Resource Allocation and Quality of Service Evaluation for Wireless Communication Systems Using Fluid Models

Lingjia Liu, Parimal Parag, *Student Member, IEEE*, Jia Tang, Wei-Yu Chen, and Jean-François Chamberland, *Member, IEEE*

***Abstract*—Wireless systems offer a unique mixture of connectivity, flexibility, and freedom. It is therefore not surprising that wireless technology is being embraced with increasing vigor. For real-time applications, user satisfaction is closely linked to quantities such as queue length, packet loss probability, and delay. System performance is therefore related to, not only Shannon capacity, but also quality of service (QoS) requirements. This work studies the problem of resource allocation in the context of stringent QoS constraints. The joint impact of spectral bandwidth, power, and code rate is considered. Analytical expressions for the probability of buffer overflow, its associated exponential decay rate, and the effective capacity are obtained. Fundamental performance limits for Markov wireless channel models are identified. It is found that, even with an unlimited power and spectral bandwidth budget, only a finite arrival rate can be supported for a QoS constraint defined in terms of exponential decay rate.**

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

I. Introduction

MOTIVATED by the emergence of wireless technologies and by the constantly increasing demand for connectivity, we study the interplay between resource allocation at the physical layer and quality of service (QoS) in wireless communication systems. Radio resources typical of a wireless communication system include spectral bandwidth and power. Much research in recent years has been devoted to developing techniques and strategies that enhance the spectral efficiency of wireless systems [1], [2]. The prevalent framework used to evaluate these techniques is information theory, with the inseparable concept of Shannon capacity. While this framework is suitable for an analysis of maximum throughput, it overlooks many aspects of user satisfaction. Queue length distribution, packet loss probability, and delay all influence the perceived quality of a communication link. These performance metrics are especially important in real-time applications where user satisfaction implies stringent QoS constraints.

Manuscript received June 27, 2006; revised February 1, 2007. The material in this paper was presented in part at the 44th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, September 2006.

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA.

Communicated by E. Modiano, Associate Editor for Communication Networks.

Digital Object Identifier 10.1109/TIT.2006.894682

Quality of service has been studied extensively in the context of wired networks [3]–[5]. Research on asynchronous transfer mode (ATM) networks and their statistical performance requirements led to the unifying concept of effective bandwidth [6], which has been applied to admission control and pricing. Effective bandwidth and the related concept of effective capacity provide a framework to identify the statistical characteristics of a stochastic process over various time scales. The literature on effective bandwidth is rich. Comprehensive discussions on the subject and its applications are provided by Kelly [7] and Chang [4].

User satisfaction and QoS also play an important role in the design of wireless systems, especially for mobile terminals that support real-time applications. Unlike its wired counterpart, a wireless connection is subject to attenuation and fading. The time-varying nature of a wireless link will affect the queue length distribution at a terminal and, consequently, it will have a significant impact on performance. The allocation of system resources for real-time applications is therefore critically important and demands a careful analysis. Performance and user satisfaction may not be captured adequately by the sole attribute of Shannon capacity. The goal of this paper is to relate power and spectral bandwidth to QoS using an evaluation framework akin to effective bandwidth. Two elements are needed to achieve this goal: a wireless channel model and a meaningful performance metric. Wireless channels have been studied in many contexts. In [8], Wang and Moayeri propose a finite-state Markov process to model a wireless fading environment. Markov models have successively been applied to Rayleigh and Nakagami fading channels [9]–[11]. Krunz and Kim [12], [13] employ independent two-state Markov processes to model the arrival traffic and offered service of a wireless connection. Based on this framework, they derive delay-bound violation probabilities for point-to-point wireless transmissions.

In [14]–[16], Wu and Negi extend the concept of effective bandwidth to effective capacity. Broadly speaking, the effective capacity characterizes the maximum arrival rate that a wireless system can sustain subject to a given QoS requirement. This concept can be viewed as the dual of the effective bandwidth. Much like its precursor, effective capacity is a useful tool to identify system limitations as a function of statistical delay-bound violation probabilities.

To relate radio resources to system performance and statistical QoS requirements, we link the behavior of the system to its physical-layer infrastructure. A mobile terminal and its asso-

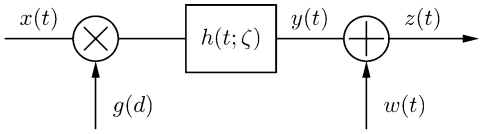

Fig. 1. Block diagram of a wireless communication channel where the transmitted signal is subject to attenuation, fading, and noise corruption.

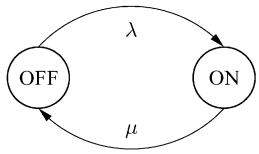

Fig. 2. Continuous-time Gilbert–Elliott Markov representation of a wireless communication link.

ciated wireless connection can be modeled as a single-server queue, provided that the receiver has the ability to acknowledge reception of the data. For example, a simple physical-layer automatic repeat request (ARQ) mechanism may be incorporated in the communication protocol to ensure that erroneous data is retransmitted. We assume that such a mechanism is in place throughout. Drawing intuition from information theory and error-control coding, the service offered to a mobile terminal can be modeled as a Markov-modulated fluid process. Previous results on Markov-modulated fluid processes and large deviations can therefore be leveraged to characterize the interplay between system resources at the physical layer and the statistical behavior of queues. In particular, we develop a framework in which spectral bandwidth, power, and code rate are tied to system performance.

The remainder of this paper is as follows. In Section II, we describe the generic wireless connection that is used as an abstraction for the physical layer. Based on a Markov assumption, we then construct a mathematical representation for the overall channel behavior. Section III contains a derivation of the equilibrium distribution for a system with a constant arrival rate and a Gilbert–Elliott wireless channel. Specifically, buffer overflow probabilities, the corresponding large deviations, and the effective capacity function are given explicit expressions in terms of physical system parameters. This analysis is subsequently extended to a variable data source. The performance analysis of the Gilbert–Elliott queueing system is presented in Section IV. We compare and contrast the statistical characteristics of the Gilbert–Elliott model with the characteristics of a continuous-state Markov channel using numerical simulations in Section V. Finally, we give our conclusions in the last section.

II. Wireless Channel

The complex baseband representation of the wireless channel under consideration is shown in Fig. 1. The term $g(d)$ accounts for the mean path attenuation, and $h(t;\zeta)$ represents the small-scale variations due to the motion of the terminals and changes in the environment [17]. The additive noise term $w(t)$ is modeled as a proper complex white Gaussian process. Note that $h(t;\zeta)$ is normalized so that the expected power gain introduced by $h(t;\zeta)$ is equal to one. The bandwidth of the transmitted signal $x(t)$ is assumed to be much smaller than the reciprocal of the delay spread. The channel is therefore purely time selective, with no frequency distortion [18]. In this case, the standard channel model of Fig. 1 can be written as

$$z(t) = g(d)h(t)x(t) + w(t)$$

where $h(t;\zeta) = h(t)\delta(\zeta)$. Furthermore, we assume that the channel is subject to purely diffuse scattering, i.e., no specular component is present. For a rich scattering environment, the multipath component $h(t)$ is well modeled as a zero-mean, proper complex Gaussian process. In particular, the envelope process $|h(t)|$ and the phase process are independent, with $|h(t)|$ having a Rayleigh probability distribution function and the phase being uniform over $[0, 2\pi)$.

