of the coefficients $|S|u_S$ [15]. Since there are at most $2^n - 1$ coefficients $|S|u_S$, the maximization in (22) can be solved in $O(n)$ (linear) time.

V. Conclusion

We considered throughput optimal control of a wireless networks with cooperative relaying. Our model applies to a general network topology and several different types of cooperative scenarios. We established the network stability region and gave a variation of the Maximum Differential Backlog policy, which we proved to be throughput optimal. We focused on a centralized implementation and showed how the structure of the underlying capacity regions can aid in implementing this policy. In practice, a distributed solution is more desirable, particularly for managing the complexity of a cooperative network. Moreover, in a large network, there may be many potential cooperative sets. A useful direction for future work would be to develop a means for determining the most "useful" of these sets.

References

[1] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for broadcast channels," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun./Jul 2002, p. 382.

[2] E. Yeh and A. Cohen, "Throughput optimal power and rate control for queued multiaccess and broadcast communications," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, 2004, p. 112.

[3] E. Yeh and A. Cohen, "Information theory, queueing, and resource allocation in multi-user fading communications," in *Proc. Conf. Information Science and Systems*, Princeton, NJ, 2004.

[4] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[5] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time varying wireless networks.," in *Proc. Infocom*, 2003.

[6] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[7] B. Schein and R. G. Gallager, "The Gaussian parallel relay network," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000, p. 22.

[8] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity - part I: System description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[9] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[10] A. Høst-Madsen and J. Zhang, "Capacity bounds and power allocation for the wireless relay channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2020–2040, Jun. 2005.

[11] G. Barriac, R. Mudumbai, and U. Madhow, "Distributed beamforming for information transfer in sensor networks," in *Proc. IPSN*, Apr. 2004.

[12] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3037–3063, Sep. 2005.

[13] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[14] E. Yeh and R. Berry, Throughput Optimal Control of Cooperative Relay Networks Dep. Elec. Eng., Yale Univ., Tech. Rep., Sep. 2006.

[15] D. N. C. Tse and S. Hanly, "Multi-access fading channels: Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2815, Nov. 1998.

[16] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. IEEE Int. Symp. Information Theory*, Ulm, Germany, 1997, p. 27.

[17] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels: Part I: Ergodic capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

Quality of Service Analysis for Wireless User-Cooperation Networks

Lingjia Liu, *Student Member, IEEE*, Parimal Parag, *Student Member, IEEE*, and Jean-François Chamberland, *Member, IEEE*

***Abstract*—A wireless communication system in which multiple users cooperate to transmit information to a common destination is considered. The traffic generated by the users is subject to a stringent quality of service requirement, which is defined in terms of the asymptotic decay-rate of buffer occupancy. The performance of this communication system is analyzed, and the corresponding achievable rate-region for the two-user scenario is identified. A simple user-cooperation scheme that improves performance is proposed. This cooperative scheme is shown to significantly enlarge the achievable rate-region of the service constrained communication system, provided that the quality of the wireless link between cooperating users is better than the individual connections from the users to the intended destination. Numerical results further indicate that the gains of cooperative strategies can be substantial. This suggests that cooperation allows for a fair distribution of the wireless resources among active users.**

***Index Terms*—Communication systems, effective bandwidth, effective capacity, fluid models, quality of service (QoS), user cooperation, wireless networks.**

I. Introduction

Recent years have been marked by a soaring demand for network access. This trend is exemplified by the constant growth of the Internet. The strong demand for network connectivity is fueled, partly, by new software applications, utility computing, and a widespread desire for real-time information access. To bridge the gap between mobile users and established communication infrastructures, wireless technology is being embraced with increasing vigor. Wireless systems offer a unique mixture of connectivity, flexibility, and freedom. Future communication networks face the dual challenge of supporting large traffic volumes and providing reliable service to delay-sensitive applications such as VoIP, video conferencing, electronic commerce, and gaming. Most of the research on physical aspects of wireless systems published in the literature today focuses on maximizing Shannon capacity [1] or spectral efficiency [2], [3]. These initiatives afford a foundation for improving throughput in wireless networks. However, the stringent service requirements typical of real-time traffic suggest that a classical capacity/throughput analysis alone does not offer a complete assessment of service quality for the communication infrastructure associated with a wireless network. Wireless channels are prone to attenuation, fading, and interference. These variations influence user satisfaction as they negatively impact queue-lengths, packet loss probabilities, and delay distributions.

Traditionally, power control and error-correcting codes have been employed to mitigate the effects of the channel fluctuations intrinsic to wireless communications. Yet, as the popularity of real-time applications increases, new paradigms that maximize throughput subject to quality of service (QoS) constraints are becoming highly desir-

Manuscript received August 15, 2006; revised February 10, 2007.

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843-3128 USA (e-mail: lingliu@ece.tamu.edu; parimal@ece.tamu.edu; chmbrlnd@ece.tamu.edu).

Communicated by R. A. Berry, Guest Editor for the Special Issue on Relaying and Cooperation.

Color versions of Figures 5–10 in this correspondence are available at http:\\ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2007.903149

able. In particular, stringent QoS requirements and end-to-end delay restrictions imposed on a communication system may preclude the use of error-correcting codes with long block-lengths, thereby limiting the benefits of coding. Although several notable contributions have improved our understanding of the subject [4], the literature on the tradeoff between throughput and service quality in wireless environments is far from being fully developed.

Very little work has been done to harness the potential of newly developed schemes at the physical layer in order to improve service quality in wireless environments. Such new schemes include multi-antenna systems [5] and user cooperation [6], [7]. The use of multiple antennas at the transmitters and receivers has been shown to substantially enhance the diversity [8] and the information theoretic capacity [5] of point-to-point wireless links. Still, practical considerations such as cost and size often limit the number of independent antennas a wireless device can utilize. Under such circumstances, user cooperation [9] has emerged as a viable alternative to multi-antenna systems. Cooperative strategies can be employed to increase the diversity [10] and the spatial multiplexing of wireless systems [11], [12] in a manner similar to a multi-antenna configuration. Furthermore, user-cooperation is found to enlarge the achievable rate-region of a multi-user system even when the transmitters only have partial channel state information [13].

In this correspondence, we propose a cross-layer approach and investigate the impact of user-cooperation on the queueing behavior of wireless communication systems. We analyze the performance of a simple cooperative strategy, and derive its achievable rate-region when the system operates under stringent service constraints. Due to the time-varying nature of wireless channels, it is difficult to provide deterministic delay guarantees to wireless users. Accordingly, we adopt a statistical QoS metric that captures the asymptotic decay-rate of buffer occupancy, i.e.,

$$\theta = -\lim_{x\to\infty} \frac{\log \Pr\{L > x\}}{x} \tag{1}$$

where L is the steady-state queue-length of the buffer present at the transmitter. The parameter θ reflects the perceived quality of a communication link; a larger θ represents a more reliable connection or a tighter QoS constraint. This metric is closely tied to the concept of effective bandwidth, which has been studied extensively in the context of wired networks [14]–[18]. Given a specific arrival process, the effective bandwidth characterizes the minimum bandwidth required for the communication system to meet a certain QoS requirement θ_0 [19], [20]. The buffer decay-rate of (1) is also related to the dual concept of effective capacity popularized by Wu and Negi [21]–[23]. Unlike wired connections where the service rates are typically constant, wireless channels are inherently unreliable and the associated service rates are usually time-varying. Assuming a constant flow of incoming data, the effective capacity characterizes the maximum arrival rate that a wireless system can support subject to a QoS requirement θ_0. When θ_0 approaches zero, the effective capacity converges to the maximum throughput supported by the wireless channel.

