Statistical Multiplexing and QoS Provisioning for Real-Time Traffic on Wireless Downlinks

Hemant M. Chaskar, *Member, IEEE*, and Upamanyu Madhow, *Senior Member, IEEE*

***Abstract*—Quality of service (QoS) provisioning in wireless networks involves accounting for the statistical fluctuations in the wireless channel quality, in addition to the traffic variability of interest in a purely wireline setting. In this paper, we consider providing QoS to packetized, delay-constrained (real-time) applications over a Rayleigh-faded wireless downlink. Since the wireless medium is prone to high error rates with typically correlated errors, it is essential to use some kind of link-layer error-recovery mechanism to provide the desired level of reliability. We call this procedure of converting a link with frequent and correlated errors into a near-lossless packet pipe "link shaping." The link-shaping scheme considered in this paper exploits the natural interleaving provided by packet-by-packet transmissions to different mobiles to break up the error correlations due to Rayleigh fading and employs forward error correction (FEC) coding on the interleaved data. In addition to considering static (peak-rate) bandwidth sharing as in conventional wireless downlinks, we propose mechanisms for statistical multiplexing of traffic, which lead to substantial capacity gains. For example, for 13 kb/s voice sources over a 1-Mb/s link, we obtain a two-fold capacity gain over static (peak-rate) bandwidth allocation.**

I. Introduction

UNLIKE wireline links, which can be modeled as lossless bit pipes, wireless links are unreliable due to their high bit error rates (BERs) and correlations among bit errors. Hence, for providing quality of service (QoS) on wireless links, it is essential to use some kind of link-layer error-recovery mechanism to increase their reliability. Here, this procedure of converting a link with frequent and correlated errors into a near lossless packet pipe is called "link shaping."

In this paper, we develop a framework for QoS provisioning with link shaping for packetized, delay-constrained (real-time) applications over a Rayleigh-faded wireless downlink. The link-shaping scheme considered in this paper exploits the natural interleaving provided by packet-by-packet transmissions to different mobiles to break up the error correlations due to fading, and employs forward error correction (FEC) coding on the interleaved data. Another widely used mechanism for link shaping is automatic repeat request (ARQ), but that is more appropriate for applications (such as data) that do not have tight delay constraints (see [1] for an example of link shaping using ARQ for data applications controlled by TCP, the Internet transport protocol). The descriptors for QoS are chosen to be the acceptable packet loss rate and the worst-case delay since these are the relevant parameters for the real-time applications such as voice or video. In addition to considering static (peak-rate) bandwidth sharing as in conventional wireless downlinks, mechanisms for statistical multiplexing of traffic that provide significant capacity gains are also considered.

Most previous studies on link-layer protocols focus on comparing the throughput efficiency of variants of FEC, ARQ, and hybrid schemes, using different models for the statistics for packet errors. Typically, these studies consider a single source which always has data to send (see [2]–[4]). Studies that address traffic variability [5] typically assume simplified traffic and loss models, and consider conventional performance measures such as mean queuing delay. Unlike these, we address link-layer recovery from the viewpoint of QoS provisioning and traffic multiplexing, which are relatively modern notions. We incorporate, in detail, the characteristics of wireless medium, the parameters of link shaping scheme, and the traffic characteristics, while paying explicit attention to QoS parameters.

The remainder of the paper is organized as follows. The formulation of the QoS provisioning problem for delay-constrained traffic on wireless links is given in Section II. The link shaping (interleaving and coding) scheme considered is described in Section III, along with guidelines for its efficient design. Mechanisms for statistical multiplexing of traffic on a wireless downlink are proposed in Section IV. Section V contains performance analysis of the link-shaping and the bandwidth-sharing schemes. Numerical results and conclusions are contained in Sections VI and VII, respectively.

II. QoS Provisioning with Link Shaping

Consider transmission of fixed-length packets over a wireless downlink. The QoS is specified by (p, D), where p and D are the acceptable packet loss rate and the worst-case delay, respectively. Packets that arrive at the base station for being forwarded to a mobile terminal encounter two main mechanisms causing delay and loss: 1) queuing delay at the base station buffer and packet loss due to buffer overflow (or packet dropping due to deadline violation while waiting for transmission) and 2) link-shaping delay at the base station, which in our case consists of interleaving and coding delay and loss due to decoding failure at the mobile.

Queuing delay can be controlled by implementing deadline-based queuing at the base station. We denote by D_1 the resulting

Manuscript received January 1, 1999; revised August 1, 2000. This work was supported by the U. S. Army Research Office under Grants DAAH04-95-1-0246 and DAAG55-98-1-0219.

H. M. Chaskar is with Nokia Research Center, Boston, MA 01803 USA (e-mail: Hemant.chaskar@nokia.com).

U. Madhow is with the Electrical and Computer Engineering Dept., University of California, Santa Barbara, CA 93106 USA (e-mail: madhow@ece.ucsb.edu).

Publisher Item Identifier S 0733-8716(01)00118-4.