While it is straightforward to describe the first-order statistics of $h(t)$, a complete characterization of this random process requires that joint distributions be specified as well. Under a Gaussian process model, it suffices to describe the correlation between any two sample points of the process. For Rayleigh flat fading, the autocorrelation function of the envelope process can be modeled using the zeroth-order Bessel function of the first kind [19]. This function is reasonable over short time horizons corresponding to terminal movements of the order of a few wavelengths. It is derived under the assumption that a mobile terminal is moving in an isotropic environment at a constant velocity. Alternatively, an autocorrelation function can be derived by assuming that the in-phase and quadrature components of $h(t)$ are independent stationary Ornstein–Uhlenbeck processes [20]. The latter model states that the correlation between two samples decays exponentially over time.

These two autocorrelation structures are useful in various contexts. However, for the sake of mathematical tractability, we consider a slightly simplified channel model. We retain the first-order statistics of the channel and assume that the marginal distribution of the envelope process is Rayleigh. Second, we assume that for a fixed threshold η, the probability of $|h(t)|$ being above or below this threshold is accurately modeled as a continuous-time Markov chain. We refer to the channel envelope exceeding η as the "ON" state; otherwise, the channel is in its "OFF" state. Such a channel structure is commonly referred to as the Gilbert–Elliott model. It is assumed to provide a sufficiently accurate representation for the statistical behavior of the quantized Rayleigh channel. This quantized channel model appears in Fig. 2. The transition rate from OFF to ON is denoted by λ; while the transition rate from ON to OFF, by μ. The generator matrix for this Markov chain is given by

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

It is easy to verify that the invariant probability for the ON state is $\lambda/(\lambda+\mu)$, while the invariant probability of being OFF is $\mu/(\lambda+\mu)$. For consistency, the stationary distribution of the Markov chain should agree with the marginal distribution of the underlying channel

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

where $f(\xi) = 2\xi e^{-\xi^2}$ with $\xi \ge 0$ is the marginal distribution of the normalized envelope process. To solve for λ and μ, two equations are needed. The first one is given by condition (1). The second condition can be derived from the Markov structure of the wireless link.

Let $P_t(t) = e^{Q_s t}$ be the probability transition matrix of the Gilbert–Elliott channel. More specifically, entry $p_{i,j}(t)$ of the matrix $P_t(t)$ represents the probability of being in state j after t seconds, when starting in state i. For a time interval t, this probability transition matrix is given by

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

We note that the channel memory of this two-state Markov process decays at a rate $\lambda + \mu$. Thus, if the memory of the underlying quantized Rayleigh channel has an exponential decay rate κ, we must have $\lambda + \mu = \kappa$. This relationship provides the second equation necessary to determine λ and μ. Solving for these parameters explicitly in terms of the channel parameters, we get

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

The quantized channel and its associated Markov structure will prove instrumental in computing the probability of buffer overflow and the effective capacity of the associated wireless connection. The more elaborate channel description whose in-phase and quadrature components are independent stationary Orstein–Uhlenbeck processes will be revisited in Section V.

A. Coding and Information Theory

A celebrated result from information theory is the Shannon capacity of the Gaussian channel

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

The variable P represents the power of the signal, $N_0/2$ denotes the power spectral density of the noise process, and W is the channel bandwidth. In theory, error-free communication can be achieved on this channel for any rate below the capacity using asymptotically long codewords [21]. Today, there exists a collection of practical codes that operate close to capacity, with minimal error rates and small delays. The capacity expression of (2) can therefore be employed as an optimistic approximation of code performance. If a code is designed to operate at a rate R, the sent information is decoded reliably whenever $R < C$; it is lost otherwise.

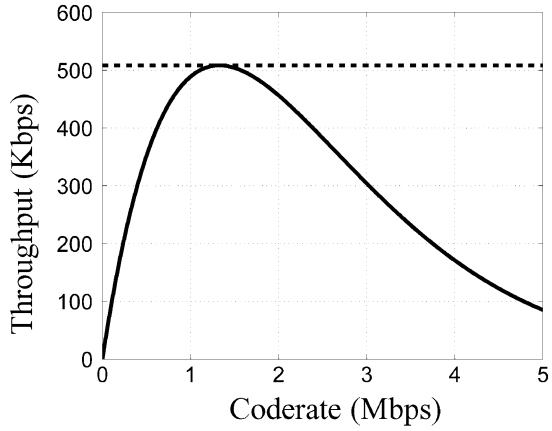

Fig. 3. Mean throughput as a function of code rate R for a Gilbert–Elliott channel model.

A similar performance description can be employed for time-varying channels such as the one introduced at the beginning of Section II. Suppose that a wireless channel varies slowly, data is assumed to reach its destination provided that $R < C(t)$, where

$$C(t) = W \log_2\left(1 + \frac{P|h(t)|^2}{N_0 W}\right) \quad \text{bits per second} \tag{3}$$

is the instantaneous capacity. On the other hand, if $R \ge C(t)$ then information is lost. This simplified characterization is valid provided that the decoding delay is small compared to the coherence time of the wireless channel. It is used in this work for mathematical convenience and because it yields useful guidelines on how to select code rates for specific systems and QoS requirements. This model can be altered to accommodate real codes and probabilities of link failures.

The state of the Gilbert–Elliott channel is related to the instantaneous capacity and the code rate as follows. Let code rate R be given. The Gilbert–Elliott channel is ON if $R < C(t)$ or equivalently

$$|h(t)| > \eta = \sqrt{\frac{N_0 W}{P}\left(2^{\frac{R}{W}} - 1\right)}. \tag{4}$$

It is OFF otherwise. We can rewrite the generator matrix for this Gilbert–Elliott channel as

$$Q_s = \begin{bmatrix}-\kappa e^{-\eta^2} & \kappa e^{-\eta^2} \\ \kappa - \kappa e^{-\eta^2} & -\kappa + \kappa e^{-\eta^2}\end{bmatrix} \tag{5}$$

where κ is the exponential decay parameter of the Markov chain and η is the threshold defined in (4). The corresponding service rate is zero when the channel is OFF and R when the channel is in its ON state.

Under these assumptions, the maximum throughput of this wireless channel is immediately seen to equal

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

This throughput can be optimized by selecting a proper code rate R. A higher rate allows more information to be transmitted when the channel is ON. However, it also implies that the channel is ON less often (larger η). Conversely, a lower rate increases the probability of the channel being ON but reduces the rate at which data is transfered. Fig. 3 plots the throughput as a function of code rate R. The parameter values for the wireless channel

TABLE I
SYSTEM PARAMETERS

$N_0 = 10^{-7}$ W/Hz	Noise power spectral density
$W = 11$ MHz	Bandwidth
$P = 100$ mW	Received Power

used in this example appear in Table I. The maximum average throughput is 508 kb/s, and it is achieved with a code rate $R = 1.33$ Mb/s.

III. QUEUEING PERFORMANCE OF MARKOV-MODULATED PROCESSES

In a wireless system, much like in a broadband network, the usage of system resources may not be well assessed by the single value of throughput or Shannon capacity. Performance measures such as queue length, packet loss probability, and delay play an instrumental role in user satisfaction. Requirements on these attributes may force a wireless system to operate much below its theoretical Shannon limit. Furthermore, the unreliable link quality intrinsic to wireless communication along with stringent delay and loss constraints may significantly alter the optimal allocation of system resources. This is exemplified below.