The remainder of the correspondence is organized as follows. Section II presents the system model we adopt, along with a precise problem formulation. It describes the wireless channel model that we employ as an abstraction for the physical layer. Section III contains a derivation of the equilibrium queue-length distribution for the underlying communication system. This distribution is used to compute the QoS metric θ associated with this system. This allows us to characterize the achievable rate-region of the cooperative scheme under study for an arbitrary QoS constraint θ_0 in Section IV. Generalizations of the system model considered in this correspondence are compared and contrasted in Section V. Conclusions and final remarks are discussed in Section VI.

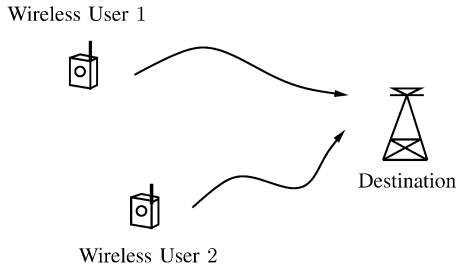

Fig. 1. Abstract model for a cooperative system with two users.

II. Problem Statement

Consider a wireless communication system where two users collaborate to transmit their respective data to a common destination, as shown in Fig. 1. The system is assumed to operate in a frequency-division multiplexing (FDM) mode. Each wireless user is subject to a mean power constraint and a finite spectral bandwidth allocation. A large buffer is available at every transmitter where outgoing packets are stored before being sent to their destination. Furthermore, we assume that the system must satisfy a global QoS constraint θ_0. That is, the asymptotic decay-rates of buffer overflow probabilities, θ_1 and θ_2, must satisfy $\min(\theta_1, \theta_2) \geq \theta_0$ where θ_i is defined in (1). Finally, we assume that channel state information is not available at the transmitters, although the channel statistics are. In practice, it is often costly for a transmitter to acquire accurate channel state information. This explains why we focus on the situation where channel state information is available only at the receiver, not at the transmitter.

A. Queueing Model

Let $a_i(t)$ denote the instantaneous arrival rate of user i at time t. Many real-time traffic sources such as voice, instant messaging, and wireless sensors can be accurately represented by on-off sources [24]. As such, we model $a_i(t)$ using a two-state Markov-modulated fluid process. We remark that a constant source can be viewed as a limiting case of an on-off source where the off-time approaches zero. For an on-off model, the instantaneous arrival rate of user i is $a_i > 0$ when the source is *on*, and zero otherwise. The arrival drift matrix for wireless user i can then be written as

$$D_{ai} = \begin{bmatrix} 0 & 0 \\ 0 & a_i \end{bmatrix}.$$

We denote the mean off-time of this user by λ_{ai}^{-1}; and its mean on-time, by μ_{ai}^{-1}. The generator matrices for the underlying continuous-time Markov chain of the arrival processes can then be expressed as

$$Q_{ai} = \begin{bmatrix} -\lambda_{ai} & \lambda_{ai} \\ \mu_{ai} & -\mu_{ai} \end{bmatrix}, \quad i = 1, 2.$$

In the situation where users do not cooperate, each wireless device transmits its data independently based on its allocated bandwidth and power budget. The connection of each user can therefore be modeled as a single-server queue, where the arrival process represents the data produced by the user and the service process is determined by the information received at the destination. Note that for the data to leave the buffer, the receiver must have the ability to acknowledge reception of the transmitted packets. For instance, a simple acknowledgment mechanism at the physical layer may be incorporated in the communication protocol to insure that erroneous data get retransmitted. We assume that such a mechanism is in place throughout. We emphasize, again, that the links between the users and their destination are orthogonal in a

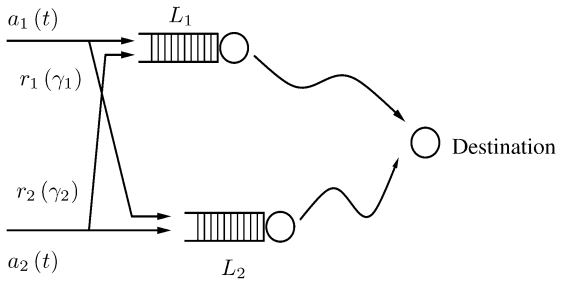

Fig. 2. User-cooperation scheme with two users.

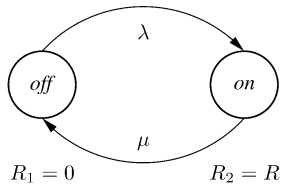

Fig. 3. Continuous-time Gilbert-Elliott Markov representation of a coded wireless communication link.

FDM system. In this case, a two-user system can effectively be decomposed into two independent point-to-point systems. More specifically, the maximum arrival rate that each user can support under QoS constraint θ_0 can be obtained separately.

In a typical wireless environment, channel conditions vary with location and time. As such, the maximum throughputs of two different users may be vastly asymmetric. For real-time traffic subject to stringent service constraints, this imbalance can be even larger. The goal of this correspondence is to design a system where cooperation among users enables them to share system resources equitably. This is accomplished by designing a communication strategy that enlarges their collective achievable rate-region under various service requirements. An expanded rate-region creates the flexibility necessary to share system resources fairly among users.

To take advantage of their mutual wireless links, the two users must first exchange data. In the proposed user-cooperation scheme, we allow each user to apply part of its own power and bandwidth to the exchange of information with its counterpart, as shown in Fig. 2. We represent the fraction of physical resources employed by user i to maintain communication with its peer by γ_i, and we let the capacity of the newly created interuser channel be denoted by $r_i(\gamma_i)$. We consider the specific scenario where the inter-user links are symmetric additive white Gaussian noise (AWGN) channels with constant gains. Thus, when generating traffic, user 1 sends data at rate $r_1(\gamma_1)$ to user 2, and stores the remaining data in its own buffer. User 2 follows a similar procedure, sending part of its data to user 1 and storing excess data locally whenever active. Based on the respective values of γ_1 and γ_2, we can characterize the achievable rate-region for the cooperative system of Fig. 2 under an arbitrary QoS parameter θ_0. The union of these rate-regions over all admissible pairs $(\gamma_1, \gamma_2) \in [0,1]^2$ yields an achievable rate-region for the proposed user-cooperation scheme. We denote this region by $\mathcal{R}(\theta_0)$, and point out that it is a function of the service requirement θ_0. As $\theta_0 \to 0$, this achievable region converges to the stability region of the system, which is characterized by its throughput optimal boundary.

