Throughput Maximization for ARQ-like Systems in Fading Channels with Coding and Queuing Delay Constraints

Nadeem Ahmed and Richard G. Baraniuk

***Abstract*— Practical delay-limited communication systems often employ retransmission algorithms such as ARQ to ensure reliable communications in fading channels. Maximizing the communications throughput in such systems can cause excessive queueing delays due to the random number of retransmission attempts required for each codeword. In this paper we consider the problem of delay-limited throughput maximization with a constraint on the expected waiting-time, which incorporates both queueing and coding delays. We explore the trade-off between queueing and coding delays and propose a novel queue management technique for fading channels.**

***Index Terms*— Throughput, delay, capacity, block-fading, rate control, outage.**

I. Introduction

The fading channels seen in many wireless systems provide a particulary hostile environment for reliable communications [10]. The transmitted signal is scattered in a time-varying manner resulting in random fluctuations of the received power making reliable communications difficult. Transmitters typically employ channel coding techniques that map sequences of input data to *codewords* that add redundancy to combat the effects of fading and other noise prior to transmission. The *coding delay* is proportional the length of codewords that are used and is often quantified by the number of fading states that affect a transmitted codeword. The communications ability of a system depends greatly on the coding delay that can be tolerated. A system is *delay-unconstrained* if the coding delay is infinite and infinite-length codewords are used. On the other hand, systems are *delay-limited* if the coding delay is finite and use finite-length codewords – clearly this includes all practical systems in use today.

Often delay-limited systems must employ codeword retransmission to ensure reliable communications between transmitter and receiver. For example, the many variants of the ARQ protocol do this in practical systems [9]. Large queues can build in such systems and many applications, such as video and voice, cannot tolerate this. In these systems, parameters must be adjusted to ensure that the queueing delay is not excessive. In this paper we focus on adjusting the coding rate and the packet arrival rate to control the queueing behavior of systems that perform retransmission in fading channels.

This paper is organized as follows. Section II overviews relevant background information and notation. In Section III, we describe the throughput maximization with both coding and queuing delay constraints. Section IV provides simulation results. Finally, we conclude and provide directions for future work in Section V.

II. Background

Consider the discrete-time block-fading additive white Gaussian noise (BF-AWGN) channel model [11]. Each "block" corresponds to the channel *coherence time*, the amount of time the channel can be assumed constant, and the transmission of N coded symbols. The system in the k^{th} block can be written as

$$\underline{y}_k = \underline{x}_k \sqrt{\alpha_k} + \underline{w}_k, \tag{1}$$

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, $\sqrt{\alpha_k} \in \mathbb{R}$, on the transmitted signal. In this work, we assume that α_k 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.

A codeword that spans K blocks of the BF-AWGN channel is said to have a K block *coding delay*. Let,

$$\underline{\alpha} = \{\alpha_0, \alpha_1, \ldots, \alpha_{K-1}\} \tag{2}$$

denote the sequence of K i.i.d. fading states affecting the transmitted codeword. We assume that the receiver has knowledge of the channel condition, or channel state information (CSI). Finally, each of the KN symbols in the codeword contain information coded at R nats/sec/Hz (nat $:= \frac{\text{bit}}{\log_2(\text{e})}$).

The *instantaneous mutual information* [8]

$$C_K(\underline{\alpha}, \mathcal{P}_{\text{av}}) := \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \alpha_k \mathcal{P}_{\text{av}}) \tag{3}$$

represents the highest reliable data rate that can be achieved using a codeword transmitted with average power $\mathcal{P}_{\text{av}}$ that is affected by the channel $\underline{\alpha}$. Clearly since the α_k's are random variables, the instantaneous mutual information is a random quantity for finite K. When the transmission rate is higher than the instantaneous mutual information an *outage* event occurs

Supported in part by NSF, AFOSR, ONR, DARPA, and the Texas Instruments Leadership University Program. Email: [nahmed,richb]@rice.edu, Web: dsp.rice.edu.

0-7803-8622-1/04/$20.00 ©2004 IEEE

which results in a decoding error at the receiver [12]. The likelihood of such events

$$P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K) := \text{Prob}[R > C_K(\underline{\alpha}, \mathcal{P}_{\text{av}})] \tag{4}$$

is known as the *outage probability*.

In [1], [2], the throughput of delay-limited systems is maximized by optimally selecting the transmission rate. Codeword retransmission is used to reduce or eliminate outage. Using this model, the maximum throughput is given by

$$T_{\max} = \sup_{R} \frac{R}{\mathbb{E}[S(R)]} \tag{5}$$

where $\mathbb{E}[S(R)]$ is the average *service time*, or number of transmission attempts, per codeword. The maximum throughput for a delay-limited system using any retransmission scheme can be determined by computing $\mathbb{E}[S(R)]$ (for that particular retransmission scheme), substituting into (5) and performing the optimization.

