Trade-off Between Energy and Delay in Wireless Packetized Systems

Todd Coleman and Muriel Médard

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
colemant@mit.edu, medard@mit.edu

October 2, 2001

Abstract

In packetized wireless systems, coding allows correct reception of multiple packets colliding at a receiver. Thus data may not need to incur delays such as those due to backoff schemes in traditional ALOHA systems. However, there is a trade-off between delay and energy consumption. Recent work in this area has considered the case where multiple users are aware of the states of other users' queues. We consider a time-slotted multiple user system with random packet arrivals. The size of the packets and probability of arrival together represent the burstiness of the system. The time slots are considered to be long enough that capacity can be achieved over a single slot in a sense we define. We consider the difference in average energy consumption when average delay is minimized, with and without knowledge of other users' queues. We also consider the case where average energy is minimized without regard for delay. We present and analyze a simple scheme with limited information sharing about queues' states. Our scheme uses a broadcast-type code for the case of low queue lengths and a multiple-acces scheme in the case of large queue lengths. We show how this scheme allows trade-offs between energy and delay.

1 Introduction

The performance of wireless nomadic data transfer systems can be characterized by a number of system qualities, including aggregate data rates, average bit transmission delay, and energy consumption. Information theoretic considerations attempt to establish ultimate limits on reliable communication. Shannon capacity assumes that a steady stream of bits is to be transmitted at all times. Many data transfer systems, however, exhibit random packet arrivals. The size of the packets and probability of arrival together represent the burstiness of the system. This violates the assumption that bits are always available for transmission.

The time-slotted ALOHA system models systems with bursty arrivals. Its use is motivated by its simplicity: users attempt to transmit data as it arrives in their transmission buffers. If two or more users transmit at the same time, a collision occurs at the receiver. Traditional ALOHA systems require users to transmit packets without explicit coordination among users. In the event of a collision, packets are discarded and users retransmit the collided packets. The capacity of such systems has generally been analyzed in terms of packet throughput.

The stability of classical ALOHA systems has been studied extensively. For an infinite number of users, it has been found in [Cap78] that the system is unstable. Stability regions have been found for systems with a finite number of users. To combat the instability, decentralized control schemes [Riv87] and conflict resolution schemes [Hay78] have been established. In general, such schemes attempt to avoid successive collisions by retransmitting with some backoff policy.

Modeling a collision as leading to loss of all packets at receivers is not always practical. The capture phenomenon, for instance, may yield correct reception of some portion of the data, for instance the data from coded slots. Moreover, users may be reliably received if, when transmitting, they take into account the worst case multiple access scenario that may arise. If coding can be implemented over sufficiently many bits, then users, when they transmit, may then use the types of codes that achieve rates on the Cover-Wyner multiple access rate region.

Stability analysis of systems with multiple-packet reception capability in the presence of channel noise has been performed in [GVS88]. The capacity region of such systems, in a sense we qualify later, that allows coding of packets *and* variable reliably received rates has recently been introduced [MMHG01]. It has been found that such a system's capacity region is the same as the capacity region of a multiple-access system where users continuously transmit. Furthermore, transmission policies that make use of detailed knowledge of users' queus, whether in a decentralized fashion or through a centralized control such as a scheduler, do not improve capacity. Hence, capacity of such systems is in general independent of burstiness and queue information availability. Many coding schemes were shown to be optimal, ensuring long-term stability while achieving rates inside the Cover-Wyner region. The impact of burstiness and queue information, however, when considering delay and energy consumption, was not illustrated in [MMHG01]. Clearly, queue information is not altogether useless. While it does not affect the type of capacity we consider, we would expect it to influence other performance parameters, such as delay.

An investigation of the trade-offs between minimum average power required to meet some quality of service cost (stringent delay or probability of buffer overflow) [Ber00] has been performed. This analysis addressed bursty multi-user channels in the presence of fading. The investigation used centralized control to combat both fading and burstiness to deliver a stringent delay constraint.

