Throughput Measures for Delay-Constrained Communications in Fading Channels *

Nadeem Ahmed and Richard G. Baraniuk

Dept. of Electrical and Computer Engineering Engineering
Rice University
Houston, TX 77005

Abstract

Fading channels, often seen in wireless systems, provide an unfavorable environment for reliable communications. Current methods for evaluating the performance of fading channels include ergodic capacity and ϵ-capacity. Ergodic capacity quantifies the ultimate reliable communication limit of the fading channel. It is only achievable with infinite coding delay, making it impossible to achieve in practice. ϵ-capacity, achievable with finite coding delay, does not provide a measure of error-free communications performance. Since practical communication systems are delay-constrained, it is possible to retransmit codewords when errors occur. We provide a new analysis framework that accounts for codeword retransmission in the analysis of fading channels. We introduce new measures, maximum zero-outage throughput and maximum ϵ-throughput, that predict the performance of practical systems and show that ergodic capacity and ϵ-capacity are special cases of our definitions. We also provide a measure that characterizes the performance of a system with more complex receiver design, using "incremental diversity" to improve throughput.

1 Introduction

Fading channels, common in wireless systems, are a particularly hostile environment for reliable communications and can adversely affect achievable information rates. The transmitted signal is scattered, in a time-varying manner, along the transmission path resulting in random fluctuations in the received power level, or *fading*. The widespread use of wireless systems makes understanding the limits of fading channels imperative.

Transmitters map information to codewords that add redundancy to the data and protect it from distortion due to fading and other noise. The *coding delay* is proportional to the length of the codeword and measured in terms of the number of fading states affecting each codeword. The highest reliable data rate achievable over a fading channel is provided by its *ergodic capacity* [8]. Ergodic capacity is an asymptotic result; it is achievable only with infinite coding delay by using infinite length codewords that are affected by infinitely many fading states and capture the ergodic nature of the channel.

Practical communication systems are *delay-constrained* and must use finite-length codes. Ergodic capacity is not achievable for such systems. The need to quantify the performance of delay-constrained systems gave rise to the notions of *outage* and *ϵ-capacity* [11]. An outage event occurs when the attempted transmission rate is greater than the *instantaneous capacity*, the capacity for a particular set of fading states affecting a codeword. Outages result in decoding errors at the receiver. ϵ-capacity is the maximum rate that can be sustained with the probability of outage no greater than ϵ.

*This research is supported in part by NSF, AFOSR, ONR, DARPA, and the Texas Instruments Leadership University Program. Email: [nahmed,richb]@rice.edu, Web: http://dsp.rice.edu.

There is a shortcoming with the current measures of performance for fading channels. Ergodic capacity bounds the performance of any communication system. However, it is only achievable for systems with infinite coding delay and is not an accurate reflection of the performance of practical delay-constrained systems. For delay-constrained systems, ϵ-capacity does not provide an estimate of error-free communication performance. Setting $\epsilon = 0$ results in an error-free estimate known as *delay-limited capacity*. However, delay-limited capacity is 0 for many common fading processes. This is clearly an ineffective estimate of communication performance as practical delay-limited systems in use today reliably communicate with non-zero data rates.

The main purpose of this paper is to propose a new framework for analyzing the performance of practical systems over fading channels. Engineers are interested in *throughput*, the effective data rate, of systems. A communications system transmitting at a rate of 100Kbps that loses 20% of codewords to outages, has a throughput of 80Kbps. It is the throughput rather than the transmission rate that more accurately reflects the performance of this system. Our goal is to find the maximum throughput that can be transmitted through a fading channel with finite length codewords, which is more representative of how systems are designed and used today. In practice, delay-limited codewords lost to outages can be retransmitted to ensure successful decoding at the receiver. Using this idea we provide a novel perspective for the performance of wireless systems, by treating the system as a single server queue and relating the throughput of the system to the transmission rate and the codeword service time. By doing so, we can more accurately predict the performance of practical systems in fading channels than current measures.

Section 2 provides background on the system model and common notions of capacity in fading channels. Maximizing throughput is described in Section 3. The case where outages are not tolerated, maximum zero-outage throughput, is in Section 4. Section 5 defines the maximum throughput with a more intelligent receiver design using *incremental diversity*. Section 6 details maximum ϵ-throughput, the case where outage probability ϵ is tolerated. We illustrate that several common notions of capacity for fading channels, ergodic capacity and ϵ-capacity, are in fact special cases of our more general definitions. We summarize our contributions in Section 7.

2 System Model and Background

The random nature of fading channels changes the notion of capacity from the conventional sense [2]. The *instantaneous capacity*, the capacity for a particular instance of the channel, itself becomes random. The *ergodic capacity* of the system depends on the amount of channel state information (CSI) available to the transmitter and receiver, which depends on the system design. For each design the ergodic capacity can differ. Our focus is on the configuration with receiver CSI. A review of capacity in fading channels can be found in [9]. Throughout this paper we use capacity and spectral efficiency, the capacity per unit bandwidth, interchangeably.