Several retransmission schemes were considered in [1], [2]. In this paper we focus on the simplest of retransmission schemes, denoted RT. In this model, the receiver requests codeword retransmission in the face of outages using a delay-less and error-free feedback link. The transmitter retransmits data until the channel condition allows successful transmission. This closely resembles practical algorithms such as stop-and-wait ARQ (SW-ARQ) with zero-delay or selective-repeat ARQ (SR-ARQ) [9]. For scheme RT the service time distribution is geometric on the positive integers [1], [2]. Solving for the expected service time and substituting the form into (5) allows us to define the *maximum zero-outage throughput* (MZT) [1], [2]

$$\text{MZT}_{\text{RT}}(\mathcal{P}_{\text{av}}, K) = \sup_{R} R[1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)]. \tag{6}$$

MZT_{RT} represents the highest *average throughput* possible for retransmission scheme RT.

III. Throughput Maximization with Queueing Delay Constraints

With systems that perform codeword retransmission it is possible for queues to build at the transmitter. This occurs since the number of transmission attempts required for each codeword in a fading channel is random; codewords can arrive at the transmitter and be enqueued for transmission while another codeword is being (re)transmitted. MZT_{RT} is the maximum throughput for scheme RT irrespective of the queueing delay [1], [2]. Implicit in the formulation for MZT_{RT} is that the *system utilization* is one – that is the transmitter is always busy (re)transmitting data. As such, some codewords may experience excessive queueing delays [7]. Many systems in practice are sensitive to such delays and in these cases it is desirable to additionally control the queueing delay.

The *expected waiting-time*, or delay, is the amount of time that a codeword spends in the system (either in the queue or under service) [5], [6]. By constraining the waiting time we can constrain the average delay codewords experience. Accounting for queuing delay, we can define

$$\begin{aligned}&\text{MZT}^{\text{D}}_{\text{RT}}(\mathcal{P}_{\text{av}}, K)\\ &= \sup_{a,R}\{\rho(a, R)R[1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K)] : \mathbb{E}[W(a, R)] \leq D\}.\end{aligned} \tag{7}$$

as the maximum zero-outage throughput with average waiting time no greater than D. In this paper, D is measured as a multiple of the channel coherence time. The formulation in (7) is similar to (6) except for the scaling factor $\rho(a, R)$, the system utilization, that accounts for the proportion of time transmitter is in use,

$$\rho(a, R) := \frac{\text{average arrival rate}}{\text{average service rate}} = a\mathbb{E}[S(R)]. \tag{8}$$

For example if the codeword arrival rate and transmission rate are such that the throughput is 2 nats/sec/Hz but $\rho = 1/2$, implying that the transmitter is busy only half of the time, then the long-term average throughput is 1 nat/sec/Hz. Throughput is maximized by optimally selecting the coding rate R and the average arrival rate of codewords a. For the purposes of our simulations we measure a as the average number of codewords that arrive in a unit of time equal to the codeword length, or K blocks of the BF-AWGN channel. The formulation in (7) is rather general and applies to any arrival process. In effect, reducing both a and R decrease the amount of information that the source transmits across the fading channel. Reducing a, decreases the frequency of transmission of codewords containing RKN nats. On the other hand, reducing R decreases the amount of information transmitted per codeword, for a fixed arrival rate a, and ultimately the amount of information that the source transmits.

Often applications are affected by the total delay irrespective of whether the delay is spent in coding or queueing. The optimization can also be performed over the coding delay K and we define

$$\text{MZT}^{\text{D}}_{\text{RT}}(\mathcal{P}_{\text{av}}) = \sup_{K}\{\text{MZT}^{\text{D}}_{\text{RT}}(\mathcal{P}_{\text{av}}, K)\} \tag{9}$$

as the highest average throughput for average power $\mathcal{P}_{\text{av}}$ and average waiting-time less than D. For small K retransmissions are less costly in terms of delay but the instantaneous capacity, the amount of information that can be reliably transmitted with each codeword, is small. For large K the opposite is true, the instantaneous capacity is larger but retransmission is more costly in terms of delay. $\text{MZT}^{\text{D}}_{\text{RT}}(\mathcal{P}_{\text{av}})$ is the right balance that maximizes the average throughput with the total average waiting-time less than D.