We consider the difference in average energy consumption when average bit delay is minimized to 0 under two different scenarios. We investigate a control scheme where the transmission policy of each user relies on detailed information about the amount of data in all users' queues. We also investigate a control scheme where each user's policy relies only on the amount of data in that particular user's queue. Hence, the impact of queue information sharing in the presence of burstiness is characterized. We also consider minimizing energy consumption without regard for delay, which turns out to be infinite. Finally, we present and analyze a simple scheme that has some optimal long-term stability properties, combats burstiness and collisions by superimposing codes anticipating different levels of interference, and is sensitive to average bit delay. The system uses a multiple access channel [Lia72] type code when all users' queue lengths are long. Otherwise, it uses a broadcast [Cov72] type code. Instead of combatting burstiness by performing probabilistic backoff policies after collisions, the system uses rate-splitting to achieve variable received rates as a function of the uncertainty of other

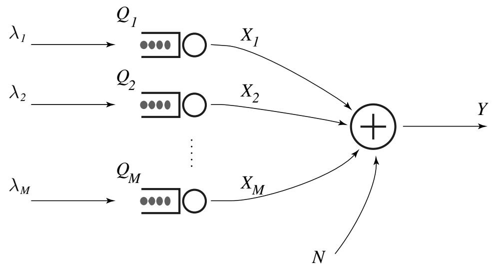

Figure 1: The M-user ALOHA model.

users' presence. As user queue lengths become long, the system switches to coding in multiple-access mode, where users code for each others' presence at optimal aggregate rates (as provided by the dominant face of the Cover-Wyner region for multiple access channels).

2 Model and Background

The multiple-access system we consider is illustrated in Figure 1. Our system differs from traditional time-slotted ALOHA systems in a number of ways:

- We model a multiple access where users share a single channel with no multiplicative attenuation but with additive white Gaussian noise (AWGN).
- Time slots are very long in terms of bits to be transmitted. Consequently, transmission data may be coded to achieve rates near information-theoretic bounds. We may also consider coding over several time slots, but note that this may complicate the discussion without providing much insight. This long time slot model may be particularly appropriate when dealing with channels with large signal-to-noise ratios, where turbo codes have been designed to achieve rates near capacity with sustainable probability of error. Coding also allows bits to be reliably received depending on the presence/absence of other users. Hence, collisions are not catastrophic and stability analysis is very different from that of traditional ALOHA.

2.1 Channel Model

The channel model we propose is a multiple access system where M users transmit to one receiver in the presence of additive white Gaussian noise (AWGN). The transmitters all share bandwidth of size W. The signal X_i of user i, along with the output, are bandlimited to W as well. Signals are sampled and synchronized. After sampling, the output and input are related at sample time t as

$$y[t] = \sum_{k=1}^{M} x[t] + N[t]$$

where $N[t]$ is a sequence of $\mathcal{N}(0, \sigma_N^2)$ i.i.d. random variables. Fading is not present in this model.

A user's queue contains all of its traffic which has not yet been successfully transmitted. Immediately before time slot n, $\vec{\triangle}(n) = (\triangle_1(n), \triangle_2(n), ..., \triangle_M(n))^T$ bits enter the users' queues. During time slot n, $\vec{u}(n) = (u_1(n), u_2(n), ..., u_M(n))^T$ bits are reliably transmitted and removed from the users' queues. If we denote the number of bits in the users' queues at the beginning of time slot n (after the new bits enter and before transmission begins) as $\vec{Q}(n) = (Q_1(n), Q_2(n), ..., Q_M(n))^T$, then user i's buffer state evolves according to

$$Q_i(n+1) = \triangle_i(n+1) + Q_i(n) - u_i(n).$$

Once a user receives a packet for transmission, the data in that packet enters the transmission queue and a portion of the data from the queue is transmitted according to a certain policy. Note that some bits may undergo unreliable transmission and need to be retransmitted. Thus the queue holds all bits that need reliable transmission.