2.1 System Model

In many applications the channel conditions change on a time scale that is much slower than the communications signalling. This motivates the discrete-time block-fading additive white Gaussian noise model (BF-AWGN) [13]. Each "block" represents the channel *coherence time*, the amount of time the channel is assumed constant, and corresponds the the transmission of N symbols. We assume the channel fading is a random process and is i.i.d. from block-to-block. The system in the k^{th} block can be modelled as

$$\underline{y_k} = \underline{x_k} h_k + \underline{w_k}, \tag{1}$$

Figure 1: (a) Channel model; (b) queuing model

with $\underline{y_k}, \underline{x_k} \in \mathbb{R}^N$ representing the system output and input. We assume a Gaussian noise process, $\underline{w_k} \sim \mathcal{N}(0, \mathbf{I}^N)$. Scattering by the environment results in reflections of the transmitted signal adding constructively or destructively with the original signal. The multipath interference due to scattering is represented by a random multiplicative gain, $h_k \in \mathbb{R}$, on the transmitted signal (See Figure 1). In this work, we assume that $|h_k|^2$ follow a χ_2^2 (chi-squared with 2 degrees of freedom) distribution. This model is commonly used for wireless communication systems without line-of-sight (LOS) between transmitter and receiver.

If the average transmitted power in the k^{th} block is $\mathcal{P}_{\text{av}}$ then the random *fading* $|h_k|^2$ results in a random received power of $|h_k|^2\mathcal{P}_{\text{av}}$. Constructive interference results in a large $|h_k|^2$ and thus a large received signal power that is conducive to communication; we term this situation a "good" fade. Destructive interference results in a small $|h_k|^2 \approx 0$ and thus a small received signal power that is not conducive to communication; we term this situation a "bad" fade.

Codewords span K blocks of the BF-AWGN channel, containing KN symbols encoded at rate R nats/sec/Hz. The *coding delay*, the amount of time required to encode data, is quantified by K, the number of blocks of the BF-AWGN channel a codeword spans. A system is considered *delay-constrained* if $K < \infty$.

2.1.1 Ergodic Capacity

In communication systems that can tolerate infinite coding delay, infinite-length codewords can be used. Such codewords are affected by infinitely many fading levels of the time-varying channel, and can "average" out is effect. With infinite coding delay, data can be transmitted reliably as long as it is encoded at a rate less than *ergodic capacity* [8]. If only the receiver has CSI the ergodic capacity is given by

$$C_{\text{ergodic}} := E_{|h|^2}\left[\log(1 + |h|^2\mathcal{P}_{\text{av}})\right], \tag{2}$$

where the subscript of fading process $|h_k|^2$ has been dropped since $K = \infty$ and all codewords are affected by all possible channel fades with the probability of each fading level determined by the fading distribution (χ_2^2 in our case). Ergodic capacity can be achieved by using random coding at the transmitter with input symbols $\sim \mathcal{N}(0, \sqrt{\mathcal{P}_{\text{av}}})$.

The infinite coding delay required to achieve ergodic capacity does not make it achievable for practical systems that have finite coding delay. However, it does serve as a useful upper bound on the performance of any communication system.

2.1.2 ϵ-Capacity

To quantify the performance of delay-constrained systems, the concept of outage [11] has been introduced. An outage occurs when the transmission rate is larger than the *instantaneous capacity*, the highest reliable data rate that can be transmitted over the set of channel fades affecting a codeword. Outages result in decoding errors at the receiver. For codewords that span K blocks of the BF-AWGN channel, the instantaneous capacity is given by $\frac{1}{K}\sum_{i=1}^{K}\log(1 + |h_i|^2\mathcal{P}_{\text{av}})$ and the *outage probability* can be written

$$P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K) = \text{Prob}[R > \frac{1}{K}\sum_{i=1}^{K}\log(1 + |h_i|^2\mathcal{P}_{\text{av}})]. \tag{3}$$

This leads to the notion of ϵ-capacity, the largest rate that can be sustained over all channel states except a subset with probability ϵ. The ϵ-capacity,

$$C_\epsilon = \sup \{R : P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K) \leq \epsilon\}, \tag{4}$$

is not a measure of error-free communication performance as there is a non-zero outage probability. By setting $\epsilon = 0$, we have an error-free estimate of performance known as *delay-limited capacity*[1], since the number of blocks in codewords is finite [9]. The delay-limit capacity is 0 when $\min(|h|^2) = 0$ is in the support of the fading process [9]. This is the case for many common fading distributions, including χ_2^2 fading.

Clearly, conventional measures for delay-constrained systems have short-comings. ϵ-capacity does not give an estimate of reliable communications performance. Delay-limited capacity, predicts zero rate for many common situations and is clearly an overly conservative estimate of performance.

3 Throughput

Currently, there is no general framework for determining how well practical, delay-constrained systems perform in fading channels. Ergodic capacity characterizes the ultimate achievable rate for delay unconstrained ($K = \infty$) systems and serves as an upper bound for practical, delay-constrained ($K < \infty$) systems. For delay-constrained systems, ϵ-capacity does not provide measure of error-free performance and delay-limited capacity yields a pessimistic estimate of performance.

The conventional measures of communication performance over fading channels characterize the amount of information that can be transmitted if codewords are transmitted on *only once*. Delay-constrained systems can transmit each codeword more than once, since codewords have finite-length. For practical systems, codeword retransmission can be used to reduce or eliminate errors. Engineers are interested in the *throughput*, or effective data rate, of communication systems. A measure that quantifies the throughput of a system and factors codeword retransmission is more representative of how communication systems in use today are designed and used.

