Performance Analysis of Wireless Hybrid-ARQ Systems with Delay-Sensitive Traffic

Nirmal Gunaseelan, Lingjia Liu, Jean-Francois Chamberland, and Gregory H. Huff

***Abstract*—The design of wireless communication schemes tailored to real-time traffic requires an analysis framework that goes beyond the traditional criterion of data throughput. This work considers an approach that relates physical system parameters to the queueing performance of wireless links. The potential benefits of multi-rate techniques such as hybrid-ARQ are assessed in the context of delay-sensitive traffic using large deviations. A continuous-time Markov channel model is employed to partition the instantaneous data-rate received at the destination into a finite number of states, each representing a mode of operation of the hybrid-ARQ scheme. The proposed methodology accounts for the correlation of the wireless channel across time, which is computed in terms of level-crossing rates. The tail asymptote governing buffer overflow probabilities at the transmitter is then used to provide a measure of overall performance. This approach leads to a characterization of the effective capacity of the system which, in turn, is applied to quantify the performance advantages of hybrid-ARQ over traditional schemes.**

***Index Terms*—Communication system, hybrid-ARQ, effective capacity, delay, quality of service (QoS), wireless networks, wireless systems.**

I. Introduction

THE rising popularity and vast deployment of wireless access points have extended the reach of the Internet beyond the traditional confines of the wired world. Wireless networking offers an added level of convenience to mobile users, and it provides an efficient means to interconnect multiple devices. In their original form, wireless routers were not designed to support delay-sensitive applications such as voice over Internet protocol (VoIP), video conferencing and utility computing. Nevertheless, these real-time applications are pushing the limits of current technology. This trend considerably raises the demand for robust, high-bandwidth data connections with stringent delay requirements. Still, offering low-latency connections remains a difficult task in distributed wireless environments, partly due to the time-varying quality of wireless channels. To meet this challenge, novel design paradigms and better analysis tools are needed [1], [2].

Paper approved by T.-S. P. Yum, the Editor for Packet Access and Switching of the IEEE Communications Society. Manuscript received February 23, 2009; revised June 16, 2009 and October 1, 2009.

N. Gunaseelan is with The MathWorks, Inc., 3 Apple Hill Dr., Natick, MA 01760 (e-mail: nirmal.gunaseelan@mathworks.com).

L. Liu is with DTL, Samsung Electronics, 1301 E Lookout Dr., Richardson, TX 75082 (e-mail: lliu@sta.samsung.com).

J.-F. Chamberland and G. H. Huff are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128 (e-mail: {chmbrlnd, ghuff}@tamu.edu).

This material is based, in part, upon work supported by the National Science Foundation (NSF) under Awards CCF-0830696 and CCF-0747363, and by the Texas Norman Hackerman Advanced Research Program under Grant No. 000512-0168-2007.

Digital Object Identifier 10.1109/TCOMM.2010.04.090104

In this paper, we favor an integrated approach and seek to identify how advanced schemes at the physical layer impact queueing performance at the link layer. The importance of our contribution is twofold. First, we construct a model that relates physical channel parameters to queueing performance. We then proceed to use this methodology to assess the potential benefits of multi-rate techniques in the context of delay-sensitive communications. We consider an environment where users operate over wireless fading channels.

The variations in channel quality experienced by every user may be attributable to multi-path effects, shadowing and inter-user interference. In certain contexts, channel fluctuations can improve performance. For instance, water-filling across time generally enhances the ergodic capacity of wireless systems [3]. However, for delay-sensitive traffic, service variations tend to hinder rapid transmissions and induce higher probabilities of buffer overflow. These adverse effects can often be mitigated by the presence of partial feedback at the transmitter, which helps ensure the reliable and swift completion of data transfers. Such feedback mechanisms are present in most communication systems [4, Ch. 4], and they are assumed to be in place throughout.

Hybrid automatic repeat-request (hybrid-ARQ) in the form of incremental redundancy is one such scheme that takes advantage of feedback for the speedy delivery of data packets [5]–[8]. Hybrid-ARQ can be employed to combat fading by having the system adapt seamlessly to changing channel conditions. In such schemes, packets remain in the transmit buffer until the corresponding information is decoded successfully at the destination. When decoding fails, the system recovers by asking the transmitter to send additional redundancy bits, thereby taking advantage of the coding gain offered by low-rate codes.

Herein, we study the benefits associated with hybrid-ARQ in queued systems, while accounting for channel correlation. We model the wireless environment using a flat-fading channel whose envelope follows a Rayleigh distribution [9]. The combined effects of fading and code performance are then approximated using a finite-state, continuous-time Markov process [10, Ch. 2]. The coherence time of the wireless link, which provides a measure of channel memory [11, Ch. 4, p. 165], is implicitly integrated into our framework through the transition rate of the Markov process. A long coherence time translates into infrequent transitions in the induced Markov process, whereas a shorter coherence time results in more rapid transitions. While more sophisticated and accurate models exist at the physical layer [12], [13], their complexities seem to preclude conducting queue-based analyses of the corresponding systems.

Previous work on resource allocation for wireless communication systems provides valuable insights on strategies to maximize throughput [14]–[16]. Yet, the literature on resource allocation in the context of delay-sensitive applications is not fully developed. One of the distinguishing characteristics between the traditional information theoretic viewpoint and our approach lies in the fact that time-correlation has a considerable impact on the performance of delay-sensitive systems, whereas the ergodic capacity of a wireless channel with partial state information can be somewhat impervious to time-dependencies [17].

Much like it affects the reliability function of block codes [18], time-dependence also influences the queueing behavior of dynamic systems. For slow fading channels, the probability of decoding failure is likely to be strongly correlated over successive blocks. To provide a true quality of service over fading channels, it is therefore imperative to consider the queueing behavior of the system [1]. The popular and convenient assumption of independent and identically distributed channel realizations across time is inadequate for most real-time applications; it results in an overly optimistic assessment of system performance.

The large-deviation principle governing buffer overflow probabilities affords a foundation for popular performance measures in queueing systems [19, Ch. 9]. One prominent asymptotic tool that often appears in wireless communications is the effective capacity. It can be employed to characterize system performance under statistical delay guarantees by identifying the maximum input data-rate that a system can support subject to an exponential constraint on the complementary cumulative distribution function of the queue [20], [21]. In this work, we use large deviations as a bridge between queueing and information theory. By conducting a queue-based analysis on systems having different transmission rates, we are able to quantify the benefits of using hybrid-ARQ over more rudimentary implementations.

We restrict our attention to the study of multi-rate systems at the physical layer. Delay-sensitive systems overcome channel variations through a number of strategies. Most notably, power control is a critical component of low-latency cellular systems, leading to a natural tradeoff between delay and energy consumption [22], [23]. Herein, we seek to identify the benefits of a multi-rate transceiver with incremental redundancy, and choose to do so in isolation by depriving the system of these other valuable mechanisms. A more encompassing analysis would jointly consider code-rate, power and frequency adaptation as functions of channel state and queue length. However, this latter approach would certainly be much more challenging and may not be amenable to the type of analysis we successfully carry below. Furthermore, it could cloud the gains attributable to hybrid-ARQ, as it may be impossible to distinguish between the various components of the system. We thus leave this more involved topic as a possible future endeavor, and restrict our attention to the case where transmit power, number of antennas and carrier frequency are fixed.

The remainder of the paper is organized as follows. In Section II, we review fading channels and concepts related to hybrid-ARQ. We describe the specific channel models we adopt for performance comparisons. We also construct a mathematical representation of the physical layer, and obtain throughput optimal data-rates for the two systems we wish to study. In Section III, we derive the equilibrium distributions for the queued systems. Based on large deviations, we obtain expressions and numerical results for buffer overflow probabilities and effective capacities. Section IV contains a numerical study of buffer occupancy for the Ornstein-Uhlenbeck channel; queueing behavior for this more elaborate model is found to be very sensitive to arrival rates; this offers supporting evidence for the theoretical findings of Section III. Overall, our analysis sheds light on the potential benefits of multi-rate systems in the context of delay-sensitive traffic. Conclusions and possible future directions are contained in Section V.

