Trade-off Between Power Consumption and Delay in Wireless Packetized Systems

by

Todd P Coleman

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

April 2002

Author .....
Department of Electrical Engineering and Computer Science
April 29, 2002

Certified by .....
Muriel Médard
Assistant Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Trade-off Between Power Consumption and Delay in Wireless Packetized Systems

by

Todd P Coleman

Submitted to the Department of Electrical Engineering and Computer Science
on April 29, 2002, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering

Abstract

In packetized wireless systems, coding allows reliable transmission of multiple packets colliding at a receiver. Thus data may not need to incur delays such as those due to back-off schemes in traditional ALOHA systems. However, there is a trade-off between delay and power 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 power consumption when average delay, in terms of slots, is minimized, with and without knowledge of other users' queues. We also consider the case where average power is minimized without regard for delay. We present and analyze a simple scheme with limited information sharing about queues' states. Our scheme uses a hybrid multiple access/broadcast-type code for the case of low queue lengths and a multiple access scheme in the case of large queue lengths. We show how this scheme allows trade-offs between power consumption and delay.

Thesis Supervisor: Muriel Médard
Title: Assistant Professor

Acknowledgments

My advisor, Professor Muriel Médard for her guidance and patience in my development as a graduate student. I truly appreciate her genuine and sincere interest in my well-being as not only a student, but also as a person. She is definitely a leader, and not only with her words, but also with her actions. Her ability to operate at such a high level of intensity and breadth, always with a smiling face and kind demeanor, constantly reminds me that without sacrificing one's happiness, the human being still has unbounded capabilities and potential.

I would also like to thank my parents for their love and support. I especially appreciate my mother's sense of humor and ability to take everything, including the good and the bad, in stride. My father's wisdom and strong sense of work ethic have particularly helped me make the most of my life while I have been here.

Contents

List of Figures

Chapter 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 power 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.

Time-slotted ALOHA [Abr70] is a protocol for 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 back-off policy.

Modelling a collision as leading to loss of all packets at receivers is not always necessary. 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 use the types of codes that achieve rates inside the information theoretic multiple access capacity region [Ahl71, Lia72].

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 [MMH+02]. 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' queues, whether in a decentralized fashion or through 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 power consumption, was not illustrated in [MMH+02]. Clearly, queue information is not altogether useless. While it may not affect capacity, 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 power 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, we characterize the impact of queue information sharing in the presence of burstiness. We also consider minimizing power 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 combating burstiness by performing probabilistic back-off policies after collisions, the system uses rate-splitting to achieve variable reliably received rates as a function of the uncertainty of other 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).

Chapter 2 provides an overview of different notions of capacity for single and multiple users. Chapter 3 discusses the multi-user time-slotted ALOHA protocol. It addresses conflict-resolution issues and recent variations that provide multiple-user reception capabilities. Chapter 4 provides the model for communication scenarios we are interested in. The burstiness model for the source, and the properties of time slots related to coding and collisions are addressed. Chapter 5 investigates power minimization and delay minimization strategies, and illustrates trade-offs amongst them. Chapter 6 analyzes a system with limited queue information and evaluates its performance. Further issues and suggestions are also addressed.

Chapter 2

Notions of Capacity

In this thesis, we are interested in understanding how multiple users may reliably communicate and obtain desirable qualities of service. Various notions of channel capacity have been discussed to answer questions about the upper limits of reliable communication. Depending on the type of system, constraints, and notion of rate, different notions may be more attractive than others. We discuss a few of them below.

2.1 Single-User Channel Models

A channel is a probabilistic mapping from a set of input messages to a set of output messages. We will focus our attention here on discrete-time channels. A random variable will be denoted with capital letters (such as X), and a sample value will be denoted with lower-case letters (such as x). The discrete-time, discrete-valued input, discrete-valued output channel

$$W = \{W^n = P_{Y^n|X^n}\left(y^n|x^n\right) : \mathcal{X}^n \to \mathcal{Y}^n\}_{n=1}^{\infty}$$

is characterized by the sequence of n-dimensional conditional probability distributions $(P_{Y^n|X^n}\left(y^n|x^n\right))$, the input alphabet $(\mathcal{X})$ and the output alphabet $(\mathcal{Y})$. A similar definition holds for discrete-time, continuous-valued input/output channels, where a conditional probability distribution acts on messages from a continuous input

alphabet and continuous output alphabet. Output sequences are used to recover the transmitted message from the input sequence. Uncontrollable noise and imperfections in signalling cause the output to be a corrupted form of the input. The transmitter and receiver utilize error-correcting codes that try to combat the uncertainty in the channel to afford reliable communication.

2.1.1 Memoryless Channels

Memoryless channels are characterized by mutually independent uses of the channel. The n-dimensional conditional probability distributions for such a channel is

$$P_{Y^n|X^n}\left(y^n|x^n\right) = \prod_{i=1}^{n} P_{Y|X}\left(y_i|x_i\right).$$

Since each use of the channel is an independent, identically distributed mapping of an input to an output, laws of large numbers and ergodic theory may be used to analyze n uses of the channel as n becomes large.

2.2 Channel coding and decoding

Channel coding is a way to introduce redundancy into the transmission of information so that transmitted messages will survive channel corruptions. We pay most of our attention here to block codes, where one of M messages is transmitted in n transmissions. Hence, a block code consists of:

i a set of possible input messages: $\mathcal{A} = \{a_1, a_2, ..., a_M\}$

ii an encoding function that maps each input message to a codeword of length n, $x : \mathcal{A} \to \mathcal{X}^n$. Each length-$n$ codeword, $x^n(a_i)$ may need to satisfy some sort of constraint.

iii a decoding function that maps each possible output sequence of the channel to one of the possible input messages, $g : \mathcal{Y}^n \to \mathcal{A}$.

The rate of the code, $R = \frac{\log_2 M}{n}$, defines the time-average number of information bits communicated per transmission. The performance criterion of a code is in general its probability of decoding error for a message, which is a function of the probability law of the channel, the probability distribution of the message set, and how encoding and decoding are performed.

An (n, M, ϵ) block code for a given channel and message set probability distribution is defined to have the following properties:

- $|\mathcal{A}| = M$
- codeword length of n
- the average probability of error across all codewords, denoted as $P_e^{(n)}$, is upper-bounded by ϵ.

A rate $R = \frac{\log_2 M}{n} \geq 0$ is ϵ-achievable if for sufficiently large n and every $\delta > 0$, there exist (n, M, ϵ) codes with rate

$$\frac{\log_2 M}{n} > R - \delta.$$

Maximum Likelihood (ML) decoding for Block Codes on Memoryless Channels

Minimum probability of error rules for codes minimize the probability of error given a particular channel output. For a memoryless channel with output y^n,

$$P_{Y^n|X^n}\left(y^n|x^n(a_i)\right) = \prod_{j=1}^{n} P_{Y|X}\left(y_j|x_j(a_i)\right)$$

Let us denote the a priori probability of message a_i being transmitted as $P(a_i)$. Then we have

$$P(a_i|y^n) = \frac{P_{Y^n|X^n}\left(y^n|x^n(a_i)\right) P(a_i)}{P_{Y^n}\left(y^n\right)}$$

A minimum probability rule chooses a_i as the decoded message if and only if

$$P(a_i|y^n) \geq P(a_j|y^n) \; \forall \; j \neq i$$

A maximum-likelihood (ML) decoding rule assigns the message a_i to be transmitted if and only if

$$P_{Y^n}(y^n|a_i) \geq P_{Y^n}(y^n|a_j) \ \forall \ j \neq i$$

We note that ML-decoding is in fact a minimum probability of error rule if the input messages are equally likely.

2.3 Coding theorems

For a large class of channels, the techniques used to prove the existence of codes that achieve capacity involve random coding arguments. Rather than trying to look for arbitrarily complex channel codes that have high rates and arbitrarily small probability of error, the techniques involve averaging over all the ensembles of codes, whose individual code symbols are chosen independently according to a probability distribution $P_X(x)$. If the ensemble average probability of error across all length-n codes, denoted as $\overline{P_e^{(n)}}$, tends to 0 with n, then one such code must have well-performing $P_e^{(n)}$ as well.

2.4 Notions of Capacity

2.4.1 Shannon capacity

The classical Shannon capacity addresses the supremum of achievable rates for memoryless channels. A rate R is defined to be achievable if it is ϵ-achievable for all $\epsilon > 0$.

Let us consider a given memoryless channel W encoding one of $M = 2^{nR}$ messages using a length-n code with individual code symbols drawn independently according to $P_X(x)$. Then we may note that by using laws of large number arguments and performing typicality decoding, the ensemble average probability of error satisfies

$$\overline{P_e^{(n)}} \ \leq \ 2^{n(R-I(X;Y))}$$

Consequently, any rate $R < I(X;Y)$ is achievable. The channel capacity in a memoryless channel has been found to be:

$$C \triangleq \sup_{P_X(x)} I(X;Y)$$

The converse to the theorem is proven on the basis of Fano's inequality, which states that for an index m chosen uniformly on the set $\mathcal{M} = \{1, ..., 2^{nR}\}$, and a codebook producing $X^n = \{X_1, ..., X_n\}$, then the message error probability, $P_e^{(n)} = P(m \neq g(\underline{Y}))$, satisfies

$$H(X^n|Y^n) \leq H\left(P_e^{(n)}\right) + P_e^{(n)} nR.$$

This inequality can be manipulated for a variety of channels to illustrate that the probability of error is in fact lower-bounded away from 0 for any rate $R > C$. For additive memoryless Gaussian noise channels with noise variance σ_N^2 and power constraint x, the capacity is found to be

$$C_{\sigma_N^2}(x) \triangleq \frac{1}{2}\log_2\left(1 + \frac{x}{\sigma_N^2}\right).$$

The proof also uses random coding arguments here, where codewords are generated according to Gaussian random variables, due to the property that the Gaussian random variable has the highest entropy of all random variables of a particular variance.

2.4.2 Error Exponents

Error exponents quantify how the increase in block length n is related to the ensemble codeword average probability of error. For a large class of channels, there is an intrinsic tradeoff between upper bounds on the probability of error and rate of an (n, M, ϵ) code. To be more precise, as illustrated in [Gal68], let us consider memoryless channels with message encoding performed by generating independent code symbols according to $P_X(x)$ and decoding performed using maximum likelihood decoding.

Then the ensemble average probability of error across all codebooks satisfies:

$$\overline{P_e^{(n)}} \leq \exp\left[-nE_r(R)\right],$$

where n is the block length, R is the codebook rate, and $E_r(R)$ is the error exponent. For any achievable rate $R < I(X;Y)$, $E_r(R) > 0$ and is defined as

$$\begin{aligned} E_r(R) &= \max_{0\leq\rho\leq 1} \max_{P_X(x)} E_0\left(\rho, P_X(x)\right) - \rho R \\ E_0\left(\rho, P_X(x)\right) &= -\ln \sum_{y=0}^{|\mathcal{Y}|-1} \left[\sum_{x=0}^{|\mathcal{X}|-1} P_X(x)\, P_{Y|X}(y|x)^{\frac{1}{1+\rho}} \right]^{1+\rho}. \end{aligned}$$

Hence, the probability of error may be made arbitrarily small for any achievable rate. Furthermore, $E_r(R)$ is a decreasing function for increasing R. Hence, for two different achievable rates, it will take a longer block length to guarantee the same upper bound on probability of error for the larger rate. This illustrates an intrinsic trade-off between delay and probability of error for any achievable rate.

2.5 General Notions of Capacity

Not all channels exhibit memoryless, stationary, or causal properties. A more general form of capacity that does not require any of the properties aforementioned to be valid has been defined in [VH94]. It relies instead on the probability density function of the sequence of normalized information random variables $\frac{1}{n} i_{X^n;Y^n}(a;b)$, where

$$i_{X^n;Y^n}(x;y) = \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)}$$

is defined as the sample mutual information.

2.5.1 Generalized Capacity

A rate R is defined to be achievable in a more general sense if the limit of cumulative distribution functions,

$$\left\{ F_n(\alpha) = P\left(\frac{1}{n} i_{X^n;Y^n}\left(x^n; y^n\right) \leq \alpha \right) \right\}_{n=1}^{\infty}$$

evaluated at $\alpha = R$, tends to 0 as $n \to \infty$. The capacity formula provided in [VH94] is defined as

$$C = \sup_{P_X(x)} \underline{I}\left(X;Y\right)$$

where $\underline{I}\left(X;Y\right)$ is defined to be *inf-information rate* between X and Y. This is the *liminf in probability* of the sequence of normalized information densities $\frac{1}{n} i_{X^n;Y^n}\left(x^n; y^n\right)$. Note that if A_n is a sequence of random variables, then its *liminf in probability* is the supremum of all reals α such that $P[A_n \leq \alpha]$ tends to 0 as n tends to infinity.

2.6 Capacity Definitions in the Context of Compound Channels

We now turn to channels that are time-varying in nature. We understand them by considering a compound channel [Wol78], $\{\Gamma(\theta) : \theta \in \Theta\}$ where the receiver knows the state θ but the transmitter does not. Each realization $\Gamma(\theta)$ is a memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$.

2.6.1 Delay-Limited Capacity

In certain systems, quality of service demands, or system needs require that the transmission of data may constrained by delay. Strict delay constraints require that a given codeword be transmitted after K blocks. We may assume that each block consists of n uses of the channel, where n is large enough to achieve rates near the capacity with sustainable probability of error. Average delay constraints are a more relaxed constraint - they require that the long-term time average delay be less than

or equal to some threshold and are useful in addressing systems that do not achieve the same rate over every transmission.

Delay-limited capacity is useful in analyzing systems with various channel realizations and stringent delay constraints. Let us consider a compound channel, $\{\Gamma(\theta) : \theta \in \Theta\}$, where θ determines the transition probabilities of the channel. Suppose the user is required at a rate of at least R bits/sec in K blocks of n transmissions, regardless of the realization of the channel. Over each block i of n transmissions, it is assumed that the channel is in some state $\Gamma(\theta_i)$. The capacity of the compound channel, for a the set of all possible input probability distributions Π and the set of all possible channel states Θ is given by [Wol78]:

$$C = \sup_{P_X(x)\in\Pi} \inf_{\theta\in\Theta} I\left(X;Y\ |\theta\right)$$

In the delay-limited [HT98] case, if we take a look at the memoryless Gaussian channel with power constraint P, we note that the input Gaussian distribution always maximizes mutual information. For a channel realization and the corresponding delay-limited capacity corresponds to the infimum of sum capacities of the compound channel:

$$C_{DL} = \inf\{C_{\underline{\theta}} : \underline{\theta} \in \Theta_K\}$$

where $\Theta_K \in \mathbf{C}^K$ is the set of all of length-K sequences of channel states that occur with nonzero probability and $C_{\underline{\theta}} = \sum_{i=1}^{k} C_{\sigma_N^2(\theta_i)}\left(P\right)$. It is interesting to note that this notion of capacity assumes that *the user's average mutual information be kept constant in time*. In terms of narrow-band block fading channels, C_{DL} is the capacity of the channel corresponding to the worst sequence of K fades in the block fading channel. It is based on the assumption that the user requires a *fixed* rate of R in each of a group of K channel blocks.

2.6.2 Capacity vs. Outage Probability

Capacity vs. Outage Probability [GCB98] brings into account the *a priori* probabilities, $P_\Theta(\theta)$ of each of the channel realizations $\Gamma(\theta)$. If one were to observe the probability of error, averaged over all channel realizations and codewords,

$$\overline{P_e^{(n)}} = \int_\Theta \overline{P_e^{(n)}}(\theta) d\pi_\Theta(\theta)$$

We next observe the following:

$$\overline{P_e^{(n)}} \leq \int_{\theta:R<C_\theta} e^{-nE_r(R,\theta)} d\pi_\Theta(\theta) + \int_{\theta:R\geq C_\theta} 1 d\pi_\Theta(\theta)$$

The intuition behind quantifying outage probabilities is based upon the error exponents from [Gal68]. So we may say that for a given outage probability q, the capacity vs outage probability q, C_q is defined as

$$C_q = \sup\{R : P(C_\theta \leq R) \leq q\}$$

This notion of capacity is basically the maximum mutual information rate that may be sent over any channel $\Gamma(\theta)$ except a subset with probability less than q.

2.6.3 Expected Capacity

In many situations, it is not necessary to transmit at a constant rate per codeword. Instead of an outage when the channel is bad, perhaps fewer bits may be received reliably. The introduction of expected capacity [EG98] quantifies this notion. We noted earlier that a rate R is ϵ-achievable if there exists a sequence of channel coding schemes such that the expected decoded rate is R and the average probability of error tends to 0. The expected capacity is the supremum of all such rates. Note that, when using compound channels, the average capacity arises from the encoder taking a broadcast channel approach for each of the channels in the collection. Decoding is performed as a function of the state of the channel. Expected capacity in this sense is

is the maximum of $\int_\Theta R(\theta) d\pi_\Theta(\theta)$ averaged over all rates $R(\theta)$. Analyzing systems in terms of average capacity becomes more reasonable when users may achieve variable reliably received rates, and a long-term average rate is of interest. It is also interesting to note that, in order to achieve rates near this notion of capacity, rather than making n large as in the delay-limited case, it is important that K be large so that enough independent channel realizations occur for laws of large numbers to apply.