We assume a delay-less, error-free feedback channel used to relay retransmission requests to the transmitter. One bit of feedback is required per codeword (to retransmit or not), and since codewords span K blocks, retransmission requires $\frac{1}{K}$ bits per block. As $K \to \infty$, feedback goes to 0. The key idea behind retransmission is that the random channel cannot stay in a bad state indefinitely, and eventually after a (possibly infinite) number of retransmission attempts, the instantaneous capacity will be higher than the attempted transmission rate, allowing successful transmission. This approach can be used to bound the performance of practical systems which retransmit codewords in error, TCP/IP or ARQ for example.

In the next section we describe the modelling of communication system with a queue. Using this model we relate the throughput of the system to the transmission rate and codeword service time.

3.1 Throughput and Codeword Service Time

The number of transmission attempts needed to send a codeword is a random variable due to the random nature of the channel. The system can be thought of as a single server queue (see

[1]Not to be confused with the delay-limited capacity of [10], which uses power control and applies to the situation with both transmitter and receiver CSI

Figure 1(b)) with expected codeword service time $E[S]$, and service rate $\frac{1}{E[S]}$. The service time distribution depends on the nature of the retransmission scheme. In general, the probability that a codeword will take s transmission attempts is

$$\text{Prob}(S=s) = \text{Prob}\left[\bigcap_{i=1}^{s-1} \text{outage}_i\right]\left[1 - \text{Prob}(\text{outage}_s | \bigcap_{i=1}^{s-1} \text{outage}_i)\right], \tag{5}$$

which is the probability it was in outage in the first $s-1$ attempts multiplied by the probability it is successfully received on the s^{th} attempt conditioned on the fact it was previously in error.

We use (5) to determine $E[S]$. Minimizing $E[S]$ (process codewords as fast as possible) can be accomplished by setting $R=0$, since there will be no outages and no retransmissions. However, the amount of information codewords contain, R (bits/sec/Hz), must be considered. We have an engineering tradeoff, trying to keep R as high as possible while keeping $E[S]$ as small as possible. We define the maximum throughput of a practical system as

$$T(\mathcal{P}_{\text{av}}, K) = \sup_R \frac{R}{E[S]}. \tag{6}$$

The service time distribution depends on the nature of the communication system and the retransmission chosen. Therefore, (6) represents the highest average throughput achievable for a particular retransmission scheme. The rest of the paper considers some special cases.

4 Maximum Zero-Outage Throughput

We introduce a new metric, *maximum zero-outage throughput* (MZT), for evaluating practical system performance in fading channels. When outages occur, the receiver requests that the codewords be retransmitted, ensuring all are successfully received. With this scheme all transmission attempts have the same probability of success. As a result, (5) becomes

$$\text{Prob}(S=s) = [P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]^{s-1}[1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)] \tag{7}$$

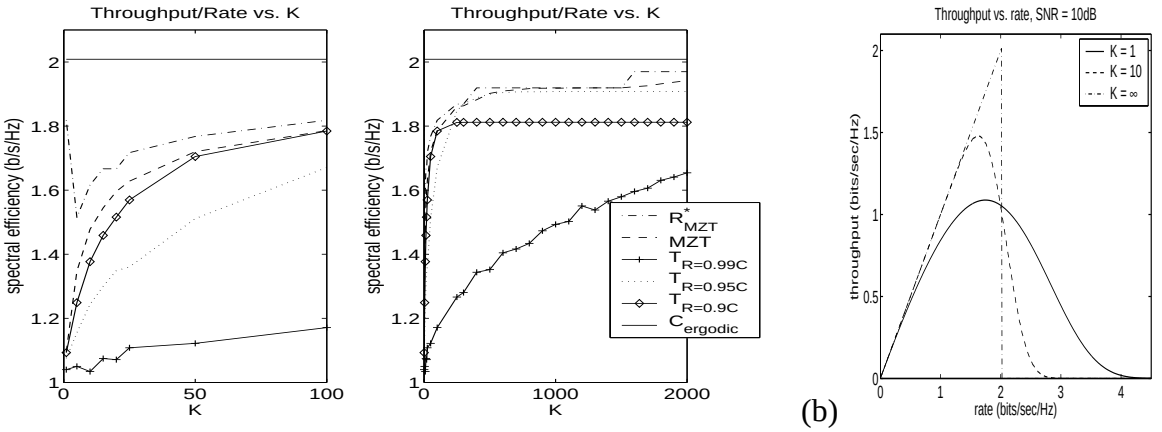
with the service time distribution being geometric, on the positive integers, with parameter $[1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]$. Therefore $E[S] = \frac{1}{1-P_{\text{out}}(R,\mathcal{P}_{\text{av}},K)}$ and the throughput is

$$\text{MZT}(\mathcal{P}_{\text{av}}, K) = \sup_R R[1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)], \tag{8}$$

where the outage probability is (3). Figure 2(b) plots $R[1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]$ for various K and assuming that the fading $|h|^2$ follows a χ^2_2 distribution. The peak of each curve represents $\text{MZT}(\mathcal{P}_{\text{av}}, \text{K})$.

In general (8) is difficult to solve explicitly, as we do not have a closed form expression for the outage probability (3). However, a semi-explicit solution for (8) can be found when $K=1$. Substituting (3) into (8) we optimize to find the maximum value. The argument that maximizes throughput is the solution to $R2^R = \frac{\mathcal{P}_{\text{av}}}{\log_e(2)}$, a transcedental function that always has a unique solution, namely $R^*_{MZT} = \frac{\mathcal{W}(\mathcal{P}_{\text{av}})}{\log_e(2)}$, where $\mathcal{W}$ is *Lambert's W function*. MZT can be found by using R^*_{MZT} in (8).