II. System Model

We seek to quantify the benefits of multi-rate techniques in the context of delay-sensitive communications. Concurrently, we wish to characterize the impact of time-dependencies on overall performance in terms of effective capacity. To cast our problem in a queueing theory framework, we develop an abstract model for the effects of the wireless channel and specify system evolution in terms of physical layer parameters.

A. Wireless Channel

Signals transmitted over a wireless medium in urban environments get reflected and scattered before reaching their intended destinations, which creates fading. Inter-user interference further exacerbates the situation, as it also alters the quality of demodulated signals at the destination. A channel model that accounts for mean path-attenuation, interference and noise is the celebrated Rayleigh fading channel model. In this setting, transmitted signal $x(t)$ is subject to multipath fading $h(t)$ and additive noise $w(t)$ [11, Ch. 4]. This can be represented mathematically by $r(t) = h(t)x(t) + w(t)$. The noise component $w(t)$ captures attenuation, interference and noise; and is often assumed to form an independent white Gaussian process. In rich scattering environments, the marginal distribution of the function $h(t)$ is well-modeled using a Rayleigh distribution for the envelope and a uniform distribution for the phase component [3].

Specifying the time statistics of the function $h(t)$ is a more difficult task. For a terminal moving at constant velocity through a zero-mean proper complex Gaussian field, the time-autocorrelation function of the envelope process $|h(t)|$ can be modeled using the zeroth-order Bessel function of the first kind [24, Ch. 3, p. 74]. An alternative autocorrelation function for $|h(t)|$ can be obtained by assuming that the in-phase and quadrature components of $h(t)$ form independent stationary Ornstein-Uhlenbeck processes [25, Ch. 5]. In this model, the correlation between two samples of $h(t)$ decays exponentially with time. Both approaches lead to reasonable, consistent channel models that may be useful for simulations, but are overly complicated for the type of queueing analysis we wish to conduct. To obtain a mathematically tractable problem formulation, we take a different approach and elect not to specify the time-evolution of the wireless channel explicitly. Rather, we model the combined effects of channel fluctuations and error-control coding using a continuous-time,

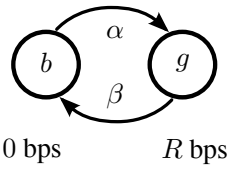

Fig. 1. A continuous-time two-state Markov process can be employed to represent the operation of a wireless link built with a traditional ARQ mechanism. Here, g denotes a *good* state where packets are transmitted at a data-rate of R bits per second, and b is a bad state where no packets get through. The Markov transition rates are represented by α and β.

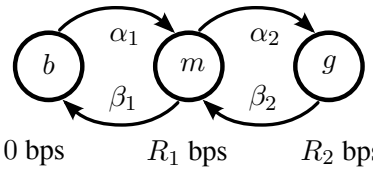

Fig. 2. The operation of a wireless link equipped with hybrid-ARQ can be modeled as a finite-state Markov process, where the instantaneous throughput from the source to the destination depends on the current quality of the physical channel. In *good* state g, the data-rate is R_2 bits per second; in *moderate* state m, the data-rate becomes R_1 bits per second; and finally in *bad* state b, no packets get through. In this figure, $\alpha_1, \beta_1, \alpha_2, \beta_2$ symbolize the Markov transition rates of the system.

finite-state Markov chain. This approach provides an adequate representation of the wireless system to evaluate the potential performance gains of multi-rate schemes.

To explain the specifics of our methodology, we must first review the fundamentals of incremental redundancy. Hybrid-ARQ generally refers to a collection of protocols that adapt to changing channel conditions through a number of techniques. A prominent class of hybrid-ARQ methods is the type of incremental redundancy proposed by Mandelbaum [26], where a terminal responds to retransmission requests by sending additional parity bits to the destination. The receiver aggregates these extra bits with the previously gathered data, allowing for enhanced error correction capability. Recent advances in error-control coding greatly facilitate the development of efficient hybrid-ARQ techniques where the transmitter adapts seamlessly to changing channel conditions, to the extent allowed by feedback. For instance, rateless codes have made the implementation of hybrid-ARQ even easier [27], [28].

B. Combined Model

Consider a non-adaptive wireless system that employs an error-correcting code with a fixed data-rate. When decoding fails, the receiver naively discards the acquired information and requests retransmission of the data block using a physical-layer ARQ mechanism. From a higher-layer perspective, this system can be in one of two modes. The wireless connection can be operating in a desirable state during which data packets are decoded successfully at the receiver and data flows through the system at a constant bit-rate. Alternatively, if the signal-to-noise ratio at the receiver drops below a certain threshold, decoding fails and the flow of data from the transmitter to the destination halts. The data transmission process only resumes once channel conditions improve again. If the physical channel exhibits correlation through time, it is natural to expect the two-state system abstraction to display memory as well. The overall behavior of this two-state model can be approximated by a continuous-time Markov chain, as illustrated in Fig. 1. When the receiver is successfully decoding data blocks, information is conveyed from the source to the destination at a constant data-rate R. On the other hand, under harsh channel conditions, the instantaneous throughput becomes zero.

The operation of a wireless link equipped with hybrid-ARQ can similarly be modeled as a continuous-time, finite-state Markov chain. Consider a simple incremental redundancy mechanism where the source transmits encoded blocks of data through a wireless channel. When the initial decoding of a packet fails, the destination requests a second block of data that contains additional parity bits. The complementary information is combined with the original message, and the destination attempts to jointly decode the aggregate message. If this process fails again, all the received samples are discarded and the transmission process starts anew. The Markov chain representing this system has three possible states: a *good* state where data blocks are decoded on first attempts, a *moderate* state where successful decoding is only possible after extra parity bits are received from the source, and a *bad* state where every decoding attempt fails. To each of these states corresponds an instantaneous throughput rate from the source to the destination. We denote the instantaneous data-rates associated with the *good* and *moderate* states by R_2 and R_1, respectively. The throughput while in the *bad* state is necessarily zero. This system abstraction appears in Fig. 2.

This characterization assumes that error-correcting codes switch abruptly in their behaviors. Probability of decoding failure is very low when the data-rate is less than the instantaneous channel capacity and, conversely, probability of failure becomes nearly equal to one when the data-rate exceeds the Shannon capacity. Second, the sizes of data packets are relatively small compared to the coherence time of the channel. Admittedly, these assumptions are somewhat conflicting in nature. From an information theoretic perspective, the block length of the codewords need to be large for the switching characteristic to come into play. Yet, the block length cannot be so large that it affects queueing behavior significantly or lasts over multiple realizations of the channel. This assumed time-scale decoupling, although not ideal, is common in the literature and can be considered a good first step in understanding the interaction between queueing and the physical layer [21], [22], [29]. A detailed study of secondary effects that may affect performance is left as a worthy future endeavor.

C. Physical Layer Parameters

The physical layer of a wireless communication system can be described in terms of a few fundamental parameters [18]. For additive white Gaussian noise, the instantaneous capacity can be written as

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

where P represents the transmit power, W is the system bandwidth and N_0 denotes the noise power spectral density.

TABLE I
SYSTEM PARAMETERS USED IN NUMERICAL EXAMPLES

$f_c = 1.9$ GHz	Carrier frequency
$W = 540$ KHz	Spectral bandwidth
$P = 1.62 \times 10^{-6}$ W	Received power
$N_0 = 10^{-12}$ W/Hz	Noise power spectral density

To ensure reliable decoding at the destination, the data-rate R of a wireless system should be less than the capacity expression of (1). This relationship provides a close approximation of code performance under the assumption that coding delay is smaller than the coherence time of the channel. This condition, $R < C(t)$, implies that packets at the destination can be decoded properly provided that

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

In the two-layer model of Section II-B, condition (2) specifies the threshold above which the channel is in its *good* state, and below which the channel is *bad*. The average throughput of the corresponding wireless connection is then given by

$$R \Pr\{|h(t)| > \eta\} = R e^{-\eta^2}.$$

Based on this relation, the maximum average throughput of this ARQ system can be obtained by optimizing over all admissible values of R.