A. Markov Fluid Model of a Queue

Consider a simple fluid queueing system with a single queue and one server. Let $a(t)$ denote the instantaneous arrival rate, and let $s(t)$ be the instantaneous service rate. The cumulative arrival function over interval $[0, t]$ is given by

$$A[0,t] = \int_{[0,t]} a(\tau) d\tau.$$

Similarly, the amount of service offered in the interval $[0, t]$ is equal to

$$S[0,t] = \int_{[0,t]} s(\tau) d\tau.$$

Under a work-conserving policy and provided that the queue is initially empty, the state of the queue is governed by the following equation [6]:

$$L_t = (A[0,t] - S[0,t]) - \inf_{0<\tau<t} \{A[0,\tau] - S[0,\tau]\}.$$

This generic model provides an appropriate framework for evaluating the performance of a queueing system subject to QoS constraints.

A natural choice to model the communication system introduced in the preceding section is a Markov-modulated fluid process. Consider a queue subject to a Markov-modulated rate process. Let Q be the generator matrix of the underlying finite-state Markov process, and assume that Q is irreducible with state space $\{1, \ldots, M\}$. In particular, the off-diagonal entry $q_{n,m}$ represents the transition rate of going from state n to m; and the corresponding diagonal entry is $q_{n,n} = -\sum_{m \neq n} q_{n,m}$, making the total row sum zero. The state m is associated with a rate d_m, which represents the difference between the instantaneous arrival rate and the instantaneous service rate. Hence, the net rate of change in the buffer while in state m is d_m when the buffer is not empty, and it is equal to $\max\{0, d_m\}$ when the buffer is empty. In other words, when the buffer is empty and the Markov process is in state m with $d_m \leq 0$ then the buffer simply remains empty. We denote the diagonal matrix $\mathrm{diag}(d_1, \ldots, d_M)$ by D.

If we use L_t to denote the level of fluid in the buffer at time t and we let u_t be the state of the underlying Markov chain at time t, then (L_t, u_t) forms a continuous-state Markov process. Define the event probability

$$F(x, m, t) = \Pr\{u_t = m, L_t \leq x\}.$$

Using the Chapman–Kolmogorov equation, we find that the function $F(x, m, t)$ satisfies

$$\frac{\partial F}{\partial t} = FQ - \frac{\partial F}{\partial x} D \tag{6}$$

where $F = (F(x,1,t), \ldots, F(x,M,t))$. The equilibrium distribution $F(x, m)$ of the continuous-state Markov process (L_t, u_t) is subject to $\partial F / \partial t = 0$, which in turn yields

$$FQ = \frac{\partial F}{\partial x} D. \tag{7}$$

We denote the invariant probability distribution of the underlying Markov chain by w, with $wQ = 0$. Then

$$\lim_{x \to \infty} F(x, m) = w_m.$$

Since the equilibrium distribution is a bounded solution to (7), it has spectral representation

$$F(x, \cdot) = w - \sum_{i=1}^{k} a_i \phi_i e^{x z_i} \tag{8}$$

where $\{(\phi_i, z_i)\}$ are the stable eigenvector–eigenvalue pairs of the eigenvalue problem

$$\phi Q = z \phi D. \tag{9}$$

If Q is reversible [22], then all such eigenvalues are real numbers [23]. Moreover, there are $k = |\{m : d_m > 0\}|$ strictly negative eigenvalues (counting multiplicity). These values are the ones included in (8). The coefficients $\{a_i\}$ are found using the boundary conditions

$$F(0, m) = 0, \quad \forall \ \{m : d_m > 0\}.$$

The unique solution to this boundary value problem is the equilibrium distribution [24]. The reader is referred to Mitra [25] and Meyn [26] for additional information about fluid models.

B. Equilibrium Distribution of Gilbert-Elliott Systems

Consider a communication system where data arrives in a buffer at a constant rate $a(t) = a$. Suppose that this buffer is serviced through a wireless connection at a rate $s(t)$, where $s(t)$ is the Markov-modulated process described in Section II. That is, $s(t)$ is equal to R when the channel is ON, and zero otherwise. We assume that the generator matrix of the underlying finite-state Markov chain is the matrix Q_s obtained in (5), i.e.,

$$Q = Q_s = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}.$$

Note that Q is reversible since Q and w are in detailed balance, $w_1 q_{1,2} = w_2 q_{2,1}$. The net arrival rate in the buffer when the buffer is not empty is

$$D = \begin{bmatrix} a & 0 \\ 0 & a-R \end{bmatrix}.$$

The eigenvalue problem $\phi Q = z\phi D$ has two solution pairs

$$(w,0) = \left(\left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu} \right), 0 \right)$$
$$(\phi, z) = \left((R-a, a), \frac{a\lambda + a\mu - R\lambda}{a(R-a)} \right).$$

The queue will be stable provided that

$$a < \frac{\lambda}{\lambda+\mu} R. \tag{10}$$

Under this condition, L_t converges in distribution to a finite random variable L. Using the boundary condition $F(0,1) = 0$, we obtain the equilibrium solution

$$\begin{aligned} F(x,\cdot) &= \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu} \right) \\ &\quad - \left(\frac{\mu}{\lambda+\mu}, \frac{a}{R-a}\frac{\mu}{\lambda+\mu} \right) \exp\left(\frac{a\lambda+a\mu-R\lambda}{a(R-a)} x \right) \\ &= \left(1 - e^{-\eta^2}, e^{-\eta^2} \right) \\ &\quad - \left(1 - e^{-\eta^2} \right) \left(1, \frac{a}{R-a} \right) \exp\left(\frac{\kappa a - R\kappa e^{-\eta^2}}{a(R-a)} x \right). \end{aligned} \tag{11}$$

Based on this equilibrium distribution, we can compute a number of performance metrics including the probability of buffer overflow, its exponential rate of convergence to zero, and the effective capacity of the system.

The probability of buffer overflow is an important performance metric. For the Gilbert–Elliott system at hand, the probability of the buffer exceeding a threshold x is given by

$$\begin{aligned} \Pr\{L > x\} &= 1 - \langle F(x,\cdot), (1,1) \rangle = 1 - \langle F(x,\cdot), \mathbf{1} \rangle \\ &= \frac{R}{R-a}\frac{\mu}{\lambda+\mu} \exp\left(\frac{a\lambda+a\mu-R\lambda}{a(R-a)} x \right) \\ &= \frac{R}{R-a}(1-e^{-\eta^2}) \exp\left(\frac{\kappa a - R\kappa e^{-\eta^2}}{a(R-a)} x \right). \end{aligned}$$

As seen in (11), the eigenvalue problem (9) applied to the present two-state system contains only one negative solution. The large deviation principle governing the probability of buffer overflow is therefore immediately seen to equal

$$\begin{aligned} -\lim_{x\to\infty} \frac{\log \Pr\{L > x\}}{x} &= -\frac{a\lambda+a\mu-R\lambda}{a(R-a)} \\ &= -\frac{\kappa a - R\kappa e^{-\eta^2}}{a(R-a)}. \end{aligned}$$

The large deviation principle governing the distribution of a queue is sometimes preferable as a design criterion because it admits a simpler form.