IV. Simulation Results

In this section we illustrate the maximum throughput achievable under both a queuing and coding delay constraint via monte carlo simulation. For our analysis we assume that the channel fading states, α_k follow a χ^2_2 distribution, which implies that $\sqrt{\alpha_k}$ are Rayleigh distributed.

As previously mentioned the queuing behavior is dependent not only on the retransmission scheme, but on the nature of the arrival process. There has been much literature on modelling of traffic pattern generated by a variety of different sources.

In this paper we consider several representative models. To model non-bursty sources we consider the constant, bernoulli and poisson arrival processes. For the constant source, the interarrival period between codewords is deterministic; the time between codeword arrivals is always $\frac{1}{a}$. For the bernoulli source, the interarrival period is a random variable that is geometrically distributed on the positive integers with parameter a. For the poisson arrival process with intensity a, the interarrival period is an exponential random variable. To model bursty sources we consider a 2-state markov modulated poisson process (MMPP) with average intensity a. This model is commonly used in literature as it approximates the correlated nature of many real-world traffic models [13]. In this model, a state variable toggles between a high traffic and low traffic state. In each state, codewords are generated according to a poisson process with high and low intensities, respectively. For the purposes of our simulations we assume that the probability of being in the low and high intensity states are $p_{\text{low}} = 0.95$ and $p_{\text{high}} = 0.05$, respectively. Additionally, the average intensity between both states is a.

Figure 1 plots the optimal $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ as a function of the maximum average waiting time D, for $K = 1$ and $\mathcal{P}_{\text{av}} =$ 10dB. $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ was found by exhaustive search over the variables R and a. We see that $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ is smaller than $\text{MZT}_{\text{RT}}(\mathcal{P}_{\text{av}}, K)$. Restricting the throughput allows the queuing delay to be constrained by reducing system utilization. For each of the arrival processes we see that the $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ approaches $\text{MZT}_{\text{RT}}(\text{K}, \mathcal{P}_{\text{av}})$ as the constraint on the waiting time is relaxed, i.e. $D \rightarrow \infty$. This figure is particularly useful is it allows us to predict the best case performance of a communication system using retransmission scheme RT with both a finite coding delay K and a finite waiting-time D. It is also interesting to note that for small D, between 10 and 20 for the different arrival processes, $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ approaches $\text{MZT}_{\text{RT}}(\mathcal{P}_{\text{av}}, K)$. Additionally we see that for the bursty MMPP process that $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ increases slower as a function of delay than the non-bursty sources. This occurs since the average delay is more difficult to constrain for bursty sources; In order to keep the expected waiting time below a threshold, $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ must be limited to a greater degree than for non-bursty sources.

For $K = 1$ and $\mathcal{P}_{\text{av}} = 10\text{dB}$, the optimal coding rate $R^*_{\text{MZT}_{\text{RT}}^{\text{D}}}$ and codeword arrival rate $a^*_{\text{MZT}_{\text{RT}}^{\text{D}}}$ that maximize (7) are shown in Figures 2 and 3, respectively. For smaller D we see that the optimal coding rate is smaller than for larger D. Clearly, as the tolerable delay is decreased, the coding rate should be reduced. We also see that the optimal arrival rate does not fluctuate nearly as much. In fact for the non-bursty sources, the optimal arrival rate is nearly constant regardless of the waiting-time constraint. This tells us that for non-bursty sources, in order to maximize the throughput, while constraining the average waiting-time, the coding rate rather than the arrival rate should be reduced; the frequency of codeword arrivals should be left unchanged while the amount of information in each codeword should be decreased. This reduces the average waiting time since a smaller coding rate results in a smaller outage probability and a smaller $\mathbb{E}[S]$. For bursty sources, reducing the coding rate alone is not sufficient to meet the waiting-time constraint. We see for smaller D, that the arrival rate must also be reduced.

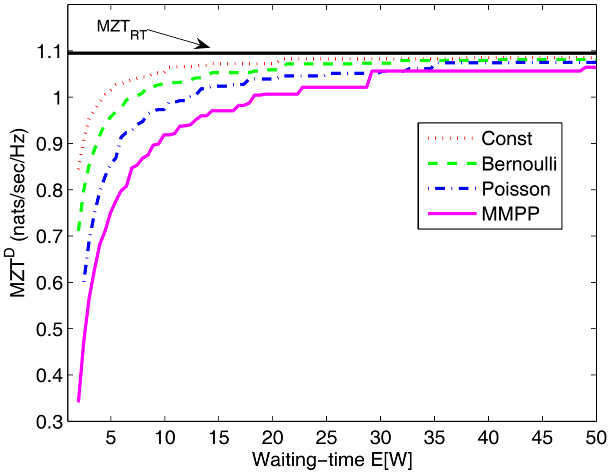

Fig. 1. The maximum throughput $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ is plotted as a function of D for $K = 1$ and $\mathcal{P}_{\text{av}} = 10$. The throughput for a finite coding delay of $K = 1$ and finite waiting-time D can be predicted.