For the three-layer link of Section II-B, there are two thresholds to choose, η_2 and η_1. These thresholds govern the transitions between *good*, *moderate* and *bad* channel states. They are related to the data-rates and physical parameters as follows,

$$\eta_1 = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R_1}{W}} - 1\right)}, \quad \eta_2 = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R_2}{W}} - 1\right)}.$$

We adopt the convention $R_2 > R_1$. The channel is in its *good* state when $|h(t)| \geq \eta_2$; it is in its *moderate* state when $\eta_2 > |h(t)| \geq \eta_1$; and it is *bad* otherwise. For fixed values of R_1 and R_2, the average throughput of this wireless connection can be expressed as $R_1 e^{-\eta_1^2} + (R_2 - R_1) e^{-\eta_2^2}$. Values for R_1 and R_2 can be selected as to optimize average throughput.

Consider a wireless link with the system parameters of Table I (e.g., 3GPP LTE system with three resource blocks of 180 KHz). Then, the optimized average throughput of the simple ARQ scheme is equal to 440.4 Kbps. In comparison, the more elaborate hybrid-ARQ mechanism provides an average throughput of 586.5 Kbps under the same conditions, a sizable gain. A throughput analysis offers an initial view of the potential benefits of multi-rate implementations. Still, we emphasize that this performance characterization can be carried out based solely on the marginal distributions of the coded systems.

More information is needed to conduct a queueing analysis of the wireless connection. The invariant distribution of the underlying Markov chain alone does not dictate the queueing behavior of the system; transition rates from one state to another are also critical. These rates are related to time-correlation at the physical layer, and they play an instrumental role in the performance of delay-sensitive communication systems. The generator matrix of a two-state continuous-time Markov chain can be expressed as

$$Q_2 = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}.$$

For the ARQ system of Section II-B, α represents the transition rate from the *bad* state to the *good* state; and β denotes the transition rate in the opposite direction. This is illustrated in Fig. 1. Markov models have been used in the context of communication systems in the past [9], [30], [31]; and techniques have been proposed to relate variables α and β to the level-crossing rates at the physical layer [11], [32], [33].

The envelope level crossing rate $\mathrm{LCR}(\eta)$ is defined as the rate (in crossings per second) at which the signal envelope crosses level η in a given direction. For a Rayleigh fading channel with a Bessel autocorrelation function, the relation between Doppler frequency f_d and the level-crossing rate is characterized by (see, e.g. [9], [34])

$$\mathrm{LCR}(\eta) = \sqrt{2\pi\eta} f_d e^{-\eta}. \tag{3}$$

Estimates for α and β can be obtained from the stationary behavior of the system by matching the average transition rates of the Markov model to the crossing rates given by (3).

From the stationary distribution, we have

$$\Pr\{|h(t)| \leq \eta\} = \frac{\beta}{\alpha + \beta} = 1 - e^{-\eta^2}$$
$$\Pr\{|h(t)| > \eta\} = \frac{\alpha}{\alpha + \beta} = e^{-\eta^2}.$$

Recall that the marginal distribution of the normalized envelope process is assumed Rayleigh. The level crossing rate of the Markov chain yields

$$\mathrm{LCR}(\eta) = \frac{\alpha\beta}{\alpha + \beta}.$$

These equations completely characterize the two-state Markov model in terms of physical channel parameters.

For our hybrid-ARQ system, the generator matrix of the continuous-time Markov chain can be written as

$$Q_3 = \begin{bmatrix} -\alpha_1 & \alpha_1 & 0 \\ \beta_1 & -(\beta_1 + \alpha_2) & \alpha_2 \\ 0 & \beta_2 & -\beta_2 \end{bmatrix}.$$

The variables α_1, β_1, α_2, and β_2 can again be specified using the parameters and properties of the underlying wireless channel, together with the selected data-rates R_1 and R_2. The stationary distribution of the envelope process yields the equations

$$\Pr\{|h(t)| \leq \eta_1\} = \frac{\beta_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} = 1 - e^{-{\eta_1}^2}$$
$$\Pr\{\eta_1 < |h(t)| \leq \eta_2\} = \frac{\alpha_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} = e^{-{\eta_1}^2} - e^{-{\eta_2}^2}$$
$$\Pr\{|h(t)| > \eta_2\} = \frac{\alpha_1\alpha_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2} = e^{-{\eta_2}^2},$$

which dictate the invariant distribution of the Markov chain. The additional relations necessary to uniquely determine Q_3

are provided by the level crossing rates

$$\mathrm{LCR}(\eta_1) = \frac{\alpha_1\beta_1\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2}$$
$$\mathrm{LCR}(\eta_2) = \frac{\alpha_1\alpha_2\beta_2}{\beta_1\beta_2 + \alpha_1\alpha_2 + \alpha_1\beta_2}.$$

Collectively, these equations specify the operation of our abstract, encoded hybrid-ARQ system. A graphical interpretation of these quantities appears in Fig. 2.

D. Packetized System and Queueing Model

We analyze below the queueing behavior of the wireless connection. In our framework, the instantaneous service rate of the queue is equal to the instantaneous throughput of the wireless channel. To capture queueing behavior, the arrival process feeding the queue must be specified. Taking into consideration previous work on this topic, and to accommodate the type of analysis we wish to carry, we assume that packets arrive in the queue according to a Poisson process. We use λ to denote the parameter of this process, and point out that the memoryless property of Poisson arrivals greatly simplifies analysis. We also assume that the size of the received packets are independent and identically distributed exponential random variables. Together, these two premises form the starting point of our queueing system. We emphasize that, conditioned on the state of the wireless channel being fixed, the packet departure process is memoryless.

As mentioned above, we take letters to represent the state of the Markov service process in both the traditional and hybrid-ARQ systems. For the two-state system, we use g to denote the *good* state and b for the *bad* state. In the three-state model, we employ additional letter m to signify that the system is operating in its *moderate* state. We represent the number of packets in the buffer using integers $n = 0, 1, 2, \ldots$ The aggregate state of the overall system can therefore be specified by a letter and a number. For example, b_0 represents a *bad* channel state with an empty queue, while g_2 is a *good* channel state with two packets in the buffer. Note that, in the later parts of the paper, we also use b_n, m_n and g_n to represent the probabilities that the system is in particular states. This intentional abuse of notation simplifies mathematical expressions derived in the paper, and it is straightforward to distinguish between a state and its probability from the context.

An abstract representation of the two-layer model appears in Fig. 3. The system forms a continuous-time Markov process. The value on each arrow represents the rate at which the process makes transitions on the link. We assume that the buffer has no limit on the number of packets it can store, and hence there are no packet losses due to drops at the transmitter. From a physical-layer point of view, the data-rate is R bits per second whenever the channel operates in its *good* state. This data-rate, along with the average packet size, determines the rate μ at which packets depart from the buffer when the channel is on.

We write the balance equations [10, p. 124] for the two-state channel of Fig. 3 as

$$b_n(\alpha + \lambda) = b_{n-1}\lambda + g_n\beta \tag{4}$$
$$g_n(\mu + \beta + \lambda) = g_{n-1}\lambda + g_{n+1}\mu + b_n\alpha. \tag{5}$$

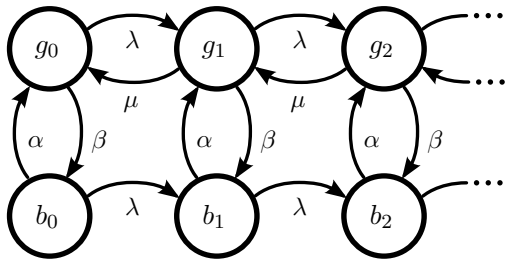

Fig. 3. Once the state of the queue is taken into consideration, the traditional ARQ system becomes a continuous-time Markov chain with a countable number of states.