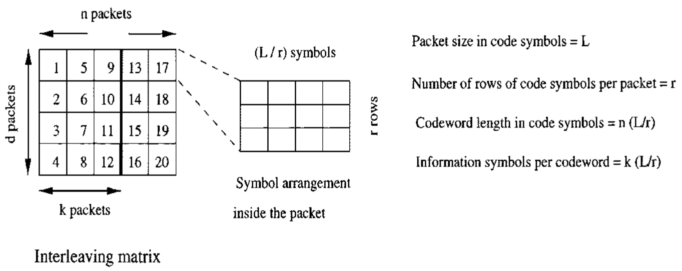
n packets
d packets
1 5 9 13 17
2 6 10 14 18
3 7 11 15 19
4 8 12 16 20
k packets
Interleaving matrix
(L / r) symbols
r rows
Symbol arrangement inside the packet
Packet size in code symbols = L
Number of rows of code symbols per packet = r
Codeword length in code symbols = n (L/r)
Information symbols per codeword = k (L/r)

Fig. 1. Interleaving and coding scheme (packet numbers in order of transmission).

upper bound on the queuing delay. The packet loss rate due to deadline violation while waiting in the base station buffer is denoted by p_1. Similarly, let D_2 denote an upper bound on the link-shaping delay, and let p_2 denote the residual packet loss rate on the wireless link after link shaping. Then, it must be ensured that $D_1 + D_2 \leq D$ and that p_1 and p_2 are both of the same order of magnitude or less than p. In the next section, we describe the link-shaping scheme considered here, followed by a discussion of design guidelines for achieving the preceding QoS objectives.

III. Wireless Link Shaping: Coding and Interleaving

In a wireless environment, the signal power received at the mobile terminal varies with time due to fading, changes in the interference level, movement of the scattering objects around the portable terminal, etc. This causes the errors on the link to be highly correlated. A practical and widely used approach to this problem is to break up these correlations by using interleaving, which ensures that successive (groups of) symbols of the same codeword are separated by a certain time interval called the "interleaving depth." It is denoted in this paper by d. The errors in the interleaved data are less correlated. This results in a more even distribution of channel errors among different codewords, thus improving the performance of the FEC coding strategy used.

Fig. 1 shows the interleaving scheme that we consider. Each square in the figure corresponds to a packet (which consists of a number of code symbols). The (information) packets in the base station buffer are transmitted on the wireless link as soon as possible and at the same time are forwarded to the interleaving matrix for computation of parity-check symbols. Thus, in the interleaving matrix, information packets are written column-wise but are encoded row-wise for the generation of parity-check packets. Different information packets in the interleaving matrix may come from different connections, depending on the bandwidth allocation strategy used. After every kd transmission slots, $(n-k)d$ parity-check packets are generated and are transmitted column-wise.

A. Improving Link Shaping Efficiency

Denoting by L the number of code symbols per packet, one possibility is to let each code symbol in a packet correspond to a different codeword. Thus, each row of n packets would actually correspond to L rows of codewords. More generally, suppose that each row of packets corresponds to r rows of codewords, where r divides L. Then, L/r symbols from each packet occur in the same codeword consecutively, so that the interleaving breaks up error correlations only partially. Since each row is n packets long, the codeword length is nL/r, and the number of information symbols per codeword is kL/r. This thus gives $(nL/r,\, kL/r)$ code. For the purpose of illustration, we consider maximum distance separable (MDS) Reed–Solomon (RS) codes [6, p. 174], so that the number of correctable errors is $t = \lfloor (L/r)((n-k)/2) \rfloor \geq L/r \lfloor (n-k)/2 \rfloor$. Note that, although RS codes are considered here, the general framework applies to other forms of FEC such as convolutional coding. We consider bounded-distance decoding [6, Ch. 7], whereby more than t symbols of the codeword are received in error, and thus, the probability of decoding into an incorrect codeword is much smaller than that of declaring decoding failure.

The following qualitative argument shows that, for fixed n, while increasing L/r leads to increased correlation in the symbol errors, it is more than offset by the following two factors: 1) better error correction capability (for the same code rate) obtained due to a larger code blocklength ($= nL/r$ symbols) and 2) large L/r, meaning that each row of packets comprises of less number of codewords. Hence, to decode a row of packets correctly, fewer of codewords need to be decoded correctly.

To see item 1, suppose first that $r = L$, so that different symbols in the same packet appear in different codewords. In other words, different symbols in the same codeword are transmitted in different packets, spaced $d-1$ transmission slots apart. For large enough d, such packets encounter independent fades. For this (n, k) RS code, the probability of decoding failure is $P[X_1 + \cdots + X_n > \lfloor (n-k)/2 \rfloor]$, where X_i is 1 if the ith code symbol is in error, and 0 otherwise. Now, suppose that $r < L$. The code in this case is an $(nL/r,\, kL/r)$ RS code, which contains n groups of L/r symbols each, with entire group being transmitted in the same packet. The probability of decoding failure is now upper bounded by $P[Y_1 + \cdots + Y_n > L/r \lfloor (n-k)/2 \rfloor]$, or $P[\hat{X}_1 + \cdots + \hat{X}_n > \lfloor (n-k)/2 \rfloor]$, where Y_i denotes the number of symbol errors within the ith group, and where $\hat{X}_i = Y_i/(L/r)$. Since the channel states seen by the code symbols in the same group are highly correlated, we assume that they are identical. Conditioned on this channel state, the average $\hat{X}_i$ has the same mean but smaller variance than the random variables X_i that appear in the expression for decoding failure probability for the case $r = L$. Thus, conditioned on the (independent) channel states seen by each of the n groups of symbols, the decoding failure probability is expected to be smaller than the corresponding conditional probability for the case $r = L$. Removing the conditioning on channel states, we expect the overall decoding failure probability to decrease as well.

Table I shows some numerical results from computer simulations, which verify the qualitative argument presented above. The simulations are run using Jakes' Rayleigh-fading simulator. RS code over GF (256) is used in simulations. Thus, each RS code symbol is byte-sized. For the purpose of relating received SNR to error probability, bits are assumed to be transmitted using binary differential phase shift keying. A detailed description of the overall simulation setup used in this study is given in Section V-A.