B. Wireless Channel

Wireless communication channels are often subject to fading. For a dense scattering environment, the fading process $h(t)$ is well-modeled as a zero-mean, proper complex Gaussian process. The envelope $|h(t)|$ and the phase $\angle(h(t))$ form stationary random processes, with $|h(t)|$ having a Rayleigh probability distribution function and the phase being uniform on $[0, 2\pi)$. If we assume that the random process $h(t)$ is normalized, then $|h(t)|$ has distribution

$$f(\xi) = 2\xi e^{-\xi^2}.$$

A Rayleigh-fading channel profile only specifies the first-order statistics of $h(t)$. A complete description of this random process requires that the higher order statistics of $h(t)$ be specified as well [25], [26]. In this correspondence, we model the overall effects of fading on encoded transmissions rather than specify higher order statistics for $h(t)$. We adopt a Gilbert-Elliott channel model for the sake of mathematical tractability. Given a certain threshold η, we assume that the probability of $|h(t)|$ being above or below this threshold is captured adequately by a continuous-time Markov chain. We refer to the channel envelope exceeding η as the *on* state; the channel is in its *off* state otherwise. This quantized channel model appears in Fig. 3. The transition rate from *off* to *on* is denoted by λ; while the transition rate from *on* to *off*, by μ. The generator matrix Q_s for this Markov chain can be written as

$$\begin{aligned} Q_s &= \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \\ &= \frac{1}{\lambda+\mu} \begin{bmatrix} 1 & \lambda \\ 1 & -\mu \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -(\lambda+\mu) \end{bmatrix} \begin{bmatrix} \mu & \lambda \\ 1 & -1 \end{bmatrix}. \end{aligned}$$

For consistency, the stationary distribution of the Markov chain should agree with the marginal Rayleigh distribution of the underlying channel

$$\begin{aligned} \Pr\{|h(t)| \le \eta\} &= \frac{\mu}{\lambda+\mu} \\ &= \int_0^\eta 2\xi e^{-\xi^2} d\xi = 1 - e^{-\eta^2} \\ \Pr\{|h(t)| > \eta\} &= \frac{\lambda}{\lambda+\mu} \\ &= \int_\eta^\infty 2\xi e^{-\xi^2} d\xi = e^{-\eta^2}. \end{aligned} \tag{2}$$

Over a time interval of duration t, the channel transition probability matrix $P_t(t)$ associated with the Gilbert–Elliott model is given by

$$P_t(t) = e^{Q_s t} = \frac{1}{\lambda+\mu} \begin{bmatrix} \mu + \lambda e^{-t(\lambda+\mu)} & \lambda - \lambda e^{-t(\lambda+\mu)} \\ \mu - \mu e^{-t(\lambda+\mu)} & \lambda + \mu e^{-t(\lambda+\mu)} \end{bmatrix}.$$

Note that the autocorrelation function of this Gilbert–Elliott model decays exponentially over time. A parameter $\kappa = \lambda + \mu$ is introduced to denote the exponential decay rate of the covariance between samples. This parameter κ designates the speed of the fading process. A large κ stands for a fast fading scenario whereas a small one implies that the fading is slow. Referring to condition (2), λ and μ can be expressed in terms of the physical channel parameters

$$\begin{aligned} \lambda &= \kappa e^{-\eta^2} \\ \mu &= \kappa - \kappa e^{-\eta^2}. \end{aligned} \tag{3}$$

For a time-invariant AWGN channel, the maximum rate at which error-free data transfer is possible is given by

$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right) \quad \text{bits per second} \tag{4}$$

where P is the power of the received signal, $N_0/2$ is the power spectral density of the noise process, and W is the spectral bandwidth. Recent developments in error-control coding allow operation near the channel

capacity with minimal error-rates and small delays. The channel capacity expression of (4) can be viewed as an optimistic approximation of code performance. If a code is designed to operate at a rate R, the sent information can be recovered reliably provided that $R < C$; otherwise it is lost.

In a fading environment, the channel gain and hence the received power are time-varying. Assuming that the channel changes slowly over time, the instantaneous capacity of the wireless link is equal to

$$C(t) = W \log_2 \left(1 + \frac{|h(t)|^2 P}{N_0 W}\right) \quad \text{bits per second.} \tag{5}$$

If the transmitted information is encoded at a rate R, it is assumed to reach its destination reliably provided that $R < C(t)$. On the other hand, if $R \geq C(t)$ then the transmitted data is lost. This simplified characterization, which we employ throughout, is valid provided that there are enough degrees of freedom available during each data transfer to permit the use of sophisticated codes. The model at hand can be altered to accommodate practical codes and their probabilities of decoding failures.

When channel state information is not available at the transmitter, a wireless user must send its data using a preselected coderate R to the destination. In this case, the state of the Gilbert–Elliott channel model is determined by the instantaneous capacity of the wireless channel defined in (5). Specifically, the Gilbert–Elliott channel is in its *on* state if $R < C(t)$, i.e.,

$$|h(t)| > \eta = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R}{W}} - 1\right)}. \tag{6}$$

It is *off* otherwise. From (3) and (6), the generator matrix for this Gilbert-Elliott channel can be rewritten in terms of the system parameters as

$$Q_s = \begin{bmatrix} -\kappa e^{-\eta^2} & \kappa e^{-\eta^2} \\ \kappa - \kappa e^{-\eta^2} & -\kappa + \kappa e^{-\eta^2} \end{bmatrix}. \tag{7}$$

The offered service rate of the channel is a Markov-modulated fluid process, with value R when the channel is *on*, and zero when it is *off*.

III. Queueing Performance Analysis

To relate the effects of the physical layer to the performance of the network, we must first understand the queueing dynamics that govern the system. We start by considering a single server fluid queue with independent arrival and service processes. We assume that these two processes are coupled by a buffer of infinite length, and that they are modulated by finite-state continuous-time Markov chains with irreducible generator matrices $Q_a \in \mathbb{R}^{M\times M}$ and $Q_s \in \mathbb{R}^{N\times N}$, respectively. We let a_m represent the instantaneous arrival rate when the arrival process is in state m, and we write $p_a(t;m)$ to denote the probability of occurrence of this event at any given time $t \geq 0$. Similarly, the offered service has instantaneous rate R_n when the underlying process is in state n, and $p_s(t;n)$ represents the probability of being in this state at time t. We can write the arrival and the service drift matrices as $D_a = \operatorname{diag}(a_1, \ldots, a_M)$ and $D_s = \operatorname{diag}(R_1, \ldots, R_N)$, respectively. In vector form, the arrival and service probability distributions become

$$p_a(t) = (p_a(t;1), p_a(t;2), \ldots, p_a(t;M))$$
$$p_s(t) = (p_s(t;1), p_s(t;2), \ldots, p_s(t;N)).$$

Using these definitions, we can write the evolution of the probability vectors in a compact fashion

$$\frac{d}{dt}p_a(t) = p_a(t)Q_a$$
$$\frac{d}{dt}p_s(t) = p_s(t)Q_s.$$

We use the vectors w_a and w_s to represent the steady-state distributions of the arrival and service processes, with $w_a Q_a = w_s Q_s = 0$.