For both bursty and non-bursty sources, a reduction in the codeword arrival rate reduces the throughput to a greater extent than a reduction in the coding rate. This is rather non-intuitive as conventional flow-control algorithms, such as TCP [9], reduce the frequency of packet generation when large queues build in communication networks [9]. The difference is reconciled by the fact that the underlying cause for the buildup of queues is different. The action that TCP takes is motivated by the assumption that queues build due to congestion in the network — that packets are being generated faster than the network can handle them. However, queues in fading channels grow due to the frequency of codeword generation and the fact that the medium itself is unreliable. For example, if the channel condition remains poor for 10 consecutive slots (resulting in outages on 10 consecutive transmission attempts) and zero new codewords arrive into the queue, then the queue size remains unchanged. However, if the link is assumed reliable and if zero codewords arrive into the queue for 10 consecutive slots, then the queue size shrinks by 10. This concept is quite intriguing as it allows for a novel method for waiting-time/delay (or queue-length) management in fading channels: If the average waiting-time is large then it can be reduced by using a smaller coding rate (codewords with a smaller amount of data) at the transmitter. Conversely, a larger coding rate (more information per codeword) can be used at the transmitter to increase communications throughput at the expense of a larger waiting-time.

The tradeoff between coding delay and queueing delay is illustrated in Figure 4, which plots $\text{MZT}_{\text{RT}}^{\text{D}}(\mathcal{P}_{\text{av}}, K)$ as a function of K for $D = 20$ and $\mathcal{P}_{\text{av}} = 10\text{dB}$. For this simulation the average waiting time is set to $D = 20$ for all K. That is, to the end user the average delay is the same for all K, however the throughput is not. By optimizing over

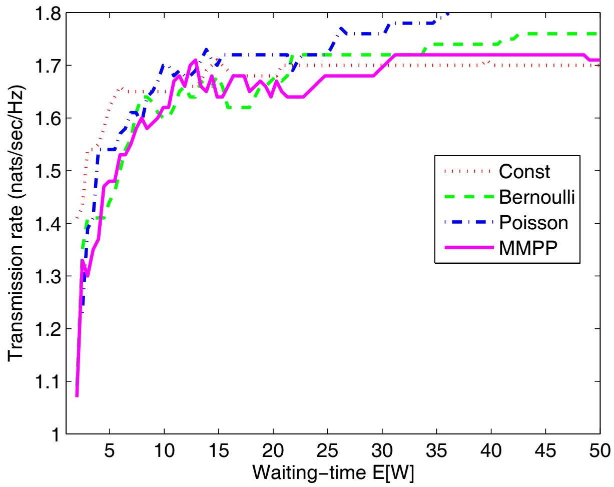

Fig. 2. Optimal transmission rate $R^*_{\mathrm{MZT}^{\mathrm{D}}_{\mathrm{RT}}}$ as function of the average waiting time D for $K = 1$ and $\mathcal{P}_{\mathrm{av}} = 10$. For small D, $R^*_{\mathrm{MZT}^{\mathrm{D}}_{\mathrm{RT}}}$ can be quite far from the asymptotic value, suggesting a reduction in the transmission rate to satisfy a waiting-time constraint.

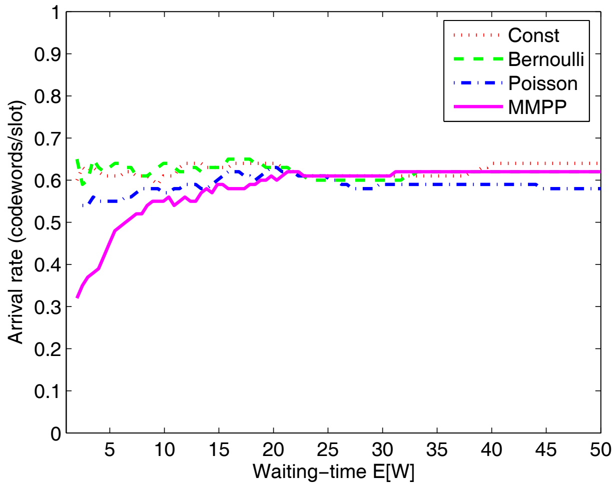

Fig. 3. Optimal arrival rate $a^*_{\mathrm{MZT}^{\mathrm{D}}_{\mathrm{RT}}}$ as a function of D for $K = 1$ and $\mathcal{P}_{\mathrm{av}} = 10$. We see that for non-bursty sources, that the optimal arrival rate remains relatively constaint. However, for the burstly MMPP source, the arrival rate drops for smaller D.