2.7 Multiple-User Channel Models

For more general situations, multiple users share the same medium over which they communicate with each other. In these situations, we may model the channel as a probabilistic mapping from multiple inputs to multiple outputs. As we shall see, a recurring idea that stems from these results is that it is better to have all users access the medium simultaneously rather than split the medium into pieces and assign each user a portion of it. We take a look at two corner cases: the multiple input, single output (multiple access) channel, and the single input, multiple output (broadcast) channel.

2.7.1 The Multiple Access Channel

The memoryless multiple access channel for M transmitters and one receiver consists of $M+1$ alphabets: $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_M, \mathcal{Y}$, and a conditional probability distribution $P_{Y|X_1,X_2}(y|x_1,x_2)$. Let us restrict our attention to two users. The capacity region, also known as the Cover-Wyner region, has been found [Lia72, Ahl71] to be the closure of the convex hull of the set of all rate tuples $(R_1, R_2, ..., R_M) \in \mathbb{R}_+^M$ satisfying:

$$\sum_{i \in \mathcal{S}} R_i \leq I\left(X(\mathcal{S}); Y | X(\mathcal{S}^\mathcal{C})\right) \qquad \mathcal{S} \subseteq \{1, 2, ..., M\}.$$

where $X(\mathcal{S}) = \{X_i : i \in \mathcal{S}\}$. Achievability of any rate-tuple within this region relies on time-sharing, where any rate for any two achievable rate-tuples $\underline{R}$ and $\underline{R}'$, the rate tuple $\lambda\underline{R} + (1-\lambda)\underline{R}', 0 \leq \lambda \leq 1$ is achievable by using the first codebook for λn symbols

and using the second codebook for $(1-\lambda)n$. Certain corner points have intuitive interpretations. For instance, when $M = 2$, the rate pair $(I(X_1;Y|X_2), I(X_2;Y))$ may be achieved by the receiver first treating X_1 as noise, and then reliably decoding X_2 for any $R_2 \leq I(X_2;Y)$. After this has been done, the receiver may eliminate the presence of X_1 in Y, and reliably decode X_1 for any $R_1 \leq I(X_1;Y|X_2)$. Similarly, by reversing the roles of X_1 and X_2, we see that the pair $(I(X_1;Y), I(X_2;Y|X_1))$ also lies in the capacity region. We note that by using time-sharing, any convex combination of the corner points may be achieved as well. It is important to note that orthogonal access schemes, such as time division multiple access (TDMA) and frequency division multiple access (FDMA) lie strictly within the aforementioned region and are in general suboptimal strategies. In the case of equal-rate equal-power users, it has been shown [Gal85] that FDMA lies on the boundary of the Cover-Wyner region.

Let us now consider the capacity region for the memoryless discrete-time two-user Gaussian multiple access channel with power constraints $\underline{P} = (P_1, P_2)$ 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

$$\sum_{i\in\mathcal{S}} R_i \leq C_{\sigma_N^2}\left(\sum_{i\in\mathcal{S}} P_i\right), \qquad \mathcal{S} \subseteq \{1,2\}.$$

The Dominant Face of the Cover-Wyner region

Let us denote the *dominant face* of the memoryless Gaussian multiple access capacity region as the subset of all rate pairs in the capacity region that satisfy $R_1 + R_2 = C_{\sigma_N^2}(P_1 + P_2)$. Figure 2-1 shows an illustration of this capacity region and its associated dominant face. 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 $\underline{P}$. Or alternatively, the power vector $\underline{P}$ delivers the minimum aggregate power for two users to transmit reliably at rates on the dominant face of the capacity region. The two endpoints of the dominant face may be achieved by the receiver considering one of them as noise, decoding the other reliably, then elim-

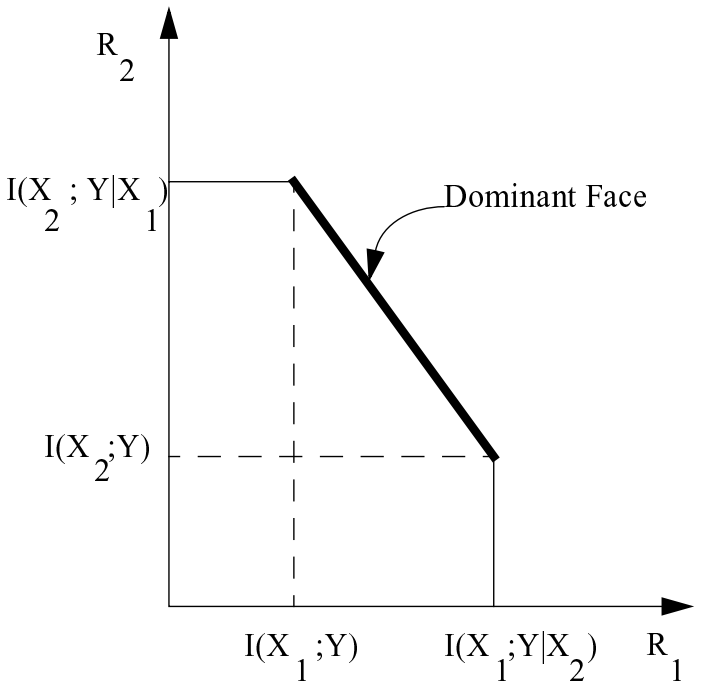

Figure 2-1: Capacity region for the two-user Gaussian multiple access channel and its associated dominant face

inating its presence to decode the first reliably. This technique allows any point on the dominant face to be achieved reliably by time-sharing between those two boundary points. It has also been found recently in [RU96] that performing rate-splitting (whereby one user splits its power and rate into virtual users who contain codebooks that appear to be noise to each other) may achieve any point in the M-user capacity region, using no more than two virtual users per physical user, with a maximum of $2M-1$ virtual users. For each user, the virtual users split power and rate to code for a single-user point-to-point channel with different signal to noise ratios that account for other users' presence. We may consider that the first user treats the second as noise. After successful decoding, the receiver may perform interference cancellation so that the second user may be decoded only in the presence of noise.

2.7.2 Broadcast Channel

The broadcast channel models the problem of a single source transmitting information simultaneously to a number of receivers. The memoryless 2-user broadcast channel consists of an input alphabet $\mathcal{X}$, two output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$, and the one-step probability transition matrix

$$P_{Y_1^n,Y_2^n|X^n}\left(y_1^n, y_2^n|x^n\right) = \prod_{i=1}^{n} P_{Y_1,Y_2|X}\left(y_{1,i}, y_{2,i}|x_i\right).$$

Suppose the channel from the transmitter to receiver 1 has capacity C_1, and the channel from transmitter to receiver 2 has capacity C_2. Without loss of generality, let us assume that $C_2 \geq C_1$. Previous achievable rate regions were found by thinking of:

- Time-sharing approaches, where a proportion λ_1 of the time was allocated to transmitting at rate C_1 to both receivers, and $\lambda_2 = 1 - \lambda_1$ of the time was allocated to transmitting at rate C_2 to receiver 2. In this case, achievable rates are given by:

  $$R_i = \sum_{j \leq i} C_j, i = 1, 2.$$

- Maximin approaches, where either both rates are reliably received by transmitting at rate

  $$R_1 = R_2 = C_{min} = min\{C_1, C_2\} = C_1$$

  where transmission rates are limited by the worst channel, or at the other extreme, by transmitting at the maximum capacity so that

  $$R_1 = 0, R_2 = C_2.$$

Achievable rate regions were found in [Cov72] that strictly dominate the previously described regions by performing *rate-splitting*, where high-rate information is simultaneously superimposed with low-rate information.

2.7.3 Degraded Broadcast Channels

A memoryless broadcast channel is said to be *stochastically degraded* if there exists a distribution $P_{Y_2|Y_1}(y_2|y_1)$ such that

$$P_{Y_2|X}(y_2|x) = \sum_{y_1} P_{Y_1|X}(y_1|x)\, P_{Y_2|Y_1}(y_2|y_1)$$

For such channels, the rate-splitting approach described above has been proven with a converse in [Gal74] to provides the capacity region. Hence, the capacity region for the degraded broadcast channel for $\{X,Y_1,Y_2\}$ is the convex hull of the closure of all (R_1, R_2) satisfying

$$R_2 \leq I(U;Y_2)$$
$$R_1 \leq I(X;Y_1|U)$$

where U serves the purpose of an auxiliary random variable. Let us denote the capacity of a memoryless Gaussian channel with signal noise variance σ_N^2 and power constraint P to be $C_{\sigma_N^2}(P) = \frac{1}{2}\log_2(1 + \frac{P}{\sigma_N^2})$ The capacity region for the Gaussian broadcast channel, with signal power constraint P, channel noise variances N_1 and N_2 where $N_2 < N_1$ has been shown with a converse in [Ber74] to be the set of all (R_1, R_2) satisfying:

$$R_1 \leq C_{(1-\alpha)P+N_1}(\alpha P)$$
$$R_2 \leq C_{N_2}((1-\alpha)P)$$

Again, we may think that the first user encodes thinking the second is noise. After successful decoding, the correct signal may be 'subtracted off' and the second user may decoded by being only in the presence of noise.

Chapter 3

Time-Slotted ALOHA

Multiple-access communication systems where the average time between packet arrivals from a single user is much larger than the time needed to transmit a single packet are said to exhibit *burstiness*. Packetized wireless networks that carry internet information are widely modelled as such. One might imagine that multi-user multiplexing methods, such as time-division multiple access (TDMA) and frequency division multiple access (FDMA), might serve as likely candidates. However, these methods have numerous disadvantages when addressing bursty multiple-user systems: they require large amounts of coordination. The traditional ALOHA system introduces a simple possible solution to the multiplexing problem: each user transmits packets over the common channel in a random-access manner.

3.1 Traditional Pure ALOHA

In traditional pure ALOHA, a large number of users attempt to use the same channel without any synchronization. The times at which users are allowed to use the channel are also unconstrained. The traditional model for users is that bits arrive according to a Poisson process of rate λ. It is assumed that the instant a new packet arrival occurs, a transmitter suddenly appears and attempts to transmit the data instantaneously. What is of interest in such a model is the *throughput*, the fraction of time that the channel is successfully transmitting information. If we denote the arrival rate to the

process as λ, and the packet transmission duration as τ, then the normalized channel traffic, G, is defined as $G = \lambda\tau < 1$. A *collision* occurs when two or more packets overlap, and no packets are successfully transmitted. If we denote the event of a collision as $\{C\}$, then we may define $\lambda' = P(C^c)\lambda$ as the rate of occurrence of packets that are received correctly. The normalized channel throughput, $S = \lambda'\tau$, may be found by considering what must occur for a packet to be successfully transmitted:

- No transmission attempt occurs within the interval of $(t - T, t)$
- No transmission attempt occurs during the interval $(t, t + T)$

Since the arrival process is Poisson of rate λ, the probability of success, which we call the throughput, is given by $P(N(t - T, t + T) = 0) = e^{-2\lambda T}$. Hence, the channel throughput is given by $S = Ge^{-G}$, and is maximized to the *capacity* of $\frac{1}{2e}$ when $G = 0.5$.

3.2 Slotted ALOHA

One of the problems with pure ALOHA is that collisions may occur due to packets barely overlapping. In *time-slotted* ALOHA, a large number of users attempt to use the same channel with a common clock. Hence, the possible times at which users may attempt to transmit data are synchronized. Each possible time interval is termed a *slot*, and is the time of transmission of one packet. An advantage of using this is that when a collision occurs, the packet transmission times overlap exactly. In this *immediate first transmission* (IFT) scheme, a user broadcasts a new packet in the next slot. The times at which users attempt to transmit packets form a discrete-time stochastic process. If we define $G_i < 1$ as the probability that the ith user will transmit a packet in some slot, then we may define the aggregate channel arrival rate as

$$G = \sum_{i=1} nG_i.$$

Although G may be greater than 1, the throughput ($S \leq G < 1$) of the channel is given by

$$S = \sum_{i=1} nS_i.$$

where S_i is the probability that user i sends the only packet in its slot. Hence, for the slotted ALOHA model with n independent users, the probability of no collision for user i is given by

$$\begin{aligned} P(C_i^c) &= \prod_{j=1, j\neq i}^{m} (1 - G_j) \\ S_i &= G_i P(C_i^c) = G_i \prod_{j=1, j\neq i}^{m} (1 - G_j) \end{aligned}$$

In the event that all users have identical statistics, we have

$$\begin{aligned} S_i &= \frac{S}{n} \\ G_i &= \frac{G}{n} \\ S &= G(1 - \frac{G}{n})^{n-1} \\ \lim_{n\to\infty} S &= Ge^{-G} \end{aligned}$$

The last equation is maximized to the *capacity* of $\frac{1}{e}$ when $G = 1$.

3.3 Collision resolution

The assumption about notifying collisions to receivers is that of *ternary feedback*, where at the end of each time slot, all users are notified whether a *collision* (2 or more users unsuccessfully transmitted), a *hole* (no one transmitted), or a *success* (one user successfully transmitted). Once a collision occurs, slotted ALOHA algorithms require each packet involved in a collision to be backlogged and queued, until it is successfully retransmitted. A probabilistic back-off mechanism is used, where at each subsequent time slot, the retransmission of a packet occurs with some probability

p. Owing to the memorylessness of the arrivals and the retransmissions, the system may be analyzed in terms of a homogeneous Markov chain, where the state is the number of backlogged packets at integer time n. If we denote the state of the system at time n as $X[n]$, then $P(X[n+1] = k + i | X[n] = k)$ may be defined, where i is the number of successful transmitted packets. Furthermore, the drift D_k, which is the expected value of the change of state conditioned on being in state k, may be appropriately defined. By observing that as the system becomes more and more backlogged, collisions are more and more likely. For a sufficiently large k, D_k will begin to tend to 0. It has been shown [Kap83] that the chain is non-ergodic, and the system is unstable. In other words, there is no justification for the assumption of statistical equilibrium in the derivations above.

The next improvement with ALOHA dealt with changing the retransmission probability as a function of k, the number of backlogged packets. In order to minimize D_k, a centralized controller would adjust $p(k)$ according to $pk + \lambda(1-p) = 1$. Assuming that a centralized controller would be able to do such, however, is clearly unreasonable. Decentralized control algorithms have been designed to attempt to estimate k and update p accordingly from the ternary feedback [HvL82]. Such algorithms provide stability for when $\lambda < e^{-}1$. It has been shown that based on our assumptions so far, whenever $\lambda > e^{-1}$, D_k is positive and the system is henceforth unstable.

ALOHA's next evolutionary stage took place by noticing that higher throughputs may be achieved if newly arrived packets are not attempted to be transmitted every time slot. Splitting algorithms [Cap78, TM78, Cap78], which are *delayed first transmission* (DFT) algorithms, were next devised to probabilistically separate the set of transmitting and non-transmitting packets, which may include newly arrived ones. These algorithms have been shown to be stable for $\lambda < 0.4871$. If the amount of feedback to the transmitters is improved to denote the number of packets involved in each collision, it has been shown [Pip81] that throughput up to unity may be achieved. Dynamic transmission probability control policies based on this *delayed first transmission* (DFT) approach have been introduced in [Riv87], where delayed packets are indistinguishable from newly arriving ones. Estimates of k are updated

by Baye's rule, and transmissions are attempted with probability $\frac{1}{k}$.

Some information-theoretic work has been done in understanding collision channels without the presence of feedback. Instead of performing retransmissions, forward error control coding allows error resiliency. Reed-Solomon codes have been proposed that allow users to code successive packets as code symbols of a Reed-Solomon code. The basic idea here is that if the decoder uses channel erasures (when collisions take place) as side information, then for arbitrarily long code sequences, the probability of successful decoding can become arbitrarily close to 1 and throughput may achieve the celebrated $\lambda e^{-\lambda}$, which is maximized to 0.368 when $\lambda = 1$. Similar Reed-Solomon (RS) coding ideas may be used to address unslotted collision channels to yield throughputs up to 0.184. A maximum channel utilization of $e^{-1} = 0.368$ has been shown to be achievable even in unslotted collision channels without feedback [MM85]. The techniques here use non-RS erasure-correcting codes to achieve maximum channel utilization of 0.368. This technique, however, assumes equal importance of all bits, which may not be practical since header bits in packets are of extreme importance. Recent work [Tho00] has taken this into account and shown achievability of throughputs up to 0.322 by performing bit-error correction across successive packets, instead of packet-erasure correction.