TABLE I
21 dB MEAN SNR, 50 Hz FADING, 25 PACKET INTERLEAVING DEPTH, 1 Mb/s LINK SPEED

Code rate	Decoding failure probability for a row of packets (10^{-3})			
	$L/r = 50$	$L/r = 10$	$L/r = 5$	$L/r = 2$
	250 symbols block	50 symbols block	25 symbols block	10 symbols block
0.2	0.001	0.001	0.001	0.002
0.4	0.043	0.106	0.162	0.287
0.6	0.188	0.304	0.318	0.588
0.8	13.846	22.063	34.974	33.631

B. *Design Guidelines*

With the above link-shaping scheme, the following guidelines can be obtained for the design of link layer at the base station.

1) *Delay Budget:* When the link-shaping scheme described above is employed, the link-shaping delay is $D_2 = (n-1)d \approx nd$ transmission slots because any transmitted packet has to wait for the transmission of the last packet in its row before the codewords associated with it can be decoded. This leaves $D_1 = D - (n-1)d$ slots as the allowable queuing delay. A natural buffering strategy with the preceding link-shaping scheme is the following. Defining a "frame" to be the duration of transmission of an $n \times d$ interleaving matrix, it is natural to express the allowable delay in terms of the number of frames. For instance, one choice might be to wait for transmission for at most one frame, i.e., all traffic that arrives during the transmission of the current frame is either transmitted in the next frame, or is dropped. This corresponds to $D_1 = D_2 \approx D/2$ transmission slots. More generally, we have $D_1 = bD_2 \approx (b/(b+1))D$ transmission slots for a nonzero integer b.

2) *Interleaving Depth:* The interleaving depth d should be large enough to break up error correlations efficiently. The value of d required for this purpose depends on the link speed and the fading frequency. For a given link speed, the larger the fading frequency, the smaller is the required value of d. The actual value of d used should be such that the correlation between the samples of the fading process separated d apart is small, e.g., less than 0.2. If d is further increased, the following two conflicting factors come into play: a) Increasing d results in smaller n for the same link-shaping delay budget. This makes it difficult to use efficient FEC parameters for link shaping. This is because smaller n results in less flexibility in the choice of code rate and smaller block lengths. b) Increasing d improves the efficiency of traffic multiplexing (see Section IV).

3) For any choice of n and d, the largest possible (limited by decoding complexity considerations) blocklength should be used. This is done by using small values of r (note that block length for the code is nL/r).

Of course, for any choice of parameters above, the code rate k/n should be such that the residual packet-loss rate $p_2 \leq p$. In view of the many variables involved in the above design, a good design is the one that gives largest admission region for the given traffic profile.

IV. TRAFFIC MULTIPLEXING

The total bandwidth on the wireless link is to be shared among M connections, each employing link shaping and seeking QoS. The previously described link-shaping mechanism can be used in conjunction with a number of bandwidth-sharing schemes. For simplicity of exposition, consider the case where $D_1 = D_2$, in detail. The general case (i.e., $b \geq 1$) is conceptually similar. When $b > 1$, the buffer has multiple partitions, each corresponding to a particular frame. It is more scalable to implement buffering in time units of frames.

A. *Static (Peak-Rate) Bandwidth Partitioning*

In this, all connections have a fixed bandwidth allocation in all frames (interleaving matrices). The basic unit of bandwidth allocation is a single row of the interleaving matrix. Thus, if c_i is the bandwidth (in terms of number of rows) given to the ith connection, then we must have $\sum_{i=1}^{M} c_i \leq d$. Note that c_i corresponds to the peak rate of the connection.

Next, we describe two mechanisms for statistical multiplexing, which are consistent with the link-shaping mechanism of Section III. The first, "packet-level multiplexing," requires that all users demodulate all packets and is impractical in many situations, as discussed in Section IV-B. Nevertheless, it provides large multiplexing gains, which serve as a performance benchmark for the more practical and versatile "row-multiplexing" scheme proposed in Section IV-C, which requires only that each user demodulate its own packets.

B. *Packet-Level Multiplexing*

We assume that each connection requires the same coding and interleaving mechanism. Since the wireless downlink is a broadcast channel, every receiver can listen to the transmissions in all time slots. Assuming that every mobile can demodulate the data in all the downlink transmission slots, consider a scheme in which packets from all connections are first multiplexed into a shared buffer and then fed into a common interleaving and coding matrix. No distinction is made at the (base station) link layer among data packets destined for different receivers. Each receiver decodes the entire matrix and, after extracting the information packets, forward only data destined to itself (identified by the packet headers) to higher layers. Thus, in this scheme, both the data and the parity-check packets within the codewords are shared across the connections. In other words, different packets in the information part of a given row of the interleaving matrix may come from different connections, and all these packets, along with the parity-check packets in the row, are used by every connection to decode the codewords in this row.

The preceding scheme relies on all mobiles being able to demodulate all packets. This broadcast assumption could easily be violated, for example, in the following scenarios: systems with forward link-power control or electronic beamforming, or systems in which mobile users are in sleep mode much of the time to conserve power and wake up only in time to receive their own packets. However, as described next, it is also possible to obtain (a smaller) multiplexing gain without the broadcast assumption underlying packet multiplexing.