Figure 2(a) shows MZT vs. K and the optimal transmission rate, R^*_{MZT}, vs. K. MZT clearly grows with K, while the relationship between R^*_{MZT} and K is not clear. R^*_{MZT} generally increases with K but not monotonically and can fluctuate greatly, especially for small K. Often system designers use codewords with a large coding delay, K, in the hope that it is "large enough" to capture the ergodic nature of the channel. If the ergodic nature of the channel is



(a) (b)

Figure 2: $\gamma = 10$dB. (a) MZT vs. K, R^*_{MZT} vs. K and effective throughput for $R = \beta C_{\mathrm{ergodic}}$ vs. K for $\beta \in \{0.9, 0.95, 0.99\}$. The plot on the left is for $1 \leq K \leq 100$, and the right for $1 \leq K \leq 2000$. Transmitting at R^*_{MZT} rather than $R = \beta C_{\mathrm{ergodic}}$ yields significant throughput gains. For $R = 0.99 C_{\mathrm{ergodic}}$, the ergodicity assumption does not hold even for $K = 2000$;(b) Throughput vs. R for various K. MZT is the maximum of each curve.

---

truly captured, then for $R < C_{\mathrm{ergodic}}$ there are zero outages, and the throughput is equivalent to the transmission rate. Figure 2(a) also shows the effective throughput if $R = \beta C_{\mathrm{ergodic}}$ for $\beta \in \{0.9, 0.95, 0.99\}$. Since these transmission rates are less than ergodic capacity, the coding delay is "large enough" if the throughput is close to these transmission rates. More specifically, the curves should plateau when the coding delay, K, is "large enough" to capture the ergodic nature of the channel. For $\beta = 0.9$ the system is close to ergodic for $K \approx 300$ and for $\beta = 0.95$ this point moves to $K \approx 700$. For $\beta = 0.99$ the system cannot be considered ergodic even for $K = 2000$. The closer the attempted transmission rate is to C_{ergodic}, the harder it is to make the ergodic assumption.

The simulations show the danger of assuming ergodicity when it is not present. If the assumption is false throughput can suffer greatly, seen by the significant difference in the MZT and effective throughput curves. For given K, throughput optimization should be performed to achieve the MZT of the system. We look $K = 25$ as an example. Here $R^*_{MZT} = 1.717$ b/s/Hz results in an MZT of 1.626 b/s/Hz. Transmitting at $R = 0.99 C_{\mathrm{ergodic}} = 1.995$ b/s/Hz yields a throughput of 1.074 b/s/Hz. Throughput optimization yields a 51.4% improvement, over assuming ergodicity, in effective data rate.

Looking at Figure 2(a) and 3(a), we see that for large K, MZT approaches C_{ergodic}. This suggests the following properties.

Theorem 1. *The outage probability, $P_{\mathrm{out}}(R, \mathcal{P}_{\mathrm{av}}, K)$, converges asymptotically, as $K \to \infty$, to the indicator function, $I(R, C_{\mathrm{ergodic}})$, which is 1 if $R > C_{\mathrm{ergodic}}$ and 0 if $R < C_{\mathrm{ergodic}}$.*

Proof. Using Chebyshev's inequality, for $R < C_{\mathrm{ergodic}}$, we bound (3) by $P_{\mathrm{out}}(R, \mathcal{P}_{\mathrm{av}}, K) \leq \frac{\alpha}{K}$. For $R > C_{\mathrm{ergodic}}$, $P_{\mathrm{out}}(R, \mathcal{P}_{\mathrm{av}}, K) \geq 1 - \frac{\alpha}{K}$. In both cases α is a constant. Then by taking $K \to \infty$ we have

$$P_{\mathrm{out}}(R, \mathcal{P}_{\mathrm{av}}, K) = I(R, C_{\mathrm{ergodic}}) \tag{9}$$

□

Theorem 2. MZT *converges to ergodic capacity as $K \to \infty$.*

Proof. By taking $K \to \infty$ and using (9), we arrive at

$$\mathrm{MZT}(\mathcal{P}_{\mathrm{av}}, \infty) = \sup_R R[1 - I(R, C_{\mathrm{ergodic}})] = C_{\mathrm{ergodic}} \tag{10}$$