For a finite number of users, the stability region of a version of slotted ALOHA with buffering and retransmissions [Ana91]. The region turns out to be the closure of the capacity region of the collision channel without feedback (found in [MM85]). Analysis performed here is made tractable by considering arrival statistics that allow the discrete-time Markov chain of the protocol to be the embedded chain of a continuous time Markov process. If we assume that user i attempts to transmit a buffered packet in the next time slot with probability p_i, then the vector of arrival rates $\underline{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_M)$ is said to be in the stability region if there exists a transmission probability vector $\underline{p} = (p_1, p_2, ..., p_M)$ such that the resulting system is stable (namely, if the queue lengths have a stationary probability distribution). The stability

region has been shown to be the subset of $\mathbb{R}_+^M$ given by

$$C = \left\{ \text{vect}\left(p_i \prod_{j \neq i} (1 - p_j) \right) : 0 \leq p_i \leq 1, 1 \leq i \leq M \right\}.$$

3.4 Multiple Packet Reception Capability

Recent ideas have revolved around removing the assumption that users have catastrophic collisions and ternary feedback. The collision channel model does not necessarily hold in practical multiple-user communication situations. Overlapping transmissions of multiple users can still result in the reliable communication, as information theory tells us. Practically, the capture phenomenon [GS87] (where one received packet's power is much stronger than the other) can sometimes lead to the strongest one's successful decoding.

A relatively new idea for multiple-packet reception strategies is motivated in using spread-spectrum multiple access ideas. Spread-ALOHA (S-ALOHA) [Abr92] assumes that a number of high-bandwidth spreading sequences, each of length L chips, is provided to all users. Whenever a user attempts to transmit, it randomly selects one of the spreading sequences to modulate the symbols in the packet. The resultant packet is transmitted in following slot, and a collision occurs when two or more users try to use the same spreading sequence in the same slot. This scheme has well-behaved peak power properties and becomes increasingly more efficient for large bandwidth and small signal-to-noise ratios.

It has been shown, however, that from an information theoretic standpoint, using S-ALOHA does not improve the information per symbol interval [Tar95]. If we consider an AWGN channel with power signal noise variance σ_N^2, and packets arriving in a time slot according to a Poisson distribution, $P[G = n] = \frac{\lambda^n}{n!} e^{-\lambda}$, then the average amount of information per symbol interval, when symbols in a packet are drawn independently from a real-valued Gaussian distribution of variance P, is given by the channel capacity in bits per symbol, $C_{\sigma_N^2}(P)$. Since the average number of successful packets per time slot is given by $S = P[G = 1] = \frac{\lambda^n}{n!} e^{-\lambda}$, the average amount of in-

formation per symbol interval is given by $\frac{\lambda^n}{n!}e^{-\lambda}C_{\sigma_N^2}(P)$ bits per symbol. This places an upper bound on any transmission scheme, since the probability of packet error is non-zero for any finite packet length. To compare slotted ALOHA and S-ALOHA in a fair manner, the amount of bandwidth used in both schemes must be equal. Next, by noting that that there are L orthogonal spreading sequences of length L, we see the number of slots in a time interval of length L is the same for both schemes. Hence, the performance is the same. By using non-orthogonal schemes, for instance fewer than L orthogonal spreading sequences (which may have desirable shift-correlation properties), the information per symbol interval is decreased. Hence, S-ALOHA with orthogonal spreading sequences may be thought of as a form of coding, but does not improve the information per symbol interval.

Stability analysis of IFT slotted ALOHA systems with multiple-packet reception capability in the presence of channel noise has been performed in [GVS88]. The channel is a generalization of the usual collision channel, owing to its ability to allow more than one packet to be received correctly in a collision. The number of packets received successfully in a slot is modelled as a random variable that depends on n, the number of packets attempting to be transmitted. Noise, capture, and code-division multiplexing may be accounted for in this model by making appropriate probabilistic mapping from packets attempted to packets successfully received. Results from this analysis show that the system is stable if the packet arrival rate is less than the limit as n tends to infinity of expected number of packets successfully received.

Multiuser detection techniques have recently been proposed to allow for increased throughput and smaller delay for idealized channel conditions [SS00]. Other constructive techniques for multiple-packet reception in the presence of Additive White Gaussian Noise (AWGN) have been performed. Performing forward-error-correction across all received packets (which are usually discarded during a collision) using maximum-likelihood decoding in spread-spectrum systems has been introduced in [BGB97]. These code-combining techniques allow for improved probability of successful decoding, along with higher throughputs and smaller packet transmission delays. The ideas presented in [CLV00] use frequency-hopping communication with interleaving to ad-

dress coherent signal-space coding schemes that have collision resistance properties. The performance of such a system, which assumes linear single-user matched filters where other users are considered as noise, is evaluated in terms of outage probabilities and information rate achievability. This system has spectral efficiency comparable to S-ALOHA and that it trades off feedback and retransmission with long interleaving delay.

Chapter 4

Model

Chapters 2 and 3 illustrated different models for multi-user communications. We have seen that, until recently, these two models have been not considered jointly. Chapter 2 typically attempts to analyze what the limits are on communication in the presence of noise and multi-user interference, and assumes a constant influx of bits arriving to each user's transmitter. Chapter 3 attempts to primarily understand the impact of burstiness on channel utilization and quality of service, and makes limits of reliable multi-user communication in the presence of noise a secondary (if even that) concern.

The multiple-access system we consider from here on forward attempts to bridge some of these ideas by observing what possibilities arise when time slots are long enough to apply error-correcting codes at rates near capacity. Figure 4-1 gives an illustration of our model. The 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 over a single slot. 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 chan-

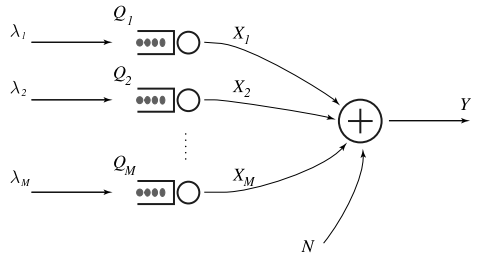

Figure 4-1: The M-user ALOHA model.

nels with small signal-to-noise ratios, where Turbo [BGT93] codes have been found 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.

Attempts to address ALOHA from an information-theoretic standpoint with the assumptions mentioned above have been performed in [MMH+02]. The capacity region for an M-user system and shared AWGN channel with given user power constraints has been found to the Cover-Wyner region for the associated multiple access channel. For a given set of arrival rates $\underline{\lambda}$ inside the multiple access capacity region, for some $\epsilon > 0, B_\epsilon < \infty$, and any $\underline{Q}(0) \subseteq \mathbb{R}_+^2$, there exists a coding policy that satisfies

$$\limsup_{j\to\infty} E[\exp(\epsilon \parallel \underline{Q}(j) \parallel)] \leq B_\epsilon, \tag{4.1}$$

which implies stability. It was also shown that any scheme that uses multiple access coding on the boundary of the dominant face of the Cover-Wyner region, at rates providing the drift of all queues to be negative, achieves stability. Furthermore, such a scheme does just as well as scheduling with queue information for long-term achievable rates.

4.1 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. User i is constrained to use an average of up to P_i units of power per transmission. 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. We may assume that this model is relevant in situations where fading is present and exhibits slow time fades, for instance in packetized indoor wireless systems. The atmospheric conditions in satellite systems also exhibit AWGN-like characteristics.

A user's queue contains all of its traffic which has not yet been successfully transmitted. Between time slot $n-1$ and n, $\underline{a}(n) = (a_1(n), a_2(n), ..., a_M(n))$ enter each of the users' transmission queues. During time slot n, $\underline{u}(n) = (u_1(n), u_2(n), ..., u_M(n))$ bits are reliably transmitted and removed from the users' queues. Let us denote $\underline{Q}(n) = (Q_1(n), Q_2(n), ..., Q_M(n))$ as the number of bits in the users' queues at time n^- (just before the nth time slot). Hence, user i's buffer state evolves according to

$$Q_i(n+1) = \max(a_i(n+1) + Q_i(n) - u_i(n), 0).$$

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. We note that a more reasonable assumption might be to assign a probability distribution $P_{L_i}(l)$ to the length of packets arriving to user i's buffer. We decide instead, to make the problem more tractable, to perform a certainty equivalent type analysis where L_i may be thought of as the expected value of the packet length with respect to distribution $P_{L_i}(l)$.

In the nth time slot:

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

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

4.2 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 $\underline{Q}(n)$. We may view this as a discrete time controlled stochastic system. Furthermore, we require that our system be stable. Issues of stability-achieving systems in this context were addressed in [MMH+02], and addressed concisely earlier with Eq. 4.1. The policies of interest must satisfy this stability condition for arrival rates of interest.

4.3 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 $\underline{Q}(n)$, the system of buffer state evolution satisfies the Markov condition:

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

$\underline{Q}$ 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. Figure 4-2 illustrates the 2-user Markov chain of queue state. In general, for the state $\underline{Q} = (Q_1, Q_2)$, there are 4 transitions leaving the state. There are numerous transition arrows entering state (Q_1, Q_2), due to general nature of the transmission policy, and the possibility of collisions.

4.4 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], as we saw in Chapter 2, quantify at what rate the probability of error decays exponentially with T. The notion of long time slots is the same as for a single user, where rates arbitrarily close to the single-user Shannon capacity can be achieved for codes with a sufficiently long block length (which corresponds to one time-slot in our model). As defined in [MMH+02], user i in time slot j sends one codeword each from a set of K_j^i codebooks $\mathcal{M}_j^{i,\kappa}$, $\kappa = 1, \ldots, K_j^i$. We denote the single-slot capacity for user i in slot j, defined below, as λ_i^j, and let $\underline{\lambda}^j = \left(\lambda_1^j, \ldots, \lambda_M^j\right)$. The set ordered set of

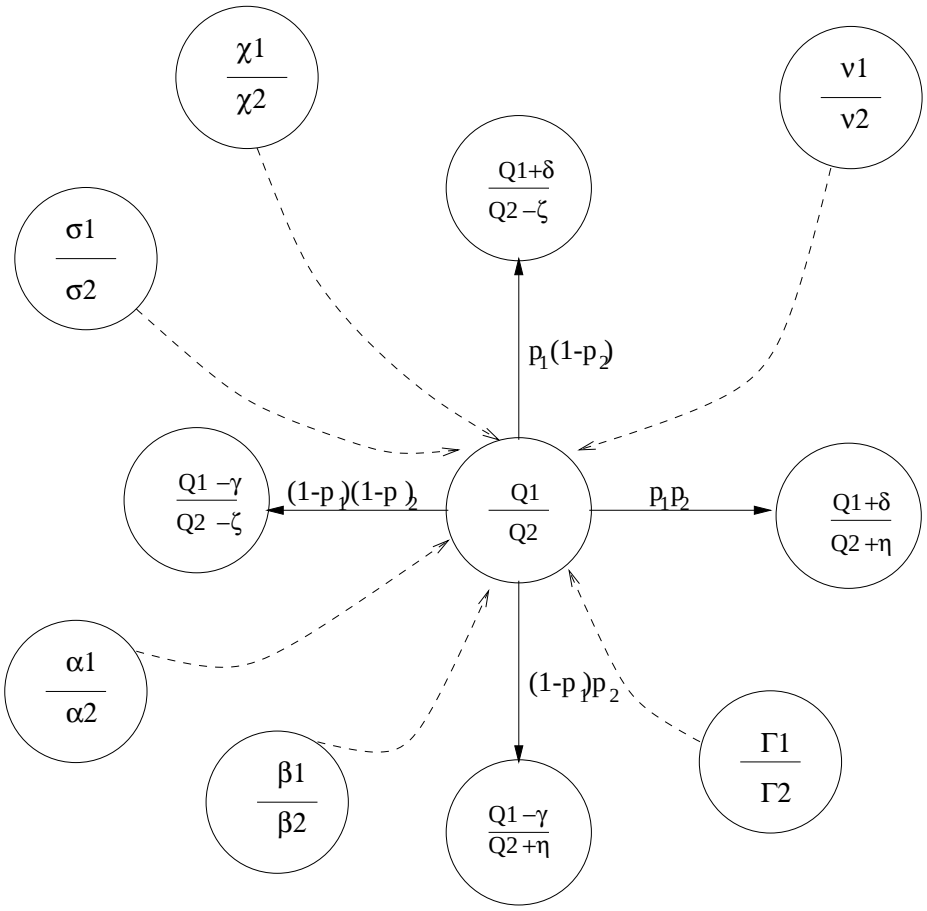

Figure 4-2: A visual illustration of the 2-user Markov chain around state (Q_1, Q_2)

codebooks $\left(\mathcal{M}_j^{i,\kappa}\right)_{\kappa=1}^{K_j^i}$ is called the codebook $\mathcal{C}_j^i$ for user i in slot j. We say that the codebook $\left(\mathcal{C}_j^1, \ldots, \mathcal{C}_j^M\right)$ achieves the single-slot capacity $\underline{\lambda}^j$ in slot j for slot-length T and error probability ξ (is $(T, \xi, \underline{\lambda}^j)$ single-slot capacity achieving) if for some sets $\mathcal{K}_j^1 \subseteq 1, \ldots K_j^1, \ldots, \mathcal{K}_j^M \subseteq (1, \ldots, K_j^M)$ known to both the transmitter and receiver there exists a decoding policy such that

(i) Every codeword from a codebook $\mathcal{M}_j^{i,\kappa}$ where $\kappa \in \mathcal{K}_j^i$ is decoded with probability of error ξ or less.

(ii) The rate associated with that codeword transmission equals the single-slot capacity, thus for $i = 1, \ldots, M$

$$\sum_{\kappa \in \mathcal{K}_j^i} \frac{\log\left(|\mathcal{M}_j^{i,\kappa}|\right)}{T} \geq \lambda_j^i.$$

A codeword that was decoded with probability ξ or less is considered to have been reliably received. We say that a codebook satisfying the conditions outlined above is $\left(T, \xi, \underline{\lambda}^j\right)$ single-slot capacity-achieving. Note that this definition differs from the standard capacity definition in that on slot j each user need not send *any* codeword in its codebook C_j^i with arbitrarily small probability: he need only send a subset of his codewords with arbitrarily small probability. This subset corresponds to a rate below the maximum associated with the full codebook, to allow for a lower rate to be reliably received in the event of a collision.

We now define multiple-slot capacity based on this single-slot capacity definition. Assume we now transmit over n slots. For a given T and $\xi > 0$, a coding and decoding policy is $(T, \xi, \underline{\lambda})$ capacity-achieving if $\forall i, \forall j$, $\exists \mathcal{C}_j^i$ that is $\left(T, \xi, \underline{\lambda}^j\right)$ single-slot capacity achieving and

$$\lim_{n \to \infty} \frac{1}{nT} \sum_{j=1}^{n} \lambda_j^i \geq \lambda_i \qquad 1 \leq i \leq M. \tag{4.2}$$

The notion of capacity described above is related to other delay-constrained and probability-of-failure notions of capacity, such as delay-limited capacity [HT98], ϵ-capacity [VH94], capacity versus outage [Sha97, GCB98], and expected capacity

[EG98]. We consider delay constraints because of finite time slot length. We use expected rates because of uncertainty regarding collisions.

Chapter 5

Minimizing Delay and Minimizing Power Consumption

We now investigate how delay minimization and power consumption minimization are affected by knowledge of user queue information. This will help to motivate the proposed scheme that will be addressed in chapter 6. We restrict our attention to a two-user scenario here, but the results may easily be extended for many users. At the beginning of time slot n,

$$\underline{a}(n) = \begin{cases} (L_1, L_2) & \text{with probability } p_1 p_2 \\ (L_1, 0) & \text{with probability } p_1(1-p_2) \\ (0, L_2) & \text{with probability } (1-p_1)p_2 \\ (0, 0) & \text{with probability } (1-p_1)(1-p_2) \end{cases}$$

5.1 Delay Minimization

We now would like to understand what the minimum amount of power is required to minimize delay. Consider the situation where the set of (p_i, L_i), burstiness pairs, for $i = 1, 2$, is known to both users. We consider the cases where users have full knowledge of each other's queues and where they only have local queue information.

Let us denote

$$C_{\sigma_N^2}(x) = \frac{1}{2}\log_2\left(1 + \frac{x}{\sigma_N^2}\right)$$

as the capacity (in bits per transmission) of a discrete-time memoryless Gaussian channel with 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,

$$C_{\sigma_N^2}^{-1}(x) = \sigma_N^2(2^{2x} - 1)$$

is the minimum amount of average power required to transmit rate x and noise variance σ_N^2 with arbitrarily vanishing error probability. In terms of signal-to-noise ratio (SNR), $C_1(x)$ is the capacity of the memoryless discrete-time Gaussian channel with SNR equal to x, and $C_1^{-1}(x)$ is the minimum SNR required to reliably achieve the rate x.