C. Row Multiplexing

If transmissions to different mobiles use different mobile-specific values for transmission parameters such as transmitted power or angular position of antenna beam (e.g., as in electronic beamforming), each mobile must be able to decode its own data independently. In that case, the basic unit of bandwidth allocation again becomes a row of the interleaving matrix. In the "row-multiplexing" strategy, a dynamic number of rows are allocated to different connections in each frame. While a number of allocation methods are possible, we consider for the purpose of illustration the following "greedy policy," which is appropriate for connections with identical characteristics (in terms of traffic, channel statistics, and QoS requirements).

At the beginning of each frame, the scheduler allocates rows to connections in a manner so as to fill up the interleaving matrix to the greatest possible extent. This greedy policy is described as follows. Let $B_i^{(l)}$ denote the number of packets of connection i remained to be assigned to interleaving matrix after the lth step of the allocation algorithm, during a particular frame. Set $B_i^{(0)} = A_i$, the number of packets of connection i that arrive during that frame. Define $i_l^* = \arg\max_i B_i^{(l)}$. If a row is available after l steps of allocation, assign it to connection i_l^*. Set

$$B_i^{(l+1)} = \begin{cases} B_i^{(l)}, & \text{if } i \neq i_l^* \\ \left(B_i^{(l)} - k\right)^+, & \text{if } i = i_l^*. \end{cases}$$

The allocation procedure during a given frame terminates when all the rows of the interleaving matrix have been allocated. Once all rows have been allocated, all data that has not been assigned to the matrix is dropped. More sophisticated policies (e.g., enforcing fairness or providing differential QoS) can be easily devised, but the purpose here is to demonstrate the multiplexing gain obtainable from this strategy.

1) Other Applications of Row Multiplexing: Before proceeding further, we point out some other scenarios where the row-multiplexing model is applicable. For example, in a code division multiplexed downlink, interference considerations may require limiting the number of downlink codes that are simultaneously active. Suppose that d downlink codes are allowed be active simultaneously. Note that, in this case, d has the meaning of total number of simultaneously active codes and not the interleaving depth. Each row of Fig. 1 would then correspond to a downlink code. Depending upon the row schedule in a given frame, only the codes of those connections which have row(s) assigned in that frame remain active, while the rest are shut off during that frame. Another example is the case of downlink transmission in an indoor environment. For indoor applications, since the fading is very slow, one has to deploy some kind of diversity techniques to avoid prolonged link outage. One solution [7] is to use frequency diversity along with an erasure correcting code, e.g., RS code. Each row in Fig. 1 then corresponds to a distinct (mobile-specific) frequency-hopping sequence. The performance evaluation of row multiplexing for the preceding scenarios is beyond the scope of this paper. In this paper, we provide numerical example for time-division multiple-access-based (TDMA) wireless downlink only (Section VI).

2) Adaptive Link Shaping: The row-multiplexing scheme can also combine traffic multiplexing with adaptive link shaping. In adaptive link shaping, the code rate for each connection could be different, depending on its QoS requirement and the quality of the channel from the base station to its destination. Such online estimates of the channel quality can be obtained by SNR measurement at the mobile terminal [8]. In row multiplexing, it is possible to change the code rate of a particular row according to the identity of the connection to which it is allocated in a given frame. To illustrate this, suppose that the code rate for connection i is R_i. In that case, when a row is allocated to connection i in a given interleaving matrix (according to some scheduling policy), $k_i = nR_i$ data packets of connection i can be transmitted in that row (with the remaining $n - k_i$ packets being the parity-check packets). Thus, the *total* bandwidth allocated to the connection in a given frame equals the number of rows allocated, while the *data* bandwidth allocated to the connection is R_i times the actual bandwidth allocated.

V. Performance Analysis

Following are the main performance issues that underlay our design of the link layer: 1) link-shaping performance, or evaluation of the packet loss rate due to decoding failure, accounting for the statistics of the symbol errors on the wireless link; 2) queuing performance, or evaluation of the packet loss rate at the base station, due to deadline violation while waiting for transmission, accounting for traffic statistics and the bandwidth-sharing scheme.

We address these issues for the case of identical connections (identical traffic characteristics) seeking the same QoS [i.e., same $(p,\, D)$ for all connections]. The downlink channels from the base station to different mobiles are modeled as independent and identically distributed Rayleigh-faded channels. While such a homogeneity is not necessary for the application of the proposed framework, this model is chosen for the purpose of illustration of basic ideas.

A. Link-Shaping Performance

In bounded distance decoding of an $(nL/r,\, kL/r)$ RS code, the decoder outputs the information symbols whether or not it manages to decode. When more than $\lfloor (L/r)\,((n-k)/2) \rfloor$ symbols of the codeword are received in error, the probability of decoding into an incorrect codeword is much smaller than that of declaring a decoding failure. From Fig. 1, note that a packet is correctly received only if every codeword associated with it is either decoded correctly, or, if there is a decoding failure, none of the information symbols in the packet are wrong. Given the complexity of accounting for the channel model and the link-shaping scheme, obtaining analytical expressions for the packet loss rate due to decoding failure at the receiver appears to be difficult. We therefore employ simulations to obtain the packet loss rates for the link-shaping scheme of Section III over a Rayleigh-fading channel.