The effective capacity is the dual concept of effective bandwidth [14]. Given specific system parameters and an exponential decay rate θ, the effective capacity is the maximum arrival rate for which the QoS requirement θ is fulfilled. Mathematically, this can be expressed as

$$\alpha(\theta) = \sup\left\{ a \geq 0 : -\lim_{x\to\infty} \frac{\log \Pr\{L > x\}}{x} \geq \theta \right\}. \tag{12}$$

Under sufficient conditions, the effective capacity function is given by

$$\alpha(\theta) = -\frac{\Lambda(-\theta)}{\theta} \tag{13}$$

where $\Lambda(\cdot)$ is defined by

$$\Lambda(-\theta) = \lim_{t\to\infty} \frac{1}{t} \log \mathrm{E}\left[e^{-\theta S[0,t]} \right].$$

For the simple problem considered in this section, (12) leads to

$$\begin{aligned} \alpha(\theta) &= \sup\left\{ a \geq 0 : \frac{a\lambda+a\mu-R\lambda}{a(R-a)} \leq -\theta \right\} \\ &= \sup\left\{ a \geq 0 : \theta a^2 - (\theta R + \lambda + \mu) a + R\lambda \geq 0 \right\}. \end{aligned} \tag{14}$$

Taking into account condition (10), this yields an explicit formula for the effective capacity of the Gilbert–Elliott channel

$$\begin{aligned} \alpha(\theta) &= \frac{\theta R + \lambda + \mu - \sqrt{(\theta R + \lambda + \mu)^2 - 4\theta R\lambda}}{2\theta} \\ &= \frac{\theta R + \kappa - \sqrt{(\theta R + \kappa)^2 - 4\theta R\kappa e^{-\eta^2}}}{2\theta}. \end{aligned} \tag{15}$$

Appendix I shows that (15) can also be obtained directly from (13).

C. On–Off Information Sources

Some traffic sources are better modeled as on–off sources. Voice, for instance, is a good example of an information process that can be accurately modeled as an on–off source. When two people are carrying a conversation, they are unlikely to speak simultaneously. On average, a person involved in a discussion speaks at most half of the time. Other data sources such as instant messaging applications and wireless sensors [27] can also be modeled as on–off sources. As such, we extend the analysis of the previous section to the case where the data source features an on–off behavior.

Suppose that data arrive in the buffer at a rate $a(t)$, where $a(t)$ is a two-state Markov-modulated source. We assume that the arrival rate is equal to $a > 0$ when the source is ON; it is equal to zero otherwise. The generator matrix of the underlying Markov chain for this arrival process can be written as

$$Q_a = \begin{bmatrix} -\lambda_a & \lambda_a \\ \mu_a & -\mu_a \end{bmatrix}.$$

Again, we assume that the service offered through the wireless channel is a Markov-modulated process with generator matrix Q_s, as defined in (5). The aggregate system is therefore a stochastic fluid process modulated by a four-state Markov chain.

The evolution of the buffer content is governed by (6), where the generator matrix Q is equal to

$$\begin{aligned} Q &= \begin{bmatrix} -\lambda_a-\lambda & \lambda_a & \lambda & 0 \\ \mu_a & -\mu_a-\lambda & 0 & \lambda \\ \mu & 0 & -\lambda_a-\mu & \lambda_a \\ 0 & \mu & \mu_a & -\mu_a-\mu \end{bmatrix} \\ &= Q_s \otimes I + I \otimes Q_a. \end{aligned}$$

Throughout, we use $A \otimes B$ to denote the Kronecker product of matrices A and B. Again, it is straightforward to verify that the generator matrix Q is reversible. The net arrival rate in the buffer is represented by

$$\begin{aligned} D &= \operatorname{diag}(0, a, -R, a-R) \\ &= -RE \otimes I + aI \otimes E, \end{aligned}$$

where the matrix E is defined by

$$E = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

We assume that the system is stable; that is, the following condition is satisfied:

$$\frac{\lambda_a}{\lambda_a+\mu_a} a < \frac{\lambda}{\lambda+\mu} R.$$

The equilibrium distribution of this system is governed by (7), and its spectral representation follows the general form of (8).

The generalized eigenvalue problem $\phi Q = z\phi D$ admits three eigenvector–eigenvalue pairs in the present case because $\det(Q - zD)$ is a third-order polynomial ($a \neq R$). We note that the underlying Markov chain is reversible since it satisfies the detailed balance condition, $w_i Q_{ij} = w_j Q_{ji}$ for $1 \leq i, j \leq 4$. This property ensures that all the eigenvalues of $\phi Q = z\phi D$ are real numbers. The first eigenvector is the invariant distribution given by

$$\begin{aligned} w &= \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu} \right) \otimes \left(\frac{\mu_a}{\lambda_a+\mu_a}, \frac{\lambda_a}{\lambda_a+\mu_a} \right) \\ &= \frac{(\mu\mu_a, \mu\lambda_a, \lambda\mu_a, \lambda\lambda_a)}{(\lambda+\mu)(\lambda_a+\mu_a)}. \end{aligned} \tag{16}$$

The associated eigenvalue is, of course, zero. To find the remaining two eigenvector–eigenvalue pairs, we make an educated guess based on a standard decomposition technique popularized by Mitra [25]. For any vector of the form $\phi = \phi_s \otimes \phi_a$, we can rewrite (9) as

$$(\phi_s Q_s + Rz\phi_s E) \otimes \phi_a = \phi_s \otimes (az\phi_a E - \phi_a Q_a). \tag{17}$$

Consider the two vectors defined by

$$\phi_v = (R - v, v) \otimes (a - v, v) \tag{18}$$

where v is either solution of the quadratic form

$$\begin{aligned} v^2(\lambda+\mu+\lambda_a+\mu_a) - vR(\lambda+\lambda_a+\mu_a) \\ -va(\lambda+\mu+\lambda_a) + aR(\lambda+\lambda_a) = 0. \end{aligned} \tag{19}$$

It is straightforward to show that the vectors jointly defined by (18) and (19) are eigenvectors of (17), with corresponding eigenvalues

$$z = \frac{v\lambda + v\mu - R\lambda}{v(R-v)} = \frac{a\lambda_a - v\lambda_a - v\mu_a}{v(a-v)}. \tag{20}$$

Incidentally, (19) is obtained by equating the preceding two expressions for z. Since (19) has two distinct real roots, the two associated eigenvectors along with the invariant distribution described in (16) completely characterize the eigenvalue problem $\phi Q = z\phi D$. We can see from the spectral representation of the equilibrium distribution (8) that the large deviation principle governing the queue occupancy is dominated by the largest negative eigenvalue of (17).