5.1.1 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: $\underline{a}(n)$. 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 power per transmission required is i.i.d. over each time slot n, and is a function of $\underline{a}(n)$. The minimum amount of aggregate power per transmission required to empty the buffers in one time slot is given by:

$$P_{min}^{(1)}(n) = \begin{cases} C_{\sigma_N^2}^{-1}\left(\frac{L_1+L_2}{T}\right) & \text{with probability } p_1p_2 \\ C_{\sigma_N^2}^{-1}\left(\frac{L_1}{T}\right) & \text{with probability } p_1(1-p_2) \\ C_{\sigma_N^2}^{-1}\left(\frac{L_2}{T}\right) & \text{with probability } (1-p_1)p_2 \\ 0 & \text{with probability } (1-p_1)(1-p_2) \end{cases}$$

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

$$\lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}P_{min}^{(1)}(n)\underset{a.s.}{\longrightarrow}p_1p_2C_{\sigma_N^2}^{-1}\left(\frac{L_1+L_2}{T}\right)+p_1(1-p_2)C_{\sigma_N^2}^{-1}\left(\frac{L_1}{T}\right)+(1-p_1)p_2C_{\sigma_N^2}^{-1}\left(\frac{L_2}{T}\right).$$

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

5.1.2 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 other's queue, then to minimize delay, all users must coordinate to transmit at the worst case scenario: when (L_1, L_2) bits enter each the queues. So user each only has access to a multiple-access codebook. Each user always anticipates the other user's presence and uses the amount of power required to empty both queues. We decompose this problem into two parts. First of all, we know from the Cover-Wyner region that the minimum amount of aggregate power required to transmit the rate pair (L_1, L_2) reliably is given by

$$C_{\sigma_N^2}^{-1}\left(\frac{L_1+L_2}{T}\right).$$

Note that many per-user power constraints may result in the rate pair lying on the dominant face of the multiple access region. Figure 5-1 shows an illustration of this for a pair of power choices. Depending on the burstiness probabilities, however, some of these power constraints may provide smaller long-term average aggregate power consumption than others. So these power constraints may be chosen as a function of the burstiness pairs, (p_i, L_i) for $i = 1, 2$, so long as they may reliably achieve the rate pair $(\frac{L_1}{T}, \frac{L_2}{T})$. If the power constraints lie in the region $\mathcal{P}$ denoted as:

$$\begin{aligned} P_1 &\geq C_{\sigma_N^2}^{-1}\left(\tfrac{L_1}{T}\right) \\ P_2 &\geq C_{\sigma_N^2}^{-1}\left(\tfrac{L_2}{T}\right) \\ P_1+P_2 &= C_{\sigma_N^2}^{-1}\left(\tfrac{L_1+L_2}{T}\right) \end{aligned}$$

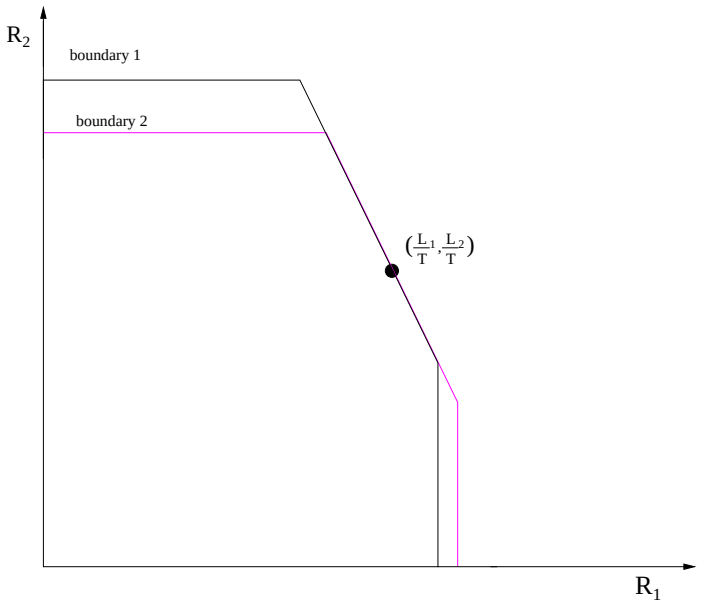

Figure 5-1: Capacity regions for different (P_1, P_2) choices that satisfy the equations above

then for block coding multiple access schemes used over a slot, there exists a time-sharing ratio ζ such that

$$\begin{aligned} \zeta C_{\sigma_N^2}(P_1) + (1-\zeta) C_{\sigma_N^2+P_2}(P_1) &= \tfrac{L_1}{T} \\ \zeta C_{\sigma_N^2}(P_2) + (1-\zeta) C_{\sigma_N^2+P_1}(P_2) &= \tfrac{L_2}{T} \end{aligned}$$

The choice of power constraints is made to minimize the long term average aggregate power consumption

$$\begin{aligned} J(P_1, P_2) &= p_1(1-p_2)P_1 + p_2(1-p_1)P_2 + p_1 p_2 (P_1 + P_2) \\ &= p_1(1-p_2)P_1 + p_2(1-p_1)\big(C^{-1}_{\sigma_N^2}\left(\frac{L_1+L_2}{T}\right) - P_1\big) + p_1 p_2 C^{-1}_{\sigma_N^2}\left(\frac{L_1+L_2}{T}\right). \end{aligned}$$

Note that the users' power constraints differ from the long-term average power consumption. This is due to the burstiness of packet arrivals. Each user's power constraints may be thought of as the amount of power used for some reliable coding mechanism over a slot for transmitting. During some time slots, however, because of burstiness and the dynamical behavior of the buffers, one user, or both may not be

transmitting. Hence, the long-term average amount of aggregate power used by both users is in fact different, and is governed by the equation above. We note that because we are constraining the sum of the users' powers to meet with equality, there is really only one degree of freedom. We have a linear objective function in one variable, and it must be minimized subject to a constraint on the value P_1:

$$C_{\sigma_N^2}^{-1}\left(\frac{L_1}{T}\right) \leq P_1 \leq C_{\sigma_N^2}^{-1}\left(\frac{L_1+L_2}{T}\right) - C_{\sigma_N^2}^{-1}\left(\frac{L_2}{T}\right).$$

The minimum is attained at either of the two boundary points, depending on the sign of $p_1 - p_2$:

$$(P_1^*, P_2^*) = \begin{cases} \left(C_{\sigma_N^2}^{-1}\left(\frac{L_1}{T}\right), C_{\sigma_N^2}^{-1}\left(\frac{L_1+L_2}{T}\right) - C_{\sigma_N^2}^{-1}\left(\frac{L_1}{\sigma_N^2}\right)\right) & \text{if } p_1 > p_2 \\ \text{any } (P_1, P_2) \in \mathcal{P} & \text{if } p_1 = p_2 \\ \left(C_{\sigma_N^2}^{-1}\left(\frac{L_1+L_2}{T}\right) - C_{\sigma_N^2}^{-1}\left(\frac{L_2}{T}\right), C_{\sigma_N^2}^{-1}\left(\frac{L_2}{T}\right)\right) & \text{if } p_1 < p_2 \end{cases}$$

To minimize long-term average aggregate power consumption, the rate pair to be achieved lies on either of the two boundary points of the dominant face of the multiple access region for unequal burstiness probabilities. For equal burstiness probabilities, any point on the dominant face leads to the same long-term average aggregate power consumption. The long term average minimum amount of power per transmission required for this scheme is given by

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} P_{min}^{(2)}(n) \underset{a.s.}{\longrightarrow} p_1(1-p_2)P_1^* + p_2(1-p_1)P_2^* + p_1p_2(P_1^* + P_2^*),$$

where P_1^* and P_2^* have been given above.

5.2 Minimizing Power Consumption

We now address the minimum amount of average power consumption needed to stabilize the bursty system. Concavity of the function $\log_2(1+x)$ provides the inequality $\frac{1}{2}\log_2(1+\frac{P}{\sigma_N^2}) \leq 2 * \frac{1}{2}\log_2\left(1+\frac{\frac{P}{2}}{\sigma_N^2}\right)$. So it is more favorable in terms of aggregate

power consumption to spread the same amount of power into multiple time slot uses rather than in just one use. Note that we may generalize this to more than two slots, so long as the system is stable. For multiple users to transmit reliably 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. Note that in terms of long-term power consumption, for certain types of ergodic arrival processes, user queue information is not necessary to perform this strategy. If each user artificially backs up its queue by not transmitting, then eventually each user will have a very large queue length. At that point, each user will have data to transmit. Afterwards, users transmit achieving the rate pair $(\frac{p_1 L_1}{T}, \frac{p_2 L_2}{T})$ lying on the dominant face of the Cover-Wyner region. As the vector of output rates tends toward the vector of input rates from above, the amount of power consumption required decreases, but average delay increases. Since the system must provide stability, the minimum amount of power 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, however, that as the vector of output tends towards the vector of input rates from above, the delay increases without bound. Hence, the delay in the case where aggregate power consumption is to be minimized is infinite. The average aggregate amount of power required is given by

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} P_{min}^{(3)}(n) \underset{a.s.}{\longrightarrow} C_{\sigma_N^2}^{-1} \left(\frac{p_1 L_1}{T} + \frac{p_2 L_2}{T} \right) = C_{\sigma_N^2}^{-1} \left(\lambda_1 + \lambda_2 \right).$$

Figure 5-2 illustrates what the minimum power *per-time slot constraint* (assuming unit noise variance) is for each scheme to stably achieve its minimization objective. We illustrate this for a fixed packet length, $\frac{L}{T}$. Note that the power constraints for full queue info are not listed, because they vary depending upon the per-time slot queue state they observe. Similarly, Figure 5-3 illustrates what the corresponding *long-term average consumption* is for each scheme to stably achieve its minimization objective. Note that depending on burstiness and queue information, a particular

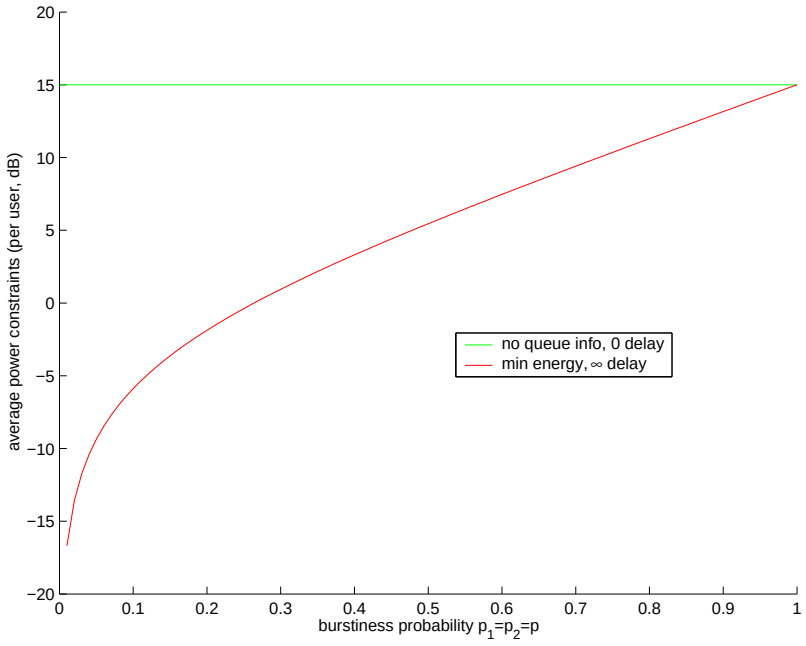

Figure 5-2: Power constraints required for minimizing average bit delay with global queue information, and for minimizing power consumption

scheme's average power constraint and consumption may differ. This is due to the fact that, when users do not have global queue information, they must code (and henceforth allocate power) for the worst-case scenario. When the worst-case scenario does not occur, the per-time-slot constraint on power will not equal the true power consumed. Nonetheless, we see that queue information has a significant effect on the performance of such systems.

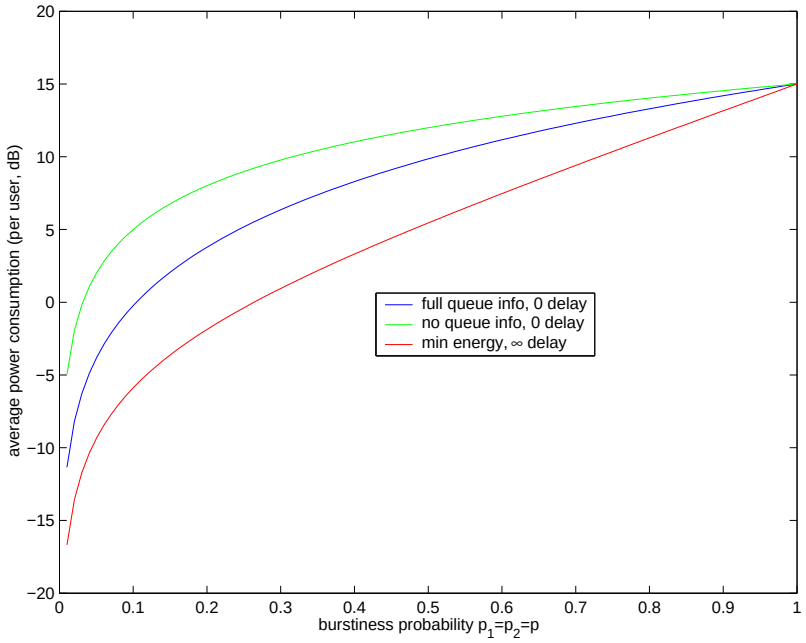

Figure 5-3: Average aggregate power consumption for minimizing average bit delay with and without detailed global queue information, and minimizing power consumption

Chapter 6

Analysis of a System with Limited Queue Information

The previous chapter illustrated not only that there is an intrinsic trade-off between power consumption and average delay in this stable ALOHA system model, but also that this trade-off is parametrized by the amount of queue information provided to users.

An information-theoretic treatment to characterize the trade-off between power consumption and buffer cost has been performed in [Ber00]. That particular piece of analysis considered bursty multi-user systems with flat narrow-band fading, and *full* queue information. It also addressed questions about error exponents and outage probability by fixing transmission of codewords to be over the time duration of the constant fade. A dynamic programming technique was used to find policies that minimize a general objective function characterizing the tradeoff:

$$\mu* = \arg\min_{\mu\in U} \limsup_{m\to\infty} \frac{1}{m} J_\mu(\underline{Q}(m)) \tag{6.1}$$

$$= \arg\min_{\mu\in U} \limsup_{m\to\infty} \frac{1}{m} \mathcal{P}\big(\mu(\underline{Q}(m))\big) + b_1(Q_1(m)) + b_2(Q_2(m)) \tag{6.2}$$

where $\mathcal{P}()$ is an appropriately defined convex power-cost function, μ is a *centralized* policy with global state information $\underline{Q}(m)$ for allocating powers to transmitting data from the buffers, U is the set of all valid control policies, and b^i's are buffer cost

functions (which might relate to buffer overflow probability or long-term average delay). This analysis, however, did not attempt to understand the role of queue information - it assumed centralized global queue information throughout its analysis. It was also illustrated in [Ber00] that a sequence of multi-user policies with only very little queue information is asymptotically optimal. The overhead of providing such information, however, was not taken into account in this model.

Motivated by the results of the previous section, we attempt to address policies that are constrained to have very limited queue information. We are also interested in the trade-off between power and delay. However, without full queue information for all users, this particular problem may not be addressed totally from a dynamic programming context. Namely, a converse has not been proven to illustrate what the set of all achievable expected rates for a multi-user time-slotted system with particular user transmission probabilities. A coding theorem providing the largest known set of achievable rates appears in [MMH$^+$02] and another scheme that generates the same set of rates using less codewords is provided in Appendix B. Such a strategy involves multiple users performing rate-splitting for both multiple-access and broadcast purposes. The multiple-access reason is evident; the broadcast reason captures the burstiness in users: it allows variable reliably received rates, depending upon whether or not other users transmit. As a consequence of the difficulty of performing optimization, we take a different approach. We offer a general class of coding schemes that allow nice opportunities for joint source/channel coding at the application layer. We present a coding scheme that addresses the burstiness of packet arrivals. We provide an analysis of its performance. Rather than attempt to combat burstiness and the possibility of collisions by probabilistically backing off transmissions, we consider using some information theoretic rate-splitting ideas to reliably communicate in the presence of multiple users, at received rates that vary depending on the number of users transmitting during any slot. Because of the burstiness of arrivals and the ensuing possibility of collisions requiring portions of the data to be retransmitted, we treat this system in a queueing context. The system utilizes a very small amount of queue information among users to operate in two modes: a multiple-access mode

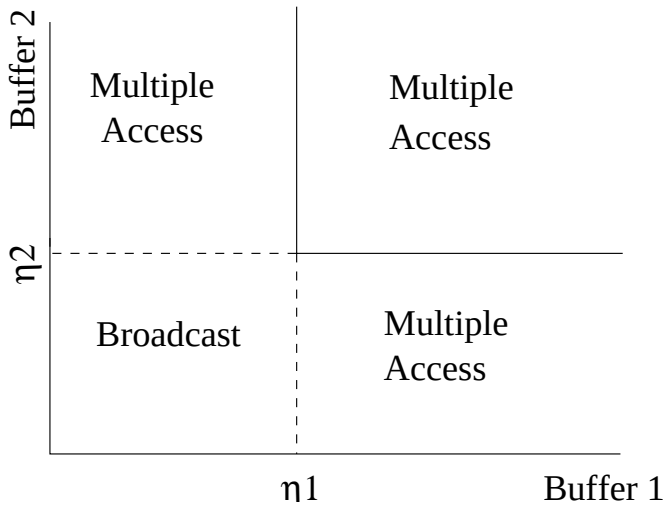

Figure 6-1: Modes of operation as a function of the queue lengths

when queue lengths are large, and a hybrid broadcast/multiple access mode otherwise (see figure 6-1). This scheme tries to address tradeoffs between average bit delay and average power consumption parameters by affording a compromise between the schemes mentioned in the previous section.

6.1 System Design

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 it attempted to transmit was received reliably, and which portion needs to be retransmitted. Hence, the amount of feedback is slightly more coarse than the ternary feedback model present in the ALOHA systems mentioned in Chapter 3.