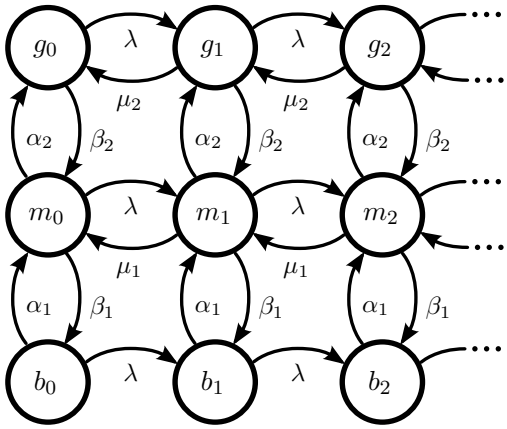

Fig. 4. The hybrid-ARQ system also becomes a Markov chain with a countable number of states after accounting for the queue occupancy.

The boundary conditions on these equations can be expressed as

$$\begin{aligned} b_0(\alpha + \lambda) &= g_0\beta \\ g_0(\beta + \lambda) &= b_0\alpha + g_1\mu. \end{aligned} \tag{6}$$

The boundary conditions on the stationary distribution for the channel states are

$$\begin{aligned} \Pr\{bad\} &= \sum_{n=0}^{\infty} b_n = \frac{\beta}{\beta + \alpha} \\ \Pr\{good\} &= \sum_{n=0}^{\infty} g_n = \frac{\alpha}{\beta + \alpha}. \end{aligned} \tag{7}$$

Under stable conditions, the balance equations along with their boundary conditions dictate the behavior of the queued system.

We next turn to the three-layer hybrid-ARQ model described previously. This system can be in a *bad* state, a *moderate* state or a *good* state. The state-space of the associated Markov model is shown in Fig. 4. In state g_n, information flows from the source to the destination at data-rate R_2; while in state m_n, the service rate is R_1. Irrespective of the channel state, the arrival process is Poisson with parameter λ. From a packet perspective, departures occur at a rate μ_1 when the system operates in its *moderate* state, and at rate μ_2 under *good* conditions.

The balance equations for the hybrid-ARQ system become

$$g_n(\mu_2+\beta_2+\lambda) = g_{n+1}\mu_2 + m_n\alpha_2 + g_{n-1}\lambda \tag{8}$$

$$m_n(\alpha_2+\beta_1+\lambda+\mu_1) = m_{n+1}\mu_1 + b_n\alpha_1 + g_n\beta_2 + m_{n-1}\lambda \tag{9}$$

$$b_n(\alpha_1+\lambda) = b_{n-1}\lambda + m_n\beta_1 \tag{10}$$

with equilibrium boundary conditions

$$\begin{aligned} g_0(\beta_2+\lambda) &= m_0\alpha_2 + g_1\mu_2 \\ m_0(\lambda+\alpha_2+\beta_1) &= g_0\beta_2 + m_1\mu_1 + \alpha_1 b_0 \\ b_0(\alpha_1+\lambda) &= m_0\beta_1. \end{aligned} \tag{11}$$

The requirements on the stationary distribution for the channel states are

$$\begin{aligned} \Pr\{bad\} &= \frac{\beta_1\beta_2}{\beta_1\beta_2+\alpha_1\alpha_2+\alpha_1\beta_2} \\ \Pr\{moderate\} &= \frac{\alpha_1\beta_2}{\beta_1\beta_2+\alpha_1\alpha_2+\alpha_1\beta_2} \\ \Pr\{good\} &= \frac{\alpha_1\alpha_2}{\beta_1\beta_2+\alpha_1\alpha_2+\alpha_1\beta_2}. \end{aligned} \tag{12}$$

This completes the definition of our abstract models for the traditional and hybrid-ARQ communication systems.

III. Performance Comparison of System Models

In this section, we study the equilibrium distributions of the two queued systems discussed above. Based on this analysis, we can compute relevant performance metrics such as effective capacity and probability of buffer overflow. These performance measures are then utilized to quantify the improvements associated with multi-rate techniques for delay-sensitive traffic over wireless links.

A. Equilibrium Distributions

The invariant distributions of the Markov models are derived from the corresponding recurrence relations using transform methods [35]. From the balance equations of the two-state ARQ model given in (4) & (5), and using probability-generating functions, we get

$$\begin{aligned} (\alpha+\lambda)(B(z)-b_0) &= \lambda z B(z) + \beta(G(z)-g_0) \\ (\mu+\beta+\lambda)(G(z)-g_0) &= \lambda z G(z) + \alpha(B(z)-b_0) \\ &\quad + \mu z^{-1}(G(z)-zg_1-g_0). \end{aligned}$$

For stable systems, the constants b_0, g_0 and g_1 can be resolved using boundary conditions (6) and (7). The procedure is explained in greater details in Appendix A. The probabilities of the system being in its various states are equal to

$$b_n = \frac{r}{2q}\left(\frac{1}{(p-q)^{n+1}} - \frac{1}{(p+q)^{n+1}}\right) \tag{13}$$

$$g_n = \frac{r}{2q\beta}\left(\frac{\frac{\alpha+\lambda}{p-q}-\lambda}{(p-q)^n} - \frac{\frac{\alpha+\lambda}{p+q}-\lambda}{(p+q)^n}\right), \tag{14}$$

where the following constants have been introduced for convenience,

$$\begin{aligned} p &= \frac{(\alpha+\beta+\mu+\lambda)}{2\lambda} \\ q &= \frac{\sqrt{(\alpha+\beta+\mu+\lambda)^2-4\mu(\alpha+\lambda)}}{2\lambda} \\ r &= \frac{\beta}{\lambda^2}\left(\frac{\alpha\mu}{\alpha+\beta}-\lambda\right). \end{aligned} \tag{15}$$

The Markov transition rates α and β are specified using the properties of the underlying wireless channel, as explained in Section II-C. Evaluating p and q reveals that $p > q > 0$. Furthermore, under stable conditions where expected arrival rate is less than expected departure rate [36, Ch. 9], we get $(p-q) > 1$. Hence, b_n and g_n vanish as n goes to infinity. The dominating decay factor in the complementary cumulative distribution function of the queue is given by [37, Ch. 1, p. 7]

$$\begin{aligned} &-\lim_{n\to\infty}\frac{1}{n}\log\left(\Pr\{L>n\}\right) \\ &= -\lim_{n\to\infty}\frac{1}{n}\log\left(\sum_{\ell=n+1}^{\infty}(b_\ell+g_\ell)\right) = \log(p-q). \end{aligned} \tag{16}$$

We have thus characterized the equilibrium distribution of the ARQ system model and its governing tail asymptote.

We can proceed in a similar manner with the three-state hybrid-ARQ model. From the balance equations (8)–(10), we get a system of linear equations in the transform domain,

$$\begin{aligned} &-\alpha_2 M(z) + (\mu_2+\beta_2+\lambda-\lambda z-\mu_2 z^{-1})G(z) \\ &\quad = (1-z^{-1})\mu_2 g_0 \\ &-\alpha_1 B(z) + (\alpha_2+\beta_1+\lambda+\mu_1-\lambda z-\mu_1 z^{-1})M(z) \\ &-\beta_2 G(z) = (1-z^{-1})\mu_1 m_0 \\ &\qquad (\alpha_1+\lambda-\lambda z)B(z) - \beta_1 M(z) = 0. \end{aligned}$$

We can then solve for the power series $B(z)$, $M(z)$ and $G(z)$ using standard linear algebra techniques. Boundary conditions on this continuous-time Markov process and its embedded jump chain can be employed to resolve constants m_0 and g_0. Details are contained in Appendix B.

