectrum Access," *IEEE/ACM Transactions on Networking*, vol. 17(6), pp. 1805–1818, 2009.

[5] V. Kanodia, A. Sabharwal, and E. Knightly, "MOAR: A Multi-channel Opportunistic Auto-rate Media Access Protocol for Ad Hoc Networks," in *Proceedings of Broadnets 2004*, 2004.

[6] B. Sadeghi, V. Kanodia, A. Sabharwal, and E. Knightly, "Opportunistic Media Access for Multirate Ad Hoc Networks," in *Proceedings of ACM MOBICOM*, 2002.

[7] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 535–547, 2000.

[8] A. Kumar, E. Altman, D. Miorandi, and M. Goyal, "New Insights from a Fixed Point Analysis of Single Cell IEEE 802.11 WLANs," in *Proceedings of IEEE INFOCOM*, 2005.

[9] G. R. Cantieni, Q. Ni, C. Barakat, and T. Turletti, "Performance Analysis under Finite Load and Improvements for Multirate 802.11," *Elsivier Computer Communications*, vol. 28(10), pp. 1095–1109, 2005.

[10] D. Malone, K. Duffy, and D. J. Leith, "Modeling the 802.11 Distributed Coordination Function in Non-saturated Heterogeneous Conditions," *IEEE/ACM Transactions on Networking*, vol. 15(1), pp. 159–172, 2007.

[11] A. Jindal and K. Psounis, "The Achievable Rate Region of 802.11-Scheduled Multi-hop Networks," *IEEE/ACM Transactions on Networking*, vol. 17(4), pp. 1118–1131, 2009.

[12] D. Leith, V. Subramanian, and K. Duffy, "Log-convexity of Rate Region in 802.11e WLANs," *IEEE Communications Letters*, vol. 14(1), pp. 57–59, 2010.

[13] A. Raniwala and T. Chiueh, "Architecture and Algorithms for an IEEE 802.11-based Multi-Channel Wireless Mesh Network," in *Proceedings of IEEE INFOCOM*, 2005.

[14] A. Mohsenian-Rad and V. Wong, "Distributed Multi-Interface Multichannel Random Access Using Convex Optimization," *Mobile Computing, IEEE Transactions on*, vol. 10, pp. 67–80, Jan. 2011.

[15] V. Anantharam, "The Stability Region of the Finite-User Slotted ALOHA Protocol," *IEEE Transactions on Information Theory*, vol. 37, pp. 535–540, 1991.

[16] Q. Wang and M. Liu, "Throughput Optimal Switching in Multi-channel WLANs," *Arxiv preprint arXiv:1201.6065v1*, 2012.

[17] K. R. Duffy, "Mean Field Markov Models of Wireless Local Area Networks," *Markov Processes and Related Fields*, vol. 16(2), pp. 295–328, 2010.

[18] E. Felemban and E. Ekici, "Single Hop IEEE 802.11 DCF Analysis Revisited: Accurate Modeling of Channel Access Delay and Throughput for Saturated and Unsaturated Traffic Cases," *IEEE Transactions on Wireless Communications*, vol. 10, no. 10, pp. 3256–3266, 2011.

[19] G. Bianchi and I. Tinnirello, "Remarks on IEEE 802.11 DCF Performance Analysis," *IEEE Communications Letters*, vol. 9, pp. 765–767, 2005.

[20] J. L. Massey and P. Mathys, "The Collision Channel Without Feedback," *IEEE Transactions on Information Theory*, vol. 31, pp. 192–204, 1985.

Qingsi Wang received his B.E. degree in electrical engineering in 2009 from Shanghai Jiao Tong University, Shanghai, China, and M.S. degree in electrical engineering: systems in 2011 from the University of Michigan, Ann Arbor, Michigan. He is currently a Ph.D. candidate in the Department of Electrical Engineering and Computer Science, University of Michigan. His research interests include resource allocation problems in wireless networks, stochastic control and game theory.

Mingyan Liu received her Ph.D. Degree in electrical engineering from the University of Maryland, College Park, in 2000 and has since been with the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, where she is currently a Professor. Her research interests are in optimal resource allocation, performance modeling and analysis, and energy efficient design of wireless, mobile ad hoc, and sensor networks. She is the recipient of the 2002 NSF CAREER Award, the University of Michigan Elizabeth C. Crosby Research Award in 2003, and the 2010 EECS Department Outstanding Achievement Award.