Simulation Set-Up: For simulations, we fix a particular packet position in all the interleaving matrices and simulate the fading process only for the codewords that are relevant for this packet position. According to the Rayleigh-fading model (see [9, Ch. 2]) for frequency nonselective (flat) fading, the received signal at the mobile surrounded by scatterers is represented as the superposition of a number of multipaths. The multiplicative gain due to fading can be written as a complex random process

$$F(t) = \sum_i \alpha_i \exp(-j2\pi f_{D,i} t)$$

where α_i denotes the amplitude of the ith multipath and $f_{D,\,i}$ its Doppler frequency shift. In the limit of a continuum of multipath components uniformly distributed in space, the fading gain $F(t)$ is approximated by (resorting to the central limit theorem) a complex Gaussian random process, so that $|F(t)|$ is a Rayleigh random variable for each t (hence the name Rayleigh fading). Also, $F(t)$ can be shown to have Bessel-functionlike autocorrelation function. This is termed as Clarke's Rayleigh-fading model. We employ the popular Jakes' simulator [9], [10] to simulate Clarke's model. In Jakes' simulator, $F(t)$ is approximated by the superposition of a finite number of spatially distributed multipath components. Specifically, in our simulations, we set $F(t) = \sum_{i=1}^{\sigma} \alpha_i \exp(-j2\pi f_{D,i} t)$ with $\sigma = 50$. α_is are sampled from the uniform distribution over [0, 1] and normalized to keep the average power $E[1/2|F(t)|^2] = (1/2)\sum_{i=1}^{\sigma} \alpha_i^2$, at a nominal value. Doppler frequency shift of the ith multipath is obtained as $f_{D,\,i} = f_D \cos\theta_i$, where f_D is the maximum Doppler shift (which is the function of the velocity of the mobile and the wavelength of the carrier), and θ_i is the direction of arrival of the ith multipath. We take $\theta_i = (i/50)(2\pi)$ for $i =$ 1 to 50. Once the values of α_is and $f_{D,i}$s are obtained in this fashion, they are not changed during the simulation. This generates a realization (i.e., a sample path) for the fading process. The signal power at time t is given by $1/2|F(t)|^2$. We assume the fading gain to be constant over the duration of transmission of an RS symbol (which could correspond to multiple bit transmissions). Then, the SNR in any bit interval, during the transmission of an RS symbol starting at time t, is given by $\gamma_b = 1/2|F(t)|^2 T_b$, where T_b is the bit duration, and the noise variance is normalized to unity. For the numerical investigation, we use the SNR-bit error probability law for binary differential PSK [11, p. 276]. Thus, for an m bit RS symbol, with bit energy-to-noise ratio of γ_b, the probability of code symbol error is given by $1-[1-(1/2)e^{-\gamma_b}]^m$. Simulations were run with 90% confidence interval.

B. Queuing Performance

To evaluate the queuing performance of the bandwidth-sharing schemes of Section IV, the packet stream of each connection arriving at the base station is modeled as generated by a discrete time Markovian source. The transitions of this Markov source occur at the slot boundaries and, depending on the new state visited, a certain number of packets arrive at the base station during that slot. Denote by P the transition probability matrix and by $\mathcal{S}$ the state space of this Markov chain. The distribution of packet arrivals in state $s \in \mathcal{S}$ is assumed to be bimodal, with probability ν_s of q_s arrivals and $1-\nu_s$ of no arrivals. The bimodal model enables the scaling of arrival rate, according to the speed of transmission.

Let the random variable A_i denote the amount of traffic (number of packets) of connection i, that arrives over a frame (of duration nd slots). Let $G(j) = P[A_i > j]$ denote the complementary cumulative distribution function (ccdf) of A_i. The ccdf $G(\cdot)$ is computed from the given traffic model by using the following recursive relation:

$$\begin{aligned} G_s^f(j) = & \nu_s \sum_{t\in\mathcal{S}} P(s,t) G_t^{f-1}(j-q_s)^+ \\ & + (1-\nu_s) \sum_{t\in\mathcal{S}} P(s,t) G_t^{f-1}(j) \end{aligned}$$

where $G_s^f(j)$ is the probability that more than j packets arrive over the frame of duration f slots, starting from state s in the first slot. $G(j)$ is then obtained as $\sum_{s\in\mathcal{S}} \pi_s G_s^{n\,d}(j)$, where π_s, for $s \in S$, is the stationary distribution for P.

1) Packet Multiplexing: For this scheme, note from Section IV-B that, if the sum of the number of packets from *all* connections arriving in a frame exceeds kd, then the excess packets are lost. An upper bound on the probability of occurrence of such a loss event (which approximates p_1), namely the event $\{\sum_{i=1}^{M} A_i > kd\}$, can be obtained via well-known Chernoff's bound. The probability of such an event is to be kept below p. Note that this is similar to the bufferless multiplexing model studied in [12] in the context of multimedia wireline networks. For the values of M and A_i considered in our numerical results (see Section VI), it turns out to be feasible to compute p_1 exactly by finding the probability mass function of A_{tot} obtained by convolution of those of A_is, where $A_{\text{tot}} = \sum_{i=1}^{M} A_i$. Then, for packet multiplexing with identical connections

$$p_1 = \frac{1}{M} \frac{E(A_{\text{tot}} - kd)^+}{EA_{\text{tot}}}.$$

2) Row Multiplexing: For row multiplexing with identical connections, the objective of the analysis is to find the expected number of total packets discarded (summed over all connections) in any frame after performing the greedy row allocation described in Section IV-C. Let the random variable A represent the number of packets of any connection that arrive per frame (all connections are assumed to be statistically identical, and hence, $A = A_i$, in distribution for all i). Assume that $A \le Jk$ with probability one, where J is an integer. We will describe the analysis for the case $J = 2$ here. Generalization to larger J is straightforward. Denote by $G(j) = P[A > j]$ the complementary cumulative distribution function (ccdf) of A. Define

$$M_1 = \sum_{i=1}^{M} I_{[A_i \le k]} \qquad M_2 = M - M_1 = \sum_{i=1}^{M} I_{[A_i > k]}.$$

The joint distribution of $(M_1,\, M_2)$ is

$$\begin{aligned} & \text{Prob}\,[M_1 = m_1,\, M_2 = m_2] \\ & \quad = \frac{M!}{m_1! m_2!} [1 - G(k)]^{m_1} [G(k) - G(2k)]^{m_2}. \end{aligned}$$

Note that $G(2k) = 0$ for $J = 2$. The expected number of packets discarded in any frame is computed by finding the same by first conditioning on M_1 and M_2, and then using the above joint distribution to remove the conditioning. To calculate the expected number of packets lost conditioned on $M_1 = m_1$ and $M_2 = m_2$, we have to consider two regimes when $J = 2$.