In the case of the hybrid-ARQ system, the dominating decay factor in the complementary cumulative distribution function of the queue is obtained by identifying the roots of the quartic form

$$\begin{aligned} q(z) &= \lambda^3 z^4 - (\alpha_1+\alpha_2+\beta_1+\beta_2+2\lambda+\mu_1+\mu_2)\lambda^2 z^3 \\ &\quad + \Big(\alpha_1\alpha_2+\alpha_1\beta_2+\alpha_1\lambda+\alpha_1\mu_1+\alpha_1\mu_2+\alpha_2\lambda+\alpha_2\mu_2 \\ &\quad + \beta_1\beta_2+\beta_1\lambda+\beta_1\mu_2+\beta_2\lambda+\beta_2\mu_1+\lambda^2+2\lambda\mu_1 \\ &\quad + 2\lambda\mu_2+\mu_1\mu_2\Big)\lambda z^2 - \Big(\alpha_1\alpha_2\mu_2+\alpha_1\beta_2\mu_1+\alpha_1\lambda\mu_2 \\ &\quad + \alpha_1\lambda\mu_1+\alpha_1\mu_1\mu_2+\alpha_2\lambda\mu_2+\beta_1\lambda\mu_2+\beta_2\lambda\mu_1 \\ &\quad + \lambda^2\mu_2+\lambda^2\mu_1+2\lambda\mu_1\mu_2\Big)z + (\alpha_1+\lambda)\mu_1\mu_2. \end{aligned}$$

The tail asymptote governing the buffer is determined by the smallest root of $q(z)$ exceeding one. This can be seen by looking at the general form of the equilibrium distribution in the transform domain, which is contained in Appendix B. For obvious reasons, we continue our analysis of the hybrid-ARQ system through a numerical study.

B. Performance Analysis

As mentioned in the introduction, the effective capacity $\lambda(\theta)$ represents the maximum input data-rate that a system can accommodate subject to a tail-asymptote requirement θ on the complementary cumulative distribution of the queue. In the case of the traditional ARQ system, it is possible to get a closed-form characterization of the effective capacity. Let L be a non-negative random variable distributed according to the invariant measure of the transmit buffer. Then, $\Pr\{L > n\}$ denotes the probability that the buffer occupancy exceeds threshold n. From (16), we know that the tail asymptote governing the complementary cumulative distribution of the buffer in the traditional ARQ system is given by

$$-\lim_{n\to\infty}\frac{1}{n}\log\Pr\{L>n\}=\log(p-q),$$

where p, q are as defined in (15). In this particular setting, the effective capacity can be expressed as

$$\lambda(\theta)=\sup_{\lambda\geq 0}\left\{-\lim_{n\to\infty}\frac{1}{n}\log\Pr\{L>n\}\geq\theta\right\}$$
$$=\sup_{\lambda\geq 0}\left\{\lambda^2\left(e^{2\theta}-e^{\theta}\right)-\lambda\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right)+\mu\alpha\geq 0\right\}.$$

The roots of this quadratic equation are

$$\lambda=\frac{\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right)}{2\left(e^{2\theta}-e^{\theta}\right)}$$
$$\pm\frac{\sqrt{\left(e^{\theta}(\alpha+\beta+\mu)-\mu\right)^2-4\mu\alpha\left(e^{2\theta}-e^{\theta}\right)}}{2\left(e^{2\theta}-e^{\theta}\right)},$$

with the smallest root providing an explicit formula for the effective capacity, $\lambda(\theta)$.

The steady-state distribution of the three-state model is harder to solve analytically. We therefore resort to a numerical characterization of the dominant roots. We identify the maximum admissible arrival rate $\lambda(\theta)$ as a function of the exponential decay on the complementary cumulative distribution of the buffer. To illustrate our findings, we provide a concrete example. We study a system with the parameters of Table I. Through this example, we can see the benefits of adding multi-rate functionalities on the performance of the system. The effective capacity is plotted as a function of the decay constraint on the complementary cumulative distribution of the buffer in Fig. 5. Vehicular speed in the figure is taken to be 15 m/s (34 mph), and expected packet size is 512 bytes. Overall behavior remains essentially unaltered at 27 m/s (60 mph).

Over the past several years, hybrid-ARQ has been shown to significantly improve the performance of wireless connections. It reduces the average number of transmissions and allows systems to operate close to capacity in a seamless fashion. The contribution of our analysis is to provide a better understanding of the benefits of hybrid-ARQ in the context of delay-sensitive systems. As the decay constraint θ nears zero, the effective capacity becomes the maximum throughput of the system. The effective capacity starts to decrease as this constraint increases. To reduce congestion in the buffer, the system is forced to limit the arrival rate. We gather from Fig. 5 that multi-rate schemes improve performance. However, it is clear that they cannot be utilized to alleviate the detrimental effects of deep fades and strong time-correlation.

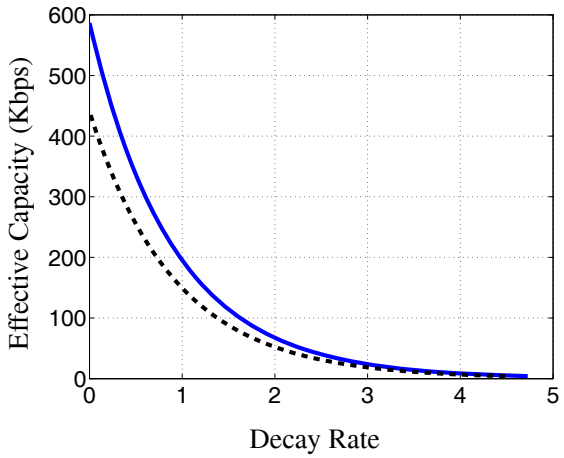

Fig. 5. This figure provides a characterization of the effective capacity as a function of the decay rate on the complementary cumulative distribution of the buffer. The hybrid-ARQ system significantly increases the effective capacity of the system (solid line) over the traditional ARQ system (dashed line). However, it fails to alter the system behavior in that the effective capacity drops very sharply as θ moves away from zero.

The Rayleigh channel under consideration is subject to deep fades where the instantaneous capacity becomes very small. Strong correlation in the channel envelop over time further exacerbates this problem. Together, these two components create long periods during which packets accumulate in the buffer, a highly undesirable tendency. Although hybrid-ARQ adapts to channel conditions, it does not change the statistics of the envelop process. Ultimately, this predicament leads to the sharp decay in effective capacity depicted in Fig. 5. This is in contrast to power control and multi-antenna implementations, which do modify the underlying profile of the channel.

IV. Ornstein-Uhlenbeck Process

The system models employed thus far are in essence Markov approximations to end-to-end wireless communication links. These models are based on two assumptions. First, the state of the system depends on whether the instantaneous capacity of the underlying channel lies above or below prescribed thresholds. Second, the stochastic process representing the time evolution of the corresponding finite-state system is well modeled as a continuous-time Markov chain. This class of Markov models permits a detailed analysis of overall system performance, which includes probability of buffer overflow and effective capacity. This acts as a major incentive in selecting a framework suitable for analysis.

While it is convenient to assume that the quantized system possesses the Markov property, a more common approach is to take the channel itself to be Markov. We emphasize that these two hypotheses are quite different. In particular, the former is much easier to handle mathematically. It is nevertheless straightforward to construct a Rayleigh-fading channel that possesses the Markov property. Below, we review the details of a Markov channel model based on the Ornstein-Uhlenbeck

process [38]. We then use this model to perform a numerical study of the natural tradeoff between delay and throughput.