For the combined arrival and service process, let $X_t \in \{(m,n) : 1 \leq m \leq M, 1 \leq n \leq N\}$ be the situation where the arrival is in state m and the offered service is in state n at time t. The probability of this event is simply equal to $p(t;m,n) = p_a(t;m)p_s(t;n)$. We employ $p(t)$ to denote the vector consisting of the elements $\{p(t;m,n)\}$ in lexicographic order. It follows that $p(t) = p_a(t) \otimes p_s(t)$, where $\otimes$ is the Kronecker product [27], [28]. The joint probability vector $p(t)$ satisfies

$$\frac{d}{dt}p(t) = p(t)Q$$

where Q is the generator matrix of the joint process. This matrix can be written as

$$Q = Q_a \otimes I_N + I_M \otimes Q_s \tag{8}$$

where I_K is a $K \times K$ identity matrix. The matrix Q is recurrent and irreducible, and $w = w_a \otimes w_s$ is the steady-state distribution for the aggregate process. The net drift matrix D of the joint process is

$$D = D_a \otimes I_N - I_M \otimes D_s. \tag{9}$$

Let L_t represent the queue-length of the user at time t. The evolution of L_t in time can be expressed as [14], [29]

$$\frac{d}{dt}L_t = (a_m - R_n)\,\mathbf{1}_{\{L_t>0\}} + (a_m - R_n)^+ \mathbf{1}_{\{L_t=0\}} \tag{10}$$

which is a stochastic differential equation on the Markov process (L_t, X_t). Define the event probability

$$F(x,m,n,t) = \Pr\{X_t = (m,n), L_t \leq x\}$$

and let $F(x,t)$ be the lexicographic arrangement of $\{F(x,m,n,t)\}$. Using this notation, we can write the Chapman-Kolmogorov forward equation in matrix form as [14], [15], [28], [29]

$$\frac{\partial}{\partial t}F + \frac{\partial}{\partial x}FD = FQ.$$

The mean arrival rate and mean service rate are given by $\bar{a} = \langle w_a D_a, \mathbf{1}\rangle$ and $\bar{R} = \langle w_s D_s, \mathbf{1}\rangle$, respectively. If the system is stable (i.e., $\bar{a} < \bar{R}$), then the underlying Markov process is positive recurrent [14]. As such, there exists a steady-state distribution for the aggregate process (L_t, X_t) [30]. Let $\pi(x,m,n)$ denote the steady-state queue-length distribution of the buffer, with

$$\frac{\partial}{\partial x}\pi(x)D = \pi(x)Q.$$

Since $\pi(x)$ is a bounded solution, it has spectral representation

$$\pi(x) = w - \sum_{l=1}^{k} \alpha_l \phi_l e^{z_l x}, \tag{11}$$

where $\{(\phi_l, z_l) : \text{Real}\{z_l\} \leq 0\}$ are k eigenvector/eigenvalue pairs that satisfy the eigenvalue problem

$$z\phi D = \phi Q. \tag{12}$$

We emphasize that $\alpha_l = 0$ for any l such that $\text{Real}\{z_l\} > 0$ because the system is stable [28]. Thus, we only need to consider eigenvalues with negative real parts. Note that the system is subject to the boundary conditions $\pi(0, m, n) = 0$ whenever $a_m - R_n > 0$. If $a_m \neq R_n$ for all m and n, then there are exactly k such boundary conditions and the steady-state distribution is uniquely determined [28].

Solving the eigenvalue problem of (12) for the whole system can be somewhat involved. However, taking advantage of the special structure of D and Q, we can decompose the original system and reduce the complexity of the problem.

Lemma 1: For any eigenvector/eigenvalue pair (ϕ_l, z_l) that satisfies $z_l\phi_l D = \phi_l Q$, there exist ϕ_{a_l}, ϕ_{s_l}, and $\nu \in \mathbb{C}$ such that

$$z_l\phi_{a_l}(D_a - \nu I_M) = \phi_{a_l} Q_a \tag{13}$$
$$z_l\phi_{s_l}(\nu I_N - D_s) = \phi_{s_l} Q_s. \tag{14}$$

Proof: For $z_l = 0$, the result is trivial. For any ν, the vector $\phi = \phi_{a_l} \otimes \phi_{s_l} = w_a \otimes w_s$ satisfies (13) and (14). Assume $z_l \neq 0$, then $z_l\phi_l D = \phi_l Q$ is equivalent to

$$\phi_l \left(D - \frac{Q}{z_l} \right) = 0. \tag{15}$$

Substituting (8) and (9) into (15), we obtain

$$\phi_l((D_a \otimes I_N - I_M \otimes D_s) - \left(\frac{Q_a}{z_l} \otimes I_N + I_M \otimes \frac{Q_s}{z_l} \right)) = 0$$

which can be rewritten as

$$\phi_l \left(\left(D_a - \frac{Q_a}{z_l} \right) \otimes I_N + I_M \otimes \left(-D_s - \frac{Q_s}{z_l} \right) \right) = 0.$$

The above equation shows that zero is an eigenvalue of the matrix

$$\left(D_a - \frac{Q_a}{z_l} \right) \otimes I_N + I_M \otimes \left(-D_s - \frac{Q_s}{z_l} \right).$$

According to [27, p. 268], zero is an eigenvalue of the above matrix if and only if there exists $\nu \in \mathbb{C}$ such that

$$\nu \in \sigma \left(D_a - \frac{Q_a}{z_l} \right)$$
$$-\nu \in \sigma \left(-D_s - \frac{Q_s}{z_l} \right) \tag{16}$$

where $\sigma(A)$ denotes the *spectrum* of matrix A. Expression (16) is equivalent to stating that there exist ϕ_{a_l} and ϕ_{s_l} such that

$$\phi_{a_l} \left(D_a - \frac{Q_a}{z_l} - \nu I_M \right) = 0$$
$$\phi_{s_l} \left(\nu I_N - D_s - \frac{Q_s}{z_l} \right) = 0. \qquad \square$$

Lemma 1 allows us to solve (12) by decomposing the original system, into two subsystems: the arrival subsystem of (13) that features a Markov-modulated arrival process and a constant service rate ν, and the subsystem described in (14) with a constant arrival rate ν and a Markov-modulated fluid service process. A similar decomposition argument can be found in [28] where (13) and (14) are shown to be sufficient conditions for z_l to be a solution to (12). Lemma 1 provides both necessary and sufficient conditions for this decomposition to exist.

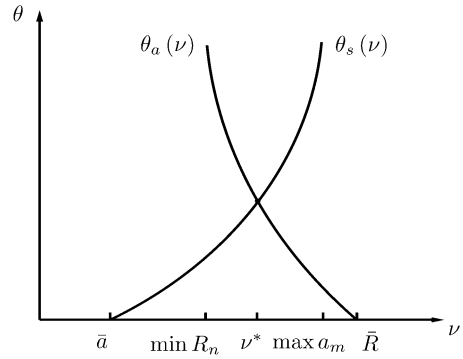

Fig. 4. $\theta_a(\nu)$ and $\theta_s(\nu)$ as a function of ν.

Based on the equilibrium queue-length distribution of the buffer (11), a number of performance metrics can be computed. A simple and important one is the probability of buffer overflow, which can be expressed as

$$\Pr\{L > x\} = 1 - \langle \pi(x), \mathbf{1} \rangle.$$

However, computing the exact probability of buffer overflow for a generic system may be difficult. A more tractable and widely adopted performance measure is the large deviations of the buffer occupancy [31]. In practice, buffers are often large and their decay rates of buffer overflow probabilities are determined primarily by the large deviation principle governing each queue. From this perspective, the QoS metric θ becomes

$$\begin{aligned} \theta &= -\lim_{x\to\infty} \frac{\log \Pr\{L > x\}}{x} \\ &= -\lim_{x\to\infty} \frac{\log(1 - \langle \pi(x), \mathbf{1} \rangle)}{x} \\ &= -\lim_{x\to\infty} \frac{\log(\sum_{l=1}^{k} \alpha_l \langle \phi_l, \mathbf{1} \rangle e^{z_l x})}{x} \\ &= -\max_{l \in \{1,\ldots,k\}} \text{Real}\{z_l\}. \end{aligned} \tag{17}$$

In other words, the QoS metric θ of the system is the absolute value of the largest negative real eigenvalue satisfying (12).