Packets arrive to each user's transmission buffers according to independent Bernoulli processes. At each time slot, user i has a probability p_i of new packet arrival. Packets are fixed to be L_i. Hence, for each user, the pair (p_i, L_i) characterizes the burstiness of the packet arrivals for user i. In the nth time slot:

$$\triangle_i(n) = \begin{cases} L_i & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i. \end{cases}$$

We assume each user has a buffer of infinite size. Time slots are of length T transmissions. The vector of arrival rates is $\vec{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_M)^T$ where $\lambda_i = \frac{p_i L_i}{T}$ is the arrival rate (in bits per transmission) to the buffer of user i.

2.2 Data Transmission Policies and State Information

At each time slot, user i must either transmit or not transmit over the time slot using codes of length T. A collision occurs if two or more receivers transmit during the same time slot. We assume that the receiver and transmitter have perfect synchronization. The receiver knows for each user, at each time slot, whether or not that user is transmitting. This may done by using coded tags on transmissions to identify users. The code on the tags is sufficient to withstand multiple access interference from all users at once. We assume that by the end of each time slot, each user knows which portion of its transmission data has been reliably received, and which portion needs to be retransmitted. We constrain the set of transmission policies of interest at time n to be a function of $\vec{Q}(n)$. We may view this as a discrete time controlled stochastic system.

2.3 Markovity and Average Bit Delay

We note that since the buffer arrival process is Bernoulli and our transmission policy at time n is a function of $\vec{Q}(n)$, the system of buffer state evolution satisfies the Markov condition:

$$P[\vec{Q}(n+1) = \vec{q} \mid \vec{Q}(n), \vec{Q}(n-1), ..., \vec{Q}(0)] = P[\vec{Q}(n+1) = \vec{q} \mid \vec{Q}(n)].$$

$\vec{Q}$ is the state variable of the homogeneous Markov chain. The transition probabilities of the state evolution are governed by the arrival processes burstiness pairs and the transmission policy.

2.4 Coding Time Slots and Collisions

As in traditional time-slotted ALOHA systems, if two or more users attempt to transmit during the same time slot, a *collision* occurs. However, because of our use of coding, a collision is not necessarily catastrophic: data may still be reliably received in the event of other users transmitting. Time slots are of length T transmissions, where T is long enough in terms of bits so that data may be transmitted with acceptable probability of error even in the event of a collision. For a large but finite T, error exponents [Gal68] for multiple access channels [Lia72] quantify at what rate the probability of error decays exponentially with T. Let n be the number of time slots during which we transmit. Each user has K sets of codewords $\mathcal{M}_j^{i,k}$, $k = 1, ..., K$. During time slot j user i has a codebook C_j^i of length T to encode at most one codeword from each $\mathcal{M}_j^{i,k}$. We denote the codebook C_j^i as (T, ξ, λ_j^i) *time-slot capacity-achieving* if there exists a decoding policy such that $\forall k \in {\mathcal{K}'}_j^i \subset \{1, ..., K_j^i\}$: a codeword from $\mathcal{M}_j^{i,k}$ was encoded, that codeword was decoded with probability of error ξ or less, and $\sum_{k \in {\mathcal{K}'}_j^i} \frac{log(|\mathcal{M}_j^{i,k}|)}{T} \geq \lambda_j^i$. Over the long term, for a given T and we say that a coding and decoding policy is $(T, \xi, \vec{\lambda})$ *capacity-achieving* if

$$\lim_{n \to \infty} \frac{1}{Tn} \sum_{j=1}^{n} E[\sum_{k \in {\mathcal{K}'}_j^i} log(|\mathcal{M}_j^{i,k}|)] = \lambda_i\,, \qquad 1 \leq i \leq M\,,$$

and the probability of error for the transmission of each user on each time-slot is upper bounded by ξ. The notion of capacity described above is related to other delay-constrained and probability-of-failure notions of capacity, such as delay-limited capcity [HT98], ϵ-capacity [VH94], capacity versus outage [Sha97, Sha00, GCB98, LOW98], and expected capacity [EG98]. We consider delay constraints because of finite time slot length. We use expected rates because of uncertainty regarding collisions. The capacity region of our above definition for users with power constraints and certain transmission policy constraints has been found in [MMHG01] to be the same as the Cover-Wyner region for multiple access channels.