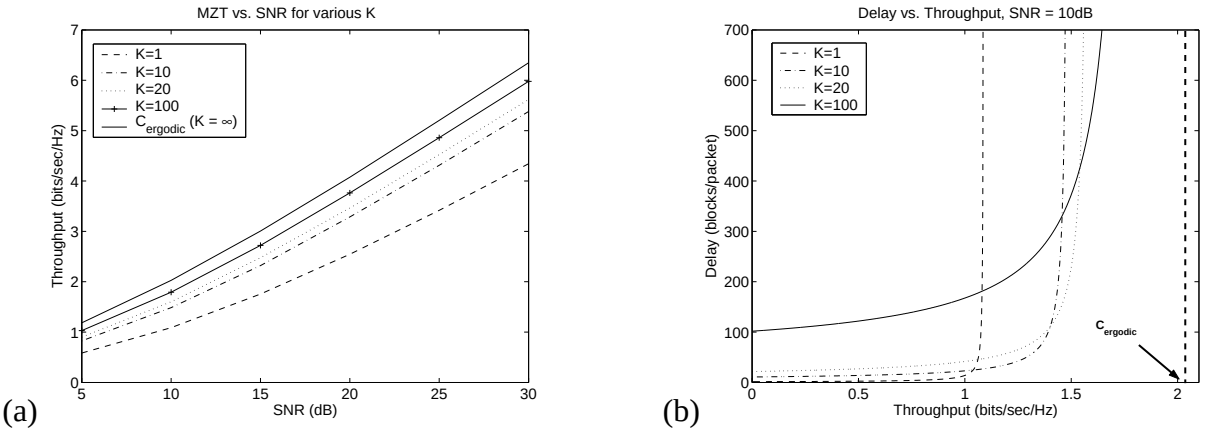


Figure 3: (a) MZT vs. SNR for $K \in \{1, 10, 20, 100\}$. As $K \to \infty$, $\mathrm{MZT} \to C_{\mathrm{ergodic}}$; (b) Delay vs. Throughput for $K \in \{1, 10, 20, 100\}$. Any throughput less than C_{ergodic} (dashed line) is achievable with finite queueing delay.

The maximization is trivial as $P_{\mathrm{out}}(R, \mathcal{P}_{\mathrm{av}}, K)$ takes only two values, 0 or 1. We see that MZT does converge to the C_{ergodic}, as $K \to \infty$. Note that $R^*_{\mathrm{MZT}} = \mathrm{MZT}(\mathcal{P}_{\mathrm{av}}, \infty) = C_{\mathrm{ergodic}}$. □

When $K = \infty$ it is impossible to get non-zero throughput if $R > C_{\mathrm{ergodic}}$. This is due to the fact that the outage probability function is the indicator function (See Theorem 1). However, when $K < \infty$ an interesting phenomenon is observed. In this case the outage probability, $P_{\mathrm{out}}(R, \mathcal{P}_{\mathrm{av}}, K)$, does not fall off abruptly as a function of R. This means that non-zero throughput can be achieved for any transmission rate.

Theorem 3. *Non-zero throughput is achievable for transmission rates higher than* C_{ergodic} *when* $K < \infty$.

Proof. Let $R = C_{\mathrm{ergodic}} + \epsilon$. By Theorem 1 we see that $P_{\mathrm{out}}(C_{\mathrm{ergodic}} + \epsilon, \mathcal{P}_{\mathrm{av}}, K) < 1$ for $K < \infty$. Using this inequality in (8), we see that the effective throughput, $R[1 - P_{\mathrm{out}}(C_{\mathrm{ergodic}} + \epsilon, \mathcal{P}_{\mathrm{av}}, K)] > 0$. □

From Figure 2(b), for $K = 1$ and $K = 10$, significant throughput is achieved for transmission rates well beyond C_{ergodic}. However as K grows, the throughput achieved when $R > C_{\mathrm{ergodic}}$ shrinks. This occurs because the larger the K, the more unlikely it is to see a sequence of K "good" fading states that support a rate higher than C_{ergodic}. When $K = \infty$, the system is ergodic and it is impossible to get a sequence of such states. Ergodicity, for $R < C_{\mathrm{ergodic}}$, improves throughput, while for $R > C_{\mathrm{ergodic}}$, limits it. In Figure 2(b), the system with the highest throughput beyond C_{ergodic} is the least ergodic one, $K = 1$. As $K \to \infty$ codewords will see all channel states in their exact proportion, and only rates lower than C_{ergodic} yield non-zero throughput.

In practical systems the random nature of the codeword service time may result in codewords arriving while another is being serviced. This results in a backlog of queued codewords awaiting transmission. For stable queuing systems, the average number of codewords in the queue must be finite. This queuing effect is not present in delay unconstrained systems transmitting at C_{ergodic}. For these systems, the delay is infinite as the transmitter must buffer an infinite amount of data to encode a single codeword that is transmitted for an infinite amount of time. There is no queueing as only a single codeword is ever transmitted, however the delay is infinite due to the codeword length.

The queueing effect for delay-constrained systems can be quantified. For example, if we assume a Poisson codeword arrival process and model the system as an M/G/1 queue [3], a closed form solution for the average queueing delay can be found. For such M/G/1 queues, the Pollaczek-Khinchin (P-K) theorem [3, 4] can be invoked to find the average total delay (sum of queueing delay and service time) for each codeword,

$$D_{\text{tot}} = K\left(\frac{\lambda E[S^2]}{2(1-\lambda E[S])} + E[S]\right) \tag{11}$$

where λ is the average arrival rate, and $E[S]$ and $E\left[S^2\right] = \frac{(1+P_{\text{out}}(R^*,\mathcal{P}_{\text{av}},K))}{(1-P_{\text{out}}(R^*,\mathcal{P}_{\text{av}},K))^2}$ are the first and second moments of the service time distribution. Scaling by K changes the units for the D_{tot} to blocks/codeword.

Figure 3(b) illustrates the throughput vs. delay profile for different K. We can infer that any rate below ergodic capacity can be achieved with finite average delay. Different models for the codeword arrival process result in different expressions for the queueing delay.

5 Incremental Diversity

MZT provides a measure of the throughput performance of practical systems using the simplest retransmission scheme. However, if we allow increased receiver functionality and a more complex retransmission scheme, it is possible to attain higher throughput than what MZT predicts. The receiver can save codewords lost to outages and use them along with subsequent retransmitted versions to jointly decode the data. The optimal method to combine the codewords, known as maximal ratio receive combining (MRRC), can improve the SNR of retransmitted codewords [5]. Since the number of codewords combined in the decoding process increases with the number of transmission attempts, we term the concept *incremental diversity* and the best throughput as *maximum zero-outage throughput-incremental diversity* ($\mathrm{MZT}_{\mathrm{ID}}$).