Consider an exponential decay rate requirement of $\theta > 0$ on the probability of buffer overflow

$$-\lim_{x\to\infty} \frac{\log \Pr\{L > x\}}{x} \geq \theta. \tag{21}$$

This requirement will be satisfied provided that the largest negative eigenvalue of (20) is less than $-\theta$. In particular, we want the following equations to hold:

$$\begin{aligned} v^2\theta - v(\theta R + \lambda + \mu) + R\lambda \geq 0 \\ v^2\theta - v(\theta a - \lambda_a - \mu_a) - a\lambda_a \geq 0. \end{aligned}$$

These conditions will be fulfilled if and only if the value of v corresponding to the largest negative eigenvalue of (20) is less than $\alpha(\theta)$ but greater than $\beta(\theta)$, where $\alpha(\theta)$ is the effective capacity introduced earlier

$$\begin{aligned} \alpha(\theta) &= \sup\left\{ \nu \geq 0 : \frac{\nu\lambda + \nu\mu - R\lambda}{\nu(R-\nu)} \leq -\theta \right\} \\ &= \sup\left\{ \nu \geq 0 : \theta\nu^2 - (\theta R + \lambda + \mu)\nu + R\lambda \geq 0 \right\} \end{aligned}$$

and $\beta(\theta)$ is the effective bandwidth of a two-state Markov-modulated fluid source [28], [3], [12]

$$\begin{aligned} \beta(\theta) &= \inf\left\{ \nu \geq 0 : \frac{a\lambda_a - \nu\lambda_a - \nu\mu_a}{\nu(a-\nu)} \leq -\theta \right\} \\ &= \inf\left\{ \nu \geq 0 : \theta\nu^2 - (\theta a - \lambda_a - \mu_a)\nu - a\lambda_a \geq 0 \right\}. \end{aligned}$$

Clearly, the QoS requirement of (21) can only be met if $\alpha(\theta) \geq \beta(\theta)$. A standard buffer decoupling argument shows that this inequality is, in fact, a necessary and sufficient condition for (21) to hold. That is, the exponential decay rate requirement $\theta > 0$ will be satisfied if and only if $\alpha(\theta) \geq \beta(\theta)$. The decoupling argument is contained in Appendix II.

This observation greatly facilitates the performance analysis contained in the next section. In particular, for a QoS requirement such as (21), an on–off source shares the same service needs as a constant source with rate $\beta(\theta)$. Thus, for a given $\theta > 0$, the allocation of system resources can be studied in terms of fixed arrival rates, whether the source rate is a constant or a Markov-modulated fluid model. This is illustrated in the next section.

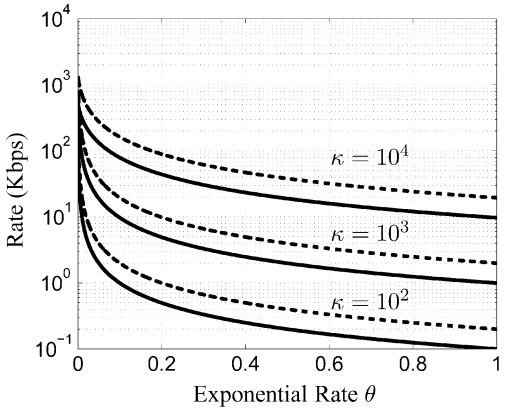

Fig. 4. Optimal code rate (dashed) and effective capacity (solid) as a function of exponential decay rate θ for various Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$.

IV. Performance Analysis of Gilbert–Elliott Systems

We proceed to analyze the performance of the Gilbert–Elliott system as a function of physical resources. Following the literature on effective bandwidth, we use the exponential decay rate as our primary performance measure.

A. Effective Capacity Analysis

The effective capacity quantifies the maximum supported arrival rate for a set of system parameters and a QoS constraint $\theta > 0$. It is an appropriate tool to quantify the optimal operating point of a wireless system. This maximum rate can either be the true rate of a constant source or the effective bandwidth of a time-varying source. Fig. 4 shows the maximal supported arrival rate $\alpha(\theta)$ as a function of the QoS constraint θ for the system parameters of Table I and the Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$. This figure also includes the optimal code rate R as a function of θ. We emphasize that for queueing constraint $\theta > 0$, the optimal code rate R differs from the throughput maximizing rate introduced in Section II-A. Not surprisingly, more stringent QoS constraints result in lower effective capacities for fixed system parameters. This is intuitive since a lower arrival rate reduces the expected queue length. More importantly, we see that the optimal code rate R is also a function of the QoS requirements. Under strict QoS constraints, an error-control code with a lower rate performs better as it reduces the probability of the channel being OFF. This analysis provides a new and systematic way to select the code rate as a function of the channel profile and the QoS requirement of a specific system.

It is interesting to note that the maximum throughput and the corresponding code rate are independent of the Markov-decay parameter κ,

$$\begin{aligned}\lim_{\theta\to 0}\alpha(\theta) &= \lim_{\theta\to 0}\frac{\theta R+\lambda+\mu-\sqrt{(\theta R+\lambda+\mu)^2-4\theta R\lambda}}{2\theta}\\ &= \lim_{\theta\to 0}\left[\frac{R}{2}-\frac{(\theta R+\lambda+\mu)R-2R\lambda}{2\sqrt{(\theta R+\lambda+\mu)^2-4\theta R\lambda}}\right]\\ &= \frac{R\lambda}{\lambda+\mu} = Re^{-\eta^2}.\end{aligned}$$

However, the effective capacity for $\theta > 0$ depends heavily on the statistical profile of the channel. Correlation impairs effective capacity. The higher the correlation coefficient, the lower the effective capacity. In other words, a throughput analysis of this system is not sufficient to provide an accurate assessment of supported rates under strict QoS constraints. This also implies that the common assumption that channel realizations are independent and identically distributed through time may lead to over-optimistic performance predictions on effective capacity.

B. Resource Requirement Analysis

The effective capacity shown in Fig. 4 decays rapidly as a function of θ. It is therefore of interest to look at the reverse problem; for a given arrival rate a, we wish to characterize the amount of physical resources necessary to meet a prescribed QoS constraint $\theta > 0$. First, we note that a necessary condition for a solution to exist is the stability criterion $a < Re^{-\eta^2}$. However, this condition may not be sufficient. Looking at (14), we see that a solution exists if and only if we can find a power P and a bandwidth W such that

$$\theta a^2 - (\theta R + \kappa)a + R\kappa e^{-\eta^2} = 0.$$

This equation can be rearranged as

$$\begin{aligned}\eta^2 &= \frac{N_0 W}{P}\left(2^{\frac{R}{W}}-1\right)\\ &= -\log\left(\frac{\theta R a+\kappa a-\theta a^2}{R\kappa}\right).\end{aligned}$$

Since $\eta^2 > 0$, the following inequality must apply:

$$0 < \theta R a + \kappa a - \theta a^2 < R\kappa.$$

A necessary and sufficient condition for a solution to exist is $\theta < \kappa/a$. We emphasize that, even with an unlimited power and spectral bandwidth budget, only a finite arrival rate can be supported for a QoS constraint $\theta > 0$. Furthermore, this bound is independent of the actual code rate R used in the system. This fact is in sharp contrast with Shannon capacity, which goes to infinity as power and spectral bandwidth grow unbounded. This limitation is partly due to the fact that, in the system under study, the transmitter has no knowledge of the channel gain. Thus, it cannot transmit at the (error-free) instantaneous channel capacity. Without channel state information, the best decay rate θ is limited by the ratio of κ to the arrival rate a. In the limit where the power and spectral bandwidth become very large, the queueing behavior of the system is increasingly dominated by the holding time of its OFF state. The queue is drained almost instantaneously when the channel is ON, while it rises linearly when the channel is OFF. The probability of the queue exceeding a threshold is then dominated by the duration of an OFF period, which is exponentially distributed.