Regime 1 ($m_2 \geq d$): In this regime, the greedy algorithm allocates d rows to the connections with the largest amount of traffic among the m_2 connections. However, the number of packets discarded during such a frame does not depend on the choice of d connections from among these m_2. We may therefore assume that d connections are chosen at random from among these m_2 connections. To proceed, define the following ccdfs:

$$G_1(j) = P[A > j \mid A \leq k], \qquad G_2(j) = P[A - k > j \mid A > k]$$
$$G_3(j) = P[A > j \mid A > k].$$

After the row allocation in such a frame, we are left with m_1, d, and $m_2 - d$ connections with ccdfs $G_1(\cdot)$, $G_2(\cdot)$, and $G_3(\cdot)$, respectively, for their unassigned traffic, all of which incurs loss. The expected number of packets lost can now be now easily computed as

$$m_1 \sum_{j\geq 0} G_1(j) + d \sum_{j \geq 0} G_2(j) + (m_2 - d) \sum_{j \geq 0} G_3(j).$$

Regime 2 ($m_2 < d$): The algorithm first allocates rows to these m_2 connections. This leaves $d_1 = d - m_2$ rows still to be allocated among M connections. The number of packets remaining to be allocated ("yet unassigned") for each of the m_1 connections which have had no allocation thus far has ccdf $G_1(\cdot)$ and the residual ("yet unassigned") number of packets among each of the m_2 connections that have received a row has ccdf $G_2(\cdot)$. d_1 rows are now allocated to the connections with the largest amount of yet unassigned traffic among M, m_1 of which have ccdf $G_1(\cdot)$ and m_2 of which have ccdf $G_2(\cdot)$ for their yet unassigned traffic. This causes $M - d_1$ connections with the smallest amount of yet unassigned traffic among M to lose their packets. We can compute the expected number of packets lost at the lth smallest position, where $1 \leq l \leq M - d_1$ (with $l = 1$ corresponding to the position with the smallest amount of yet unassigned traffic among M), as

$$\sum_{j \geq 0} \mathrm{Prob}[l\text{th smallest position has more than } j \text{ packets}]$$

For the event $\{l$th smallest position has more than j packets$\}$ to occur, at least $M - l + 1$ among M must have more than j yet unassigned packets. Thus, its probability is

$$\sum_{\substack{s_1 + s_2 \geq M - l + 1;\\ 0 \leq s_i \leq m_i}} \prod_{i=1}^{2} \binom{m_i}{s_i} [G_i(j)]^{s_i} [1 - G_i(j)]^{m_i - s_i}.$$

The expected number of total packets lost is obtained by summing the expected number of packets lost at each of the $M - d_1$ positions with the smallest amount of yet unassigned traffic among M. Hence, it is given by

$$\sum_{l=1}^{M-d_1} \sum_{j \geq 0} \sum_{\substack{s_1 + s_2 \geq M - l + 1;\\ 0 \leq s_i \leq m_i}} \prod_{i=1}^{2} \binom{m_i}{s_i} [G_i(j)]^{s_i} [1 - G_i(j)]^{m_i - s_i}.$$

Remark: The analysis for the case $J > 2$ is conceptually similar with the only difference that, in that case, there will be more than two regimes. For example, when $J = 3$, there are three regimes, namely $2m_3 \geq d$; $2m_3 < d$, $2m_3 + m_2 \geq d$; and $2m_3 + m_2 < d$, where

$$M_1 = \sum_{i=1}^{M} I_{[A_i \leq k]}, \qquad M_2 = \sum_{i=1}^{M} I_{[k < A_i \leq 2k]}$$
$$M_3 = M - M_1 - M_2 = \sum_{i=1}^{M} I_{[A_i > 2k]}.$$

VI. Numerical Example

Consider traffic from a number of 13 Kb/s voice sources traversing a TDMA-based wireless downlink. Each voice source is modeled statistically using a two-state Markov chain with a voice activity factor of 40%. This voice-source model is derived from the model in [13] for 32 Kb/s packetized voice by scaling down the peak rate in the active state of the source. The nominal channel model for all connections is the Rayleigh model for frequency nonselective fading with a (maximum) Doppler frequency of 50 Hz. We assume packetized transmission in terms of ATM cells, each consisting of 50 bytes. Thus, each transmission slot on the downlink corresponds to the transmission of 50 bytes. We use byte-sized symbols (so that $L = 50$), and therefore employ RS codes on GF(256). This limits the maximum block length for the code to be 256 symbols (see [6, p. 174]), any smaller block length being obtained by using the truncations of the RS code on GF(256). The values of (transmitted) SNR employed in the simulations range from 15 to 24 dB. Intercell interference can be modeled explicitly, but can be considered in our framework by suitably adjusting the SNR values.