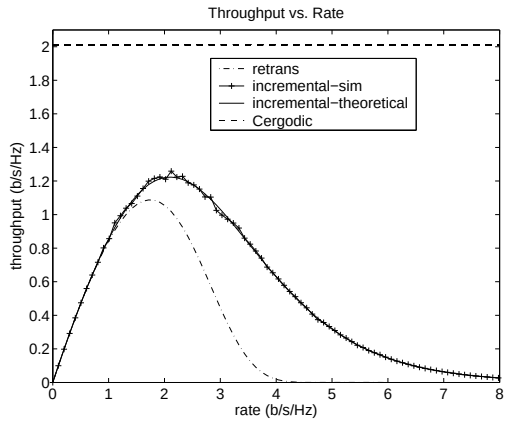
With this system configuration, the service time distribution for the server in the queueing model changes from Section 4. The probability of success on the s^{th} attempt does not have a simple form, due to MRRC, which makes the success probability dependent on the previous errors. However (5) can be used to determine the first and second moments of the service time distribution and determine $\mathrm{MZT}_{\mathrm{ID}}$ and its delay profile.

Although it is difficult to determine the service time distribution for arbitrary coding delays, an analytical expression can be obtained when $K = 1$. The probability that a codeword is in error on the first s attempts is $\mathrm{Prob}[\bigcap_{i=1}^{s} \mathrm{outage}_i] = \mathrm{Prob}[\sum_{i=1}^{s} \|h_i\|^2 \leq \alpha]$ where $\alpha = \frac{2^R-1}{\mathcal{P}_{\text{av}}}$. We realize that $\sum_{i=1}^{s} \|h_i\|^2$, a χ^2 random variable with $2s$ degrees of freedom, then by applying Bayes rule to (5) we have $\mathrm{Prob}(S = s) = e^{-\alpha}\frac{\alpha^{s-1}}{(s-1)!}$. This can be used to determine $E[S] = 1 + \alpha$ and $E[S^2] = \alpha^2 + 3\alpha + 1$. The throughput is then

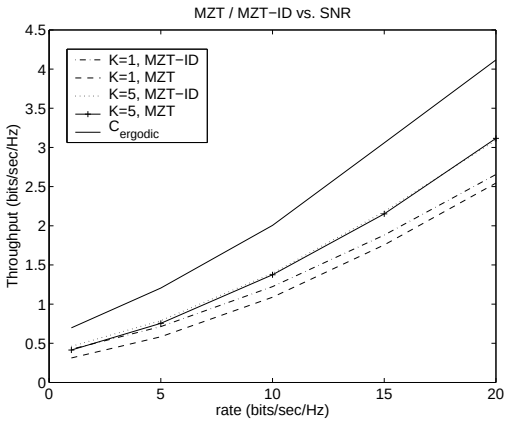
$$\mathrm{MZT}_{\mathrm{ID}}(\mathcal{P}_{\text{av}}, 1) = \sup_R \frac{R\mathcal{P}_{\text{av}}}{2^R + \mathcal{P}_{\text{av}} - 1}. \tag{12}$$

The argument which maximizes this throughput is the solution to $2^R(R\log_e(2) - 1) = \mathcal{P}_{\text{av}} - 1$, which can be shown to be $R^*_{\mathrm{MZT-ID}} = \frac{\mathcal{W}(\frac{\mathcal{P}_{\text{av}}-1}{e})+1}{\log_e(2)}$. The system is stable, meaning that we have finite queuing delay, for any attempted throughput less than $\mathrm{MZT}_{\mathrm{ID}}$. The average delay can be determined by using $E[S]$ and $E[S^2]$ in the P-K theorem. Figure 4(a) shows the performance of retransmission with incremental diversity against conventional retransmission. The theoretical throughput, and the simulated value is illustrated. We see that $\mathrm{MZT}_{\mathrm{ID}}$ is higher than MZT.

Theorem 4. *When $K = 1$ and assuming $|h|^2$ follows a χ^2_2 distribution,* $\mathrm{MZT}(\mathcal{P}_{\text{av}}, 1) \leq \mathrm{MZT}_{\mathrm{ID}}(\mathcal{P}_{\text{av}}, 1)$.



(a)



(b)

Figure 4: (a) Throughput vs. rate using simple retransmission and incremental diversity. The theoretical and simulated values are shown with $\gamma = 10$dB. Incremental diversity has higher throughput for all transmission rates; (b) MZT and $\mathrm{MZT}_{\mathrm{ID}}$ vs. SNR for $K = 1$ and $K = 5$. For smaller K, incremental diversity yields significant gains in throughput.