Fig. 5 shows the target power P as a function of the QoS constraint θ for an arrival rate $a = 14.4$ kb/s and the parameters of Table I. Note that power P can be obtained in closed form as

$$P = -\frac{N_0 W\left(2^{\frac{R}{W}}-1\right)}{(\log(\theta R a+\kappa a-\theta a^2)-\log(R\kappa))}.$$

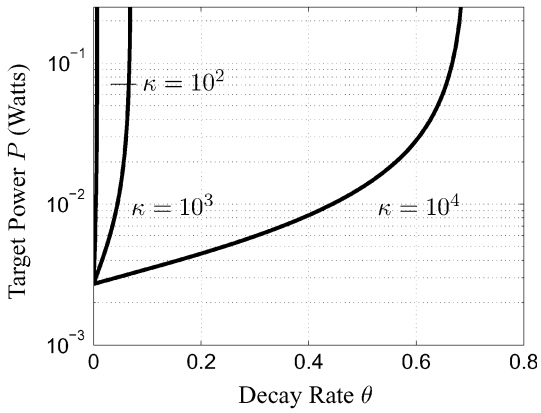

Fig. 5. Signal power as a function of exponential decay rate θ for various Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$.

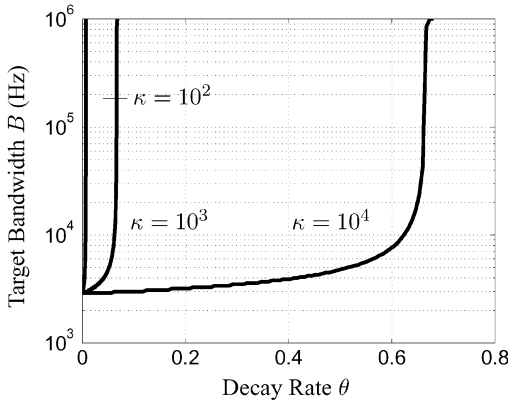

Fig. 6. Spectral bandwidth as a function of exponential decay rate θ for various Markov decay parameters $\kappa \in \{10^2, 10^3, 10^4\}$.

As expected, the required power goes to infinity for a finite θ. The set of supported exponential decay rates is intimately connected to the Markov decay factor κ. Not only does correlation decrease effective capacity, it also limits the rates and qualities of service that can be sustained on a given wireless channel. Similar findings can be obtained for the spectral bandwidth requirement as a function of arrival rate a and QoS constraint θ. Fig. 6 shows minimum spectral bandwidth as a function of decay rate θ for an arrival rate of $a = 14.4$ kb/s. Again, the amount of resource required goes to infinity for a finite θ.

The speech process of an interlocutor involved in an English language conversation can be modeled as an on–off information source [29]. Exponentially distributed talk spurts with a mean duration of $\mu_a^{-1} \approx 352$ ms are followed by silent periods with mean $\lambda_a^{-1} \approx 650$ ms. Using advanced signal processing techniques, active speech can be compressed to a rate of 14.4 kb/s [30]. The average throughput of an encoded speech process is therefore 5.06 kb/s. However, for a delay constraint of 20 ms or an approximately equivalent QoS constraint $\theta = 0.5$, the effective bandwidth of speech is essentially equal to 14.4 kb/s. The very strict delay constraint imposed on speech traffic forces the effective bandwidth to be nearly equal to its peak rate, which is much higher than the average throughput. As seen on Figs. 5 and 6, voice traffic cannot be successfully transmitted over highly correlated channels without sophisticated power control. This partly explains why power control is critical to cellular telephony [31]–[33].

V. Numerical Analysis

In Section II, the Gilbert–Elliott channel model is introduced as a first-order approximation to the autocorrelation of a Rayleigh-fading channel. This simplified channel model permits the derivation in closed form of many important quantities, including the probability of buffer overflow and the effective capacity. Recall that the Gilbert–Elliott model is based on two assumptions. First, the state of the Gilbert–Elliott channel identifies whether the instantaneous realization of the underlying Rayleigh channel lies above or below a prescribed threshold. Second, the stochastic process representing the time evolution of this quantized channel is accurately modeled as a two-state, continuous-time Markov chain.

While it is mathematically convenient to assume that the quantized channel possesses the Markov property, a more common approach is to assume that the channel itself is Markov (not the quantized version). Furthermore, we note that it is straightforward to construct a Rayleigh-fading channel that possesses the Markov property. In particular, consider the Ornstein–Uhlenbeck equation

$$dX_t = -\kappa X_t dt + \sigma dB_t$$

where κ, σ are real constants and B_t is a one-dimensional Brownian motion. The solution to this stochastic differential equation is called the Ornstein–Uhlenbeck process. This solution has the Markov property and it is given by [20], [34]

$$X_t = X_0 e^{-\kappa t} + \int_0^t e^{-\kappa(t-s)} \sigma dB_s.$$

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

$$\begin{aligned}
&\mathrm{E}\left[(X_t - \mathrm{E}[X_t])^2\right] \\
&\quad = \mathrm{E}\left[\left(X_0 e^{-\kappa t} + \int_0^t e^{-\kappa(t-s)} \sigma dB_s \; - \mathrm{E}[X_0] e^{-\kappa t}\right)^2\right] \\
&\quad = \mathrm{E}\left[(X_0 - \mathrm{E}[X_0])^2\right] e^{-2\kappa t} + \mathrm{E}\left[\left(\int_0^t e^{-\kappa(t-s)} \sigma dB_s\right)^2\right] \\
&\quad = \mathrm{E}\left[(X_0 - \mathrm{E}[X_0])^2\right] e^{-2\kappa t} + \mathrm{E}\left[\int_0^t e^{-2\kappa(t-s)} \sigma^2 ds\right] \\
&\quad = \mathrm{E}\left[(X_0 - \mathrm{E}[X_0])^2\right] e^{-2\kappa t} + \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa t}\right).
\end{aligned}$$

If $X_0 \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ and $\sigma^2 = \kappa$, then $X_t \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ for all $t \geq 0$. A Rayleigh-fading channel that possesses the Markov property can therefore be obtained by assigning independent stationary Ornstein–Uhlenbeck processes to the in-phase and quadrature component of the channel. The first-order statistic of the corresponding $h(t)$ is a zero-mean, proper complex Gaussian process 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 various results and techniques including the Chapman–Kolmogorov equation of Section III and the large-deviation principle for Markov

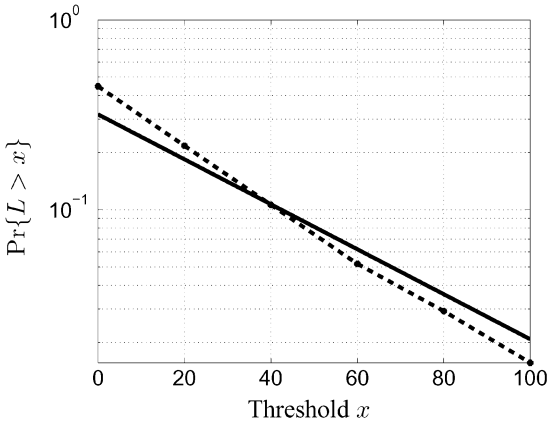

Fig. 7. Comparison of $\Pr\{L > x\}$ for the two-state Markov channel (solid) along with the empirically measure probabilities of buffer overflow for the Markov Rayleigh-fading model (dashed) for decay parameters $\kappa = 10^3$.

fluid processes. These limitations and the vast literature on Markov-modulated processes explain our early adoption of the Gilbert–Elliott model in Section II.