Let the absolute values of the maximum negative eigenvalues for the aggregate system and its individual components be denoted by

$$\theta = -\max\{\text{Real}\{z\} < 0 : \det(zD - Q) = 0\}$$
$$\theta_a(\nu) = -\max\{\text{Real}\{z\} < 0 : \det(zD_a - z\nu I_M - Q_a) = 0\}$$
$$\theta_s(\nu) = -\max\{\text{Real}\{z\} < 0 : \det(z\nu I_N - zD_s - Q_s) = 0\}.$$

It is well-known [32]–[34] that for an irreducible generator matrix Q_a and a real positive diagonal matrix D_a, $\theta_a(\nu)$ is continuous and monotonically increasing from zero to infinity as ν ranges from the mean rate to the peak rate, i.e., $\nu \in [\bar{a}, \max_m a_m]$. Similarly, $\theta_s(\nu)$ is continuous and monotonically decreasing from infinity to zero for $\nu \in [\min_n R_n, \bar{R}]$. Therefore, if $\max_m a_m > \min_n R_n$ and $\bar{a} < \bar{R}$ then there exists a $\nu^* \in [\bar{a}, \max_m a_m]$ such that $\theta_a(\nu^*) = \theta_s(\nu^*)$, as illustrated in Fig. 4. On the other hand, if $\max_m a_m \leq \min_n R_n$, the buffer is always empty and hence $\theta = \infty$. The following theorem asserts that, for a stable system, the large deviation principle associated with the joint system is identical to that governing the two subsystems with parameter ν^*.

Theorem 1: Let Q_a and Q_s be irreducible, recurrent generator matrices, and let D_a and D_s be nonnegative diagonal matrices. If the

system is stable (i.e., $\bar{a} < \bar{R}$), then there exists a $\nu^* \in [\bar{a}, \bar{R}]$ such that

$$\theta = \theta_a(\nu^*) = \theta_s(\nu^*).$$

Proof: Denote the value where these two functions meet by $\theta^* = \theta_a(\nu^*) = \theta_s(\nu^*)$. Clearly, $\nu^* \in [\bar{a}, \bar{R}]$. We need to show that $\theta^* = \theta$. Assume not, then $\theta < \theta^*$ by the minimality of θ. In addition, lemma 1 implies that there exists a $\nu_0 \in \mathbb{C}$ such that z_0 is an eigenvalue of both decoupled systems and $\theta = \mathrm{Real}\{z_0\}$. It follows from the minimality of $\theta_a(\nu)$ and $\theta_s(\nu)$ that $\theta_a(\nu_0) \leq \theta < \theta_a(\nu^*)$ and $\theta_s(\nu_0) \leq \theta < \theta_s(\nu^*)$. From the monotonicity of $\theta_a(\nu)$, we conclude that $\nu_0 > \nu^*$. However, from the monotonicity of $\theta_s(\nu)$, we get $\nu_0 < \nu^*$. This is a contradiction. We then conclude that $\theta = \theta^*$. □

From theorem 1, we find that once ν^* is determined, the QoS metric θ of the system can be obtained by analyzing the behavior of the two independent subsystems. Define

$$\beta(\theta) = \theta_a^{-1}(\theta)$$
$$\alpha(\theta) = \theta_s^{-1}(\theta).$$

For a specific QoS parameter θ^*, $\beta(\theta^*)$ is the effective bandwidth of the arrival process [20], and $\alpha(\theta^*)$ is the effective capacity of the service process [21]. Under the conditions of Theorem 1, the QoS parameter θ^* is the unique solution to the equation

$$\nu^* = \beta(\theta^*) = \alpha(\theta^*). \tag{18}$$

Note that in (18), ν^* is the effective bandwidth of the arrival process and the effective capacity of the service process under the QoS constraint θ^*. A QoS constraint θ_0 for the aggregate system is said to be achievable if and only if $\theta^* \geq \theta_0$. Since $\beta(\theta)$ is monotonically increasing in θ and $\alpha(\theta)$ is monotonically decreasing, the QoS constraint θ_0 can be fulfilled if and only if

$$\beta(\theta_0) \leq \alpha(\theta_0). \tag{19}$$

IV. Achievable Rate-Regions For A Two-User System

In this section, we characterize the achievable rate-region of the user-cooperation system depicted in Fig. 2 when operating under QoS constraint θ_0. Because the system employs FDM, the wireless links between the users and their common destination can be modeled as independent Gilbert-Elliott channels. Assume that both wireless channels have the same expected power gain. The generator matrix Q_{si} corresponding to the modulating Markov process of user i is given by

$$Q_{si} = \begin{bmatrix} -\kappa_i e^{-\eta_i^2} & \kappa_i e^{-\eta_i^2} \\ \kappa_i - \kappa_i e^{-\eta_i^2} & -\kappa_i + \kappa_i e^{-\eta_i^2} \end{bmatrix}$$

the drift matrix D_{si} is equal to

$$D_{si} = \begin{bmatrix} 0 & 0 \\ 0 & R_i \end{bmatrix}$$

where κ_i denotes the exponential decay rate of the channel of user i, and R_i and η_i are respectively the selected coderate and decoding threshold of that same user, as defined in (6). When the two users do not cooperate, user i sets up a wireless connection to the destination using its own physical resources, power P_i and spectral bandwidth allocation W_i. The effective bandwidth of source i, as described in Section II can then be expressed as [35]

$$\beta_i(\theta_0, a_i) = \frac{\theta_0 a_i - \lambda_{ai} - \mu_{ai} + \sqrt{(\theta_0 a_i - \lambda_{ai} - \mu_{ai})^2 + 4\theta_0 a_i \lambda_{ai}}}{2\theta_0} \tag{20}$$

where a_i denotes the peak rate of the underlying on-off source. Similarly, the effective capacity of the wireless channel of user i is [37]

$$\alpha_i(\theta_0, W_i, P_i) = \max_{R_i} \left\{ \frac{\theta_0 R_i + \kappa_i - \sqrt{(\theta_0 R_i + \kappa_i)^2 - 4\theta_0 R_i \kappa_i e^{-\eta_i^2}}}{2\theta_0} \right\}.$$

Recall that the value of η_i depends implicitly on P_i, W_i, and R_i, as seen in (6). According to (19), the peak rate pair (a_1, a_2) is achievable under QoS parameter θ_0 if and only if the effective bandwidth of the traffic generated by user i is less than the effective capacity of the corresponding wireless channel, i.e.,

$$\beta_i(\theta_0, a_i) \leq \alpha_i(\theta_0, W_i, P_i) \tag{21}$$

for $i = 1, 2$. Since the wireless channels are orthogonal, using (20) and (21) we can solve for the maximum supported arrival peak-rate a_i^* for user i subject to the QoS constraint θ_0. The achievable rate-region of the noncooperative FDM system is in the form of a rectangle limited by the maximum supported peak-rates of the two links under QoS parameter θ_0

$$a_i \leq a_i^* = \alpha_i(\theta_0, W_i, P_i) \left(1 + \frac{\mu_{ai}}{\theta_0 \alpha_i(\theta_0, W_i, P_i) + \lambda_{ai}} \right). \tag{22}$$

Now, consider the situation where the two wireless users cooperate by taking advantage of the AWGN interuser channels. We assume that user i assigns a fraction γ_i of its power and bandwidth to the exchange of information with its counterpart. If the expected gain of the inter-user channel is G, then its Shannon capacity is given by

$$r_i(\gamma_i) = \gamma_i W_i \log \left(1 + \frac{G P_i}{N_0 W_i} \right).$$

The power and bandwidth remaining for the uplink connection between user i and the destination become $(1 - \gamma_i)P_i$ and $(1 - \gamma_i)W_i$, respectively. The effective capacity for the resulting wireless channel can be expressed as

$$\nu_i(\gamma_i) = \alpha_i(\theta_0, (1 - \gamma_i)W_i, (1 - \gamma_i)P_i).$$

It is clear from the system model described in Section II that the inter-user traffic originating from user i is an on-off process with peak-rate $r_i(\gamma_i)$. This traffic is modulated by the same two-state Markov chain that modulates the original source. Therefore, the effective bandwidth of this traffic can be expressed as $\beta_i(\theta_0, r_i(\gamma_i))$. Similarly, the portion of the traffic generated by user i which is stored locally and sent directly to the destination is an on-off process with peak-rate $a_i - r_i(\gamma_i)$. The effective bandwidth of this local traffic then becomes $\beta_i(\theta_0, a_i - r_i(\gamma_i))$.