First, a set of results is presented for the downlink transmission speed of 1 Mb/s. These results are intended to demonstrate the various tradeoffs involved in link shaping. For 1 Mb/s transmission speed, the time duration of transmission of one packet (cell) on the link turns out to be 0.4 ms. In one case, the QoS required by all connections is taken to be a cell loss ratio of no more than 10^{-3} and a worst-case delay of 100 ms, while in the other case, the QoS required by all connections is taken to be a cell loss ratio of no more than 10^{-2} and a worst-case delay of 100 ms. The delay requirement of 100 ms translates into $D = 250$ transmission slots.

Beyond a certain minimum value, there are two factors governing the choice of d. Smaller d (and hence larger n) allows the use of more efficient codes due to the greater flexibility in the choice of code rates (allowing the designer to use the code rate that is just right) and also due to the possibility of using larger

TABLE II
LOW SNR DESIGN: 18 dB TRANSMITTED MEAN SNR, TARGET CELL LOSS RATIO = 10^{-3}

d	n	Code parameters	p_2	Number of connections		
slots	packets	in symbols	(10^{-3})	Static	Packet Mux.	Row Mux.
62	2	(100,50)	22.4*	31	66	66
41	3	(150,50)	4.4*	20	38	38
31	4	(200,50)	3.8	15	26	26
25	5	(250,100)	4.1	25	49	38

TABLE III
HIGH SNR DESIGN: 28 dB TRANSMITTED MEAN SNR, TARGET CELL LOSS RATIO = 10^{-3}

d	n	Code parameters	p_2	Number of connections		
slots	packets	in symbols	(10^{-3})	Static	Packet Mux.	Row Mux.
62	2	(100,50)	2.9	31	66	66
41	3	(150,100)	2.8	41	94	72
31	4	(200,150)	2.5	31	108	64
25	5	(250,200)	2.0	25	117	54

TABLE IV
LOW SNR DESIGN: 15 dB TRANSMITTED MEAN SNR, TARGET CELL LOSS RATE = 10^{-2}

d	n	Code parameters	p_2	Number of connections		
slots	packets	in symbols	(10^{-2})	Static	Packet Mux.	Row Mux.
62	2	(100,50)	10.6*	31	82	82
41	3	(150,50)	2.0	20	50	50
31	4	(200,50)	1.9	15	37	37
25	5	(250,150)	3.0	25	110	60

TABLE V
HIGH SNR DESIGN: 24 dB TRANSMITTED MEAN SNR, TARGET CELL LOSS RATE = 10^{-2}

d	n	Code parameters	p_2	Number of connections		
slots	packets	in symbols	(10^{-2})	Static	Packet Mux.	Row Mux.
62	2	(100,50)	1.6	31	82	82
41	3	(150,100)	1.9	41	115	90
31	4	(200,150)	2.9	31	130	80
25	5	(250,200)	3.0	25	140	70

TABLE VI
ROW MULTIPLEXING GAIN FOR THE TARGET CELL LOSS RATE OF 10^{-2}

Downlink Transmission Speed	Row Multiplexing Gain	
	Low SNR	High SNR
125 Kb/s	1.0	1.2
250 Kb/s	1.5	1.5
500 Kb/s	1.9	1.8
1 Mb/s	2.4	2.2

block lengths. On the other hand, d is the number of "independent" bandwidth bins available in each frame for allocation to different connections in the row multiplexing scheme of Section IV-C. From this viewpoint, larger d is desirable as it allows more flexibility in the bandwidth assignment, which may result in more efficient bandwidth usage. In the light of the preceding tradeoffs, a good design is one that gives the largest admission region for the given traffic profile.

These tradeoffs are seen in Tables II–V. Note in Tables II and IV that the desired link-shaping performance cannot be obtained for $d = 62$ (marked by * in Tables II and IV). In Table II, the required code rate with $d = 31$ is much smaller compared to that required with $d = 25$, so that the admission region is bigger for $d = 25$ (for packet- and row-multiplexing schemes). Also, the number of independent bandwidth bins for row multiplexing with $d = 41$ is larger than that with $d = 31$, and hence, the admission region is bigger for $d = 41$ despite it having smaller code rate than the latter. Similar tradeoffs are seen in Table IV as well.

When the SNR is high (Tables III and V), it is easier to do link shaping. In particular, observe that all the codes considered in Tables III and V give the desired link-shaping performance. Also, the codes at the bottom of the table have higher code rates than those at the top. However, the admission region is smaller with row multiplexing for the bottom entries of Tables III and V because reduction in d reduces the efficiency of row multiplexing.

Another set of results is presented in Table VI to compare the statistical multiplexing gain obtainable with row-multiplexing scheme at various downlink transmission speeds. For a given transmission speed, the multiplexing gain is quantified as the ratio of maximum number of voice sources admissible with row multiplexing to that admissible with static (peak-rate) bandwidth partitioning. Note that, at low transmission speed (125 Kb/s), there is little difference between the capacities with static-bandwidth partitioning and row multiplexing. This is due to the lack of sufficient number of independent bandwidth bins (rows) for row multiplexing. However, as the transmission speed increases, row multiplexing offers significant multiplexing gain over static-bandwidth partitioning.

VII. CONCLUSION

We have developed a framework for QoS provisioning for delay constrained traffic on a Rayleigh-faded wireless downlink, with detailed consideration of traffic as well as wireless channel characteristics. As seen from the numerical results, the requirement for wireless link shaping has a significant impact on QoS provisioning. Therefore, it is crucial to account for physical and link layer choices in network layer designs. We also proposed mechanisms for statistical multiplexing, compatible with the link-shaping strategy considered, that provide significant gains in system capacity. Our model-based approach facilitates the identification of various tradeoffs and approximate system sizing. However, as described next, there are a number of issues that are not fully addressed in this paper. These furnish interesting topics for further research.