Let X_t be the solution to the following continuous-time stochastic differential equation

$$dX_t = -\nu X_t dt + \sigma dW_t, \tag{17}$$

where ν, σ are real constants and W_t is a one-dimensional Brownian motion. This solution is a special case of an Ornstein-Uhlenbeck process and possesses the Markov property. It can be expressed in integral form as

$$X_t = X_0 e^{-\nu t} + \int_0^t e^{-\nu(t-s)} \sigma dW_s. \tag{18}$$

The variance of this process at time t can be computed explicitly,

$$\mathrm{Var}[X_t] = \mathrm{E}[(X_0 - \mathrm{E}[X_0])^2] e^{-2\nu t} + \frac{\sigma^2}{2\nu}\left(1 - e^{-2\nu t}\right).$$

By properly selecting the distribution of X_0 and for suitable values of ν and σ, X_t becomes a stationary Gauss-Markov process [38]. A Rayleigh-fading channel that possesses the Markov property can therefore be constructed by assigning independent stationary Ornstein-Uhlenbeck processes to the in-phase and quadrature components of the channel. The marginal distribution of the corresponding fading process $h(t)$ is a zero-mean, proper complex Gaussian measure, as desired.

The caveat in this approach is that the quantized version of the channel becomes a hidden Markov process. This precludes the application of standard queueing results and popular techniques from large-deviation theory, and hence renders analysis much more challenging, if not intractable. These limitations and the vast body of literature on Markov modulated processes explain our early adoption of the simpler system models. On the other hand, the Gauss-Markov problem formulation lends itself to a numerical study, which we conduct next.

The system parameters used in the study are again taken from Table I. The vehicular speed is set to 15 m/s. The Ornstein-Uhlenbeck processes are discretized by sampling at 2 ms intervals, and the correlation coefficient is determined by matching an exponential curve to the first two maximums of the zeroth-order Bessel function of the first kind. The arrival rate is constant, and the buffer is updated every 16 ms. The maximum throughput of the traditional system is equal to 710 Kbps, whereas the hybrid-ARQ link remains stable for arrival rates up to 848 Kbps.

Although the specifics of the more intricate Gauss-Markov channel differ from those of our previous framework, we expect the relative performance of the traditional and hybrid-ARQ systems to be comparable to what we have observed so far. That is, we anticipate hybrid-ARQ to improve throughput, but it is unlikely to alter the very sharp decay in effective capacity as a function of parameter θ. We use numerical simulations to assess the validity of these conjectures.

A large-deviation characterization may be convenient for asymptotic analysis, however it is hardly fitting to a numerical study. In this section, we consider average queue length and mean delay as functions of arrival rate, two performance criteria appropriate for numerical simulations. The average queue length as a function of arrival rate is shown in Fig. 6 for the traditional and hybrid-ARQ systems. Not so surprisingly, the hybrid-ARQ scheme performs better by supporting a higher arrival rate for a given average queue length. In terms of buffer overflow probability, the most significant gains are found at rates close to the maximum throughput of the traditional system. Also, we point to the very sharp increase in queue length as the arrival rate approaches the corresponding optimal throughput. This hints at a very sharp decay in effective capacity as a function of parameter θ, as predicted.

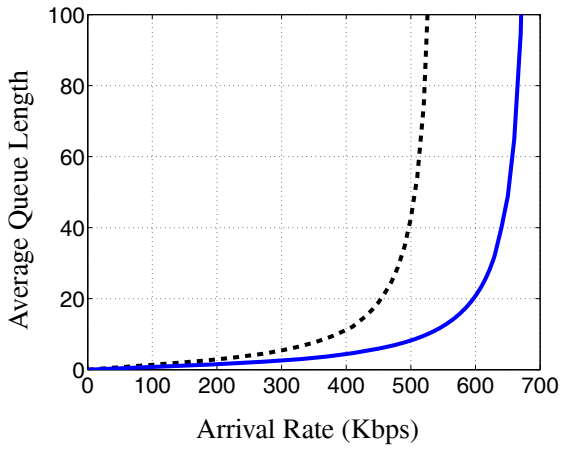

Fig. 6. This figure shows the average queue length as a function of arrival rate for hybrid-ARQ (solid line) and a traditional ARQ system (dashed line). Packets are assumed to be 640 bits in size.

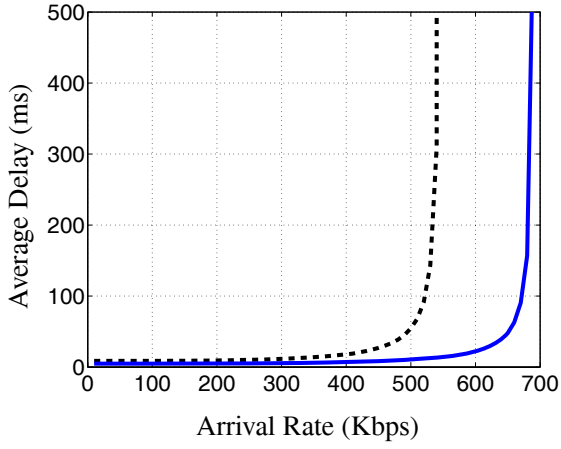

Fig. 7. This figure plots the average delay as a function of arrival rate for hybrid-ARQ (solid line) and a traditional ARQ system (dashed line). The expected delay is obtained by applying Little's law.

We can use Little's law to derive the mean delays associated with our two communication schemes. Recall that the average number of packets in a queue is equal to the arrival rate multiplied by the expected delay. We can therefore compute mean delay as a function of arrival rate in a straightforward manner from the results displayed above. Fig. 7 shows expected delay as a function of arrival rate for the traditional and hybrid-ARQ systems. To put this figure in perspective, a single-mode VoIP system with a five percent probability of delay exceeding 100 ms and a 30 Kbps data rate (including

overhead) leads to an average delay of about 40 ms. We should mention that these parameters depend heavily on the codec employed. The traditional system can support arrival rates up to 480 Kbps while maintaining the average delay below 40 ms, and the hybrid-ARQ implementation can push this number to 640 Kbps. Overall, the multi-rate scheme can support higher arrival rates for a same probability of buffer overflow or delay requirement.

V. Conclusions

Multi-rate systems offer significant performance improvements over traditional ARQ schemes. Hybrid-ARQ has long been recognized as a means to approach ergodic capacity with relatively short code-words by seamlessly adapting to various channel conditions. This feedback mechanism prevents the transmitter from having to code over very long periods of time, a definite advantage for delay-sensitive traffic. Our study points to the benefits of multi-rate schemes in the context of delay-sensitive wireless communications.

Two types of occurrences are especially detrimental to the performance of queued wireless systems. If the marginal distribution of the channel has a high variance, with a significant probability of the received signal being weak, then the distribution of the queue is likely to have a larger mean thereby creating undue delays. This problem gets compounded when the channel is strongly correlated over time. Deep fades in strongly correlated channels lead to large queue build-ups and necessarily longer latencies.

Hybrid-ARQ is beneficial in that it allows the system to operate close to the instantaneous capacity of a channel at all times. Unfortunately, it cannot be employed to alter the profile of this same channel. This is in contrast to multi-antenna implementations, power control loops and frequency-hopping patterns which inherently change the statistical properties of the underlying channel and lead to performance enhancements. Power control can be employed to combat deep fades, whereas multi-antenna arrays produce channel hardening and thus reduce variations in service. The inability of hybrid-ARQ to modify channel statistics explains why the effective capacity curves of Section III-B portray the same very sharp decay for the traditional and hybrid-ARQ systems. The same observation holds for the average queue-length characterization of Fig. 6.

A very sharp decay in effective capacity as a function of decay parameter θ is trouble for delay-sensitive communications. This implies that hybrid-ARQ is unlikely to have a major impact on power consumption for a wireless system that carries real-time traffic. For low-power systems, alternative means will have to be explored to reduce transmit power levels in the context of real-time systems. This is especially important for ad hoc and multi-hop systems because the channel fade levels of two collocated transmitters with distinct destinations may be quite different, leading to excessive interference. The type of centralized interference management used in cellular systems, which include rapid power-control loops, is unlikely to be sufficient for ad hoc networks. Future research directions point to the study of communication schemes that improve effective capacity without inducing strong variations in transmit power. Likely candidates for low-latency communication schemes include multi-antenna implementations and schemes that reduce time-correlation in the instantaneous capacity.