Independence of the traffic generated by the two users and the additivity property of the effective bandwidth for independent sources [15] imply that the total effective bandwidth of the input process to buffer i is the sum of the effective bandwidths of the local traffic and the inter-user traffic coming from its counterpart. Equation (19) states that a QoS constraint θ_0 is achievable if and only if the total effective bandwidth of the incoming traffic is smaller than the effective capacity of the offered service. In the present case, this condition yields two inequalities

$$\begin{aligned} \beta_1(\theta_0, a_1 - r(\gamma_1)) + \beta_2(\theta_0, r(\gamma_2)) &\leq \nu_1(\gamma_1) \\ \beta_2(\theta_0, a_2 - r(\gamma_2)) + \beta_1(\theta_0, r(\gamma_1)) &\leq \nu_2(\gamma_2). \end{aligned} \tag{23}$$

Since $\beta_1(\theta_0, a_1 - r(\gamma_1))$ and $\beta_2(\theta_0, a_2 - r(\gamma_2))$ are both nonnegative, the values of the parameter pair (γ_1, γ_2) are further constrained by

$$\begin{aligned} \beta_2(\theta_0, r(\gamma_2)) &\leq \nu_1(\gamma_1) \\ \beta_1(\theta_0, r(\gamma_1)) &\leq \nu_2(\gamma_2). \end{aligned}$$

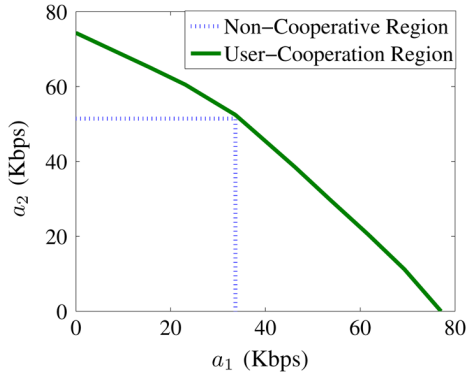

Fig. 5. Comparison of the achievable rate-regions when $\theta = 0.001$.

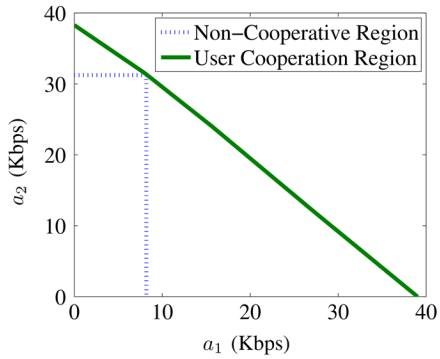

Fig. 6. Comparison of the achievable rate-regions when $\theta = 0.01$.

Let $\mathcal{C}$ denote the set of pairs of the form (γ_1, γ_2) for which the above inequalities hold. For any $(\gamma_1, \gamma_2) \in \mathcal{C}$, the achievable rate-region of the cooperative system, which we denote by $\mathcal{R}(\theta_0, \gamma_1, \gamma_2)$, is found to be

$$\begin{aligned} a_1 &\le r_1(\gamma_1) + (\nu_1(\gamma_1) - \beta_2(\theta_0, r_2(\gamma_2))) \\ &\quad \times \left(1 + \frac{\mu_{a1}}{\theta_0(\nu_1(\gamma_1) - \beta_2(\theta_0, r_2(\gamma_2))) + \lambda_{a1}}\right) \\ a_2 &\le r_2(\gamma_2) + (\nu_2(\gamma_2) - \beta_1(\theta_0, r_1(\gamma_1))) \\ &\quad \times \left(1 + \frac{\mu_{a2}}{\theta_0(\nu_2(\gamma_2) - \beta_1(\theta_0, r_1(\gamma_1))) + \lambda_{a2}}\right). \end{aligned} \tag{24}$$

The achievable rate-region of the user-cooperation scheme under QoS constraint θ_0 is then given by

$$\mathcal{R}(\theta_0) = \bigcup_{(\gamma_1,\gamma_2)\in\mathcal{C}} \mathcal{R}(\theta_0, \gamma_1, \gamma_2).$$

Note that the achievable rate-region of the noncooperative system is given by $\mathcal{R}(\theta_0, 0, 0)$. It is therefore a subset of $\mathcal{R}(\theta_0)$. The boundary of $\mathcal{R}(\theta_0)$ can be obtained by maximizing a_2 over (γ_1, γ_2) while keeping a_1 fixed in (24) and *vice versa*. The solution of the boundary problem can be obtained by standard optimization techniques such as the Lagrange multiplier method.

The comparison between the achievable rate-regions is illustrated through an example. Numerical values of the parameters for the wireless channels and the arrival processes used in this example appear in Table I. The two wireless channels are assumed to have the same expected power gains. However, the channel of user 2 changes faster than that of user 1, with $\kappa_2 > \kappa_1$. Suppose that the gain of the AWGN interuser channel is one ($G = 1$). The achievable rate-region of the system under the cooperative scheme is compared to that of the noncooperative scheme in Figs. 5 and 6, where the numerical values for θ are equal to 0.001 and 0.01, respectively. From these figures, we see that the achievable rate-region of the cooperative system is strictly

TABLE I
SYSTEM PARAMETERS

$N_0 = 10^{-6}$ W/Hz	Noise power spectral density
$W_1 = W_2 = 11$ MHz	Bandwidth
$P_1 = P_2 = 100$ mW	Received power
$\kappa_1 = 10^2$ sec^{-1}	Decay parameter of channel 1
$\kappa_2 = 10^3$ sec^{-1}	Decay parameter of channel 2
$\lambda_{a1}^{-1} = \lambda_{a2}^{-1} = 650$ ms	Average silent period
$\mu_{a1}^{-1} = \mu_{a2}^{-1} = 352$ ms	Average talk burst

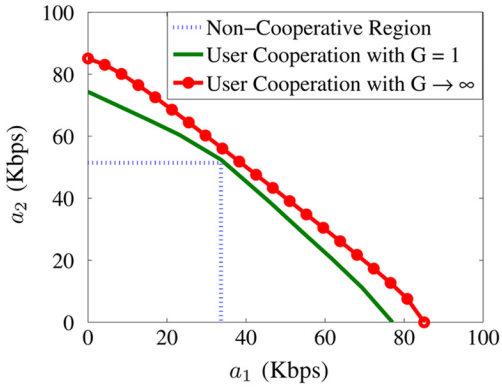

Fig. 7. Comparison of the achievable rate-regions when $\theta_0 = 0.001$.

larger than the region of the traditional FDM system. We note that, even though the expected channel gains of the two wireless channels are the same, there is a large imbalance between the maximum supported peak-rates of the two users under QoS constraint θ_0. This can be explained by the fact that the channel memory of user 2 decays faster than the channel memory of user 1, resulting in a higher order of time-diversity for user 2 [36]. Futhermore, the asymmetry between the maximum achievable rates of the two users increases as the QoS constraint becomes more stringent. Both figures suggest that, under strict QoS requirements, user cooperation provides an efficient means to share radio resources fairly among users.