6.1.1 Mode I (Large Queue Lengths): Multiple Access

The results in [MMH$^+$02] show that the capacity region of the time-slotted ALOHA system with power-constrained users in the presence of AWGN is the same as the

capacity region of the corresponding information-theoretic multiple-access channel. For any vector of arrival rate $\underline{\lambda}$ lying inside the Cover-Wyner region, there exist $(T, \xi, \underline{\lambda})$ *capacity-achieving* coding schemes that will provide system stability. As described in [MMH+02], 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 $\underline{Q}$, we denote the vector-valued drift of the state to be

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

In our case, based on our model of the data transmission policies state information in Section 4.2, $D_i(\underline{q}) = \lambda_i T - \mu_i(\underline{q})T$, where $\mu_i(\underline{q})$ is a function of $\underline{q}$. If there is only a finite set of states such that the drift inequality $D_i(\underline{q}) < 0 \ \forall i$ is not satisfied, then from we note via Pake's Lemma [Pak69] that 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 $\underline{\eta} = (\eta_1, \eta_2, ..., \eta_M)$. Thus, given a set of burstiness pairs and per transmission power constraints, as all users' queues states become backed up (cross this threshold $\underline{\eta}$), they transmit data out of the queues at rates $\underline{\mu}_{MA} = (\mu_{MA,1}, \mu_{MA,2}, ..., \mu_{MA,M}) = E[\underline{\mu} \mid \underline{Q} \geq \underline{\eta}]$. To transmit optimally aggregate data subject to the power constraints, $\underline{\mu}_{MA}$ must lay on the dominant face of the multiple access region:

$$\sum_{j=1}^{M} \mu_{MA,j} = C_{\sigma_N^2}\left(\sum_{j=1}^{M} P_j\right)$$

To ensure stability that operating point should provide negative drift for all queues:

$$\lambda_j - \mu_{MA,j} < 0 \ \forall j.$$

We denote the *reasonable power constraint region* as the set of multiple access mode power constraints for our system in that satisfy:

$$C_{\sigma_N^2}^{-1}\left(\frac{p_1 L_1}{T}+\frac{p_2 L_2}{T}\right) \leq P_1+P_2 \leq C_{\sigma_N^2}^{-1}\left(\frac{L_1+L_2}{T}\right). \tag{6.3}$$

If the power constraints were to lie below the lower bound, the system would not be stable ($\underline{\lambda} > \underline{\mu}$). On the other hand, if the power constraints were to lie above the upper bound, then the corresponding power constraints of a delay-minimizing scheme would be less, and so would the average delay (which is 0). Equivalently, from a more queueing theory perspective, we may say that it is suitable for

$$\rho = \frac{\sum_{i=1}^{M} \lambda_i}{\sum_{j=1}^{M} \mu_{MA,j}} = \frac{\sum_{i=1}^{M} \frac{p_i L_i}{T}}{C_{\sigma_N^2}\left(\sum_{j=1}^{M} P_j\right)}$$

to take on values only within a subset of $(0,1)$. The set of feasible power constraints requires ρ to satisfy

$$\frac{\sum_{i=1}^{M} \frac{p_i L_i}{T}}{\sum_{j=1}^{M} \frac{L_j}{T}} < \rho < 1.$$

6.1.2 Mode II (Small Queue Lengths): Hybrid Broadcast/Multiple Access

When user queue lengths are small (below the threshold $\underline{\eta}$), they switch to a combined multiple access/broadcast mode. Even in the event of a collision, data is reliably received from all users. The coding scheme used in this mode allows combating burstiness by achieving variable reliably received rates. 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 eliminated for future decoding). Similar techniques are used for achieving capacity on the multiple access channel. We provide a coding strategy (discussed in detail in Appendix B) that employs both techniques. The rationale for the multiple-access reason portion of the technique is evident; the broadcast reason is

to allow variable reliably received rates, depending upon whether or not other users transmit. In the analysis in the previous chapter, we note that the nature of the problem (either minimizing delay to 0 or stably minimizing the amount of aggregate power consumption) implies that the amount of power used per time slot is either some fixed value or 0. Consequently, to afford a power-delay trade-off for our system, we observe that users should not use the same amount of transmit power at all times. Certain modes of operation afford more use of transmit power than others, and the system parameters may be tuned to vary so that the fraction of time the system is in one mode versus another. We capture this by introducing the parameter γ, which is the ratio of the mode II power to the mode I power:

$$P_{BC,i} = \gamma_i P_i.$$

Each user has a simple *deterministic* transmission policy: if a user has data in its queue to transmit, it attempts to do so. A fraction α_i of user i's power $P_{BC,i}$, 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. All virtual users generate independent zero-mean Gaussian codewords with variance equal to the respective powers (see figure B-2). 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. Likewise, it generates codewords according to independent zero-mean Gaussian random variables with variance $(1 - \alpha)P_{BC,i}$. More rate-splitting is performed to achieve higher rates when users do both transmit. Motivated by the multiple access results in [RU96], this is performed by further rate-splitting amongst the low-resolution users. Figure B-2 (in the appendix section) illustrates the rate-splitting process for users.

For the two-user case, each signal of the LR and HR type has a rate such that it

can be decoded within the required probability of error if the SNR is at least:

$$\begin{aligned}
SNR'_{LR,1} &= \frac{\alpha_1\beta P_{BC,1}}{P_{BC,2}+(1-\alpha_1\beta)P_{BC,1}+\sigma_N^2} && \text{for } \mathrm{LR}'_1 \\
SNR_{LR,2} &= \frac{\alpha_2 P_2}{(1-\alpha_2)P_{BC,2}+(1-\alpha_1\beta)P_{BC,1}+\sigma_N^2} && \text{for } \mathrm{LR}_2 \\
SNR''_{LR,1} &= \frac{\alpha_1(1-\beta)P_{BC,1}}{(1-\alpha_2)P_{BC,2}+(1-\alpha_1)P_{BC,1}+\sigma_N^2} && \text{for } \mathrm{LR}''_1 \\
SNR_{HR,1} &= \frac{(1-\alpha_1)P_{BC,1}}{\sigma_N^2} && \text{for } \mathrm{HR}_1 \\
SNR_{HR,2} &= \frac{(1-\alpha_2)P_{BC,2}}{\sigma_N^2} && \text{for } \mathrm{HR}_2
\end{aligned}$$

Hence, the low-resolution and high-resolution rates that may be achieved are as follows:

$$\begin{aligned}
\mu_{LR,1} &= \frac{1}{2}\log_2(1+SNR'_{LR,1}) + \frac{1}{2}\log_2(1+SNR''_{LR,1}) \\
\mu_{HR,1} &= \frac{1}{2}\log_2(1+SNR_{HR,1}) \\
\mu_{LR,2} &= \frac{1}{2}\log_2(1+SNR_{LR,2}) \\
\mu_{HR,2} &= \frac{1}{2}\log_2(1+SNR_{HR,2})
\end{aligned}$$

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:

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

6.2 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 hybrid broadcast/multiple access 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 (FSA) 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 are below η_i, each user employs the hybrid broadcast/multiple access scheme. Otherwise, users employ multiple access mode encoding. We do not model the communication link between users for this communication, but note that the shared information is not substantial.

6.3 Performance

We note that for any set of burstiness pairs $\{(p_i, L_i)\}_{i=1}^2$ with corresponding rate-tuples $\left(\frac{p_1 L_1}{T}, \frac{p_2 L_2}{T}\right)$ lying inside the Cover-Wyner region, proper coding of our scheme during multiple access mode will result in the Markov chain being ergodic. Let us consider the two-user scenario and note how the analysis may easily extended for more users. The steady-state probabilities $\pi_{\underline{q}} = \lim_{n\to\infty} P[\underline{Q}(n) = \underline{q}]$ for the Markov chain are governed by:

- The burstiness pairs (p_i, L_i) of each user
- The average per-transmission multiple access power constraints P_i for each user,
- The ratio of the mode II (hybrid broadcast/multiple access) power constraint to the mode I (multiple access) power constraint γ_i for each user
- The broadcast rate-splitting power ratio α_i and low-resolution multiple access rate-splitting ratio β_i for each user
- The operation mode thresholds η_i

We note that the long-term average queue size

$$N(\underline{P}, \underline{p}, \underline{L}, \underline{\alpha}, \underline{\beta}, \underline{\eta}, \underline{\gamma}) = \sum_{\underline{q} \subseteq \mathbb{R}_+^M} \left(\sum_{i=1}^{M} q_i \pi_i(\underline{P}, \underline{p}, \underline{L}, \underline{\alpha}, \underline{\beta}, \underline{\eta}, \underline{\gamma}) \right)$$

may be used to calculate the long-term average bit delay $\overline{T}(\underline{P}, \underline{p}, \underline{L}, \underline{\alpha}, \underline{\beta}, \underline{\eta}, \underline{\gamma})$ via Little's Result: $\overline{T} = \frac{N}{\lambda_1+\lambda_2}$. We might imagine that for a set of time-slot power constraints $\underline{P}$, we could attempt to choose the power splitting ratios $\underline{\alpha}$ to minimize the long-term average bit delay,

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

For an open-loop policy of this form for a particular $\underline{\alpha}$ and $\underline{\beta}$, we may attempt to understand the Markov chain describing state transitions and the steady-state probabilities (see Appendix A for a characterization of the global balance equations for steady-state probabilities). These steady-state probabilities are in fact quite difficult to analyze, due to the logarithms and the regimes where balance equations are of different forms. We would like to instead interpret the choice of allocating high-resolution and low-resolution is more of an application-specific tunable parameter than something to optimize. The source-channel separation theorem does not hold in this type of scheme, because of the broadcasting mechanism employed. Thus, applications using this scheme may allocate data with different levels of quality of service to these different streams. The low-resolution stream is a constant influx of information that may be reliably achieved. This information may be decoded for users with very stringent quality of service constraints - such as voice, streaming audio/video, etc. On the other hand, the high-resolution information may be coded and transmitted to users in a variable amount of time. Since the Markov chain is ergodic for all arrival rates inside the multiple access region, we may analyze our system by truncating the state space and perform an approximation [Fre71] using simulations on a state space of a finite number of states.

Since we use a very limited amount of queue information sharing, and are willing to accept a small but non-zero average bit delay, this proposed scheme affords a compromise between the delay minimizing scheme with no queue information and the power consumption minimizing system of the previous section.

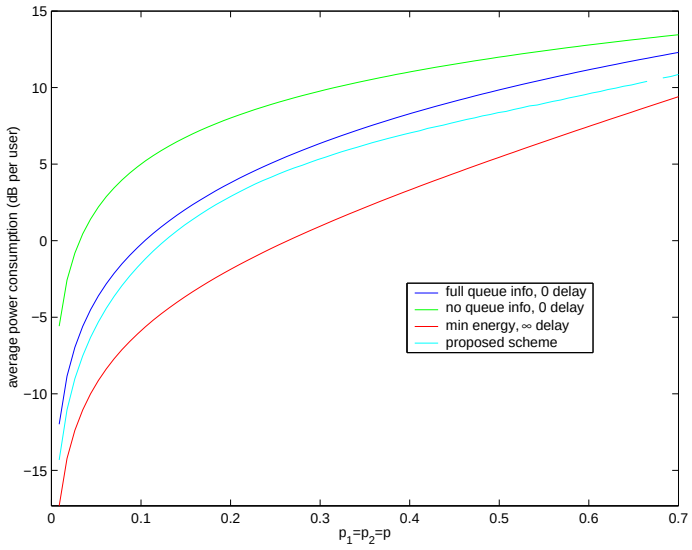

Figure 6-2: Average aggregate power consumption as a function of burstiness for fixed packet length and varying probabilities with $\alpha_1 = \alpha_1 = 0.5$, $\sigma_N^2 = 1$

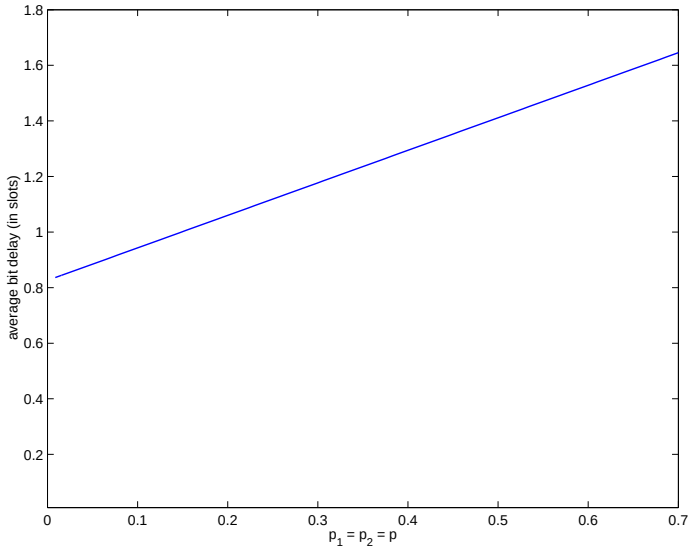

Figure 6-3: Average bit delay as a function of burstiness for fixed packet length and varying probabilities with $\alpha_1 = \alpha_1 = 0.5$, $\sigma_N^2 = 1$

Figures 6-2 and 6-3 show simulation results for the average power consumption and average bit delay for our proposed scheme with power constraints being a fixed convex combination of the boundaries of the reasonable power constraint region for each value of $p = p_1 = p_2$. The average power consumption of the systems mentioned in the previous section are superimposed in the figure as well. In the regime of small yet nonzero burstiness probabilities (which is where most bursty packetized systems operate), the impact of allowing a small yet nonzero tolerable delay along with a small amount of queue information sharing is illustrated: both average bit delay and average power consumption are near their respective minimal boundaries. Our scheme uses less energy than that of a system with no queue information and 0 delay because that system obtains no large benefit from one of the two users being empty. In our scheme, however, most of the time the system is in broadcast mode and if one of the two users is empty, that user consumes no power to transmit while other user may reliably transmit the low-resolution *and* high-resolution data during that time slot.

6.3.1 Placement of η and γ

We next attempt to understand the relation between placement of the boundary η_i between the two modes of operation for each user (see figure 6-1), and how this relates to how much power is allocated in the mode II. We note that there is a trade-off between power consumption and delay in moving η_i towards 0 or ∞. We note that from section 5.2, to minimize power consumption, users should not transmit (using no power) for an arbitrary long time, and afterwards transmit at multiple access, affording transmission rates on the boundary of the Cover-Wyner region. In our case, while in the broadcast-multiple access bursty mode, users transmit with less power, and due to uncertainty regarding each other's transmissions, afford reliably received rates less than those in the multiple access mode. There exists a power-delay trade-off in terms of allocating system parameters. Note that the amount of power used in broadcast mode per time slot, which is γP_{MA}, trades off power consumption and delay. Allowing γ to decrease affords a smaller amount of aggregate power consump-

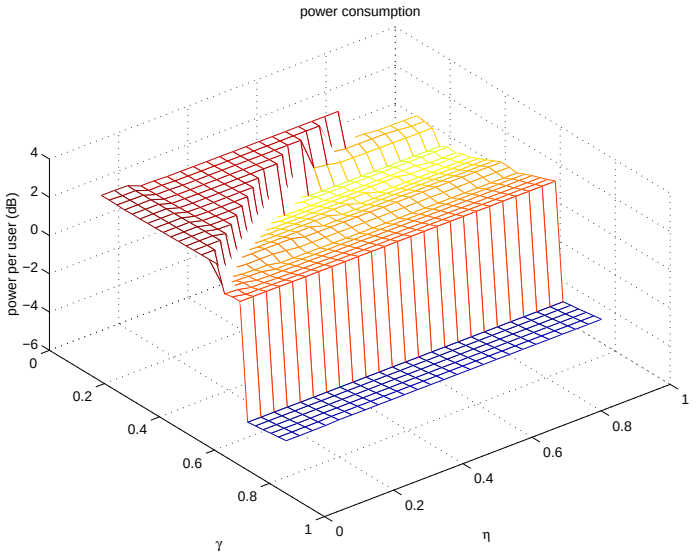

Figure 6-4: average power consumption for $p = 0.1$

tion. However, on the flip side, the amount of delay incurred is penalized. Also, the operation mode boundary η_i, trades off power consumption and delay as well.

Small Packet Lengths

For small packet lengths, we note that it may be possible for users to use enough power in a single time-slot to transmit a whole packet when there is no collision. In this case, it is very advantageous from both a power and delay perspective to do so. Figures 6-4 and 6-5 illustrate these results.