Appendix A
Invariant Distribution of Two-State System

In this section, we compute the equilibrium distribution of the two-state Markov model. We start with the balance equations,

$$b_n(\alpha+\lambda) = b_{n-1}\lambda + g_n\beta$$
$$g_n(\mu+\beta+\lambda) = g_{n-1}\lambda + g_{n+1}\mu + b_n\alpha,$$

and use probability-generating functions to get the desired distribution. The power series are obtained by multiplying the equations above by the proper powers of z, and then summing over all values of $n \geq 1$. This yields

$$(\alpha+\lambda)(B(z)-b_0) = \lambda z B(z) + \beta(G(z)-g_0)$$
$$(\mu+\beta+\lambda)(G(z)-g_0) = \lambda z G(z) + \alpha(B(z)-b_0) + \mu z^{-1}(G(z)-zg_1-g_0)$$

where we have implicitly defined

$$B(z) = \sum_{n=0}^{\infty} b_n z^n \qquad G(z) = \sum_{n=0}^{\infty} g_n z^n.$$

Using boundary conditions defined in (6), we get

$$(\alpha+\lambda-\lambda z)B(z) - \beta G(z) = 0$$
$$-\alpha B(z) + (\mu+\beta+\lambda-\lambda z-\mu z^{-1})G(z) = (1-z^{-1})\mu g_0.$$

For a given value of g_0, this forms a pair of linear equations in two unknowns. It can therefore be solved in a straightforward manner. The value of g_0 is obtained through the last boundary condition, $B(1) = \frac{\beta}{\alpha+\beta}$. After some cancellation, the equilibrium distribution in the transform domain becomes

$$B(z) = \frac{\left(\frac{\alpha\mu}{\alpha+\beta} - \lambda\right)\beta}{z^2\lambda^2 - z\lambda(\alpha+\beta+\mu+\lambda) + \mu(\alpha+\lambda)}$$
$$G(z) = -\frac{\left(\frac{\alpha\mu}{\alpha+\beta} - \lambda\right)(\alpha+\lambda-\lambda z)}{z^2\lambda^2 - z\lambda(\alpha+\beta+\mu+\lambda) + \mu(\alpha+\lambda)}.$$

The steady-state distribution of the system can be derived through partial fraction expansion, followed by taking inverse transforms. First, we express the power series in simpler forms

$$B(z) = \frac{r}{2q}\left(\frac{1}{z-(p+q)} - \frac{1}{z-(p-q)}\right)$$
$$G(z) = \frac{r}{2q\beta}\left(\frac{\lambda(p-q)-\alpha-\lambda}{z-(p-q)} - \frac{\lambda(p+q)-\alpha-\lambda}{z-(p+q)}\right),$$

using the convenient notation defined in (15). The probabilities of the system being in its various states are equal to

$$b_n = \frac{r}{2q}\left(\frac{1}{(p-q)^{n+1}} - \frac{1}{(p+q)^{n+1}}\right)$$
$$g_n = \frac{r}{2q\beta}\left(\frac{\frac{\alpha+\lambda}{p-q}-\lambda}{(p-q)^n} - \frac{\frac{\alpha+\lambda}{p+q}-\lambda}{(p+q)^n}\right),$$

where $n \geq 0$. We note that the dominating decay factor in the complementary cumulative distribution function of the queue is given by the smallest solution to the quadratic form

$$z^2\lambda^2 - z\lambda(\alpha+\beta+\mu+\lambda) + \mu(\alpha+\lambda) = 0.$$

APPENDIX B
INVARIANT DISTRIBUTION OF THREE-STATE SYSTEM

We wish to characterize the equilibrium distribution of the three-state Markov model. However, as seen below, this involves finding the roots of a fourth-order polynomial equation and carrying out partial fraction expansion with respect to these roots. Rather than computing these roots explicitly, we find the form of the quartic equation and use it to capture the asymptotic behavior of the queue. Through large deviations, we know that the asymptotic behavior of the complementary cumulative distribution function of the queue is dominated by the smallest stable root of the characteristic polynomial [37].

We begin with the balance equations for the hybrid-ARQ system given in (8)–(10). Using power series representations, these three equations become

$$\begin{aligned}
&-\alpha_2 M(z) + (\mu_2 + \beta_2 + \lambda - \lambda z - \mu_2 z^{-1})G(z) \\
&\quad = (1 - z^{-1})\mu_2 g_0 \\
&-\alpha_1 B(z) + (\alpha_2 + \beta_1 + \lambda + \mu_1 - \lambda z - \mu_1 z^{-1})M(z) \\
&-\beta_2 G(z) = (1 - z^{-1})\mu_1 m_0 \\
&\qquad (\alpha_1 + \lambda - \lambda z)B(z) - \beta_1 M(z) = 0.
\end{aligned}$$

Note that the boundary conditions listed in (11) have already been leveraged to simplify this system of linear equations. We find the characteristic polynomial by computing the determinant of the 3×3 matrix shown in (19).

After cancellation, taking away a factor of $(1 - z^{-1})$, the characteristic function becomes

$$\begin{aligned}
q(z) &= \lambda^3 z^4 - (\alpha_1 + \alpha_2 + \beta_1 + \beta_2 + 2\lambda + \mu_1 + \mu_2)\lambda^2 z^3 \\
&\quad + \Big(\alpha_1\alpha_2 + \alpha_1\beta_2 + \alpha_1\lambda + \alpha_1\mu_1 + \alpha_1\mu_2 + \alpha_2\lambda + \alpha_2\mu_2 \\
&\quad + \beta_1\beta_2 + \beta_1\lambda + \beta_1\mu_2 + \beta_2\lambda + \beta_2\mu_1 + \lambda^2 + 2\lambda\mu_1 \\
&\quad + 2\lambda\mu_2 + \mu_1\mu_2\Big)\lambda z^2 - \Big(\alpha_1\alpha_2\mu_2 + \alpha_1\beta_2\mu_1 + \alpha_1\lambda\mu_2 \\
&\quad + \alpha_1\lambda\mu_1 + \alpha_1\mu_1\mu_2 + \alpha_2\lambda\mu_2 + \beta_1\lambda\mu_2 + \beta_2\lambda\mu_1 \\
&\quad + \lambda^2\mu_2 + \lambda^2\mu_1 + 2\lambda\mu_1\mu_2\Big)z + (\alpha_1 + \lambda)\mu_1\mu_2.
\end{aligned}$$

Using the adjugate matrix formula, we can solve this matrix equation. The equilibrium distributions in the transform domain can be written as

$$\begin{aligned}
B(z) &= -\frac{\beta_1\beta_2 z\mu_2 g_0}{q(z)} \\
&\quad - \frac{\beta_1(\mu_2 + \beta_2 + \lambda - \lambda z - \mu_2 z^{-1})z\mu_1 m_0}{q(z)} \\
M(z) &= -\frac{\beta_2(\alpha_1 + \lambda - \lambda z)z\mu_2 g_0}{q(z)} \\
&\quad - \frac{(\alpha_1 + \lambda - \lambda z)(\mu_2 + \beta_2 + \lambda - \lambda z - \mu_2 z^{-1})z\mu_1 m_0}{q(z)} \\
G(z) &= -\frac{\alpha_2(\alpha_1 + \lambda - \lambda z)z\mu_1 m_0}{q(z)} + \frac{\alpha_1\beta_1 z\mu_2 g_0}{q(z)} \\
&\quad - \frac{(\alpha_1 + \lambda - \lambda z)(\alpha_2 + \beta_1 + \lambda + \mu_1 - \lambda z - \mu_1 z^{-1})z\mu_2 g_0}{q(z)}.
\end{aligned}$$

The values of m_0 and g_0 can be obtained using the remaining boundary conditions. The normalization property, $B(1) + M(1) + G(1) = 1$, gives one condition. Stability and its requirements, $\lim_{n\to\infty} g_n = \lim_{n\to\infty} m_n = 0$, provide the second equation. Since Q is irreducible and recurrent, the invariant measure is unique [10, Ch. 3, p. 124].