3 Minimizing Delay and Minizing Energy

Before we proceed to introduce an optimal coding strategy, we attempt to understand how delay and energy minimization are affected by knowledge of queue information. We restrict our attention to a two-user scenario, but the results may easily be extended for many users.

3.1 Delay Minimization

We now would like to understand what the minimum amount of energy is required to minize delay. Consider the situation where the set of (p_i, L_i) burstiness pairs is known to all users. We consider the case where users have full knowledge of each other's queues, and when they only have local queue information. At the beginning of time slot n,

$$\vec{\Delta}(n) = \begin{cases} (L_1, L_2)^T & \text{with probability } p_1 p_2 \\ (L_1, 0)^T & \text{with probability } p_1(1-p_2) \\ (0, L_2)^T & \text{with probability } (1-p_1)p_2 \\ (0, 0)^T & \text{with probability } (1-p_1)(1-p_2) \end{cases}$$

Let us denote $C(x) = \frac{1}{2}\log(1 + \frac{x}{\sigma_N^2})$ as the capacity (in bits per transmission) of a discrete-time memoryless Gaussian channel with signal noise variance σ_N^2 and average power per transmission constraint x. It is the maximum rate at which information may be transmitted with arbitrarily vanishing error probability. Similarly, $C^{-1}(x) = \sigma_N^2(2^{2x} - 1)$ is the minimum amount of average energy required to transmit rate x with arbitrarily vanishing error probability.

We now comment on the capacity region for the discrete-time two-user AWGN multiple access channel with power constraint $\vec{P} = (P_1, P_2)^T$ and noise variance σ_N^2. It is defined to be the subset of $\mathbb{R}_+^2$ with rate pairs (R_1, R_2) satisfying

$$\sum_{i \in \mathcal{S}} R_i \leq C(\sum_{i \in \mathcal{S}} P_i), \qquad \mathcal{S} \subseteq \{1, 2\}.$$

Let us denote the *dominant face* of the capacity region as the subset of all rate pairs in the capacity region that satisfy $R_1 + R_2 = C(P_1 + P_2)$. For all rate pairs (R_1, R_2) that are not dominant, there exists a (R_{dom1}, R_{dom2}) that satisfies $R_{dom1} \geq R_1$ and $R_{dom2} \geq R_2$. We note that any dominant rate pair delivers the maximum aggregate rate of reliable transmission for a given power constraint $\vec{P}$. Or alternatively, the power vector $\vec{P}$ delivers the minimum aggregate power for two users to transmit reliably at rates on the dominant face of the capacity region.

3.1.1 Full Knowledge of Other Users' Queues

Immediately before time slot n, users have full knowledge of the number of bits that have just entered everyone's queue: $\vec{\triangle}(n)$. To minimize delay to 0, each user must empty the total contents of everyone's queues each time slot of length T transmissions. Every user has access to two codebooks: a multiple-access codebook that may achieve the rate pair $(\frac{L_1}{T}, \frac{L_2}{T})$ reliably, and a single-user codebook that may achieve the rate $\frac{L_i}{T}$ reliably for user i when no multiple access interference is present. We note that the minimum amount of aggregate energy per transmission required is i.i.d. over each time slot n, and is a function of $\vec{\triangle}(n)$. The minimum amount of aggregate energy per transmission required to empty the buffers in one time slot is given by:

$$P_{min}^{(1)}(n) = \begin{cases} C^{-1}(\frac{L_1+L_2}{T}) & \text{with probability } p_1p_2 \\ C^{-1}(\frac{L_1}{T}) & \text{with probability } p_1(1-p_2) \\ C^{-1}(\frac{L_2}{T}) & \text{with probability } (1-p_1)p_2 \\ 0 & \text{with probability } (1-p_1)(1-p_2) \end{cases}$$