In the idealized scenario where $G \to \infty$, the users can exchange an arbitrary amount of information at no extra cost in terms of power and bandwidth. Let r_i be the rate at which user i sends information to its counterpart through the interuser channel when its source is *on*. The effective bandwidth of the interuser traffic is $\beta_i(\theta_0, r_i)$, while the effective bandwidth of the excess traffic stored in the local buffer becomes $\beta_i(\theta_0, a_i - r_i)$. From (19), we know that the rate-pair (a_1, a_2) is achievable under QoS constraint θ_0 provided that

$$\begin{aligned} \beta_1(\theta_0, a_1 - r_1) + \beta_2(\theta_0, r_2) &\le \alpha_1(\theta_0, W_1) \\ \beta_2(\theta_0, a_2 - r_2) + \beta_1(\theta_0, r_1) &\le \alpha_2(\theta_0, W_2). \end{aligned} \tag{25}$$

We note that the effective bandwidth is a concave function, with strict inequality over a nontrivial set of values [15]. There are therefore situations where peak-rates a_1 and a_2 can be supported through user cooperation, but not by a traditional FDM system.

These results are easier to understand through an example. For the system parameters listed in Table I, but with $G \to \infty$, the achievable rate-region of the cooperative system is plotted along with that of the noncooperative system in Fig. 7 for $\theta = 0.001$, and in Fig. 8 for $\theta = 0.01$.

As shown in the figures, user-cooperation provides a significant statistical gain over noncooperative system in terms of achievable rates. This gain becomes larger as the QoS constraint becomes more stringent. We can infer from Fig. 7 that the sum peak-rate, $a_1 + a_2$, increases when the two users are cooperating through a perfect inter-user channel, as discussed above. The achievable rate-region of the user-cooperation scheme gets larger as the quality of the inter-user channel

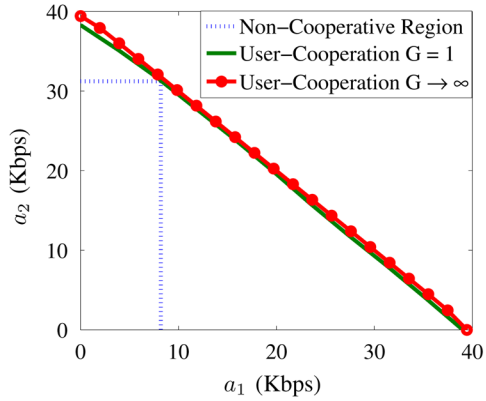

Fig. 8. Comparison of the achievable rate-regions when $\theta_0 = 0.01$.

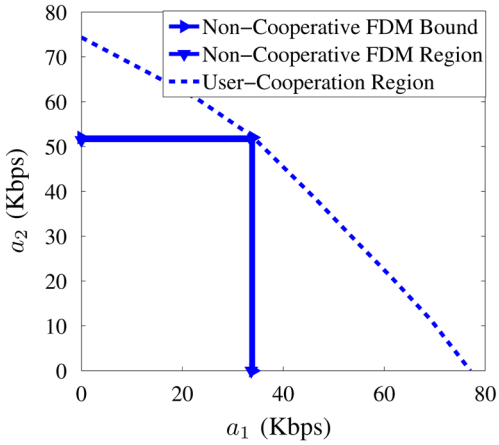

Fig. 9. Comparison of the rate-regions for $\theta_0 = 0.01$.

improves. Yet this gain is less significant when the QoS constraint θ_0 becomes large. This can be explained by the fact that, as θ_0 increases, the effective capacity of each variable channel decreases dramatically [37]. Because the throughput of the AWGN inter-user channel does not vary with θ_0, this latter channel behaves more like an idealized channel to the users as the QoS constraint becomes increasingly stringent. Thus, user-cooperation seems to be beneficial as long as the channel gain of the AWGN inter-user channel is adequate.

V. Alternative System Models

So far we have shown that, under various QoS constraints, the achievable rate-region of the cooperative strategy is significantly larger than that of the noncooperative FDM system. The FDM model for the noncooperative system is employed to circumvent mathematical difficulties that arise from interuser interference. Moreover, to keep our abstract model simple, we assume that the interuser traffic is transmitted instantaneously to the other users. That is, there is no buffer associated with the interuser channel. A valid criticism of our model is that the FDM assumption may unfairly penalize the performance of the noncooperative system, as compared to its performance when successive interference cancellation is used at the destination. Another observation regarding our model is the fact that having a queue for the interuser channel may improve the performance of the cooperative system. In this section, we consider these more elaborate systems and discuss their impacts on performance analysis. In particular, we argue that the intuition gained from the simpler model holds for these more intricate models as well.

A. Successive Interference Cancelation

From [1], we know that the maximum achievable rate-region for the multi-access channel encompasses the region achieved by a FDM system. This greater flexibility is obtained by using successive interference cancellation at the receiver. For any fading realization (h_1, h_2), the achievable rate-region of a multi-access channel is the polyhedron bounded by the inequalities

$$\begin{aligned} C_1 &\le W \log_2\left(1 + \frac{|h_1|^2 P_1}{N_0 W}\right) \\ C_2 &\le W \log_2\left(1 + \frac{|h_2|^2 P_2}{N_0 W}\right) \\ C_1 + C_2 &\le W \log_2\left(1 + \frac{|h_1|^2 P_1 + |h_2|^2 P_2}{N_0 W}\right). \end{aligned} \tag{26}$$

Here, P_i is the mean received power of user i and W is the total spectral bandwidth available to the two users. We note that this region is upper bounded by the rate-region of a FDM system with twice the spectral bandwidth ($2W$). In particular, consider a FDM system with allocation $W_1 = W_2 = W$, the achievable rate-region of this alternate system is specified by

$$\begin{aligned} C_1 &\le W \log_2\left(1 + \frac{|h_1|^2 P_1}{N_0 W}\right) \\ C_2 &\le W \log_2\left(1 + \frac{|h_2|^2 P_2}{N_0 W}\right). \end{aligned} \tag{27}$$

Clearly, the region defined by (26) is a subset of (27). Thus, we can upper bound the rate-region of a noncooperative multiple-access system that uses successive interference cancelation by that of a FDM system that has double the spectral bandwidth of the original system.

Assume the total system bandwidth is $W = 11$ MHz. We compare the achievable rate-region of the user-cooperation system to the region corresponding to a FDM system with twice the bandwidth in Fig. 9. The latter region is an absolute upper bound for the region of a noncooperative system that supports successive interference cancellation. The result suggests that user-cooperation may offer significant gains in performance over a noncooperative system that uses successive interference cancelation. This behavior is explained, partly, by the fact that additional spectral bandwidth offers diminishing returns in terms of effective capacity. The effective capacity of a QoS constrained system appears to level off even before the system enters its information theoretic wideband regime [37].