When the ratio of the broadcast mode power to multiple access mode power, γ, is above a particular threshold, the average power consumption and average bit delay both drop steeply. This threshold corresponds to when the combined $\mu_{LR,i} + \mu_{HR,i} = \frac{L_i}{T}$. In other words, this threshold corresponds to when a full packet arrives into the system and can be emptied out by one user when the other user is not transmitting. Another steep change in the curve, which is a function of both η_i and γ_i, takes place at the threshold where the operation mode boundary satisfies $\eta_i = \frac{L_i}{T} - (\mu_{LR,i} + \mu_{HR,i})$. At this threshold, when a packet arrives to the system, where all other queue lengths are 0, $\mu_{LR,i} + \mu_{HR,i}$ bits are emptied reliably and the system transitions into multiple

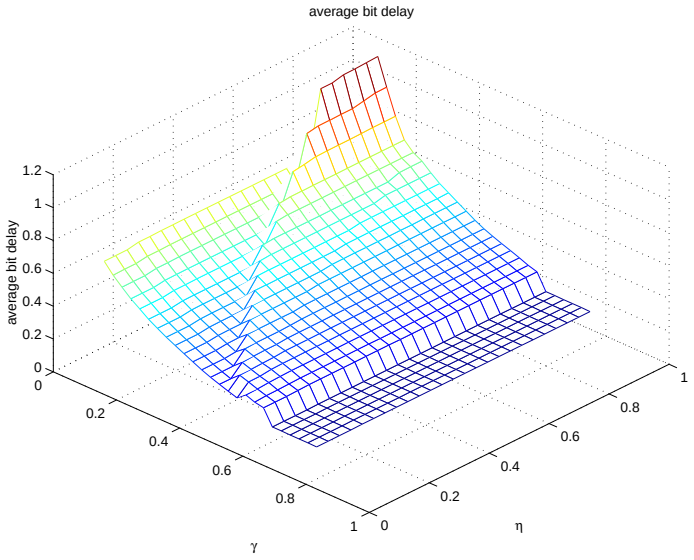

Figure 6-5: average bit delay for $p = 0.1$

access in the next slot. From a power consumption point of view, this incurs extra cost for all η_is less than this value. But from a delay perspective, this has the opposite effect.

Large Packet Lengths

We now consider large packet lengths, which require multiple time slots for transmission for any reasonable SNR. In this case, we note that the trade-off between power consumption and delay can readily be illustrated for our proposed scheme. Simulation results illustrate the power consumption and average bit delay, at a particular value of ρ, plotted as functions of varying η and γ. We note from the simulations that, for appropriate values of γ, increasing η improves power consumption while increasing delay. However, as figure 6-6 illustrates, the effect of increasing η for poor choices of γ (namely near 1) reverses. It becomes increasingly worse from both power consumption and delay perspectives to increase η. This is because the system is not exploiting any reduction in power consumption for being in a mode that delivers smaller arrived data rates. Hence, from a system design perspective, as long as the system operates in a regime where γ is reasonable, power consumption and delay are traded off by

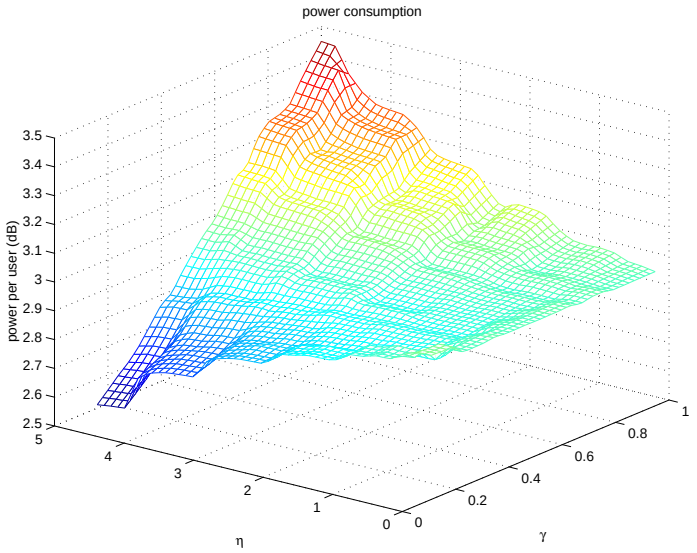

Figure 6-6: average power consumption for $p = 0.1$, $\rho = 0.75$

varying η.

We note that, by choosing appropriate γ, η, and ρ, we may be able to characterize the trade-off for delays lying between 0 and ∞. Hence, our scheme allows our power consumption curves to lie anywhere between the power consumption minimizing, and delay-minimizing curves. We note that for small values of p (less than 0.1 is where a system demonstrating burstiness usually lies), there is a tremendous gap between the two boundary curves and henceforth large room for improvement of delay with small queue information. By tuning such parameters of our system as ρ, γ, and η, we may afford very reasonable trade-offs. Interestingly, the power and delay trade-off for a particular values of ρ, in our system, as we may see in 6-8 is approximately linear.

6.4 Conclusions

We have considered multiple users with bursty data simultaneously communicating in the presence of noise. We have studied an aspect of the trade-off between power consumption and delay in multi-user systems, and more importantly, how it is parametrized by queue information. To perform this analysis, we relied on ideas

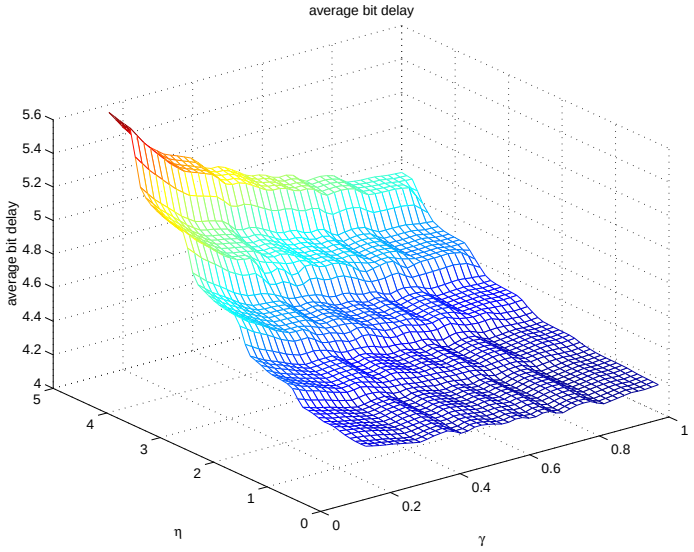

Figure 6-7: average bit delay for $p = 0.1$, $\rho = 0.75$

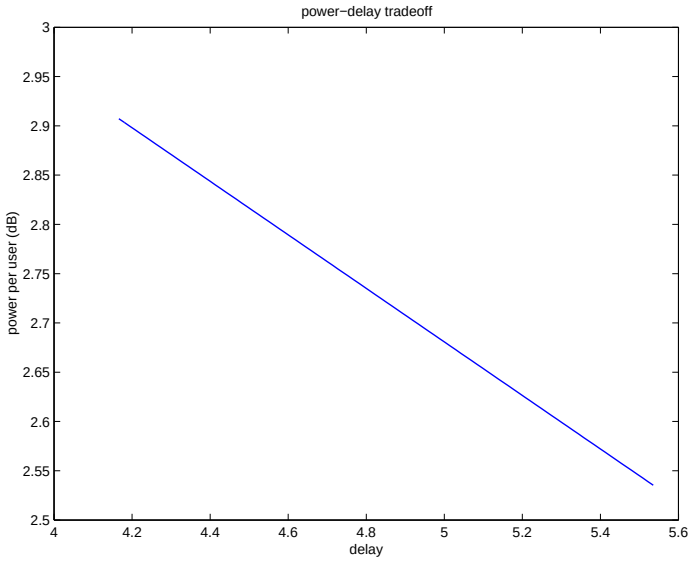

Figure 6-8: average bit delay for $p = 0.1$, $\rho = 0.75$

from the information theoretical literature along with those from ALOHA. We next proposed coding schemes that use limited queue information to better afford this trade-off. We did not attempt to optimally design such schemes, because it appears to be somewhat of a difficult problem. Instead, we proposed coding schemes that allow a larger class of design parameters that perhaps could be exposed to higher layers. This inter-layer tunability appears to be necessary when the breakdown of the source-channel separation theorem takes place so that applications may better use the communication system to achieve desired quality of service. Allowing for buffering and variable reliably received rates has enabled us to allow a larger class of QoS requirements to be addressed. If we address this type of coding system in terms of a wireless networking system, the scheme proposed above allows for users to prioritize data, where stringent delay-constrained data may be sent reliably without the need for retransmission via low-resolution components, and data which is not as delay-constrained may be reliably sent via high-resolution, which possibly may need retransmissions. We consider the parameter α in our system to be tunable and exposed to higher layers, so that it may match with refinable source encoders to afford better possible end-to-end performance. Listed below are several areas for further work:

- How do unequal arrival rates affect the system?
- Should we use different threshold crossings depending on the direction of the crossing? For instance, while in multiple access, should users continue to empty their queues until they are all empty?
- What may be the benefits of a policy with more possible modes, relying on more information? In particular, would it benefit much to exploit these modes to yield different power control strategies while being within each? It would also be very interesting to incorporate the cost of broadcasting network management information. This would characterize how meaningful extra state information by introducing power cost penalties for conveying more of it. Including this with the power consumption would help to illustrate that.

Appendix A

Steady-State Probability Equations For Queue Lengths

$$
\begin{aligned}
SNR'_{LR,1} &= \frac{\alpha_1 \beta P_1}{P_2 + (1-\alpha_1\beta)P_1 + \sigma_N^2} \\
SNR_{LR,2} &= \frac{\alpha_2 P_2^2}{(1-\alpha_2)P_2 + (1-\alpha_1\beta)P_1 + \sigma_N^2} \\
SNR''_{LR,1} &= \frac{\alpha_1(1-\beta)P_1}{(1-\alpha_2)P_2 + (1-\alpha_1)P_1 + \sigma_N^2} \\
SNR_{HR,1} &= \frac{(1-\alpha_1)P_1}{\sigma_N^2} \\
SNR_{HR,2} &= \frac{(1-\alpha_2)P_2}{\sigma_N^2} \\
R_{LR,1} &= \frac{1}{2}\log_2(1+SNR'_{LR,1}) + \frac{1}{2}\log_2(1+SNR''_{LR,1}) \\
R_{HR,1} &= \frac{1}{2}\log_2(1+SNR_{HR,1}) \\
R_{LR,2} &= \frac{1}{2}\log_2(1+SNR_{LR,2}) \\
R_{HR,2} &= \frac{1}{2}\log_2(1+SNR_{HR,2})
\end{aligned}
$$

$$
\begin{aligned}
\mu_{MA,1} + \mu_{MA,2} &= C_{\sigma_N^2} \left(\sum_{j=1}^{M} P_j \right) T \\
\mu_{LR,1}(\alpha_1) &= \mu_{LR,1} T \\
\mu_{HR,1}(\alpha_1) &= \mu_{HR,1} T \\
\mu_T(\alpha_1) &= \mu_{LR,1}(\alpha_1) + \mu_{HR,1}(\alpha_1) \\
\mu_{LR,2}(\alpha_2) &= \mu_{LR,2} T \\
\mu_{HR,2}(\alpha_2) &= \mu_{HR,2} T \\
\mu_{MA,1} &= \mu_{MA,1} T \\
\mu_{MA,2} &= \mu_{MA,2} T
\end{aligned}
$$

$$
\begin{aligned}
\bullet\ \pi(0,0) &= (1-p_1)(1-p_2)\Bigg[\pi(0,0) + \sum_{1\le i\le \mu_{LR,1}(\alpha_1)}\ \sum_{1\le j\le \mu_{LR,2}(\alpha_2)} \pi(i,j) + \\
&\qquad \sum_{1\le i\le \mu_T(\alpha_1)} \pi(i,0) + \sum_{1\le j\le \mu_T(\alpha_2)} \pi(0,j)\Bigg] \\
\bullet\ \underline{Q} &\in (1,1) : (L_1 - \mu_{LR,1}(\alpha_1), L_2 - \mu_{LR,2}(\alpha_2)) \ \cup \\
&\quad (L_1 - \mu_{LR,1}(\alpha_1), 1) : (\infty, L_2 - \mu_{LR,2}(\alpha_2)) \ \cup \\
&\quad (1, L_2 - \mu_{LR,2}(\alpha_2)) : (L_1 - \mu_{LR,1}(\alpha_1), \infty) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big) \\
\bullet\ \underline{Q} &\in (L_1 - \mu_{LR,1}(\alpha_1) + 1, 1) : (\infty, L_2 - \mu_{LR,2}(\alpha_2)) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 + (\mu_{LR,2}(\alpha_2))\big) \\
\bullet\ \underline{Q} &\in (1, L_2 - \mu_{LR,2}(\alpha_2) + 1) : (L_1 - \mu_{LR,1}(\alpha_1), \infty) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + (\mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) \\
\bullet\ \underline{Q} &\in (L_1 - \mu_{LR,1}(\alpha_1) + 1, L_2 - \mu_{LR,2}(\alpha_2) + 1) : (\eta_1 - \mu_{MA,1}, \eta_2 - \mu_{MA,2}) \ \cup \\
&\quad (\eta_1 - \mu_{MA,1} + 1, L_2 - \mu_{LR,2}(\alpha_2) + 1) : (\infty, \eta_2 - \mu_{MA,2}) \ \cup \\
&\quad (L_1 - \mu_{LR,1}(\alpha_1) + 1, \eta_2 - \mu_{MA,2} + 1) : (\eta_1 - \mu_{MA,1}, \infty) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\quad p_1p_2\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) \\
\bullet\ \underline{Q} &\in (1,0) : (L_1 - \mu_T(\alpha_1), 0) : \\
\pi\big(Q_1, 0\big) &= (1-p_1)(1-p_2)\left[\pi\big(Q_1 + \mu_T(\alpha_1), 0\big) + \pi\big(Q_1 + \mu_{LR,1}(\alpha_1), \mu_{LR,2}(\alpha_2)\big)\right]
\end{aligned}
$$

$$
\begin{aligned}
\bullet\, \underline{Q} &\in (L_1 - \mu_T(\alpha_1) + 1, 0) : (L_1 - \mu_{LR,1}(\alpha_1), 0) : \\
\pi\big(Q_1, 0\big) &= (1-p_1)(1-p_2)\left[\pi\big(Q_1 + \mu_T(\alpha_1), 0\big) + \pi\big(Q_1 + \mu_{LR,1}(\alpha_1), \mu_{LR,2}(\alpha_2)\big)\right] + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_T(\alpha_1)), \mu_{LR,2}(\alpha_2)\big) \\
\bullet\, \underline{Q} &\in (L_1 - \mu_{LR,1}(\alpha_1) + 1, 0) : (\infty, 0) : \\
\pi\big(Q_1, 0\big) &= (1-p_1)(1-p_2)\left[\pi\big(Q_1 + \mu_T(\alpha_1), 0\big) + \pi\big(Q_1 + \mu_{LR,1}(\alpha_1), \mu_{LR,2}(\alpha_2)\big)\right] + \\
&\quad p_1(1-p_2)\Bigg[\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), \mu_{LR,2}(\alpha_2)\big) + \\
&\qquad \pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)) + \mu_{HR,1}(\alpha_1), 0\big)\Bigg] \\
\bullet\, \underline{Q} &\in (0, 1) : (0, L_2 - \mu_T(\alpha_2)) : \\
\pi\big(0, Q_2\big) &= (1-p_1)(1-p_2)\left[\pi\big(0, Q_2 + \mu_T(\alpha_2)\big) + \pi\big(\mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big)\right] \\
\bullet\, \underline{Q} &\in (0, L_2 - \mu_T(\alpha_2) + 1) : (0, L_2 - \mu_{LR,2}(\alpha_2)) : \\
\pi\big(0, Q_2\big) &= (1-p_1)(1-p_2)\left[\pi\big(0, Q_2 + \mu_T(\alpha_2)\big) + \pi\big(\mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big)\right] + \\
&\quad (1-p_1)p_2\pi\big(\mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_T(\alpha_2))\big) \\
\bullet\, \underline{Q} &\in (0, L_2 - \mu_{LR,2}(\alpha_2) + 1) : (0, \infty) : \\
\pi\big(0, Q_2\big) &= (1-p_1)(1-p_2)\left[\pi\big(0, Q_2 + \mu_T(\alpha_2)\big) + \pi\big(\mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big)\right] + \\
&\quad (1-p_1)p_2\Bigg[\pi\big(\mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\qquad \pi\big(0, Q_2 - (L_2 - \mu_{LR,2}(\alpha_2)) + \mu_{HR,2}(\alpha_2)\big)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\bullet\, \underline{Q} \;&\in\; (\eta_1 - \mu_{MA,1} + 1, \eta_- \mu_{MA,+} 1) : (\eta_1 - \mu_{LR,1}(\alpha_1), \eta_2 - \mu_{LR,2}(\alpha_2)) \;\cup \\
&\quad (\eta_1 - \mu_{LR,1}(\alpha_1) + 1, \eta_2 - \mu_{MA,2} + 1) : (\infty, \eta_2 - \mu_{LR,2}(\alpha_2)) \;\cup \\
&\quad (\eta_1 - \mu_{MA,1} + 1, \eta_2 - \mu_{LR,2}(\alpha_2) + 1) : (\eta_1 - \mu_{LR,1}(\alpha_1), \infty) : \\
\pi\big(Q_1, Q_2\big) \;&=\; (1-p_1)(1-p_2)\left[\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \pi\big(Q_1 + \mu_{MA,1}, Q_2 + \mu_{MA,2}\big)\right] + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\quad p_1 p_2 \pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) \\
\bullet\, \underline{Q} \;&\in\; (\eta_1 - \mu_{LR,1}(\alpha_1) + 1, \eta_2 - \mu_{LR,2}(\alpha_2) + 1) : (\eta_1, \eta_2) \;\cup \\
&\quad (\eta_1 + 1, \eta_2 - \mu_{LR,2}(\alpha_2) + 1) : (\infty, \eta_2) \;\cup \\
&\quad (\eta_1 - \mu_{LR,1}(\alpha_1) + 1, \eta_2 + 1) : (\eta_1, \infty) : \\
\pi\big(Q_1, Q_2\big) \;&=\; (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{MA,1}, Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\quad p_1 p_2 \pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) \\
\bullet\, \underline{Q} \;&\in\; (\eta_1 + 1, \eta_2 + 1) : (\eta_1 + L_1 - \mu_{MA,1}, \eta_2 + L_2 - \mu_{MA,2}) : \\
\pi\big(Q_1, Q_2\big) \;&=\; (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{MA,1}, Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\quad p_1 p_2 \pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) \\
\bullet\, \underline{Q} \;&\in\; (\eta_1 + L_1 - \mu_{MA,1} + 1, \eta_2 + 1) : (\eta_1 + L_1 - \mu_{LR,1}(\alpha_1), \eta_2 + L_2 - \mu_{MA,2}) : \\
\pi\big(Q_1, Q_2\big) \;&=\; (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{MA,1}, Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\Big[\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\qquad \pi\big(Q_1 - (L_1 - \mu_{MA,1}), Q_2 + \mu_{MA,2}\big)\Big] + \\
&\quad p_1 p_2 \pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big)
\end{aligned}
$$