Since the arrival processes are independent and Bernoulli, ergodicity holds and we have:

$$\lim_{m\to\infty} \sum_{n=1}^{m} \frac{1}{n} P_{min}^{(1)}(n) \to_{a.s.} p_1p_2C^{-1}\Big(\frac{L_1+L_2}{T}\Big) + p_1(1-p_2)C^{-1}\Big(\frac{L_1}{T}\Big) + (1-p_1)p_2C^{-1}\Big(\frac{L_2}{T}\Big).$$

We note that similar results hold for any set of ergodic arrival processes $\vec{\triangle}$.

3.1.2 No Knowledge of Other Users' Queues

If we now assume that each user still has access to the (p_i, L_i) burstiness pairs of everyone, but does not have access to the amount of data entering the others' queues, then to minimize delay, all users must coordinate to transmit at the worst case scenario: when (L_1, L_2) enters each others' queues. If we re-examine how Hence, user i transmits either no

set their per The minimum amount of aggregate energy to empty the buffers in one time slot is given by:

$$P_{min}^{(2)}(n) = \begin{cases} C^{-1}(\frac{L_1+L_2}{T}) & \text{with probability } 1-(1-p_1)(1-p_2) \\ 0 & \text{with probability } (1-p_1)(1-p_2) \end{cases}$$

Hence, the long term average minimum amount of energy per transmission required is given by

$$\lim_{m\to\infty} \sum_{n=1}^{m} \frac{1}{n} P_{min}^{(2)}(n) \to_{a.s.} \big((1-(1-p_1)(1-p_2)\big)C^{-1}(\frac{L_1+L_2}{T}).$$

3.2 Minimizing Energy Consumption

Concavity of the function $\log(1+x)$ provides the inequality $\frac{1}{2}\log(1+\frac{P}{\sigma_N^2}) \geq 2*\frac{1}{2}\log(1+\frac{\frac{P}{2}}{\sigma_N^2})$. So it is more favorable in terms of aggregate energy consumption to transmit the same amount of energy over multiple time slots rather than in just one use. We may recursively perform the same argument for more than two time-slots. We also know that for multiple users to reliably transmit at a prescribed rate-tuple, using a coding scheme where that rate-tuple lies on the dominant face of the multiple access capacity region minimizes the amount of aggregate power required. We note that in terms of a long-term power utilization, user queue information is not necessary to perform this strategy. If each user artificially backs up its queue by not transmitting, then after a while each user will have very large queue lengths. At that point, each user will have data to transmit. Afterwards, users transmit on the dominant face of a multiple access region determined by their powers. As the vector of output rates tends toward the vector of input rates from above, the amount of energy consumption required decreases, but average delay increses. Since the system must provide stability, the minimum amount of energy required will correspond to when the vector of output rates matches the vector of input rates with equality. For ergodic processes, the proportion of time users spent artificially backing up queues tends to 0. We note, that since each of the server utilizations is exactly 1, and the arrival process is Bernoulli, the expected delay is infinite. The average aggregate amount of energy required is given by

$$\lim_{m\to\infty} \sum_{n=1}^{m} \frac{1}{n} P_{min}^{(3)}(n) \to_{a.s.} C^{-1}(\sum_i \frac{p_i L_i}{T}).$$

Similar results hold for many ergodic arrival processes.

4 Energy and Delay Tradeoffs for a System with Limited Queue Information

We now present a coding scheme that addresses the burstiness of packet arrivals and analyze how it performs. The system operates in two modes: a multiple-access mode when all queue lengths are large, and a broadcast channel mode otherwise (see Figure 2). The scheme utilizes a very small amount of queue information sharing among users.