Proof. Let $\mathrm{MZT}(\mathcal{P}_{\mathrm{av}}, 1) = R_1[1 - P_{out}(R_1, \mathcal{P}_{\mathrm{av}}, 1)] = R_1 e^{-(\frac{2^{R_1}-1}{\mathcal{P}_{\mathrm{av}}})}$ and let $\mathrm{MZT}_{\mathrm{ID}}(\mathcal{P}_{\mathrm{av}}, 1) = \frac{R_2 \mathcal{P}_{\mathrm{av}}}{2^{R_2}+\mathcal{P}_{\mathrm{av}}-1}$, where $R_1 = R^*_{\mathrm{MZT}}$ and $R_2 = R^*_{\mathrm{MZT-ID}}$. Then,

$$\begin{aligned} \log \mathrm{MZT}(\mathcal{P}_{\mathrm{av}}, 1) &= \log R_1 - (\frac{2^{R_1}-1}{\mathcal{P}_{\mathrm{av}}}) \leq \log R_1 - \log(\frac{2^{R_1}-1+\mathcal{P}_{\mathrm{av}}}{\mathcal{P}_{\mathrm{av}}}) \\ &\leq \log R_2 - \log(\frac{2^{R_2}-1+\mathcal{P}_{\mathrm{av}}}{\mathcal{P}_{\mathrm{av}}}) = \log \mathrm{MZT}_{\mathrm{ID}}(\mathcal{P}_{\mathrm{av}}, 1). \end{aligned} \quad (13)$$

where the first inequality comes from the fact that $x \geq \log(1+x), \forall x \geq 0$ as well as $R_1, R_2 > 0$ and $\mathcal{P}_{\mathrm{av}} > 0$. Since $\log(x)$ is a monotonic increasing function of x, $\mathrm{MZT}(\mathcal{P}_{\mathrm{av}}, 1) \leq \mathrm{MZT}_{\mathrm{ID}}(\mathcal{P}_{\mathrm{av}}, 1)$, completing the proof. □

Figure 4(b) shows both MZT and $\mathrm{MZT}_{\mathrm{ID}}$ vs. SNR for $K = 1$ and $K = 5$. Like MZT, the larger the K, the closer $\mathrm{MZT}_{\mathrm{ID}}$ is to ergodic capacity. This is intuitive as the transmitter functionality is identical to the system in Section 4.

The difference between the systems in Sections 4 and 5 is the receiver functionality. We can see that for $K = 1$, there is a significant gain in throughput by using the MRRC at the receiver, while for $K = 5$ the gain is still present but not as large. The MRRC increases throughput when there are larger numbers of retransmission attempts for each codeword, which occurs when outage are more likely, for smaller K (Recall Theorem 1, which shows that outage probability goes to 0 for $R < C_{\mathrm{ergodic}}$ as $K \to \infty$).

6 Maximum ϵ-Throughput

Systems making a possibly infinite number of transmission attempts per codeword can be generalized to one which makes at most L attempts. Doing so can result in a tighter bound on codeword delay and *jitter*, the variance of the delay. Many applications, such as video and voice, are delay and jitter sensitive; excessive amounts of either can render data useless. If L is finite, every codeword is not guaranteed to arrive successfully as some may require more than L attempts. We will illustrate this concept by using the retransmission scheme of Section 4 as a closed form expression exists for the service time distribution. We define *maximum ϵ-throughput* ($\mathrm{M}\epsilon\mathrm{T}$) as the highest throughput with ϵ probability of outage. Clearly as $L \to \infty$, $\mathrm{M}\epsilon\mathrm{T}$ becomes MZT.

The probability of successful transmission on the s^{th} attempt is (7) for $s \leq L$, and 0 for $s > L$. The service time, for $s \leq L$, follows a geometric distribution, with parameter $[1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]$. The mass in the tail of the geometric distribution for $s = L+1, \ldots, \infty$, represents the overall outage probability and is $[P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]^L$. The expected service time is $E[S] = \frac{1-[P_{\text{out}}(R,\mathcal{P}_{\text{av}},K)]^L}{[1-P_{\text{out}}(R,\mathcal{P}_{\text{av}},K)]}$, and the $\text{M}\epsilon\text{T}$ is then defined as

$$\text{M}\epsilon\text{T}(\mathcal{P}_{\text{av}}, K) = \sup_R \left\{ R \frac{(1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K))}{1 - [P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]^L} : P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)^L \leq \epsilon \right\} \tag{14}$$

Note if $L = 1$, that the definition for $\text{M}\epsilon\text{T}$ becomes precisely of ϵ-capacity in (4). Clearly ϵ-capacity is a special case of $\text{M}\epsilon\text{T}$. If we limit ourselves to $L_{\max}$ transmission attempts, we find that for each value of $L \in \{1, 2, \ldots, L_{\max}\}$, there is a different value of R which maximizes the throughput subject to the outage constraint. In general, we must perform the optimization for all $L \in \{1, 2, \ldots, L_{\max}\}$ and pick the best throughput from among the $L_{\max}$ candidates.

The benefit of allowing multiple transmission attempts is potentially higher throughput for a given ϵ. For example, if $\epsilon = 0.05$, $K = 1$ and $\mathcal{P}_{\text{av}} = 10$dB then for $L = 1$, $C_\epsilon = \text{M}\epsilon\text{T}(\mathcal{P}_{\text{av}}, 1) = 0.413$ nats/sec/Hz, while for $L = 2$, $\text{M}\epsilon\text{T}(\mathcal{P}_{\text{av}}, 1) = 1.031$ nats/sec/Hz. The throughput benefit of allowing multiple transmission attempts, for the same outage constraint, can be significant.