A. Application to Existing Systems

The system parameters employed here were chosen for convenience of illustration, rather than for comparison with existing systems such as the TDMA-based GSM, IS-136, or EDGE cellular systems. Currently, these systems do not employ statistical multiplexing, but rather assign fixed TDMA slots corresponding to the peak rate of the voice connection, which amounts to the static allocation strategy in our terminology. It would be interesting to see whether our row-multiplexing strategy can lead to substantial multiplexing gain in these systems.

B. Heterogeneity

While we have restricted attention to identical connections with identical QoS requirements, extension of the scheduling and link-shaping strategies to heterogeneous connections (in terms of traffic, channel statistics, and QoS requirements) is an important topic of future research.

C. Measurement-Based Procedures

Our approach here is to use nominal traffic and channel models, and is therefore only applicable for approximate system sizing. Since wireless channels can vary considerably from the nominal model, and since assumption of a worst-case model would be inefficient, an important topic for future research is that of measurement-based resource allocation and admission control. Unlike in the wireline context, measurement-based procedures for wireless links would need to measure channel parameters in addition to traffic parameters. Including such measurements in the QoS provisioning framework will result in fully adaptive link shaping.

References

[1] H. Chaskar, T. Lakshman, and U. Madhow, "TCP over wireless with link level error control: Analysis and design methodology," *IEEE ACM Trans. Networking*, vol. 7, pp. 605–615, Oct. 1999.

[2] R. Comroe and D. Costello, "ARQ schemes for data transmission in mobile radio systems," *IEEE J. Select. Areas Commun.*, vol. SAC-2, pp. 472–481, July 1984.

[3] T. Sato, M. Kawabe, T. Kato, and A. Fukasawa, "Throughput analysis method for hybrid ARQ schemes over burst error channels," *IEEE Trans. Veh. Technol.*, vol. 42, pp. 110–118, Feb. 1993.

[4] M. Zorzi, "Performance of FEC and ARQ error control in bursty channels under delay constraints," in *Proc. IEEE Veh. Technol. Conf.*, Ottawa, Canada, 1998, pp. 1390–1394.

[5] D. Towsley and J. Wolf, "On the statistical analysis of queue lengths and waiting times for statistical multiplexers with ARQ retransmission schemes," *IEEE Trans. Commun.*, vol. COMM-27, pp. 693–702, Apr. 1979.

[6] R. Blahut, *Theory and Practice of Error Control Codes*. Reading, MA: Addison-Wesley, 1984.

[7] A. Saleh and L. Cimini, "Indoor radio communications using time-division multiple access with cyclical slow frequency hopping and coding," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 59–70, Jan. 1989.

[8] K. Balachandran, S. Kadaba, and S. Nanda, "Channel quality estimation and rate adaptation for cellular mobile radio," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1244–1256, July 1999.

[9] G. Stüber, *Principles of Mobile Communication*. Norwell, MA: Kluwer, 1996.

[10] W. Jakes, *Microwave Mobile Communications*. Piscataway, NJ: IEEE, 1993.

[11] J. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.

[12] R. Guérin, H. Ahmadhi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968–981, Sept. 1991.

[13] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 833–846, Sept. 1986.

Hemant M. Chaskar (S'96–M'97) was born in Pune, India, on January 26, 1972. He received the B.Eng. degree from the University of Pune in 1993, the M. Eng. degree from the Indian Institute of Science, Bangalore, in 1995, both in electronics and telecommunication engineering. He received the Ph.D. degree in electrical engineering from the University of Illinois, Urbana-Champaign, in 1999.

From August 1995 to February 1999, he was a Research Assistant with the Coordinated Sciences Laboratory, University of Illinois, Urbana-Chamapign. During the summer of 1997, he was an Intern with the Performance Analysis Department, Lucent Technologies-Bell Labs, Holmdel, NJ. Since May 1999, he was been with the Communication Systems Laboratory, Nokia Research Center, Boston, MA. With Nokia, he is currently working on architecture, services, mobility protocols, and radio resource management for third generation wireless IP networks. He has worked on areas related to multimedia Internet with emphasis on QoS provisioning using differentiated services and MPLS, high-speed packet scheduling, and congestion control. His research interests are in wireless communication, high-speed networks, communication theory, and applied mathematics.

Upamanyu Madhow (S'86–M'90–SM'96) received the B.S. degree in engineering from the Indian Institute of Technology, Kanpur, in 1985. He received the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois, Urbana-Champaign in 1987 and 1990, respectively.

From 1990 to 1991, he was a Visiting Assistant Professor with the University of Illinois. From 1991 to 1994, he was a Research Scientist with Bell Communications Research, Morristown, NJ. He was with the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign as an Assistant Professor from 1994 to 1998, and as an Associate Professor from 1998 to 1999. Since 1999, he has been an Associate Professor with the Electrical and Computer Engineering Department, University of California, Santa Barbara. His current research interests are in wireless communications and high-speed networks.

Dr. Madhow is a recipient of the NSF CAREER award. He is an Associate Editor for Spread Spectrum for the IEEE TRANSACTIONS ON COMMUNICATIONS and an Associate Editor for Detection and Estimation for the IEEE TRANSACTIONS ON INFORMATION THEORY.