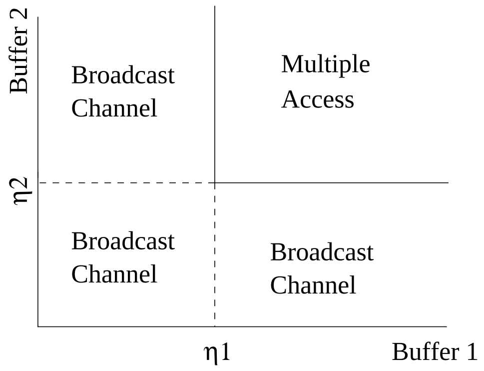

Figure 2: Modes of operation for our proposed scheme

4.1 System Design

Along with the modeling assumptions defined in 2, we limit each user to satisty a per time-slot energy constraint. In each time slot, user j may consume up to P_j units of energy. We assume a limited information sharing scheme where, at time slot n, each user does not know the contents of the newly arrived packets in other users queues. By the end of the time slot, perhaps through feedback from the receiver, each user knows which portion of the data he attempted to transmit was received reliably, and which portion needs to be retransmitted.

4.1.1 Large Queue Lengths: Multiple Access Mode

The results in [MMHG01] show that the capacity region of the time-slotted ALOHA system with power-constrained users is the same as the capacity region of the multiple-access channel. For any vector of arrival rates lying inside the Cover-Wyner region, there exist coding schemes that will provide system stability. As described in [MMHG01], as all users' queues become very backed up, they are able to transmit simultaneously at rate-tuples lying on the dominant face of the multiple access region while sustaining a small upper bound on probability of error. Error exponents we discussed previously provide bounds to the probability of error for a given slot length, T.

For our system Markov chain with state variable $\vec{Q}$, we denote the vector-valued drift of the state to be

$$\vec{D}(\vec{q}) = E[\vec{Q}(n+1) - \vec{Q}(n) \mid \vec{Q}(n) = \vec{q}].$$

In our case, based on our model of the data transmission policies state information in 2.2, $D_i(\vec{q}) = \lambda_i - \mu_i(\vec{q})$, where $\mu_i(\vec{q})$ is a function of $\vec{q}$. If $D_i(\vec{q}) < 0 \ \forall i$, then via Pak's Lemma [Pak69], the chain is ergodic, and steady-state probabilities exist. Hence, a sufficient condition for stability of our system model is to is to eventually transmit at multiple access after the queue states cross a finite threshold $\vec{\eta} = (\eta_1, \eta_2, ..., \eta_M)^T$. Thus, given a set of burstiness pairs and per transmission power constraints, as all users' queues states become backed up (cross this threshold $\vec{\eta}$), they transmit data out of the queues at rates $\vec{\mu}_{ma} = (\mu_{ma1}, \mu_{ma2}, ..., \mu_{maM})^T = E[\vec{\mu} \mid \vec{Q} \geq \vec{\eta}]$. To transmit optimally aggregate data subject to the power constraints, $\vec{\mu}_{ma}$ must lay on the dominant face of the multiple access region: $(\sum_j \mu_{maj} = C(\sum_j P_j))$. To ensure stability that operating point should provides negative drift for all queues ($\lambda_j - \mu_{maj} < 0 \ \forall j$). We note that the particular threshold is given by $\eta_j = T\mu_{maj}$, where T is the time slot length in transmissions.

4.1.2 Small Queue Lengths: Broadcast Mode

When all users' queues lengths are not above the threshold $\vec{\eta}$, any rate tuple $\vec{\lambda}$ lying on the dominant face of the multiple access region is not $(T, \xi, \vec{\lambda})$ time-slot achievable. Let us restrict our attention to two users, but note that the ideas may naturally be extended. Capacity on the degraded AWGN broadcast channel is achieved by rate-splitting (where a user superimposes two independent virtual-user codes) at the transmitter and successive decoding (signals are iteratively decoded and subtracted out for future decoding). One (low-resolution) signal is coded to be received reliably by both users. The other (high-resolution) signal is coded to be received reliably by the receiver with the stronger SNR. The receiver decodes a virtual user, eliminates its contribution, and then decodes the next.