For the system parameters listed in Table I, the effective capacities $\alpha_i(\theta_0, W, P)$ of the two wireless channels are plotted as a function of spectral bandwidth W in Fig. 10. We can see from the figure that the effective capacities of the two channels level off rapidly once W is large enough. This explains why doubling the spectral bandwidth of a FDM system does not necessarily improve the effective capacity by much. This limitation is also partly due to the underlying assumption that channel state information is not available at the transmitters. Incidentally, users cannot transmit at the (error-free) instantaneous Shannon capacity, and therefore they do not benefit from the additional degrees of freedom associated with a larger spectral bandwidth. When the available spectrum is large enough, the queueing behavior of the system is dominated by the holding time of the service *off* state, which is independent of the channel bandwidth [37].

B. Cooperation With Interuser Buffers

A straightforward generalization of the user-cooperation scheme proposed in Section II is to add buffers for the interuser traffic of both transmitters. In this case, the inter-user traffic can be buffered locally and, as such, data can be sent to the other user even when the source is in its *off* state. This more flexible setup can only improve system

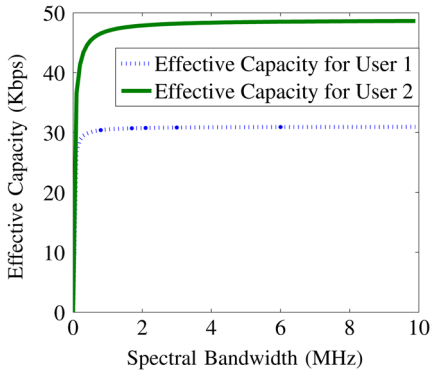

Fig. 10. Effective capacity for $\theta_0 = 0.01$.

performance and thereby enlarge the achievable rate-region of the user-cooperation system. To characterize the achievable rate-region of this communication scheme, we need to derive the effective bandwidth of the departure process of the interuser traffic. Since the gain of the interuser channel is constant, the effective bandwidth of the interuser traffic can be analytically characterized [20]. However, we elect not to compute the exact achievable rate-region for the problem at hand when buffers are used for the interuser channels. Rather, we provide tight upper and lower bounds for the periphery of the achievable rate-region.

The user-cooperation system without interuser buffers can be thought of as a special case of the cooperative system described above. It corresponds to the situation where the interuser buffers remain empty at all times. In this sense, the achievable rate-region of the user-cooperation system without interuser buffers serves as a lower bound for the achievable rate-region of the user-cooperation system with interuser buffers. On the other hand, the achievable rate-region of the idealized user-cooperation system ($G \to \infty$) serves as an upper bound for the user-cooperation system with inter-user buffers. Indeed, as G approaches infinity, the constant service rates of the interuser channels become increasingly large. This insures that these buffers remain empty. The boundary of the achievable rate-region for the buffered user-cooperation system must lie between the dashed line and the solid line in Fig. 7 and in Fig. 8. Since the gap between the upper and lower bounds is quite narrow, the gains associated with using interuser buffers for the system under study must be somewhat marginal. The tedious analysis of the more elaborate buffered scheme provides little additional insight about the possible benefits of user-cooperation in wireless systems, it is therefore not included in this correspondence.

VI. Conclusion

In this correspondence, we proposed a simple user-cooperation scheme that works under the assumption that channel state information is only available at the receivers, not at the transmitters. A Markov model was introduced to capture the unreliable nature of the wireless environment. For a fixed coderate, the overall performance of the wireless channel is modeled as a two-state Gilbert–Elliott model. Based on this Markov assumption, we introduced a cross-layer approach to analyze the performance of the wireless communications systems under strict QoS constraints.

The achievable rate-region of the proposed user-cooperation scheme is characterized and it is compared to the region of a noncooperative system. Numerical results suggest that cooperation yields a large gain over traditional systems. User-cooperation can therefore provide wireless users with the flexibility to better share system resources. Our queueing analysis also hints at the fact that overall performance depends heavily on the time correlation of the underlying physical channel. In that sense, effective capacity is much more sensitive to higher order statistics than, say, ergodic capacity or outage capacity. It is therefore imperative to use channel models that are amenable to analysis while providing an accurate representation of reality.

Acknowledgment

The authors would like to thank Prof. Randall Berry and two anonymous referees for their comments and suggestions, which have helped improve the presentation of the correspondence.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.
[2] S. Verdu and S. Shamai, "Spectral efficiency of cdma with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.
[3] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.
[4] R. Berry and R. Gallagar, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.
[5] E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–596, Nov. 1999.
[6] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—Part I: System description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.
[7] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity - part II: Implementation aspects and performance analysis," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1939–1948, Nov. 2003.
[8] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block code from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999.
[9] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IIEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.
[10] J. N. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.
[11] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3037–3063, Sept. 2005.
[12] A. Høst-Madsen, "Capacity bounds for cooperative diversity," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1522–1544, Apr. 2006.
[13] L. Liu, J.-F. Chamberland, and S. Miller, "The uplink achievable rate region of a user cooperation scheme," in *Proc. IEEE Canadian Workshop on Information Theory*, Jun. 2005, pp. 163–166.
[14] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type uas channel," *Queueing Syst.*, vol. 9, pp. 17–28, 1991.
[15] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 329–343, Jun. 1993.
[16] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass markov fluids and other atm sources," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 424–428, Aug. 1993.
[17] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Transa. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.
[18] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.
[19] F. P. Kelly, "Effective bandwidths at multi-type queues," *Queueing Syst.*, vol. 9, pp. 5–15, 1991.
[20] C.-S. Chang, *Performance Guarantees in Communication Networks*. New York: Springer-Verlag, 2000.
[21] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.
[22] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assrance," *IEEE Trans. Veh. Technol.*, vol. 53, no. 5, pp. 1547–1557, Sept. 2004.
[23] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Trans. Veh. Technol.*, vol. 54, no. 3, pp. 1198–1206, May 2005.
[24] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996.

[25] H. S. Wang and N. Moayeri, "Finite-state markov channel — A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[26] W. Turin and R. vanNobelen, "Hidden markov modelling of flat fading channels," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 9, pp. 1809–1817, Dec. 1998.

[27] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge University Press, 1994.

[28] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Prob.*, vol. 20, pp. 646–676, 1993.

[29] T. E. Stern and A. I. Elwalid, "Analysis of separable markov-modulated rate models for information-handling systems," *Adv. Appl. Prob.*, vol. 23, pp. 105–139, 1991.

[30] J. R. Norris, *Markov Chains, ser. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge, U.K.: Cambridge University Press, 1998.

[31] A. Ganesh, N. O'Connell, and D. Wischik, *Big Queues*. New York: Springer, 2004.

[32] G. Debreu and I. N. Herstein, "Nonnegative square matrices," *Econometrica*, vol. 21, pp. 597–607, 1953.

[33] J. E. Cohen, "Random evolutions and the spectral radius of a nonnegative matrix," *Math. Proc. Cambridge Philosoph. Soc.*, vol. 86, pp. 345–350, 1979.

[34] J. E. Cohen, "Convexity of the dominant eigenvalue of an essentially nonnegative matrix," *Proc. Amer. Math. Soc.*, pp. 657–658, 1981.

[35] F. P. Kelly, S. Zachary, and I. B. Ziedins, *Stochastic Networks: Theory and Applications*, ser. Royal Statistical Society Lecture Notes. Oxford, U.K.: Oxford University Press, 1996.

[36] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge University Press, 2005.

[37] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1767–1777, May 2007.