The cost of the improved throughput is queueing delay that is not present for ϵ-capacity. As in Section 4, the delay can be quantified. If we assume a Poisson codeword arrival process, then using $E[S]$ and $E[S^2] = \frac{(1-[P_{\text{out}}(R,\mathcal{P}_{\text{av}},K)]^L)[1+P_{\text{out}}(R,\mathcal{P}_{\text{av}},K)]}{[1-P_{\text{out}}(R,\mathcal{P}_{\text{av}},K)]^2} - 2L[P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]^L$, the delay can be computed by the P-K theorem. The delay and variance of the service time, $\sigma_s = E[S^2] - E[S]^2$, is smaller than a system with zero outage, since $L < \infty$. This makes a limited retransmission system, transmitting at $\text{M}\epsilon\text{T}$, attractive for delay and jitter sensitive applications.

7 Conclusion

We presented a framework for analyzing the communications performance of practical systems in fading channels. Communications engineers are interested in the throughput, or effective data rate, of a system. We posed the throughput problem in terms of a single server queue and provided a novel framework in which throughput is related to the transmission rate and average codeword service time. In this manner, we gain a zero-outage measures of communications performance for fading channels for any type of fading distribution. This is not possible with the conventional analysis that limits the number of transmission attempts to one, and predicts a delay-limited capacity of zero in many situations of interest.

Using this framework we illustrated some special cases based on different system designs. In the first configuration, the transmitter codes codewords spanning K blocks and the receiver sends retransmission requests. At most L transmission attempts are attempted. In the case $L = \infty$, we defined maximum zero-outage throughput (MZT) as, as the highest throughput achievable with zero outage. We obtained a semi-explicit solution when $K = 1$, and showed that asymptotically as $K \to \infty$, MZT approaches C_{ergodic}. When $L < \infty$, we defined maximum ϵ-throughput ($\text{M}\epsilon\text{T}$) for at most L transmission attempts with outage probability ϵ. We showed that ϵ-capacity is a special case of $\text{M}\epsilon\text{T}$, when $L = 1$.

In the second configuration the transmitter again codes over K blocks, but the receiver functionality is different. It saves codewords that are in error, rather than discarding them, and combines them with subsequent retransmissions to jointly decode information. We termed this incremental diversity, and defined maximum zero-outage throughput with incremental diversity (MZT_{ID}) as the highest throughput with this configuration. We obtained a semi-explicit

solution for $K = 1$. Simulations show that $\mathrm{MZT}_{\mathrm{ID}}$ is greater than MZT for practical systems, and we prove this result explicitly for $K = 1$.

Simulations showed the significant performance gains obtained by optimally selecting the transmission rate when $K < \infty$. We illustrated that the closer the transmission rate is to C_{ergodic}, the harder it is to assume ergodicity. The notion of capacity is "softened" when $K < \infty$; non-zero throughput can be obtained for transmission rates higher than C_{ergodic}. The delay that is introduced with retransmission based systems is an average delay constraint. Rather than having a hard limit on the amount of time that is spent in the the queueing system, as would be the case for delay-limited capacity, we have an average delay of D_{tot} blocks for each codeword. By loosening the delay constraint we gain throughput measures for all fading distributions, that provide a more realistic measure of performance of practical systems in fading channels.

References

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

[2] T. Cover and J. Thomas, "Elements of Information Theory," *New York: Wiley, 1991.*

[3] L. Kleinrock, "Queuing Systems - Volume I: Theory,"*New York: Wiley, 1975.*

[4] D. Bertsekas and R. Gallager, "Data Networks 2nd ed,"*New Jersey: Prentice Hall, 1992.*

[5] T.S. Rappaport, *Wireless Communications. Principles and Practices*. New Jersey: Prentice Hall, 1996.

[6] R. Berry and R. Gallager, "Communication over Fading Channels with Delay Constraints," *IEEE Trans. Info. Theory., v. 48, no. 5, May 2002.*

[7] J. Wolfowitz, "Coding Theorems of Information Theory, 2nd ed.," *New York: Springer-Verlag, 1964.*

[8] A. Goldsmith and P. Varaiya, "Capacity of Fading Channels with Channel Side Information," *IEEE Trans. Info. Theory., v. 43, no. 6, Nov. 1997.*

[9] E. Biglieri, J. Proakis and S. Shamai, " Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Trans. Comm., v. 44, no. 6, Oct. 1998.*

[10] S. Hanly and D. Tse, " Multiaccess Fading Channels - Part II: Delay-Limited Capacities," *IEEE Trans. Info. Theory., v. 44, no. 7, Nov. 1998.*

[11] S. Ozarow, S. Shamai and A. Wyner, " Information Theoretic Considerations for Cellular Mobile Radio," *IEEE Trans. Vehic. Tech., v. 43, no. 2, May 1994.*

[12] I. Bettesh and S. Shamai, "Queueing Analysis of the Single User Fading Channel," *in Proc. 21st IEEE Conv. Electrical and Electronic Engineers in Israel, 2000 pp. 274-277*

[13] R. J. McEliece and W. E. Stark, " Channels with Block Interference," *IEEE Trans. Info. Theory., v. 30, no. 1, Jan. 1984.*

[14] G. Taricco, E. Biglieri and G. Caire, "Limiting Performance of Block-Fading Channels with Multiple Antennas," *in Proc. ITW 1999, pp. 27-29*

[15] G. Caire, G. Taricco and E. Biglieri, " Optimum Power Control Over Fading Channels," *IEEE Trans. Info. Theory., v. 45, no. 5, Jul. 1999.*

[16] E. Teletar and R. G. Gallager, "Combining Queueing Theory with Information Theory for Multiaccess," *IEEE Jour. Sel. Areas in Comm., v.13, no.6, Aug. 1995.*