In the small queue length regime of our system, we use this broadcast idea to combat burstiness: each user splits into virtual users which code anticipating different users' presence that time slot. Even in the event of a collision, data is reliably received from all users. Each user has a simple *deterministic* transmission policy: if a user has data in its queue to transmit, it attempts to do so. The only time a collision does not take place is when one of the two users' queues is empty. Each user codes to transmit over two possibe channels: a channel with the other user present, and a channel without the other user. A fraction α of user i's power P_i, is allocated to a virtual low-resolution user that codes anticipating not only the presence of the virtual user counterpart for user i, but also the other physical user's presence. The high-resolution virtual user for user i does not anticipate the other physical users' presence: it is only received reliably when user $j \neq i$ does not transmit. The rates at which each virtual user attempts to transmit reliably via successive decoding are as follows:
$\mu_{LR1} = \frac{1}{2}log(1 + \frac{\alpha_1 P_1}{\sigma_N^2 + P_2 + (1-\alpha_1)P_1})$
$\mu_{HR1} = \frac{1}{2}log(1 + \frac{(1-\alpha_1)P_1}{\sigma_N^2})$
$\mu_{LR2} = \frac{1}{2}log(1 + \frac{\alpha_2 P_2}{\sigma_N^2 + P_1 + (1-\alpha_2)P_2})$
$\mu_{HR2} = \frac{1}{2}log(1 + \frac{(1-\alpha_2)P_2}{\sigma_N^2})$.
In each time slot, a collision occurs when more than one user transmits. If a collision occurs, only the low-resolution component of each user is reliably received. Otherwise, the low and high-resolution components of the sole transmitting user are reliably received. Hence, when the system is in this mode, the reliably received rate pair is as follows:

$$\vec{\mu}_{BC} = \begin{cases} (\mu_{LR1}, \mu_{LR2})^T & \text{if a collision occurs} \\ (\mu_{LR1} + \mu_{HR1}, 0)^T & \text{if only user 1 transmits} \\ (0, \mu_{LR2} + \mu_{HR2})^T & \text{if only user 2 transmits} \end{cases}$$

4.2 Delay

We note that for any set of burstiness pairs with corresponding rate-tuple lying inside the Cover-Wyner region for our power constraints, the Markov chain is ergodic. Let us consider the two-user scenario and note how the analysis may easily extended for more users. The steady-state probabilities $\pi_{\vec{q}} = \lim_{n \to \infty} P[\vec{Q}(n) = \vec{q}]$ for the Markov chain are governed by:

- The per-transmission power constraints P_i for each user,
- The burstiness pairs (p_i, L_i) of each user, and

- The rate-splitting power ratio α_i for each user

We note that the long-term average queue size $N(\vec{P}, \vec{p}, \vec{L}, \vec{\alpha}) = \sum_{\vec{q} \in \mathbb{R}_+^M} \sum q_i \pi(\vec{P}, \vec{p}, \vec{L}, \vec{\alpha})$ may be used to calculate the long-term average bit delay $\overline{T}(\vec{P}, \vec{p}, \vec{L}, \vec{\alpha})$ via Little's Result. We choose the pair power spliting ratios $\vec{\alpha}$ to minimize the long-term average bit delay,

$$\vec{\alpha}^* = arg \min_{\alpha \in [0,1]^M} \overline{T}(\vec{P}, \vec{p}, \vec{L}, \vec{\alpha}) = \frac{1}{\lambda_1 + \lambda_2} arg \min_{\alpha \in [0,1]^M} N(\vec{P}, \vec{p}, \vec{L}, \vec{\alpha}).$$

Since the chain is ergodic for all arrival rates inside the multiple access region, we may truncate the state space and perform an approximation [Fre71] using simulations on a state space of finite number of states. For each set of burstiness pairs and a given power constraint, we perform a discretization over the possible splitting ratios and select $\vec{\alpha}$ that minimizes delay. We note that the delay and choice of $\vec{\alpha}$ is not just a function of the vector of arrival rates, but a function of the individual burstiness pairs.