In this section, we use numerical simulations to assess the validity of the Gilbert–Elliott channel model in approximating the behavior of a Markov Rayleigh-fading channel. Since most of our results are based on the equilibrium distributions of queues, we compare the analytic probability of buffer overflow for the Gilbert–Elliott channel with the empirical distribution of the queue length for the Markov Rayleigh-fading channel. Recall that the probability of buffer overflow for the Gilbert–Elliott channel is given by

$$\Pr\{L > x\} = \frac{R}{R-a}\left(1 - e^{-\eta^2}\right)\exp\left(\frac{\kappa a - R\kappa e^{-\eta^2}}{a(R-a)}x\right).$$

Fig. 7 shows $\Pr\{L > x\}$ for the Gilbert–Elliott channel along with the empirically measured probabilities of buffer overflow for the Markov Rayleigh-fading model.

As seen on the graph, there is a noticeable difference between the two systems. Nevertheless, the exponential decay rate associated with the Gilbert–Elliott channel seems to provide an upper bound for the decay rate of the Orstein–Uhlenbeck system. This is encouraging as the Gilbert–Elliott model appears to provide a conservative measure of effective capacity.

VI. Conclusion

We considered the allocation of system resources in the context of wireless communications under QoS constraints. A Markov model was introduced to capture the unreliable nature of wireless systems. For a given error-correcting code, the behavior of the overall wireless connection is assumed equivalent to a continuous-time Markov chain. Code rate selection affects system throughput. A higher code rate allows more information to be transmitted when the channel is ON, but it also reduces the probability of this event occurring. Conversely, a lower code rate increases the probability of the channel being ON, yet it decreases the rate at which information flows when the wireless link is ON. The throughput of a system can be optimized by proper code selection.

While throughput represents a major component of user satisfaction; queue length distribution, packet loss probability, and delay are also important factors that influence the QoS perceived by the users. One of the main QoS measures present in the literature is the large deviation principle governing the probability of buffer overflow, i.e.,

$$-\lim_{x\to\infty}\frac{\log\Pr\{L > x\}}{x}.$$

This definition of QoS was used to conduct a performance analysis of the Gilbert–Elliott system as a function of physical resources. The effective capacity is found to decay sharply as a function of QoS constraint $\theta > 0$. Furthermore, the optimal code selection for a wireless system depends on its QoS requirement. A more stringent constraint on θ lowers the optimal code rate R.

Correlation is also found to have a major impact on performance. The effective capacity of a slowly varying channel can be very small. For communications under QoS constraints, the popular assumption of independent and identically distributed channel realizations results in an over-optimistic assessment of system performance. The impact of correlation on system performance is perhaps best exemplified by the fact that arrival rate a can only be supported if $\theta < \kappa/a$. That is, even with unlimited amount of physical resources, the maximum arrival rate supported under QoS constraint θ is bounded.

The numerical analysis section suggests that alternative Markov models for the underlying wireless channel should be explored. For instance, a finite-state Markov model can be used to represent the channel itself, rather than modeling the ability of the decoder to recover data reliably. The performance evaluation method presented in this work provides, nonetheless, an elegant framework to quantify the amount of physical resources necessary to support rate a under QoS constraint θ. Alternatively, this framework can be used in conjunction with effective capacity to characterize the maximal arrival rate a subject to specific resource and QoS constraints.

Appendix I
Effective Capacity of Service Process

In this appendix, we use the Kolmogorov backward equation to derive a formula for the effective capacity of a Markov-modulated service process. We parallel an argument by Kesidis *et al.* [28], albeit in the context of effective capacity.

Consider the stationary fluid process introduced in Section III. Recall that a process is said to be Markov fluid if its time derivative is a function of a continuous-time, finite-state Markov chain. Let $S[0,t]$ be the amount of service offered to a user during the interval $[0,t]$, and suppose that $S[0,t]$ is a Markov fluid process. Let u_t denote the state of the modulating Markov chain, taking value in $\{1,2,\ldots,M\}$. Using previously established notation, u_t has generator matrix Q_s and invariant distribution w. When the modulating chain u_t is in state m, we denote the offered service rate by s_m. We assume that

$0 \leq s_m \leq s_{m+1} < \infty$ for all $m \in \{1, 2, \ldots, M-1\}$. Given that the generator matrix Q_s is irreducible and reversible, we can write the effective capacity of this channel as

$$\alpha(\theta) = \lim_{t\to\infty} -\frac{1}{\theta t} \log \mathrm{E}\left[e^{-\theta S[0,t]}\right], \qquad 0 < \theta < \infty.$$

We proceed to evaluate $\alpha(\theta)$ explicitly. Define the function

$$\psi_j(\theta, t) = \mathrm{E}_j\left[e^{-\theta S[0,t]}\right] = \mathrm{E}\left[e^{-\theta S[0,t]}\middle| u_0 = j\right].$$

For positive $\epsilon \ll 1$, the transition matrix $P_t(\epsilon)$ can be written as

$$P_t(\epsilon) = e^{Q_s \epsilon} = I + \epsilon Q_s + o(\epsilon).$$

Using this notation, the standard backward equation becomes

$$\begin{aligned}\psi_j(\theta, t) &= \mathrm{E}\left[\mathrm{E}[e^{-\theta S[0,t]}\middle| u_\epsilon]\middle| u_0 = j\right] \\ &= \sum_{i=1}^{M} e^{-\theta\epsilon s_j}\psi_i(\theta, t-\epsilon)e^{\epsilon Q_s}(j,i) + o(\epsilon) \\ &= \sum_{i=1}^{M}(1-\theta\epsilon s_j)\psi_i(\theta, t-\epsilon)(I+\epsilon Q_s)(j,i) + o(\epsilon).\end{aligned}$$

Rearrange the preceding equation, we get

$$\begin{aligned}&\frac{\psi_j(\theta,t) - \psi_j(\theta, t-\epsilon)}{\epsilon} \\ &\quad = \psi_j(\theta, t-\epsilon)(Q_s(j,j) - \theta s_j) \\ &\qquad + \sum_{i\neq j}(1-\theta\epsilon s_j)\psi_i(\theta, t-\epsilon)Q_s(j,i) + \frac{o(\epsilon)}{\epsilon}. \quad (22)\end{aligned}$$

As $\epsilon \to 0$, (22) becomes

$$\frac{\partial \psi_j(\theta,t)}{\partial t} = \psi_j(\theta,t)(Q_s(j,j) - \theta s_j) + \sum_{i\neq j}\psi_i(\theta,t)Q_s(j,i).$$

Defining the diagonal matrix $S = \mathrm{diag}(s_1, \ldots, s_M)$ and the vector

$$\Psi(\theta, t) = (\psi_1(\theta,t), \ldots, \psi_M(\theta,t)),$$

we can write the above equations in matrix form as

$$\frac{\partial \Psi(\theta,t)}{\partial t} = (Q_s - \theta S)\Psi(\theta,t). \quad (23)$$

This differential equation is subject to the boundary conditions $\Psi(\theta, 0) = \mathbf{1}$. It follows that

$$\Psi(\theta, t) = \exp((Q_s - \theta S)t)\mathbf{1}.$$

We can rewrite the effective capacity as

$$\begin{aligned}\alpha(\theta) &= \lim_{t\to\infty} -\frac{1}{\theta t}\log \mathrm{E}\left[e^{-\theta S[0,t]}\right] \\ &= \lim_{t\to\infty} -\frac{1}{\theta t}\log(w\exp((Q_s - \theta S)t)\mathbf{1}).\end{aligned}$$

Using the Perron–Frobenius theorem [35], we obtain

$$\alpha(\theta) = -\frac{1}{\theta}\max_i \gamma_i$$

where $\{\gamma_i\}$ are the eigenvalues of the matrix $Q_s - \theta S$.