$$
\begin{aligned}
\bullet\, \underline{Q} \;&\in\; (\eta_1+1, \eta_2+L_2-\mu_{MA,2}+1) : (\eta_1+L_1-\mu_{MA,1}, \eta_2+L_2-\mu_{LR,2}(\alpha_2)) : \\
\pi\big(Q_1,Q_2\big) \;&=\; (1-p_1)(1-p_2)\pi\big(Q_1+\mu_{MA,1}, Q_2+\mu_{MA,2}\big) + \\
&\quad\; p_1(1-p_2)\pi\big(Q_1+(L_1-\mu_{LR,1}(\alpha_1)), Q_2-\mu_{LR,2}(\alpha_2)\big) + \\
&\quad\; p_1p_2\pi\big(Q_1-(L_1-\mu_{LR,1}(\alpha_1)), Q_2-(L_2-\mu_{LR,2}(\alpha_2))\big) + \\
&\quad\; (1-p_1)p_2\Big[\pi\big(Q_1+\mu_{LR,1}(\alpha_1), Q_2-(L_2-\mu_{LR,2}(\alpha_2))\big) + \\
&\qquad\qquad \pi\big(Q_1+\mu_{MA,1}, Q_2-(L_2-\mu_{MA,2})\big)\Big] \\
\bullet\, \underline{Q} \;&\in\; (\eta_1+L_1-\mu_{LR,1}(\alpha_1)+1, \eta_2+1) : (\infty, \eta_2+L_2-\mu_{MA,2}) : \\
\pi\big(Q_1,Q_2\big) \;&=\; (1-p_1)(1-p_2)\pi\big(Q_1+\mu_{MA,1}, Q_2+\mu_{MA,2}\big) + \\
&\quad\; p_1(1-p_2)\pi\big(Q_1-(L_1-\mu_{MA,1}), Q_2+\mu_{MA,2}\big) + \\
&\quad\; (1-p_1)p_2\left[\pi\big(Q_1+\mu_{LR,1}(\alpha_1), Q_2-(L_2-\mu_{LR,2}(\alpha_2))\big)\right] + \\
&\quad\; p_1p_2\pi\big(Q_1-(L_1-\mu_{MA,1}), Q_2-(L_2-\mu_{MA,2})\big)
\end{aligned}
$$

$$
\begin{aligned}
\bullet\ \underline{Q} &\in (\eta_1+1, \eta_2+L_2-\mu_{LR,2}(\alpha_2)+1) : (\eta_1+L_1-\mu_{MA,1}, \infty) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1+\mu_{MA,1}, Q_2+\mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\pi\big(Q_1-(L_1-\mu_{LR,1}(\alpha_1)), Q_2+\mu_{LR,2}(\alpha_2)\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1+\mu_{MA,1}, Q_2-(L_2-\mu_{MA,2})\big) + \\
&\quad p_1p_2\pi\big(Q_1-(L_1-\mu_{MA,1}), Q_2-(L_2-\mu_{MA,2})\big) \\
\bullet\ \underline{Q} &\in (\eta_1+L_1-\mu_{MA,1}+1, \eta_2+L_2-\mu_{MA,2}+1) : \\
&\quad (\eta_1+L_1-\mu_{LR,1}(\alpha_1), \eta_2+L_2-\mu_{LR,2}(\alpha_2)) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1+\mu_{MA,1}, Q_2+\mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\Big[\pi\big(Q_1-(L_1-\mu_{LR,1}(\alpha_1)), Q_2+\mu_{LR,2}(\alpha_2)\big) + \\
&\qquad \pi\big(Q_1-(L_1-\mu_{MA,1}), Q_2+\mu_{MA,2}\big)\Big] + \\
&\quad p_1p_2\Big[\pi\big(Q_1-(L_1-\mu_{LR,1}(\alpha_1)), Q_2-(L_2-\mu_{LR,2}(\alpha_2))\big) + \\
&\qquad \pi\big(Q_1-(L_1-\mu_{MA,1}), Q_2-(L_2-\mu_{MA,2})\big)\Big] + \\
&\quad (1-p_1)p_2\Big[\pi\big(Q_1+\mu_{LR,1}(\alpha_1), Q_2-(L_2-\mu_{LR,2}(\alpha_2))\big) + \\
&\qquad \pi\big(Q_1+\mu_{MA,1}, Q_2-(L_2-\mu_{MA,2})\big)\Big]
\end{aligned}
$$

$$
\begin{aligned}
\bullet\ \underline{Q} &\in (\eta_1 + L_1 - \mu_{LR,1}(\alpha_1) + 1, \eta_2 + L_2 - \mu_{MA,2} + 1) : (\infty, \eta_2 + L_2 - \mu_{LR,2}(\alpha_2)) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{MA,1}, Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_{MA,1}), Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1 p_2 \Bigg[\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\qquad \pi\big(Q_1 - (L_1 - \mu_{MA,1}), Q_2 - (L_2 - \mu_{MA,2})\big)\Bigg] + \\
&\quad (1-p_1)p_2 \Bigg[\pi\big(Q_1 + \mu_{LR,1}(\alpha_1), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\qquad \pi\big(Q_1 + \mu_{MA,1}, Q_2 - (L_2 - \mu_{MA,2})\big)\Bigg] \\
\bullet\ \underline{Q} &\in (\eta_1 + L_1 - \mu_{MA,1} + 1, \eta_2 + L_2 - \mu_{LR,2}(\alpha_2) + 1) : (\eta_1 + L_1 - \mu_{LR,1}(\alpha_1), \infty) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{MA,1}, Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\Bigg[\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 + \mu_{LR,2}(\alpha_2)\big) + \\
&\qquad \pi\big(Q_1 - (L_1 - \mu_{MA,1}), Q_2 + \mu_{MA,2}\big)\Bigg] + \\
&\quad p_1 p_2 \Bigg[\pi\big(Q_1 - (L_1 - \mu_{LR,1}(\alpha_1)), Q_2 - (L_2 - \mu_{LR,2}(\alpha_2))\big) + \\
&\qquad \pi\big(Q_1 - (L_1 - \mu_{MA,1}), Q_2 - (L_2 - \mu_{MA,2})\big)\Bigg] + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + \mu_{MA,1}, Q_2 - (L_2 - \mu_{MA,2})\big) \\
\bullet\ \underline{Q} &\in (\eta_1 + L_1 - \mu_{LR,1}(\alpha_1) + 1, \eta_2 + L_2 - \mu_{LR,2}(\alpha_2) + 1) : (\infty, \infty) : \\
\pi\big(Q_1, Q_2\big) &= (1-p_1)(1-p_2)\pi\big(Q_1 + \mu_{MA,1}, Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1(1-p_2)\pi\big(Q_1 - (L_1 - \mu_{MA,1}), Q_2 + \mu_{MA,2}\big) + \\
&\quad p_1 p_2\pi\big(Q_1 - (L_1 - \mu_{MA,1}), Q_2 - (L_2 - \mu_{MA,2})\big) + \\
&\quad (1-p_1)p_2\pi\big(Q_1 + \mu_{MA,1}, Q_2 - (L_2 - \mu_{MA,2})\big)
\end{aligned}
$$

Appendix B

Another Hybrid Broadcast/Multiple Access Coding Mechanism with Less Virtual Codewords

B.1 Multiple Access Rate-Splitting followed by Broadcast Rate-Splitting

The results in [MMH+02] combines concepts from *multiple-access communications* [Ahl71, Lia72]; *broadcast channels* [Cov72, Cov75, Cov98]; and *rate splitting* [RU96] to combat burstiness in a multiple-user AWGN channel. The basic idea behind this approach springs from the following observation. In multiple access channels, capacity is achieved through rate splitting. This involves first constructing *virtual users* that share available rate and power and that transmit independently. The receiver then decodes the received signals consecutively, so that some users are regarded as noise to other users during decoding. After a user is decoded, the user's contribution to the signal is eliminated, and the noise for the remaining undecoded signal is reduced. A similar approach is taken to achieve capacity in the degraded

AWGN broadcast channel. For broadcast AWGN channels, we superimpose two codes, a low resolution and a high resolution code. The low resolution code is decoded by considering the high resolution code as noise. Once the low resolution code is decoded, its contribution is eliminated. Hence, there is a similarity between the decoding mechanism used to achieve capacity in multiple-access channels and that used in degraded broadcast channels. In the system considered here, a user codes to transmit over two possible channels: a channel with the other user present and a channel without the other user. Thus, the problem bears some traits of both degraded broadcast channels and of multiple access channels. We exploit this fact along with our observations to construct our capacity-achieving coding scheme. For the model considered in [MMH+02], rate splitting is performed as follows: We divide user 1 into two independent users, U_1' and U_1'', which send independent WGN signals with variance $\beta\sigma_1^2$ and $(1-\beta)\sigma_1^2$, respectively. There is no rate splitting for user 2, which maps to a single user, U_2. As in broadcast channels, each of the users we have constructed sends two messages on two separate signals. That is, U_1' sends signals LR_1' and HR_1', which are independent WGN signals with variance $\alpha_1'\beta\sigma_1^2$ and $(1-\alpha_1')\beta\sigma_1^2$, respectively. U_1'' sends signal LR_1'' and HR_1'', which are independent WGN signals with variance $\alpha_1''(1-\beta)\sigma_1^2$ and $(1-\alpha_1'')(1-\beta)\sigma_1^2$, respectively. U_2 sends signal LR_2 and HR_2, which are independent WGN signals with variance $\alpha_2\sigma_2^2$ and $(1-\alpha_2)\sigma_2^2$, respectively. Each $\alpha_1, \alpha_2, \beta$ lies in $[0,1]$. Figure B-1 illustrates this coding scheme.

The notations LR and HR are the abbreviations of *low resolution* and *high resolution*, respectively, since we are in effect using a broadcast code within our multiple access scheme. We decode signals one after the other in the following order:

$$\text{First } \mathrm{LR}_1', \text{ then } \mathrm{LR}_2, \mathrm{LR}_1'', \mathrm{HR}_1'', \mathrm{HR}_2, \text{ and finally } \mathrm{HR}_1'. \tag{B.1}$$

If one of the six signals is not present, the receiver proceeds to the next one. Each signal is decoded so that all signals not yet decoded are considered noise, and signals that have been decoded and reconstructed are cancelled. Here we assume the

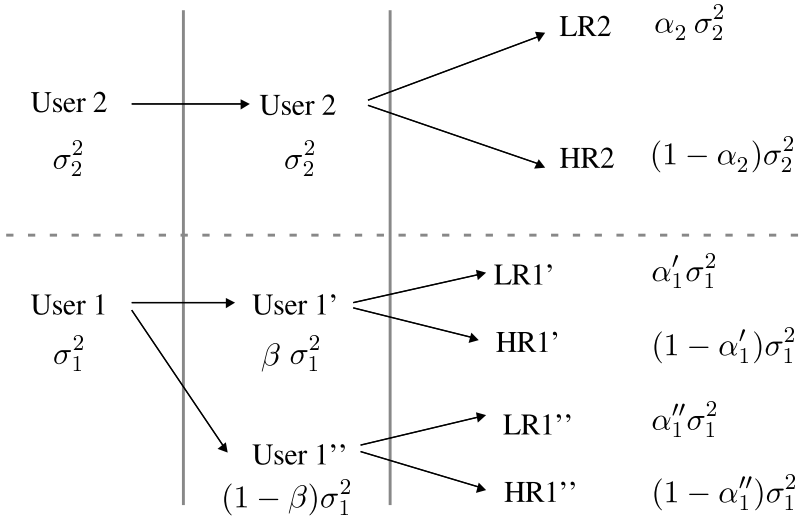

Figure B-1: Representation of the coding scheme.

signal can be decoded with over a time slot of length T with probability of error less than or equal to ξ.

We may now present the three possible cases that arise and the corresponding decoding rules. Each signal of the LR and HR type has a rate such that it can be decoded within the required probability of error if the SNR is at least: Each signal of the LR and HR type has a rate such that it can be decoded within the required probability of error if the SNR is at least:

$$\begin{array}{ll}
\dfrac{\alpha_1'\beta\sigma_1^2}{\sigma_2^2+(1-\alpha_1'\beta)\sigma_1^2+\sigma_N^2} & \text{for } \mathrm{LR}_1' \\
\dfrac{\alpha_2\sigma_2^2}{\sigma_2^2(1-\alpha_2)+(1-\alpha_1'\beta)\sigma_1^2+\sigma_N^2} & \text{for } \mathrm{LR}_2 \\
\dfrac{\alpha_1''(1-\beta)\sigma_1^2}{\sigma_2^2(1-\alpha_2)+(1-\alpha_1'\beta-\alpha_1''(1-\beta)))\sigma_1^2+\sigma_N^2} & \text{for } \mathrm{LR}_1'' \\
\dfrac{(1-\alpha_1'')(1-\beta)\sigma_1^2}{(1-\alpha_1')\beta\sigma_1^2+\sigma_N^2} & \text{for } \mathrm{HR}_1'' \\
\dfrac{(1-\alpha_2)\sigma_2^2}{\sigma_N^2} & \text{for } \mathrm{HR}_2 \\
\dfrac{\sigma_1^2\beta(1-\alpha_1')}{\sigma_N^2} & \text{for } \mathrm{HR}_1'
\end{array}$$

Our coding and decoding scheme is defined so that all LR signals above will always have a sufficiently large SNR. These signals are therefore always received reliably. For the HRs, they will not have sufficient SNR if user 1 and user 2 send at the same

time. If the minimum SNR is not met for any one of the HR signals, that signal is not decoded. We consider the following cases:

Case 1: Only user 2 transmits.

- First, we decode LR_2, which yields a rate of $\frac{1}{2}\log\left(1+\frac{\alpha_2\sigma_2^2}{\sigma_2^2(1-\alpha_2)+(1-\alpha_1'\beta)\sigma_1^2+\sigma_N^2}\right)$.
- Next, we decode signal HR_2, which yields a rate of $\frac{1}{2}\log\left(1+\frac{(1-\alpha_2)\sigma_2^2}{\sigma_N^2}\right)$.

The total rate is the sum of the above two rates.

Case 2: Only user 1 transmits.

- First, we decode LR_1', which yields a rate of $\frac{1}{2}\log\left(1+\frac{\alpha_1'\beta\sigma_1^2}{\sigma_2^2+(1-\alpha_1'\beta)\sigma_1^2+\sigma_N^2}\right)$.
- Second, we decode signal LR_1'', which yields a rate of $\frac{1}{2}\log\left(1+\frac{\alpha_1''(1-\beta)\sigma_1^2}{\sigma_2^2(1-\alpha_2)+(1-\alpha_1'\beta-\alpha_1''(1-\beta))\sigma_1^2+\sigma_N^2}\right)$.
- Third, we decode the signal HR_1'', which yields a rate of $\frac{1}{2}\log\left(1+\frac{(1-\alpha_1'')(1-\beta)\sigma_1^2}{(1-\alpha_1')\beta\sigma_1^2+\sigma_N^2}\right)$.
- Finally, we decode HR_1', yielding rate $\frac{1}{2}\log\left(1+\frac{\sigma_1^2\beta(1-\alpha_1')}{\sigma_N^2}\right)$.

The total rate is the sum of the above four rates.

Case 3: User 1 and 2 both transmit.

- First, we decode LR_1', yielding rate $\frac{1}{2}\log\left(1+\frac{\alpha_1'\beta\sigma_1^2}{\sigma_2^2+(1-\alpha_1'\beta)\sigma_1^2+\sigma_N^2}\right)$.
- Second, we decode LR_2, which yields a rate of $\frac{1}{2}\log\left(1+\frac{\alpha_2\sigma_2^2}{\sigma_2^2(1-\alpha_2)+(1-\alpha_1'\beta)\sigma_1^2+\sigma_N^2}\right)$.
- Third, we decode signal LR_1'', which yields a rate of $\frac{1}{2}\log\left(1+\frac{\alpha_1''(1-\beta)\sigma_1^2}{\sigma_2^2(1-\alpha_2)+(1-\alpha_1'\beta-\alpha_1''(1-\beta))\sigma_1^2+\sigma_N^2}\right)$.