4.3 Queue Information Sharing

We note that users have different sets of codebooks for which they transmit information, a set of codebooks for when they transmit in multiple access mode, and a set of codebooks for when they transmit in broadcast mode. Users notify each other when their queue state crosses the threshold η_i. Hence, each user has total knowledge of a synchronized finite-state automaton that denotes whether or not each user's queue length has crossed η_i. When the FSA is in the state where all users thresholds have are greater than η_i, each user switches to multiple access mode. Otherwise, they operate in their broadcast channel mode. We do not model the communication link between users for this communication, but note that it is not a substantial amount of information that is shared.

References

[Ber00] R. A. Berry. *Power and Delay Trade-offs in Fading Channels.* PhD dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, June 2000.

[Cap78] J. I. Capetanakis. The multiple access broadcast channel: Protocol and capacity considerations. *MIT Tech. Rep ESL-R-806*, 1978.

[Cov72] T. M. Cover. Broadcast channels. *IEEE Transactions on Information Theory*, 18:2–14, 1972.

[EG98] M. Effros and A. Goldsmith. Capacity definitions and coding strategies for general channels with receiver side information. *Proceedings of ISIT*, page 39, 1998.

[Fre71] D. Freedman. *Approximating Countable State Markov Chains.* Holden-Day, San Francisco, CA, 1971.

[Gal68] R. Gallager. *Information Theory and Reliable Communication.* John Wiley & Sons, New York, NY, 1968.

[GCB98] G. Taricco G. Caire and E. Biglieri. Minimum outage probability for slowly-fading channels. *Proceedings of the International Symposium on Information Theory*, page 7, 1998.

[GVS88] S. Ghez, S. Verd'u, and S. Schwartz. Stability properties of slotted ALOHA with multipacket reception capability. *IEEE Transactions on Automatic Control*, 33:640–649, 1988.

[Hay78] J. Hayes. An adaptive technique for local distribution. *IEEE Transactions on Communications*, 26:1178–1186, 1978.

[HT98] S.V. Hanly and D.N.C. Tse. Multiaccess fading channels. ii. delay-limited capacities. *IEEE Transactions on Information Theory*, 44(7):2816 – 2831, 1998.

[Lia72] H. Liau. *Multiple Access Channels*. PhD dissertation, University of Hawaii, Department of Electrical Engineering and Computer Science, June 1972.

[LOW98] S. Shamai L.H. Ozarow and A.D. Wyner. Information theoretic considerations for cellular mobile radio. *IEEE Transactions on Vehicular Technology*, 43(2):359–378, 1998.

[MMHG01] M. Médard, S. P. Meyn, J. Huang, and A. J. Goldsmith. Capacity of time-slotted ALOHA packetized multiple-access systems. 2001.

[Pak69] A. G. Pakes. Some conditions for ergodicity and recurrence of Markov chains. *Oper. Res.*, 17:1059–1061, 1969.

[Riv87] R. Rivest. Network control by Bayesian broadcast. *IEEE Transactions on Information Theory*, 33:323–328, 1987.

[Sha97] S. Shamai. A broadcast strategy for the Gaussian slowly fading channel. *International Symposium on Information Theory*, page 150, 1997.

[Sha00] S. Shamai. A broadcast strategy for the multiple-access slowly fading channel. *Proceedings of the International Symposium on Information Theory*, page 128, 2000.

[VH94] S. Verdú and T. S. Han. A general formula for channel capacity. *IEEE Transactions on Information Theory*, 40(7), 1994.