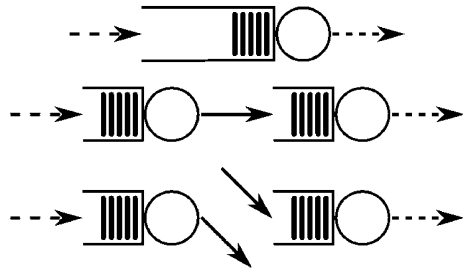

Fig. 8. Illustration of the three systems introduced in the buffer decoupling argument.

A. Two-State Markov Fluid Example

Consider the Gilbert–Elliott channel model introduced in Section II. For this channel, the generator matrix Q_s is given by

$$Q_s = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}.$$

The characteristic equation of the matrix $(Q_s - \theta S)$ is equal to

$$\begin{aligned}\det[\gamma I - (Q_s - \theta S)] &= \det\begin{bmatrix} \gamma+\lambda & -\lambda \\ -\mu & \gamma+\mu+\theta R \end{bmatrix} \\ &= \gamma^2 + (\theta R + \lambda + \mu)\gamma + \theta R\lambda.\end{aligned}$$

The maximum eigenvalue of the matrix $(Q_s - \theta S)$ is immediately found to be

$$\max_i \gamma_i = \frac{-(\theta R + \lambda + \mu) + \sqrt{(\theta R + \lambda + \mu)^2 - 4\theta R\lambda}}{2}.$$

Thus, the effective capacity of the Gilbert–Elliott channel model is

$$\begin{aligned}\alpha(\theta) &= \lim_{t\to\infty} -\frac{1}{\theta t}\log\left(w\exp\left(\begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu-\theta R\end{bmatrix}t\right)\mathbf{1}\right) \\ &= \frac{(\theta R + \lambda + \mu) - \sqrt{(\theta R + \lambda + \mu)^2 - 4\theta R\lambda}}{2\theta}. \quad (24)\end{aligned}$$

Using the relations $\lambda = \kappa e^{-\eta^2}$ and $\mu = \kappa - \kappa e^{-\eta^2}$, the effective capacity function can be expressed as

$$\alpha(\theta) = \frac{\theta R + \kappa - \sqrt{(\theta R + \kappa)^2 - 4\theta R\kappa e^{-\eta^2}}}{2\theta}$$

which coincides with (15).

APPENDIX II
BUFFER DECOUPLING ARGUMENT

Consider a queueing system with a Markov-modulated arrival process $a(t)$ and a Markov-modulated service process $s(t)$. Compare this single queue with a system that contains two queues. The arrival rate in the first queue is again $a(t)$ and the service offered to the second queue is $s(t)$. Moreover, the first queue is serviced at a constant rate v whenever it is nonempty, and the departing packets from the first queue are immediately placed in the second queue. This is illustrated in Fig. 8. Because of the additional constraint present in the second scenario, the queue length in the first system is always less than or equal to the sum of the queues in the latter system.

Now compare the second system with a network composed of two independent queues. The arrival process in the first queue is $a(t)$, and this queue is served at a constant rate v when it is nonempty. Packets arrive in the second queue at a constant rate v, and they are served at a rate $s(t)$. Note that the length of the first queue in the third system is always equal to the length of the first queue in the second system. Furthermore, the length of the second queue in the second system is always less than or equal to the length of the second queue in the third system. It follows that the large deviation principle governing the queue length in the first system is always less than or equal to the large deviation principle governing the sum of the queues in the third system. As such the QoS constraint θ as defined in (21) will be fulfilled whenever there exists a positive v such that $\alpha(\theta) \geq v \geq \beta(\theta)$. In particular, $\alpha(\theta) \geq \beta(\theta)$ is a sufficient condition for the QoS requirement of (21) to be satisfied. This is the desired property.

REFERENCES

[1] S. Verdú and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[2] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.

[3] 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.

[4] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[5] 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.

[6] C.-S. Chang, *Performance Guarantees in Communication Networks*, ser. Telecommunication Networks and Computer Systems. New York: Springer, 1995.

[7] F. P. Kelly, S. Zachary, and I. Ziedins, *Stochastic Networks: Theory and Applications*, ser. Royal Statistical Society Lecture Note Series. Oxford, U.K.: Oxford Univ. Press, 1996.

[8] 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.

[9] 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.

[10] 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.

[11] C.-D. Iskander and P. T. Mathiopoulos, "Fast simulation of diversity Nakagami fading channels using finite-state Markov models," *IEEE Trans. Broadcast.*, vol. 49, no. 3, pp. 269–277, Sep. 2003.

[12] M. M. Krunz and J. G. Kim, "Fluid analysis of delay and packet discard performance for QoS support in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 2, pp. 384–395, Feb. 2001.

[13] J. G. Kim and M. M. Krunz, "Bandwidth allocation in wireless networks with guaranteed packet-loss performance," *IEEE/ACM Trans. Netw.*, vol. 8, no. 3, pp. 337–349, Jun. 2000.

[14] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of services," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[15] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assurance," *IEEE Trans. Veh. Technol.*, vol. 53, no. 5, pp. 1547–1557, Sep. 2004.

[16] 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.

[17] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall PTR, 2001.

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

[19] V. V. Veeravalli and A. Sayeed, Wideband Wireless Channels: Statistical Modeling, Analysis and Simulation, Univ. Illinois, 2004.

[20] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, ser. Universitext, 6th ed. Berlin, Germany: Springer-Verlag, 2003.

[21] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

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

[23] F. P. Kelly, *Reversibility and Stochastic Networks*. New York: Wiley, 1979.

[24] B. Hajek, Analysis of Computer Networks Univ. Illinois, 2003.

[25] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Probab.*, vol. 20, pp. 646–676, Sep. 1988.

[26] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. New York: Springer-Verlag, 1996.

[27] 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.

[28] 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.

[29] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Sel. Areas Commun.*, vol. 4, no. 6, pp. 856–868, Sep. 1986.

[30] *Mobile Station—Base Station Compatability Standard for Dual-Mode Wideband Spread Spectrum Cellular Systems*, Std. TIA/EIA/IS-95, Telecommunications Industry Association, Jul. 1993.

[31] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.

[32] J.-F. Chamberland and V. V. Veeravalli, "Decentralized dynamic power control for cellular CDMA systems," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 549–559, May 2003.

[33] N. Bambos and S. Kandukuri, "Power-controlled multiple access schemes for next-generation wireless packet networks," *IEEE Wireless Commun.*, vol. 9, no. 3, pp. 58–64, Jun. 2002.

[34] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, ser. Graduate Texts in Mathematics, 2nd ed. New York: Springer-Verlag, 1997.

[35] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, ser. Stochastic Modeling and Applied Probability, 2nd ed. New York: Springer-Verlag, 1998.

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

[37] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 7, pp. 968–981, Sep. 1991.

[38] S. Shamai and S. Verdú, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1302–1327, May 2001.