The total rate for user 1 is the sum of the rates of LR_1' and LR_1''. The total rate for user 2 is the rate LR_2.

Case 4: Neither user transmits, so the total rate is 0.

B.2 Broadcast Rate-Splitting followed by Multiple Access Rate-Splitting

The proposed scheme that follows uses similar types of rate-splitting, but in reverse order. Namely, a broadcast rate-splitting approach is performed across both users with ratios α_1 and α_2. Afterwards, the low-resolution rate of one of the two users is further rate-split β_1 into two more virtual users for multiple access purposes. This yields a total of 5 virtual codewords rather than the aforementioned amount of 6. The basic idea in this approach is the realization that the high-resolution user for each is user is thought of to *only be reliably received when the other user is not present*. Hence, it does not appear worthwhile to rate-split according to multiple access before rate-splitting according to the high-resolution user, according to the structure of the problem.

In this approach, we reverse the order of the rate-splitting and do not rate-split multiple access for the high-resolution users. As in broadcast channels, each of the users we have constructed sends two messages on two separate signals. That is, U_1 sends signals LR_1 and HR_1, which are independent WGN signals with variance $\alpha_1\sigma_1^2$ and $(1-\alpha_1)\sigma_1^2$, respectively. U_2 sends signal LR_2 and HR_2, which are independent WGN signals with variance $\alpha_2\sigma_2^2$ and $(1-\alpha_2)\sigma_2^2$, respectively. The low-resolution component of user 1 is then divided into two independent virtual users for multiple access purposes. Hence, we have $U_{LR1'}$ and $U_{LR1''}$, which send independent WGN signals with variance $\beta_1\alpha_1\sigma_1^2$ and $(1-\beta_1)\alpha_1\sigma_1^2$, respectively. There is no rate splitting for the low-resolution component of user 2. Each $\alpha_1, \alpha_2, \beta_1$ lies in $[0, 1]$ Figure B-2 illustrates this coding scheme.

The notations LR and HR are the abbreviations of *low resolution* and *high resolution*, respectively, since we are in effect using a broadcast code for the initial rate splitting. We decode signals one after the other in the following order:

$$\text{First } \widetilde{LR}_1', \text{ then } \widetilde{LR}_2, \widetilde{LR}_1'', \widetilde{HR}_1, \text{ and finally } \widetilde{HR}_2. \tag{B.2}$$

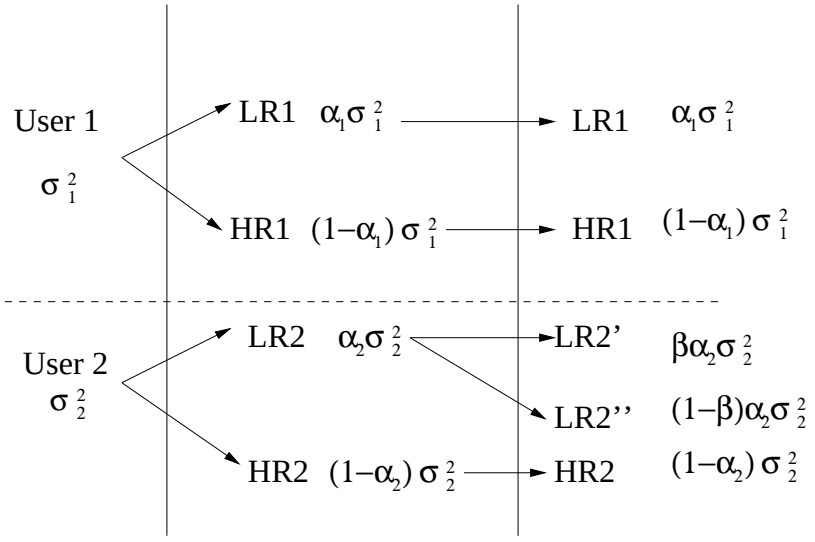

Figure B-2: Representation of the proposed new coding scheme.

If one of the five signals is not present, the receiver proceeds to the next one. Each signal is decoded so that all signals not yet decoded are considered noise, and signals that have been decoded and reconstructed are cancelled. Here we assume the signal can be decoded with over a time slot of length T with probability of error less than or equal to ξ.

Each signal of the LR and HR type has a rate such that it can be decoded within the required probability of error if the SNR is at least:

$$
\begin{aligned}
SNR'_{\widetilde{LR},1} &= \frac{\widetilde{\alpha}_1\beta\sigma_1^2}{\sigma_2^2 + (1-\widetilde{\alpha}_1\beta)\sigma_1^2 + \sigma_N^2} && \text{for } \widetilde{\mathrm{LR}}_1' \\
SNR_{\widetilde{LR},2} &= \frac{\widetilde{\alpha}_2\sigma_2^2}{(1-\widetilde{\alpha}_2)\sigma_2^2 + (1-\widetilde{\alpha}_1\widetilde{\beta})\sigma_1^2 + \sigma_N^2} && \text{for } \widetilde{\mathrm{LR}}_2 \\
SNR''_{\widetilde{LR},1} &= \frac{\widetilde{\alpha}_1(1-\widetilde{\beta})\sigma_1^2}{(1-\widetilde{\alpha}_2)\sigma_2^2 + (1-\widetilde{\alpha}_1)\sigma_1^2 + \sigma_N^2} && \text{for } \widetilde{\mathrm{LR}}_1'' \\
SNR_{\widetilde{HR},1} &= \frac{(1-\widetilde{\alpha}_1)\sigma_1^2}{\sigma_N^2} && \text{for } \widetilde{\mathrm{HR}}_1 \\
SNR_{\widetilde{HR},2} &= \frac{(1-\widetilde{\alpha}_2)\sigma_2^2}{\sigma_N^2} && \text{for } \widetilde{\mathrm{HR}}_2
\end{aligned}
$$

Hence, the low-resolution and high-resolution rates that may be achieved are as follows:

$$\widetilde{R_{LR,1}} = \frac{1}{2}\log_2(1+SNR'_{\widetilde{LR},1}) + \frac{1}{2}\log_2(1+SNR''_{\widetilde{LR},1})$$
$$\widetilde{R_{HR,1}} = \frac{1}{2}\log_2(1+SNR_{\widetilde{HR},1})$$
$$\widetilde{R_{LR,2}} = \frac{1}{2}\log_2(1+SNR_{\widetilde{LR},2})$$
$$\widetilde{R_{HR,2}} = \frac{1}{2}\log_2(1+SNR_{\widetilde{HR},2})$$

It can be shown that any rate achieved w/ the aforementioned scheme can be achieved with this scheme. Consider a set of allocations of $\alpha'_1, \alpha''_1, \alpha_2, \beta$ for the first scheme. We may choose $\widetilde{\alpha}_1, \widetilde{\alpha}_2, \widetilde{\beta}$ so that

$$\widetilde{R_{LR,1}} = R_{LR,1}$$
$$\widetilde{R_{HR,1}} = R_{HR,1}$$
$$\widetilde{R_{LR,2}} = R_{LR,2}$$
$$\widetilde{R_{HR,2}} = R_{HR,2}$$

Proof:

Let us first set $\widetilde{R_{LR,2}} = R_{LR,2}$ by setting

$$\widetilde{\alpha}_2 = \alpha_2. \tag{B.3}$$

Next, we may attempt to set $\widetilde{R_{HR,1}} = R_{HR,1}$. Note that from the properties of the AWGN channel capacity function, $C_{\sigma_N^2}(P_1) + C_{\sigma_N^2+P_1}(P_2) = C_{\sigma_N^2}(P_1+P_2)$, where $C_N(x) = \frac{1}{2}\log_2\left(1+\frac{x}{N}\right)$. Hence,

$$R_{HR,1} = R'_{HR,1} + R''_{HR,1} = C_{\sigma_N^2}\left((\beta(1-\alpha'_1) + (1-\beta)(1-\alpha''_1))\,\sigma_1^2\right)$$

and we may allow $\widetilde{R_{HR,1}} = R_{HR,1}$ by setting $1-\widetilde{\alpha}_1 = \beta(1-\alpha'_1) + (1-\beta)(1-\alpha''_1)$,

or equivalently,

$$\widetilde{\alpha}_1 = (1-\beta)\alpha_1'' + \beta\alpha_1'. \tag{B.4}$$

Next, we may attempt to set $\widetilde{R_{LR,2}} = R_{LR,2}$ by choose our final degree of freedom, $\widetilde{\beta}$, accordingly. Note that

$$\widetilde{R_{LR,2}} = C_{\sigma_N^2+(1-\widetilde{\alpha}_1)\sigma_1^2+(1-\widetilde{\alpha}_2)\sigma_2^2+\widetilde{\alpha_1}(1-\widetilde{\beta})\sigma_1^2}\left(\widetilde{\alpha}_2\sigma_2^2\right) \text{ and}$$
$$R_{LR,2} = C_{\sigma_N^2+(1-\alpha_1)\sigma_1^2+(1-\alpha_2)\sigma_2^2+(1-\beta)\alpha_1''\sigma_1^2}\left(\alpha_2\sigma_2^2\right)$$

Hence, we may equate these two rates by setting $\widetilde{\alpha_1}(1-\widetilde{\beta})\sigma_1^2 = (1-\beta)\alpha_1''\sigma_1^2$, or equivalently,

$$\widetilde{\beta} = \frac{\beta\alpha_1'}{(1-\beta)\alpha_1'' + \beta\alpha_1'}. \tag{B.5}$$

But we still must check that with these values of $\widetilde{\beta}, \widetilde{\alpha}_1$, and $\widetilde{\alpha_2}$, we in fact arrive at our final set of rates matching: $\widetilde{R_{LR,1}} = R_{LR,1}$. Let us now verify this:

$$\begin{aligned}
\widetilde{R_{LR,1}} &= C_{\sigma_N^2+\sigma_2^2+\widetilde{\alpha}_1(1-\widetilde{\beta})\sigma_1^2}\left(\widetilde{\alpha}_1\widetilde{\beta}\sigma_1^2\right) + C_{\sigma_N^2+(1-\widetilde{\alpha}_2)\sigma_2^2+(1-\widetilde{\alpha}_1)\sigma_1^2}\left(\widetilde{\alpha}_1(1-\widetilde{\beta})\sigma_1^2\right) \\
&= C_{\sigma_N^2+\sigma_2^2+\left(\widetilde{\alpha}_1(1-\widetilde{\beta})+(1-\widetilde{\alpha}_1)\right)\sigma_1^2}\left(\widetilde{\alpha}_1\widetilde{\beta}\sigma_1^2\right) + C_{\sigma_N^2+(1-\alpha_2)\sigma_2^2+(1-\widetilde{\alpha}_1)\sigma_1^2}\left(\widetilde{\alpha}_1(1-\widetilde{\beta})\sigma_1^2\right) \\
&= C_{\sigma_N^2+\sigma_2^2+\left(\alpha_1''(1-\beta)+(1-\widetilde{\alpha}_1)\right)\sigma_1^2}\left(\beta\alpha_1'\sigma_1^2\right) + C_{\sigma_N^2+(1-\alpha_2)\sigma_2^2+(1-\widetilde{\alpha}_1)\sigma_1^2}\left(\alpha_1''(1-\beta)\sigma_1^2\right) \\
&= C_{\sigma_N^2+\sigma_2^2+\left(\alpha_1''(1-\beta)+\beta(1-\alpha_1')+(1-\beta)(1-\alpha_1'')\right)\sigma_1^2}\left(\beta\alpha_1'\sigma_1^2\right) + \\
&\quad C_{\sigma_N^2+(1-\alpha_2)\sigma_2^2+\left(\beta(1-\alpha_1')+(1-\beta)(1-\alpha_1'')\right)\sigma_1^2}\left(\alpha_1''(1-\beta)\sigma_1^2\right) \\
&= C_{\sigma_N^2+\sigma_2^2+\left(1-\alpha_1'\beta\right)\sigma_1^2}\left(\beta\alpha_1'\sigma_1^2\right) + C_{\sigma_N^2+(1-\alpha_2)\sigma_2^2+\left(1-\alpha_1'\beta-\alpha_1''(1-\beta)\right)\sigma_1^2}\left(\alpha_1''(1-\beta)\sigma_1^2\right) \\
&= R_{LR,1}' + R_{LR,1}'' \\
&= R_{LR,1}
\end{aligned}$$

where the third equation is due to (B.3), the third equation is due to the fact that $\widetilde{\alpha}_1\widetilde{\beta} = \beta\alpha_1'$ and $\widetilde{\alpha}_1(1-\widetilde{\beta}) = \alpha_1''(1-\beta)$ (combine ((B.4) and (B.5)), the fourth is due to (B.4). Note that this collapses the previous $2(2M-1) = 4M-2$ virtual users

into $M + 2M - 1 = 3M - 1$ virtual users.

Bibliography

[Abr70] N. Abramson. The ALOHA system - another alternative for computer communications. *Fall Joint Computer Conference*, 1970.

[Abr92] N. Abramson. Fundamentals of packet multiple access for satellite networks. *IEEE Journal on Selected Areas in Communications*, 10(2):309–316, February 1992.

[Ahl71] R. Ahlswede. Multi-way communication channels. *ISIT*, pages 23–52, 1971.

[Ana91] V. Anantharam. The stability region of the finite-user, slotted ALOHA system protocol. *IEEE Transactions on Information Theory*, 37:535–540, 1991.

[Ber74] P. Bergmans. A simple converse for broadcast channels with additive white Gaussian noise. *IEEE Transactions on Information Theory*, 20(2):279–280, 1974.

[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.

[BGB97] A. M. Y. Bigloo, T. A. Gulliver, and V. K. Bhargava. Maximum-likelihood decoding and code combining for DS/SSMA slotted aloha. *IEEE Transactions on Communications*, 45(12):1602–1612, December 1997.

[BGT93] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error-correcting codes and decoding: Turbo codes. *Proc. IEEE International Communications Conference*, 1993.

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

[CLV00] G. Caire, E. Leonardi, and E. Viterbo. Modulation and coding for the Gaussian collision channel. *IEEE Transactions on Information Theory*, 46(6):2007–2026, September 2000.

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

[Cov75] T. M. Cover. An achievable rate region for the broadcast channel. *IEEE Transactions on Information Theory*, 21:399–404, 1975.

[Cov98] T. M. Cover. Comments on broadcast channels. *IEEE Transactions on Information Theory*, 44:2524–2530, 1998.

[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.

[Gal74] R. G. Gallager. Capacity and coding for degraded broadcast channels. *Probl. Inform. Transm.*, 10(3):185–193, 1974.

[Gal85] R. Gallager. A perspective on multiaccess channels. *IEEE Transactions on Information Theory*, 31:124–142, 1985.

[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.

[GS87] D. J. Goodman and A. A. M. Saleh. The near/far effect in local ALOHA radio communications. *IEEE Transactions on Vehicular Technology*, 36(1):19–27, February 1987.

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

[HvL82] B. Hajek and T. van Loon. Decentralized dynamic control of multiaccess broadcast channels. *IEEE Transactions on Automatic Control*, 3:559–569, June 1982.

[Kap83] M. Kaplan. A sufficient condition for non-ergodicity of a markov chain. *IEEE Transactions on Information Theory*, 25:470–471, 1983.

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

[MM85] J.L. Massey and P. Mathys. The collision channel without feedback. *IEEE Transactions on Information Theory*, 31:192–204, 1985.

[MMH+02] M. Médard, S. P. Meyn, J. Huang, A. J. Goldsmith, and T.P. Coleman. Capacity of time-slotted ALOHA packetized multiple-access systems. Submitted to IEEE Transactions on Wireless Communications, 2002.

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

[Pip81] N. Pippenger. Bounds on the performance of protocols for a multiple access broadcst channel. *IEEE Transactions on Information Theory*, 27:45–151, 1981.

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

[RU96] B. Rimoldi and R. Urbanke. A rate-splitting approach to the Gaussian multiple access channel. *IEEE Transactions on Information Theory*, 42(2):364–375, 1996.

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

[SS00] T. Shinomiya and H. Suzuki. Slotted ALOHA mobile packet communication systems with multiuser detection in a base station. *IEEE Transactions on Vehicular Technology*, 49(3):948–955, May 2000.

[Tar95] F. Tarköy. Information-theoretic aspects of spread ALOHA. *IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, pages 1318–1320, 1995.

[Tho00] G. Thomas. Capacity of the wireless packet collision channel without feedback. *IEEE Transactions on Information Theory*, 46(3):1141–1144, May 2000.

[TM78] B. S. Tsybakov and V. A. Mikhailov. Free synchronous packet access in a broadcast channel with feedback. *Problemy Peredachi Informatsii*, 14:32–59, 1978.

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

[Wol78] J. Wolfowitz. *Coding Theorems of Information Theory*. Springer-Verlag, Berlin, Germany, 1978.