K, the throughput can be maximized without any effect on the average waiting-time. In this case we see that there is a unique coding delay, $K = 16$, that corresponds to $\mathrm{MZT}^{\mathrm{D}}_{\mathrm{RT}}(\mathcal{P}_{\mathrm{av}})$. This indicates that for a total waiting-time of $D = 20$ that the coding delay should be set to $K = 16$ and the codeword arrival and transmission rates found by solving (7). Also note that $\mathrm{MZT}^{\mathrm{D}}_{\mathrm{RT}}(\mathcal{P}_{\mathrm{av}}, K) = 0$ for $K = 20$ and $D = 20$. This occurs since in this situation only a single transmission attempt is permitted and non-zero throughput is not possible (delay-limited capacity is zero) [10].

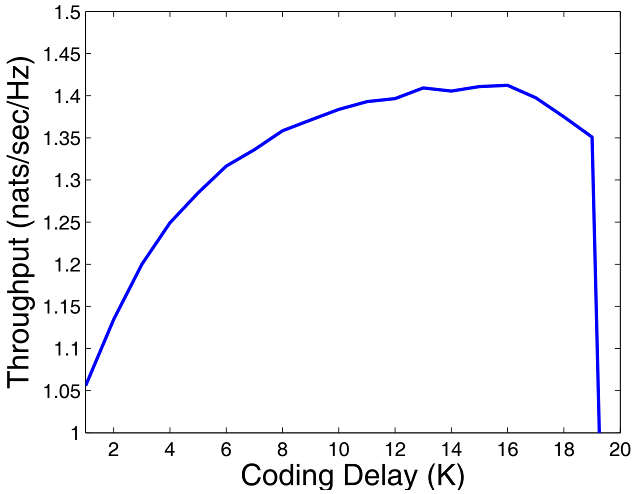

Fig. 4. $\mathrm{MZT}^{\mathrm{D}}_{\mathrm{RT}}(\mathcal{P}_{\mathrm{av}}, K)$ as a function of K for a waiting time of $D = 20$ and for $\mathcal{P}_{\mathrm{av}} = 10$dB. Clearly there is a tradeoff between coding delay and queueing delay. In this example a coding delay of $K = 16$ achieves maximum throughput for $D = 20$.

V. Conclusions

Our analysis has led to several important contributions. We developed a technique that maximizes the communications throughput and constrains both a coding and queuing delays. This has importance since all practical communication systems are coding delay constrained and many systems of interest, especially those transmitting multimedia traffic, additionally need to constrain queuing delay. Our formulation and analysis can predict the best-case throughput of practical retransmission protocols in fading channels. We also explore the trade-off between queuing and coding delays. For small K retransmissions are less costly in terms of delay but the instantaneous mutual information, the amount of information that can be reliably transmitted with each codeword, is small. Conversely, for large K the opposite is true.

Our analysis suggests a novel technique for queue management and a promising area for future research – the management of delay using rate control based on queue status. When queueing delays in practical systems become excessive, the queue-length status can be used to reduce the coding rate and hence the congestion in the system. This is counter to common congestion control technique since the nature of queuing in fading channels is different than in reliable wired links.

References

[1] N. Ahmed and R. G. Baraniuk, "Throughput Measures for Delay-Constrained Communications in Fading Channels," *in Proc. 41st Allerton Conf. Comm., Cont. and Comp. Oct 1-3, 2003*

[2] N. Ahmed and R. G. Baraniuk, "Delay-limited Throughput Maximization for Fading Channels Part I: Optimal Rate Selection," *In preparation for IEEE Trans. Info. Theory.*

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

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

[5] J. J. Hunter, “Mathematical Techniques of Applied Probability. Vol. 1. Discrete Time Models: Basic Theory,”*Academic Press, 1983.*

[6] J. J. Hunter, “Mathematical Techniques of Applied Probability. Vol. 1. Discrete Time Models: Techniques and Applications,”*Academic Press, 1983.*

[7] L. Kleinrock, “Queuing Systems - Volume I: Theory,”*New York: Wiley, 1975.*

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

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

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

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

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

[13] H. Heffes and D. Lucantoni,“ A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance,”*IEEE Jour. Sel. Areas in Comm., pp. 856-868, 1986.*