REFERENCES

[1] R. A. Berry and E. M. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 59-68, Sept. 2004.

[2] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," *IEEE Wireless Commun.*, vol. 12, no. 1, pp. 3-11, Feb. 2005.

[3] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619-2692, Oct. 1998.

[4] M. Guizani, *Wireless Communications Systems and Networks*, ser. Information Technology: Transmission, Processing and Storage. Springer, 2004.

[5] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971-1988, July 2001.

[6] M. Zorzi and R. R. Rao, "On the use of renewal theory in the analysis of ARQ protocols," *IEEE Trans. Commun.*, vol. 44, no. 9, pp. 1077-1081, Sept. 1996.

[7] K. Narayanan and G. Stuber, "A novel ARQ technique using the turbo coding principle," *IEEE Commun. Lett.*, vol. 1, no. 2, pp. 49-51, Mar. 1997.

[8] S. Sesia, G. Caire, and G. Vivier, "Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes," *IEEE Trans. Commun.*, vol. 52, no. 8, pp. 1311-1321, Aug. 2004.

[9] 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.

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

[11] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Prentice Hall PTR, 2001.

[12] V. Erceg, D. Michelson, S. Ghassemzadeh, L. Greenstein, A. Rustako Jr., P. Guerlain, M. Dennison, R. Roman, D. Barnickel, S. Wang, and R. Miller, "A model for the multipath delay profile of fixed wireless channels," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 399-410, Mar. 1999.

[13] A. Ganesan and A. Sayeed, "A virtual input-output framework for transceiver analysis and design for multipath fading channels," *IEEE Trans. Commun.*, vol. 51, no. 7, pp. 1149-1161, July 2003.

[14] W. Wu, A. Arapostathis, and S. Shakkottai, "Optimal power allocation for a time-varying wireless channel under heavy-traffic approximation," *IEEE Trans. Automatic Control*, vol. 51, no. 4, pp. 580-594, Apr. 2006.

[15] M. Zafer and E. Modiano, "Optimal rate control for delay-constrained data transmission over a wireless channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4020-4039, Sept. 2008.

[16] M. Neely, "Intelligent packet dropping for optimal energy-delay tradeoffs in wireless downlinks," *IEEE Trans. Automatic Control*, vol. 54, no. 3, pp. 565-579, Mar. 2009.

[17] A. J. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1986-1992, Nov. 1997.

[18] R. G. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.

[19] C.-S. Chang, *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.

[20] G. de Veciana, G. Kesidis, and J. Walrand, "Resource management in wide-area ATM networks using effective bandwidths," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1081-1090, Aug. 1995.

[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] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135-1149, May 2002.

[23] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058-3068, Aug. 2007.

[24] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[25] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., ser. Universitext. Springer, 2003.

[26] D. Mandelbaum, "An adaptive-feedback coding scheme using incremental redundancy," *IEEE Trans. Inf. Theory*, vol. 20, no. 3, pp. 388-389, May 1974.

$$\begin{bmatrix} 0 & -\alpha_2 & (\mu_2+\beta_2+\lambda-\lambda z-\mu_2 z^{-1}) \\ -\alpha_1 & (\alpha_2+\beta_1+\lambda+\mu_1-\lambda z-\mu_1 z^{-1}) & -\beta_2 \\ (\alpha_1+\lambda-\lambda z) & -\beta_1 & 0 \end{bmatrix} \tag{19}$$

[27] M. Luby, T. Gasiba, T. Stockhammer, and M. Watson, "Reliable multimedia download delivery in cellular broadcast networks," *IEEE Trans. Broadcasting*, vol. 53, no. 1, pp. 235-246, Mar. 2007.

[28] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551-2567, June 2006.

[29] R. Prakash and V. V. Veeravalli, "Centralized wireless data networks with user arrivals and departures," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 695-713, Feb. 2007.

[30] M. Zorzi and R. R. Rao, "On the statistics of block errors in bursty channels," *IEEE Trans. Commun.*, vol. 45, no. 6, pp. 660-667, June 1997.

[31] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688-1692, Nov. 1999.

[32] C.-D. Iskander and P. T. Mathiopoulos, "Analytical level crossing rates and average fade durations for diversity techniques in Nakagami fading channels," *IEEE Trans. Commun.*, vol. 50, no. 8, pp. 1301-1309, Aug. 2002.

[33] W. Turin and R. van Nobelen, "Hidden Markov modelling of flat fading channels," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 9, pp. 1809-1817, Dec. 1998.

[34] F. Babich and G. Lombardi, "A Markov model for the mobile propagation channel," *IEEE Trans. Veh. Technol.*, vol. 49, no. 1, pp. 63-73, Jan. 2000.

[35] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*, 4th ed., ser. Probability and Statistics. Wiley, 2008.

[36] P. Bremaud, *Markov Chains*. Springer, 2008.

[37] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., ser. Stochastic Modelling and Applied Probability. Springer, 1998.

[38] 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.

Nirmal Gunaseelan (S'04, M'09) received the B.Eng. degree in Electronics and Communications Engineering from PSG College of Technology, Anna University, India, in 2005, and the M.S. degree in Electrical Engineering from Texas A&M University, in 2008. He is currently working at The MathWorks in the Signal Processing and Communications modeling space.

Lingjia Liu received the Ph.D. degree from Texas A&M University in Electrical Engineering, in 2007; and the B.S. degree from Shanghai Jiao Tong University in Electronic Engineering, in 2003. He joined the Dallas Telecommunication Laboratory, Samsung Electronics, in 2008, where he is currently a senior research engineer engaged in the design of physical-layer communication schemes. Lingjia Liu is the recipient of a Texas Telecommunications Engineering Consortium (TxTEC) Fellowship through the Department of Electrical and Computer Engineering at Texas A&M University. He was also awarded the Global Samsung Best Paper Award during the Global Samsung Technical Conference, in 2008.

Jean-François Chamberland (S'98, M'04, SM'09) received the Ph.D. degree, in 2004, from the University of Illinois at Urbana-Champaign; the M.S. degree, in 2000, from Cornell University; and the B.S. degree, in 1998, from McGill University, Canada. He joined Texas A&M University, in 2004, where he is currently an Assistant Professor in the Department of Electrical and Computer Engineering. Among the awards he has received for research and teaching are a Young Author Best Paper Award from the IEEE Signal Processing Society, in 2006; and an Early Career Development (CAREER) Award from the National Science Foundation, in 2008.

Gregory H. Huff (S'01, M'06) received the B.S., M.S., and Ph.D. degrees, all in Electrical Engineering, from the University of Illinois at Urbana-Champaign in 2000, 2003, and 2006, respectively. He has been an Assistant Professor with the Electromagnetics and Microwave Laboratory in the Department of Electrical and Computer Engineering at Texas A&M University since 2006. He also participated in the Lewis' Educational and Research Collaborative Internship Program (LERCIP) in 2005 at the NASA Glenn Research Center. In addition to several other honors, Gregory Huff is the recipient of a 2008 Presidential Early Career Award for Scientists and Engineers (PECASE) through the Army Research Office's Young Investigator Program and a 2008 National Science Foundation CAREER Award. His current research includes multifunctional antenna techniques, biologically-inspired reconfigurable antennas and deformable smart skins, and integrative/adaptive antennas for sensor networks and high speed communications.